Machine Learning and Technostress as Important Aspects for Improving the Performance of Data Scientists in Contemporary Marketing Contexts

Dissertation

zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

der Rechts- und Wirtschaftswissenschaftlichen Fakultät

der Universität Bayreuth

Vorgelegt

von

Nicholas Daniel Derra

aus

Fürth

Dekan: Herr Prof. Dr. Jörg Schlüchtermann

Erstberichterstatter: Herr Prof. Dr. Daniel Baier

Zweitberichterstatter: Herr Prof. Dr. Claas Christian Germelmann

Tag der mündlichen Prüfung: 12. April 2021

Dedication

Für Corinna, Cleo und Mira,

die mir täglich das Lächeln auf mein Gesicht zaubern.

„Witzigkeit im Übermaß ist des Menschen größter Schatz."

- Rowena Ravenclaw

Acknowledgements

Eine Dissertation neben beruflichen Pflichten sowie denen eines Familienvaters verlangt neben der akademischen Weiterentwicklung insbesondere nach Durchhaltevermögen, Flexibilität, Organisationstalent, Souveränität und Stressresistenz. Dabei ist es unabdingbar, dass das jeweilige persönliche Umfeld dem Doktoranden in den schwierigen Phasen der Arbeit – insbesondere bei eventuellem Aufkommen von Rastlosigkeit, Frust und Verzweiflung – zur Seite steht. Ich selbst habe das große Glück, unglaublich viel Zuspruch erfahren zu haben, ohne den diese Arbeit nicht entstehen hätte können.

Ich danke meinem Betreuer im Rahmen dieser Dissertation, Herrn Prof. Dr. Daniel Baier, für die Begleitung und Unterstützung meines akademischen Werdegangs, welche bereits während meines Studiums begann und nun in meiner Dissertation ihren Höhepunkt findet. Niemand hat mich während meiner betriebswirtschaftlichen Ausbildung derart geprägt wie Sie, und es war mir stets eine Ehre, von Ihnen zu lernen. In diesem Zusammenhang danke ich außerdem meinem Zweitgutachter, Herrn Prof. Dr. Class Christian Germelmann, für dessen Bereitschaft zur Bewertung meiner Arbeit sowie dem Präsidenten meines Arbeitgebers BF/M-Bayreuth, Herrn Prof. Dr. Torsten M. Kühlmann, für die Übernahme des Prüfungsvorsitzes. Weiterhin gilt mein Dank meinen hervorragenden Ko-Autoren Prof. Dr. Daniel Baier, Prof. Dr. Torsten M. Kühlmann, Tobias Albrecht, Julia Becker, Theresa Rausch, Christian Regal und Lukas Wolf für die starke Zusammenarbeit bei der Erstellung der wissenschaftlichen Beiträge für diese Dissertation.

Der Grundstein für das spätere Leben wird insbesondere in der Kindheit und Jugend gelegt. Ich bin meiner Großmutter Carola sowie meinen Eltern Dagmar und Daniel ausgesprochen dankbar für die vielen Dinge, die sie mir beigebracht haben, für die Werte, die mir vermittelt wurden, und den Lebensweg, den sie mir ermöglicht haben. Ich danke außerdem meinen Freunden Bernd, Cornelia, Daniel, Kati, Klaus und Sandra für den vielen Zuspruch, die offenen Ohren in schwierigen Zeiten, die Ablenkung in der Freizeit und die Bereitstellung unzähliger Tassen Kaffee mit einem Schuss Milch.

Niemand jedoch musste mich häufiger und intensiver unterstützen (oder ertragen) als meine Familie. Ich bin meiner Ehefrau Corinna unendlich dankbar für die Freiheit, die ich bei der Erstellung dieser Arbeit genießen durfte. Du hast mir den Rücken freigehalten, standest hinter, neben und vor mir, wann immer ich Dich gebraucht habe – dafür danke ich Dir von ganzem Herzen ebenso wie unseren beiden Kindern Cleo und Mira. Ihr drei habt mir zu jeder Zeit und in jeder noch so herausfordernden Situation das Lachen und die Leichtigkeit erhalten. Ich liebe Euch, Ihr seid mein Zuhause.

Abstract

Based on recent developments caused by the big data revolution, data science has massively increased its importance for businesses. Within the marketing context, various types of customer data have become available in enormous amounts and need to be processed as efficiently as possible for creating valuable knowledge. Therefore, data scientists' performance has become crucial for marketing departments to achieve competitive advantages in the modern highly digitalized economy.

Within the raising field of data science, machine learning has become an outstanding trend since these approaches are able to automatically solve numerous classification and prediction problems with enormous performance. Thus, machine learning is seen as a key technology which will radically transform business practice in the future. Even though machine learning has already been applied to various marketing tasks, research is still at an early stage requiring further investigations of how marketing can successfully benefit from machine learning applications.

Besides these data-driven opportunities provided by digitalization, technostress has evolved into an enormous downside of digitalized workplaces, leading to a significant decrease in employees' performance. However, existing research lacks to provide evidence about different coping strategies and their potential to support employees in overcoming technostress. Furthermore, research currently fails to consider technostress regarding both highly digitalized occupational groups like data scientists and respective workplace environments for providing a deeper understanding of how employees suffer from stress caused by the use of digital technologies.

Due to these recent challenges for data scientists, this cumulative thesis provides useful insights and new opportunities by focusing on machine learning and technostress issues as two aspects which promise major potentials for enhancing data scientists' performance in today's marketing contexts. Five research papers are included for effectively tackling both fields of research: three papers deliver both methodological and empirical findings for extending machine learning in marketing research by examining model architectures as well as applying machine learning to recent marketing problems. In addition, two research papers contribute to research by providing knowledge about technostress issues of data scientists as a heterogeneous and highly digitalized occupational group as well as examining different coping strategies for effectively overcoming stress due to the use of digital technologies. Beyond that, the findings deliver practical implications for marketing managers who aim to improve the performance of data scientists in a contemporary marketing environment.

Table of Contents

I Introduction

1 Motivation

Due to its enormous economic value, relevant data has become the oil of businesses (van der Aalst, 2014). During the last decade, the big data revolution has provided numerous opportunities and challenges for applying data science to create valuable knowledge out of customer data (Erevelles et al., 2016; Lukosius & Hyman, 2019). Based on enormous accompanying developments regarding the availability, collectability, and storage of huge amounts of various data, nearly every department within a company has got new opportunities of developing improvements in decision making: various recent studies have already confirmed the importance of this data-driven decision making (see, e. g., Ferraris et al. (2019); Müller et al. (2018); Wamba et al. (2017)), showing that the application of data science for analysing big data increases the performance of organisations and, thus, builds competitive advantages. I. e., it is particularly important to perform data science instead of just storing the data as well as the contained information (Chen et al., 2012; Davenport, 2006). In this context, marketing has always been a popular application field of this data-driven decision making (Provost & Fawcett, 2013; Wedel & Kannan, 2016).

For efficiently meeting these big data developments of today's marketing environment and, therefore, creating competitive advantages caused by data-driven decision making, it is indispensable for companies to employ experts who are capable of fulfilling the numerous data science tasks concerning working with and creating knowledge out of data (Davenport & Patil, 2012; Erevelles et al., 2016; van der Aalst, 2014). Hence, the relevance of such employees – so-called data scientists – has exceedingly raised during the last decade due to the availability, capture, and storage of huge amounts of data due to the digital transformation and, thus, has led to a major demand for these employees (Davenport, 2020; Ismail & Abidin, 2016; Mauro et al., 2018; Murawski & Bick, 2017). Due to their massive importance for building competitive advantages out of data-driven decision making, it is crucial for companies to aim for improving the performance of data scientists. In this context, broad methodological as well as domain (e. g., marketing) knowledge for effectively solving data science problems is highly required (Ayankoya et al., 2014; Manieri et al., 2015; Waller & Fawcett, 2013). Therefore, data scientists need to constantly train their skills and competences by adopting recent trends and innovative technologies for further improving their task-specific performance and, consequently, building competitive advantages.

Within the wide field of data science research, machine learning has become the outstanding trend which has reached particular importance for gaining competitive advantages due to the developments of the big data revolution (Cui et al., 2006; Hazen et al., 2014; Ma & Sun, 2020; Saura, 2020). In the following years, machine learning will fundamentally transform core processes within nearly all companies' business practice (Brynjolfsson & Mcafee, 2017). In this context, it has been proven that marketing may also strongly benefit from machine learning applications as they represent the state of the art within marketing analytics (Hagen et al., 2020; Huang & Rust, 2018; Jordan & Mitchell, 2015; Rust, 2020; Wedel & Kannan, 2016). Therefore, it is highly important to understand how these models are composed for effectively tackling marketing tasks by applying machine learning models (Ma & Sun, 2020). However, machine learning within marketing research is still at an early stage, requiring further studies and enhancements in the future for constantly extending this promising area of research (Chintagunta et al., 2016a; Chintagunta et al., 2016b; Dimetreska et al., 2018; Ma & Sun, 2020; Saura, 2020).

Besides the opportunities and changes offered by the digital transformation and, in particular, the big data revolution, these developments may also enormously demand employees (Okkonen et al., 2019; Schwemmle & Wedde, 2012; Timonen & Vuori, 2018). In this context, a massive psychological dark side of digitalization has been risen next to its advantages, affecting both productivity and well-being of employees: using information and communication technologies (ICT) at work causes technostress which represents a specific form of stress induced by the frequent use of digital technologies at work (Ayyagari et al., 2011; Ragu-Nathan et al., 2008; Tarafdar et al., 2007; Tarafdar et al., 2010). Technostress has become of particular relevance due to the rapid implementation of countless ICT during the last two decades (Hartl, 2019; Osmundsen et al., 2018), leading to the consideration of the digital transformation at work as a double-edged sword (Apt et al., 2016). For overcoming technostress at work, employees require suitable resources like, e. g., organisational factors (Ragu-Nathan et al., 2008), environmental aspects (Galluch et al., 2015), and inhibitors at employee's level (Srivastava et al., 2015; Sumiyana & Sriwidharmanely, 2020). However, the application of various coping strategies which may be actively and autonomously implemented by the employee is inadequately examined in technostress research and, therefore, requires further investigations (Pirkkalainen et al., 2019; Tarafdar et al., 2019).

Moreover, research currently fails to consider technostress in the context of specific occupational groups but focuses on general relationships between technostress constructs instead (Ayyagari et al., 2011; Fischer & Riedl, 2020; Ragu-Nathan et al., 2008; Tarafdar et al., 2007; Tarafdar et al., 2010;

Tarafdar et al., 2011; Tarafdar et al., 2015). Since the investigation of stress within separate occupational groups in order to create knowledge regarding their specific regularities is widely established in psychological research (see, e. g., Grace & van Heuvelen (2019); Rees & Cooper (1992); Travers & Cooper (1993)), the examination of technostress in the context of data scientists as a highly digitalized job group appears to be necessary for supporting data scientists in overcoming technostress and, thus, improving their performance.

Due to these necessities and recommendations for future research, this doctoral thesis aims to provide useful knowledge for further improving the performance of data scientists in modern marketing contexts. To achieve this goal, the focus is on both machine learning applications to marketing problems and employees' technostress issues as these topics have been proven to act as crucial aspects for creating competitive advantages in today's digitalized business world. Based on this general goal, the prevailing research questions are proposed as follows:

> *RQ1: How can data scientists improve their performance by successfully applying machine learning algorithms in contemporary marketing contexts?*
>
> *RQ2: How can data scientists improve their performance by effectively overcoming technostress at work?*

As provided in Figure 1, this thesis contains five research papers which tackle either RQ1 regarding the topic of machine learning applications (research papers #1 - #3) or RQ2 regarding technostress issues (research papers #4 - #5) as important aspects of data scientists' marketing performance. These papers have already been published or are currently under review within sophisticated academic journals.

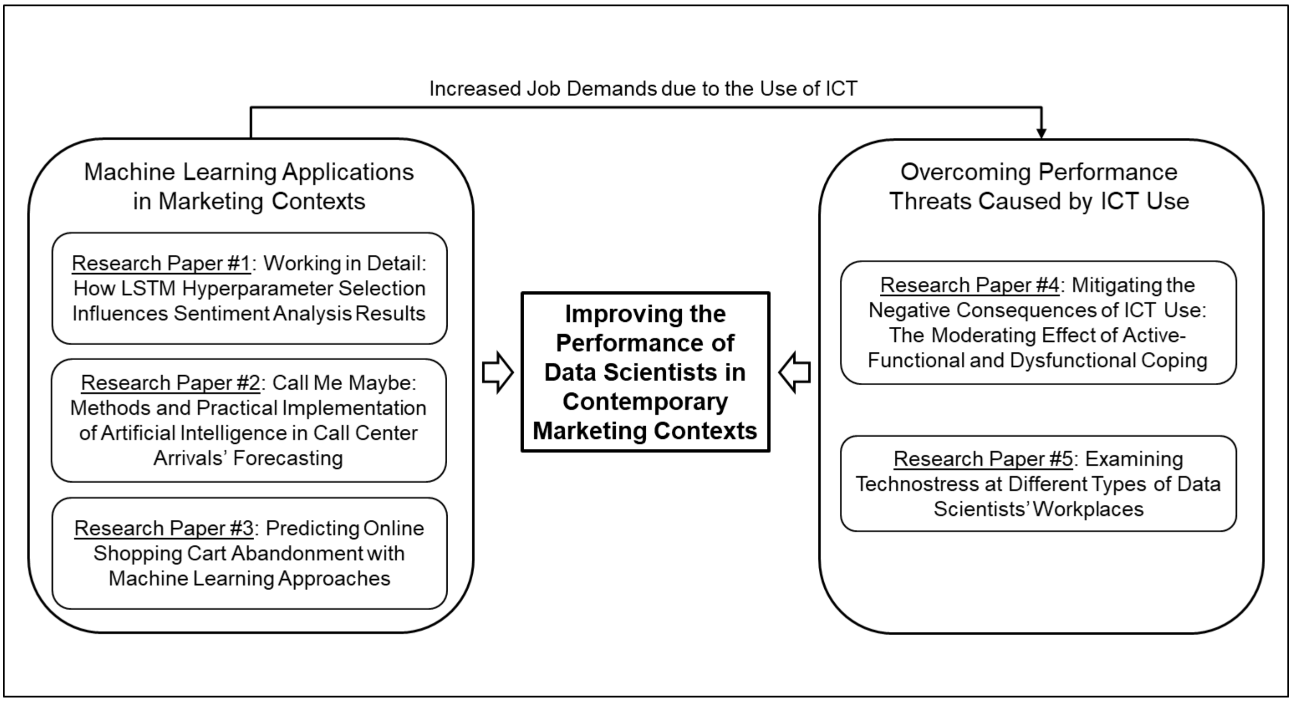

Figure 1. Contextual Framework of this Thesis

In the context of machine learning applications, this thesis delivers new opportunities of applying supervised machine learning models to marketing and, further, compares various algorithms regarding their performance at solving a specific task. Moreover, a deeper understanding of how these algorithms may be successfully compiled is offered so marketers are able to receive important knowledge for creating models which achieve high task-specific performance in order to enhance return on investment.

In addition, this thesis also provides insights into technostress as an enormous downside of digitalization data scientists have to struggle with. By that, technostress knowledge regarding data scientists as a specific occupational group as well as the examination of different coping strategies to successfully overcome technostress is to be examined for enabling data scientists to overcome performance threats caused by ICT use.

To achieve this, this thesis is structured as follows: in chapter I, the theoretical background of data science and machine learning in today's marketing, data scientists as a highly digitalized occupational group as well as technostress research and, further, the research agenda including a detailed overview of the included research papers is outlined. These research papers are then provided in the following chapters II to VI within this thesis. Finally, chapter VII provides a summarizing conclusion of the compiled findings.

2 Theoretical Background

2.1 Data Science in Contemporary Marketing

Within the marketing context, the systematic utilization of quantitative data has an impressive history of more than 100 years (Wedel & Kannan, 2016). Within this bright history, the founding of the Marketing Science Institute by the initiative of the Ford Foundation and the Harvard Institute of Basic Mathematics for Applications in Business in 1961 is seen as the major impact for successfully applying analytics to marketing issues (Winer & Neslin, 2014). Since then, the field of data science has been widely used for extending marketing research (Wedel & Kannan, 2016).

In modern business environments, both the opportunities and challenges for applying data science to create valuable knowledge out of customer data have been massively raised due to the big data revolution (Erevelles et al., 2016; Lukosius & Hyman, 2019). Overall, big data is defined as huge datasets containing structured and/or unstructured data that can be processed and analysed for creating knowledge such as patterns and trends out of it (Hazen et al., 2014). In this context, the big data revolution is differing from conventional data collection by several characteristics called the *three Vs*: volume, i. e., huge amounts of available data; velocity, i. e., rapid processes of data creation in real-time; and variety, i. e., the creation of numerous types of unstructured data (Chintagunta et al., 2016a; Erevelles et al., 2016; Lycett, 2013). Furthermore, the collection and analysis of big data is also associated with two other characteristics called veracity and value (Lycett, 2013; Wedel & Kannan, 2016): while veracity is described as the importance of considering the quality of collected data regarding reliability and validity (IBM, 2012; Wedel & Kannan, 2016), value represents the focus on data which is valuable for gaining domain-specific knowledge (Lycett, 2013).

In the context of marketing, the big data revolution has transformed consumers into permanent generators of both traditional, structured, and transactional data as well as more contemporary, unstructured, and behavioural data leading to a transformation of marketing decision making (Erevelles et al., 2016). Digital data which is collected through online and mobile applications provides valuable insights on consumers' feelings, behaviours, and interactions around products, services, and marketing actions (Wedel & Kannan, 2016). The analysis of such data enables marketers to gain knowledge out of complex and dynamic data of consumers' behaviour and markets (Chintagunta et al., 2016a): while surveys and experiments may enable rapid and diverse data collection as well, big data mostly exhibits observational characteristics (Ma & Sun, 2020; Wedel & Kannan, 2016).

Due to these developments, companies aim for processing the collected data in order to create valuable insights (Provost & Fawcett, 2013). In this context, research has already proven the success of data-driven decision making by showing that applying data science to big data – so-called big data analytics – increases the performance of organisations (Ferraris et al., 2019; Müller et al., 2018; Wamba et al., 2017). Consequently, the conduct of data analysis instead of just storing the data and its contained information is of special relevance for building competitive advantages (Chen et al., 2012; Davenport, 2006). Therefore, the field of data science is closely related to big data, both massively increasing in popularity within both research and business practice (Waller & Fawcett, 2013).

Generally, data science represents the application of quantitative and qualitative methods to extract valuable information for solving relevant problems and predicting outcomes (Waller & Fawcett, 2013). In doing so, the term data analytics is used interchangeably (Agarwal & Dhar, 2014). Data science utilizes numerous data mining techniques which perform the extraction of knowledge from data, aiming for the overarching goal of improving the quality of businesses' decision making (Provost & Fawcett, 2013). For performing high-quality data science, very broad domain knowledge, e. g., for solving marketing problems, is mandatory as well (Ayankoya et al., 2014; Manieri et al., 2015; Waller & Fawcett, 2013).

Since big data is massively changing marketing processes, many of the methods developed by marketing academics in the past support today's decision making in customer relationship management, marketing mix, and personalization leading to an increased financial performance (Wedel & Kannan, 2016). The application of data science methods on big data has become crucial for decision making in marketing (Amado et al., 2018), realising that big data is only able to offer valuable insights if it is efficiently analysed. Thus, bringing together data science and marketing research has evolved an essential interdisciplinary field within marketing analytics, using a broad set of methods for measuring, analysing, predicting, and managing marketing performance in order to maximise effectiveness and return on investment (Wedel & Kannan, 2016).

The usage of knowledge extracted out of big data for marketing decision making also helps marketing managers to receive credibility within companies (Rogers & Sexton, 2012): marketers may take advantage of collected big data in various ways, e. g., for interaction with customers via chatbots (Luo et al., 2019), for product and service personalization (Anshari et al., 2019), and automatic implementation of real-time marketing actions like online advertising (Jabbar et al., 2020) in order to increase perceived customer value, satisfaction, and loyalty which leads to higher success of these marketing actions (Wedel & Kannan, 2016). Furthermore, data science has been broadly applied for performing targeted marketing, online advertising, customer relationship management, and cross-selling recommendations (Provost & Fawcett, 2013). To achieve this, big data offers many different types of data including clickstream, social media, video, image, text, and location data as sources of useful knowledge (Ma & Sun, 2020; Wedel & Kannan, 2016). In this context, direct marketing has particularly gained benefits out of data science, i. e., in terms of collecting, analysing, and interpreting data (Palacios-Marqués et al., 2016; Provost & Fawcett, 2013; Tiago & Veríssimo, 2014).

Consequently, marketing research deals with the benefits of analysing these kinds of data via data science approaches aiming to provide useful knowledge out of it, i. e., online reviews for identifying customers' suggestions for improvements and, thus, increasing product and service quality (Qi et al., 2016), social media data for evaluating brand equity and competitive positions (Godey et al., 2016), mobile retail data for better recommendations and personalized offerings (Portugal et al., 2018), GPS data for geo-targeting customers with contextual promotions (Banerjee et al., 2013), keyword search for improving the design of companies' websites and advertising (Ghose & Yang, 2009), and clickstream data for recognizing segments of customers (Schellong et al., 2017).

Due to the opportunities provided by the big data revolution, marketing research constantly moves away from conventional approaches and focuses on dynamic and analytical decision making (Li et al., 2018). More specifically, the availability of big data has enormously increased interest in the empirical-then-theoretical approach which aims to develop marketing theory based on observed empirical findings. In this context, modern marketers require advanced analytical skills for handling big data, i. e., data mining tools, cognitive computing, and machine learning approaches (Lukosius & Hyman, 2019). Consequently, future marketing research needs to extend the application of data science and, in particular, machine learning approaches on various types of data for gaining new competitive advantages by further improving marketing decision making in modern digitalized environments (Chintagunta et al., 2016a; Chintagunta et al., 2016b).

2.2 Machine Learning in Marketing

Basically, machine learning represents a subgroup within the artificial intelligence paradigm (Goodfellow et al., 2017) which is described as programming a digital computer for acting comparable to humans and animals who apply the process of learning (Samuel, 1959). Within machine learning, the concept of learning represents the automatic search for more suitable representations of input data with respect to a given task (Chollet & Allaire, 2018). I. e., such algorithms improve their performance in solving a specific (marketing) problem by collecting relevant experience out of other examples and, therefore, are rather trained than programmed.

Machine learning models may be distinguished between supervised, unsupervised, and reinforcement learning approaches (Jordan & Mitchell, 2015; Ma & Sun, 2020; Stinis, 2019). Within supervised learning, the algorithm is trained via labelled training data, i. e., the training examples contain both input values and the accompanying output value. The supervised model defines a classifier or predictor function which denotes the output based on the given input by processing the given training data. During training, the model is optimised by processing a validation set after each iteration (Ma & Sun, 2020). After the training section is finished, the model can classify unknown data based on the pattern information detected during the learning process. The most popular supervised machine learning approaches comprise decision trees (Breiman et al., 1984), support vector machines (Cortes & Vapnik, 1995), naïve bayes (Duda et al., 1973), k-nearest neighbour (Cover & Hart, 1967), and artificial neural networks (Jain et al., 1996), which have been further developed into numerous high-performing variants, e. g., tree-based ensemble learning methods (Opitz & Maclin, 1999; Rokach, 2010), convolutional neural networks (LeCun et al., 1989), and long short term memory neural networks (Hochreiter S. & Schmidhuber, 1997). In the marketing context, important supervised learning problems comprise natural language processing tasks like, e. g., sentiment classification of online texts (Dhaoui et al., 2017), customer churn prediction (Vafeiadis et al., 2015), and customer loyalty evaluation (Ansari & Riasi, 2016).

For performing unsupervised learning, the training data is unlabelled and does not contain any output variables. The algorithm aims to detect useful features and patterns which have not been identified yet (Dimitrieska et al., 2018; Ma & Sun, 2020; Saura, 2020). Unsupervised machine learning models are, inter alia, clustering algorithms (Xu & Wunsch, 2005) and topic models like latent dirichlet allocation (Blei et al., 2003), which are already well-established in marketing research (Ma & Sun, 2020).

Recent developments within unsupervised learning particularly deal with unsupervised artificial neural network architectures such as deep autoencoders (Vincent et al., 2010) and deep belief networks (Hinton et al., 2006). Typical unsupervised marketing issues constitute customer segmentation (Tsai et al., 2015) or discovering topics in online communities (Reisenbichler & Reutterer, 2019).

Finally, reinforcement learning represents a class of algorithms where the model aims to optimize a learning function which is connected to its environment (Jordan & Mitchell, 2015; Kaelbling et al., 1996). The model (or agent) is performing a reaction to a given input and, thereby, changes the current state of the environment. This change is announced to the agent as a feedback signal indicating whether the action impacts the state positively or negatively. The agent is then aiming to increase the long-term sum of these feedbacks by systematic trial and error. In this context, the main distinction to supervised learning is that the model is told the new current state, but not which action would have been the best choice for enhancing it (Kaelbling et al., 1996). Reinforcement learning problems are usually implemented for control-theoretic settings where the agent learns a control strategy for acting in an unknown dynamical environment (Jordan & Mitchell, 2015). Reinforcement learning has raised relevance due to the successful implementation within artificial neural networks which are able to process large amounts of input data and, subsequently, discover complex relationships between actions and environments (Bruyn et al., 2020). However, even though reinforcement learning enhanced relevance within overall business practice (Ma & Sun, 2020), it merely plays a minor role in marketing contexts due to the popularity of supervised learning approaches (Bruyn et al., 2020).

Based on the existence and availability of big data within online marketing contexts, machine learning applications in marketing research particularly address digital marketing (Saura, 2020). More specifically, machine learning approaches are particularly suitable within e-commerce marketing since it has been proven to be both easy and cheap to collect online customer behaviour data in such an automated environment (Kohavi & Provost, 2001). In this context, it is highly important to perform classification and prediction in real-time since the Internet has been shown to be a very fast-paced environment (Jabbar et al., 2020). Due to the automatised nature of machine learning, these algorithms can perform such real-time reactions and, hence, are capable of influencing customer behaviour.

Even though marketing research has already dealt with machine learning models in great detail, the rapid developments within the digital revolution and, in particular, both the infinity of countless types of customer data as well as the possibility of creating new algorithms or improving existing models lead to a high necessity of constantly expanding this area of marketing research. Therefore, further research regarding innovative and successful machine learning approaches as well as new marketing applications is highly recommended for creating competitive advantages out of companies' marketing activities (Ma & Sun, 2020; Saura, 2020). Overall, the utilization of machine learning in the marketing context is still at an early stage which will strongly enhance in the future (Dimitrieska et al., 2018; Ma & Sun, 2020). Therefore, further studies which successfully apply machine learning to new marketing issues and, particularly, shed light on practical implementations of such models are highly important for enhancing modern marketing research and practice (Chintagunta et al., 2016a; Chintagunta et al., 2016b). In this context, the excellent performance of supervised learning approaches in complex marketing tasks particularly strikes which in turn implies focusing on innovative supervised models.

2.3 The Data Scientist

From a global perspective, a data scientist may be described as an expert who extracts knowledge from collected data as well as manages both the whole data lifecycle and relevant IT infrastructures (Manieri et al., 2015). However, research has proven that the occupational group of data scientists appears to be very heterogeneous in the context of required skills and tasks (Davenport, 2020; Ismail & Abidin, 2016; Mauro et al., 2018) and, therefore, has to be considered in more detail. In this context, research has already defined job profiles (Costa & Santos, 2017) and educational curricula (Richards & Marrone, 2014), or collected information from experts (Mikalef et al., 2018; Stanton & Stanton, 2016) to identify a data scientist's required skills and occupational roles.

Regarding the job-related skill variety as proposed by Hackman & Oldham (1976), data scientists require a wide field of both hard and soft skills, i. e., specific knowledge due to the use of numerous ICT as well as advanced skills in mathematics, statistics, machine learning, and communication skills (Costa & Santos, 2017; Doyle, 2019; Ismail & Abidin, 2016; Richards & Marrone, 2014).

Besides this variety of skills, data scientists also exhibit heterogeneous work profiles which occur due to the various application fields, structures within the respective company, and various data science objectives: several studies have pointed out different occupational profiles associated with 'data scientist' as the generic term, e. g., business analysts, data engineers, statisticians, and data analysts (Baškarada & Koronios, 2017; Ho et al., 2019; Mauro et al., 2018). These job titles occur due to the separate process stages of the data lifecycle the respective employees are then working at.

Considering this variety of skills, roles, and tasks within the occupational group of data scientists, both business practice and research stated that it seems to be unrealistic to find employees fulfilling all the required demands and, hence, created the term "Unicorn Data Scientist" for such experts (Baškarada & Koronios, 2017; Davenport, 2020; Davenport & Patil, 2012). Therefore, defining a data scientist as an overall expert who extracts knowledge from collected data as well as manages the whole data lifecycle and relevant IT infrastructures as proposed by Manieri et al. (2015) appears inappropriate.

Furthermore, the tasks of the data lifecycle which aim to create knowledge out of collected data are fulfilled by several employees working in various affiliations due to the presence of huge amounts of data in many departments within a company (Janssen et al., 2017) and, moreover, the necessity of advanced domain knowledge for performing data science (Ayankoya et al., 2014; Manieri et al., 2015; Waller & Fawcett, 2013). These employees do not work as full-time data scientists but, at the same time, require data science skills for answering specific questions. However, such workers who fulfil analytical work tasks of data scientists are often not classified as one but keep other job titles which are closely related to their respective department. This wide spreading of employees who perform data science within companies leads to difficulties in detecting these employees within a company: due to the given heterogeneity of skills, roles, and tasks, they can neither be detected by job titles nor department affiliations.

Overall, managers need to be able to detect data scientists within the company for significantly enhancing their performance. However, research currently lacks to provide a more practically based definition of data scientists as an occupational group because the focus is on both a universal but unrealistic definition as well as numerous job titles around different tasks within the data lifecycle. Furthermore, since various employees within different departments of a given company fulfil data science tasks by holding other occupational names, a title-based definition appears to be inappropriate for detecting them. Consequently, since employees who frequently fulfil data scientists’ tasks appear to be a crucial source for creating competitive advantages and, at the same time, detecting them is an indispensable prerequisite for improving their job performance, a definition with a strong reference to reality appears to be necessary.

2.4 Technostress

As already pointed out, the digital transformation and the big data revolution offer enormous opportunities and chances for improving the performance of businesses. However, the rapid velocity of these developments enormously demands employees to adopt new capabilities for efficiently handling work tasks as well (Okkonen et al., 2019; Schwemmle & Wedde, 2012; Timonen & Vuori, 2018), resulting in a massive psychological dark side of digitalization: using ICT at work causes technostress as a specific form of stress induced by the frequent use of digital technologies at work which affects both productivity and well-being of employees (Ayyagari et al., 2011; Ragu-Nathan et al., 2008; Tarafdar et al., 2007; Tarafdar et al., 2010). Conceptually introduced as employees' inability to handle the use of digital technologies in a healthy way by Brod (1984), technostress became of particular importance due to the rapid implementation of numerous ICT (Hartl, 2019; Osmundsen et al., 2018), leading to an ambivalence of digital transformation at work (Apt et al., 2016).

Overall, technostress is induced if employees perceive an inability to successfully establish numerous requirements and trends regarding digital technologies. Such feelings may occur with regard to, e. g., skills which are no longer required, an information overload, frequent interruptions during tasks at work, or the overlap of work and leisure time (Tarafdar et al., 2010). In this context, technostress is triggered by several specific stimuli called technostress creators which have been defined by Tarafdar et al. (2007) as follows:

- *Techno-uncertainty* – employees' confusion caused by new technological developments at work.
- *Techno-insecurity* – the fear of being replaced by either other employees with higher ICT affinity or by a digital technology itself.
- *Techno-overload* – requirements to work faster, longer, and more which are induced by ICT.
- *Techno-invasion* – blurring boundaries between work and leisure matters or time periods.
- *Techno-complexity* – employees` feelings of missing skills regarding ICT use at work.

Besides this well-established distinction, technical problems like system failures during ICT use represented by techno-unreliability (Riedl et al., 2012) and workflow disruptions due to ICT usage described by techno-interruptions (Galluch et al., 2015) have been classified as additional stressors due to the use of ICT as well.

If an employee’s perceptions of these technostress creators go beyond given personal and job-related resources, the upcoming technostress leads to technostress-related strains which represent individual's psychological, physical, or behavioural responses to technostress creators (Atanasoff & Venable, 2017). Examples for such strains are, inter alia, mental exhaustion (Ayyagari et al., 2011; Srivastava et al., 2015) and psychological detachment (Barber et al., 2019; Santuzzi & Barber, 2018). Technostress is also related to negative job-related consequences for employees, e. g., lower productivity at work (Tarafdar et al., 2007; Tarafdar et al., 2015), less job satisfaction and loyalty to the employer (Tarafdar et al., 2011) as well as serious health issues like higher burnout rates (Srivastava et al., 2015).

For reducing technostress and its negative consequences, it is necessary to have access to resources which may inhibit the negative effects of occurring technostress creators (Pirkkalainen et al., 2019; Tarafdar et al., 2011; Tarafdar et al., 2019). In this context, several organisational technostress inhibitors have been discovered, i. e., providing technical support, literacy facilitation, and involvement facilitation (Ragu-Nathan et al., 2008). Furthermore, other factors have been proven as successfully stemming technostress, e. g., timing and method control (Galluch et al., 2015) at environmental level and technology self-efficacy (Tarafdar et al., 2015) as well as personality traits (Srivastava et al., 2015; Sumiyana & Sriwidharmanely, 2020) at the employee’s level. In contrast, the adaption of different ways of coping which are of particular importance in overcoming stress due to individuals’ abilities to implement such strategies on their own are insufficiently investigated in the technostress context (Tarafdar et al., 2019).

Coping strategies are generally defined as cognitive and behavioural attempts which aim to manage specific external or internal demands which are perceived as challenging an individual’s resources (Lazarus & Folkman, 1984). Coping strategies are often distinguished in different types, e. g., problem-focused and emotion-focused coping (Folkman et al., 1986), functional and dysfunctional coping (Erschens et al., 2018), proactive and reactive coping (Pirkkalainen et al., 2019), or, in more detail, up to 14 different ways to overcome stress (Carver, 1997). Nevertheless, there is not a clear consensus considering the role of coping: while information systems research has followed the transactional theory of stress (Lazarus & Folkman, 1984) for a long time and, therefore, considered coping as a mediator (see, e. g., Gaudioso et al. (2016); Hauk et al. (2019); Zhao et al. (2020)), a few recent information systems studies (Nisafani et al., 2020; Pirkkalainen et al., 2019) as well as studies from industrial and organisational psychology (Lewin & Sager, 2009; Searle & Lee, 2015; Yip et al., 2008) assume coping as moderating the relationship between job-related stressors and strains. At the same time, coping strategies in technostress contexts are highly under-studied and need further interdisciplinary investigation (Pirkkalainen et al., 2019; Tarafdar et al., 2019).

Besides these general issues regarding technostress, related research also lacks to create a deeper connection between technostress and specific job groups: prior studies primarily focus on general relationships between technostress constructs (Ayyagari et al., 2011; Fischer & Riedl, 2020; Ragu-Nathan et al., 2008; Tarafdar et al., 2007; Tarafdar et al., 2010; Tarafdar et al., 2011; Tarafdar et al., 2015) but, at the same time, do not consider specific job titles in order to get a more individual understanding of employees' technostress and further to examine whether there is a need to define different strategies to overcome technostress even within a job class. While various psychological studies investigate stress within occupational groups in order to gain a deeper understanding of their respective specificities (see, e. g., Grace & van Heuvelen (2019); Rees & Cooper (1992); Travers & Cooper (1993)) and, further, examine relationships between several workplace attributes and work stress (i. e., customer contact (Hartline & Ferrell, 1996), leadership function (Ganster, 2005; Hambrick et al., 2005), and educational background (Golubic et al., 2009) as job-related characteristics and company size (Dekker & Barling, 1995; van Dijkhuizen & Reiche, 1980) as well as different dimensions of organisational culture within enterprises (Lansisalmi et al., 2000; Thompson et al., 1996) as company-related characteristics), current technostress research fails to offer job-specific findings. However, such investigations are of prominent relevance regarding job categories with a high level of digitalization at work since technostress and ICT use are closely related. In this context, data scientists are both particularly suitable and important for examining technostress due to their highly digitalized workplaces and their crucial role in gaining competitive advantages for companies.

Overall, both employees and employers are highly recommended to pay high attention to technostress issues and, moreover, to aim to reduce technostress. Consequently, further interdisciplinary as well as context-related technostress research is highly required.

3 Research Agenda

Considering the circumstances explained above, there is a great importance for marketing business practice to improve the performance of data scientists in marketing contexts leading to a high recommendation of further research within this topic. On the one hand, the countless opportunities for more purposeful and personalized marketing activities provided by machine learning and, specifically, supervised learning approaches are of enormous importance for marketers who aim to extract knowledge out of various customer data and, subsequently, use this information in order to increase the performance of marketing activities. On the other hand, the danger of increased technostress caused by the rapid developments of digitalization that employees have to deal with has to be closely observed as well. Considering the particular importance of people who work as data scientists due to their crucial role in data-driven decision making, employers are highly recommended to avoid high levels of technostress within this highly digitalized occupational group. However, research still lacks both job-specific and coping-related investigations regarding negative consequences of ICT use. For effectively meeting these issues and, consequently, solving the research questions provided in this thesis' motivation, five research papers are included in the following chapters II to VI. In doing so, research papers #1 - #3 meet RQ1 regarding machine learning applications in marketing and, further, research papers #4 - #5 tackle RQ2 by considering technostress issues as important aspects for improving the performance of data scientists in contemporary marketing contexts.

Research paper #1 meets RQ1 by investigating the potentials of deep neural networks (long short term memory networks, specifically) in the context of sentiment analysis tasks. By precisely performing the sentiment analysis task of the widely utilized IMDB large movie dataset (Maas et al., 2011), the paper provides an examination of 8 hyperparameters within the model and how these hyperparameters influence network performance. The hyperparameters were separately varied within their characteristic values for investigating the influence of the respective hyperparameter on the overall network performance. While 5 hyperparameters have been shown to increase classification accuracy, 3 other variants surprisingly lowered the network performance. Furthermore, the improvements could not be cumulated within the network which leads to the assumption of various interaction effects between the hyperparameters. Hence, research paper #1 contributes to the deeper understanding of the functioning within machine learning applications for automatically analysing online reviews.

Since the expansion of machine learning for improving data-driven decision making in marketing is highly recommended, research paper #2 and research paper #3 both address RQ1 by focusing on new applications of machine learning within marketing tasks. At first, research paper #2 successfully implements various practically relevant machine learning models for automatically predicting call centre arrivals and compares these approaches with conventional time series models regarding prediction accuracy. For doing this, the models were trained with two call centre datasets provided by a German online retailer containing half-hourly time series samples of 174.5 weeks, i. e., 31,410 observations each. For comparing these models, four different lead times were implemented as well as cross-validation with an expanding rolling window which constitutes an iterative process where the training data is rolled forward during model training. Results show that machine learning algorithms may outperform traditional models with a random forest approach delivering the strongest performance. Furthermore, this paper enhances the practical implementations of machine learning by providing a methodological walk-through encoding of the comparison process.

In contrast, research paper #3 focuses on the important e-commerce problem of online shopping cart abandonment by utilizing different machine learning algorithms for automatically predicting such abandoners based on their clickstream behaviour. With a sample of 821,048 aggregated clickstreams, numerous machine learning approaches were trained and compared with standard logistic regression as a conventional benchmark model regarding predictive performance and practicability. In doing so, the paper provides a deep methodological contribution on successfully applying machine learning to online shopping cart abandonment, proving that machine learning approaches are able to deliver stronger prediction accuracy as classic models. Within the implemented approaches, gradient boosting with regularization yielded the best results for unknown test data but, at the same time, a decision tree approach as well as boosted logistic regression provided comparable accuracy with clearly less model complexity. Hence, these methods have proven to be interesting alternatives due to their successful trade-off between performance and practicability.

For considerably contributing to RQ2, research papers #4 and #5 both focus on technostress issues as the striking downside of digitalization at work. Research paper #4 meets the appeal for further interdisciplinary technostress research regarding the role of different coping strategies for overcoming technostress at work (Pirkkalainen et al., 2019; Tarafdar et al., 2019): based on a sample of 3,362 German knowledge workers collected by an external panel during a larger technostress research project, a moderated mediation model via covariance-based structural equation modelling was developed for investigating the effectiveness of two reactive coping strategies (active-functional and dysfunctional) as moderating the relationship between stressors due to the use of ICT and employees' exhaustion, with exhaustion mediating the influence of technology-related stressors on productivity.

Thereby, this paper brings together psychological and information systems research by applying the job-demands resources model (Demerouti et al., 2001) to technostress research, conceptualizing coping as a personal resource. The results provided valuable findings, showing that a higher level of technostress-related job demands is associated with higher levels of both exhaustion and productivity, proving that employees should be demanded by ICT use at a medium level. Furthermore, while active-functional coping is associated with less exhaustion and, in contrast, dysfunctional coping is related to a higher level of it, both coping strategies have been found to buffer the effects of technostress on exhaustion contradicting prior results regarding the effects of dysfunctional coping. This means that, besides the negative consequences in long-term, dysfunctional coping like drinking alcohol or refusing to accept existing problems may help overcoming technostress under certain conditions which, in turn, has to be carefully considered by both employers and employees.

At last, research paper #5 meets the lack of research regarding job-specific knowledge of technostress by examining technostress within the heterogeneous and highly digitalized occupational group of data scientists. At first, the paper tackles the problems of classifying data scientists due to their heterogeneity of roles and tasks by delivering a definition approach of data scientists based on their use of ICT. Subsequently, four different groups of data scientists' workplaces were detected by performing latent class analysis via job- and company-related workplace attributes which are associated with general work stress on a sample of 486 German data scientists. These groups were then compared via global and pairwise van der Waerden normal score tests for gaining insights into how different types of data scientists perceive the challenges of technostress. Results show that data scientists working at different workplaces exhibit significant distinctions of technostress creators, strains due to the use of ICT, and job performance. In this context, the technostress-related findings partially contradict results of work stress studies. Thus, the paper contributes to technostress research by examining findings of work stress research in technostress context: it provides evidence that data scientists as an important occupational group which has been shown to be crucial for creating competitive advantages must not be unified in the context of technostress but, instead, differ in their perception of technostress with respect to their workplace environment. Managers are therefore recommended to implement more specific strategies to provide support for data scientists in overcoming technostress at work.

An overview of the described research papers can be seen in Table 1. Subsequently, the described research papers are provided in the following chapters II to VI within this thesis.

Paper	Title	Authors	Content	Methodology	Data
#1	Working in Detail: How LSTM Hyperparameter Selection Influences Sentiment Analysis Results	Nicholas Daniel Derra Daniel Baier	• performs the IMDB large movie dataset sentiment analysis task via deep long short term memory (LSTM) networks • analyses the effects of 8 hyperparameters via separate variation • investigates the potential of cumulating positive effects of hyperparameter variants on overall network performance	Deep LSTM networks	50,000 online movie reviews
#2	Call Me Maybe: Methods and Practical Implementation of Artificial Intelligence in Call Center Arrivals' Forecasting	Tobias Albrecht Theresa Maria Rausch Nicholas Daniel Derra	• successfully implements machine learning models to call centre arrivals' forecasting • compares machine learning approaches and conventional time series models via cross-validation with an expanding rolling window • enhances practical implementation of machine learning by providing a methodological walk-through example of the developed comparison process	Various machine learning approaches; conventional time series models	2 datasets of call centre arri-vals (31,410 observations each)
#3	Predicting Online Shopping Cart Abandonment with Machine Learning Approaches	Theresa Maria Rausch Nicholas Daniel Derra Lukas Wolf	• successfully implements machine learning models to online shopping cart abandonment prediction • compares machine learning approaches with standard logistic regression as a conventional benchmark model regarding prediction performance and practicability	Various machine learning approaches; standard logistic regression	821,048 aggregated clickstream observations
#4	Mitigating the Negative Consequences of ICT Use: The Moderating Effect of Active-Functional and Dysfunctional Coping	Julia Becker Nicholas Daniel Derra Christian Regal Torsten M. Kühlmann	• brings together psychology and information systems research • conceptualizes coping as a personal resource within the JD-R model, moderating the relationship of stressors due to ICT use and exhaustion • investigates the role of active-functional and dysfunctional coping as reactive strategies for overcoming technostress, focusing on both organisational and individual outcomes	Covariance-based structural equation modelling	3,362 German knowledge workers
#5	Examining Technostress at Different Types of Data Scientists' Workplaces	Nicholas Daniel Derra Christian Regal Simon Henrik Rath Torsten M. Kühlmann	• defines employees who work as data scientists via the specific usage of digital technologies • classifies different types of data scientists' workplaces based on 8 general workplace attributes which are related to overall work stress • examines technostress within the occupational group of data scientists by comparing the detected subclasses in terms of tech-nostress creators, strains due to the use of ICT, and overall job performance	Latent class analysis; van der Waerden normal score test	486 German data scientists

Table 1. Overview of Included Research Papers within this Thesis

References

Agarwal, R., & Dhar, V. (2014). Editorial —Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research*, *25*(3), 443–448. https://doi.org/10.1287/isre.2014.0546

Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research Trends on Big Data in Marketing: A Text Mining and Topic Modeling Based Literature Analysis. *European Research on Management and Business Economics*, *24*(1), 1–7. https://doi.org/10.1016/j.iedeen.2017.06.002

Ansari, A., & Riasi, A. (2016). Modelling and Evaluating Customer Loyalty Using Neural Networks: Evidence from Startup Insurance Companies. *Future Business Journal, 2*(1), 15–30. https://doi.org/10.1016/j.fbj.2016.04.001

Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer Relationship Management and Big Data Enabled: Personalization & Customization of Services. *Applied Computing and Informatics*, *15*(2), 94–101. https://doi.org/10.1016/j.aci.2018.05.004

Apt, W., Bovenschulte, M., Hartmann, E. A., & Wischmann, S. (2016). *Foresight-Studie „Digitale Arbeitswelt"* [Foresight Study: Digital Working World]. https://www.bmas.de/SharedDocs/Downloads/DE/PDF-Publikationen/Forschungsberichte/f463-digitale-arbeitswelt.pdf

Atanasoff, L., & Venable, M. A. (2017). Technostress: Implications for Adults in the Workforce. *The Career Development Quarterly*, *65*(4), 326–338. https://doi.org/10.1002/cdq.12111

Ayankoya, K., Calitz, A., & Greyling, J. (2014). Intrinsic Relations between Data Science, Big Data, Business Analytics and Datafication. In *Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology* (pp. 192-198). Association for Computing Machinery, New York City, New York, USA. https://doi.org/10.1145/2664591.2664619

Ayyagari, R., Grover, V., & Purvis, R. (2011). Technostress: Technological Antecedents and Implications. *MIS Quarterly, 35*(4), 831-858. https://doi.org/10.2307/41409963

Banerjee, S., Viswanathan, V., Raman, K., & Ying, H. (2013). Assessing Prime-Time for Geotargeting with Mobile Big Data. *Journal of Marketing Analytics, 1*(3), 174–183. https://doi.org/10.1057/jma.2013.16

Barber, L. K., Conlin, A. L., & Santuzzi, A. M. (2019). Workplace Telepressure and Work-Life Balance Outcomes: The Role of Work Recovery Experiences. *Stress and Health: Journal of the International Society for the Investigation of Stress*, *35*(3), 350–362. https://doi.org/10.1002/smi.2864

Baškarada, S., & Koronios, A. (2017). Unicorn Data Scientist: The Rarest of Breeds. *Program*, *51*(1), 65–74. https://doi.org/10.1108/PROG-07-2016-0053

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Chapman & Hall.

Brod, C. (1984). *Technostress: The Human Cost of the Computer Revolution*. Addison-Wesley.

Bruyn, A. de, Viswanathan, V., Beh, Y. S., Brock, J. K.-U., & Wangenheim, F. von (2020). Artificial Intelligence and Marketing: Pitfalls and Opportunities. *Journal of Interactive Marketing*, *51*, 91–105. https://doi.org/10.1016/j.intmar.2020.04.007

Brynjolfsson, E., & Mcafee, A. (2017). The Business of Artificial Intelligence. *Harvard Business Review*, 1–20. https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence

Carver, C. S. (1997). You Want to Measure Coping but Your Protocol‘s too Long: Consider the Brief COPE. *International Journal of Behavioral Medicine*, *4*(1), 92–100. https://doi.org/10.1207/s15327558ijbm0401_6

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188. https://doi.org/10.2307/41703503

Chintagunta, P., Hanssens, D. M., & Hauser, J. R. (2016a). Marketing and Data Science: Together the Future is Ours. *Marketing Intelligence Review*, *8*(2), 18–23. https://doi.org/10.1515/gfkmir-2016-0011

Chintagunta, P., Hanssens, D. M., & Hauser, J. R. (2016b). Editorial - Marketing Science and Big Data. *Marketing Science*, *35*(3), 341–342. https://doi.org/10.1287/mksc.2016.0996

Chollet, F., & Allaire, J. J. (2018). *Deep Learning with R*. Manning.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Costa, C., & Santos, M. Y. (2017). The Data Scientist Profile and its Representativeness in the European E-Competence Framework and the Skills Framework for the Information Age. *International Journal of Information Management*, *37*(6), 726–734. https://doi.org/10.1016/j.ijinfomgt.2017.07.010

Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Cui, G., Wong, M. L., & Lui, H.-K. (2006). Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. *Management Science*, *52*(4), 597–612. https://doi.org/10.1287/mnsc.1060.0514

Davenport, T. H. (2006). Competing on Analytics. *Harvard Business Review*, *84*(1), 98–107.

Davenport, T. H., & Patil, D. J. (2012). Data Scientist. *Harvard Business Review*, *90*(5), 70–76.

Davenport, T. (2020). Beyond Unicorns: Educating, Classifying, and Certifying Business Data Scientists. *Harvard Data Science Review*. Advance online publication. https://doi.org/10.1162/99608f92.55546b4a

Dekker, I., & Barling, J. (1995). Workforce Size and Work-Related Role Stress. *Work & Stress*, *9*(1), 45–54. https://doi.org/10.1080/02678379508251584

Demerouti, E., Bakker, A. B., Nachreiner, F., & Schaufeli, W. B. (2001). The Job Demands-Resources Model of Burnout. *The Journal of Applied Psychology*, *86*(3), 499–512. https://doi.org/10.1037//0021-9010.86.3.499

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social Media Sentiment Analysis: Lexicon Versus Machine Learning. *Journal of Consumer Marketing*, *34*(6), 480–488. https://doi.org/10.1108/JCM-03-2017-2141

Dimitrieska, S., Stankovska, A., & Efremova, T. (2018). Artificial Intelligence and Marketing. *Entrepreneurship*, *6*(2), 298–304.

Doyle, A. (2019). *Important Job Skills for Data Scientists*. https://www.thebalancecareers.com/list-of-data-scientist-skills-2062381

Duda, R. O., Hart, P. E., & Stork, D. G. (1973). *Pattern Classification and Scene Analysis*. Wiley.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data Consumer Analytics and the Transformation of Marketing. *Journal of Business Research*, *69*(2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001

Erschens, R., Loda, T., Herrmann-Werner, A., Keifenheim, K. E., Stuber, F., Nikendei, C., Zipfel, S., & Junne, F. (2018). Behaviour-Based Functional and Dysfunctional Strategies of Medical Students to Cope with Burnout, *Medical Education Online, 23*(1), 1-11. https://doi.org/10.1080/10872981.2018.1535738

Ferraris, A., Mazzoleni, A., Devalle, A., & Couturier, J. (2019). Big Data Analytics Capabilities and Knowledge Management: Impact on Firm Performance. *Management Decision*, *57*(8), 1923–1936. https://doi.org/10.1108/MD-07-2018-0825

Fischer, T., & Riedl, R. (2020). On the Stress Potential of an Organisational Climate of Innovation: A Survey Study in Germany. *Behaviour & Information Technology, 67*(1), 1–22. https://doi.org/10.1080/0144929X.2020.1836258

Folkman, S., Lazarus, R. S., Dunkel-Schetter, C., DeLongis, A., & Gruen, R. J. (1986). Dynamics of a Stressful Encounter: Cognitive Appraisal, Coping, and Encounter Outcomes. *Journal of Personality and Social Psychology*, *50*(5), 992–1003. https://doi.org/10.1037/0022-3514.50.5.992

Galluch, P., Grover, V., & Thatcher, J. (2015). Interrupting the Workplace: Examining Stressors in an Information Technology Context. *Journal of the Association for Information Systems*, *16*(1), 1–47. https://doi.org/10.17705/1jais.00387

Ganster, D. C. (2005). Response: Executive Job Demands: Suggestions from a Stress and Decision-Making Perspective. *Academy of Management Review*, *30*(3), 492–502. https://doi.org/10.5465/amr.2005.17293366

Gaudioso, F., Turel, O., & Galimberti, C. (2016). Explaining Work Exhaustion from a Coping Theory Perspective: Roles of Techno-Stressors and Technology-Specific Coping Strategies. *Studies in Health Technology and Informatics*, *219*, 14–20. https://doi.org/10.3233/978-1-61499-595-1-14

Ghose, A., & Yang, S. (2009). An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets. *Management Science, 55*(10), 1605–1622. https://doi.org/10.1287/mnsc.1090.1054

Godey, B., Manthiou, A., Pederzoli, D., Rokka, J., Aiello, G., Donvito, R., & Singh, R. (2016). Social Media Marketing Efforts of Luxury Brands: Influence on Brand Equity and Consumer Behavior. *Journal of Business Research*, *69*(12), 5833-5841. https://doi.org/10.1016/j.jbusres.2016.04.181

Golubic, R., Milosevic, M., Knezevic, B., & Mustajbegovic, J. (2009). Work-Related Stress, Education and Work Ability Among Hospital Nurses. *Journal of Advanced Nursing*, *65*(10), 2056–2066. https://doi.org/10.1111/j.1365-2648.2009.05057.x

Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep Learning*. MIT.

Grace, M. K., & van Heuvelen, J. S. (2019). Occupational Variation in Burnout Among Medical Staff: Evidence for the Stress of Higher Status. *Social Science & Medicine*, *232,* 199–208. https://doi.org/10.1016/j.socscimed.2019.05.007

Hackman, J.R., & Oldham, G. R. (1976). Motivation Through the Design of Work: Test of a Theory. *Organizational Behavior and Human Performance*, *16*(2), 250–279. https://doi.org/10.1016/0030-5073(76)90016-7

Hagen, L., Uetake, K., Yang, N., Bollinger, B., Chaney, A. J. B., Dzyabura, D., Etkin, J., Goldfarb, A., Liu, L., Sudhir, K., Wang, Y., Wright, J. R., & Zhu, Y. (2020). How can Machine Learning Aid Behavioral Marketing Research? *Marketing Letters*, *31*(4), 361–370. https://doi.org/10.1007/s11002-020-09535-7

Hambrick, D. C., Finkelstein, S., & Mooney, A. C. (2005). Executive Job Demands: New Insights for Explaining Strategic Decisions and Leader Behaviors. *Academy of Management Review*, *30*(3), 472–491. https://doi.org/10.5465/amr.2005.17293355

Hartl, E. (2019). A Characterization of Culture Change in the Context of Digital Transformation. In *Proceedings of the 25th Americas Conference on Information Systems AMCIS* (pp. 1-10). Symposium conducted at the meeting of the Association for Information Systems, Cancun, Mexico.

Hartline, M. D., & Ferrell, O. C. (1996). The Management of Customer-Contact Service Employees: An Empirical Investigation. *Journal of Marketing*, *60*(4), 52–70. https://doi.org/10.1177/002224299606000406

Hauk, N., Göritz, A. S., & Krumm, S. (2019). The Mediating Role of Coping Behavior on the Age-Technostress Relationship: A Longitudinal Multilevel Mediation Model. *PloS One*, *14*(3), 1-22. https://doi.org/10.1371/journal.pone.0213349

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data Quality for Data Science, Predictive Analytics, and Big Data in Supply Chain Management: An Introduction to the Problem and Suggestions for Research and Applications. *International Journal of Production Economics*, *154*, 72–80. https://doi.org/10.1016/j.ijpe.2014.04.018

Hinton, G., Osindro, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, *18*(7), 1527–1554.

Ho, A., Nguyen, A., Pafford, J. L., & Slater, R. (2019). A Data Science Approach to Defining a Data Scientist. *SMU Data Science Review*, *2*(3), Article 4.

Hochreiter S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

IBM (2012). *What is Big Data?* http://www-01.ibm.com/software/data/bigdata/

Ismail, N. A., & Abidin, A. Z. (2016). Data Scientist Skills. *Journal of Mobile Computing & Application*, *3*(4), 52–61. https://doi.org/10.9790/0050-03045261

Jabbar, A., Akhtar, P., & Dani, S. (2020). Real-Time Big Data Processing for Instantaneous Marketing Decisions: A Problematization Approach. *Industrial Marketing Management*, *90*, 558–569. https://doi.org/10.1016/j.indmarman.2019.09.001

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial Neural Networks: A Tutorial. *Computer*, *29*(3), 31–44. https://doi.org/10.1109/2.485891

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors Influencing Big Data Decision-Making Quality. *Journal of Business Research*, *70*(1), 338–345.

Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science*, *349*(6245), 253–255. https://doi.org/10.1126/science.aac4520

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kohavi, R., & Provost, F. (2001). Applications of Data Mining to Electronic Commerce. In F. Provost & R. Kohavi (Eds.), *Provost, Kohavi (Hg.) 2001 – Applications of Data Mining to Electronic Commerce* (pp. 5–10). Springer.

Lansisalmi, H., Peiro, J. M., & Kivimaki, M. (2000). Collective Stress and Coping in the Context of Organizational Culture. *European Journal of Work and Organizational Psychology*, *9*(4), 527–559. https://doi.org/10.1080/13594320050203120

Lazarus, R. S., & Folkman, S. (1984). *Stress, Appraisal, and Coping*. Springer.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten Digit Recognition with a Back-Propagation Network. *Advances in Neural Information Processing Systems*, *2*, 396–404.

Lewin, J. E., & Sager, J. K. (2009). An Investigation of the Influence of Coping Resources in Sales-Persons‘ Emotional Exhaustion. *Industrial Marketing Management*, *38*(7), 798–805. https://doi.org/10.1016/j.indmarman.2008.02.013

Li, J., Ni, X., Yuan, Y., & Wang, F. Y. (2018). A Hierarchical Framework for Ad Inventory Allocation in Programmatic Advertising Markets. *Electronic Commerce Research and Applications*, *31*(May), 40–51. https://doi.org/10.1016/j.elerap.2018.09.001.

Lukosius, V., & Hyman, M. R. (2019). Marketing Theory and Big Data. *The Journal of Developing Areas*, *53*(4), 1-9. https://doi.org/10.1353/jda.2018.0082.

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science*, *38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192

Lycett, M. (2013). ‘Datafication’: Making Sense of (Big) Data in a Complex World. *European Journal of Information Systems*, *22*(4), 381–386. https://doi.org/10.1057/ejis.2013.10

Ma, L., & Sun, B. (2020). Machine Learning and AI in Marketing – Connecting Computing Power to Human Insights. *International Journal of Research in Marketing*, *37*(3), 481–504. https://doi.org/10.1016/j.ijresmar.2020.04.005

Maass, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142-150). Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA.

Manieri, A., Brewer, S., Riestra, R., Demchenko, Y., Hemmje, M., Wiktorski, T., Ferrari, T., & Frey, J. (2015). Data Science Professional Uncovered: How the EDISON Project will Contribute to a Widely Accepted Profile for Data Scientists. In *Proceedings of the 7th IEEE International Conference on Cloud Computing Technology and Science (CloudCom’15)* (pp. 588–593). IEEE, Washington, DC, USA. https://doi.org/10.1109/CloudCom.2015.57

Mauro, A. de, Greco, M., Grimaldi, M., & Ritala, P. (2018). Human Resources for Big Data Professions: A Systematic Classification of Job Roles and Required Skill Sets. *Information Processing & Management*, *54*(5), 807–817. https://doi.org/10.1016/j.ipm.2017.05.004

Mikalef, P., Giannakos, M. N., Pappas, I. O., & Krogstie, J. (2018). The Human Side of Big Data: Understanding the Skills of the Data Scientist in Education and Industry. In *2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 503–512). IEEE, New York City, New York, USA. https://doi.org/10.1109/EDUCON.2018.8363273

Müller, O., Fay, M., & vom Brocke, J. (2018). The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. *Journal of Management Information Systems*, *35*(2), 488–509. https://doi.org/10.1080/07421222.2018.1451955

Murawski, M., & Bick, M. (2017). Digital Competences of the Workforce – a Research Topic?. *Business Process Management Journal*, *23*(3), 721–734. https://doi.org/10.1108/BPMJ-06-2016-0126

Nisafani, A. S., Kiely, G., & Mahony, C. (2020). Workers' Technostress: A Review of its Causes, Strains, Inhibitors, and Impacts. *Journal of Decision Systems, 61*(10), 1–16. https://doi.org/10.1080/12460125.2020.1796286

Okkonen, J., Vuori, V., & Palvalin, M. (2019). Digitalization Changing Work: Employees' View on the Benefits and Hindrances. In Á. Rocha, C. Ferrás, & M. Paredes (Eds.), *Information Technology and Systems: Proceedings of ICITS 2019* (pp. 165–176). Springer, Cham, Switzerland.

Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, *11*, 169–198. https://doi.org/10.1613/jair.614

Osmundsen, K., Iden, J., & Bygstad, B. (2018). Digital Transformation: Drivers, Success Factors, and Implications. In *Proceedings of the 12th Mediterranean Conference on Information Systems MCIS*. Symposium conducted at the meeting of the Association for Information Systems, Corfu, Greece. https://aisel.aisnet.org/mcis2018/37/

Palacios-Marqués, D., Welsh, D. H., Méndez-Picazo, M. T., Palacios-Marqúes, D., Welsh, D. H. B., & Méndez-Picazo, M. T. (2016). The Importance of the Activities of Innovation and Knowledge in the Economy: Welcome to the Journal of Innovation and Knowledge. *Journal of Innovation and Knowledge*, *1*(1), 1–2. https://doi.org/10.1016/j.jik.2016.02.002

Pirkkalainen, H., Salo, M., Tarafdar, M., & Makkonen, M. (2019). Deliberate or Instinctive? Proactive and Reactive Coping for Technostress. *Journal of Management Information Systems*, *36*(4), 1179–1212. https://doi.org/10.1080/07421222.2019.1661092

Portugal, I., Alencar, P., & Cowan, D. (2018). The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review. *Expert Systems with Applications*, *97*, 205–227. https://doi.org/10.1016/j.eswa.2017.12.020

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Qi, J., Zhang, Z., Jeon, S., & Zhou, Y. (2016). Mining Customer Requirements from Online Reviews: A Product Improvement Perspective. *Information & Management*, *53*(8), 951–963. https://doi.org/10.1016/j.im.2016.06.002

Ragu-Nathan, T. S., Tarafdar, M., Ragu-Nathan, B. S., & Tu, Q. (2008). The Consequences of Technostress for End Users in Organizations: Conceptual Development and Empirical Validation. *Information Systems Research*, *19*(4), 417–433. https://doi.org/10.1287/isre.1070.0165

Rees, D., & Cooper, C. L. (1992). Occupational Stress in Health Service Workers in the UK. *Stress Medicine*, *8*(2), 79–90. https://doi.org/10.1002/smi.2460080205

Reisenbichler, M., & Reutterer, T. (2019). Topic Modeling in Marketing: Recent Advances and Research Opportunities. *Journal of Business Economics*, *89*(3), 327–356. https://doi.org/10.1007/s11573-018-0915-7

Richards, D., & Marrone, M. (2014). Identifying the Education Needs of the Business Analyst: An Australian Study. *Australasian Journal of Information Systems*, *18*(2). https://doi.org/10.3127/ajis.v18i2.803

Riedl, R., Kindermann, H., Auinger, A., & Javor, A. (2012). Technostress from a Neurobiological Perspective. *Business & Information Systems Engineering*, *4*(2), 61–69. https://doi.org/10.1007/s12599-012-0207-7

Rogers, D., & Sexton, D. (2012). *Marketing ROI in the Era of Big Data: The 2012 BRITENYAMA Marketing in Transition Study*. Columbia Business School’s Center on Global Brand Leadership. www.iab.net/media/file/2012-BRITE-NYAMA-Marketing-ROI-Study.pdf

Rokach, L. (2010). Ensemble-Based Classifiers. *Artificial Intelligence Review*, *33*(1-2), 1–39. https://doi.org/10.1007/s10462-009-9124-7

Rust, R. T. (2020). The Future of Marketing. *International Journal of Research in Marketing*, *37*(1), 15–26. https://doi.org/10.1016/j.ijresmar.2019.08.002

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Santuzzi, A. M., & Barber, L. K. (2018). Workplace Telepressure and Worker Well-Being: The Intervening Role of Psychological Detachment. *Occupational Health Science*, *2*(4), 337–363. https://doi.org/10.1007/s41542-018-0022-8

Saura, J. R. (2020). Using Data Sciences in Digital Marketing: Framework, Methods, and Performance Metrics. *Journal of Innovation & Knowledge.* Advance online publication. https://doi.org/10.1016/j.jik.2020.08.001

Schellong, D., Kemper, J., & Brettel, M. (2017). Generating Consumer Insights from Big Data Clickstream Information and the Link with Transaction-Related Shopping Behavior. In *Proceedings of the 25th European Conference on Information Systems (ECIS).* Association for Information Systems, Atlanta, Georgia, USA.

Schwemmle, M., & Wedde, P. (Eds.). (2012). *Digitale Arbeit in Deutschland: Potenziale und Problemlagen*. Friedrich-Ebert-Stiftung, Medienpolitik.

Searle, B. J., & Lee, L. (2015). Proactive Coping as a Personal Resource in the Expanded Job Demands–Resources Model. *International Journal of Stress Management*, *22*(1), 46–69. https://doi.org/10.1037/a0038439

Srivastava, S. C., Chandra, S., & Shirish, A. (2015). Technostress Creators and Job Outcomes: Theorising the Moderating Influence of Personality Traits. *Information Systems Journal, 25*(4), 355–401. https://doi.org/10.1111/isj.12067

Stanton, A. D.'A., & Stanton, W. W. (2016). The Relationship Between Big Data, Data Science, Digital Analytics and the Skills and Abilities Needed to Optimise Marketing Decisions. *Applied Marketing Analytics*, *2*(3), 265–279.

Sumiyana, S., & Sriwidharmanely, S. (2020). Mitigating the Harmful Effects of Technostress: Inducing Chaos Theory in an Experimental Setting. *Behaviour & Information Technology, 39*(10), 1079–1093. https://doi.org/10.1080/0144929X.2019.1641229

Stinis, P. (2019). Enforcing Constraints for Time Series Prediction in Supervised, Unsupervised and Reinforcement Learning. *ArXiv:1905.07501*.

Tarafdar, M., Cooper, C. L., & Stich, J.-F. (2019). The Technostress Trifecta - Techno Eustress, Techno Distress and Design: Theoretical Directions and an Agenda for Research. *Information Systems Journal*, *29*(1), 6–42. https://doi.org/10.1111/isj.12169

Tarafdar, M., Pullins, E., & Ragu-Nathan, T. S. (2015). Technostress: Negative Effect on Performance and Possible Mitigations. *Information Systems Journal*, *25*(2), 103–132. https://doi.org/10.1111/isj.12042

Tarafdar, M., Tu, Q., & Ragu-Nathan, T. S. (2010). Impact of Technostress on End-User Satisfaction and Performance. *Journal of Management Information Systems*, *27*(3), 303–334. https://doi.org/10.2753/MIS0742-1222270311

Tarafdar, M., Tu, Q., Ragu-Nathan, B. S., & Ragu-Nathan, T. S. (2007). The Impact of Technostress on Role stress and Productivity. *Journal of Management Information Systems*, *24*(1), 301–328. https://doi.org/10.2753/MIS0742-1222240109

Tarafdar, M., Tu, Q., Ragu-Nathan, T. S., & Ragu-Nathan, B. S. (2011). Crossing to the Dark Side: Examining Creators, Outcomes and Inhibitors of Technostress. *Communications of the ACM*, *54*(9), 113–120. https://doi.org/10.1145/1995376.1995403

Thompson, N., Stradling, S., Murphy, M., & O'neill, P. (1996). Stress and Organizational Culture. *The British Journal of Social Work*, *26*(5), 647–665. https://doi.org/10.1093/oxfordjournals.bjsw.a011139

Tiago, M. T. P., & Veríssimo, J. M. C. (2014). Digital Marketing and Social Media: Why Bother? *Business Horizons*, *57*(6), 703–708. https://doi.org/10.1016/j.bushor.2014.07.002

Timonen, H., & Vuori, J. (2018). Visibility of Work: How Digitalization Changes the Workplace. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. Association for Information Systems, Atlanta, Georgia, USA. https://doi.org/ 10.24251/HICSS.2018.634

Travers, C. J., & Cooper, C. L. (1993). Mental Health, Job Satisfaction and Occupational Stress Among UK Teachers. *Work & Stress*, *7*(3), 203–219. https://doi.org/10.1080/02678379308257062

Tsai, C. F., Hu, Y. H., & Lu, Y. H. (2015). Customer Segmentation Issues and Strategies for an Automobile Dealership with Two Clustering Techniques. *Expert Systems with Applications*, *32*(1), 65–76. https://doi.org/10.1111/exsy.12056

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A Comparison of Machine Learning Techniques for Customer Churn Prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003

van der Aalst, W. M. (2014). Data Scientist: The Engineer of the Future. In Mertins K., Bénaben F., Poler R., & Bourrières JP (Eds.), *Enterprise Interoperability VI: Proceedings of the I-ESA Conferences* (Vol. 7, pp. 13–26). Springer International Publishing, Basel, Switzerland. https://doi.org/10.1007/978-3-319-04948-9_2 van Dijkhuizen, N., & Reiche, H. M. (1980). Psychosocial Stress in Industry: A Heartache for Middle Management? *Psychotherapy and Psychosomatics*, *34*(2-3), 124–134.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, *11*, 3371–3408.

Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: a Revolution that will Transform Supply Chain Design and Management. *Journal of Business Logistics*, *34*(2), 77–84. https://doi.org/10.1111/jbl.12010

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J.-f., Dubey, R., & Childe, S. J. (2017). Big Data Analytics and Firm Performance: Effects of Dynamic Capabilities. *Journal of Business Research*, *70*(1), 356–365. https://doi.org/10.1016/J.JBUSRES.2016.08.009

Wedel, M., & Kannan, P. K. (2016). Marketing Analytics for Data-Rich Environments. *Journal of Marketing*, *80*(6), 97–121. https://doi.org/10.1509/jm.15.0413

Winer, R. S., & Neslin, S. A. (2014). *The History of Marketing Science*. World Scientific Publishing. https://doi.org/10.1142/9128

Xu, R., & Wunsch, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141

Yip, B., Rowlinson, S., & Siu, O. L. (2008). Coping Strategies as Moderators in the Relationship between Role Overload and Burnout. *Construction Management and Economics*, *26*(8), 871–882. https://doi.org/10.1080/01446190802213529

Zhao, X., Xia, Q., & Huang, W. (2020). Impact of Technostress on Productivity from the Theoretical Perspective of Appraisal and Coping Processes. *Information & Management*, *57*(8), 103265. https://doi.org/10.1016/j.im.2020.103265

II Working in Detail: How LSTM Hyperparameter Selection Influences Sentiment Analysis Results

Authors: Derra, Nicholas Daniel; Baier, Daniel

Published in: Archives of Data Science, Series A (2020), Vol. 6, No. 1, pp. 1-22,
DOI: 10.5445/KSP/1000098011/10

Abstract: Sentiment analysis of written customer reviews is a powerful way to generate knowledge about customer attitudes for future marketing activities. Meanwhile, Deep Learning as the most powerful machine learning method is of particular importance for sentiment analysis tasks. Due to this current relevance, an LSTM network based on a literature review to solve the challenging classification task of the IMDB Large Movie Dataset is created. Hyperparameters are varied separately from each other to better understand their single influences on the overall model accuracy. Furthermore, we transformed variants with positive impacts into a final model in order to investigate whether the impacts can be cumulated. While preparing the amount of training data and the number of iteration steps resulted in a higher accuracy, pre-trained word vectors and higher network capacity did not work well separately. Even though implementing the variants with positive influences together raised the model´s performance,
the improvement was lower than some single variants.

Keywords: Sentiment Analysis, Deep Learning, LSTM, Hyperparameter, Optimization

1 Introduction

Sentiment analysis (SA) has been one of the largest fields of research in natural language processing (NLP), data mining, text mining and information retrieval since the beginning of the 21st century. Due to the ever-increasing use of internet and online activities (e-commerce, forums, blogs and social networks) for presenting personal opinions about products and services, the analysis of the resulting huge amounts of data (Big Data) is of particular importance for marketing managers (Zhang et al, 2018b). Meanwhile, Deep Learning (DL) algorithms deliver stronger results in processing sequential text data for SA tasks than other Machine Learning (ML) methods do (LeCun et al, 2015). For this, the current literature focuses on the development of models that classify popular benchmark datasets (IMDB Large Movie Dataset by Maas et al (2011); Yelp Dataset by Zhang et al (2015)) with a new accuracy high score. We argue that in this context only the overall performance of an architecture is observed while the various influences of individual hyperparameters on the model performance are insufficiently analysed. For this reason, the separate effects of various hyperparameters within an LSTM network for the IMDB Large Movie Dataset sentiment analysis task are observed through separate variation. Simultaneously, after a short introduction (Section 1), the discussion of theoretical backgrounds including SA (Section 2.1) and DL models for SA (Section 2.2) as well as the description of the IMDB dataset (Section 3.1) and related work (Section 3.2), an LSTM which is able to solve the IMDB SA task with high accuracy is constructed (Section 3.3). Within this model, 8 hyperparameters are separately varied to investigate their impact on classification performance. Subsequently, the variations with a positive impact on validation accuracy are transformed into a final model in order to cumulate the effects. This final model is then compared with the single hyperparameter variants by test accuracy. In addition, the machine times required are also measured (Section 3.4). Finally, the results are discussed in Section 4.

2 Theoretical Background

2.1 Sentiment Analysis

SA, also called mood analysis, is the field of computational studies of emotions as well as opinions, feelings, evaluations and attitudes towards objects such as products, services, organizations, individuals, events, topics and issues as well as their characteristics (Ain et al, 2017; Medhat et al, 2014). They are analysed in forums, blogs, social networks, e-commerce websites, reports and other internet sources (Ravi and Ravi, 2015). SA is a subset of both NLP and affective computing (AC) (Yadollahi et al, 2017; Zhang et al, 2018a) and can therefore be seen as an intersection of both areas of research.

It is carried out by methods of information retrieval and data mining (Ravi and Ravi, 2015). While the different SA tasks can be correctly subdivided into the subareas of opinion mining (analysis of contained opinions in texts) and emotion mining (analysis of contained emotions in texts) (Yadollahi et al, 2017), a more comprehensive approach summarizing opinions and emotions (Ravi and Ravi, 2015) seems to be more effective. Since the concept of a sentiment encompasses both opinions and emotions, a precise SA can only be achieved by analysing both areas simultaneously (Ain et al, 2017; Medhat et al, 2014).

While SA is often used as a synonym for sentiment or polarity classification, it is considered to be the central SA task (Cambria et al, 2013). However, in this article this trend of literature is taken into account (see inter alias Araque et al (2017) and Medhat et al (2014)) so the term SA is used interchangeably after the various fields of SA tasks were shown. A sentiment, respectively polarity classification, is the recognition of the sentiment orientation within a text and the classification into one of at least two classes. As the most common task in SA, the polarity classification classifies texts according to their opinion into a predefined sentiment polarity, whereby both binary, tertiary and finer n-grade classifications are possible (Ravi and Ravi, 2015). Polarity classification can take place on three granularity levels, regardless of the classification object (opinion, emotion or both). For this, the document, sentence and aspect level are differentiated (Medhat et al, 2014; Yadollahi et al, 2017; Zhang et al, 2018a) where the polarity classification at the document level is considered to be the most common. At the document level, a complete text document is considered as the smallest unit. This document expresses an overall positive or negative opinion or emotion and it is usually assigned either to the positive or the negative class (Aggarwal and Aggarwal, 2017; Medhat et al, 2014; Yadollahi et al, 2017; Zhang et al, 2018a). Yet, the length of the document is irrelevant (Yadollahi et al, 2017). At this level it is assumed that not every single sentence contains an opinion relating to the subject so the document contains irrelevant sentences (Aggarwal and Aggarwal, 2017). Since the IMDB sentiment classification task is to classify film ratings of different lengths and without focusing on specific aspects with respect to their polarity, the IMDB task is performed at the document level.

Approach	**Reference**	**Accuracy**
Support Vector Machines (SVM)	Wang and Manning (2012)	89.16%
Maximum Entropy (ME)	Brychcín and Habernal (2013)	92.24%
Naive Bayes (NB)	Narayanan et al (2013)	88.80%
NB-SVM	Mesnil et al (2014)	91.87%
Decision Trees (DT)	Zhou and Feng (2017)	89.16%
Deep Learning (DL)	Howard and Ruder (2018)	95.40%

Table 1: *ML methods for the IMDB sentiment classification task*

The polarity classification approaches can be divided into ML-based, lexicon-based, while hybrid approaches are ultimately a combination of ML with a previously created lexicon (Maynard and Funk, 2011; Ravi and Ravi, 2015). ML techniques treat sentiment classifications as text classification tasks and use syntactic and linguistic properties to solve problems (Medhat et al, 2014). They clearly outperform the semantic approaches in dealing with specific tasks (Ravi and Ravi, 2015). They are divided into methods of supervised, unsupervised and semi-supervised learning, whereby unsupervised ML methods only play a minor role in SA research and are only marginally or not explained in the relevant overview literature (Medhat et al, 2014; Ravi and Ravi, 2015; Yadollahi et al, 2017).

For polarity classification with supervised learning, probabilistic classifiers such as Bayesian Networks, Naive Bayes classifiers and Maximum Entropy classifiers, linear classifiers such as Support Vector Machines and Artificial Neural Networks (Deep Learning), Decision Trees and Rule-based classifiers are frequently used (Medhat et al, 2014; Ravi and Ravi, 2015). In terms of the IMDB Large Movie Dataset, the classification performance of the different methods is shown in table 1. Thereby, DL models have achieved a large number of correct classification rates higher than 92% in recent years, massively outperforming other ML approaches (compare our literature review in Section 4). The current IMDB benchmark performed with DL achieves 95.40% accuracy (Howard and Ruder, 2018). To summarize, while ML approaches have task-specific higher accuracy than lexicon production, DL outperforms conventional ML.

2.2 Deep Learning

DL approaches are part of the research field of artificial intelligence (AI) (Arel et al, 2010) as well as a methodologically emerging area of ML called Representation Learning. Within DL methods, several stages of representation transformation take place in succession (LeCun et al, 2015). Meanwhile, DL is defined as a class of ML techniques based on Artificial Neural Networks (ANN) that use numerous (hidden) process layers in hierarchical architectures to learn characteristics and recognize patterns from data (Deng, 2011, 2014). However, the depth required for the concept of DL is not uniformly defined in research (Schmidhuber, 2015).

In the context of ANNs, the concept of learning describes a process for updating the network architecture and the weights of neuron connections to efficiently handle a specific task (Jain et al, 1996). In DL, the most commonly used supervised learning algorithm is the backpropagation method for error minimization which allowed to map direct connections of neurons over several layers so that the weights within the ANNs were efficiently learned (Deng, 2014; Schmidhuber, 2015). In general, backpropagation is a special case of the general gradient descent process (Schmidhuber, 2015). This approach by Rumelhart et al (1986) repeatedly adjusts the weights within an ANN to minimize the difference between the actual output vector and the known output vector setpoint for finding an optimal set of weights. The quality of the weights is described by the difference between the actual and target output vectors in a quadratic error function.

Basically, Deep Neural Networks are classified in Feed Forward (FNN) and Recurrent (RNN) Neural Networks (Jain et al, 1996; Schmidhuber, 2015). Furthermore, the forward models are divided into Deep Autoencoders (DAE), Deep Belief Networks (DBN) and Convolutional Neural Networks (CNN) (Deng, 2014; Zhang et al, 2018b). The recurrent networks were later developed into so-called Long Short Term Memories (LSTM) (Gers et al, 1999; Hochreiter and Schmidhuber, 1997). While DAEs and DBNs are only used for (unsupervised) pre-training in polarity classification tasks, one-dimensional CNNs, but especially RNNs and their powerful relatives LSTMs are able to classify text data very well (LeCun et al, 2015).

RNNs are more powerful than any forward DL model because of their ability to create memories (Schmidhuber, 2015). Due to the backward links, they can account for time sequences and are therefore perfect in processing sequential data, e.g. natural language. RNNs have a cyclic architecture and are able to learn the data properties through a memory from previous inputs (Jain et al, 1996; Zhang et al, 2018a). The memory of an RNN is its ability to process all the elements of a sequence where the input of a unit thus consists of two parts, the current input and the output of previous calculations (Zhang et al, 2018a). This is possible because the information from previous calculations is stored as an internal state within the RNN (LeCun et al, 2015; Zhang et al, 2018a).

However, especially at deep RNNs, the vanishing or exploding gradients during backpropagation training has proved to be very problematic due to long-term dependencies (Bengio et al, 1994; Hochreiter, 1991; Schmidhuber, 2015; Zhang et al, 2018a). To address this phenomenon called fundamental DL problem, LSTMs were developed (Gers et al, 1999; Hochreiter and Schmidhuber, 1997). Today, the most successful RNNs are based on this architecture (Deng, 2014; Schmidhuber, 2015). By using so-called constant error carousels, also known as memory cells, LSTMs are able to remember processes that already took place many time steps ago. These units are connected to themselves with a weight of 1 and thus copy their own state. This connection is linked to another unit, called gate unit, which decides when to erase the learned memory, which information is erased, and which new information is stored in the memory (Gers et al, 1999; Hochreiter and Schmidhuber, 1997;

LeCun et al, 2015; Zhang et al, 2018a). Accordingly, a distinction is made between input, forget and output gate units (Hochreiter and Schmidhuber, 1997; Zhang et al, 2018a). The additional possibility of forgetting information and the associated influence on the internal memory enables the effective use of long-term dependencies without vanishing or exploding gradients.

Conventional RNNs and LSTMs can only use the information of previous time steps and therefore do not use all available information of sequential data (Zhang et al, 2018a). For this reason, Bidirectional LSTMs (BiLSTM) have been developed. They consist of two opposing LSTMs stacked on top of each other and are thereby able to process text sequences forward and backward at the same time. Finally, the internal states of both networks are taken into account for calculating the output of the bidirectional network (Schuster and Paliwal, 1997). The bidirectional architecture often provides better sentiment classification results than its unidirectional counterparts since the context between a given word within a text and its subsequent words might be as important as the context to previous words for classifying the sentiment of this word (see, e.g., Howard and Ruder (2018), Johnson and Zhang (2016)).

The danger in supervised learning processes, so-called overfitting, is often caused by a limited amount of training data, too many parameters to be learned (the network capacity) or a large number of training epochs. In such a case, the network learns to identify specific characteristics of the training data which are irrelevant or even obstructive for classifying unknown data (Srivastava et al, 2014). Thus, the task-specific generalization decreases with additional training epochs so the model loses massive usefulness in the analysis of unknown data. RNNs -particularly their bidirectional variants- are quite susceptible to overfitting due to their huge capacity (memory architecture and additional backward neuron connections) so that such models are usually trained with fewer epochs than other architectures in order to learn cumbersome specific features (Hong and Fang, 2015).

In addition, to avoid overfitting, another hyperparameter can be integrated into the model. This method, known as dropout regularization, randomly sets a share of its output per layer to zero, thus extracting a thinned net from the original complex model. The size of this eliminated share is determined by the dropout rate. As a result, the network does not learn any irrelevant patterns contained in the training data which improves unknown data performance a lot (Srivastava et al, 2014). The additional implementation of a recurrent dropout rate makes this method implementable for RNNs (Gal and Ghahramani, 2015).

Since DL algorithms (like other ML methods as well) can not use text data as input, datasets in text form have to be converted into numerical vectors (Zhang et al, 2018a). This results in very high-dimensional property vectors (called One Hot Encoding (OH)) since each word contained must be assigned its own value. ML applications, therefore, require a feature selection step that removes unimportant properties or words for the task to be performed and thus reduces the dimensionality without reducing the quality of the subsequent classification (Rui et al, 2016; Yang and Pedersen, 1997).

An advantage of sentiment classification via DL is that, in contrast to other ML methods, no feature selection is necessary to avoid these high-dimensional feature vectors since DL models are able to handle high-dimensional data very well and process a feature selection by using the embedding layer for training so-called word embeddings. Using a specific algorithm, it generates smaller numerical vectors and at the same time more information contained by removing the words which are irrelevant for the classification task. Examples of such word embedding algorithms are Word2Vec (Mikolov et al, 2013) and GloVe (Pennington et al, 2014). The word embeddings and the weights are learned simultaneously based on the present training data. If there is insufficient training data for a classification task, pre-trained word embeddings calculated using one of the two algorithms can be used. Such pre-trained vectors are freely available via internet (for Word2Vec: see Google (2013), for GloVe: see Stanford (2014)).

3 Experiments

3.1 Dataset

The IMDB Large Movie Dataset was developed by Maas et al (2011). It was designed to meaningfully test and compare binary sentiment classification methods. This dataset contains 100,000 film ratings from the Internet Movie Database (50,000 labeled and 50,000 unlabeled samples), with each movie represented by a maximum of 30 ratings (Maas et al, 2011). The goal of the IMDB SA task is to correctly classify whether a movie rating is positive or negative. The average length of a review document is 231 words (Wang and Manning, 2012). Within the labeled data, there are 25,000 positive and 25,000 negative reviews each, with only clearly polarized contributions taken into account. Therefore, neutral reviews are not included. The labeled dataset is also divided into 25,000 reviews for training and testing each (Maas et al, 2011). The unlabeled training dataset with 50,000 reviews is intended to, e.g., train a semi-supervised architecture with unsupervised pre-training. This dataset contains positive, neutral and negative sentiments (Maas et al, 2011). In general, is has to be mentioned that the particular difficulty of classifying film ratings presents a major challenge for all ML methods (Turney, 2002). The basic difficulties and challenges in text analysis, including irony, sarcasm, various word diffractions, synonyms, stop words, etc. are just as demanding as the different lengths of the evaluation documents.

3.2 Related Work

Reference	Architecture	Specific Architecture	Test Accuracy
Le and Mikolov (2014)	FNN	PV-FNN	92.58%
Dai and Le (2015)	LSTM	SA-LSTM	92.80%
Johnson and Zhang (2015)	CNN	RE-CNN	93.49%
Dieng et al (2016)	RNN	Topic-RNN	93.72%
Johnson and Zhang (2016)	LSTM	OH-BiLSTM	94.06%
Miyato et al (2016)	LSTM	VA-LSTM	94.09%
Gray et al (2017)	LSTM	Block-Sparse LSTM	94.99%
Radford et al (2017)	LSTM	Byte-Level LSTM	92.88%
Xu et al (2017)	RNN	SSVAE-RNN	92.77%
Howard and Ruder (2018)	LSTM	ULMFiT	95.40%

Table 2: *DL models for the IMDB sentiment classification task*

The IMDB Large Movie Dataset classification task has already been solved by a variety of high-performance models, especially during the last 4 years the accuracy of the task has been improved regularly. The currently best architecture was set up by Howard and Ruder (2018) with their ULMFiT model and achieves an accuracy of 95.40% in classifying the IMDB test data. The 10 most powerful DL architectures are listed in Table 2. Within these models, it is noticeable that LSTMs were used disproportionately (6 out of 10). Also, Merity et al (2017) describe these architectures as particularly advantageous for language modelling tasks, as LSTMs are more resistant to the fundamental deep learning problem of the vanishing gradient than other architectures. In addition, Johnson and Zhang (2015) also demonstrated the efficient use of CNNs for sentiment classification. Although they do not match the accuracy of the best LSTM models, they are convincing due to their competitive classification rates and comparatively low computational effort. However, LSTMs seem to be more promising in setting a new accuracy high score. The literature review also shows that the implementation of unsupervised elements, especially for pre-training, has positive effects on the performance of deep learning models (8 out of 10 models contained unsupervised learning structures). Nevertheless, due to the question regarding the influences of individual hyperparameters on the overall classification performance, the implementation of unsupervised pre-training is superfluous.

3.3 Model

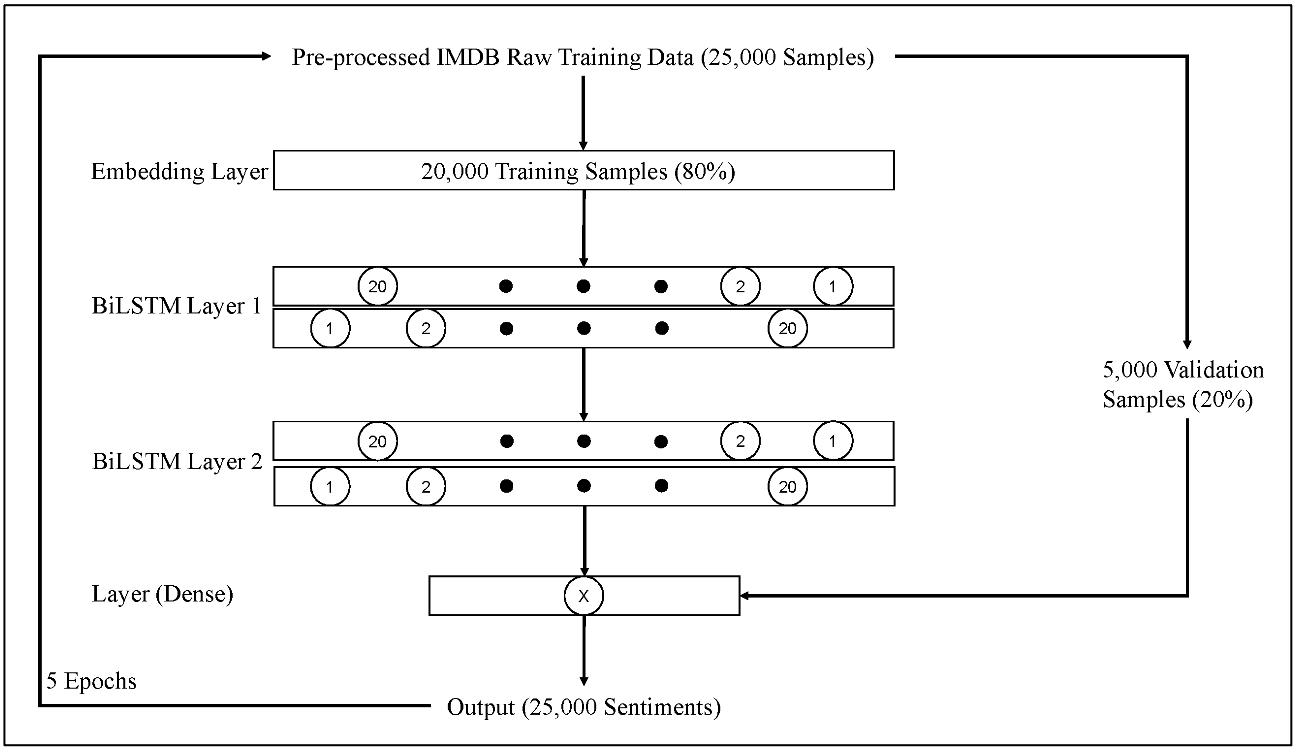

Figure 1: The LSTM model

As Figure 1 shows, our LSTM model has a bidirectional architecture, similar to Howard and Ruder (2018) model, but it initially contains only 2 BiLSTM layers and 20 units per layer and direction. After the initial embedding layer which is used for training the word embeddings, both BiLSTM layers are utilized for learning representations. The final, fully-connected dense layer executes the binary classification of the 25,000 training (respectively test) samples with a sigmoid function. As an optimizer the "RMSprop" algorithm (Hinton et al, 2012) is used, as a loss function a binary cross entropy. The number of words used as features is 10,000, and the maximum review length is 500 words. The model is trained for 5 epochs (which is a good number of epochs compared to the results of Hong and Fang (2015) for highly regularized LSTMs) with a batch size of 100, the validation split is 20% (5,000 samples, respectively).

In this model, the following hyperparameters are now to be varied to examine their single impact on the correct classification rate: The number of words considered as features, the sequence length of the comments, the proportion of validation data, the use of pre-trained GloVe word embeddings, the number of hidden BiLSTM layers, the number of units per hidden layer, the dropout and recurrent dropout rates (for preventing overfitting), and the size of the data batches (during training, the training data is divided into batches of a fixed size which are given successively through the network; the weights of the network are updated after every batch). For each hyperparameter, a specific value is set and a variant is selected that suggests a greater learning performance. Within the experiment, only one hyperparameter is chosen into its variant value at the same time. The other hyperparameters stay at their default value. The selected values are summarized in table 3.

The hyperparameters "validation data" and "batch size" were chosen lower in the variant since a larger amount of training data as well as smaller batches suggest a better classification performance. Since an adaptation of the network parameters takes place after each batch, smaller batches mean a higher number of such adjustments and thus deeper learning processes. For all other values, however, a stronger performance is assumed if the values are higher. The values are changed separately while the other hyperparameters maintain their default configuration. The determined values are then compared with the global default variant using the validation data performance in order to show their single impacts on the network performance. In this way, 8 comparison pairs are created (1 for each hyperparameter). If a hyperparameter variation has a positive effect on the validation performance, it will be transformed into a final model which will be compared to the default configuration for investigating whether the effects on accuracy can be cumulated to a high-performing model. The hyperparameter "dropout" is tested for preventing overfitting during training. At the same time, the machine times are observed. The computations are accomplished with Amazon Web Services (m4.2xlarge, 32GB).

3.4 Results

Model	Default	Variant	Train Acc	Train Loss	Val Acc	Val Loss	Machine Time
Standard	——	——	94.60%	0.1541	87.84%	0.2905	18 min 20 sec
Max features	10,000	20,000	95.89%	0.1258	88.68%	0.3106	18 min 7 sec
Max len	500	1,000	95.49%	0.1326	88.66%	0.2947	18 min 1 sec
Val split	0.2	0.1	95.18%	0.1395	88.48%	0.2931	19 min 26 sec
GloVe	no	Yes	50.23%	0.6932	50.14%	0.6930	16 min
Units / Layer	20	100	94.52%	0.1574	86.12%	0.2992	44 min 17 sec
Layer	2	3	94.98%	0.1421	87.74%	0.3195	26 min 31 sec
Dropout	no	Yes	91.65%	0.2275	86.74%	0.3513	21 min 28 sec
Batch size	100	50	95.10%	0.1431	88.90%	0.2916	28 min 16 sec

Table 3: *Training and validation results of the default configuration and the variants*

Without any hyperparameter variation, the default model reaches 94.60% training and 87.84% validation accuracy. The values of the loss function were 0.1541 for training and 0.2905 for validation. Due to overfitting in unregularized BiLSTMs, this value is already reached during the 2nd training epoch. Nevertheless, our model performs on a quite respectful level since there is no pre-training integrated. The training session required 18 minutes and 20 seconds.

The increase of word features (10,000 to 20,000 words) provided 95.89% training and 88.68% validation accuracy (loss function: 0.1258 resp. 0.3106) which means an increase of 0.84% in validation performance compared to the default model. This result was reached in the 2nd epoch as well, another rapid overfitting was observed. The training required 18 minutes and 7 seconds which was surprisingly less than the default model. Since the accuracy rate was higher, this hyperparameter variant was implemented in the final model.

The increase of the maximum sample length (500 to 1,000) also improved the performance (95.49% for training and 88.66% for validation accuracy, 0.1326 resp. 0.2947 for the loss function values), this time an increase of 0.82% in validation accuracy compared to the default model was observed. Not surprisingly, the 2nd training epoch performed best, this variation needed 18 minutes and 1 second training time. This variant was also implemented into the final model.

Changing the ratio of training and validation data from 80:20 to 90:10 resulted in a further increase in validation accuracy to 88.48% (+ 0.64%) which was already achieved during the 2nd epoch (training accuracy: 95.18%; loss function values were 0.1395 for training resp. 0.2931 for validation). Subsequently, overfitting could be observed again. This was accompanied by an increase in computing time to 19 minutes and 26 seconds. Since the accuracy increased due to the greater amount of training data, the final model will also be trained with the higher number of samples.

The use of pre-trained word embeddings from the GloVe database caused a massive loss of accuracy. While computing time was clearly the shortest at precisely 16 minutes, a training accuracy of 50.23% and a validation accuracy of only 50.14% could be achieved (loss function values: 0.6932 for training resp. 0.6930 for validation), which corresponds to a validation accuracy loss of 37.70% compared to the default model. This very poor performance is due to the lack of task-specific training of the word embeddings, which means that the values remained almost constant over the 5 epochs. The strong benefits of pre-training in literature, as found in Howard and Ruder’s (2018) model, are achieved through huge datasets used to learn the word embeddings and weights. At the same time, the word embeddings are not frozen, but constantly adapted during the learning process. While the GloVe word embeddings used here is also based on just 400,000 words, for example, WIKITEXT-103 incorporates embedding vectors for about 103,000,000 words. Thus, the word embeddings used are far from having enough information to precisely solve the specific classification task of the IMDB dataset. The pre-trained embedding vectors are therefore not integrated into the final model. However, unsupervised pre-training is indispensable for creating a particularly powerful architecture if it is carried out with very large amounts of useful information and the parameters found are then further adapted to the task.

Increasing the units per hidden layer from 20 to 100 led to a massive increase in computational time to 44 minutes and 17 seconds. This is the consequence of the higher computational effort since the additional units also process a large amount of information during the training. However, the validation accuracy fell by 1.72% to 86.12% (training: 94.52%) and the values of the loss function were worse as well (0.1574 for training and 0.2992 for validation). This result is particularly surprising given the fact that the most powerful LSTM models from the literature have clearly greater capacities. However, the performance can not be explained by overfitting, since the training data were not classified very well and the best validation performance was not achieved until the 5th epoch. This result indicates additional influences between different hyperparameters, which go beyond separate variations of individual parameters. Due to the inadequate outcome of this study, the final model does not require an increase in the number of units as the higher storage capacity should lead to an increase in classification performance, which was clearly missed.

The integration of a third BiLSTM layer, similar to Howard and Ruder (2018) network, also resulted in a lower validation accuracy of 87.74% (training accuracy: 94.98%) and worse values of the loss function (0.1421 for training and 0.3195 for validation), however, this difference is lower compared to the higher number of units (-0.1% vs default configuration). This ratio was reached during the 3rd epoch so overfitting can be observed another time (presumably by additional network capacity). The machine time increased to 26 minutes and 31 seconds. Although this result does not necessarily preclude the inclusion of a third hidden layer to the final model, due to the increased machine time and the simultaneous (minor) deterioration of the accuracy, the third BiLSTM layer will not be included.

Using a dropout / recurrent dropout regularization with the values 0.2 / 0.2 reduced the validation accuracy of the model by 1.1% to 86.74% (training accuracy: 91.65%) with simultaneous increase of the calculation time to 21 minutes and 28 seconds. The values of the loss function were 0.2275 for training and 0.3513 for validation. However, the dropout was introduced to avoid overfitting and thus increase the stability of the model. Since the top value was reached in the fifth epoch, the dropout was successful so that the regularization is to be evaluated advantageously and integrated into the final model.

The use of a smaller batch size (50 versus 100 samples) brought the highest validation accuracy gain of a single changed hyperparameter (1.06% to 88.90%). The training performance was 95.10% and the values of the loss function were 0.1431 for training and 0.2916 for validation. It is also positive that the validation quota could be reached twice (epoch 2 and 3) before overfitting begins. In this case, the model benefits from a higher number of parameter adjustments regarding the smaller batch size. However, machine time was quite high at 28 minutes and 16 seconds, due to the smaller denomination of the training data. By increasing the performance, the final model will be trained with smaller batches as well.

Model	Train Acc	Train Loss	Val Acc	Val Loss	Test Acc	Test Loss	Machine Time
Standard model	94.60%	0.1541	87.84%	0.2905	87.01%	0.3652	18 min 20 sec
Max features	95.89%	0.1258	88.68%	0.3106	86.54%	0.3603	18 min 7 sec
Max len	95.49%	0.1326	88.66%	0.2947	86.87%	0,3301	18 min 1 sec
Val split	95.18%	0.1395	88.48%	0.2931	87.60%	0,3361	19 min 26 sec
GloVe	50.23%	0.6932	50.14%	0.6930	52.81%	1.2817	16 min
Units / Layer	94.52%	0.1574	86.12%	0.2992	86.04%	0.3741	44 min 17 sec
Layer	94.98%	0.1421	87.74%	0.3195	85.90%	0.3415	26 min 31 sec
Dropout	91.65%	0.2275	86.74%	0.3513	85.53%	0.3599	21 min 28 sec
Batch size	95.10%	0.1431	88.90%	0.2916	87.51%	0.3534	28 min 16 sec
Final model	93.01%	0.1948	88.36%	0.3042	87.46%	0.3779	83 min 28 sec

Table 4: *Results of the hyperparameter variants incl. the final model and test data performances*

On the basis of the discussed validation results of the hyperparameter variations, the default configuration should now be modified seeking for a more powerful final model. A total of 5 single hyperparameter variants could be identified as well-working, including the higher number of words considered as features, the larger comment length, the use of smaller batch size, the greater amount of training data, and the integration of dropout regularization to avoid overfitting (as the validation results showed, overfitting in LSTMs is a big issue to deal with). The tested pre-trained GloVe word embeddings, on the other hand, could not be taken into account due to the massive loss of accuracy. Also, the implementation of additional layers and units could not improve the network.

The training of the final model was highly more computationally intensive than the variants of individual hyperparameters (83 minutes and 28 seconds). This observation is not surprising due to the observed calculation times of the individual variations, the computational effort of the individual hyperparameters just adds up in the final model. The accuracy, however, reached 93.10% for training and 88.36% for validation which corresponds to an increase of 0.52% in validation performance compared to the default configuration (reached in the 5th epoch so the dropout implementation was successful in avoiding overfitting). But, at the same time, it is highly noteworthy that the performance in the validation data is worse than in the variants of the individual positive-acting hyperparameters which were set to improve the network accuracy (dropout regularization was implemented to avoid overfitting). This means that LSTM hyperparameters do not just work on their own but seem to interact with the other hyperparameter settings. In fact, this experimental design is well-suited for understanding the effects of the various hyperparameters on the network in general, but it is not optimal for finding the strongest setting within an LSTM. Nonetheless, the final model has achieved a higher validation performance than the already well-performing default configuration.

For further evaluating the variants and the final model compared to the default configuration, the test dataset of the IMDB dataset was classified. For this, the raw test data was preprocessed as well as the training data (vectorization and word embeddings learned by embedding layer). The default configuration achieved 87.01% test accuracy while the created final model achieved a comparatively stronger accuracy of 87.46%. Compared with the single variants, the separate variation of the validation split and the batch size were even outperforming the final model while the variant of the validation split reached the highest test accuracy (87.60%) with a machine time of 19 minutes and 26 seconds. To summarize, the test classification performance could be increased by 0.45% (resp. approximately 113 additionally correctly classified comments) through varying the 5 hyperparameters classified as positive and by 0.59% (resp. approximately 148 additionally correctly classified comments) through the separate variation of the validation split compared to the default model. The 0.45% increase in classification performance represents an improvement associated with highly increased computational time requirements while the higher increase of 0.59% could be reached with only a small gain of machine time.

4 Discussion / Conclusion

The aim of this work was to investigate the impacts of single hyperparameter variants within an LSTM network to perform the IMDB Large Movie Dataset SA task. For this purpose, an LSTM network based on the task-specific DL models from the literature of recent years was created. A total of 8 hyperparameters contained in this network were separately varied and compared with the default configuration by their validation performance. In this way, 5 hyperparameters (maximum number of words taken into account as characteristics, maximum comment length, dropout regularization, use of a larger training dataset, and the use of a smaller batch size) could be demonstrated as positive influences while implementing additional hidden layers, additional units per layer and pre-trained GloVe word embeddings could not achieve any positive effects. The variants which improve the validation accuracy were then transformed into a final model to see whether the impacts of the separate hyperparameters could be added. While the validation data performance of the final model was higher than the default model, some single variants outperformed the final model so the effects of single variants were not able to be cumulated. In addition, comparing the default configuration, the separate variants and the final model based on test data accuracy, the default model achieved 87.01% with a machine time of 18 minutes and 20 seconds, while the final model achieved 87.46% at a clearly higher computation time of 83 minutes and 28 seconds. At the same time, the separate variants of the validation split and the batch size even outperformed the final model due to test accuracy and machine time (with the separate variation of the validation split as the overall best configuration performing a test accuracy of 87.60%). In this way, the already precisely classifying default configuration could be increased by a further 0.45% (approximately 113 additional comments) through creating the final model and even 0.59% with a separate variant. In fact, the separate influences of the hyperparameter variants on accuracy could not be cumulated but, at the same time, the machine time did.

Looking at the separate variants, it is striking that the better performance was not achieved by increasing the network capacity (additional layers or units per layer) but the consideration of a larger number of features, longer comments and a larger number of training samples were able to raise accuracy, even the use of smaller batch sizes contributed to a stronger performance. In particular, the network was able to benefit from larger amounts of data and a greater number of iteration steps. At the same time, the results for the variants that result in an increase in capacity (number of units / layers) are surprisingly negative and should not be implemented as a single variant in BiLSTMs which already have a large network capacity. Overall, the results indicate interactions between the various hyperparameters that can not be observed in this experimental setup with separate variants. This is supported by the current literature who use a much higher capacity than the model configured here. Accordingly, higher capacities should definitely not be excluded from the construction of DL models for performing SA, rather such changes should be examined together with other hyperparameter variants in order to possibly further increase the classification performance.

In spite of the lower classification performance compared to the currently best models, it was possible to clearly demonstrate how the single hyperparameters of an LSTM model influence the performance of the overall architecture. In comparison to the architecture by Howard and Ruder (2018), the performances of the model used here are significantly lower. This is due to the fact that Howard and Ruder (2018) use a huge unlabeled dataset for efficiently pre-training their network. In addition, they combined different hyperparameters for increasing the network capacity (additional BiLSTM layers and more units per layer) while preventing overfitting with dropout regularization. The results of their ULMFiT model indicate interactions between the different hyperparameters as our experiment with separate variants did.

While the separate influences of the hyperparameter variants on overall accuracy could be shown precisely, the experiment has to be limited due to the fact that the validation split during the training epochs has been set randomly so small variances due to different validation samples can not be excluded. Though, since every configuration is trained for 5 epochs with 5 different validation splits, the risk of a variance at the validation results is negligible. Furthermore, no effects between the individual parameter variants were analyzed. These effects could be observed by the surprisingly poor classification results for those variants which increase network capacity and the accuracy of the final model compared to different single hyperparameter variants. In this respect, the investigation is limited, and we would like to encourage further research in the field of hyperparameter variants in LSTM networks. In particular, studies that use this paper as a first step to understand the single hyperparameter effects on the network and go on investigating combinations of variants (i.e. using a fractional factorial design (see, e. g., Gunst and Mason (2009)) can further advance the currently still fragile state of research. We believe that a deeper understanding of hyperparameter influences in LSTMs will definitely help to outperform the current IMDB Large Movie Dataset highscore with new and innovative LSTM models.

References

Aggarwal U, Aggarwal G (2017) Sentiment Analysis: A Survey. International Journal of Computer Sciences and Engineering 5(5):222–225, SivaKumar K (ed).

Ain TQ, Ali M, Riaz A, Noureen A, Kamranz M, Hayat B, Rehman A (2017) Sentiment Analysis Using Deep Learning Techniques: A Review. International Journal of Advanced Computer Science and Applications 8(6):424–433. DOI: 10. 14569/IJACSA.2017.080657.

Araque O, Corcuera-Platas I, Sánchez-Rada JF, Iglesias CA (2017) Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications 77:236–246. DOI: 10.1016/j.eswa.2017.02.002.

Arel I, Rose DC, Karnowski TP (2010) Deep Machine Learning - A New Frontier in Artificial Intelligence Research. IEEE Computational Intelligence Magazine 5(4):13–18. DOI: 10.1109/MCI.2010.938364.

Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 5(2):157–166. DOI: 10. 1109/72.279181.

Brychcín T, Habernal I (2013) Unsupervised Improving of Sentiment Analysis Using Global Target Context. Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP) 2013, pp. 122–128, Mitkov R, Angelova G, Bontcheva K (eds), INCOMA Ltd. Shoumen.

Cambria E, Schuller B, Xia Y, Havasi C (2013) New Avenues in Opinion Mining and Sentiment Analysis. IEEE Intelligent Systems 28(2):15–21. DOI: 10.1109/MIS. 2013.30.

Dai AM, Le QV (2015) Semi-supervised Sequence Learning. In: Advances in Neural Information Processing Systems 28, Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds). Curran Associates, Inc., pp. 3079–3087. URL: http://papers. nips.cc/paper/5949-semi-supervised-sequence-learning.pdf.

Deng L (2011) An Overview of Deep-Structured Learning for Information Processing. Proceedings of the Asian-Pacific Signal & Information Processing Annual Summit & Conference (APSIPA-ASC) 2011, pp. 1–14, EI Compendex, Xi’an, China. URL: http://www.apsipa.org/proceedings_2011/pdf/ APSIPA336.pdf.

Deng L (2014) A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Transactions on Signal and Information Processing 3:1–29, Cambridge University Press, Cambridge. DOI: 10.1017/ATSIP.2013.9.

Dieng AB, Wang C, Gao J, Paisley J (2016) TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency. arXiv 2016(11):1–13. URL: http:// arxiv.org/pdf/1611.01702v2.

Gal Y, Ghahramani Z (2015) A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. arXiv 2015(12):1–14. URL: http://arxiv.org/pdf/1512.05287v5.

Gers FA, Schmidhuber J, Cummins F (1999) Learning to forget: continual prediction with LSTM. 9th International Conference on Artificial Neural Networks: ICANN ’99, pp. 850–855, Edinburgh, United Kingdom. DOI: 10.1049/cp:19991218.

Google (2013) word2vec. URL: https://code.google.com/archive/p/ word2vec/ [accessed 02.11.2018].

Gray S, Radford A, Kingma DP (2017) Gpu kernels for block-sparse weights. arXiv 2017(11):1–12. URL: http://arxiv.org/pdf/1711.09224.

Gunst RF, Mason RL (2009) Fractional factorial design. Computational Statistics 1(2):234–244. DOI: 10.1002/wics.27.

Hinton G, Srivastava N, Swersky K (2012) Neural Networks for Machine Learning. video lectures Coursera, pp. 1–31.

Hochreiter S (1991) Untersuchungen zu dynamischen Neuronalen Netzen. Diploma thesis, Institut für Informatik, München, Technische Universität München.

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computation 9(8):1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Hong J, Fang M (2015) Sentiment Analysis with Deeply Learned Distributed Representations of Variable Length Texts. Stanford University Technology Report, pp. 1–9. URL: https://cs224d.stanford.edu/reports/HongJames.pdf.

Howard J, Ruder S (2018) Universal Language Model Fine-tuning for Text Classification. arXiv 2018(05):1–12. URL: http://arxiv.org/pdf/1801.06146v5.

Jain AK, Mao J, Mohiuddin KM (1996) Artificial neural networks: A tutorial. Computer 29(3):31–44. DOI: 10.1109/2.485891.

Johnson R, Zhang T (2015) Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding. Advances in Neural Information Processing Systems 28 (NIPS 2015), pp. 919–927, Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds), Montreal, Canada.

Johnson R, Zhang T (2016) Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings. arXiv 2016(05):1–9. URL: http://arxiv.org/ pdf/1602.02373v2.

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning (ICML) 32:1188–1196, Xing EP, Jebara T (eds), JMLR.org, Beiing, China.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. DOI: 10.1038/nature14539.

Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies 1:142–150, Association for Computational Linguistics, Portland, Oregon, USA.

Maynard D, Funk A (2011) Automatic Detection of Political Opinions in Tweets. Extended Semantic Web Conference (ESWC) 2011: ESWC 2011 Workshops, pp. 88–99, Garcia-Castro R, Fensel D, Antoniou G (eds), Heraklion, Greece. DOI: 10.1007/ 978-3-642-25953-1.

Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5(4):1093–1113. ISSN: 2090-4479, DOI: 10.1016/j.asej.2014.04.011.

Merity S, Keskar NS, Socher R (2017) Regularizing and Optimizing LSTM Language Models. arXiv 2017(08):1–10. URL: http://arxiv.org/pdf/1708. 02182v1.

Mesnil G, Mikolov T, Ranzato M, Bengio Y (2014) Ensemble of Generative and Discriminative Techniques for Sentiment Analysis of Movie Reviews. arXiv 2014(05):1–5. URL: http://arxiv.org/pdf/1412.5335v7.

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of Word Representations in Vector Space. arXiv 2013(09):1–12. URL: http://arxiv.org/pdf/ 1301.3781v3.

Miyato T, Dai AM, Goodfellow I (2016) Adversarial Training Methods for SemiSupervised Text Classification. arXiv 2016(05):1–11. URL: http://arxiv.org/ pdf/1605.07725v3.

Narayanan V, Arora I, Bhatia A (2013) Fast and accurate sentiment classification using an enhanced Naive Bayes model. International Conference on Intelligent Data Engineering and Automated Learning (IDEAL) 2013, pp. 194–201, Yin H, Tang K, Gao Y, Klawonn F, Lee M, Weise T, Li B, Yao X (eds), Hefei, China.

Pennington J, Socher R, Manning C (2014) Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, Moschitti A, Pang B, Daelemans W (eds), Association for Computer Linguistics, Doha, Qatar. DOI: 10.3115/v1/d14- 1162.

Radford A, Jozefowicz R, Sutskever I (2017) Learning to Generate Reviews and Discovering Sentiment. arXiv 2017(04):1–9. URL: http://arxiv.org/pdf/ 1704.01444v2.

Ravi K, Ravi V (2015) A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems 89:14–46. ISSN: 0950-7051, DOI: 10.1016/j.knosys.2015.06.015.

Rui W, Liu J, Jan Y (2016) Unsupervised feature selection for text classification via word embedding. Proceedings of the 2016 IEEE International Conference on Big Data Analysis (ICBDA), pp. 1–5, Hangzhou, China.

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by backpropagating errors. Nature 323(6088):533–536. DOI: 10.1038/323533a0. Schmidhuber J (2015) Deep Learning in Neural Networks: An Overview. Neural Networks 61:85–117. ISSN: 0893-6080, DOI: 10.1016/j.neunet.2014.09.003. Schuster M, Paliwal KK (1997) Bidirectional Recurrent Neural Networks. IEEE Transactions on Signal Processing 45(11):2673–2681. DOI: 10.1109/78.650093.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1):1929–1958. DOI: 10.5555/2627435.2670313.

Stanford (2014) GloVe. URL: https://nlp.stanford.edu/projects/glove [accessed 02.11.2018].

Turney PD (2002) Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th annual meeting on association for computational linguistics (ACL), pp. 417–424, Isabelle P, Charniak E, Lin D (eds), Association for Computer Linguistics, Philadelphia, Pennsylvania, USA. DOI: 10.3115/1073083.1073153.

Wang S, Manning CD (2012) Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of the 50th annual meeting of the Association for Computational Linguistics 2:90–94, Li H, Lin CY, Osborne M, Lee GG, Park JC (eds), Association for Computer Linguistics, Jeju Island, Korea. URL: https:// www.aclweb.org/anthology/P12-2018.

Xu W, Sun H, Deng C, Tan Y (2017) Variational Autoencoder for Semi-Supervised Text Classification. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), pp. 3358–3364, San Francisco, California, USA. URL: https:// aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14299.

Yadollahi A, Shahraki AG, Zaiane OR (2017) Current State of Text Sentiment Analysis from Opinion to Emotion Mining. ACM Computing Surveys 50(2):1–33. DOI: 10. 1145/3057270.

Yang Y, Pedersen JO (1997) A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, ICML ’97, pp. 412–420. ISBN: 15-5860-486-3, DOI: 10.5555/645526.657137.

Zhang L, Wang S, Liu B (2018a) Deep learning for sentiment analysis: A survey. Data Mining and Knowledge Discovery 8(4):e1253. DOI: 10.1002/widm.1253.

Zhang Q, Yang LT, Chen Z, Li P (2018b) A survey on deep learning for big data. Information Fusion 42:146–157. ISSN: 1566-2535, DOI: 10.1016/j.inffus.2017.10. 006.

Zhang X, Zhao J, LeCun Y (2015) Text Understanding from Scratch. arXiv 2015(02):1–9. URL: http://arxiv.org/pdf/1502.01710v5.

Zhou ZH, Feng J (2017) Deep Forest. arXiv 2017(02):1–34. URL: http://arxiv. org/pdf/1702.08835v3.

III Call Me Maybe: Methods and Practical Implementation of Artificial intelligence in Call Center Arrivals' Forecasting

Authors: Albrecht, Tobias; Rausch, Theresa Maria; Derra, Nicholas Daniel

Published in: Journal of Business Research (2021), Vol. 123, pp. 267-278,
DOI: 10.1016/j.jbusres.2020.09.033

Abstract: Machine learning (ML) techniques within the artificial intelligence (AI) paradigm are radically transforming organizational decision-making and businesses' interactions with external stakeholders. However, in time series forecasting for call center management, there is a substantial gap between the potential and actual use of AI-driven methods. This study investigates the capabilities of ML models for intra-daily call center arrivals' forecasting with respect to prediction accuracy and practicability. We analyze two datasets of an online retailer's customer support and complaints queue comprising half-hourly observations over 174.5 weeks. We compare practically relevant ML approaches and the most commonly used time series models via cross-validation with an expanding rolling window. Our findings indicate that the random forest (RF) algorithm yields the best prediction performances. Based on these results, a methodological walk-through example of a comprehensive model selection process based on cross-validation with an expanding rolling window is provided to encourage implementation in individual practical settings.

Keywords: Artificial Intelligence, Machine Learning, Call Center Forecasting, Predictive Analytics

1 Introduction

Artificial Intelligence (AI) is considered the key enabler for the digitalization of a company in a broad spectrum of areas (Brynjolfsson and McAfee, 2017). Today, in the course of increasing availability of data, affordable as well as scalable processing power, and the development of advanced machine learning (ML) techniques, AI is about to radically transform how firms make decisions (Agrawal et al., 2019). It is expected to facilitate the internal decision-making processes of organizations by making it smarter, faster, and overall more efficient. To benefit from this potential competitive advantage, companies need to identify existing domain problems, find compatible AI solutions, and put an implementation concept into practice (Overgoor et al., 2019). This requires a thorough understanding of the task-specific capabilities and feasibility of AI methods like ML. So far, a lack of expertise in this area paired with a high level of perceived complexity is often preventing the implementation of ML solutions in practical settings (Tambe et al., 2019).

Particularly the interaction of companies with external stakeholders, such as customers, is about to be fundamentally transformed by AI (Kaplan and Haenlein, 2019). Fueled by an almost unlimited flow of information about their customers, service-oriented companies in particular, can capitalize on AI-driven decision support. Based on latent characteristics and previous customer behavior, ML techniques can predict future interactions (Wedel and Kannan, 2016). One of the most prevalent and dominant points of interaction between many organizations and their customers and therefore, critical for managing customer experience, are call centers or customer service centers (Whiting and Donthu, 2006). To constantly provide high service quality in the form of short waiting times at this touchpoint, a sufficient number of call center agents is needed (Atlason et al., 2008). Consequently, the process of predicting call arrival volumes and deciding on the required staffing level is a critical success factor in this area. In this connection, the capabilities of innovative ML techniques promise more flexible and precise predictions and thus, the possibility of enhanced organizational planning and better customer service.

Despite the encouraging prospects for service improvement and cost savings, a comprehensive understanding of the potential of ML models for creating additional value in call center forecasting is lacking. In order to gain more profound insights into the performance and practicability of such AI-driven models in this context, research comprising a methodological perspective with a focus on prediction accuracy as well as a practical angle on the selection and implementation of models is required.

This study proposes a two-step approach that, in the first step, provides a thorough understanding of the forecast accuracy of ML methods in call arrival forecasting and, in the second step, makes the underlying process of method comparison and selection feasible to decision-makers in practice. Specifically, we conduct an in-depth analysis of the forecast accuracy of viable ML models based on the call arrival data of a real German online retailer. Using two different datasets, i.e., the customer support and customer complaints queue of the corresponding call center, we perform a comprehensive method comparison opposing selected ML models to the three most commonly used time series models in this field. In the second step, we provide a methodological walk-through example for a valid model selection process based on cross-validation with an expanding rolling window. We illustrate the practical implementation of the process in a programming environment that is accessible to non-machine learning experts and practitioners using the random forest (RF) algorithm as the best-performing model for an in-depth example.

This paper therefore aims to present a starting point for shifting traditional call center forecasting towards a paradigm drawing on AI-driven methods. By systematically evaluating the predictive potential of ML models in comparison to commonly used methods, new sophisticated but yet applicable models for practical use are identified. In a business setting, following the explicated implementation in a reproducible programming environment is supposed to empower practitioners to develop insights on the use of ML for forecasting call center arrivals in individual data environments.

The remainder of this paper is structured as follows: In Section 2, we present the theoretical background of AI-driven methods in customer analytics and review the state of research in call center arrivals' forecasting, before adequate ML models for this field are introduced. Subsequently, in Section 3, we describe the methodology of our research. In Section 4, we present the results of our analysis for two different customer service channels and in Section 5, we discuss the theoretical contribution and the limitations of our study. We then illustrate the implementation of the best-performing RF model by giving a detailed code and walk-through example and demonstrate methodological as well as practical implications of the proposed approach. Finally, Section 7 presents a summary and concluding remarks.

2 Theoretical Background

2.1 Artificial Intelligence in Customer Analytics

For businesses, the strategic challenge of understanding and managing customer relationships is becoming increasingly important and demanding at the same time. While organizations today have easy access to enormous amounts of data about their customers, extracting relevant information to support prospective decision-making and thus, standing out from competitors in the long term has become a difficult hurdle to overcome for many of them (Kitchens et al., 2018). In the course of these changing market dynamics, businesses slowly realize the potential of AI in predictive analytics to enhance organizational decision-making by forecasting customer-related data and, therefore, effectively infer their future behavior (Huang and Rust, 2018). Predictive analytics techniques generally comprise statistical models and other empirical methods aimed at creating predictions as well as approaches for assessing the quality of those predictions in practice (Shmueli and Koppius, 2011). More recently, ML as a subset of AI has been added to the domains contributing effectively to business prediction problems as they provide a way to handle complex problems by forecasting future data based on more extensive sets of historical values (Chen et al., 2012). In literature, innovative ML approaches have been successfully applied to various customer analytics problems such as customer preferences analysis (Yang and Allenby, 2003), customer retention (Donkers et al., 2003), and customer profitability management (Reinartz et al., 2005).

However, so far, the practical implementation of ML models in predictive customer analytics is limited (Wedel and Kannan, 2016). Drawing on the early distinction between forecasting methods and forecasting systems proposed by Harrison and Stevens (1976) may explain this slow adoption. While the former transforms input data into output information in a mere technical way, the latter in addition includes the people concerned with the forecast and the resulting actions. Based on that view, the evaluation and selection of a forecasting system explicitly go well beyond the accuracy of its prediction model and includes meaning and usability in practical implementation. In terms of this applicability, many ML approaches still exhibit shortcomings as they do not provide much insight into the influence and dynamics of the underlying factors that lead to the prediction results (Martens et al., 2011; Najafabadi et al., 2015). Due to this lack of comprehensibility and interpretability, many ML techniques are commonly considered as black box models (Doshi-Velez and Kim, 2017; Guidotti et al., 2019). Moreover, such models are frequently perceived as complex regarding the implementation. A high number of hyperparameters gives models the flexibility of adapting to a multitude of business problems but, at the same time, makes it complex for the user to build and optimize the ML algorithm. This especially applies to the broad class of artificial neural networks (Bergstra et al., 2011; Paliwal and Kumar, 2009). For the above reasons, other categories of ML approaches come into the focus for practical use.

Widely established methods like support vector machines and Bayesian approaches promise ease of use while maintaining good performance levels on data sets characterized by moderate complexity (Arora et al., 1998; Verbeke et al., 2011). Tree-based models, and in this field especially ensemble learning methods like RF and gradient boosting, gained popularity for their robustness and flexibility in modeling input–output relationships of various types and volumes of highly complex data (Fang et al., 2016; Lemmens and Croux, 2006). Research found them to provide high prediction accuracy as well as descriptive results in diverse customer analytics problems such as churn analysis (Burez and van den Poel, 2009) and credit risk management (Fantazzini and Figini, 2009). In addition, a small number of hyperparameters makes their construction, customization, and optimization more manageable and comprehensible (Breiman, 2001).

2.2 Call Center Arrivals' Forecasting

In recent years, the role of call centers has fundamentally changed in many organizations and across all industries. While call centers previously only had an information function which did not exceed simple order processes, nowadays, more and more complex tasks and customer demands need to be fulfilled across multiple communication channels using modern digital technology (Aksin et al., 2007). However, instead of experiencing declining importance in the course of this transformational process, the opposite is the case. Call centers are increasingly transforming into customer interaction centers that form the basis for an efficient and value-oriented customer relationship management (Gans et al., 2003). They constitute an interface to the customer and provide complex services, while, at the same time, giving companies the opportunity of collecting large amounts of otherwise inaccessible customer data (Ibrahim et al., 2016). Subsequently, it is possible to anticipate customer needs and behavior through data analysis and forecasting techniques (Taylor, 2008). Based on those insights, internal processes and external expectations can be aligned to optimize business performance as well as customer experience.

One of the most important internal processes in call centers is the staffing of agents as customer service representatives who directly handle tasks such as order taking, complaint resolution, information, and help desk functions as well as after-sales and supplementary services (Dean, 2007; Koole and Pot, 2005). While overstaffing results in high personnel costs, understaffing can lead to extended waiting times for customers and consequently causes lower perceived service quality, decreasing customer satisfaction, and a lack of customer loyalty (Brady and Cronin, 2001). To determine the optimal staffing level, an accurate and robust prediction of call arrival volumes based on historical data is needed (Weinberg et al., 2007). Hence, the search for appropriate forecasting methods is the focus of scholars and practitioners alike. However, preceding literature so far mainly investigated traditional statistical models without taking into account the substantial changes coming along with the transforming role of call centers in organizations (Gans et al., 2003). Today, the increasing volume and variety of data through a multitude of channels as well as the necessity of realtime analysis and predictions call for more flexible and powerful methods.

Call center arrivals are count data limited to non-negative integers. Such discrete data are frequently estimated as Poisson arrival rates (see e.g., Cezik and L'Ecuyer (2008); Taylor (2012)). However, with arrival rates not being easily predictable, other researchers point out ascertained randomness of arrivals in real call centers (see e.g., Aksin et al. (2007); Shen and Huang (2008)). Generally, call center arrivals data exhibit specific characteristics and challenges that affect the forecasting process. Firstly, an important feature of call arrival rates is their time dependence that typically manifests itself in intraday (or subdaily), daily, weekly, monthly, or yearly seasonalities as repeating patterns in the arrival counts (Ibrahim et al., 2016). Secondly, the data are often high-dimensional and sensitive to contextual factors. Hence, additional information like holidays, promotional activities, and other special events may improve model predictions by indicating variations and outliers in the data (Barrow and Kourentzes, 2018). Thirdly, procedural characteristics are affecting the forecasting of incoming calls, such as (a) the specific call type (e.g., complaints, order taking, or after-sales service) associated with the forecast, (b) the length of forecast intervals, which may commonly range from monthly or weekly to daily or even sub-daily (i.e., hourly, half-hourly etc.) time spans, and (c) the period between the creation of the forecast and the first interval of the prediction, i.e., the lead time. Lead time is an organizational parameter resulting from staffing regulations and is assumed to strongly affect forecast accuracy as more recent data promise better predictions (Aksin et al., 2007; Rausch and Albrecht, 2020). Given these properties, the need for methods with high modeling flexibility, while being able to handle time dependencies and complex data structures, becomes evident.

With time dependence often being considered as one of the predominant features of the call arrival data, common forecasting techniques in research mostly originate from the field of time series analysis with call arrivals being a set of contiguous, dependent observations $y(t) = 0, 1, 2, \ldots$, each one being recorded sequentially at time t (Box et al., 2015). The most widely investigated and compared methods in literature include simple stationary time series models as well as the nonstationary seasonal autoregressive integrated moving average (ARIMA) model (Box and Jenkins, 1970), Holt Winters' exponential smoothing models (Holt, 2004; Winters, 1960), and random walk methods (Taylor, 2008). While ARIMA and exponential smoothing provide sophisticated complementary solutions to the general forecasting problem (Hyndman and Athanasopoulos, 2018) and constitute the most commonly used approaches in call center forecasting due to their high prediction accuracy (Andrews and Cunningham, 1995; Barrow and Kourentzes, 2018; Mabert, 1985; Taylor, 2012; Thompson and Tiao, 1971), the random walk model is frequently utilized as a benchmark within literature due to its naïve forecasts and its informative value for model comparisons (Taylor, 2008). Besides, regression analysis in the form of generalized linear models (GLM), linear fixed-effects, random-effects, and mixed-effects models is implemented for call arrivals' forecasting (Avramidis et al., 2004; Ibrahim and L'Ecuyer, 2013; Nelder and Wedderburn, 1972).

In contrast, research on ML techniques in call center arrivals' forecasting is still in its infancy. Ebadi Jalal et al. (2016) first indicate time-sensitive ML models to be eligible for forecasting call volumes in call centers. To improve short-term accuracy in call arrivals' forecasting, Barrow (2016) developed a hybrid method adjusting seasonal moving average predictions by means of nonlinear artificial neural networks and found it to outperform traditional time series models like ARIMA and Holt Winters'. Moreover, ML is shown to be capable of modeling complex outliers and thus, to improve call arrival prediction accuracy and to yield better results than ARIMA and an innovation state space model (ETS) (Barrow and Kourentzes, 2018). Recently, Rausch and Albrecht (2020) investigated RF algorithm as another ML method in their comparison of novel time series and regression models for call center arrivals forecasting. RF was found to yield higher prediction accuracy for nearly all of the considered lead time constellations. Despite first promising findings and the investigation of several approaches, current research lacks a comprehensive understanding of the full capabilities of ML in call center forecasting. To close this gap, an extensive assessment of the forecast accuracy of ML models in comparison to the most commonly used methods is still to be done. However, according to comparisons of common methods conducted on call center data, the selection of the best forecasting method can ultimately be highly dependent on the characteristics of the specific prediction problem (Andrews and Cunningham, 1995; Taylor, 2008). Therefore, a feasible process of model comparison and selection needs to be established to give methodological guidelines to practitioners and to match the set of researched forecasting methods with those considered in practice. Today, although a lot of progress has been made regarding the development of advanced methods, call arrivals' forecasting in real business environments is frequently still done based on experience or ordinary stochastic models with limited predictive capabilities (Ibrahim et al., 2016).

2.3 Machine Learning Approaches

Models from the field of ML are assumed to improve call center arrivals' forecasting and extend the range of feasible methods by providing additional robustness and accuracy to predictions. As the practicability of models play a central role in this field of application, non-parametric ML algorithms, that are comprehensible and comparatively easy to implement, such as tree-based models, k-nearest neighbor (KNN) algorithm, and support vector machines, come to the fore (Coussement and van den Poel, 2008; Li et al., 2010; Singh et al., 2017).

2.3.1 Bagging: Random Forest

Tree-based methods are frequently utilized in business prediction problems since they yield desirable accuracies despite their ease of use (Breiman, 2001). In bagging, successive decision trees are grown independently from earlier trees, i.e., each tree is constructed using a bootstrap sample of the data (Breiman, 1996). A subclass of bagging methods are RFs, as proposed by Breiman (2001), which add an additional layer of randomness to bagging and change how the trees are constructed. Thereby, non-parametric the RF algorithm is one of the most widely used ML algorithms, supported by its robustness towards outliers and its moderate computation time compared with boosting and other bagging methods (Breiman, 2001).

The algorithm draws n_{tree} bootstrap samples from the training data and then grows an unpruned regression tree for each bootstrap sample by randomly sampling m_{try} of the predictors at each node and choosing the best split among them. More formally, the resulting RF is an ensemble of B trees $\{T_1(X), \ldots, T_B(X)\}$, where $X = \{x_1, \ldots, x_p\}$ is a p-dimensional vector of predictors associated with a dependent variable; the ensemble produces B outputs $\{\hat{y}_1 = T_1(X), \ldots, \hat{y}_B = T_B(X)\}$, where $\hat{y}_b, b = 1, \ldots, B$ is the prediction for the dependent variable by the bth tree (Svetnik et al., 2003). The outputs of all n_{tree} trees are aggregated to produce one final prediction $\hat{Y}$; for regression trees it is the average of the single tree predictions (Liaw and Wiener, 2002). I.e., the RF prediction is the unweighted average over the ensemble:

$$\hat{Y} = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_B\ (T_B(X))$$

To tune the hyperparameters, an estimate of the error rate based on training data can be obtained: at each bootstrap iteration, the data which is not in the bootstrap sample, i.e., the out-of-bag (OOB) data n, is predicted by using the tree grown with the bootstrap sample. Then the OOB predictions are aggregated, and the error rate is calculated (Liaw and Wiener, 2002). In each bootstrap training set, about one-third of the sample is left out, i.e., is used for OOB predictions (Breiman, 2001).

2.3.2 Boosting: Gradient Boosting Machines

In contrast to bagging, boosting constructs successive weak learners (e.g., decision trees) to produce a final strong learner. Each sequentially added weak learner intends to correct the preceding learners (Schapire, 1990). Thereby, gradient boosting (machines) fits the new predictor or learner to the residual errors made by the preceding predictors or learners and uses gradient descent to identify the errors in the previous predictions, i.e., gradient boosting allows the optimization of an arbitrary differentiable loss function (Friedman, 2001; Friedman, 2002). Formally, J_m are the number of leave and the tree partitions the input space into J_m joint regions $R_{1m}, \dots, R_{J_m m}$ and predicts a constant value in each region. γ_{im} is the multiplier chosen as an optimal value for each of the tree's regions to minimize the loss function L. Then the generic gradient tree boosting model can be defined as

$$\mathrm{F_m(x)}=\mathrm{F_{m\text{-}1}}(\mathrm{x}) + \sum_{i=1}^{J_m} \gamma_{im}\, 1(x \in R_{im}), \text{with } \gamma_{\mathrm{im}} = \arg\min_{\gamma} \sum_{x_i \in R_{im}} L(y_i, F_{m-1}(x_i) + \gamma).$$

Since gradient boosting frequently leads to overfitting, regularization techniques can be included to constrain the fitting procedure. E.g., dropout regularization – inspired by neural networks in a deep learning context – grows consecutive trees from the residual errors of a subset or sample of previous trees instead of using all previous trees (Rashmi and Gilad-Bachrach, 2015).

2.3.3 K-Nearest Neighbor

The KNN algorithm is frequently considered due to its simplicity in comparison with other ML approaches. The algorithm was first formalized by Cover and Hart (1967) for classification tasks: given an unlabeled instance, the algorithm finds a group of k most similar objects (or nearest neighbors respectively) given its features by computing the distance $d(.,.)$ (e.g., Euclidean distance) between them and further, assigns a class label which matches the class of the majority of the k neighbors. This concept can easily be extended to regression tasks where the output is the average of the k nearest neighbors, i.e.,

$$\hat{Y} = \frac{1}{K}\sum_{i=1}^{K} y_i$$

where y_i is the ith case of the nearest neighbors.

2.3.4 Support Vector Regression

Suppose we are given a space of input patterns $\mathcal{X}$, i.e., the training data, $\{(x_1, y_1), \dots, (x_k, y_k)\} \subset \mathcal{X} \times \mathbb{R}$ with y_k being the output vectors and x_k are the input vectors. The basic support vector machine is a non-probabilistic binary linear classifier and it non-linearly maps input vectors into a higher dimensional feature space in which a linear decision surface, i.e., a separating hyperplane, is constructed (Cortes and Vapnik, 1995; Vapnik, 1982). Thus, its representation of the training data as points in the feature space is separated into categories by the hyperplane and predictions of new instances are classified into those categories. The main aim in ε-support vector regression (SVR) (Vapnik, 1995) is based on the same principles but with minor differences: the function $f(x)$ should have at most ε deviation from the actual targets for the training data and simultaneously, should be as flat as possible (Smola and Schölkopf, 2004). In the linear and most basic case, f is taking the form

$$f(x) = \langle \omega, x \rangle + b \text{ with } \omega \in \mathcal{X}, b \in \mathbb{R}$$

where $\langle \cdot, \cdot \rangle$ is the dot product in the space of input patterns $\mathcal{X}$. To ensure flatness, a small ω can be obtained by a convex optimization problem:

$$\text{minimize } \frac{1}{2}\|\omega\|^2$$

$$\text{subject to } \begin{cases} y_t - \langle \omega, x_i \rangle - b \leq \varepsilon \\ \langle \omega, x_i \rangle + b - y_i \leq \varepsilon \end{cases}$$

It assumes that function $f$approximates all pairs $\langle x_i, y_i \rangle$ with ε precision. Slack variables ξ_i, ξ_i^* can cope with such otherwise infeasible constraints of the optimization problem. Moreover, kernels can be used to make SV algorithms nonlinear by transforming the data into a higher dimensional feature space (Smola and Schölkopf, 2004).

3 Methodology

3.1 Preliminary Data Analysis

We analyze call center data of a leading German online retailer for fashion that were gathered and selected iteratively and in close exchange with the local data experts and department managers. Overall, the retailer's call center comprises four different queues: customer complaints, customer support, personal consultation service, and order taking. In this paper, we investigate two datasets describing the call arrival volume of the customer support and customer complaints queue. Both are open from 7 a.m. to 10 p.m. from Monday through Saturday. The half-hourly datasets comprise 31,410 observations or 174.5 weeks of data from January 2, 2016 to May 7, 2019. One day comprises 30 observations, one week consists of 180 observations, and one year comprises 9,367.5 observations considering leap years. We exclude two weeks of data (or 360 observations) since these values are missing due to an internal system change for interval capturing.

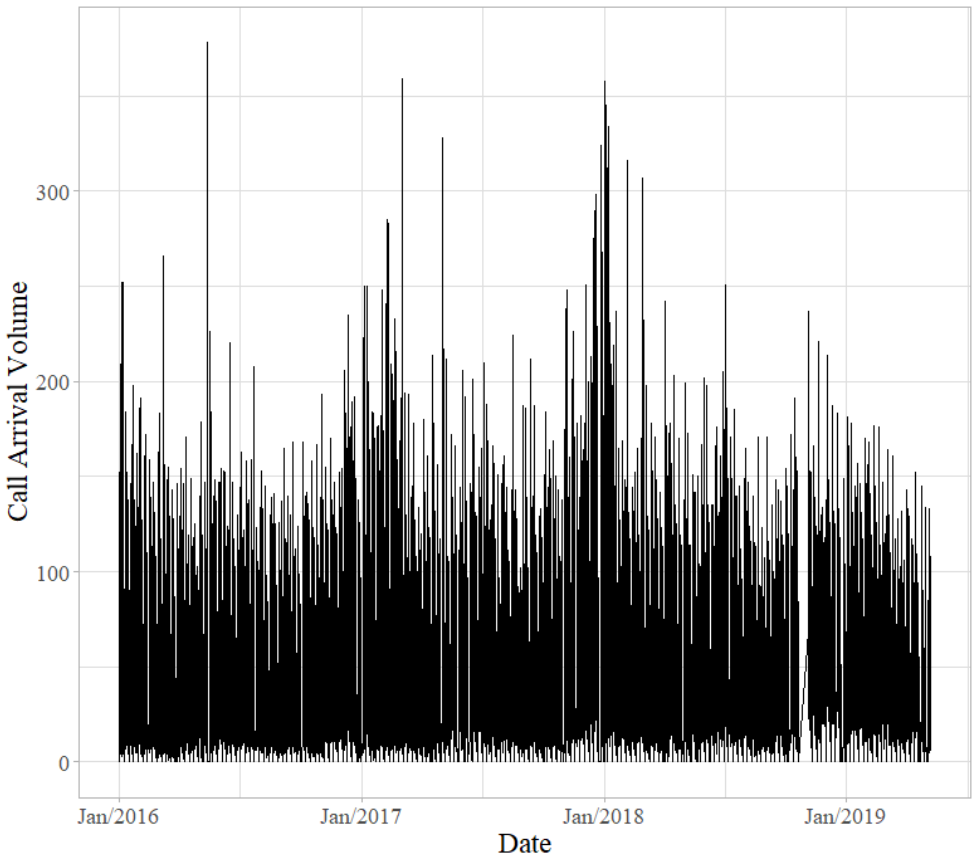

Fig. 1. Overall call arrival volume of customer support queue.

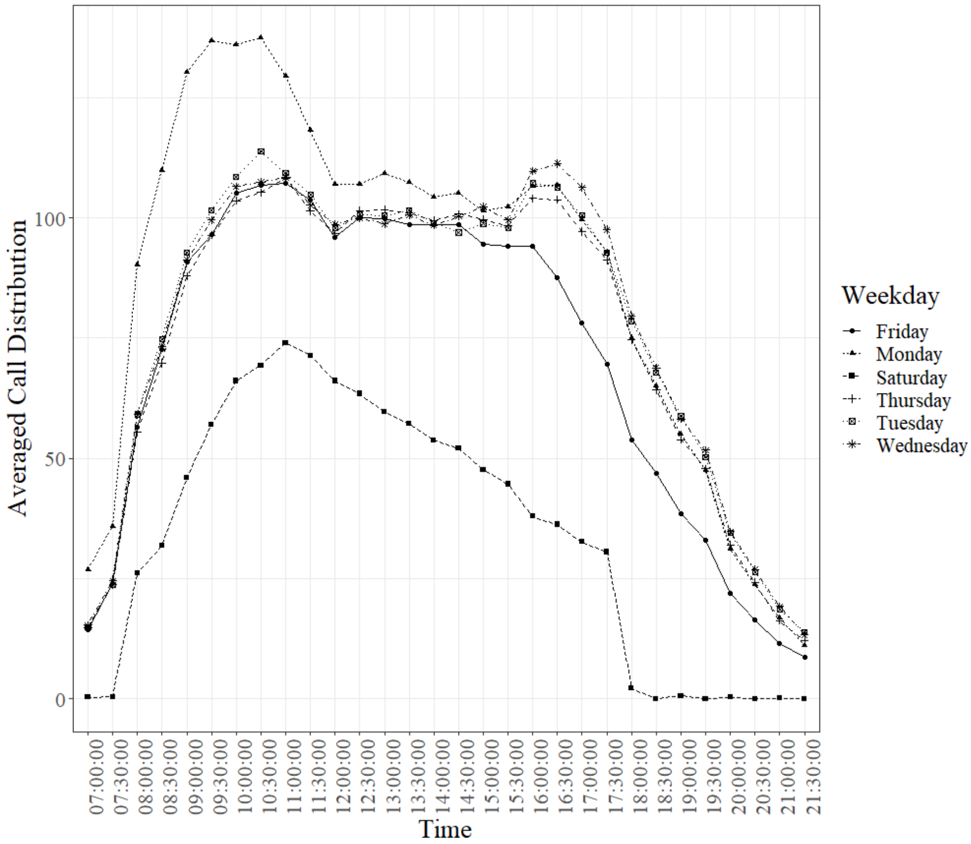

Fig. 2. Averaged call distribution per day for customer support queue.

For the customer support queue, the maximum number of call arrivals per half hour is 378, and the data comprise 2,218 zeroes, i.e., intervals without call arrivals. The customer support data are over-dispersed, exhibiting a mean of 70.9539 and a variance of 2,181.6742. We conducted an Augmented Dickey Fuller (ADF) test to check whether the data have unit root and hence, are nonstationary: we cannot reject the null hypothesis of unit root in the data with a p-value of 0.9798 at lag order 9,360 (value of test statistic -0.5469) and thus, assume that our data are nonstationary. Consequently, we have to apply time series decomposition to our time series models. Drawing on seasonal-trend decomposition based on Loess (STL) (Cleveland et al., 1990), the time series is detrended and deseasonalized resulting in a seasonal component Sˆ t and a seasonally adjusted component Aˆ t , i.e., the data without a seasonal component. The latter can be forecasted with any non-seasonal forecasting method, whereas the seasonal component is forecasted by using the last period of the estimated component, i.e., a seasonal na¨ıve method. Finally, inverting the decomposition's transformations yields the forecasts of the original time series (Brockwell et al., 2002).

Figure 1 depicts the arrival volume of the customer support queue during the 174.5 weeks of our data. Apparently, the call arrival volume remains more or less constant throughout the considered period.

With respect to the averaged call distribution per day in Figure 2, Mondays are the busiest days with an extremely high peak in the morning hours. The remaining weekdays exhibit a relatively similar course with a peak in the morning and a second peak during the afternoon. In contrast, there are few call arrivals on Saturdays.

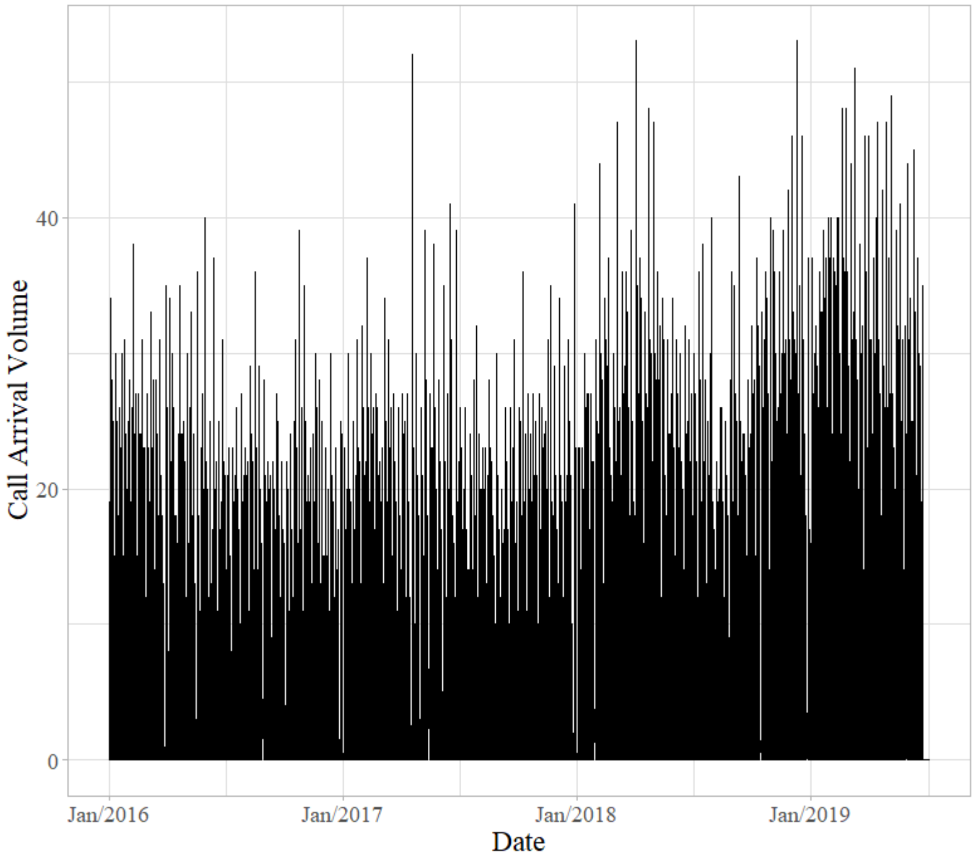

Fig. 3. Overall call arrival volume of customer complaints queue.

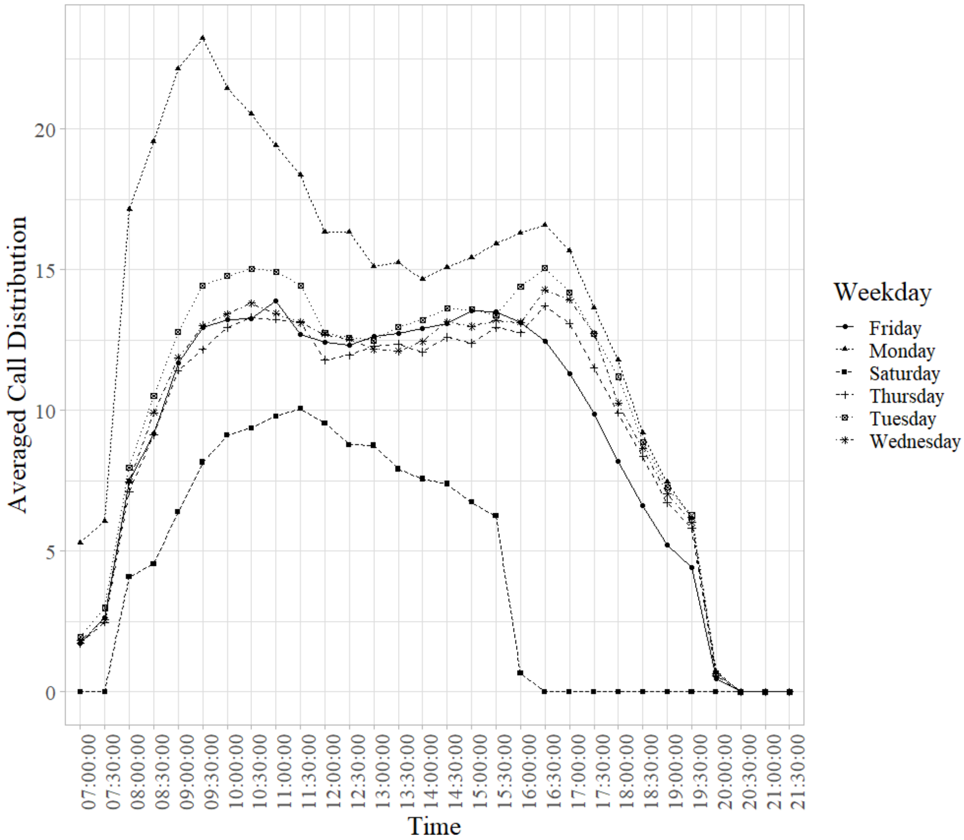

Fig. 4. Averaged call distribution per day for customer complaints queue.

Regarding the customer complaints queue, the maximum number of call arrivals per interval is 53, and the dataset contains 6,551 intervals without call arrivals. Since we cannot reject the null hypothesis of the ADF test with a p-value of 0.7905 at lag order 9,360 (value of test statistic -1.5009) we assume our data to have unit root and, consequently, to be nonstationary. Accordingly, time series decomposition is applied to the time series model. Similar to the customer support queue, Figure 3 shows the overall arrival volume of the customer complaints queue: the call arrival volume remains relatively constant over time, but there is a slight increase towards the end of the dataset. Figure 4 reveals that the customer complaints' averaged call distribution per day is similar to the customer support queue on a lower level.

We model predictor variables (summarized in Table 1) to yield more accurate forecasts. Largely, our variables align with those of extant literature such as weekdays and billing periods (Aldor-Noiman et al., 2009) or holidays and catalog mailings (Andrews and Cunningham, 1995).

Table 1: Predictor Variables.

Variable	Description
Time-of-the-day	Nominal variable capturing the time-of-the-day-effect; 30 half-hourly values ranging from 7 a.m. to 9:30 p.m.
Day-of-the-week	Nominal variable capturing the day-of-the-week-effect; six values ranging from Monday to Saturday
Holiday	Nominal variable capturing the effect of German public holidays; 16 values for public holidays and ordinary weekdays
Day after holiday	Dummy variable capturing the effect of days after German public holidays, two values for days after holidays and ordinary weekdays
Outlier	Nominal variables capturing the effect of outliers; four values for extreme outliers as well as outliers (marked by the management), days on which the call center is closed, and ordinary weekdays
School holidays	Metric variable capturing the effect of German school holidays; specifying the number of German states having school holidays
Year	Nominal variable capturing the effect of busier seasons; eight values for semiannual sections from January 2016 to May 2019
CW0-3	Four dummy variables capturing the effect of catalog mailings on the first weekend, as well as the first, the second, and the third week after release
MMail1-2, MPost1-2, DMail1-2	Six dummy variables capturing the effect of reminders via e-mail (MMail) as well as via mail (MPost) and due date e-mails (DMail) on the day of delivery and the day after

3.2 Research Design

To evaluate the predictive power of adequate ML approaches and to ensure the practical value of our study, we follow a two-step approach. It comprises the analysis of prediction performance in the form of a method comparison in line with extant forecasting research (see e.g., Taylor (2008); Cao and Parry (2009)) and, as proposed by Buitinck et al. (2013), an in-depth walk-through example of the process of model comparison and selection to make the practical implementation accessible to decision-makers and non-experts.

In the first step, we conduct a model comparison of selected ML methods, presented in Section 2.3 (i.e., gradient boosting with dropout (GBD), gradient boosting with L1 and L2 regularization (GBR), KNN, RF, and SVR) with the three most commonly used time series models identified in Section 2.2 (i.e., ARIMA, ETS, and RW, for further formal information on these time series approaches readers are referred to the Appendix). The included methods summarized in Table 2 cover sophisticated ML and time series models as well as standard benchmark techniques. The model performance is evaluated based on the two datasets described in Section 3.1, and we include four different lead times in our experimental setup (three weeks, two weeks, one week, and no lead time from the forecast origin). This is done to validate our results as well as to assess the flexibility of the investigated models in an authentic forecasting situation that is comparable to real call center settings with specific organizational requirements like staffing regulations. We thereby aim to provide an extensive and robust assessment of the prediction accuracy of feasible ML models in call center arrivals' forecasting.

For model validation, we apply cross-validation with an expanding rolling window. Thereby, the initial model is fitted with its optimized hyperparameters using 118 weeks or 21,270 observations from January 2, 2016 to April 7, 2018 as training data. We then predict one week or 180 observations (i.e., forecast horizon $h = 180$). For the next iteration k, we roll the training data one week forward, re-optimize the model's hyperparameters or re-estimate the model respectively, and predict one week further. We repeat this step 52 times, i.e., for one year, and thus, $k = 52$. As stated earlier, we have to exclude two weeks of data from October 22, 2018 to November 4, 2018 and thus, we predict 9,000 observations. We evaluate the models' performance by comparing the predictions with the actual values, i.e., the test data, and hence, compute forecast accuracy.

Table 2: Models for comparison.

Model type	Model	Description
ML approaches	GBD	Algorithm builds an ensemble of weak tree learners, minimizes the model's loss by adding weak learners sequentially using a gradient descent like procedure, and randomly drops boosting tree members
	GBR	Algorithm builds an ensemble of weak linear base learners and utilizes L1 (Lasso Regression) as well as L2 (Ridge Regression) regularization
	KNN	Algorithm predicts an observation by averaging the values of the k nearest neighbors
	RF	Algorithm builds an ensemble of decision trees using a bootstrap sample of the data for each tree and averages the aggregated prediction of the trees
	SVR	Algorithm builds a separating hyperplane into the feature space of output and input vectors which should have at most ε deviation from the actual targets and should be as flat as possible
Time series models	STL + ARIMA	Time series is decomposed based on the Loess procedure and the seasonally adjusted component is fore- casted based on the time series' lagged values and lagged errors
	STL + ETS	Time series is decomposed based on the Loess procedure and the seasonally adjusted component is fore- casted based on previous level and error
	STL RW - DRIFT	Time series is decomposed based on the Loess procedure and the seasonally adjusted component is forecasted based on the time series' last observation and the average of changes between consecutive observations

Note: ARIMA = autoregressive integrated moving average, ETS = error, trend, seasonal (innovation state space model), GBD = gradient boosting with dropout, GBR = gradient boosting with regularization, KNN = k-nearest neighbor, RF = random forest, RWDRIFT = random walk with drift, STL = seasonal-trend decomposition based on Loess, SVR = support vector regression.

As performance measures for forecast accuracy, we draw on the mean absolute error (MAE) and the root mean squared error (RMSE)

$$\text{MAE} = \frac{1}{T}\sum_{i=1}^{T}|Y_i - \hat{Y}_i| \qquad \text{RMSE} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(Y_i - \hat{Y}_i)^2}$$

where the test subset is given by Y_i, the predicted values are $\hat{Y}_i$, and T is the number of predicted values. Both error measures are frequently utilized by literature to determine accuracy (e.g., see (Aldor-Noiman et al., 2009; Ibrahim et al., 2016; Taylor, 2008; Weinberg et al., 2007)) since they are easy to interpret and further, scale-dependent and therefore, suitable to compare forecasts on the same scale. Complementary, we report the computation time of both the benchmark time series models as well as the ML approaches to capture computational complexity and add practical value to the results.

In the second step, in Section 6, we provide a methodological walk-through example for a valid model selection process based on cross-validation with an expanding rolling window. By illustrating different sequences of the implemented programming code used in the experimental design of the first step, we conduct the comparison and selection of the most suitable forecasting method comprehensible to organizational decision-makers and detach the study's value from specific characteristics of our datasets by making the implemented approach reproducible. Additionally, we aim to provide further evidence for the practical applicability of adequate ML algorithms in call center forecasting. Therefore, we do not only describe the generic programming of time series cross-validation with an expanding rolling window but further give detailed insights into the implementation of RF algorithm as the best-performing ML model in our preceding analysis. We also provide guidance on how to measure MAE, RMSE, and computation time in the process. For the methodological walkthrough, we make use of the open-source statistical programming language R (Ihaka and Gentleman, 1996). Drawing on the combined results of both method evaluation and overall implementation process, we then derive practical implications for organizations.

4 Results

Drawing on the results for the customer support queue in Table 3 and Table 4, the RF algorithm outperforms the remaining approaches in every lead time constellation: with respect to both MAE and RMSE, the model yields the most accurate forecasts. The GBD, GBR, and SVR models yield comparable results, whereas the KNN approach was the most inaccurate forecasting method. Generally, every considered ML approach is superior to the benchmark time series models for all lead time constellations (except for the KNN method). Among the time series models, the ETS model is the best-performing approach. Overall, the models' performances worsen slightly with increasing lead time.

Regarding computation time, the RWDRIFT model was excelling with an estimation time of 39 seconds[1] for 52 iterations of the expanding rolling window. The remaining time series models yield comparable low computation times with 142.41 seconds for ETS and 1,260.94 seconds for ARIMA. The AI-driven methods are computationally more intensive with 61,423.71 seconds estimation time for GBR, 93,861.33 seconds for KNN, 171,380.33 seconds for SVR, and 184,367.70 seconds for GBD. With 75,185.62 seconds for the estimation procedure of the rolling window, the RF algorithm provides an acceptable trade-off between accuracy and computation time: for the prediction of one iteration k (i.e., of the forecast horizon h = 180 observations), the model takes 24.1 minutes.

Table 3: MAE results for customer support arrivals' forecasts.

Lead time / Model	No lead time	One week	Two weeks	Three weeks
GBD	13.4601	13.6603	13.9540	14.2203
GBR	12.9393	13.1488	13.3987	13.7386
KNN	18.2068	18.8704	19.2332	19.8064
RF	**11.7544**	**11.8129**	**12.0648**	**12.8134**
SVR	13.2325	13.2063	13.2256	13.6019
STL+ARIMA	14.5263	14.7407	15.5448	15.8520
STL+ETS	14.5152	14.5382	15.2424	15.7428
STL+RW	14.6651	14.6334	15.2941	15.7877

Note: The best accuracy results for each lead time are marked in bold. ARIMA = autoregressive integrated moving average, GBR = gradient boosting with regularization, GBD = gradient boosting with dropout, ETS = error, trend, seasonal (innovation state space model), KNN = k-nearest neighbor, RF = random forest, RWDRIFT = random walk with drift, STL seasonal-trend decomposition based on Loess, SVR = support vector regression.

[1] With 40 GB RAM.

Table 4: RMSE results for customer support arrivals' forecasts.

Lead time / Model	No lead time	One week	Two weeks	Three weeks
GBD	18.8706	19.0480	19.3644	19.9452
GBR	18.1216	18.3299	18.6043	19.3079
KNN	24.9867	25.9203	26.3528	27.6355
RF	**15.5678**	**16.6541**	**16.8929**	**18.4903**
SVR	18.3313	18.4081	18.3059	18.9199
STL+ARIMA	22.7009	23.1810	24.2187	25.0726
STL+ETS	22.9251	23.0876	23.9239	24.7768
STL+RW	23.0503	23.1555	23.9506	24.7793

Note: The best accuracy results for each lead time are marked in bold. ARIMA = autoregressive integrated moving average, GBR = gradient boosting with regularization, GBD = gradient boosting with dropout, ETS = error, trend, seasonal (innovation state space model), KNN = k-nearest neighbor, RF = random forest, RWDRIFT = random walk with drift, STL seasonal-trend decomposition based on Loess, SVR = support vector regression.

To check the robustness of our results, we further consider the queue for customer complaints call arrivals. Since there are less call arrivals compared to the customer support queue, the MAE and RMSE are generally lower. Similar to the previous findings, the RF yields the most accurate forecasts compared with the remaining approaches for all considered lead times except for the MAE result with two weeks lead time for which GBR is found to be superior (see Table 5 and 6). Aside from RF, GBR is outperforming the RWDRIFT model.

The remaining models (i.e., GBD, KNN, and SVR) generate slightly more inaccurate forecasts. Moreover, with the lead time extending, the MAE and RMSE results worsen steadily in most cases.

Table 5: MAE results for customer complaints arrivals' forecasts.

Lead time / Model	No lead time	One week	Two weeks	Three weeks
GBD	3.7668	3.8067	3.8933	3.9694
GBR	3.6058	3.6962	**3.3783**	3.8362
KNN	4.5016	4.7366	4.8095	4.8350
RF	**3.3561**	**3.4348**	3.5629	**3.6746**
SVR	4.3283	4.2826	4.3224	4.2830
STL+ARIMA	3.7197	3.7639	3.8297	3.9073
STL+ETS	3.6990	3.7475	3.8199	3.9163
STL+RW	3.6589	3.7460	3.7968	3.9017

Note: The best accuracy results for each lead time are marked in bold. ARIMA = autoregressive integrated moving average, GBR = gradient boosting with regularization, GBD = gradient boosting with dropout, ETS = error, trend, seasonal (innovation state space model), KNN = k-nearest neighbor, RF = random forest, RWDRIFT = random walk with drift, STL seasonal-trend decomposition based on Loess, SVR = support vector regression.

Table 6: RMSE results for customer complaints arrivals' forecasts.

Lead time / Model	No lead time	One week	Two weeks	Three weeks
GBD	5.3580	5.4212	5.5593	5.6714
GBR	5.2140	5.3527	5.4871	5.5734
KNN	6.4708	6.8279	6.9224	6.9549
RF	**4.9422**	**5.0672**	**5.2338**	**5.3791**
SVR	5.9909	6.0502	6.1240	5.9487
STL+ARIMA	5.5152	5.5807	5.6783	5.8243
STL+ETS	5.4833	5.5559	5.6635	5.8210
STL+RW	5.3958	5.4754	5.5949	5.7647

Note: The best accuracy results for each lead time are marked in bold. ARIMA = autoregressive integrated moving average, GBR = gradient boosting with regularization, GBD = gradient boosting with dropout, ETS = error, trend, seasonal (innovation state space model), KNN = k-nearest neighbor, RF = random forest, RWDRIFT = random walk with drift, STL seasonal-trend decomposition based on Loess, SVR = support vector regression.

To gain further insights regarding the models' performance, we plotted the last predicted week (i.e., 180 observations) for the customer support queue. Figure 5 depicts the time series models' predictions, whereas Figure 6 illustrates the machine learning models' predictions. On the first day of the week (i.e., Monday), the call center was closed, and consequently, this led to an exceptionally high arrival volume on the day after. Apparently, the time series models cannot capture such special days due to the lack of additional information, i.e., predictor variables indicating e.g. holidays and days after. The remaining ML models capture ordinary weekdays and further, such special days more accurately since they allow for the inclusion of explanatory variables for prediction. Consequently, the ML approaches exceed the time series models regarding predictive performance.

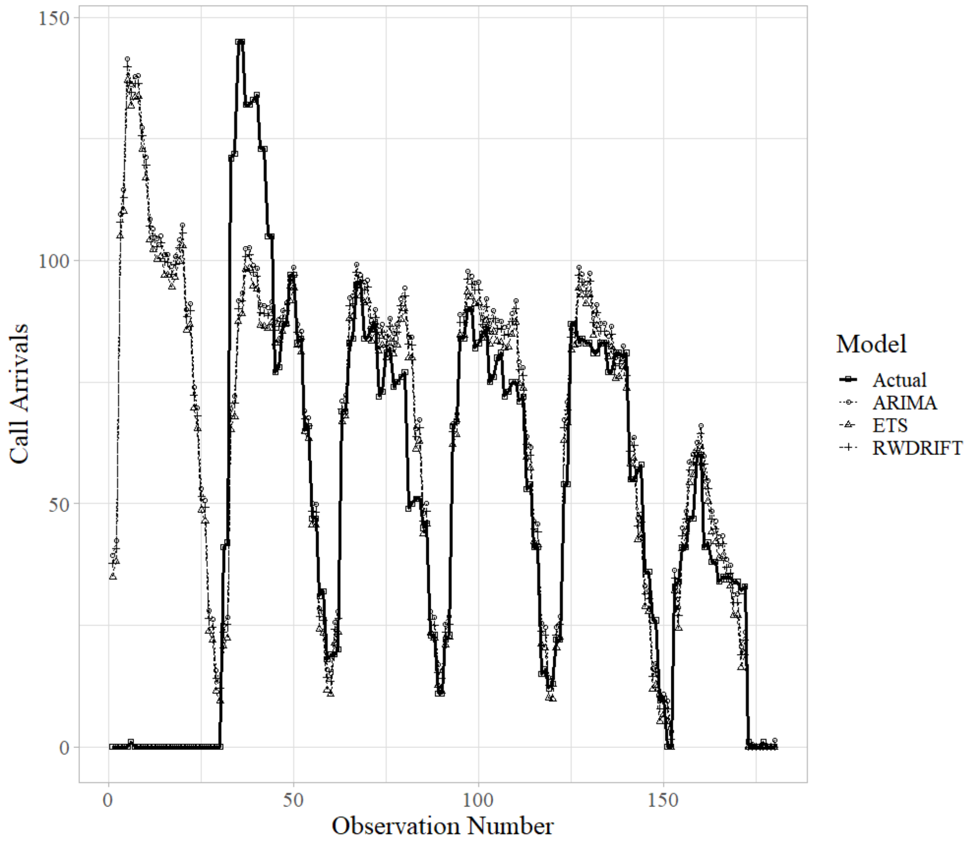

Fig. 5. Last predicted week of the time series models.

Note: The bold line represents the actual values. ARIMA = autoregressive integrated moving average, ETS = error, trend, seasonal (innovation state space model), RWDRIFT = random walk with drift.

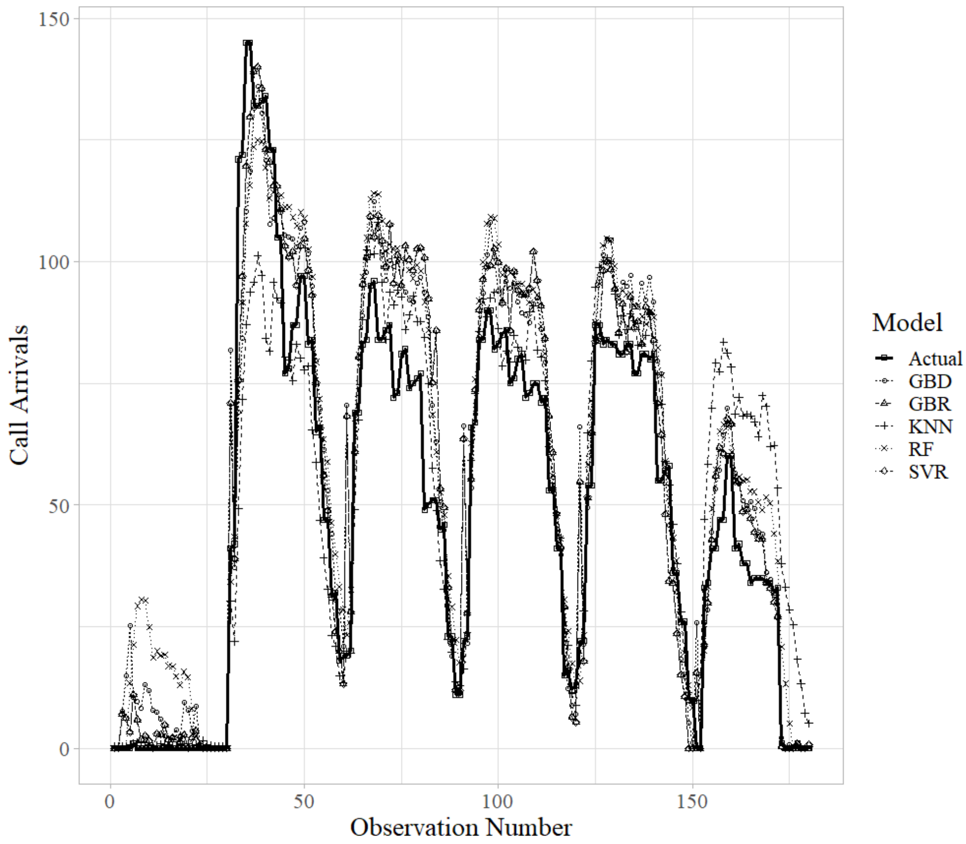

Fig. 6. Last predicted week of the machine learning models.

Note: The bold line represents the actual values. GBD = gradient boosting with drop-out, GBR = gradient boosting with regularization, KNN = k-nearest neighbor, RF = random forest, SVR = support vector regression.

5 Discussion

The underlying investigation entails several theoretical implications and contributions made to literature. We present a starting point for shifting traditional call center forecasting literature towards a new paradigm drawing on AI-driven methods by providing a comprehensive understanding of the predictive potential of ML models. As traditional forecasting literature (Mabert, 1985; Thompson and Tiao, 1971; Andrews and Cunningham, 1995; Taylor, 2008, 2012; Barrow and Kourentzes, 2018) is predominantly characterized by the use of time series models, we intend to broaden this perspective: Across the two datasets examined, our investigated ML algorithms outperform benchmark models as well as more sophisticated time series models that prior studies most commonly focused on (e.g., ARIMA, exponential smoothing, etc.) in nearly all lead time constellations. Thereby, extending the research on call arrival forecasting techniques with ML approaches like GBR, GBD, KNN, RF, and SVR in this analysis leads to a wider range of methods to generate predictions that are more accurate.

Our comprehensive model comparison underpins the preliminary findings of previous studies (Barrow, 2016; Ebadi Jalal et al., 2016; Rausch and Albrecht, 2020), which used single AI-driven methods like RF or neural networks, indicating that ML techniques are capable of improving the accuracy of call center arrivals' forecasts. Our results prove that tree-based methods and particularly the RF algorithm yield the highest potential for significantly improving forecast accuracy. This finding is replicated for a considerably lower level of call arrival volume in the customer complaints queue.

With regard to the models' practicability, which was neglected by extant literature so far, we are first to consider simultaneously different lead times (i.e., three weeks, two weeks, one week, and no lead time), the trade-off between complexity (i.e., estimation time and computation effort), and forecast accuracy in the model comparison. Extant call center forecasting literature focused mainly on forecast accuracy as a primary decision criterion or considered varying forecast horizons (Aldor-Noiman et al., 2009; Barrow, 2016; Taylor, 2008, 2012) while keeping lead times constant and neglecting complexity. Results prove the leading ML models, and especially RF, to be highly relevant for practical use as their forecast accuracy is less affected by lead time extension. Computation effort, on the other hand, is moderate, and implementation is feasible.

Additionally, we took a closer look at the main reasons for the superiority of ML models. Shedding light on the predictions of special days, such as days after holidays, indicates that ML methods excel in coping with anomalous values as predictor variables are included in 19 the generated ex-post forecasts. Hence, one of the main aspects of ML approaches outperforming traditional time series models is assumed to be the ability to capture additional information on the predicted date or customer contact activities by businesses with the inclusion of predictor variables. Thereby, this characteristic of ML techniques makes them not only stand out in terms of forecast accuracy when it comes to outliers (Barrow and Kourentzes, 2018) but also positively affects the overall prediction performance over longer time periods. Nevertheless, albeit few suggestions regarding useful predictor variables have been made (e.g., catalog mailings and holidays (Andrews and Cunningham, 1995) or billing cycles (Aldor-Noiman et al., 2009)), research still lacks a comprehensive understanding on suitable predictor variables for call center arrivals' forecasting. We thus add to the existing body of literature by highlighting that variables such as the time of the day, day of the week, holidays, days after holidays, catalog mailings, and reminders provide valuable information for modeling ex-post forecasts.

The empirical results reported herein should be considered in the light of some limitations. The primary limitation to the generalization of these results accompanies one of the strengths of the study.

Keeping in mind the required balance between prediction accuracy and model complexity, we focus on practical relevance in our model selection and neglect models like e.g. sophisticated types of artificial neural networks since such models are time-consuming in estimation, and thus, inadequate for practical use. We also refrain from developing and testing an own method. With an abundance of different ML methods and modifications in literature, we apply ready-to-use methods that are comparatively easy to implement and present a methodological extension to research in the form of a novel implementation focus. Second, the models' prediction performances are depending on the underlying data and, thus, are assumed to vary slightly for different datasets. Therefore, we validate the models' forecast accuracy on two datasets to prove the robustness of our results and further provide the methodological tutorial for testing the identified ML models' performance on other datasets. We do not distinguish between different forecast horizons like several other studies as we re-estimated our models for every week rolling forward from forecast origin, and thus, the forecast horizon constantly remains one week, i.e., 180 observations.

6 Practical Implications: Methodological Walk-Through for Call Center Arrivals' Forecasting

Based on the results of the conducted model comparison, organizations are suspected of benefiting from including ML approaches in their process of evaluating and selecting the most suitable method for forecasting call center arrivals and therefore, to support their staffing decisions. To make the underlying process of method comparison and selection accessible to decision-makers in practice as well as to overcome its perceived high complexity and organizations' lack of expertise, we provide a methodological walk-through example based on cross-validation with an expanding rolling window. In doing so, we propose to view the question of method in call center forecasting as the overall issue of implementing a forecasting system that includes prediction accuracy as well as practicability for the user. By presenting a methodological tutorial, we aim to overcome the dependence of method comparisons on data characteristics and, at the same time, accelerate the adoption of ML techniques in this field. On these grounds, we provide a description of the generic cross-validation approach in the programming environment R as well as an in-depth example of RF algorithm as the best-performing model of our previous analysis.

```
results <- vector("numeric")
tic("Looptime")
for(i in 1:k){
   train_subset <- data[1:((n_train − h)+(i*h)),]
   test_subset <- data[((n_test − h)+(i*h)):(((n_test − h) − 1)+((i+1)*h)),]
   ## Insert Model here
}
toc(log = TRUE)
timelog <- tic.log(format = TRUE)
results <- pmax(results,0)

mae(data[n_test: (n_test + k ∗ h), m_calls], results)
rmse(data[n_test: (n_test + k ∗ h), m_calls], results)
```

Fig. 7. R Code for rolling expanding window with generic for-loop. Note: The bold variables have to be replaced depending on the specific dataset.

Figure 7 illustrates a generic for-loop for the expanding rolling window that can be utilized to identify the most accurate model. Let n be the n^{th} observation (i.e., row) of the dataset, m be the m^{th} variable (i.e., column) of the dataset, and h be the forecast horizon.

After analyzing and preprocessing the data as described in Section 3.1, we define an empty numeric vector the results are stored in sequentially during the loop. The for-loop itself can be iterated k times: let the forecast horizon h be e.g. one week and out-of-sample predictions with cross validation shall be generated for one year, then $k = 52$, i.e., 52 weeks. For each iteration $k = 1,2, \dots, K$ during the loop we define the training and test subset which roll forward for one unit of the forecast horizon h, i.e., $i * h$. Since $1 * h$ observations are added during the first iteration for syntax reasons, h observations are subtracted from the training and test subsets (n_{train} and n_{test} respectively) to yield the intended initial training and test subsets.

After the loop finishes, the looptime is reported with the *toc()* function to survey the models' computation time as a potentially crucial aspect for decision-makers. Further, in case some models might generate negative predictions we set the minimum value for predictions to zero with *pmax()*. The MAE and RMSE are both calculated by inserting the vector of actual values as the first argument and the vector of predicted values as the second argument.

```
results <- vector("numeric")
tic("Looptime")
for(i in 1:k){
  train_subset <- data[1:((n_train − h)+(i*h)),]
  test_subset <- data[((n_test − h)+(i*h)):(((n_test − h) − 1)+((i+1)*h)),]
  task <- makeRegrTask(data = train_subset, target = "m_calls")
  RF <- tuneRanger(task, num.trees = t, iters = j)
  pred_rf <- predict(RF$model, newdata = test_subset)
  pred_data <- pred_rf$data
  pred_response <- pred_data$response
  results <- append(results, pred_response)
  }
toc(log = TRUE)
timelog <- tic.log(format = TRUE)
results <- pmax(results,0)

mae(data[n_test: (n_test + k * h), m_calls], results)
rmse(data[n_test: (n_test + k * h), m_calls], results)
```

Fig. 8. R Code for rolling expanding window with random forest. Note: The bold variables have to be replaced depending on the specific dataset.

To test a model's predictive ability, it can be integrated into the generic for-loop. Figure 8 demonstrates the R Code for the loop with the implemented RF. To achieve ease of use as well as to guarantee high model accuracy, we make use of R's *tuneRanger* package, which automatically tunes the forest's hyperparameters (i.e., m_{try}, minimum node size, and sample fraction) by creating a regression task with makeRegrTask() (Probst et al., 2019).

The package is favorable since it utilizes sequential model-based optimization (SMBO)[2] as a tuning strategy, which is faster and moreover, better regarding its performance than standard tuning packages (Probst et al., 2019). It conducts an SMBO with 30 random points for the initial design (i.e., random points drawn from the hyperparameter space) and 70 iterative steps in the optimization procedure. Optionally, the number of iterations i can be inserted manually. m_{try} values are sampled from $[0, p]$ with p being the number of predictors. Sample size values are sampled from $[0.2 *$

[2] For detailed information on the SMBO procedure, readers are referred to Probst et al. (2019).

$n, 0.9 * n]$ with the number of observations n. Node size values are sampled with higher probability for smaller values by sampling x from $[0,1]$ and hence, transforming the value by $[(n * 0.2)^x]$. Further, out-of-bag predictions during the fitting procedure can be evaluated with several different error measures (mean squared error (MSE^{OOB}) as default for regression). The number of trees t can be inserted optionally: research found the model's performance peak to be reached during the construction of the first 100 trees (Probst et al., 2012).

Subsequent to the fitting procedure, predictions based on unknown test data are generated. By using the *append()* function, the predictions with length *h* are attached sequentially for *k* iterations. As described in Section 3.2, the MAE and RMSE results allow for a practically valid comparison of different models.

The methodological walk-through combined with the preceding results of the model comparison lead to several practical implications for businesses and organizational decisionmakers. First, the hypothesized improvement of call center arrivals' forecasting accuracy was confirmed for the selection of feasible ML methods. Thereby, the range of applicable methods providing robust and accurate predictions in this field is extended to suitable ML algorithms. In comparison with commonly used forecasting techniques, ML models generate more precise forecasts in almost every case. That way, unnecessary costs caused by overstaffing as well as customer dissatisfaction originating from long waiting times due to understaffing can be avoided: In case the forecasts overestimate the actual customer support call arrival volume, decision-makers can save approximately 1.83[3] call center agents per day on average if RF (best-performing ML model) compared to ETS (best-performing time series model) is employed. Vice versa, in case the forecasts underestimate the actual call arrival volume, customers would need to wait approximately 0.41[4] minutes less on average if RF is implemented instead of ETS. Furthermore, the findings also indicate that decisionmakers are recommended to minimize lead time in case it is possible in the scope of staffing regulations.

Overall, we exclusively investigated models standing out due to the favorable trade-off between accuracy and practicability, especially in terms of complexity regarding estimation time as well as computation effort. The comprehensibility and ease of implementation of treebased models as best-performing methods is further verified by the applied example above. From a general perspective, organizations are encouraged to use the demonstrated process of cross-validation with an expanding

---

[3] If the processing time is 10 minutes per call arrival and the working hours per call center agent are 8 hours per day.
[4] If the processing time is 10 minutes per call arrival and there are 70.95 call arrivals per interval on average.

rolling window not only to test and implement different approaches for call center arrivals' forecasting but also to adapt it for any forecasting task based on sequential data (e.g., e-mail arrivals, product sales, etc.). The implementation of this approach in an accessible programming environment further fills the need of practitioners for a task-specific guideline for the selection of AI-driven methods and helps to overcome the practicability issues identified in literature.

7 Conclusion

The process of forecasting call center arrival volumes in an increasingly complex data environment is predestinated to capitalize on AI-driven methods by improving internal decision making. Accurate forecasts generated by ML algorithms are assumed to generate cost savings and service improvements through precise staffing. However, insights on and practical use of ML in call center arrivals' forecasting are limited.

Acting on the assumed potential of ML in this field as well as on the existing constraints regarding practicability in organizational use, this paper follows a two-step approach of model performance evaluation and practical implementation. The first step constitutes an extensive model comparison of selected feasible ML methods with common as well as sophisticated time series models using the call center arrival data of a large online retailer. In the second step, the implementation of the model evaluation and selection process based on cross validation with an expanding rolling window is made accessible for practitioners by providing a methodological walk-through example.

The results of the method comparison confirm the hypothesized high potential of ML models for accuracy improvements based on two datasets and various lead times investigated. Tree-based methods and particularly RF algorithm yield the best prediction performances and therefore approve as preferable alternatives to commonly used methods. These findings are substantiated by the implementation example using RF as the best-performing model. By providing an efficient and reproducible way of assessing the case-specific value of ML methods in forecasting for organizations within a programming environment, the dependence of method comparison results on data characteristics as well as the lack of comprehensibility and methodological expertise in practical settings are mitigated or even eliminated.

This paper therefore presents a starting point for shifting traditional call center forecasting towards a new paradigm drawing on AI-driven methods by demonstrating the high predictive potential of ML models in comparison to commonly used methods. From a practical perspective, this study contributes to an improved understanding for businesses on how to deal with the increasingly complex task of forecasting call center arrivals caused by the datafication of customer relationships. Being aware of the general applicability of ML models to yield high forecast accuracy, organizations are now enabled to test ML techniques in individual practical settings by adapting the proposed implementation of a valid model selection process in time series forecasting. Improvements in prediction accuracy achieved by this approach can directly be capitalized on through optimized staffing. Future research is encouraged to extend the predictions to concrete staffing recommendations incorporating average service times. As a whole, this work suggests that taking the next step in call center arrivals' forecasting research towards advanced ML, such as deep neural networks and hybrid approaches, is likely to be beneficial. In this case, the evaluation of these methods beyond forecasting accuracy is recommended to ensure the practical value of future findings.

Appendix: Time Series Models

Overall, one strength of time series models is their ability to generate predictions only based on the time series' previous values without any other contextual information and thus, they are adequate models if information is scarce. The non-seasonal ARIMA (p, d, q) model (Box and Jenkins, 1970) assumes that a with d degrees differenced time series depends on its past values being p periods apart and on a finite number q of prior forecast errors ε with p, d, and q being non-negative integers. Thus, it consists of an autoregressive process as well as moving average process

$$y_t' = c + \phi_1 y_{t-1}' + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

with y_t' being the differenced time series, ϕ_p being the parameter for autoregressive process, θ_q being the parameter for moving average process, and ε being the error. Since its development in the 1970s, the ARIMA model is among the most popular forecasting approaches across numerous application contexts, as it was found to perform well in the short-term (Barrow, 2016) and further, is suitable for a variety of data types with different characteristics as there are stationary as well as nonstationary ARIMA methods (Hyndman and Athanasopoulos, 2018).

While ARIMA models intend to capture autocorrelations in the data, exponential smoothing models draw on trend and seasonality in the data (Hyndman and Athanasopoulos, 2018). Holt-Winters' exponential smoothing model (Holt, 2004; Winters, 1960) was proposed in the late 1950s and weight the averages of the time series' previous observations. Thereby, the weights are decreasing exponentially the further the observations lie in the past. The component form of simple exponential smoothing can be defined as

$$\ell_t = (1 - \alpha)\ell_{t-1} + \alpha y_t$$

$$\hat{y}_{t+h|t} = \ell_t$$

with horizon $h = 1, 2, ..$, smoothing parameter $0 \leq \alpha \leq 1$ and series level (or smoothed value) ℓ_t at time t. If the exponential smoothing model further allows for additive or multiplicative errors, it evolves into an innovations state space model ETS$(\cdot,\cdot,\cdot)$ for (Error= {Additive (A), Multiplicative (M)}, Trend= {None (N), A}, Seasonal= {N, A, M}):

$$\ell_t \ = \ell_{t-1} + \alpha\varepsilon_t$$

$$y_t \ = \ell_{t-1} + \varepsilon_t$$

where α is the smoothing parameter and $\ell_t$is the series level (or smoothed value) at time t. Random walk models are frequently used for nonstationary data as random walks typically consist of long periods of apparent (upward or downward) trends and exhibit sudden changes in direction (Hyndman and Athanasopoulos, 2018). The forecasts from the random walk model are equal to the time series' last observation:

$$\hat{y} = y_{t-1} + \varepsilon_t$$

As an extension to the basic model, the drift parameter c is frequently added which is the average of changes between consecutive observations:

$$\hat{y} = c + y_{t-1} + \varepsilon_t$$

If c is positive, there is an increase in the average of changes between consecutive observations and thus, the prediction $\hat{Y}$ will tend to drift upwards and vice versa for negative values of c.

References

Agrawal, A., Gans, J., and Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, (Working Paper 24626):1–17.

Aksin, Z., Armony, M., and Mehrotra, V. (2007). The modern call center: A multidisciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688.

Aldor-Noiman, S., Feigin, P. D., and Mandelbaum, A. (2009). Workload forecasting for a call center: Methodology and a case study. *The Annals of Applied Statistics*, 3(4):1403–1447.

Andrews, B. H. and Cunningham, S. M. (1995). L. l. bean improves call-center forecasting. *Interfaces*, 25(6):1–13.

Arora, N., Allenby, G. M., and Ginter, J. L. (1998). A hierarchical bayes model of primary and secondary demand. *Marketing Science*, 17(1):29–44.

Atlason, J., Epelman, M. A., and Henderson, S. G. (2008). Optimizing call center staffing using simulation and analytic center cutting-plane methods. *Management Science*, 54(2):295–309.

Avramidis, A. N., Deslauriers, A., and L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908.

Barrow, D. and Kourentzes, N. (2018). The impact of special days in call arrivals forecasting: A neural network approach to modelling special days. *European Journal of Operational Research*, 264(3):967–977.

Barrow, D. K. (2016). Forecasting intraday call arrivals using the seasonal moving average method. *Journal of Business Research*, 69(12):6088–6096.

Bergstra, J. S., Bardenet, R., Bengio, Y., and K´egl, B. (2011). Algorithms for hyperparameter optimization. *Advances in Neural Information Processing Systems*, pages 2546– 2554.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, New Jersey.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Brady, M. K. and Cronin, J. J. (2001). Some new thoughts on conceptualizing perceived service quality: A hierarchical approach. *Journal of Marketing*, 65(3):34–49.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brockwell, P. J., Davis, R. A., and Calder, M. V. (2002). *Introduction to time series and forecasting*. Springer, New York.

Brynjolfsson, E. and McAfee, A. (2017). The business of artificial intelligence. *Harvard Business Review*, 7:1–20.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). Api design for machine learning software: experiences from the scikit-learn project. arXiv.

Burez, J. and van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636.

Cao, Q. and Parry, M. E. (2009). Neural network earnings per share forecasting models: A comparison of backward propagation and the genetic algorithm. *Decision Support Systems*, 47(1):32–41.

Cezik, M. T. and L'Ecuyer, P. (2008). Staffing multiskill call centers via linear programming and simulation. *Management Science*, 54(2):310–323.

Chen, H., Chiang, R. H. L., and Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4):1165–1188.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonaltrend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273– 297.

Coussement, K. and van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, 34(1):313–327.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Dean, A. M. (2007). The impact of the customer orientation of call center employees on customers' affective commitment and loyalty. *Journal of Service Research*, 10(2):161–173.

Donkers, B., Franses, P. H., and Verhoef, P. C. (2003). Selective sampling for binary choice models. *Journal of Marketing Research*, 40(4):492–497.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Ebadi Jalal, M., Hosseini, M., and Karlsson, S. (2016). Forecasting incoming call volumes in call centers with recurrent neural networks. *Journal of Business Research*, 69(11):4811– 4814.

Fang, K., Jiang, Y., and Song, M. (2016). Customer profitability forecasting using big data analytics: A case study of the insurance industry. *Computers & Industrial Engineering*, 101:554–564.

Fantazzini, D. and Figini, S. (2009). Random survival forests models for sme credit risk measurement. *Methodology and Computing in Applied Probability*, 11(1):29–45.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, (29):1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1– 42.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):205–228.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10.

Huang, M.-H. and Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, 21(2):155–172.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia.

Ibrahim, R. and L'Ecuyer, P. (2013). Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management*, 15(1):72– 85.

Ibrahim, R., Ye, H., L'Ecuyer, P., and Shen, H. (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting*, 32(3):865–874.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Kaplan, A. and Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1):15–25.

Kitchens, B., Dobolyi, D., Li, J., and Abbasi, A. (2018). Advanced customer analytics: Strategic value through integration of relationship-oriented big data. *Journal of Management Information Systems*, 35(2):540–574.

Koole, G. and Pot, A. (2005). Approximate dynamic programming in multi-skill call centers. *Proceedings of the Winter Simulation Conference 2005*, pages 576–583.

Lemmens, A. and Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, 43(2):276–286.

Li, H., Sun, J., and Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications*, 37(8):5895–5904.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Mabert, V. A. (1985). Short interval forecasting of emergency phone call (911) work loads. *Journal of Operations Management*, 5(3):259–271.

Martens, D., Vanthienen, J., Verbeke, W., and Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems*, 51(4):782–793.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):886.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Royal Statistical Society: Series A (General)*, 135(3):370–384.

Overgoor, G., Chica, M., Rand, W., and Weishampel, A. (2019). Letting the computers take over: Using ai to solve marketing problems. *California Management Review*, 61(4):156– 185.

Paliwal, M. and Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, 36(1):2–17.

Probst, P., Wright, M., and Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):281–300.

Rashmi, K. V. and Gilad-Bachrach, R. (2015). Dart: Dropouts meet multiple additive regression trees. *AISTATS*, pages 489–497.

Rausch, T. M. and Albrecht, T. (2020). The impact of lead time and model selection on the accuracy of call center arrivals' forecasts. *Proceedings of the 28th European Conference on Information Systems (ECIS)*, pages 1–16.

Reinartz, W., Thomas, J. S., and Kumar, V. (2005). Balancing acquisition and retention resources to maximize customer profitability. *Journal of Marketing*, 69(1):63–79.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, (5):197–227.

Shen, H. and Huang, J. Z. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, 10(3):391–410.

Shmueli, G. and Koppius, O. (2011). Predictive analytics in information systems research. *MIS Quarterly*, (35):553–572.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., and Kumar Roy, P. (2017). Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, 70:346–355.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958.

Tambe, P., Cappelli, P., and Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4):15– 42.

Taylor, J. W. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science*, 54(2):253–265.

Taylor, J. W. (2012). Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science*, 58(3):534–549.

Thompson, H. E. and Tiao, G. C. (1971). Analysis of telephone data: A case study of forecasting seasonal time series. *The Bell Journal of Economics and Management Science*, 2(2):515.

Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data*. Springer, New York, addendum 1 edition.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer, New York.

Verbeke, W., Martens, D., Mues, C., and Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38(3):2354–2364.

Wedel, M. and Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, 80(6):97–121.

Weinberg, J., Brown, L. D., and Stroud, J. R. (2007). Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association*, 102(480):1185–1198.

Whiting, A. and Donthu, N. (2006). Managing voice-to-voice encounters. *Journal of Service Research,* 8(3):234–244.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342.

Yang, S. and Allenby, G. M. (2003). Modeling interdependent consumer preferences. *Journal of Marketing Research*, 40(3):282–294.

IV Predicting Online Shopping Cart Abandonment with Machine Learning Approaches

Authors: Rausch, Theresa Maria; Derra, Nicholas Daniel; Wolf, Lukas

Published in: International Journal of Market Research (2020), Online First, pp. 1-24,
DOI: 10.1177/1470785320972526

Abstract: Excessive online shopping cart abandonment rates constitute a major challenge for e-commerce companies and can inhibit their success within their competitive environment. Simultaneously, the emergence of the Internet's commercial usage results in steadily growing volumes of data about consumers' online behavior. Thus, data-driven methods are needed to extract valuable knowledge from such big data to automatically identify online shopping cart abandoners. Hence, this contribution analyzes clickstream data of a leading German online retailer comprising 821,048 observations to predict such abandoners by proposing different machine learning approaches. Thereby, we provide methodological insights to gather a comprehensive understanding of the practicability of classification methods in the context of online shopping cart abandonment prediction: our findings indicate that gradient boosting with regularization outperforms the remaining models yielding an F_1-Score of 0.8569 and an AUC value of 0.8182. Nevertheless, as gradient boosting tends to be computationally infeasible, a decision tree or boosted logistic regression may be suitable alternatives, balancing the trade-off between model complexity and prediction accuracy.

Keywords: E-Commerce; Shopping Cart Abandonment; Prediction; Classification; Machine Learning; Supervised Learning

1 Introduction

To strengthen a company's position within its competitive environment, marketers need to be able to precisely predict potential customers regarding their purchase and, further, non-purchase behavior. Considering this in the context of online shopping environment, customers frequently place items in their virtual shopping cart for reasons other than immediate purchase. This phenomenon is known as shopping cart abandonment and is particularly apparent in the context of e-commerce: it is the behavioral outcome of consumers placing item(s) in their online shopping cart without making a purchase by completing the checkout process during that online session (Huang et al., 2018; Kukar-Kinney and Close, 2010). Extant literature investigated the behavioral perspective of online shopping cart abandonment by identifying inhibitors to the purchase process: financial risks and concerns about delivery and return policies (Kukar-Kinney and Close, 2010) the usage of shopping carts as organization tools or for entertainment purposes (Kukar-Kinney and Close, 2010), and inhibitors at the checkout stage like perceived transaction inconvenience and privacy intrusion (Rajamma et al., 2009) are – inter alia – the main factors leading to online shopping cart abandonment.

With the spread of the Internet's commercial usage, the ability to track consumers' online activities allows companies to collect unbiased information about consumers' behavior. The detailed records of past usage behaviors comprised by log files and resulting clickstream data can be analyzed by marketers to gain valuable insights. In this context, clickstream data have frequently been modeled to derive implications for website design or advertising efforts (see, for example, Chatterjee et al. (2003) and Montgomery et al. (2004)) and further, to predict consumers' future behaviors, e.g. regarding purchase (see, for example, Bucklin and Sismeiro (2003) and Moe and Fader (2004a)).

Thus, the antecedents of online shopping cart abandonment are well understood by behavioral literature and clickstream data has been studied by methodological research to analyze consumers' behavior. The rise of the Internet and the era of big data resulted in an excessive 'datafication' (Kelly and Noonan, 2017; Lycett, 2013) of the organizational environment yielding the field of business intelligence comprising data analytics and predictive analytics approaches (Chen et al., 2012). However, despite the richness of clickstream data, prior shopping cart abandonment literature still lacks data-driven methods based on machine learning which make use of this information source to predict such abandoning customers. This might be due to the insufficient awareness of suitable intelligent approaches to extract knowledge from the steadily growing volumes of data (Fayyad et al., 1996).

To address this research gap, we utilize clickstream data of a leading German online retailer to train and subsequently compare different machine learning approaches for the prediction of online shopping cart abandonment (i.e., tree-based methods (more specifically, adaptive boosting, boosted logistic regression, decision tree, gradient boosting with regularization, gradient boosting, gradient boosting with dropout, random forest, and stochastic gradient boosting), k-nearest neighbor, naïve bayes, multi-layer perceptron with dropout, and a support vector machine with radial basis kernel). We successfully implement these machine learning methods for online shopping cart abandonment prediction and compare them with logistic regression as a standard non-machine learning benchmark model regarding their predictive performance.

Our paper makes several key contributions to the preceding literature. By combining the research fields of both shopping cart abandonment as well as clickstream data analysis with machine learning approaches, we particularly shed light on the practicability of machine learning methods in this application context, as this was neglected by prior research. Further, we provide insights into the characteristics of customers abandoning their shopping cart based on clickstream data that is unsusceptible to self-selection, relatively unobtrusive, and easy to gather. We extensively review literature on classification methods to identify shopping cart abandonments and present validation procedures as well as performance metrics for such methods. Our findings can be useful both for marketing intelligence research by extending the field of machine learning applications in marketing contexts through automatically predicting online shopping cart abandoners and for practitioners to actively prevent such abandonments by several real time reactions, e. g. providing real-time purchase incentives, and moreover, to gain insights into machine learning methods.

The remainder of this paper is organized as follows: the subsequent section describes the related work on online shopping cart abandonment and clickstream data. Further, Section 3 summarizes the background on machine learning approaches for classification. Section 4 outlines the methodology comprising a preliminary data analysis and the research design. In Section 5 and 6, we present the findings and discuss both theoretical and practical implications, limitations, as well as directions for future research. Finally, Section 7 draws a conclusion.

2 Related Work

2.1 Online Shopping Cart Abandonment

The online shopping cart abandonment phenomenon causes substantial losses of turnover for online retailers (Huang et al., 2018; Rajamma et al., 2009) resulting in a weakened position within their competitive environment. Therefore, extant marketing literature addressed this problem by drawing on a behavioral perspective to identify and understand essential determinants of online shopping cart abandonment: Rajamma, Paswan, and Hossain (2009) focused on potential inhibitors at the checkout stage and found increased perceived transaction inconvenience (e.g., long registration forms) and high perceived risk (e.g., perceived security of information asked) to enhance online shopping cart abandonment. Partially, these findings seem to be applicable to new customers which are unfamiliar with the checkout process. Similarly, (Kukar-Kinney and Close, 2010) findings indicate that privacy intrusion and security concerns rather lead to the consumers' decision to buy the product from a stationary offline store. Further, they found the entertainment value of shopping carts, the use of shopping carts as an organization tool, the wait for sale, and the concerns about costs to be antecedents of shopping cart abandonment (Kukar-Kinney and Close, 2010). Their identified determinants were supported by Close and Kukar-Kinney (2010) proving that customers' tendencies to add items to the online shopping cart for reasons other than immediate purchase are – inter alia – due to organizational purposes. Huang, Korfiatis, and Chang (2018) focused on mobile shopping cart abandonment in their study. They found intrapersonal (i.e., conflicts regarding mobile shopping attributes and low self-efficacy regarding mobile shopping) and interpersonal (i.e., discrepancies from the other's attitudes to self-attitudes) conflicts to disturb consumers' emotions during mobile shopping, and in turn, implying shopping cart abandonment. Overall, their findings indicate that the utilized device for online shopping might impact purchase behavior as well. Cho, Kang, and Cheon (2006) proved that consumers' confusion by information overload, high value-consciousness, negative past experiences, intention to conduct price comparisons, and unreliable websites are likely to trigger online shopping cart abandonment[1].

[1] Cho, Kang, and Cheon (2006) defined online shopping cart abandonment rather as a hesitation reaction which implies that the customer actively drops items placed in his/her shopping cart. Thus, their definition differs slightly from the definition of Kukar-Kinney and Close (2010), which was used in this study for an understanding of shopping cart abandonment.

2.2 Clickstream Data

Drawing on a more holistic perspective of online shopping behavior, further literature shifted away from explanatory behavioral approaches to data-driven methods predicting online purchase behavior in general. Typically, such predictions are based on clickstream data (see, e.g., Moe and Fader (2004a), Sismeiro and Bucklin (2004), or van den Poel and Buckinx (2005)). Clickstream data model the navigation path a customer takes through the online shop (Montgomery, 2001; Montgomery et al., 2004) and can be extracted from log files which register all requests and information transferred between the customer's computer and the company's commercial web server (Bucklin and Sismeiro, 2003).

Examples for using clickstream data to predict online shopping behavior are – inter alia - Moe and Fader (2004a) who proposed a conversion model predicting each customer's probability of making a purchase based on purchase and visit history. The same authors (Moe and Fader, 2004b) also developed a model for evolving visiting behavior and further, they examined the relationship between visiting frequency and purchasing propensity. They found consumers visiting an e-commerce site more frequently to have a greater propensity to buy (Moe and Fader, 2004b). Van den Poel and Buckinx (2005) predicted purchase behavior and investigated the contribution of different variables: they proved (1) general clickstream variables (i.e., number of days since last visit, and speed of clickstream behavior during last visit), (2) more detailed clickstream variables (i.e., number of accessories (and personal pages and products respectively) viewed during last visit), (3) demographic variables (i.e., gender and the fact of supplying personal information), and (4) historical purchase behavior (i.e., number of days since last purchase and number of past purchases) to be meaningful predictors. Montgomery et al. (2004) set up different models to predict purchase conversion probability by modeling path information.

Moreover, clickstream data was frequently utilized by research to predict not only purchase behavior but further similar outcome variables. For instance, Bucklin and Sismeiro (2003) investigated drivers affecting the length of time spent viewing a website and the visitor's decision to continue browsing or to exit the website. Sismeiro and Bucklin (2004) decomposed the purchase process into sequences that must be completed for a purchase to take place (i.e., completion of product configuration, input of personal information, and order confirmation with provision of credit card data) and predicted the probability of completion for each task with covariates of browsing behavior, repeat visitation, use of decision aids, input effort, and information gathering.

3 Machine Learning Approaches for Classification

Overall, e-commerce as a research subject is suitable for the application of machine learning approaches as proposed by Kohavi and Provost (2001): online retailers can easily and inexpensively collect rich data with respect to the online behavior of customers (i.e., clickstream data) and, further, implement data mining and machine learning applications since political and social barriers are substantially lower than for traditional businesses. Consequently, typical problems for successfully applying machine learning (i.e., the need for a large volume of controlled and reliable data, data with sufficient descriptions, the ability to evaluate results, and to integrate applications successfully) are reduced by the characteristics of e-commerce environment (Kohavi and Provost, 2001).

Machine learning constitutes a new paradigm within data science research and emerged in the course of the artificial intelligence era, which, in turn, was first coined by Samuel (1959) describing it as "the programming of a digital computer to behave it in a way which, if done by human beings […], would be described as involving the process of learning". In this context, learning may be understood as the automatic search for more useful representations of data regarding a specific task (Chollet and Allaire, 2018). Machine learning algorithms and systems are consequently trained rather than explicitly programmed. During this process, these systems find statistical structure in given examples which are relevant to the task and derive rules for automating the task using guidance from a feedback signal (Cui et al., 2006). Thereby, classification algorithms are types of supervised learning approaches within machine learning which predict a qualitative response for an observation, i.e., they assign an observation to a category (James et al., 2013): Formally, let $\{y_k, x_k\}_{k=1}^{N}$ bet a training set, where $y_k \in \{0,1,2,\ldots,K-1\}$ is the class membership and $x_k = \mathbb{R}^n$ is the vector of predictor values, then the task is to learn a function to predict the class label y_k from x_k. Thereby, $K = 2$ in case of binary classification and $K > 2$ in case of multi-class classification tasks.

Drawing on the online shopping cart abandonment problem, the prediction of purchasers and non-purchasers (i.e., customers abandoning their shopping cart) can be considered a binary classification task. Common machine learning approaches for binary classification include – inter alia – tree-based methods, support vector machines, naïve bayes, k-nearest neighbor, and neural networks. The approaches are explained in detail hereinafter.

3.1 Tree-Based Approaches

One of the most common machine learning approaches are tree-based methods which descend from single decision trees, as proposed by Breiman et al. (1984). Basically, decision trees are flowchart-like structures that generate "if-else" rules and thereby allow for prediction of observation classes. Thereby, classification and regression tree models follow a recursive top-down approach in which binary trees aim to partition the predictor space with predictor variables $x_1, \ldots, x_k$ into subsets in which the distribution of the dependent variable y is successively more homogeneous (Chipman et al., 1998).

Generally, single decision trees have the advantage of being easy to interpret and to understand (Moro et al., 2014). However, they frequently lead to overfitting, i.e., the model learns to identify specific characteristics of the training data which are irrelevant or even obstructive for the classification of unknown data (Friedman, 2001; Srivastava et al., 2014). This results in drawbacks of predictive performance and less expressiveness of the models. Ensemble learning methods that construct several individually trained decision trees and combine their results into a classifier outperforming the single predictions (Opitz and Maclin, 1999; Rokach, 2010) may offer a solution to this problem. In this context, two widely used methods of aggregating trees are boosting and bagging.

In boosting, a family of algorithms converts weak learners (i.e., models that achieve accuracy just above random guessing) to strong learners with a powerful predictive capacity. The idea is to train weak learners sequentially with each weak learner trying to correct its predecessor (Schapire et al., 1998). Thus, each decision tree is built using feedback from previously grown trees (James et al., 2013). Popular boosting algorithms include adaptive boosting "AdaBoost" (Freund and Schapire, 1997), boosted logistic regression "LogitBoost" (Friedman et al., 2000), gradient boosting machines "GB" (Friedman, 2001, 2002), and stochastic gradient boosting "SGB" (Friedman, 2002)[2]. For instance, AdaBoost as a basic boosting algorithm makes predicitions by combining the output of weak learners to a weighted sum and putting higher weights on incorrectly classified instances

$$\hat{y} = \operatorname{sign}\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right)$$

[2] The concepts of AdaBoost, LogitBoost, and gradient boosting are closely related as all approaches produce an ensemble of weak learners but – in contrast to AdaBoost and LogitBoost – gradient boosting models minimize the model's loss by adding weak learners sequentially using a procedure similar to gradient descent, i.e., it allows arbitrary differentiable loss functions to be used.

with the weak hypothesis h_m detected by the weak learner and its importance α_m.

In contrast to boosting, bagging (or bootstrap aggregating) grows successive trees independently from earlier trees, i.e., each tree is constructed using a bootstrap sample of the data and, hence, a majority vote is taken for prediction (Breiman, 1996). Random forests add an additional layer of randomness to bagging and change how the trees are constructed: in standard decision trees each node is split using the best split among all predictor variables whereas in random forests the nodes are split using the best among a subset of predictors randomly chosen at that node (Breiman, 2001; Liaw and Wiener, 2002). Due to the recursive structure of tree-based methods they often capture interaction effects between variables. However, since we focus on the performane of the models and not the importance of specific variables, we will not consider interaction effects further in our study.

Overall, tree-based methods have been found to outperform other established approaches across a variety of different classification tasks such as IP traffic flow classification (Williams et al., 2006), customer churn prediction (Vafeiadis et al., 2015), or – similar to our context – prediction of online purchase intention (Bogina et al., 2019; Boroujerdi et al., 2014; Zheng and Liu, 2018). They are particularly favorable since ensemble methods are able to reduce both bias and variance of the single learning algorithms: While individual models may get stuck in local minima, a weighted combination of several different local minima – produced by ensemble methods – are able to minimize the risk of choosing the wrong local minimum (Dietterich, 2002).

3.2 Support Vector Machines

Aside from tree-based methods, support vector machines are powerful tools for classification tasks (James et al., 2013). The basic support vector machine is solving pattern recognition problems by mapping data into a multidimensional input space and constructing an optimal hyperplane that separates the space into homogenous partitions[3] (Cortes and Vapnik, 1995; Vapnik, 1982). Predictions of new instances are then classified into those partitions. The support vector machine aims at constructing a classifier in the form of

$$\hat{y} = \operatorname{sign}\left[\sum_{i=1}^{N} \alpha_k y_k \psi(x, x_k) + b\right]$$

[3] A hyperplane is defined as a flat affine subspace of dimension $p-1$ with p being the number of dimensions (i.e., the number of considered predictor variables) (James et al. 2013). Basically, the ‘hyperplane’ is a line if the feature space is two-dimensional (i.e., two predictor variables) and a simple plane if the space is three-dimensional (i.e., three predictor variables).

where α_k are positive real constants, b is a real constant, and $\psi(\cdot,\cdot)$ represents the hyperplane (e.g., $\psi(x, x_k) = x_k^T x$ in case of a linear support vector machine) (Suykens and Vandewalle, 1999). Aside from the linear case, Boser, Guyon, and Vapnik (1992) proposed a non-linear classifier by applying the so-called kernel trick which allows the algorithm to fit the hyperplane in a transformed feature space.

We used a support vector machine with radial basis kernel for the comparison of machine learning models. However, support vector machines may become computationally infeasible on very large datasets like clickstream data (L'Heureux et al., 2017).

3.3 Naïve Bayes

The naïve bayes approach is a basic classifier based on applying the Bayes' theorem with the naïve assumption that the attributes are conditionally independent (Duda et al., 1973). The classifier assigns a new case to a class label $\hat{y} = C_k$ by deriving the maximum a posteriori probability:

$$\hat{y} = \underset{k \epsilon \{1,\dots,K\}}{\arg\max}\, p(C_k) \prod_{i=1}^{n} p(x_i |\, C_k)$$

Naïve bayes as a generative classifier is frequently utilized for classification tasks due to its simplicity, efficiency, and efficacy (Muhammad and Yan, 2015).

3.4 K-Nearest Neighbor

Another basic approach, the k-nearest neighbor algorithm, classifies an observation by a majority vote of the observation's neighbors (Cover and Hart, 1967). The underlying assumption of the algorithm is that observations which lay closely together within the predictor space (i.e., neighbors) will have the same class label. Thus, the classifier weights the class of the nearest neighbors strikingly high in order to predict the class label of an unclassified sample (Cover and Hart, 1967). The class is thereby assigned by taking the majority vote of the *k* nearest neighbors, with *k* being the number of neighbors that are considered during the classification task. The nearest neighbors are determined with the help of arbitrary distance functions (e.g., Euclidian distance $d(.,.)$). For new observations (y, x) the nearest neighbor $(y_{(1)}, x_{(1)})$ within the training set is defined by

$$d\big(x, x_{(1)}\big) = \min_k (d(x, x_k))$$

and $\hat{y} = y_{(1)}$ – the class of the nearest neighbor – is selected as prediction for y. $x_{(j)}$ and $y_{(j)}$ describe the jth nearest neighbor of x and its class membership y.

K-nearest neighbor as a local learning approach may be suitable for online shopping cart abandonment prediction tasks since it is able to alleviate the challenge of imbalanced data (L'Heureux et al., 2017).

3.5 Artificial Neural Networks

Artificial neural networks are highly parallelized computer systems comprising process units (i.e., neurons) located on process layers with numerous weighted interconnections performing a learning process to create meaningful data representations (Jain et al., 1996). Regarding the concept of deep learning, artificial neural networks may use a number of hidden process layers (the depth of a network) between input and output layer containing non-linear operations in hierarchical architectures to learn characteristics and recognize patterns from given data (Bengio, 2009; Deng, 2011; Hinton et al., 2006). The concept of learning within deep learning (or artificial neural networks, respectively) describes a process of updating the network architecture and the weights of the neuron connections (Jain et al., 1996). To improve the performance, the optimizer is implementing a backpropagation algorithm to minimize the discrepancy between the actual and the target output vector (i.e., the loss score) by adjusting the weights (Rumelhart et al., 1986; Schmidhuber, 2015). To avoid overfitting, a regularization method called dropout can be integrated in the network which randomly sets a share of its output per layer to zero (Srivastava et al., 2014).

Concerning their connection structure (i.e., topology), neural network architectures can be distinguished between feedforward networks (e.g., multi-layer perceptrons (Deng, 2011; Zhang et al., 2018)) with neuron connections running to the output layer acyclically and recurrent networks (e.g., long short-term memories (Hochreiter and Schmidhuber, 1997)) containing backward connections to build cyclic architectures (Jain et al., 1996; Schmidhuber, 2015). The most commonly used feedforward neural networks – multi-layer perceptrons – can be defined as

$$\hat{y} = \beta_0 + \sum_{h=1}^{H} \beta_h g\left(\gamma_{0i} + \sum_{i=1}^{I} \gamma_{hi} p_i\right)$$

where I denotes the number of inputs p_i, H is the number of hidden nodes in the network, the weights $\omega = (\beta, \gamma)$ with $\beta = [\beta_1, \dots, \beta_H]$ and $\gamma = [\gamma_{11}, \dots, \gamma_{HI}]$ are for the hidden and output layer respectively, $g(\cdot)$ is the transfer function (e.g., sigmoid logistic), and β_0 as well as γ_{0i} are the biases of each node (Zhang et al., 1998).

Multi-layer perceptrons were found to outperform other machine learning approaches for purchase intention prediction only after balancing the class distribution with oversampling (Sakar et al., 2019) since deep learning approaches are frequently sensitive to class imbalance (L'Heureux et al., 2017).

4 Methodology

4.1 Preprocessing and Preliminary Data Analysis

The purpose of this study is to predict shopping cart abandonment by making use of machine learning. The machine learning models explained in Section 3 are compared to find the best classifier for this task. The clickstream data were gathered from server log files of a leading German online retailer which primarily distributes fashion. The data were created by the online retailer through extracting the customers' chronological online shop activities out of sequential log files. Each log file observation comprised one action or activity (e.g., a click) of a certain customer such as adding a product to the cart or clicking on a product to view its details. Subsequently, each customer's activities during a session were assigned to summarizing variables. Hence, all activities of a customer were aggregated to one observation with different variables describing the session. Thereby, a session is a period of sustained web browsing or a sequence of the user's page viewings until the user exits the online shop (Montgomery et al., 2004). The data comprise 3,511,037 observations or sessions between February 1, 2019 and April 30, 2019, i.e., three months. Further, the data contain 18 explanatory variables for each observation or session listed in Table 1 many of which are consistent with van den Poel and Buckinx' (2005) findings. We are only interested in visitors who made use of the virtual shopping cart during the session, i.e., who placed item(s) in their cart. In line with Close and Kukar-Kinney (2010), shopping cart usage is thus defined as necessary precondition for shopping cart abandonment. Thus, we filtered out customers which did not add any items to their shopping cart during the session, so-called just-browsing customers, and 821,048 observations (23,38%) remained. We modeled the dependent variable – shopping cart abandonment – as a dummy variable using the information about the customer's compiled and ordered shopping carts (variables BASKETS_BB and BASKETS) during the session:

$$\mathrm{Y} = \begin{cases} 1 \text{ if number of compiled shopping carts>0 \& number of ordered shopping carts=0;} \\ 0 \text{ if number of compiled shopping carts>0 \& number of ordered shopping carts>0.} \end{cases}$$

Our data contain 520,653 (63.41%) observations of shopping cart abandonments (or non-purchasers respectively) and 300,395 (36.59%) observations of purchasers. Hence, the dataset is relatively balanced. We excluded the variable for the number of ordered shopping carts (BASKETS_BB) and the value of ordered shopping carts (VALUE_BB) further for prediction[4].

Table 1: Variables of Clickstream Data.

Variable	Index	Description
Shopping Cart Abandonment	SCA	Dependent dummy variable capturing customer's shopping cart abandonment $Y = \begin{cases} 1 \text{ if customer abandoned;} \\ 0 \text{ otherwise.} \end{cases}$
Number of Ordered Shopping Carts	BASKETS_BB	Metric predictor variable capturing the number of shopping carts ordered during the customer's session
Number of Compiled Shopping Carts	BASKETS	Metric predictor variable capturing the number of shopping carts compiled during the customer's session
Number of Logins	LOGS	Metric predictor variable capturing the number of logins during the customer's session
Number of Existing Customers' Logins to the Second Step of the Ordering Process	LOGS_CUST_STEP2	Metric predictor variable capturing the number of logins of existing customers to the second step of the purchasing process during the customer's session
Number of New Customers' Logins to the Second Step of the Ordering Process	LOGS_NEWCUST_STEP2	Metric predictor variable capturing the number of logins of new customers to the second step of the purchasing process during the customer's session
Number of Overall Page Viewings	PIS	Metric predictor variable capturing the number of overall page viewings during the customer's session
Number of Shopping Cart Page Viewings	PIS_AP	Metric predictor variable capturing the number of shopping carts page viewings during the customer's session
Number of Detailed Product Page Viewings	PIS_DV	Metric predictor variable capturing the number of detailed product page viewings during the customer's session
Number of Category Overview Page Viewings	PIS_PL	Metric predictor variable capturing the number of category overview page viewings (i.e., all products within a category) during the customer's session
Number of Department Page Viewings	PIS_SHOPS	Metric predictor variable capturing the number of department page viewings (i.e., all categories within a department) during the customer's session
Number of Detailed Product Page Viewings Using Search Function	PIS_SDV	Metric predictor variable capturing the number of detailed product page viewings after using the search function during the customer's session

[4] These variables are values referring to the customers' order and, thus, they would not be known ex-ante for prediction.

Variable	Index	Description
Number of Search Results Page Viewings	PIS_SR	Metric predictor variable capturing the number of overall search results page viewings during the customer's session
Number of Product Types in Shopping Cart	POSITIONS	Metric predictor variable capturing different product types in the shopping cart during the customer's session
Number of Items in Shopping Cart	QUANTITY	Metric predictor variable capturing the number of items in the shopping cart during the customer's session
Value of Ordered Shopping Carts	VALUE_BB	Metric predictor variable capturing the value of shopping carts ordered during the customer's session
New Customer	NEW_CUST	Predictor dummy variable capturing new customers $X_{16} = \begin{cases} 1 \text{ if new customer;} \\ 0 \text{ otherwise.} \end{cases}$
Accessing Online Shop via Desktop	WEB_CUST	Predictor dummy variable capturing customers that access the online shop via desktop $X_{17} = \begin{cases} 1 \text{ if accessing via desktop;} \\ 0 \text{ otherwise.} \end{cases}$
Accessing Online Shop via Mobile Phone	MOBILE_CUST	Predictor dummy variable capturing customers that access the online shop via mobile phone $X_{18} = \begin{cases} 1 \text{ if accessing via mobile phone;} \\ 0 \text{ otherwise.} \end{cases}$

Figure 1 illustrates the relationship between the page viewing and login variables by demonstrating the customer's clickstream in the online shop: the customer typically starts browsing departments (PIS_SHOPS), then selects a certain category within a department (PIS_PL), and further, chooses a certain product within a category (PIS_DV). Optionally, the customer uses the shop's search engine (PIS_SR) to look systematically for a specific product (PIS_SDV). To make a purchase, the customer can either directly sign in (LOGS) or check the items in the shopping cart (PIS_AP) first and then sign in and hence, proceed to the second step of the purchasing process (LOGS_CUST_STEP2 or LOGS_NEWCUST_STEP2). However, signing in to the second step of the purchasing process does not necessarily lead to a purchase of the customer.

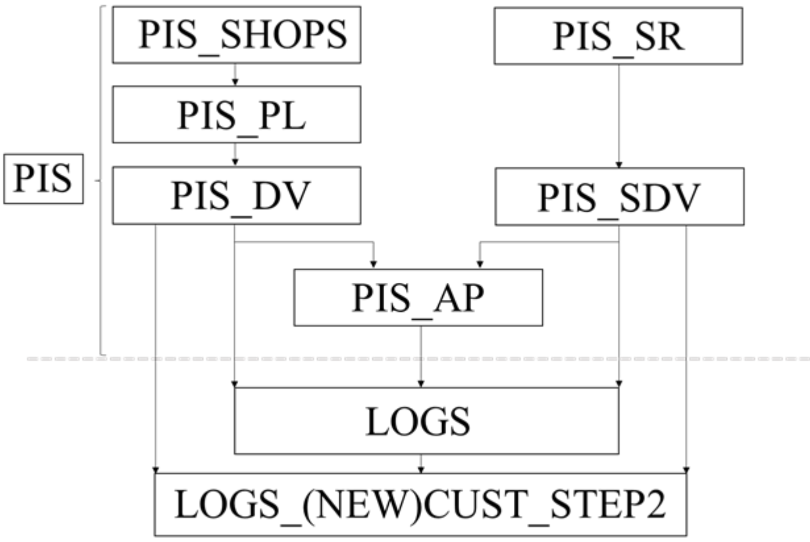

Figure 1: Main Clickstream of Customers in the Online Shop.

Note: LOGS = Number of Logins, LOGS_CUST_STEP2 = Number of Existing Customers' Logins to the Second Step of the Ordering Process, LOGS_NEWCUST_STEP2 = Number of New Customers' Logins to the Second Step of the Ordering Process, PIS = Number of Overall Page Viewings, PIS_AP = Number of Shopping Cart Page Viewings, PIS_DV = Number of Detailed Product Page Viewings, PIS_PL = Number of Category Overview Page Viewings, PIS_SDV = Number of Detailed Product Page Viewings Using Search Function, PIS_SHOPS = Number of Department Page Viewings, PIS_SR = Number of Search Results Page Viewings.

Nevertheless, with respect to the descriptive statistics in Table 2, we find that existing customers (or new customers respectively) which subsequently make a purchase sign in to the second step of the ordering process approximately 5.93 times (or 4.46 times respectively) more often than non-purchasers. Generally, purchasers sign in more often (1.03 logins on average) than non-purchasers (0.93 logins on average). This might indicate that the cause for shopping cart abandonment frequently occurs before the customer proceeds to the checkout stage.

Furthermore, the number of purchasers' overall page viewings is 2.09 times higher than of non-purchasers on average. Overall, customers abandoning their shopping cart browse less pages than purchasers – regardless of the pages' type. Particularly, the median reveals that there are significant differences regarding the number of page viewings between purchasers and abandoners: the median of abandoners' overall page viewings is 12, 1 for department viewings, and 0 for all other types of page viewings. In contrast, purchasers' median for overall page viewings is 35, 6 for department viewings, and for example, 2 for shopping cart viewings.

On average, purchasers add more items and different product types (3.48 and 3.38 respectively) to their shopping cart than non-purchasers (2.95 and 2.88 respectively).

There is a larger absolute (48,839) and relative (9.38%) proportion of new customers among the observations of shopping cart abandonments than among those making a purchase (15,387 observations or 5.12% respectively). Moreover, there is a larger proportion of mobile shoppers among customers abandoning their shopping cart (45.85%) compared to the observations of purchasers (28.1%). The latter descriptive findings are consistent with the results of preceding (behavioral) research: e.g., as argued earlier, Huang, Korfiatis, and Chang (2018) proved that online shopping cart abandonment occurs more frequently for customers using a mobile device due to high emotional ambivalence. Moe and Fader (2004a) found that – among new customers – online conversion rate is lower as purchasing thresholds and perceived risks are high for unexperienced visitors.

Table 2: Descriptive Statistics of Clickstream Data.

Variable	Observations of Shopping Cart Abandonments (n=520,653)					Observations of Purchasers (n=300,395)				
	Mean	**SD**	**Median**	**Min**	**Max**	**Mean**	**SD**	**Median**	**Min**	**Max**
BASKETS	0.99	0.11	1	0	2	1.09	0.40	1	0	49
LOGS	0.93	0.27	1	0	2	1.03	0.21	1	0	2
LOGS_CUST_STEP2	0.06	0.23	0	0	1	0.32	0.47	0	0	1
LOGS_NEWCUST_STEP2	0.02	0.13	0	0	1	0.07	0.26	0	0	1
PIS	22.26	27.47	12	1	513	46.45	37.06	35	2	593
PIS_AP	1.05	2.17	0	0	71	3.06	3.42	2	0	57
PIS_DV	3.42	7.48	0	0	200	6.53	9.76	3	0	203
PIS_PL	3.99	11.37	0	0	279	8.75	17.00	1	0	315
PIS_SHOPS	7.68	17.55	1	0	405	15.87	25.14	6	0	396
PIS_SDV	1.40	3.92	0	0	142	3.13	5.46	1	0	127
PIS_SR	2.82	7.48	0	0	222	5.71	10.18	2	0	208
POSITIONS	2.88	3.31	2	1	66	3.38	3.31	2	1	111
QUANTITY	2.95	3.55	2	1	143	3.48	3.49	2	1	143
	Counts		**Proportion**			**Counts**		**Proportion**		
NEW_CUST	48,839		9.38%			15,387		5.12%		
WEB_CUST	214,455		41.29%			171,789		57.19%		
MOBILE_CUST	238,694		45.85%			84,401		28.1%		

Note: BASKETS = Number of Carts Compiled, LOGS = Number of Logins, LOGS_CUST_STEP2 = Number of Existing Customers' Logins to the Second Step of the Ordering Process, LOGS_NEWCUST_STEP2 = Number of New Customers' Logins to the Second Step of the Ordering Process, MOBILE_CUST = Customer Accessing via Mobile Phone**,** NEW_CUST = New Customer, PIS = Number of Overall Page Viewings, PIS_AP = Number of Shopping Cart Page Viewings, PIS_DV = Number of Detailed Product Page Viewings, PIS_PL = Number of Category Overview Page Viewings, PIS_SDV = Number of Detailed Product Page Viewings Using Search Function, PIS_SHOPS = Number of Department Page Viewings, PIS_SR = Number of Search Results Page Viewings, POSITIONS = Number of Product Types, QUANTITY = Number of Items, WEB_CUST = Customer Accessing via Desktop.

4.2 Experimental Setup

Since each machine learning approach and its subsequent refinements and modifications exhibit individual strengths and weaknesses in dependence of the underlying data and the requested task it is highly recommended in the machine learning literature to compare and test different algorithms

(Moro et al., 2014; Razi and Athappilly, 2005). Thus, we compared different models of those proposed in Section 3 to predict shopping cart abandonment for our data, listed in Table 3. Additionally, we included a standard logistic regression model in our comparison serving as a non-machine learning benchmark method.

Table 3: Machine Learning Approaches for Comparison.

Approach	**Description**
Adaptive Boosting (AdaBoost)	Ensemble of weak learners, algorithm puts higher weights on incorrectly classified instances
Boosted Logistic Regression (LogitBoost)	Algorithm applies logistic regression techniques to the AdaBoost method by minimizing the logistic loss
Decision Tree (DT)	Algorithm recursively partitions the predictor space into subsets in which the distribution of the dependent variable is successively more homogeneous
Gradient Boosting (Linear Base Learner) with L1 and L2 Regularization (GBReg)	Ensemble of weak learners (with linear base learners), algorithm applies L1 (Lasso Regression) and L2 (Ridge Regression) Regularization
Gradient Boosting (Tree Base Learner) (GBTree)	Ensemble of weak learners (with tree base learners), algorithm minimizes the model's loss by adding weak learners sequentially using a gradient descent like procedure
Gradient Boosting (Tree Base Learner) with Dropout (GBDropout)	See GBTree, but the algorithm randomly drops boosting tree members
k-Nearest Neighbor (KNN)	Algorithm classifies an observation by assigning it to the class most common among its k nearest neighbors
Multi-Layer Perceptron Network with Dropout (MLPDropout)	Feedforward Neural Network with dropout regularization technique
Naïve Bayes (NB)	Algorithm is based on the Bayes' theorem and classifies an observation by deriving the maximum a posteriori probability
Random Forest (RF)	Ensemble of decision trees, algorithm predicts new data by aggregating the predictions of the trees
Stochastic Gradient Boosting (SGB)	Algorithm fits base learner at each iteration on the subsample of the data – instead of the full – drawn at random without replacement
Support Vector Machine with Radial Basis Kernel (SVMRadial)	Support vector machine implementation with radial basis kernel

To estimate and, hence, validate the models, we randomly partitioned the data into a training and a test subset in a 67/33 ratio, i.e., 67% (or 550,098 observations respectively) of the data are used as training data and 33% (or 270,950 observations respectively) are used as test data.

We performed k-fold cross-validation with the training data to fit the models and optimized their hyperparameters respectively (Geisser, 1975; Stone, 1974): the sample, i.e., the training data, is randomly split into k equal sized subsamples $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_k$. Of the k subsamples, one single subsample is retained as validation data to test the fitted model subsequently and the remaining $k-1$ subsamples are used as training data to fit the model. This step is repeated k times with each of the k subsamples serving as validation data once. Drawing on machine learning literature, $k = 10$ is frequently utilized since it provides an adequate trade-off between method's variance and method's bias (i.e., trade-off between the estimated parameter's expected value and the estimated value) (Bradley, 1997; Breiman, 1996; Kohavi, 1995; Tibshirani and Tibshirani, 2009; Zhang, 1993). Thus, we applied 10-fold cross-validation.

Further, to validate and evaluate our models' performance, we considered different performance metrics that indicate the models' predictive ability. In a binary decision problem, the classifier labels observations as either positive or negative. Consequently, the classification procedure yields four different outputs in a $2x2$ confusion matrix: the sample is either correctly classified as positive (true positive (TP)), correctly classified as negative (true negative (TN)), falsely classified as positive (false positive (FP) or Type II error), or falsely classified as negative (false negative (FN) or Type I error). Thereby, accuracy is one of the most commonly used measures for classification performance due to its simplicity (see e.g., Kohavi (1995)). It is the ratio between correctly classified samples to the total number of samples:

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{P+N}}$$

However, recent research shifted away from solely presenting accuracy results since accuracy assumes balanced class distribution and equal error costs (i.e., Type I errors are equivalent to Type II errors) which is rarely the case in real world applications (Davis and Goadrich, 2006; Provost and Fawcett, 1997). To address these problems, a receiver operating characteristics (ROC) curve and thus, the area under the ROC curve (AUC)[5] have been increasingly used by the machine learning community since they are insensitive to changes in class distributions and scale-invariant (Bradley, 1997; Fawcett, 2006). A ROC graph is a two-dimensional depiction of classification performance to measure different classifiers' performances and captures the trade-off between benefits (i.e., true positives) and costs (i.e., false positives) (Fawcett, 2006). It is created by plotting the true positive rate (TPR)

[5] In literature, the area under the ROC curve is frequently referred to as AUROC instead of AUC.

(or sensitivity or recall respectively) against the false positive rate (FPR) (or $1 - $ specificity respectively) (Bradley, 1997; Fawcett, 2006; Hand, 2009; Provost and Fawcett, 2001):

$$\text{TPR}=\text{Sensitivity}=\text{Recall}=\frac{\text{TP}}{\text{P}};\quad \text{FPR}=1\text{-Specificity}=\frac{\text{FP}}{\text{N}};\quad \text{Specificity}=\frac{\text{TN}}{\text{N}}$$

The classifier's AUC value is a portion of the area of the unit square and its value ranges from 0.0 to 1.0 (perfect classification). It should be higher than 0.5 which equals the AUC of an uninformative classifier (Bradley, 1997; Fawcett, 2006). An important statistical property of the AUC is that a classifier's AUC is equivalent to the probability that the classifier will rank a randomly chosen positive observation higher than a randomly chosen negative observation (Fawcett, 2006).

An alternate performance measure is the F_1-Score comprising both precision and recall:

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}};\quad F_1=2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}}$$

Ideally, the performance measure is chosen by properly reflecting the investigation's aims to avoid misleading conclusions. Since our data is relatively balanced it seems reasonable to consider accuracy as a basic performance metric. However, as we intend to convert customers abandoning their shopping carts into purchasers our main aim is to correctly classify actual positives (i.e., observations of shopping cart abandonments) by minimizing the Type I error. Consequently, the higher the recall the less false negatives (i.e., shopping cart abandonments classified as purchasers) have been predicted. Besides, we intend to maximize the proportion of actual positives among the predicted positives by minimizing the Type II error, i.e., purchasing customers should not be classified as non-purchasers. Thus, the higher the precision the less false positives have been predicted. The F_1-Score considers the trade-off between recall and precision. Therefore, we determined the F_1-Score, recall, and precision as our main performance metrics for the test data. Additionally, to yield valid results, we considered the ROC curve or the AUC respectively as a performance metric since it is a common measure of separability capturing the trade-off between both TPR (or sensitivity or recall respectively, analogous to F_1-Score) and FPR (i.e., how many negative instances are falsely classified as positive among the negative instances). For the training data, the best classifier during hyperparameter optimization was automatically chosen based on the AUC values.

Although prediction accuracy (i.e., AUC, F_1-Score, and accuracy) is frequently the main decision criterion when comparing different machine learning models, the models' complexity in terms of computation time and computation effort (e.g., numbers of hyperparameters to be optimized) is of similar importance regarding the application in practice and should therefore be considered as well (Doshi-Velez and Kim, 2017; Guidotti et al., 2019; Tambe et al., 2019).

5 Findings

Drawing on the training results in Table 4, gradient boosting with regularization outperformed the remaining approaches with an AUC of 0.9008. The final gradient boosting model's fitted hyperparameters did not include the lasso regression technique (L1 regularization) but made use of the ridge regression technique (L2 regularization). The gradient boosting with tree base learners and random forest yielded comparable results (AUC of 0.8953 and 0.8954 respectively) whereas naïve bayes and boosted logistic regression realized the lowest AUC values (0.8218 and 0.8381 respectively).

Regarding estimation time, the benchmark logistic regression, decision tree, and boosted logistic regression performed the fastest 10-fold cross validation to optimize the hyperparameters (20.3, 225.07, and 380.0 seconds respectively). The support vector machine and adaptive boosting were the most time-consuming models to estimate (1,306,838.6 and 703,903.9 seconds respectively). Gradient boosting with regularization yielded a moderate estimation time (4,021.28 seconds) and thus, provides an appropriate trade-off between AUC and estimation time.

Table 4: Training Data Results.

Model	Fitted Parameters	AUC

	Parameter	Fitted Value		Estimation Time (Seconds)[6]
Logistic Regression			0.8003	20.3
AdaBoost	Number of Trees	50	0.8698	703,903.9
	Method	Adaboost.M1		
LogitBoost	Number of Boosting Iterations	21	0.8381	380.0
DT	Complexity Parameter	0.0129	0.7988	225.07
GBReg	Number of Boosting Iterations	150	**0.9008**	4,021.28
	L2 Regularization	0.1		
	L1 Regularization	0		
	Learning Rate	0.3		
GBTree	Number of Boosting Iterations	150	0.8953	6,701.14
	Maximum Tree Depth	3		
	Shrinkage	0.4		
	Minimum Loss Reduction	0		
	Subsample Ratio of Columns	0.8		
	Minimum Sum of Instance Weight	1		
	Subsample Percentage	1		
GBDropout	Number of Boosting Iterations	150	0.8952	49,794.27
	Maximum Tree Depth	3		
	Shrinkage	0.4		
	Minimum Loss Reduction	0		
	Subsample Ratio of Columns	0.8		
	Minimum Sum of Instance Weight	1		
	Subsample Percentage	0.75		
	Fraction of Trees dropped	0.01		
	Probability of Skipping Dropout	0.95		
KNN	Maximum Number of Neigh-bors	30	0.8828	127,773.4
	Distance	2		
	Kernel	Optimal		
MLPDropout	Number of Hidden Units	768	0.8807	218,894.0
	Dropout Rate	0.35		
	Batch Size	64		
	Learning Rate	0.000006		
	Rho	0.2		
	Learning Rate Decay	0		
	Activation Function	Sigmoid		
	Epochs	30		
NB	Laplace Correction	0	0.8218	5,757.49
	Distribution Type	Kernel Density Estimation		
	Bandwidth Adjustment	0.3		
RF	Number of Randomly Selected Predictors	14	0.8954	171,587.7
	Splitting Rule	Gini		
	Minimal Node Size	35		
SGB	Number of Boosting Iterations	150	0.8800	2,033.17

[6] With 40 GB RAM.

	Maximum Tree Depth	3		
	Shrinkage	0.1		
	Minimum Terminal Node Size	10		
SVMRadial	Sigma Cost	0.1818 0.5	0.8808	1,306,838.6

Note: The highest AUC value is marked in bold. AdaBoost = Adaptive Boosting, DT = Decision Tree, GBDropout = Gradient Boosting with Dropout, GBReg = Gradient Boosting with L1 and L2 Regularization, GBTree = Gradient Boosting with Tree Base Learners, KNN = k-Nearest Neighbor, LogitBoost = Boosted Logistic Regression, MLPDropout = Multi-Layer Perceptron Network with Dropout, NB = Naïve Bayes, RF = Random Forest, SGB = Stochastic Gradient Boosting, SVMRadial = Support Vector Machine with Radial Basis Kernel.

Since we are rather interested in the fitted models' performances on new and unknown data, the test data results in Table 5 exhibit a higher practical relevance than the preceding results: similarly to the training data results, the gradient boosting model with regularization was superior to the remaining models regarding the test data. It yielded the best AUC (0.8182) and accuracy (82.29%) results. In line with these findings, the F_1-Score (0.8569) proves that the model is the most suitable approach in our comparison to balance the trade-off between precision and recall. With respect to its confusion matrix in the Appendix, the gradient boosting model classified 28,209 abandonments falsely as purchasers (16.42% of all abandonments) and 19,767 purchasers as abandonments respectively (19.94% of all purchasers). This is further reflected by the model's precision (0.8790) and recall (0.8358), i.e., there is a high proportion of both correctly predicted abandonments among all correctly and falsely predicted abandonments (87.90%) and correctly predicted abandonments among all actual abandonments (83.58%).

Although naïve bayes realized an extremely high recall (0.9996), its precision (0.6351) is just slightly better than random guessing. This is due to its negligible Type I error (i.e., 68 abandonments classified as purchasers (0.0004% of all abandonments)) and its substantial Type II error (i.e., 98,677 purchasers classified as abandonments (99.52% of all purchasers)). Consequently, by focusing exclusively either on precision or recall, one could draw misleading conclusions regarding model selection. The F_1-Score of the naïve bayes model (0.7767) reveals that it constitutes a suboptimal choice.

Similarly, albeit the decision tree classified a high proportion of purchasers correctly and only 12,688 (i.e., 12.80% of all purchasers) wrong, it categorized 55,634 cart abandonments as purchasers (i.e., 32.38% of all abandonments). Thus, due to its high Type I error, its recall is extremely low (0.6762), but it realized the highest precision value of all models (0.9015).

Generally, our results indicate a substantial predictive ability of the most tree-based methods (i.e., gradient boosting with regularization (and linear base learners), gradient boosting (with tree base learners), gradient boosting with dropout (and tree base learners), and random forest) compared with the remaining machine learning approaches. The latter were outperformed by tree-based models with regard to all relevant performance metrics (AUC, accuracy, and F_1-Score).[7]

Logistic regression as a non-machine learning benchmark approach yielded the lowest F_1-Score but realized a higher AUC value than several other machine learning approaches like boosted logistic regression, k-nearest neighbor, multi-layer perceptron, naïve bayes, and support vector machine. Nevertheless, it did not perform better than the tree-based methods (except for adaptive boosting, decision tree, and stochastic gradient boosting) with regard to AUC.

Moreover, the k-nearest neighbor algorithm as a basic machine learning approach outperformed more sophisticated algorithms like the multi-layer perceptron, the stochastic gradient boosting, and adaptive boosting with respect to its AUC value (0.7962).

Table 5: Test Data Results.

Model	Performance Metrics				
	AUC	**Accuracy**	**Precision**	**Recall**	**F_1-Score**
Logistic Regression	0.8012	78.94%	0.6677	0.8454	0.7461
AdaBoost	0.7516	78.54%	0.8024	0.8777	0.8384
LogitBoost	0.7623	77.19%	0.8349	0.7981	0.8161
DT	0.7741	74.78%	**0.9015**	0.6762	0.7728
GBReg	**0.8182**	**82.29%**	0.8790	0.8358	**0.8569**
GBTree	0.8105	81.78%	0.8701	0.8377	0.8536
GBDropout	0.8123	81.84%	0.8731	0.8350	0.8536
KNN	0.7962	80.5%	0.8585	0.8290	0.8435
MLPDropout	0.7911	80.36%	0.8503	0.8378	0.8440
NB	0.5022	63.56%	0.6351	**0.9996**	0.7767
RF	0.8108	81.75%	0.8711	0.8359	0.8531
SGB	0.7902	80.08%	0.8521	0.8299	0.8409
SVMRadial	0.7956	81.23%	0.8479	0.8578	0.8528

Note: For each column, the highest value is marked in bold. AdaBoost = Adaptive Boosting, DT = Decision Tree, GBDropout = Gradient Boosting with Dropout, GBReg = Gradient Boosting with L1 and L2 Regularization, GBTree = Gradient Boosting with Tree Base Learners, KNN = k-Nearest Neighbor, LogitBoost = Boosted Logistic Regression, MLPDropout = Multi-Layer Perceptron Network with Dropout, NB = Naïve Bayes, RF = Random Forest, SGB = Stochastic Gradient Boosting, SVMRadial = Support Vector Machine with Radial Basis Kernel.

[7] Tree-based approaches are typically not subject to multicollinearity (Climent et al. 2019). Thus, we did not remove any correlated variables during the training process.

6 Discussion

Our findings contribute to a deeper understanding regarding the successful implementation of machine learning methods for predicting online shopping cart abandoners with a strong forecast performance in order to apply marketing techniques in real-time to convert them to purchasers. Thus, we discuss our findings' theoretical contribution and practical implications in this Section. We also discuss limitations and propose suggestions for future research.

6.1 Theoretical Contribution

Overall, we fill a research gap by identifying suitable machine learning approaches for online shopping cart abandonment prediction not only in terms of accuracy but, further, in terms of practicability. Thereby, we contribute to literature in several ways. First, we are able to characterize customers abandoning their shopping cart descriptively with our data. Preceding literature on shopping cart abandonment (e.g., Close and Kukar-Kinney (2010), Huang et al. (2018), or Kukar-Kinney and Close (2010)) primarily shed light on behavioral aspects of the abandonment process with experimental designs. In contrast, our research deals with unbiased clickstream data comprising an exceptionally high number of observations. Our data indicate that there is a higher proportion of new customers and mobile shoppers among customers abandoning their shopping carts compared to purchasers whereas the latter add more items to their shopping cart and view an increased number of pages on average.

Second, we contribute to literature by proposing a broad range of machine learning models to compare their performance regarding online shopping cart abandonment prediction and, thus, to predict future customers abandoning their shopping carts in real-time. Prior literature either drew on a behavioral perspective to understand the antecedents of shopping cart abandonment or predicted – more generally – purchase behaviors with conservative approaches and less observations (see e.g., Huang et al. (2018), Kukar-Kinney and Close (2010), or Sismeiro and Bucklin (2004)). For our data, the gradient boosting with regularization yielded the highest accuracy (82.29%). However, with respect to our main aim, to minimize the Type I error (i.e., abandoners falsely classified as purchasers) and the Type II error (i.e., purchasers falsely classified as abandoners), we focused on the F_1-Score capturing the trade-off between precision and recall. Consistent with the accuracy results, the gradient boosting with regularization outperformed the remaining models regarding the F_1-Score (0.8569). Additionally, it realized the highest AUC value (0.8182) compared to the other models.

Overall, we found tree-based methods to be superior to the remaining machine learning approaches and logistic regression as a benchmark non-machine learning approach aligning with prior research comparing machine learning approaches in different application fields like customer churn prediction or phishing detection (Abu-Nimeh et al., 2007; Caruana and Niculescu-Mizil, 2006; Vafeiadis et al., 2015) and – similar to our context – prediction of online purchase intention (Bogina et al., 2019; Boroujerdi et al., 2014; Zheng and Liu, 2018). Thus, we complement the literature on machine learning comparisons in a marketing context.

Moreover, despite the striking importance of prediction accuracy as a decision criterion for appropriate machine learning approaches, the models' practicability with respect to modeling complexity as an essential criterion is of particular importance (Doshi-Velez and Kim, 2017; Guidotti et al., 2019; Tambe et al., 2019) but, at the same time, is often neglected by current research. Thus, we considered the models' complexity in terms of computation time and computation effort (e.g., numbers of hyperparameters to optimize) to add to literature. Thereby, the decision tree approach and boosted logistic regression yielded only slightly worse AUC results compared to gradient boosting with regularization and, simultaneously, their complexity in terms of both computation effort and time was rather low. Hence, in case of online shopping cart abandonment prediction, a decision tree model and boosted logistic regression perform well in balancing the trade-off between accuracy and complexity. Further, as stated by prior literature, we found the support vector machine approach to be extremely computationally infeasible (L'Heureux et al., 2017) despite its acceptable prediction accuracy.

6.2 Practical Implications

Our research may help to gather a comprehensive understanding of machine learning approaches for prediction or classification, particularly with regard to online shopping cart abandonment prediction. More specifically, our research provides multifold practical implications for decision makers.

Since research about advanced machine learning approaches in marketing contexts is still in its infancy (e.g., Cheung et al. (2003) and Cui et al. (2006)) we reviewed relevant literature to provide an introduction to such models, its potential applications, as well as performance metrics, and common methods for validation: for machine learning models, k-fold cross-validation is a common method to optimize the models' hyperparameters. Decision makers should draw on either recall as a performance measure if their main aim is to correctly classify abandonments or precision if they intend to avoid falsely classified purchasers. The F_1-Score considers the trade-off between both. Besides, the

AUC is a common measure of separability since it is insensitive to skewed class distributions. Overall, tree-based approaches and particularly boosting methods are superior to the remaining machine learning models regarding forecast accuracy within online shopping cart abandonment prediction. Random forest yields comparable results but is rather time-consuming to estimate (171,587.7 seconds estimation time). The support vector machine and adaptive boosting are computationally intensive with estimation times of 1,306,838.6 and 703,903.9 seconds respectively.

Aside from pointing out methodological aspects, we drew on an economical perspective to enhance an organization's turnover: with regard to our data, the mean value of purchasers' ordered shopping carts (VALUE_BB) is 271.73 euro and they add 3.479 items into their shopping cart on average and thus, we expect the online retailer's sales loss for each shopping cart abandonment to be around 230 euro with 2.945 items in their shopping cart on average. Therefore, we determined a suitable approach to correctly identify shopping cart abandonments as well as purchasers: our findings indicate that gradient boosting with regularization outperformed the remaining approaches. Organizations can implement this method to predict non-purchasers in real-time when a sufficient amount of information about the customer's activities during the session has been collected. Overall, we found particularly tree-based machine learning approaches such as random forest or gradient boosting to outperform traditional classification approaches such as logistic regression and decision tree, which are frequently utilized by practitioners.

Drawing on an overall practicability perspective, decision makers may take a slight loss in prediction accuracy into account if, instead, the model's complexity in terms of computation time and effort is substantially lower: in our application context, decision tree and boosted logistic regression yielded acceptable prediction results and their computation effort was substantially lower compared to gradient boosting methods.

6.3 Limitations and Future Research

Our research is subject to limitations which stimulate further research. First, the set of useful variables for prediction was limited. With respect to extant literature (see e.g., Bucklin and Sismeiro (2003), Moe and Fader (2004a), or van den Poel and Buckinx (2005)), we expect e.g. demographic variables, historical purchase behavior, or the time customers spend on the single pages to be informative variables. Further, we did not have information about the customers' identity and thus, could not determine whether there were recurring customers. However, this information could be of great interest for analyzing online behavior and predicting shopping cart abandonment. For instance, Huang et al.

(2018) anticipated that some customers might use the mobile phone for initial purchase stages (i.e., browsing and collecting information) and then switch to the computer for completing the purchase. However, such customers are listed as two distinct sessions in the current data. Another missing information concerns the value of abandoned shopping carts. While there is a variable that indicates the value of ordered carts (i.e., VALUE_BB), the value of abandoned carts can only be estimated. In line with extant literature on shopping cart abandonment (e.g., Close and Kukar-Kinney (2010); Kukar-Kinney and Close (2010)), it can be assumed that the value of ordered items influences abandoning rates and, thus, could aid the prediction of such. Moreover, if detailed information about spent time and further, the chronological order of customers' actions in the online shop would be available, we could decompose the session into sequences or segments. Then, we could determine a critical point in the customer's session in which abandonment can be predicted reliably with the F_1-Score or the AUC exceeding a defined threshold (see e.g., Sismeiro and Bucklin (2004)). Hence, future research could replicate the present study with more detailed data, e.g. between-site clickstream data (i.e., panel data collected by media measurement company), that are typically more comprehensive and frequently used in clickstream analyses (see e.g., Moe and Fader (2004a)).

Second, we excluded just-browsing customers from our investigation. A possible direction for future research could be to conduct a multi-class classification by differentiating between purchasers, abandonments, and just-browsing customers, similar to the cluster analysis of Moe (2003).

Third, the models' performance strongly depends on the optimized hyperparameters which may be a time-consuming procedure for some of the models. Therefore, we considered only a limited range of possible hyperparameter values. Moreover, other values of k in cross-validation could lead to different results.

Lastly, a real-time implementation requires a certain amount of data to be collected before the model can make a reliable decision.

By implementing these models, companies may detect shopping cart abandoners in real-time and subsequently convert some of them into purchasers by making use of targeted marketing measures such as individual chat pop-ups, coupons or special discounts. For instance, Close and Kukar-Kinney (2010) suggest human-human interactions (i.e., live chats with employees or other online shoppers) to avoid shopping cart abandonment. These could pop-up on the website if the online user is predicted to abandon by the machine learning model. Therefore, future research is recommended to test whether pop-up messages and offers impact customers' online shopping behavior and can prevent online shopping cart abandonment.

7 Conclusion

Online shopping cart abandonment can inhibit corporate growth and hence, harm a company's success within its competitive environment. Simultaneously, the emergence of the Internet's commercial usage leads to the ability to track consumers' online activities and online behavior resulting in clickstream data.

Thus, to identify online shopping cart abandoners by extracting valuable knowledge from such clickstream data we proposed different machine learning approaches. We analyzed data of a German online retailer comprising 821,048 observations and fitted the models using 10-fold cross validation. Thereby, our paper contributes to extant literature by combining research fields of both online shopping cart abandonment and clickstream data with machine learning approaches.

Our data indicate that among customers abandoning their shopping carts there is a higher proportion of new customers and mobile shoppers compared to purchasers whereas the latter add more items to their shopping cart and have a higher number of page viewings on average. Moreover, our comparison results prove that gradient boosting with regularization is a suitable method to distinguish between abandonments and purchasers yielding an AUC of 0.8182, an F_1-Score of 0.8569, and an accuracy of 82.29%. Nevertheless, a decision tree or boosted logistic regression may be suitable alternatives yielding only slightly less accurate prediction results and being computationally more feasible.

Nevertheless, research on clickstream data combined with machine learning approaches is still in its infancy – particularly in a marketing context. Thereby, machine learning will be inevitable for e-commerce businesses to be successful in the long-term and the analysis provided in this paper shall stimulate further research on this topic.

Appendix: Confusion Matrices

Model	Prediction	Actual	
		0 (Purchaser)	**1 (Abandonment)**
Logistic Regression	**0 (Purchaser)**	83,817	41,722
	1 (Abandonment)	15,335	130,076
AdaBoost	**0 (Purchaser)**	62,009	21,005
	1 (Abandonment)	37,143	150,793
LogitBoost	**0 (Purchaser)**	72,036	34,692
	1 (Abandonment)	27,116	137,106
DT	**0 (Purchaser)**	86,464	55,634
	1 (Abandonment)	12,688	116,164
GBReg	**0 (Purchaser)**	79,385	28,209
	1 (Abandonment)	19,767	143,589
GBTree	**0 (Purchaser)**	77,662	27,875
	1 (Abandonment)	21,490	143,923
GBDropout	**0 (Purchaser)**	78,294	28,352
	1 (Abandonment)	20,858	143,446
KNN	**0 (Purchaser)**	75,687	29,383
	1 (Abandonment)	23,465	142,415
MLPDropout	**0 (Purchaser)**	73,803	27,869
	1 (Abandonment)	25,349	143,929
NB	**0 (Purchaser)**	475	68
	1 (Abandonment)	98,677	171,730
RF	**0 (Purchaser)**	77,903	28,197
	1 (Abandonment)	21,249	143,601
SGB	**0 (Purchaser)**	74,409	29,217
	1 (Abandonment)	24,743	142,581
SVMRadial	**0 (Purchaser)**	72,724	24,427
	1 (Abandonment)	26,428	147,371

References

Abu-Nimeh S, Nappa D, Wang X, et al. (2007) A comparison of machine learning techniques for phishing detection. In: *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit:* New York, NY: ACM.

Bengio Y (2009) *Learning deep architectures for AI*. Boston: Now Publishers Inc.

Bogina V, Mokryn O and Kuflik T (2019) Will this session end with a purchase? Inferring current purchase intent of anonymous visitors. *Electronic Commerce Research and Applications* 34: 100836.

Boroujerdi EG, Mehri S, Sadeghi Garmaroudi S, et al. (2014) A study on prediction of user's tendency toward purchases in websites based on behavior models. In: *2014 6th Conference on Information and Knowledge Technology (IKT): 27 - 29 May 2014, Shahrood*. Piscataway, NJ: IEEE.

Boser BE, Guyon IM and Vapnik VN (1992) A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*: 144–152.

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7): 1145–1159.

Breiman L (1996) Bagging predictors. *Machine Learning* 24(2): 123–140.

Breiman L (2001) Random Forests. *Machine Learning* 45(1): 5–32.

Breiman L, Friedman J, Olshen R, et al. (1984) *Classification and regression trees*. New York: Chapman & Hall.

Bucklin RE and Sismeiro C (2003) A Model of Web Site Browsing Behavior Estimated on Clickstream Data. *Journal of Marketing Research* 40(3): 249–267.

Caruana R and Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: Cohen W (ed.) *Proceedings of the 23rd international conference on Machine learning:* New York, NY: ACM.

Chatterjee P, Hoffman DL and Novak TP (2003) Modeling the Clickstream: Implications for Web-Based Advertising Efforts. *Marketing Science* 22(4): 520–541.

Chen H, Chiang RHL and Story VC (2012) Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly* 36(4): 1165–1188.

Cheung K-W, Kwok JT, Law MH, et al. (2003) Mining customer product ratings for personalized marketing. *Decision Support Systems* 35(2): 231–243.

Chipman HA, George EI and McCulloch RE (1998) Bayesian CART Model Search. *Journal of the American Statistical Association* 93(443): 935–948.

Cho C-H, Kang J and Cheon HJ (2006) Online Shopping Hesitation. *CyberPsychology & Behavior* 9(3): 261–274.

Chollet F and Allaire JJ (2018) *Deep Learning with R*. Shelter Island: Manning Publications.

Climent F, Momparler A and Carmona P (2019) Anticipating bank distress in the Eurozone: An Extreme Gradient Boosting approach. *Journal of Business Research* 101: 885–896.

Close AG and Kukar-Kinney M (2010) Beyond buying: Motivations behind consumers' online shopping cart use. *Journal of Business Research* 63(9-10): 986–992.

Cortes C and Vapnik V (1995) Support-vector networks. *Machine Learning* 20(3): 273–297.

Cover T and Hart P (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1): 21–27.

Cui G, Wong ML and Lui H-K (2006) Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. *Management Science* 52(4): 597–612.

Davis J and Goadrich M (2006) The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*: 233–240.

Deng L (2011) An Overview of Deep-Structured Learning for Information Processing. *Proceedings of Asian-Pacific Signal & Information Processing Annual Summit and Conference (APSIPA-ASC)*: 1–14.

Dietterich TG (2002) Ensemble Learning. In: Arbib MA (ed.) *The Handbook of Brain Theory and Neural Networks:* Cambridge, MA: The MIT Press, pp. 110–125.

Doshi-Velez F and Kim B (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Duda RO, Hart PE and Stork DG (1973) *Pattern classification and scene analysis*. New York: Wiley.

Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognition Letters* 27(8): 861–874.

Fayyad U, Piatetsky-Shapiro G and Smyth P (1996) From Data Mining to Knowledge Discovery in Databases. *AI magazine* 17(3): 37–54.

Freund Y and Schapire RE (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55: 119–139.

Friedman J (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29(5): 1189–1232.

Friedman J (2002) Stochastic Gradient Boosting. *Computational statistics & data analysis* 38(4): 367–378.

Friedman J, Hastie TH and Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. *The Annals of Statistics* 28(2): 337–407.

Geisser S (1975) The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association* 70(350): 320–328.

Guidotti R, Monreale A, Ruggieri S, et al. (2019) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* 51(5): 1–42.

Hand DJ (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* 77(1): 103–123.

Hinton GE, Osindero S and Teh Y-W (2006) A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation* 18(7): 1527–1554.

Hochreiter S and Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8): 1735–1780.

Huang G-H, Korfiatis N and Chang C-T (2018) Mobile shopping cart abandonment: The roles of conflicts, ambivalence, and hesitation. *Journal of Business Research* 85: 165–174.

Jain AK, Mao J and Mohiuddin KM (1996) Artificial Neural Networks: A Tutorial. *Computer* 29(3): 31–44.

James G, Witten D, Hastie T, et al. (2013) *An Introduction to Statistical Learning*. New York, NY: Springer New York.

Kelly S and Noonan C (2017) The doing of datafication (and what this doing does): Practices of edification and the enactment of new forms of sociality in the Indian public health service. *Journal of the Association for Information Systems* 18(2): 872–899.

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artifical Intelligence (IJCAI)*: 1137–1145.

Kohavi R and Provost F (2001) Applications of Data Mining to Electronic Commerce. In: *Applications of Data Mining to Electronic Commerce:* Springer, Boston, MA, pp. 5–10.

Kukar-Kinney M and Close AG (2010) The determinants of consumers' online shopping cart abandonment. *Journal of the Academy of Marketing Science* 38(2): 240–250.

L'Heureux A, Grolinger K, Elyamany HF, et al. (2017) Machine Learning With Big Data: Challenges and Approaches. *IEEE Access* 5: 7776–7797.

Liaw A and Wiener M (2002) Classification and Regression by randomForest. *R News* 2(3): 18–22.

Lycett M (2013) 'Datafication': making sense of (big) data in a complex world. *European Journal of Information Systems* 22(4): 381–386.

Moe WW (2003) Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream. *Journal of Consumer Psychology* 13(1-2): 29–39.

Moe WW and Fader PS (2004a) Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing* 18(1): 5–19.

Moe WW and Fader PS (2004b) Dynamic Conversion Behavior at E-Commerce Sites. *Management Science* 50(3): 326–335.

Montgomery AL (2001) Applying Quantitative Marketing Techniques to the Internet. *Interfaces* 31(2): 90–108.

Montgomery AL, Li S, Srinivasan K, et al. (2004) Modeling Online Browsing and Path Analysis Using Clickstream Data. *Marketing Science* 23(4): 579–595.

Moro S, Cortez P and Rita P (2014) A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62: 22–31.

Muhammad I and Yan Z (2015) Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing* 05(03): 946–952.

Opitz D and Maclin R (1999) Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research* 11: 169–198.

Provost F and Fawcett T (1997) Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. *Proc of the 3rd International Conference on Knowledge Discovery and Data Mining*: 43–48.

Provost F and Fawcett T (2001) Robust Classification for Imprecise Environments. *Machine Learning* 42(3): 203–231.

Rajamma RK, Paswan AK and Hossain MM (2009) Why do shoppers abandon shopping cart? Perceived waiting time, risk, and transaction inconvenience. *Journal of Product & Brand Management* 18(3): 188–197.

Razi MA and Athappilly K (2005) A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications* 29(1): 65–74.

Rokach L (2010) Ensemble-based classifiers. *Artificial Intelligence Review* 33(1-2): 1–39.

Rumelhart DE, Hinton GE and Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088): 533–536.

Sakar CO, Polat SO, Katircioglu M, et al. (2019) Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications* 31(10): 6893–6908.

Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM Journal of research and development* 3(3): 210–229.

Schapire RE, Freund Y, Bartlett P, et al. (1998) Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics* 26(5): 1651–1686.

Schmidhuber J (2015) Deep Learning in Neural Networks: An Overview. *Neural Networks* 61: 85–117.

Sismeiro C and Bucklin RE (2004) Modeling Purchase Behavior at an E-Commerce Web Site: A Task-Completion Approach. *Journal of Marketing Research* 41(3): 306–323.

Srivastava N, Hinton G, Krizhevsky A, et al. (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15: 1929–1958.

Stone M (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* 36(2): 111–133.

Suykens JAK and Vandewalle J (1999) Least Squares Support Vector Machine Classifiers. *Neural Processing Letters* 9(3): 293–300.

Tambe P, Cappelli P and Yakubovich V (2019) Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. *California Management Review* 61(4): 15–42.

Tibshirani RJ and Tibshirani R (2009) A Bias Correction for the Minimum Error Rate in Cross-Validation. *The Annals of Applied Statistics* 3(2): 822–829.

Vafeiadis T, Diamantaras KI, Sarigiannidis G, et al. (2015) A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory* 55: 1–9.

van den Poel D and Buckinx W (2005) Predicting online-purchasing behaviour. *European Journal of Operational Research* 166(2): 557–575.

Vapnik VN (1982) *Estimation of Dependences Based on Empirical Data*. New York: Springer Publishing Company, Incorporated.

Williams N, Zander S and Armitage G (2006) A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *ACM SIGCOMM Computer Communication Review* 36(5): 5–16.

Zhang G, Patuwo BE and Hu MY (1998) Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting* 14(1): 35–62.

Zhang P (1993) Model Selection Via Multifold Cross Validation. *The Annals of Statistics* 21(1): 299–313.

Zhang Q, Yang LT, Chen Z, et al. (2018) A survey on deep learning for big data. *Information Fusion* 42: 146–157.

Zheng B and Liu B (2018) A scalable purchase intention prediction system using extreme gradient boosting machines with browsing content entropy. In: *2018 IEEE International Conference on Consumer Electronics (ICCE):* Piscataway, NJ: IEEE.

V Mitigating the Negative Consequences of ICT Use: The Moderating Effect of Active-Functional and Dysfunctional Coping

Authors: Julia Becker, Nicholas Daniel Derra, Christian Regal, and Torsten M. Kühlmann

Submitted to: Journal of Decision Systems (2020)

Abstract: Through the course of rapid digitalization, negative consequences and strain resulting from the use of information and communication technologies at work have become an important topic of debate. With this paper, we contribute to the current discourse by examining how employees mitigate technostress. We transfer theory from psychology to information systems literature by investigating a moderated mediation model where coping was conceptualized as a personal resource in line with the job demands-resources model. The moderating effects of two different reactive coping strategies—active-functional and dysfunctional—were investigated within a final sample of 3,362 German knowledge workers. By using covariance-based structural equation modelling, we found that technology-related demands are associated with higher level of both strain and productivity. We found a competitive mediation effect where the direct effect of demands on productivity is of opposite direction as the indirect mediated effect via strain. These effects are buffered by both active-functional and dysfunctional coping. They reduce the extent to which demands lead to strain. Further, active-functional coping is associated with lower strain whereas dysfunctional coping is associated with higher strain. The contribution of this paper for technostress research is discussed and implications for future research are given. The recommendations for employers and employees are highlighted.

Keywords: Negative Consequences of ICT Use; Technostress; Strain; Coping; Active-Functional coping; Dysfunctional coping

1 Introduction

Digitalization's rapid progression leads to comprehensive and ubiquitous change that affects individuals, economies, and society (Gimpel et al., 2018). Digital transformation is driven by a wide variety of digital technologies and their adoption (Hartl, 2019; Osmundsen, Iden, & Bygstad, 2018). Even though many opportunities and chances accompany this development (e.g., products and services can be offered in less time or with better quality), there are some downsides. In particular, the use of information and communication technologies (ICT) in occupational settings may cause stress. During the last years, research has noted this as a specific form of stress called technostress (Ayyagari, Grover, & Purvis, 2011; Tarafdar, Tu, & Ragu-Nathan, 2010; Tarafdar, Tu, Ragu-Nathan, & Ragu-Nathan, 2007). The term technostress itself was coined in the 80s by Brod (1984, p. 16), who designated it as a person´s “inability to cope with the new computer technologies in a healthy manner”. This is the case if individuals do not feel able to adapt to or keep pace with the increasing technological changes, for example, when e-mails are perceived as constant interruptions or the boundaries between the work life and private life become blurred due to the need for constant availability (Tarafdar et al., 2010). Hence, the impact of digitalization on an employee’s working environment must be regarded as ambivalently (Apt, Bovenschulte, Hartmann, & Wischmann, 2016).

It has been shown that technology-related factors that induce stress are associated with a reduction in productivity, job satisfaction, and loyalty to the employer as well as an increased risk of burnout and a poor work-life balance (Ayyagari et al., 2011; Califf, Sarker, & Sarker, 2020; Khaoula, Khalid, & Omar, 2020; Srivastava, Chandra, & Shirish, 2015; Tarafdar et al., 2010; Tarafdar, Tu, Ragu-Nathan, & Ragu-Nathan, 2011). Research has also identified several organizational and individual factors that positively moderate the relationship between techno stressors and health and organizational outcomes (Srivastava et al., 2015; Tarafdar, Pullins, & Ragu-Nathan, 2015).

All these beneficial factors have in common that they are outside the individual’s scope of influence. They are either organizational factors (Ragu-Nathan, Tarafdar, Ragu-Nathan, & Tu, 2008) or inherent stable personality traits (Sumiyana & Sriwidharmanely, 2020). But little is known about actual behaviours or thoughts that the individual deploys to mitigate the harmful effects of technostress. A few studies are concerned with coping, but these conceptualize coping as a mediator between technostress and strain in line with the transactional model of stress (Lazarus & Folkman, 1984). In contrast, research from industrial and organizational psychology emphasizes the role that coping plays as a personal resource (Searle & Lee, 2015), moderating the relationship between job demands and strain (Bakker & Demerouti, 2017). Accordingly, the neglect of coping as a moderator proposed by Frese (1986) is still present within the research field of technostress in information systems (IS) and was only recently addressed by few research articles (Nisafani, Kiely, & Mahony, 2020; Pirkkalainen, Salo, Tarafdar, & Makkonen, 2019) focusing on the role of proactive and reactive coping (Pirkkalainen et al., 2019). Hence, coping responses to technostress are under-studied and interdisciplinary enrichment between psychological literature and IS research is needed (Pirkkalainen et al., 2019; Tarafdar, Cooper, & Stich, 2019). The disciplines share a common and joint research interest but yet, most articles about technostress are published in IS journals and only few in psychological journals disrupting the flow of information, knowledge and exchange of theories from one discipline to the other (Bondanini, Giorgi, Ariza-Montes, Vega-Muñoz, & Andreucci-Annunziata, 2020).

In this paper, we aim to provide evidence that coping as a personal resource moderates the relationship between (techno)stress and strain as proposed by the psychological theory of job demands-resources (JD-R) model (Demerouti, Bakker, Nachreiner, & Schaufeli, 2001) complementing the perspective of coping as mediator of emotional responses that is grounded in the transactional theory of stress (Lazarus & Folkman, 1984). Thereby, we contribute to research by investigating the influence of technostress on organizational and individual-level outcomes while modelling coping as a moderator in line with the workplace-specific JD-R. This includes the conceptualization of strain mediating the influence of technology-related demands on work productivity. Furthermore, we emphasize the importance of distinguishing between functional and dysfunctional coping, two forms of reactive coping, to gather insights about the differentiation of effective and less effective ways to overcome strain related to digital technology use.

The present manuscript is structured as follows: first, we will address the theoretical background and give an overview of the current research streams in IS and psychology regarding the negative consequences of ICT use. Subsequently, based on the existing literature, we designed a conceptual model that integrates the relationships between techno stressors, their impact on strain, well-being, organizational outcomes, and the moderating effect of individual coping behaviours. This model guided our empirical study on the impacts of technostress. Lastly, we will summarize and carefully discuss the empirical findings and give an outlook for future research.

2 Theoretical Background

2.1 Technostress

The concept of technostress is anchored in the transactional theory of stress (Lazarus & Folkman, 1984). Stress is a process where individuals appraise the demands of a given situation as taxing or exceeding their resources while interacting with their environment. Consequently, technostress refers to stress that arises during ICT usage (Tarafdar et al., 2019). Tarafdar et al. (2007) emphasize that "in the organizational context, technostress is caused by individuals' attempts and struggles to deal with constantly evolving ICT and the changing physical, social, and cognitive requirements related to their use" (p. 304). Hence, employees might experience technostress due to an increased usage of ICT at the workplace (Ragu-Nathan et al., 2008).

Previous research has identified several factors that may induce technostress. The five most-cited techno stressors are those characterized by Tarafdar et al. (2007): *Complexity* refers to situations where employees do not feel able to handle job-related technologies due to a perceived lack of skills. *Insecurity* relates to employees´ fear of being replaced by new technologies or other employees, resulting in losing their job. *Invasion* is connected to blurred boundaries between work-related and private periods. Situations where employees have to work faster, longer, and even more due to ICT usage represent *overload*. At last, *uncertainty* describes employees´ confusion in ICT use caused by new developments regarding the organization's technologies. Besides these well-established techno stressors there are other aspects which are discussed as demanding: Riedl, Kindermann, Auinger, and Javor (2012) investigated *unreliability*, which refers to ICT troubles like system breakdowns. Furthermore, a disturbed workflow through *interruptions* has been considered as another technology-related stressor (Galluch, Grover, & Thatcher, 2015). Too many interruptions in the leisure time via mobile technologies are a source of stress leading to work-family conflict and lower adoption of IS in the workplace (Tams, Ahuja, Thatcher, & Grover, 2020). This is in line with current findings where technical problems, disruptions (in the workflow or meetings), communication overload, and continuing work tasks at home were identified as stressful events related to ICT use (Braukmann, Schmitt, Ďuranová, & Ohly, 2018).

The described factors may lead to strain, which is defined as an employee´s psychological, physical, or behavioural response to techno stressors (Atanasoff & Venable, 2017). In this context, several studies have already dealt with different facets of strain like mental exhaustion (i.e., feeling burned out and drained (Ayyagari et al., 2011; Srivastava et al., 2015)), or problems of psychological detachment (Barber, Conlin, & Santuzzi, 2019; Santuzzi & Barber, 2018). Furthermore, technostress is also associated with adverse organizational outcomes (i.e., lower productivity (Tarafdar et al., 2007; Tarafdar et al., 2015), lower user satisfaction (Fischer & Riedl, 2020), and lower employee's loyalty to the employer (Tarafdar et al., 2011).

To reduce technostress, Ragu-Nathan et al. (2008) investigated three situational factors and organizational mechanisms: technical support, literacy facilitation (users are encouraged to share their experiences with and knowledge about new technologies), and involvement facilitation (users are consulted in the implementation of new technologies and are actively encouraged to try them out). These technostress-inhibitors operated as moderators of the relationship between technostress and job-satisfaction, organizational commitment, and continuance commitment. Other factors that influence the relationship between techno stressors and outcomes are timing control and method control (Galluch et al., 2015). Furthermore, individual moderating variables like technology self-efficacy (Tarafdar et al., 2015) and personality traits like openness, agreeableness, neuroticism, and extraversion (Srivastava et al., 2015) have been identified.

2.2 Coping with Technostress

According to the transactional theory of stress (Lazarus & Folkman, 1984, p. 141), coping is defined "as constantly changing cognitive and behavioural efforts to manage specific external and/or internal demands that are appraised as taxing or exceeding the resources of the person". These efforts are commonly classified into different styles of coping. Besides the broadly acknowledged distinction between problem-focused coping (directed at the problem itself in terms of modifying or improving the person-environment relation) and emotion-focused coping (comprising strategies which aim at regulating stressful emotions) proposed by Folkman, Lazarus, Dunkel-Schetter, DeLongis, and Gruen (1986), more fine-grained taxonomies include active coping, seeking instrumental social support, religion, positive reinterpretation, mental disengagement or behavioural disengagement—only to name a few (Carver, Scheier, & Weintraub, 1989). In a more detailed approach, 14 different coping styles have been differentiated (Carver, 1997). Thereby, active coping and seeking instrumental social support can be subsumed under problem-focused coping, whereas positive reinterpretation and turning to religion are examples of positively related emotion-focused coping. Hence, these two higher-level categories reflect active-functional strategies (Prinz, Hertrich, Hirschfelder, & Zwaan, 2012). In contrast, coping strategies where individuals try to avoid the overall issue and escape from the problem instead of tackling it at source are considered dysfunctional. Examples are mental and behavioural disengagement as well as alcohol and drug consumption (Carver et al., 1989).

Research using this more fine-grained taxonomy found that active coping is associated with lower exhaustion (Gaudioso, Turel, & Galimberti, 2017). The use of active-functional strategies, such as seeking social support, is negatively associated with burnout (Erschens et al., 2018). It has also been observed that maladaptive, dysfunctional coping like behavioural disengagement is associated with increased work exhaustion (Gaudioso et al., 2017; Prinz et al., 2012) and strain (Hauk, Göritz, & Krumm, 2019). In total, there is some evidence that active-functional coping strategies positively influence employees' well-being and organizational outcomes, whereas dysfunctional coping negatively impacts those outcomes. However, it is not clear how coping moderates the relationship between techno stressors and organizational as well as health outcomes. Active-functional coping should be beneficial, whereas dysfunctional coping may be seen as a malfunctioning strategy to overcome the long-term consequences of stress.

There is no consensus in research whether coping strategies should be considered a moderator or mediator. Frese (1986) mentioned this issue in his study and highlights that this specific distinction is often neglected. As emphasized above, the technostress framework from IS literature is based on Lazarus and colleagues (Folkman et al., 1986; Lazarus & Folkman, 1984), where coping is modelled as a mediator. Several studies have already addressed this in the context of technostress research (Gaudioso, Turel, & Galimberti, 2016; Hauk et al., 2019; Zhao, Xia, & Huang, 2020). Maladaptive coping, for example, translates invasion and overload through the strain facets of work-family conflict and distress into higher exhaustion. In contrast, adaptive coping strategies mediate the same relationship resulting in lower work exhaustion (Gaudioso et al., 2017). Behavioural disengagement mediates the relationship between age and technology-induced strain operationalized as emotional and physical exhaustion (Hauk et al., 2019).

At the same time, stressors and work demands, which also include stress resulting from the use of ICT, constitute a typical subject of matter in psychological investigations (Barber et al., 2019; Braukmann et al., 2018; Day, Paquet, Scott, & Hambley, 2012; Day, Scott, & Kelloway, 2010; Golden, 2012; Sonnentag, Kuttler, & Fritz, 2010). In this context, coping strategies have been discovered numerous times as a moderating variable: Lewin and Sager (2009) found that problem-focused coping strategies moderate the impact of stressors on emotional exhaustion. Yip, Rowlinson, and Siu (2008) provide evidence that coping buffers the negative effects of job stressors on burnout. Similarly, Searle and Lee (2015) found that pro-active coping moderates the relationship between demands and burnout. Ashill, Rod, and Gibbs (2015) show in their study that self-directed coping mitigates dysfunctional effects of job demand stressors on emotional exhaustion while other-directed coping buffers the relationship between job demands and job performance. Recently published articles in IS also started to model coping as a moderator (Nisafani et al., 2020; Pirkkalainen et al., 2019).

Investigating coping as a moderator, psychological research widely uses the JD-R model (Demerouti et al., 2001) as the theoretical foundation which has been applied and expanded to explain the relationship between job demands, personal resources, and strain (e.g., exhaustion as one facet of burnout (Demerouti, Mostert, & Bakker, 2010)). In keeping with the JD-R model, “job resources refer to those physical, psychological, social, or organizational aspects of the job that may do any of the following: be functional in achieving work goals, reduce job demands and the associated physiological and psychological costs, stimulate personal growth and development” (Demerouti et al., 2001, p. 501). ”Personal resources can be seen as the beliefs individuals have in their ability to act on the environment” (Bakker & Demerouti, 2017, p. 275). How people cope with stress can be treated as a personal resource as well (Searle & Lee, 2015). Personal resources can buffer the impact of job demands on strain, while strain variables like exhaustion negatively affect employees’ job performance (Bakker & Demerouti, 2017). According to Ninaus, Diehl, Terlutter, Chan, and Huang (2015) and Patel, Ryoo, and Kettinger (2012), it can also be differentiated between demands and resources within ICT use. Employees may benefit from ICT use, but it also increases demands and causes strain (Bakker & Demerouti, 2017). These resources also include coping strategies to mitigate strain directly (Ângelo & Chambel, 2014). The JD-R model has also been used as a theoretical foundation for conceptualizing technostress (Christ-Brendemühl & Schaarschmidt, 2020; Florkowski, 2019; Mahapatra & Pati, 2018; Ninaus et al., 2015; Wang, Kakhki, & Uppala, 2017) but it has not been applied in investigating coping strategies as a moderator in the technostress context yet. We aim to close this theoretical gap.

3 Research Model and Hypotheses Development

We are referring to the agenda postulated by Tarafdar et al. (2019) who claim a lack of research on coping strategies and its effects on the relationships between techno stressors and outcomes. Simultaneously, other researchers (Nisafani et al., 2020; Pirkkalainen et al., 2019) call for further investigations of coping strategies and how they might lead to different coping outcomes. To fill this gap, the respective moderating effects of active-functional and dysfunctional coping behaviour are the focus of our examination. Another reason for this is that Pirkkalainen et al. (2019) focus on the effects of proactive (i.e., strengthening one’s ability to cope) and reactive coping, neglecting the different types of reactive coping. Based on the findings above, we developed a research model (the simplified moderated mediation model is displayed in Figure 1) building on both psychological literatures regarding job demands as well as negative consequences of ICT use and technostress literature from the IS domain.

The model establishes a relation between job demands, strain (represented through exhaustion), and job performance (represented through productivity) - with strain mediating the impact of job demands on job performance - as well as the moderating effect of coping as a resource which is in line with the JD-R model (Bakker & Demerouti, 2017). Furthermore, the direct effect of coping on strain, as proposed by Ângelo and Chambel (2014), is included. To our understanding, the techno stressors described above represent technology-related job demands resulting from the use of ICT for work purposes. The wording 'demands' will be subsequently used. Therefore, in the model, the second-order construct job demands comprises the five techno stressors (Tarafdar et al., 2007) mentioned and explained above: complexity, insecurity, invasion, overload, and uncertainty. Also, interruptions and unreliability (ICT hassles) were identified as affective events related to ICT use that may have negative consequences for well-being (Braukmann et al., 2018).

In line with the proposed model, we deduct hypotheses for the relationships between job demands, exhaustion, productivity, and coping. It has been shown that technostress is associated with lower productivity and simultaneously, techno stressors can induce strain. Further, the JD-R model proposes that strain translates into lower job performance, so we assume:

Hypothesis 1a: Job demands are negatively associated with the productivity of employees.

Hypothesis 1b: The relationship between job demands and productivity is mediated by exhaustion.

Even though the psychological framework of the JD-R model has already been applied in the technostress context (Day et al., 2010; Florkowski, 2019; Mahapatra & Pati, 2018; Ninaus et al., 2015; Patel et al., 2012; Wang et al., 2017), there is no research concerning coping strategies moderating the relationship between techno stressors and outcomes yet. For investigating these effects in our model, we differentiate between active-functional and dysfunctional coping. First, active-functional coping (like support-seeking behaviour and searching for solutions or improvements in a stressful situation) is associated with a lower level of exhaustion. In contrast, dysfunctional coping (like displacing reality, escaping behaviour, and the consumption of alcohol or drugs) is related to an increased level of exhaustion; we propose accordingly:

Hypothesis 2a: Active-functional coping is negatively related to employees' level of exhaustion.

Hypothesis 2b: Active-functional coping acts as a moderator, mitigating the negative impact of techno stressors on exhaustion.

Hypothesis 3a: Dysfunctional coping is positively related to employees' level of exhaustion.

Hypothesis 3b: Dysfunctional coping acts as a moderator reinforcing the negative consequences of techno-stressors on exhaustion.

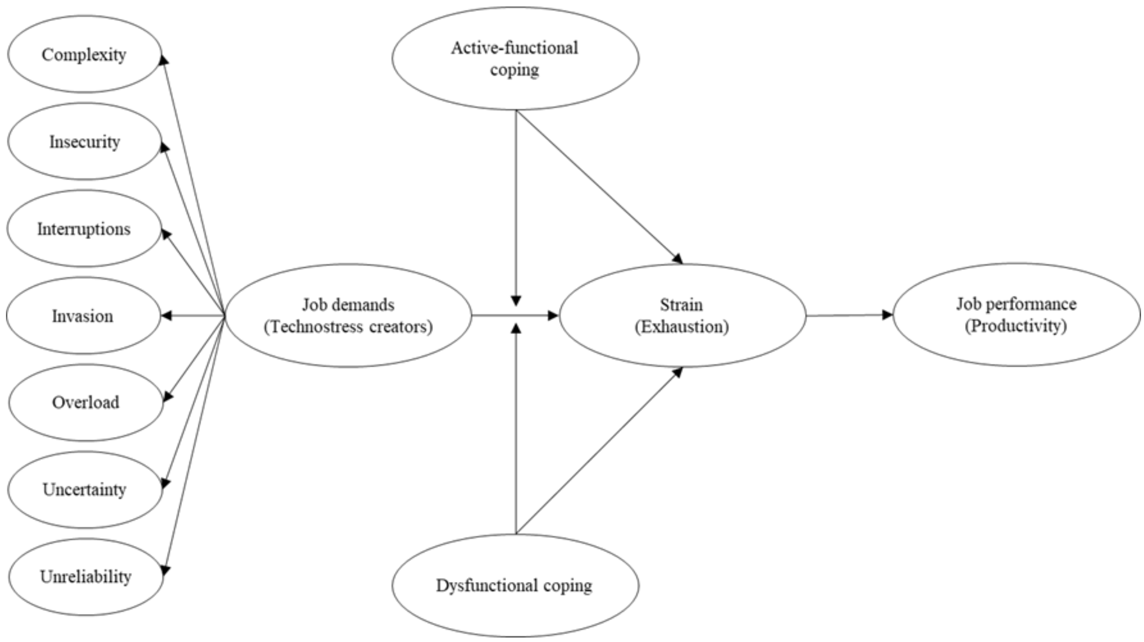

Figure 1. The proposed research model of the assumed relationships in accordance with Nisafani et al. (2020).

4 Method

4.1 Sample

Data for this study was collected within the setting of a larger research project supervised by an interdisciplinary committee from which ethical approval for the survey was obtained. For more information concerning ethics, please see the declaration at the end of this manuscript. Except for one published paper and one paper in preparation where the five techno stressors proposed by Tarafdar et al. (2007), which were assessed with the scales developed by Ragu-Nathan et al. (2008), were used, none of the variables (and scales, respectively) utilized in the present study have already been used in previous publications based on data from the same project. Respondents were acquired via an external research panel and paid a small incentive for participation in the study. Participants gave informed consent, which means they actively agreed that they are over 18 years of age, have read the information on intentions of the research project, ethics and processing of data and data protection by ticking a box. A contact person was listed, and they were informed that they had the possibility to withdraw their consent to participate without giving reasons or incurring disadvantages at any time. Subjects were guaranteed that their answers were collected anonymously as far as possible. “Protecting respondent anonymity and reducing evaluation apprehension” helps to reduce possible common method bias (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003, p. 888). To do so, we reminded participants that there are no right or wrong answers and that we are interested in their honest opinion at the introduction of each subsection, trying to minimize method bias. The panellist company was instructed to collect answers from German knowledge workers. Knowledge workers are defined as employees working in an occupation where information is a resource, tool and result of work (Klotz, 1997). Examples for relevant professions are technicians, engineers, scientists, finance, controlling, managers, journalists, consultants, and lawyers. The questionnaire included control variables to test our sample's representativeness, namely age, sex, employment status, occupational title and sector, number of hours worked per week, and education. Further, intensity of technology use for work purposes was assessed. In the first step, the answers of n = 445 participants were collected for a quantitative pre-test of the scales. In a second step, answers for the main study were collected. This final sample consisted of n = 3,362 respondents. Preliminary analysis showed that the distribution of participants according to the control variables age, sex, and sectors (Federal Statistical Office of Germany, 2018a, 2018b) is representative of the German working population. About 46% percent of participants were female and 54% male. The mean age was 42.44 years (SD = 11.39). 23% of the participants have a secondary school education, 27% finished a vocational apprenticeship, 19% had a bachelor’s degree, 27% finished with a master’s degree, and 4% percent completed a Ph.D. Most participants (30%) worked in the public or private service sector, followed by 15% who worked in the trade, transport or hotel sector, followed by the producing sector without construction industry (15%), business services industry (14%), information and communication (11%), finance- and insurance services (7%), construction sector (4%), land- and housing sector (2%), and agriculture, forestry, and fishing (< 1%).

4.2 Measures

We relied on established, validated scales in the survey. All questions were administered in German. If necessary, the items were translated from the original language. Therefore, three German native speakers translated the questions in parallel. They met afterwards to resolve discrepancies and agreed on the best translation. In this step, we tried to avoid common method bias. The following rules were applied to all items in the translation procedure: “keep questions simple, specific, and concise; avoid double-barrelled questions; decompose questions relating to more than one possibility into simpler, more focused questions; and avoid complicated syntax.” (Podsakoff et al., 2003, p. 888). The measures were subjected to extensive testing with participants who had not been involved in the research process previously to identify ambiguous terms and to ensure understanding of the translated items. In this quantitative pre-test, the scales' quality and psychometric properties were evaluated based on the answers of n = 445 participants.

Complexity, insecurity, invasion, overload, and uncertainty were assessed with the scales developed by Ragu-Nathan et al. (2008). Complexity was measured using five items, for example: *“I need a long time to understand and use new technologies”*. The scale for insecurity encompasses five items, including *“I have to constantly update my skills to avoid being replaced.”* Invasion comprises three items (e.g., *“I have to be in touch with my work even during my vacation due to this technology”*). Overload was measured with four items. An example is *“I am forced by this technology to work with very tight time schedules”*. Lastly, uncertainty was measured with four items (e.g., *“There are constant changes in computer software in our organization”*). Additionally, interruptions were assessed with three items published by Galluch et al. (2015), for example, *“I experienced many distractions during the task”* and finally, unreliability (Ayyagari et al., 2011) was also measured with three items (e.g., *“The features provided by digital technologies are dependable”*). We used a five-point Likert-type rating scale from 0 = *I do not agree at all* to 4 = *I totally agree* for all items.

Exhaustion was measured with a subscale of the Maslach Burnout Inventory (Maslach & Jackson, 1986). It contains nine items, for example, *“I feel emotionally drained by my work”*. A five-point Likert-type rating scale ranging from 0 = *I do not agree at all* to 4 = *I totally agree* was used.

Productivity was measured with four items (Chen & Karahanna, 2014). It describes self-evaluated work performance (fulfilment of work tasks and general demands). An example item is *“I have a reputation in this organization for doing my work very well”*. Ratings were made on a five-point Likert-type rating scale ranging from 0 = *I do not agree at all* to 4 = *I totally agree*.

Coping was assessed with a selection of 15 items from the Brief COPE (Carver, 1997). We used the existing German translation of the inventory (Knoll, Rieckmann, & Schwarzer, 2005). While the original scale contains 28 items paired up in 14 subscales with two items each, the subscales from Prinz et al. (2012) that build on the Brief COPE consist of nine items for active-functional coping and six items for dysfunctional coping. Active-functional coping comprises for example, *"I've been taking action to try to make the situation better"*. An example for dysfunctional coping is *"I've been using alcohol or other drugs to make myself feel better"*. Answers were assessed on a three-point frequency scale ranging from 0 = *never* to 2 = *often*. The items are displayed in Table 6 in the Appendix.

The covariate technology use was assessed with one self-developed item: "*How often do you use digital technologies for your work?"*. Frequency answers were given from 0 = *never* to 4 = *several times a day*.

4.3 Means of Analysis

After running descriptive analyses, we subjected the items for the two coping subscales identified by Prinz et al. (2012) to an exploratory factor analysis (EFA) with varimax rotation (see Appendix) to see whether the expected two factors are extracted because the authors of the original scale did not provide this clustering (see Table 5 in the Appendix). The relationships of the variables we propose in our research model were analysed using covariance-based structural equation modelling (Jöreskog, 1970). We utilized the widely used open-source software R and the integrated development environment R-Studio (R Development Core Team, 2019; RStudio Team, 2019). For specific analyses, we used complementary packages in addition to the R base program (i. e., lavaan (Rossel, 2012), psych (Revelle, 2019), GPARotation (Bernaards & Jennrich, 2005), and semTools (Jorgensen, Pornprasertmanit, Schoemann, & Rossel, 2019)).

To test nonlinear and interactive effects in structural equation models, Kenny and Judd (1984) proposed the product indicator (PI) approach. The products of the observed variables are used as indicators for the latent interaction term in the measurement model. To create the product term, the indicator with the highest reliability should be chosen (Saris, Batista-Foguet, & Coenders, 2007), while the product shows optimal reliability as an indicator of the latent interaction variable, whereby the power of the test of the latent moderator increases by an increase in the reliability of the indicator (Saris et al., 2007). When using product indicators, missing independency of higher-order indicators from the lower-level indicators due to the multiplication of the two variables is a problem. Statistical procedures have been introduced to deal with this dependency of higher-order indicators to lower-order indicators. Lin, Wen, Marsh, and Lin (2010) propose a double mean centring strategy. This approach performs well and eliminates the need for the constraint of the inclusion of a mean structure, as introduced by Jöreskog and Yang (1996). Double mean centring also performs better with non-normal data than (single) mean centring and orthogonalization. It can be combined with different matching strategies of indicators and is available with most commercial SEM software. Hence, to create the indicators for the latent interaction term between techno stressors and coping, we used the PI approach in which indicators were chosen and matched according to reliability. The product terms were double mean centred (Lin et al., 2010).

5 Results

5.1 Measurement Models

Preceding the analysis of the proposed relationships in our hypothesis, we tested the measurement models of the endogenous (strain and productivity) and exogenous (job demands and coping) latent variables. Job demands were modelled as a second-order construct (reflected in the seven technology-related stressors) with both first-order and second-order indicators being reflective. For more information about the choice of measurement model please compare Ragu-Nathan et al. (2008, p. 428). The moderated mediation was set up as Hayes (2013) described and based on the in-depth explanations (Stride, Gardner, Catley, & Thomas, 2019). Coping moderates the relationship between the independent variable (IV) job demands and the mediator exhaustion (IV–Mediator path) and, further, has a direct effect on exhaustion.

We first assessed means and standard deviations, item reliabilities (loadings), and internal consistency (Cronbach's alpha). Table 1 shows an overview of the scales' properties. For brevity of presentation, the values in the table reflect the final measurement model after deletion of single indicators.

Table 1. Descriptive statistics, factor loadings, and reliability of the scales in the study.

Scale	Items	*M*	*SD*	Loadings	α
Complexity	5	1.22	1.04	0.77–0.87	0.91
Insecurity	5	1.23	1.03	0.72–0.82	0.87
Interruptions	3	1.59	1.16	0.85–0.90	0.90
Invasion	3	1.28	1.12	0.64–0.88	0.82
Overload	4	1.62	1.10	0.70–0.85	0.88
Uncertainty	4	1.80	1.04	0.74–0.85	0.87
Unreliability	3	1.82	1.10	0.85–0.92	0.91
Exhaustion	9	1.50	1.09	0.76–0.91	0.96
Productivity	4	2.62	0.85	0.81–0.83	0.89
Active-functional coping (A)	6	0.73	0.60	0.68–0.76	0.86
Dysfunctional coping (D)	4	0.28	0.45	0.62–0.79	0.80

Cronbach's alpha was above 0.70 for all constructs, as recommended (Nunnally & Bernstein, 1994). The test of item reliability showed good results. The factor loadings for each indicator should be above the value of 0.70, indicating that the underlying latent factor accounts for more than 50% of the variance in the respective indicator (Fornell & Larcker, 1981). Most loadings met this threshold. For the items of the two coping constructs and one item of invasion, values below the threshold of 0.70 were observed. The reliability of constructs is evaluated by the average variance extracted (AVE). It determines whether the latent construct accounts for more than 50% of its indicator's variance on average. This threshold was met by invasion and dysfunctional coping, whereas it was below 0.50 for active-functional coping due to very low loadings, even below 0.60. The two items with the lowest loading were removed, which improved the AVE of active-functional coping to 0.51. Further, two items of the latent interaction term between active functional coping and technostress displayed loadings below 0.60. Hence, they were taken out of the model as well.

Internal consistency measures like Cronbach's alpha are not sufficient to imply homogeneity and unidimensionality of constructs (Tavakol & Dennick, 2011). Hence, in addition, we analysed the discriminant validity of the latent endogenous constructs with the Fornell-Larcker criterion (Fornell

& Larcker, 1981) based on AVE and the correlations among the latent constructs. It is considered as given if the square root of the AVE (printed along the diagonal of the correlation matrix) is higher than the correlations with the other latent variables (off-diagonal elements) (Fornell & Larcker, 1981). The results are displayed in Table 2. All correlations between the latent variables were significant at the level $p < 0.001$. The square root of the AVE printed along the diagonal is higher than the correlations with respective other components for each of the latent factors. This suggests that the discriminant validity of the endogenous constructs in our model is given.

In addition to the procedural remedies which we have taken to avoid common method bias, which is described in the method section, we conducted Harman‘s single factor test (Harman, 1967) to infer whether common method variance that potentially results in common method bias seems a problem in our data set. Results of an unrotated principal component analysis to which we subjected all study items show that about 14% is the highest proportion of variance attributed to the first factor. Accordingly, common method variance and, hence, common method bias is not considered a problem.

Table 2. Discriminant validity according to the Fornell-Larcker criterion.

Scale	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Complexity (1)	**0.82**										
Insecurity (2)	0.68	**0.76**									
Interruptions (3)	0.60	0.54	**0.87**								
Invasion (4)	0.62	0.72	0.57	**0.78**							
Overload (5)	0.67	0.71	0.71	0.66	**0.81**						
Uncertainty (6)	0.43	0.62	0.41	0.50	0.55	**0.80**					
Unreliability (7)	0.54	0.52	0.63	0.48	0.64	0.44	**0.88**				
Exhaustion (8)	0.49	0.41	0.50	0.42	0.53	0.21	0.42	**0.85**			
Productivity (9)	-0.12	-0.02	-0.04	0.02	-0.01	0.11	-0.04	-0.18	**0.82**		
Active-functional coping (10)	0.19	0.15	0.27	0.14	0.27	0.13	0.26	0.20	0.12	**0.71**	
Dysfunctional coping (11)	0.49	0.49	0.38	0.50	0.42	0.31	0.34	0.43	-0.02	0.45	**0.71**

5.2 Structural Model

After validating the measurement model, we analysed the structural model to test our hypotheses. Unweighted least squares (ULS) were used as an estimator for the evaluation of the model because ULS perform better with non-normal and ordinal data as they do not make assumptions about the distribution (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009). Standard errors were obtained through bootstrapping with 1,000 iterations. We tested the models stepwise: First, only the covariate was included, then the IV was added. Next, the mediator variable strain was included and in the last step, we set up the full moderated mediation model. The results are displayed in Table 3.

We assessed the root mean square error of approximation (RMSEA), the square root mean residual (SRMR), the Tucker-Lewis index (TLI), and the comparative fit index (CFI) as indicators of model fit. The χ^2 test statistic is not available with ULS estimation. The absolute fit index RMSEA indicates a good model fit at values smaller than 0.05, just like the SRMR. CFI and TLI indicate satisfactory model fit greater than 0.95 and a good fit at values above 0.97 (Geiser, 2011). Strict cut-off values were applied to check the model's suitability since it has been shown that in ULS estimations, the indices tend not to detect model–data misfit or misspecifications as efficiently as in maximum likelihood (ML) estimations (Xia & Yang, 2019). Overall, the moderated mediation model showed a good fit. SRMR was 0.05, indicating only a small divergence between the empirically observed and model-implied covariance matrix. RMSEA was 0.05 slightly above the strict threshold of 0.05. CFI and TLI indicate a good fit of the model (both, CFI = 0.98, TLI = 0.98) with values higher than 0.97. Even with the strict cut-off criteria, the model seems to fit the data well. Next, we regarded the regression paths of model 4 to evaluate our hypotheses (cf. Table 7 in the Appendix for standard errors and *z* values of the moderated mediation model).

Table 3. Results of the model estimation: direct and moderation effects.

	Model 1		Model 2		Model 3		Model 4	
	Exhaustion	Productivity	Exhaustion	Productivity	Exhaustion	Productivity	Exhaustion	Productivity
Intensity of technology use	-0.03*	0.08***	-0.03*	0.07***	-0.03*	0.07***	-0.03*	0.07***
Job demands (TS creators)	-	-	-	-0.02	0.57***	0.11***	0.44***	0.12***
Strain (exhaustion)	-	-	-	-0.18***		-0.25***		-0.25***
Active-functional coping (A)	-	-	-	-			-0.05*	
Dysfunctional coping (D)	-	-	-	-			0.31***	
Coping (A) × job demands	-	-	-	-	-		-0.05**	
Coping (D) × job demands	-	-	-	-	-		-0.12***	
R^2	<0.00	0.01	<0.00	0.03	0.32	0.05	0.36	0.05
ΔR^2	0.00	0.01	0.00	0.02	0.32	0.02	0.04	0.00

Note. Standardized path coefficients are displayed. Bootstrapped standard errors were used for the interpretation of the results. * $p < .05$, ** $p < .01$, *** $p < .001$.

Results of the mediation analysis show that job demands are significantly related to productivity as well as exhaustion. Further, exhaustion is significantly related to productivity. At the same time, the calculated total effect of job demands on productivity ($c = c' + (a \times b)$) was not significant (c = 0.01 (0.03), z = 0.57, p = .568) while the total indirect effect ($ab = a \times b$) of job demands on productivity via exhaustion was significant (ab = -0.11 (0.02), z = -7.61, p < .001). Thus, Hypothesis 1a must be rejected, whereas the results support Hypothesis 1b. Contrary to our expectations, job demands are positively related to job performance and higher productivity. Furthermore, job demands are positively associated with exhaustion as expected and higher levels of exhaustion go along with lower productivity. When both effects are significant but the indirect effect (*ab*) and the direct effect *c'* point to different directions, we speak of competitive mediation (Zhao, Lynch, & Chen, 2010).

The direct effect of active-functional coping on exhaustion was significant, as well as the direct effect of dysfunctional coping on exhaustion (see Table 3). The results support the assumptions in Hypotheses 2a and 3a. The use of active-functional coping strategies like support-seeking or actively trying to change the stressful situation is associated with lower levels of exhaustion. In contrast, trying to deal with a threatening situation through denial or consumption of alcohol or drugs to overcome negative feelings is associated with higher levels of exhaustion.

Active-functional coping significantly moderates the relationship between job demands and exhaustion. The negative sign of the path coefficient of the latent interaction term indicates that the negative consequences of ICT use are mitigated. The same applies to dysfunctional coping. The sign of the path estimate for the latent interaction term is also negative. Contrarily to our expectations, the use of dysfunctional coping strategies does not reinforce the effect of job demands on exhaustion but buffers it instead (see Table 3). Hence, Hypothesis 2b is supported by the data, whereas Hypothesis 3b must be rejected.

Additionally, indirect effects were calculated based on the path coefficients and low, medium, and high levels of the two moderator variables ($M \pm 1$ SD). This analysis differentiates between the total indirect and conditional indirect effects (simple slopes for each combination of conditions). The results are displayed in Table 8 in the Appendix. All combinations of low, medium, and high values for each moderator variable point to the same direction. Coping may reduce the detrimental effect of job demands on exhaustion as well as mitigate the negative impact of ICT use on strain. The analyses also show that the effect of dysfunctional coping is larger than the effect of active-functional coping (compare Table 4).

Table 4. Conditional indirect effects from the moderated mediation model.

	Low A (–1 *SD*)	Medium A (M)	High A (+1 *SD*)
High D (+1 *SD*)	-0.07***	-0.06***	-0.05**
Medium D (M)	-0.09***	-0.09***	-0.08***
Low D (–1 *SD*)	-0.12***	-0.11***	-0.10***

Note. Standardized path coefficients are displayed. Bootstrapped standard errors were used for the interpretation of the results of the conditional indirect effects. * $p < .05$, ** $p < .01$, ***$p < .001$.

6 Discussion

Our results from the covariance-based structural equation model revealed several unexpected insights. First, besides the negative indirect effect between job demands and productivity (through mediation via exhaustion), there is a positive direct effect. This positive effect means that, with increasing job demands, productivity rises, which intuitively seems contradictory. This kind of relationship is described in the goal setting theory (Locke & Latham, 2002). Difficulties and hard to achieve goals motivate people to do their best for goal achievement until their capability or commitment reaches a limit. Accordingly, a curvilinear relationship between general stress or work pressure and performance is observed (Hofmans, Debusscher, Dóci, Spanouli, & Fruyt, 2015; Leung, Huang, Su, & Lu, 2011): people who feel fewer demands are not able to utilize their full potential, and productivity is low. With increasing demands, productivity raises until a specific turning point is reached: if this level is exceeded and the perceived demands are demanded too much, productivity drops. The curvilinear relationship between demands and productivity refers to a rather short period, so there is a temporal aspect. The temporal consideration is reinforced by the fact that long-term increased strain, (i. e., chronic strain), can ultimately lead to burnout (Janssen, 2001).

Another reason for the positive effect of job demands on productivity is a potential suppressor effect, which occurs when the direct and indirect effects on a dependent variable have opposite signs and, therefore, an inconsistent mediation is present (Tzelgov & Henik, 1991). In the literature, it is considered to be realistic that two opposing direct and indirect effects with similar magnitude almost neutralize each other so the total effect is not significant (MacKinnon, Krull, & Lockwood, 2000).

For example, let's take the hypothetical example of workers making widgets, where X is intelligence, M is boredom, and Y is widget production (McFatter, 1979). Intelligent workers tend to get bored and produce less, but smarter workers also tend to make more widgets. Therefore, the overall relation between intelligence and widgets produced may actually be zero, yet there are two opposing mediational processes. Therefore, besides the observed positive relationship between technostress and productivity, an increase in demands may simultaneously lead to a higher level of exhaustion, resulting in lower productivity. Hence, we argue that, despite the positive relationship between job demands and productivity, technostress may lower productivity in a long-term view or have no positive impact on productivity. On the other hand, however, technological job demands increase the strain, leading to long-term health effects and negatively impact organizational objectives from a long-term perspective. Therefore, technostress should be reduced for organizational and human reasons.

Considering the role of coping for overcoming technostress, our results initially confirm prior research regarding the direct effects: a broad application of active-functional strategies is negatively related to exhaustion. In contrast, a broad application of dysfunctional coping may increase it. In doing so, dysfunctional coping exhibits a stronger direct impact on exhaustion. A possible explanation for this could be the nature of active-functional coping: strategies from the active-functional category (such as actively seeking to change the stressful situation) require individuals' energy and cause cognitive effort in implementation, which, in turn, may reduce the buffering effect on exhaustion.

In contrast, surprisingly, both active-functional and dysfunctional coping reduces the relationship between job demands and exhaustion. Furthermore, we even observed considerably higher values for dysfunctional coping regarding the buffering effect on the relationship between job demands and exhaustion. This implies that even though dysfunctional strategies go along with higher exhaustion, their moderating effect on the relationship between job demands and strain is stronger compared to active-functional strategies. This is particularly interesting because dysfunctional coping is said to be detrimental. The consumption of alcohol or drugs, for example, may lead to long-term adverse effects on physical and mental health (Kahler, Ramsey, Read, & Brown, 2002). Moreover, passive denial of a given situation has been proven to be a concept that is related to the development of depression (Kortte, Wegener, & Chwalisz, 2003; Naditch, Gargan, & Michael, 1975) - another reason why dysfunctional coping seems to be a bad strategy to tackle strain.

Nevertheless, these dysfunctional coping strategies seem to help reduce the harmful effects of strain resulting from modern technologies in our sample. The reasons for this relationship emerge when the time perspective is taken into account: coping strategies from the dysfunctional category, such as alcohol or denial of the problem, may result in short-term cognitive and emotional relief. From a long-term perspective, however, alcohol consumption naturally leads to other serious health consequences. The low level of content-related involvement with job demands leads to a reduced competence build-up, which ultimately means that resources are not strengthened. Therefore, we argue that dysfunctional coping, despite its short-term positive effects, would reinforce the consequences of demands in the long-term and, thus, should be avoided for efficiently overcoming technostress.

In conclusion, we see in Table 4 that a broad portfolio of coping strategies consisting of both active-functional and dysfunctional coping reduces the indirect negative effect of technostress via strain on productivity and, thus, also the suppressor effect. This implies that employees who use many different coping strategies from both categories would experience less exhaustion, ultimately leading to more productivity due to the additional direct effect of demands. On the other hand, the data show that employees with generally few different coping strategies can benefit from the suppressor effect as the total effect of the demands on productivity diminishes. However, they are still exposed to the negative consequences in terms of exhaustion. Employees who focus on a broad portfolio in one of the two categories reduce the negative indirect effects of demands on productivity via strain to such an extent that the positive direct effect of demand on productivity potentially remains significant, although the negative health effects - even if in reduced form - should not be neglected. In this context, it is shown that employees who utilize dysfunctional coping strategies can reduce the indirect effect more strongly, resulting in overall higher productivity, while, at the same time, causing more exhaustion than with active-functional coping, which in turn leads to less increase in productivity. The long-term consequences of dysfunctional copying have already been discussed in the previous paragraph.

6.1 Theoretical Contribution

Our research provides three important contributions to research on technostress and coping, namely: (1) investigating the influence of technostress and coping on organizational and individual-level outcomes; (2) modeling coping as a moderator applying the workplace-specific JD-R model as a meta-lens; and (3) emphasize the importance of the distinction between functional and dysfunctional coping of technostress concerning organizational and individual-level outcomes. We will discuss each contribution in detail in die following paragraphs.

In addition to the aspects discussed previously, our research addresses the call by Sarker, Chatterjee, Xiao, and Elbanna (2019) that most manuscripts in high-quality journals are concerned merely with the organizational outcomes. In a socio-technical system – i.e., a system focusing on the reciprocal interaction between technology as the technical component and employee as the social component (Lee, Thomas, & Baskerville, 2015; Ryan, Harrison, & Schkade, 2002) - it is important to consider both organizational and individual-level outcomes to create synergies (Griffith, Fuller, & Northcraft, 1998; Pava, 1983; Wallace, Keil, & Rai, 2004). Therefore, our research addresses the influence of functional and dysfunctional coping on both organizational (productivity) and individual-level outcomes (exhaustion).

Furthermore, in the context of technostress, we have applied the JD-R model as a theoretical meta-lens, in which both organizational and individual-level outcomes play a key role and which has not been applied in this context before (Bondanini et al., 2020). Thus, in comparison to the transactional model of Lazarus and Folkman (1984), which is usually used in the technostress literature, we applied a model that is explicitly focused on the working context. In this, we have also decided to model coping as a moderator, which has also been applied in recently published studies on coping and technostress (Nisafani et al., 2020; Pirkkalainen et al., 2019) and is in line with the JD-R model. Hence, according to our opinion and recent literature, coping can also act as a moderator and have a buffering effect on the relationship between technostress creators and long-term outcomes. This emphasizes the difference to "coping […] as a mediator of short-term emotional reactions" known from Lazarus and Folkman (1987, p. 147).

In addition to modelling coping as a moderator, we also distinguished the specific nature of coping and examined the influence of different coping styles. Thus, we extend recent literature (Nisafani et al., 2020; Pirkkalainen et al., 2019) which focused on a distinction between proactive coping (i.e., strengthening one's ability to cope) - and reactive coping, neglecting the different types of reactive coping. Dysfunctional coping like alcohol or drug consumption as a reactive form of coping has not been thoroughly investigated. For example, addiction in the context of ICT use is most salient in behavioural addiction like consumption of pornography or extensive gaming (Tarafdar, Maier, Laumer, & Weitzel, 2020) while there is less focus on substance abuse. We were able to provide evidence that this aspect should not be neglected in IS research.

Furthermore, we shed light on the role of coping mechanisms used to reduce technostress and, therefore, provide knowledge for the conceptual model of Nisafani et al. (2020) that is in its current form solely covering causal effects of technostress. By doing this, we expand the current knowledge of the existing technostress literature dealing with coping, which is an as-yet less studied research area (Pirkkalainen et al., 2019; Tarafdar et al., 2019).

Overall, technostress research is a highly interdisciplinary field, while it simultaneously is the very essence of IS research community (Sarker et al., 2019). Such plurality of research perspectives is important to create a deeper understanding of emerging threats due to ICT use. Accordingly, this paper brings together psychology and IS research by successfully applying the JD-R model to investigate the relationships between job demands, exhaustion, and productivity and examining the role of coping in the context of ICT use. Within our study, we extend the synthesis of these research fields by particularly meeting the recommendations for further investigating the under-researched role of strategies that individuals deploy to overcome strain caused by ICT used in an occupational setting.

6.2 Practical Implications

Our results provide valuable insights for practitioners who aim to meet technostress efficiently. Therefore, we extend the recently published conceptual model of work-related technostress by Nisafani et al. (2020) by adding active-functional and dysfunctional coping to the list of existing inhibitors, thus addressing the gap mentioned by the authors. In doing this, we support organizations to better deal with the organizational and individual-level outcomes of using ICTs and provide three suggestions, namely: (1) the appropriate level of demands; (2) the effect of different types of coping strategies; and (3) a categorization of employees with different coping styles.

First, for optimizing employees' job performance, employers should ensure that their employees are exposed to the right level of demands for achieving a high level of productivity. A very low as well as an excessive level of job demands should be avoided. Otherwise, the employee would be under- or overcharged which may result in lower job performance.

Second, regarding coping strategies for meeting technostress, both employees and employers have to carefully deal with the temptations of dysfunctional coping due to the stronger influence on the relationship of job demands and exhaustion: dysfunctional strategies may induce serious consequences in a long-term perspective, e. g., alcohol consumption naturally leading to negative health consequences which disturb employees' life as a whole, or a low level of content-related involvement with job demands leading to a reduced competence build-up. In this context, employers have to be aware of both their economic as well as social responsibilities: they may increase the support for their employees in applying active-functional coping in order to reduce its effort and, hence, increase the beneficial effects of these strategies in overcoming technostress. Simultaneously, even though dysfunctional coping may seem to be an adequate strategy to overcome technostress, it is crucial to convey the fact that other problems, like addiction, could arise in the long run as well. Employers should be aware of this double-edged sword and take preventive measures to identify individuals with addiction risk. In practice, there are some common measures to identify and support employees with addictive behaviour, e. g., companies and work councils hold regular information events to sensitize both managers and employees to the subject of addiction. Besides, managers should participate in training programs to provide them with the necessary know-how to identify and support potentially addicted employees. Overall, stakeholders like companies, works councils, managers, employees, company doctors, occupational safety specialists, among others, should ensure this is put to practice and promote appropriate handling of dysfunctional coping.

Third, to reinforce the mitigating effect of coping strategies to overcome technostress, companies should further support their employees regarding their specific coping behaviour: employees who use few different ways of coping should be encouraged to acquire a broader repertoire of various coping strategies for effectively tackling different kinds of stressful situations. At the same time, employees who predominantly use one kind of strategy (active-functional or dysfunctional) are recommended to adopt the other category as well and should be supported by their employer in expanding their respective coping behaviour. In this context, it appears highly important to be aware of the long-term health issues of dysfunctional coping, especially if employees often use dysfunctional strategies (predominantly or in combination with active-functional strategies). Hence, employers should ensure to provide know-how regarding these long-term issues by establishing specific health initiatives.

6.3 Limitations and Future Research

Besides the provided insights, our study has several limitations that have to be considered. We used a cross-sectional study design to investigate coping as a moderator where the relationships are based on covariance analysis. Thereby, it is important to note that this does not imply causality. We cannot infer whether dysfunctional coping leads to higher exhaustion from the cross-sectional data assessed at one point in time. Causality may just flow the other way round. For example, individuals who feel exhausted might tend to cope with stressful situations in a dysfunctional manner by consuming alcohol, drugs, or behavioural disengagement, respectively. This would mean that dysfunctional coping is not that dysfunctional at all. Besides, we have looked at coping strategies in general instead of actual coping actions to derive broader findings. In doing so, we took Prinz et al. (2012) as a reference and looked at two possible coping strategies - namely active-functional coping and dysfunctional coping. Although we could already derive compelling contributions and implications from this distinction, a differentiated consideration regarding coping strategies could lead to further insights. Finally, we have focused our analyses only on one component of strain - exhaustion. In addition to this, there are further options such as other burnout facets, absence duration, or general health complaints, which may be taken into account.

To summarize, applying the JD-R model within the technostress context by considering coping as moderating the relationship of technostress creators and strain delivers interesting insights contradicting prior results. For future research, we argue that coping as a moderator should be further investigated. Our results extend current knowledge in the IS in terms of coping for overcoming technostress while arguing for further interdisciplinary studies necessary to provide useful knowledge. In doing so, it might be particularly interesting to provide longitudinal and cross-level designs to investigate the effects of dysfunctional coping. The evidence suggests that causality flows in both directions (Hauk et al., 2019). Behavioural disengagement leads to increased strain, and, in turn, a higher level of strain leads to increased behavioural disengagement at a later point in time. Further coping responses are dynamic und users shift from one strategy to another in the process of coping (Salo, Makkonen, & Hekkala, 2020). Hence, it would be interesting to understand coping processes better across time. Furthermore, considering a broader set of different coping strategies could lead to more sophisticated results and enable practitioners to design and support more specific measures to address the negative consequences of ICT use.

Overall, since we successfully put together both IS and psychological stress literature and therefore address the call for further studies proposed by Tarafdar et al. (2019), this paper enriches technostress research regarding the moderating effects of coping strategies and, building on this, further studies which examine coping as moderating the effects of technostress on various outcomes are highly recommended.

Appendix

Table 5. Rotated component matrix from exploratory factor analysis of the two coping subscales.

	Factor	
Item	1	2
Brief COPE 2	0.57	
Brief COPE 3		0.67
Brief COPE 4		0.74
Brief COPE 5	0.58	
Brief COPE 7	0.72	
Brief COPE 8		0.59
Brief COPE 10	0.72	
Brief COPE 11		0.75
Brief COPE 13	**0.49**	**0.48**
Brief COPE 14	0.75	
Brief COPE 15	0.62	
Brief COPE 21	**0.53**	**0.41**
Brief COPE 23	0.67	
Brief COPE 25	0.65	
Brief COPE 26	**0.41**	**0.53**

Note. Results of a principal axis factoring with varimax rotation. Number of factors was determined through parallel criterium. Factor loadings < .35 are not printed. Cross-loadings are in boldface, these items were excluded for the analysis of the measurement and the structural model.

Table 6. Items of the coping scales: wording, descriptive statistics, and factor loadings.

	M	*SD*	Loading
Active-functional coping			
Brief COPE 7: I've been taking action to try to make the situation better.	0.88	0.84	0.70
Brief COPE 10: I've been getting help and advice from other people.	0.76	0.77	0.76
Brief COPE 14: I've been trying to come up with a strategy about what to do.	0.86	0.84	0.72
Brief COPE 15: I've been getting comfort and understanding from someone.	0.50	0.69	0.70
Brief COPE 23: I've been trying to get advice or help from other people about what to do.	0.63	0.73	0.72
Brief COPE 25: I've been thinking hard about what steps to take.	0.69	0.84	0.68
Dysfunctional coping			
Brief COPE 3: I've been saying to myself "this isn't real".	0.34	0.61	0.69
Brief COPE 4: I've been using alcohol or other drugs to make myself feel better.	0.24	0.54	0.77
Brief COPE 8: I've been refusing to believe that it has happened.	0.34	0.59	0.63
Brief COPE 11: I've been using alcohol or other drugs to help me get through it.	0.22	0.53	0.79

Note. Items which were excluded during the analysis of the measurement model are omitted. Factor loadings were obtained from confirmatory factor analysis in SEM.

Table 7. Detailed results of the moderated-mediation model

	Productivity			Exhaustion		
Predictor	*Est*	*SE*	z^a	*Est*	*SE*	z^a
Job demands	0.12***	0.04	4.19	0.44***	0.06	14.64
Exhaustion	-0.25***	0.02	-9.22			
Active-functional coping (A)				-0.05*	0.05	-2.25
Dysfunctional coping (D)				0.31***	0.09	8.10
Coping (A) × job demands				-0.05**	0.03	-2.61
Coping (D) × job demands				-0.12***	0.06	-4.85
R^2		0.05			0.36	

Note. Standardized path coefficients are displayed. [a]Bootstrapped standard errors were used for the interpretation of the results.

Table 8. Conditional indirect effects from the moderated mediation model.

Moderator values		Indirect effect		
A	D	*Est*	*SE*	z^a
Low A (–1 *SD*)	Low D (–1 *SD*)	-0.12***	0.02	-8.22
Medium A (*M*)	Low D (–1 *SD*)	-0.11***	0.02	-8.04
High A (+1 *SD*)	Low D (–1 *SD*)	-0.10***	0.02	-7.58
Low A (–1 *SD*)	Medium D (*M*)	-0.09***	0.02	-6.80
Medium A (*M*)	Medium D (*M*)	-0.09***	0.02	-6.51
High A (+1 *SD*)	Medium D (*M*)	-0.08***	0.02	-5.96
Low A (–1 *SD*)	High D (+1 *SD*)	-0.07***	0.02	-3.93
Medium A (*M*)	High D (+1 *SD*)	-0.06***	0.02	-3.58
High A (+1 *SD*)	High D (+1 *SD*)	-0.05**	0.02	-3.13

Note. Standardized path coefficients are displayed. [a]Bootstrapped standard errors were used for the interpretation of the results of the indirect effects. * $p < .05$, ** $p < .01$, *** $p < .001$.

References

Ângelo, R. P., & Chambel, M. J. (2014). The role of proactive coping in the Job Demands–Resources Model: A cross-section study with firefighters. *European Journal of Work and Organizational Psychology, 23*(2), 203–216. https://doi.org/10.1080/1359432X.2012.728701

Apt, W., Bovenschulte, M., Hartmann, E. A., & Wischmann, S. (2016). *Foresight-Studie „Digitale Arbeitswelt"* [Foresight Study "Digital Working World"]. Retrieved from https://www.bmas.de/SharedDocs/Downloads/DE/PDF-Publikationen/Forschungsberichte/f463-digitale-arbeitswelt.pdf

Ashill, N. J., Rod, M., & Gibbs, T. (2015). Coping with stress: A study of retail banking service workers in Russia. *Journal of Retailing and Consumer Services, 23*, 58–69. https://doi.org/10.1016/j.jretconser.2014.12.006

Atanasoff, L., & Venable, M. A. (2017). Technostress: Implications for adults in the workforce. *The Career Development Quarterly, 65*(4), 326–338. https://doi.org/10.1002/cdq.12111

Ayyagari, R., Grover, V., & Purvis, R. (2011). Technostress: Technological antecedents and implications. *MIS Quarterly, 35*(4), 831. https://doi.org/10.2307/41409963

Bakker, A. B., & Demerouti, E. (2017). Job demands-resources theory: Taking stock and looking forward. Journal of Occupational Health Psychology, 22(3), 273–285. https://doi.org/10.1037/ocp0000056

Barber, L. K., Conlin, A. L., & Santuzzi, A. M. (2019). Workplace telepressure and work-life balance outcomes: The role of work recovery experiences. Stress and Health: Journal of the International Society for the Investigation of Stress, 35(3), 350–362. https://doi.org/10.1002/smi.2864

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, *65*(5), 676–696. https://doi.org/10.1177/0013164404272507

Bondanini, G., Giorgi, G., Ariza-Montes, A., Vega-Muñoz, A., & Andreucci-Annunziata, P. (2020). Technostress dark side of technology in the workplace: A scientometric analysis. *International Journal of Environmental Research and Public Health*, *17*(21). https://doi.org/10.3390/ijerph17218013

Braukmann, J., Schmitt, A., Ďuranová, L., & Ohly, S. (2018). Identifying ICT-related affective events across life domains and examining their unique relationships with employee recovery. *Journal of Business and Psychology, 33*(4), 529–544. https://doi.org/10.1007/s10869-017-9508-7

Brod, C. (1984). *Technostress: The Human Cost of the Computer Revolution*. Reading, MA: Addison-Wesley.

Califf, C. B., Sarker, S., & Sarker, S. (2020). The bright and dark sides of technostress: A mixed-methods study involving healthcare IT. *MIS Quarterly*, *44*(2), 809–856. https://doi.org/10.25300/MISQ/2020/14818

Carver, C. S. (1997). You want to measure coping but your protocol's too long: Consider the brief COPE. *International Journal of Behavioral Medicine, 4*(1), 92–100. https://doi.org/10.1207/s15327558ijbm0401_6

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology, 56*(2), 267–283. https://doi.org/10.1037//0022-3514.56.2.267

Chen, A., & Karahanna, E. (2014). Boundaryless technology: Understanding the effects of technology-mediated interruptions across the boundaries between work and personal life. *AIS Transactions on Human-Computer Interaction, 6*(2), 16–36. https://doi.org/10.17705/1thci.00059

Christ-Brendemühl, S., & Schaarschmidt, M. (2020). The impact of service employees' technostress on customer satisfaction and delight: A dyadic analysis. *Journal of Business Research*, *117*, 378–388. https://doi.org/10.1016/j.jbusres.2020.06.021

Day, A., Paquet, S., Scott, N., & Hambley, L. (2012). Perceived information and communication technology (ICT) demands on employee outcomes: The moderating effect of organizational ICT support. *Journal of Occupational Health Psychology, 17*(4), 473–491. https://doi.org/10.1037/a0029837

Day, A., Scott, N., & Kevin Kelloway, E. (2010). Information and communication technology: Implications for job stress and employee well-being. In P. L. Perrewé & D. C. Ganster (Eds.), *Research in Occupational Stress and Well Being: Vol. 8. New developments in theoretical and conceptual approaches to job stress* (pp. 317–350). https://doi.org/10.1108/S1479-3555(2010)0000008011

Demerouti, E., Bakker, A. B., Nachreiner, F., & Schaufeli, W. B. (2001). The job demands-resources model of burnout. *The Journal of Applied Psychology, 86*(3), 499–512. https://doi.org/10.1037//0021-9010.86.3.499

Demerouti, E., Mostert, K., & Bakker, A. B. (2010). Burnout and work engagement: A thorough investigation of the independency of both constructs. *Journal of Occupational Health Psychology, 15*(3), 209–222. https://doi.org/10.1037/a0019408

Erschens, R., Loda, T., Herrmann-Werner, A., Keifenheim, K. E., Stuber, F., Nikendei, C., Zipfel, S., & Junne, F. (2018). Behaviour-based functional and dysfunctional strategies of medical students to cope with burnout. *Medical Education Online, 23*(1), 1–11. https://doi.org/10.1080/10872981.2018.1535738

Federal Statistical Office of Germany (2018a). Erwerbstätige und Arbeitnehmer nach Wirtschaftsbereichen (Inlandskonzept) 1 000 Personen [Employees by economic sector (national concept) by unit of 1000 persons]. Retrieved from https://www.destatis.de/DE/Themen/Arbeit/Arbeitsmarkt/Erwerbstaetigkeit/Tabellen/arbeitnehmer-wirtschaftsbereiche.html

Federal Statistical Office of Germany (2018b). Erwerbstätige und Erwerbstätigenquote nach Geschlecht und Alter 2008 und 2018: Ergebnis des Mikrozensus [Employment and employment rate by gender and age in 2008 and 2018: results of the microcensus]. Retrieved from https://www.destatis.de/DE/Themen/Arbeit/Arbeitsmarkt/Erwerbstaetigkeit/Tabellen/erwerbstaetige-erwerbstaetigenquote.html

Fischer, T., & Riedl, R. (2020). On the stress potential of an organisational climate of innovation: A survey study in Germany. *Behaviour & Information Technology*, *67*(1), 1–22. https://doi.org/10.1080/0144929X.2020.1836258

Florkowski, G. W. (2019). HR technologies and HR-staff technostress: an unavoidable or combatable effect? *Employee Relations: The International Journal*, *41*(5), 1120–1144. https://doi.org/10.1108/ER-08-2018-0214

Folkman, S., Lazarus, R. S., Dunkel-Schetter, C., DeLongis, A., & Gruen, R. J. (1986). Dynamics of a stressful encounter: Cognitive appraisal, coping, and encounter outcomes. *Journal of Personality and Social Psychology, 50*(5), 992–1003. https://doi.org/10.1037/0022-3514.50.5.992

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(4), 625–641. https://doi.org/10.1080/10705510903203573

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.2307/3151312

Frese, M. (1986). Coping as a moderator and mediator between stress at work and psychosomatic complaints. In D. Meichenbaum (Series Ed.) & M. H. Appley & R. Trumbull (Vol. Eds.), *The Plenum Series on Stress and Coping: Vol. 2. Dynamics of stress* (pp. 183–206). https://doi.org/10.1007/978-1-4684-5122-1_10

Galluch, P., Grover, V., & Thatcher, J. (2015). Interrupting the workplace: Examining stressors in an information technology context. *Journal of the Association for Information Systems, 16*(1), 1–47. https://doi.org/10.17705/1jais.00387

Gaudioso, F., Turel, O., & Galimberti, C. (2016). Explaining work exhaustion from a coping theory perspective: Roles of techno-stressors and technology-specific coping strategies. *Studies in Health Technology and Informatics, 219*, 14–20. https://doi.org/10.3233/978-1-61499-595-1-14

Gaudioso, F., Turel, O., & Galimberti, C. (2017). The mediating roles of strain facets and coping strategies in translating techno-stressors into adverse job outcomes. *Computers in Human Behavior, 69*, 189–196. https://doi.org/10.1016/j.chb.2016.12.041

Geiser, C. (2011). *Datenanalyse mit Mplus: Eine anwendungsorientierte Einführung* [Data analysis with Mplus. An application-oriented introduction.] (2nd rev. ed.). https://doi.org/10.1007/978-3-531-93192-0

Gimpel, H., Hosseini, S., Huber, R. X. R., Probst, L., Röglinger, M., & Faisst, U. (2018). Structuring digital transformation: A framework of action fields and its application at ZEISS. *Journal of Information Technology Theory and Application, 19*(1), 31–54.

Golden, T. D. (2012). Altering the effects of work and family conflict on exhaustion: Telework during traditional and nontraditional work hours. *Journal of Business and Psychology, 27*(3), 255–269. https://doi.org/10.1007/s10869-011-9247-0

Griffith, T. L., Fuller, M. A., & Northcraft, G. B. (1998). Facilitator influence in group support systems: Intended and unintended effects. *Information Systems Research, 9*(1), 20–36.

Harman, H. H. (1967). *Modern factor analysis* (2. rev. ed.). Chicago, IL: University of Chicago Press.

Hartl, E. (2019). A characterization of culture change in the context of digital transformation. In *Proceedings of the 25th Americas Conference on Information Systems AMCIS*. Symposium conducted at the meeting of the Association for Information Systems, Cancun, Mexico. Retrieved from https://aisel.aisnet.org/amcis2019/national_cultures_and_is/national_cultures_and_is/2/

Hauk, N., Göritz, A. S., & Krumm, S. (2019). The mediating role of coping behavior on the age-technostress relationship: A longitudinal multilevel mediation model. *PloS One, 14*(3), 1-22. https://doi.org/10.1371/journal.pone.0213349

Hofmans, J., Debusscher, J., Dóci, E., Spanouli, A., & Fruyt, F. de (2015). The curvilinear relationship between work pressure and momentary task performance: The role of state and trait core self-evaluations. *Frontiers in Psychology, 6*, 1–11. https://doi.org/10.3389/fpsyg.2015.01680

Janssen, O. (2001). Fairness perceptions as a moderator in the curvilinear relationships between job demands, and job performance and job satisfaction. *Academy of Management Journal, 44*(5), 1039–1050. https://doi.org/10.2307/3069447

Jöreskog, K. G. (1970). A general method for estimating a linear structural equation system. *ETS Research Bulletin Series, 1970*(2), i-41. https://doi.org/10.1002/j.2333-8504.1970.tb00783.x

Jöreskog, K. G., & Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling. Issues and techniques* (pp. 57–88).

SemTools: Useful Tools for Structural Equation Modeling (Version 0.5-2) [Computer software] (2019). Retrieved from https://CRAN.R-project.org/package=semTools

Kahler, C. W., Ramsey, S. E., Read, J. P., & Brown, R. A. (2002). Substance-induced and independent major depressive disorder in treatment-seeking alcoholics: Associations with dysfunctional attitudes and coping. *Journal of Studies on Alcohol, 63*(3), 363–371. https://doi.org/10.15288/jsa.2002.63.363

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96*(1), 201–210. https://doi.org/10.1037/0033-2909.96.1.201

Khaoula, S., Khalid, C., & Omar, T. (2020). The impact of stress due to digital communication on productivity: The exploratory study. *International Journal of Business and Administrative Studies*, *6*(4). https://doi.org/10.20469/ijbas.6.10003-4

Klotz, U. (1997). Informationsarbeit und das Ende des Taylorismus [Information work and the end of Taylorism]. In J. Krämer, J. Richter, J. Wandel, & G. Zinsmeister (Eds.), Talheimer Sammlung kritisches Wissen: Vol. 23. *Schöne neue Arbeit: Die Zukunft der Arbeit vor dem Hintergrund neuer Informationstechnologien* (1st ed., pp. 113–134). Mössingen-Talheim: Talheimer.

Knoll, N., Rieckmann, N., & Schwarzer, R. (2005). Coping as a mediator between personality and stress outcomes: A longitudinal study with cataract surgery patients. *European Journal of Personality, 19*(3), 229–247. https://doi.org/10.1002/per.546

Kortte, K. B., Wegener, S. T., & Chwalisz, K. (2003). Anosognosia and denial: Their relationship to coping and depression in acquired brain injury. *Rehabilitation Psychology*, *48*(3), 131–136. https://doi.org/10.1037/0090-5550.48.3.131

Lazarus, R. S., & Folkman, S. (1984). *Stress, Appraisal, and Coping* (11. [print.]). New York: Springer.

Lazarus, R. S., & Folkman, S. (1987). Transactional theory and research on emotions and coping. *European Journal of Personality*, *1*(3), 141–169.

Lee, A. S., Thomas, M., & Baskerville, R. L. (2015). Going back to basics in design science: from the information technology artifact to the information systems artifact. *Information Systems Journal*, *25*(1), 5–21.

Leung, K., Huang, K.-L., Su, C.-H., & Lu, L. (2011). Curvilinear relationships between role stress and innovative performance: Moderating effects of perceived support for innovation. *Journal of Occupational and Organizational Psychology, 84*(4), 741–758. https://doi.org/10.1348/096317910X520421

Lewin, J. E., & Sager, J. K. (2009). An investigation of the influence of coping resources in salespersons' emotional exhaustion. *Industrial Marketing Management, 38*(7), 798–805. https://doi.org/10.1016/j.indmarman.2008.02.013

Lin, G.-C., Wen, Z., Marsh, H., & Lin, H.-S. (2010). Structural equation models of latent interactions: Clarification of orthogonalizing and double-mean-centering strategies. *Structural Equation Modeling: A Multidisciplinary Journal, 17*(3), 374–391. https://doi.org/10.1080/10705511.2010.488999

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717. https://doi.org/10.1037/0003-066X.57.9.705

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, *1*(4), 173.

Mahapatra, M., & Pati, S. P. (2018). Technostress creators and burnout: A Job Demands-Resources Perspective. In *Sigmis-CPR'18: Proceedings of the 2018 ACM SIGMIS Conference on Computers and People Research* (pp. 70–77). New York, NY, USA: ACM Association for Computing Machinery. https://doi.org/10.1145/ 3209626.3209711

Maslach, C., & Jackson, S. E. (1986). *Maslach Burnout Inventory* (2nd ed.). Palo Alto, CA: Consulting Psychologist Press.

McFatter, R. M. (1979). The use of structural equation models in interpreting regression equations including suppressor and enhancer variables. *Applied Psychological Measurement*, *3*(1), 123–135. https://doi.org/10.1177/014662167900300113

Naditch, M. P., Gargan, M. A., & Michael, L. B. (1975). Denial, anxiety, locus of control, and the discrepancy between aspirations and achievements as components of depression. *Journal of Abnormal Psychology*, *84*(1), 1–9. https://doi.org/10.1037/h0076254

Ninaus, K., Diehl, S., Terlutter, R., Chan, K., & Huang, A. (2015). Benefits and stressors - perceived effects of ICT use on employee health and work stress: An exploratory study from Austria and Hong Kong. *International Journal of Qualitative Studies on Health and Well-Being, 10*, 28838. https://doi.org/10.3402/qhw.v10.28838

Nisafani, A. S., Kiely, G., & Mahony, C. (2020). Workers' technostress: A review of its causes, strains, inhibitors, and impacts. *Journal of Decision Systems*, *61*(10), 1–16. https://doi.org/10.1080/12460125.2020.1796286

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory. McGraw-Hill Series in Psychology*. New York, NY: McGraw-Hill.

Osmundsen, K., Iden, J., & Bygstad, B. (2018). Digital transformation: Drivers, success factors, and implications. In *Proceedings of the 12th Mediterranean Conference on Information Systems MCIS*. Symposium conducted at the meeting of the Association for Information Systems, Corfu, Greece. Retrieved from https://aisel.aisnet.org/mcis2018/37/

Patel, J., Ryoo, S., & Kettinger, W. (2012). Theorizing the dual role of information technology in technostress research. In *Proceedings of the 18th Americas Conference on Information Systems AMCIS*. Symposium conducted at the meeting of the Association for Information Systems, Seattle, Washington, USA.

Pava, C. H. P. (1983). Designing managerial and professional work for high performance: A sociotechnical approach. *National Productivity Review*, *2*(2), 126–135. https://doi.org/10.1002/npr.4040020204

Pirkkalainen, H., Salo, M., Tarafdar, M., & Makkonen, M. (2019). Deliberate or instinctive? Proactive and reactive coping for technostress. *Journal of Management Information Systems*, *36*(4), 1179–1212.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. The *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Prinz, P., Hertrich, K., Hirschfelder, U., & Zwaan, M. de (2012). Burnout, Depression und Depersonalisation – Psychologische Faktoren und Bewältigungsstrategien bei Studierenden der Zahn- und Humanmedizin [Burnout, depression and depersonalisation -psychological factors and coping strategies in dental and medical students]. *GMS Zeitschrift Für Medizinische Ausbildung*, *29*(1), 1-14. https://doi.org/10.3205/zma000780

R: A Language and Environment for Statistical Computing (Version 3.6.1) [Computer software] (2019). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Ragu-Nathan, T. S., Tarafdar, M., Ragu-Nathan, B. S., & Tu, Q. (2008). The consequences of technostress for end users in organizations: Conceptual development and empirical validation. *Information Systems Research*, *19*(4), 417–433. https://doi.org/10.1287/isre.1070.0165

Psych: Procedures for Psychological, Psychometric, and Personality Research (Version 1.8.12) [Computer software] (2019). Evanston, IL: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Riedl, R., Kindermann, H., Auinger, A., & Javor, A. (2012). Technostress from a neurobiological perspective. *Business & Information Systems Engineering*, *4*(2), 61–69. https://doi.org/10.1007/s12599-012-0207-7

Rossel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

RStudio: Integrated Delopment Environment for R (Version 1.2.5001) [Computer software] (2019). Boston, MA: PCB. Retrieved from http://www.rstudio.com/

Ryan, S. D., Harrison, D. A., & Schkade, L. L. (2002). Information-technology investment decisions: when do costs and benefits in the social subsystem matter? *Journal of Management Information Systems*, *19*(2), 85–127.

Salo, M., Makkonen, M., & Hekkala, R. (2020). The interplay of IT users' coping strategies: Uncovering momentary emotional load, routes, and sequences. *MIS Quarterly*, *44*(3), 1143–1175. https://doi.org/10.25300/MISQ/2020/15610

Santuzzi, A. M., & Barber, L. K. (2018). Workplace telepressure and worker well-being: The intervening role of psychological detachment. *Occupational Health Science, 2*(4), 337–363. https://doi.org/10.1007/s41542-018-0022-8

Saris, W. E., Batista-Foguet, J. M., & Coenders, G. (2007). Selection of indicators for the interaction term in structural equation models with interaction. *Quality & Quantity, 41*(1), 55–72. https://doi.org/10.1007/s11135-005-3956-2

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the IS discipline: Its historical legacy and its continued relevance. *MIS Quarterly*, *43*(3), 695–720.

Searle, B. J., & Lee, L. (2015). Proactive coping as a personal resource in the expanded job demands–resources model. *International Journal of Stress Management, 22*(1), 46–69. https://doi.org/10.1037/a0038439

Sonnentag, S., Kuttler, I., & Fritz, C. (2010). Job stressors, emotional exhaustion, and need for recovery: A multi-source study on the benefits of psychological detachment. *Journal of Vocational Behavior*, *76*(3), 355–365. https://doi.org/10.1016/j.jvb.2009.06.005

Srivastava, S. C., Chandra, S., & Shirish, A. (2015). Technostress creators and job outcomes: Theorising the moderating influence of personality traits. *Information Systems Journal*, *25*(4), 355–401. https://doi.org/10.1111/isj.12067

Stride, C. B., Gardner, S. E., Catley, N., & Thomas, F. (2019). *Mplus code for mediation, moderation and moderated mediation models*. Retrieved from http://www.offbeat.group.shef.ac.uk/FIO/models_and_index.pdf

Sumiyana, S., & Sriwidharmanely, S. (2020). Mitigating the harmful effects of technostress: Inducing chaos theory in an experimental setting. *Behaviour & Information Technology*, *39*(10), 1079–1093. https://doi.org/10.1080/0144929X.2019.1641229

Tams, S., Ahuja, M., Thatcher, J., & Grover, V. (2020). Worker stress in the age of mobile technology: The combined effects of perceived interruption overload and worker control. *The Journal of Strategic Information Systems*, *29*(1), 101595. https://doi.org/10.1016/j.jsis.2020.101595

Tarafdar, M., Cooper, C. L., & Stich, J.-F. (2019). The technostress trifecta - techno eustress, techno distress and design: Theoretical directions and an agenda for research. *Information Systems Journal, 29*(1), 6–42. https://doi.org/10.1111/isj.12169

Tarafdar, M., Maier, C., Laumer, S., & Weitzel, T. (2020). Explaining the link between technostress and technology addiction for social networking sites: A study of distraction as a coping behavior. *Information Systems Journal*, *30*(1), 96–124. https://doi.org/10.1111/isj.12253

Tarafdar, M., Pullins, E., & Ragu-Nathan, T. S. (2015). Technostress: Negative effect on performance and possible mitigations. *Information Systems Journal, 25*(2), 103–132. https://doi.org/10.1111/isj.12042

Tarafdar, M., Tu, Q., & Ragu-Nathan, T. S. (2010). Impact of technostress on end-user satisfaction and performance. *Journal of Management Information Systems, 27*(3), 303–334. https://doi.org/10.2753/MIS0742-1222270311

Tarafdar, M., Tu, Q., Ragu-Nathan, B. S., & Ragu-Nathan, T. S. (2007). The impact of technostress on role stress and productivity. *Journal of Management Information Systems, 24*(1), 301–328. https://doi.org/10.2753/MIS0742-1222240109

Tarafdar, M., Tu, Q., Ragu-Nathan, T. S., & Ragu-Nathan, B. S. (2011). Crossing to the dark side: Examining creators, outcomes and inhibitors of technostress. *Communications of the ACM*, *54*(9), 113–120. https://doi.org/10.1145/1995376.1995403

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tzelgov, J., & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin, 109*(3), 524–536. https://doi.org/10.1037/0033-2909.109.3.524

Wallace, L., Keil, M., & Rai, A. (2004). How software project risk affects project performance: An investigation of the dimensions of risk and an exploratory model. *Decision Sciences*, *35*(2), 289–321.

Wang, W., Kakhki, M. D., & Uppala, V. (2017). The interaction effect of technostress and non-technological stress on employees' performance. In *Proceedings of the Americas Conference on Information Systems: Organizational Transformation & Information Systems (SIGORSA)*. [S. l.]: Association for Information Systems.

Xia, Y., & Yang, Y. (2019). Rmsea, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods, 51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yip, B., Rowlinson, S., & Siu, O. L. (2008). Coping strategies as moderators in the relationship between role overload and burnout. *Construction Management and Economics, 26*(8), 871–882. https://doi.org/10.1080/01446190802213529

Zhao, X., Xia, Q., & Huang, W. (2020). Impact of technostress on productivity from the theoretical perspective of appraisal and coping processes. *Information & Management*, *57*(8), 103265. https://doi.org/10.1016/j.im.2020.103265

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research, 37*(2), 197–206. https://doi.org/10.1086/651257

VI Examining Technostress at Different Types of Data Scientists' Workplaces

Authors: Derra, Nicholas Daniel; Regal, Christian; Rath, Simon Henrik; Kühlmann, Torsten M.

Submitted to: Scandinavian Journal of Information Systems (2020)

Abstract: Data scientists represent a heterogeneous occupational group that has reached high relevance due to the wide-spread availability of quantitative data generated in the rapid progress of digital transformation. These employees play a crucial role in gaining competitive advantages for companies out of such big data. In this context, employees who frequently analyse data often occupy different job titles and, therefore, are difficult to detect. At the same time, a psychological downside of digitalization, which is called technostress, has risen. However, these issues caused by the use of information and communication technologies are rarely examined in the context of specific occupational groups and workplace attributes. Considering these challenges, this article extends current technostress research by focusing on technostress within the specific job class of data scientists. We classify different types of data scientists' workplaces through performing latent class analysis using several workplace attributes within a sample of n=486 German data scientists. Subsequently, we reveal considerable distinctions between these classes regarding the intensity of technostress creators, strains due to ICT use, and job performance. We discuss our empirical findings and deliver theoretical contributions as well as practical implications for both employees and employers and starting points for future research.

Keywords: Technostress, Strain, Digitalization, Workplace, Data Scientist

1 Introduction

Digitalization has already changed numerous aspects of individuals, economies, and society (Fitzgerald et al. 2013; Gimpel et al. 2018). Doing this with an enormous progression also greatly changes the architectures and environments of workplaces and, therefore, challenging employees in requiring new capabilities for efficiently handling work tasks (Okkonen et al. 2019; Schwemmle & Wedde 2012; Timonen & Vuori 2018). Digitalization also created new digital jobs or excessively raised their relevance for companies, for example, information security officers (Botta et al. 2007), software developers (Britto et al. 2018), and data scientists (Murawski & Bick 2017). Considering data scientists, this job class has been proven to be very heterogeneous in the context of tasks and required skills (Davenport 2020; Ismail & Abidin 2016; Mauro et al. 2018). Furthermore, employees who fulfil analytical work tasks of data scientists are often not classified as such but occupy other job titles. Thus, data scientists are hard to detect within companies. Therefore, further knowledge regarding the attributes of data scientists' workplaces is required.

Further, the digital transformation is mainly characterized by implementing a vast number of digital technologies (Hartl 2019; Osmundsen et al. 2018)). Besides the advantages of those innovative digital tools, the work-related use of information and communication technologies (ICT) may induce a specific form of stress called technostress (Ayyagari et al. 2011; Ragu-Nathan et al. 2008; Tarafdar et al. 2007; Tarafdar et al. 2010). This concept has already been introduced during the 1980s as the inability to healthfully handle ICT use (Brod 1984). Technostress may occur if employees feel unable to successfully adapt or to keep up with the multiple developments regarding digital technologies due to skills which are no longer required because of new software, an abundance of information, frequent interruptions through numerous communication channels, or the overlap of work and leisure time through continuous availability (Tarafdar et al. 2010). Considering this, the changes through digital transformation at work have to be classified as ambivalently (Apt et al. 2016).

Several studies have also shown that, in general, technostress is related to lower productivity, job satisfaction, and loyalty to the employer as well as negative consequences regarding health outcomes (Ayyagari et al. 2011; Srivastava et al. 2015; Tarafdar et al. 2010; Tarafdar et al. 2011). Hence, there is a particular importance of further investigating technostress at work. Existing research, however, primarily focuses on the general circumstances of technostress perceived by employees (Ayyagari et al. 2011; Ragu-Nathan et al. 2008; Tarafdar et al. 2007; Tarafdar et al. 2019) but, at the same time, lacks to take a more in-depth look at specific job profiles in order to get a more specific understanding of employees' technostress and, further, to examine whether there is a need to define different strategies to overcome technostress. Such investigations are of prominent relevance regarding job categories with a high level of digitalization at the workplace due to the close connection of technostress and ICT use.

Considering psychological research, numerous studies have already dealt with specific occupational groups regarding their respective level of work stress (see, for example, Grace & van Heuvelen (2019), Rees & Cooper (1992), Travers & Cooper (1993)). Furthermore, it has been shown that various job-related (see, for example, Hambrick et al. (2005), Hartline & Ferrell (1996)) and company-related (see, for example, Dekker & Barling (1995), Thompson et al. (1996)) are associated with different levels of work stress as well. However, current technostress research neglects this job-related context and particularly focuses on general relationships of technostress constructs (Ayyagari et al. 2011; Ragu-Nathan et al. 2008; Tarafdar et al. 2007; Tarafdar et al. 2019). Therefore, we argue that a substantial gap in technostress research has to be filled to gain a deeper understanding of stress due to ICT use in an organisational context. For overcoming this deficit, we propose an extension of technostress research by examining technostress in the context of specific job profiles. Thereby, we aim to investigate technostress within the occupational group of data scientists, which has been proven to play a crucial role in gaining competitive advantages for companies in today’s business environments (Costa & Santos 2017; Ismail & Abidin 2016; Mauro et al. 2018). To achieve this, our study is structured as follows: first, we outline the theoretical background of data scientists and technostress research. Building on this, we define categories of data scientists’ workplaces based on job attributes using a sample of $n = 486$ employees who fulfil data science work tasks. The empirically generated categories are analysed with respect to technostress creators, strains due to ICT use, and job performance. To conclude, we will discuss our results and provide theoretical contributions and practical implications as well as approaches for future research.

2 Theoretical Background

2.1 Data Scientists

Various studies have already confirmed the importance of data-driven managerial decision-making (Ferraris et al. 2019; Müller et al. 2018; Wamba et al. 2017), showing that big data analytics increase the performance of organisations and, thus, build competitive advantages. Employees who are able to efficiently handle and create knowledge out of data have reached particular relevance through the increased availability, compilation, and storage of huge amounts of data provided by the digital transformation of businesses, leading to a great demand for these employees (Davenport 2020; Ismail & Abidin 2016; Mauro et al. 2018). Though, it has been challenging to pinpoint tasks and responsibilities of these so-called data scientists: researchers have explored job profiles (Costa & Santos 2017), educational curricula (Richards & Marrone 2014), or gathered key insights from experts (Mikalef et al. 2018; Stanton & Stanton 2016) to identify a data scientist's required skills and knowledge.

Regarding the occupational dimension of skill variety proposed by Hackman & Oldham (1976), the data scientist's job is associated with a wide variety of required skills and knowledge domains. In this context, analytical and statistical skills are particularly relevant (Costa & Santos 2017; Doyle 2019; Ismail & Abidin 2016; Richards & Marrone 2014). Following this skill variety, analyses of occupational profiles have shown that under the single term "data scientist", many different occupational roles have been developed in business practice, e. g., business analysts, data engineers, statisticians, and data analysts (Baškarada & Koronios 2017; Ho et al. 2019; Mauro et al. 2018). This variety of roles exists due to the heterogeneous application domains, organisational structures, and purposes of data processing. Therefore, a data scientist's job can be regarded as more of an umbrella term comprising heterogeneous tasks and requirements (Doyle 2019; Mauro et al. 2018).

Considering this variety of tasks and requirements, research has summarized that a person fulfilling all the requirements of a data scientist can hardly be found in the labour market – rendering the person a "Unicorn Data Scientist" (Baškarada & Koronios 2017; Davenport 2020; Davenport & Patil 2012). Therefore, defining a data scientist as an expert who extracts knowledge from collected data as well as manages the whole data lifecycle and regarded IT infrastructures as proposed by (Manieri et al. 2015) seems to be unrealistic in the context of real company environments. Furthermore, recent sudies focus on examining data scientists' tasks and roles but, at the same time, little is known about the workplace environments of data scientists.

In addition, employees who fulfil some of a data scientist's tasks are also hard to find within a company. The tasks of efficiently analysing data are spread on several employees with various job titles since large datasets occur in nearly every department of a company (Janssen et al. 2017) and, moreover, due to the necessity of exhibiting broad domain knowledge for efficiently performing data science (Waller & Fawcett 2013). Consequently, these employees do not work at the analysis of data full-time and do not hold related occupational titles, but, at the same time, their job descriptions require data science skills and they frequently fulfil data science tasks. Hence, employees who frequently work as part-time data scientists can not be detected through classifications based on job titles but have to be identified by their tasks. Nevertheless, for enhancing data scientists' performance by tackling technostress, it is crucial to detect employees who frequently fulfil data scientist tasks.

In this context, considering the job description for data scientists proposed by the German Federal Employment Agency (2020), data scientists do screen work, comprising both customer interaction and teleworking. Their work is mostly dependent on the usage of numerous ICT: they frequently use a variety of hard- and software including operating systems, the internet, telephone, network systems, information and knowledge management systems, development software, and statistical software. However, almost all of these ICTs are not job-specific since their use is generally common in office workplaces. Yet, the frequent application of statistical software seems to be a well-performing attribute to classify a data scientist's workplace since data science work implicates the analysis of various data. Therefore, we define an employee working as a data scientist not as a person holding specific job titles but based on the everyday use of statistical software programs.

2.2 Technostress

According to the Transactional Model of Stress, stress refers to a process where individuals appraise a given situation's demands as going beyond their resources (Lazarus & Folkman 1984). For the use of digital technologies, technostress represents a specific type of stress emerging from ICT use (Tarafdar et al. 2019). Technostress results from individuals' efforts to handle ICTs' progression and the shift of physical, social, and cognitive demands within the working context (Tarafdar et al. 2007). Hence, employees are likely to experience technostress (Ragu-Nathan et al. 2008).

Technostress is induced by several specific factors known in the literature. A wide-spread categorization of these factors (or technostress creators, respectively) has been proposed by Tarafdar et al. (2007), distinguishing:

- *Techno-uncertainty* as employees' confusion created by new developments regarding the technologies used at the workplace;
- *Techno-insecurity* as employees' fear of being replaced by other employees with higher knowledge in ICT use or by ICT itself;
- *Techno-overload* as employees` requirements to work faster, longer, and even more due to ICT usage;

- *Techno-invasion* as blurred boundaries between work-related and private issues and time periods;
- *Techno-complexity* as employees`feelings of having a lack of skills in handling job-related technologies.

Suppose individuals appraise the intensity or frequency of the technostress creators above-mentioned as exceeding their existing resources. In that case, these technostress creators culminate in technostress-related strain, defined as the generic term for an individual's psychological, physical, or behavioural responses to technostress creators (Atanasoff & Venable 2017). Since strains are strongly related to a reduced level of job performance (Bakker et al. 2008; Bakker & Demerouti 2017; Taris 2006), overcoming strain at work is of particular importance to both employers and employees. In this context, technostress research distinguishes several facets of strain due to ICT use, for example, mental exhaustion (Srivastava et al. 2015) or psychological detachment (Barber et al. 2019; Santuzzi & Barber 2018).

Since strain is a general construct with various manifestations (Cooper et al. 2001), it is crucial to focus on emerging strains related to technostress creators (Ayyagari et al. 2011; Salanova et al. 2007). Thus, for examining the overall level of strain induced by ICT use, an analysis of strain associated with specific technostress creators is required. The relationships between technostress creators and strains due to the use of ICT as well as job performance are depicted in Figure 1.

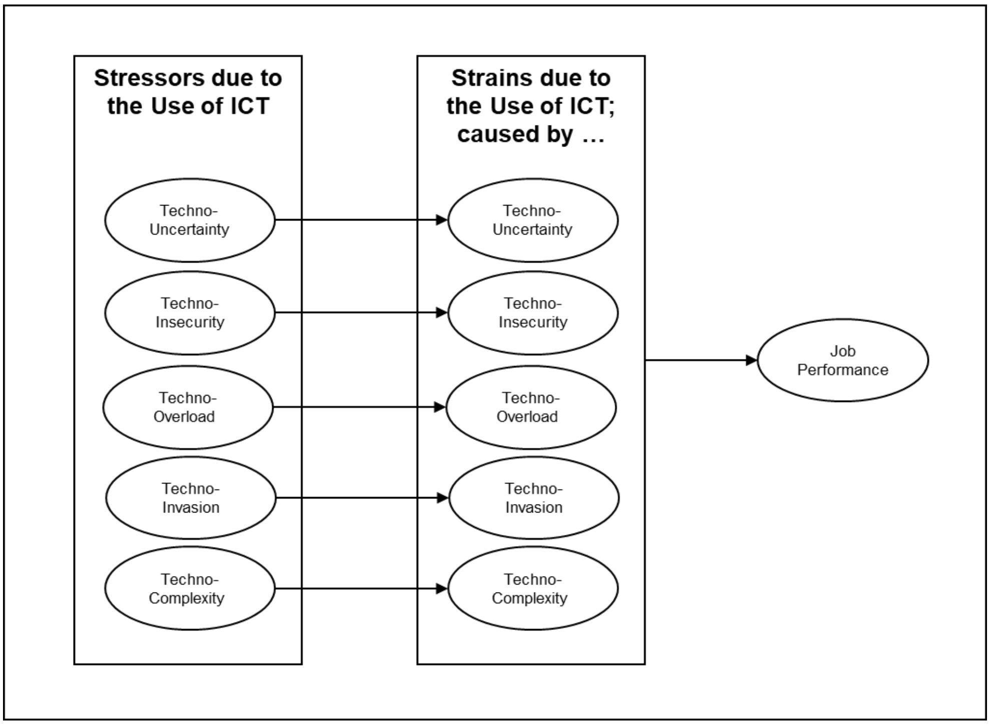

Figure 1. The relationships between technostress creators, strains due to the use of ICT, and job performance

In the past, studies that dealt with technostress have focused on the relationships between these established constructs in general (Ayyagari et al. 2011; Fischer & Riedl 2020; Ragu-Nathan et al. 2008; Tarafdar et al. 2007; Tarafdar et al. 2010; Tarafdar et al. 2011; Tarafdar et al. 2015) but, at the same time, little is known about technostress in the context of occupational groups as well as specific workplace-related attributes. In contrast, numerous psychological studies investigate occupational groups separately in order to gain a deeper understanding of their respective specificities (see, for example, Grace & van Heuvelen (2019), Rees & Cooper (1992), Travers & Cooper (1993)). In addition, psychological research has already proven relationships between several workplace characteristics and stress caused by occupational settings: work stress is therefore associated with jobs that exhibit customer contact (Hartline & Ferrell 1996) or a leadership function (Ganster 2005; Hambrick et al. 2005). Furthermore, Golubic et al. (2009) have provided empirical evidence that lower educational background is related to higher work stress levels. Considering company-related characteristics, work stress is also related to the company size (Dekker & Barling 1995; van Dijkhuizen & Reiche 1980) and different dimensions of organisational culture within a company (Lansisalmi et al. 2000; Thompson et al. 1996). More specifically, higher levels of work stress are associated with large enterprises (Dekker & Barling 1995), less perceived support culture (Dekker & Barling 1995), and greater bureaucracy (Chan et al. 2000).

Due to these findings, we argue that it is necessary to investigate technostress within specific occupational groups and, moreover, in the context of various workplace attributes to create a more detailed understanding of occurring technostress for being able to assist employees with overcoming it. Due to their importance for modern businesses and their highly digitalized workplaces, we consider the occupational group of data scientists suitable for examining technostress. Since research has proven several relationships between workplace attributes and work stress, we aim to determine whether different classes of data scientists' workplaces differ in terms of technostress creators, technostress-related strains, and overall job performance.

3 Methodology

3.1 Sample

The data we used for our examination were collected within a large research project examining technostress among German employees and developing preventive measures to efficiently reduce technostress at work. After running a quantitative pre-test containing $n_{pre} = 445$ participants, the data of the main study with a sample size of $n_{final} = 4{,}560$ participants was collected by an external panel provider. The applied questionnaire included numerous control variables to test representativity (age, sex, industry, employment status, number of hours worked per week). Regarding the control variables age, sex, and industry, preliminary analysis showed that the main study sample represents the German workforce (Federal Statistical Office of Germany 2018a, 2018b). The participants confirmed they were over 18 years old and have read the information about the research project itself, data processing, and data protection. Participants have further been informed that withdrawal from their approval of participation anytime without any negative consequences. After completing the questionnaire, an expense allowance was paid to the participants.

For identifying data scientists within our sample, we subsampled full-time workers ($number\ of\ hours\ worked\ per\ week \geq 35$) who utilize statistical software daily. After data cleaning (invalid responses and outliers), the subsample consisted of $n = 486$ data scientists. Within the sample, the female-male ratio is 32.30% to 67.70% and about 55.14% of the participants possess an academic background (see Table 1).

Gender	N	%	Digital Literacy	n	%
Male	329	67.70	Low	74	15.22
Female	157	32.30	High	412	84.78

Age	N	%	Education	n	%
<21	23	4.73	Primary School Education	5	1.03
21-24	112	23.04	Secondary School Education	43	8.85
25-39	147	30.25	High School	60	12.35
40-59	128	26.34	Completed Apprenticeship	110	22.63
60-64	46	9.47	College Degree (Bachelor)	104	21.40
>65	30	6.17	College Degree (Master)	141	29.01
			Dissertation (PhD)	23	4.73

Table 1. Demographic properties of the sample (n=486)

3.2 Measures

The questionnaire was phrased in German. Three German native speakers translated the questions which were originally formulated in English and established a final wording for translating the items. The questions were kept simple, specific, concise, without ambiguous questions, comprehensible for avoiding common method bias (Podsakoff et al. 2003). Since decreasing evaluation apprehension reduces common method bias as well (Podsakoff et al. 2003), participants were further informed that the items could not be answered right or wrong. Finally, the measures were carefully validated with a quantitative pre-test with $n_{pre} = 445$ external respondents. Besides the construction requirements described above, we additionally performed Harman's single factor test (Harman 1967) to consider possible common method bias within our data. For this, we conducted an unrotated principal component analysis with all items we used for group comparisons (Chang et al. 2010; Podsakoff et al. 2003; Tehseen et al. 2017). Since the highest proportion of variance attributed to one factor was about 17.07%, common method bias is not considered as a problem within the examined data.

Considering the heterogeneity of data scientists' workplaces, we asked the participants for both job-related and company-related attributes in order to develop a general picture of their respective workplaces. For this, we focused on attributes which have already been proven to be related to employees' stress at work, i. e., customer contact (Hartline & Ferrell 1996), the required educational level (Golubic et al. 2009) and leadership function (Ganster 2005; Hambrick et al. 2005) for the job dimension and, further, company size (Dekker & Barling 1995; van Dijkhuizen & Reiche 1980) and organisational culture (Chan et al. 2000; Dekker & Barling 1995; Lansisalmi et al. 2000; Thompson et al. 1996) representing the company dimension. Since the use of ICT at work also represents a highly relevant characteristic in the context of technostress, we added the workplace's degree of digitalization as another job-related attribute.

Customer contact, leadership function, the level of requirement, and company size were asked in a binary format (see Table 2). Following Gimpel et al. (2019), we measured the degree of digitalization via the number of technologies used at work and their frequency of use. In doing so, we asked for the use of 40 widely used technologies (Gimpel et al. 2018), using a 5-point rating scale ranging from $0 = never$ to $4 = several\ times\ a\ day$. The number and the frequency of technologies at work were then combined to a degree of digitalization, which is classified into four categories through median splits: few technologies rarely used, few technologies frequently used, many technologies rarely used, and many technologies frequently used. For describing organisational culture, we used the organisational culture index with its elements innovativeness, support, and bureaucracy as proposed by (Wallach 1983) with a 5-point rating scale from $0 = not\ at\ all$ to $4 = entirely$. Median splits transformed the answers into binary categories.

Aspect	*Indicators*	*Characteristics*
Job	Customer Contact	Yes; No
	Leadership Function	Yes; No
	Requirement Level	Non-academic; Academic
	Degree of Digitization	Few, Rarely; Few, Often; Many, Rarely; Many, Often
Company	Company Size	less than 250; 250 or more
	Innovative Culture	low; high
	Supportive Culture	low; high
	Bureaucratic Culture	low; high

Table 2. Overview of the measures and their ranges for LCA

The five technostress creators – techno-uncertainty, techno-insecurity, techno-overload, techno-invasion, and techno-complexity – were assessed by established and validated scales proposed by Ragu-Nathan et al. (2008): techno-uncertainty was measured with four items (e.g., "There are constant changes in computer software in our organisation."); techno-insecurity is captured by five items (e. g., "I have to constantly update my skills to avoid being replaced."); techno-overload was measured with four items (e. g., "I am forced by this technology to work with very tight time schedules."); techno-invasion encompasses 3 items (e.g., "I have to be in touch with my work even during my vacation due to this technology."); techno-complexity includes five items (e. g., "I need a long time to understand and use new technologies"). All items for measuring technostress creators, strain due to the use of ICT, and job performance, were asked using a 5-point Likert-type rating scale ranging from $0\ = I\ do\ not\ agree\ at\ all$ to $4\ = I\ totally\ agree$. For measuring strain due to ICT use, the participants responded to the question "And how much does that strain you?" after every item regarding the respective technostress creator. For measuring the strain due to the use of ICT, we used a 5-point Likert-type rating scale from $0 = not\ at\ all$ to $4 = very\ largely$. By that, we measured the overall level of strain due to ICT use and determined the level of strain caused by the respective technostress creator. In addition, job performance was measured by four self-report items regarding work performance as proposed by Chen & Karahanna (2014). The items asked for both fulfilling general workplace demands and success in handling work tasks (e. g., "I have a reputation in this organisation for doing my work very well.").

3.3 Means of Analysis

For analysing the data, we utilized the open-source software R (R Development Core Team 2019) and the R Studio user interface (RStudio Team 2019). After subsampling the daily-users of statistical software and examining the data through descriptive analysis, we performed a Latent Class Analysis (LCA) using the workplace attributes explained above to identify subgroups of data scientists.

We used the attributes – customer contact, leadership function, required educational level, degree of digitization, company size, level of innovativeness, level of support, and level of bureaucracy – as indicators and conducted LCAs that specified 2 to 8 classes each while repeating these computations ten-times for robustness. We applied well-established fit measures for evaluating LCA models using log-likelihood-ratio G^2 test for goodness of fit, which has been proven to work better than χ^2 test for LCA (Nylund et al. 2007) and both the Akaike Information Criterion AIC (Akaike 1974) and the Bayesian Information Criterion BIC (Schwarz 1978) for model comparison. We implemented LCA using the specific R package 'poLCA' (Linzer & Lewis 2011).

After identifying the best latent class model, we compared the discovered classes of data scientists regarding their perceived level of technostress creators, strain due to ICT use, and job performance through running group comparisons. Since descriptive analysis showed that the data is both not normally distributed and contains heterogeneity of variance, we implemented the van der Waerden normal score test (van der Waerden 1952) since it has proven to deliver superior results compared to both parametric (ANOVA test) and nonparametric (Kruskal-Wallis test) test irrespective of whether the assumptions of normality and homogeneity of variance apply for the samples (Hageman 1992; Tucker 1994).

Similar to the Kruskal-Wallis test, the van der Waerden normal score test replaces ranks with so-called normal scores $W_{i,j}$ which are inverse normal statistics calculated from quantiles within the standard normal distribution through

$$W_{i,j} = \Phi^{-1}(\frac{R(X_{i,j})}{N+1})$$

where $\Phi - 1$ denotes the normal quantile function, $X_{i,j}$ is the ith value within the jth group, $R(X_{i,j})$ is the the assigned rank of $X_{i,j}$, n_i is the size of sample i, and $N = \sum n_i$ is the size of all samples combined. The van der Waerden normal score test statistic W is then defined as

$$W = \frac{(N-1)\sum_{i=1} \frac{(\sum_j W_{i,j})^2}{n_i}}{\sum_{i=1}\sum_{j=1} {W_{i,j}}^2}$$

with $W_{i,j}$ as the jth expected normal score in the ith sample (Feir-Walsh & Toothaker 1974; van der Waerden 1952).

We first examined global comparisons for every technostress creator, strain variable, and job performance (α = 0.05). If a global test was significant, we further implemented pairwise comparisons with controlling for family-wise error rates via Holm-Bonferroni method (Holm 1979) for investigating the specific differences between the data scientist workplaces. For investigating the effect sizes, we further considered Vargas and Delaney's A (Vargha & Delaney 2000).

Since perceived technostress is also related to employees' age (Ragu-Nathan et al. 2008; Şahin & Çoklar 2009), we further tested for homogeneity of the latent groups regarding age through another van der Waerden normal score test. The result was not significant ($p = 0.275$), so the groups' differences regarding technostress cannot be explained by age differences.

4 Results

4.1 Latent Class Analysis

Considering LCA's results regarding data scientists' workplace attributes, we first exclude the model with two classes of workspaces since this model is significant for log-likelihood-ratio G^2 (compare Table 3). Regarding goodness of fit, the model with eight classes achieves the best values. Simultaneously, the model with four classes shows the best (or rather lowest) value for AIC, while the model with three classes performs best for BIC. Thus, these models have to be examined in more detail.

Given our goal of detecting explainable workplace classes, a split into eight types would separate the sample into sparse groups and, further, seems rather complex, which can be seen at high scores in both AIC and BIC. Therefore, the model with eight groups is rejected. Regarding AIC, the model with four classes is preferred, while model three performs best in BIC, so both models seem to be comparable in balancing fit and complexity. Based on these results, we compare the models' goodness of fit where the model with four classes outperforms regarding log-likelihood-ratio G^2. Furthermore, considering the sample's distribution among the different types of workplaces within the models, we find a noticeable imbalance through a very dominant type containing more than 50% of the sample within the 3-type model. Hence, we select the model with four classes.

Classes	*df*	*log-likelihood*	*G²*	*p (G²)*	*χ²*	*p (χ²)*	*AIC*	*BIC*
2	465	-2626.427	536.590	0.012	675.376	0.001	5294.854	5382.764
3	454	-2578.857	441.149	0.659	500.940	0.063	5221.713	**5355.672**
4	443	-2556.279	395.993	0.947	480.029	0.109	**5198.558**	5378.565
5	432	-2545.930	375.296	0.977	472.970	0.085	5199.860	5425.915
6	421	-2537.080	358.023	0.988	459.732	0.094	5204.588	5476.691
7	410	-2529.055	340.752	0.992	421.515	0.375	5209.316	5527.468
8	399	**-2523.295**	**329.021**	**0.997**	**392.895**	**0.882**	5219.586	5583.786

Table 3. The goodness of fit measures of the LCA for the varying number of assumed classes

Table 4 shows the impact of the indicators' characteristics on the respective association of a data scientist with a type of workplace as well as the distribution of the sample. We consider influences with a probability of $\geq \frac{2}{3}$ for binary and $\geq \frac{1}{3}$ for quaternary indicators as a major characteristic.

Dimension	*Indicators*	*Characteristic*	*Latent Class* Type 1 (83)	Type 2 (91)	Type 3 (225)	Type 4 (87)
Job	Customer Contact	Yes	**0.906**	**0.857**	**0.834**	0.508
		No	0.094	0.143	0.166	0.492
	Leadership Function	Yes	0.651	**0.669**	**0.906**	0.200
		No	0.349	0.332	0.094	**0.800**
	Requirement Level	Non-academic	**0.794**	0.380	0.295	0.475
		Academic	0.207	0.620	**0.705**	0.526
	Degree of Digitization	Few, Rarely	0.106	0.200	0.070	0.150
		Few, Often	**0.586**	0.281	0.182	**0.660**
		Many, Rarely	0.153	**0.346**	**0.421**	0.081
		Many, Often	0.155	0.172	0.327	0.110
Company	Company Size	less than 250	**0.914**	0.614	0.319	0.055
		250 or more	0.087	0.386	**0.681**	**0.945**
	Innovativeness	low	0.209	**0.923**	0.064	0.342
		high	**0.792**	0.077	**0.936**	0.659
	Support	low	0.167	**0.783**	0.094	0.526
		high	**0.833**	0.217	**0.906**	0.474
	Bureaucracy	low	0.194	**1.000**	0.074	0.239
		high	**0.806**	0.000	**0.926**	**0.761**

Table 4. The probabilities that one class holds a specific characteristic; bold values are remarkable for the respective type of workplace compared to the other types

Considering these results, we are now able to distinguish classes of data scientists' workplaces as follows:

Type 1 – Customer Service Management within SMEs (CSM-SME): workplaces that require direct contact to the customer; furthermore, the data scientists working here use only a few ICT but make often use of them; this workplace is particularly common in small and medium-sized companies and does not require academic know-how; employees tend to work in innovative companies with a strong supportive culture but, at the same time, have to deal with high bureaucracy.

Type 2 – Customer Interaction Lead Position with Low Levels of Innovativeness, Support, and Bureaucracy (CIL-noISB): workplaces with leadership function that also require customer contact; these workplaces tend to appear within enterprises exhibiting a low culture of innovation and support as well as bureaucracy; in addition, a broad range of ICT is exploited while the individual technologies are rarely used.

Type 3 – Customer Interaction Lead Position within Large Enterprises (CIL -LE): workplaces comprising direct contact with customers which are also associated with both academic background and a leadership position; herein, a large number of ICT is utilized in different frequencies; this workplace type often occurs in large enterprises having a high level of innovative, support, and bureaucratic culture.

Type 4 – Back Office Expertise within Large Enterprises (BOE-LE): workplaces that are not associated with management responsibilities; only a few ICT are used here but, at the same time, these technologies are frequently utilized; this type of workplace is particularly common in large companies holding a dominant bureaucratic culture.

Considering the distribution of data scientists in this context, it is notable that the highest percentage of data scientists are assigned to CIL-LE with about 46.3% ($n_{CIM-LE} = 225$) while the other types of workplaces are comparably distributed with 17.0% to 18.70% each (for a detailed view of the respective group structures, see Appendix 3).

4.2 Van Der Waerden Normal Score Test

We now compare the four types in terms of both technostress creators and strains caused by ICT as well as their perceived job performance. As already pointed out, we explicitly distinguish technostress creators and strain due to ICT use as proposed in technostress literature (Ayyagari et al. 2011; Salanova et al. 2007). Table 5 shows the results for the five technostress creators and their related strains as well as the perceived job performance. The 25%, 50%, and 75% quantiles, as well as mean and standard deviation, are given for the four types of workplaces each.

Data scientists working at CIL-LE workplaces report the highest values regarding the technostress creators uncertainty, insecurity, overload, and invasion compared to the other classes and, further, the highest cumulated demands regarding the five technostress creators as well ($mean_{cum} = 1.965$). Concerning the remaining facet techno-complexity, data scientists from CIL-noISB workplaces report the highest value.

Regarding technostress-related strains, CIL-LE data scientists only hold the highest values for strains from two technostress creators, namely insecurity and invasion. However, these data scientists generally have the highest strains across all facets in total ($mean = 1.441$). The highest values for both overload- and uncertainty-related strains is now at CIL-noISB-type and no longer for CIL-LE. Furthermore, CIL-noISB occupies the highest value for complexity-related strain, consistent with the respective technostress creator. Interestingly, data scientists report the highest value for CIL-LE workplaces' job performance, despite overall highest values for technostress creators and strains due to digital technologies. In contrast, CIL-noISB report a clearly worse job performance compared to the other classes. Besides these issues, data scientists of the other workplace classes (CSM-SME and BOE-LE) do not show any apparent peculiarities in both technostress creators and strains due to the use of ICT as well as job performance.

Type	Construct	*Technostress Creator* 25% Quantile	50% Quantile	75% Quantile	M	SD	*Strain due to ICT use* 25% Quantile	50% Quantile	75% Quantile	M	SD
CSM-SME (n = (83)	Uncertainty	1.125	2.000	2.875	1.925	1.112	0.000	0.750	2.000	1.051	1.080
	Insecurity	0.400	1.200	1.900	1.282	1.082	0.000	0.600	1.600	0.872	0.992
	Overload	1.000	2.000	2.750	1.795	1.185	0.000	1.000	2.000	1.211	1.119
	Invasion	0.333	1.667	2.500	1.546	1.288	0.000	0.667	2.000	1.108	1.165
	Complexity	0.200	1.000	2.100	1.241	1.127	0.000	0.800	2.000	1.063	1.080
	Job Performance	-	-	-	-	-	2.500	3.000	3.250	2.825	0.775
CIL-noISB (n = 91)	Uncertainty	1.125	2.000	2.750	1.926	1.059	0.500	1.500	2.000	**1.412**	1.031
	Insecurity	0.600	1.800	2.200	1.523	0.968	0.000	1.400	2.000	1.266	1.029
	Overload	1.125	2.000	2.500	1.852	1.005	0.625	2.000	2.500	**1.646**	1.108
	Invasion	0.500	1.667	2.333	1.557	1.077	0.500	1.667	2.000	1.451	1.004
	Complexity	0.700	1.800	2.20	**1.527**	0.964	0.200	1.600	2.000	**1.336**	1.043
	Job Performance	-	-	-	-	-	1.750	2.000	2.750	2.159	0.824
CIL-LE (n = 225)	Uncertainty	2.000	2.750	3.000	**2.546**	0.885	0.500	1.500	2.000	1.392	1.053
	Insecurity	1.000	2.000	2.400	**1.800**	1.066	0.200	1.200	2.200	**1.308**	1.101
	Overload	1.500	2.250	2.750	**2.080**	1.049	0.500	1.750	2.500	1.623	1.126
	Invasion	1.000	2.000	2.667	**1.887**	1.146	0.333	1.667	2.333	**1.596**	1.189
	Complexity	0.400	1.400	2.400	1.514	1.197	0.000	1.000	2.200	1.284	1.155
	Job Performance	-	-	-	-	-	2.500	3.000	3.500	**3.027**	0.693
BOE-LE (n = 87)	Uncertainty	1.625	2.000	2.500	2.060	0.879	0.000	1.000	2.000	1.129	1.076
	Insecurity	0.600	1.200	2.000	1.411	1.028	0.000	0.200	1.400	0.795	0.992
	Overload	1.250	2.000	2.875	1.948	1.186	0.125	1.250	2.250	1.385	1.169
	Invasion	0.000	1.333	2.000	1.215	1.081	0.000	0.333	1.667	0.935	1.126
	Complexity	0.200	1.000	1.900	1.136	1.025	0.000	0.600	1.600	0.887	0.990
	Job Performance	-	-	-	-	-	2.000	2.500	3.000	2.641	0.771

Table 5: 25%-Quantile, 50%-Quantile, 75%-Quantile, mean (M), and standard deviation (SD) of both technostress creators and related strain, for four classes of data scientists' workplaces; bold values indicate the highest value for a technostress creators

For examining whether the detected types of workplaces differ in their levels of technostress creators and strains, we first conducted global van der Waerden normal score tests on the four classes of workplaces. Table 6 shows the results of these global tests.

Dependent Variable	*Technostress Creator*	*Strain*
Techno-Uncertainty	**< 0.001**	**0.036**
Techno-Insecurity	**0.001**	**< 0.001**
Techno-Overload	0.155	**0.020**
Techno-Invasion	**< 0.001**	**< 0.001**
Techno-Complexity	**0.028**	**0.007**
Job Performance	**< 0.001**	

Table 6. p-values for global van der Waerden normal score tests comparing the workplace classes of data scientists

Considering technostress creators, there are global differences within the subgroups for the factors techno-uncertainty, techno-insecurity, techno-invasion, and techno-complexity. Concerning the technostress-related strains, the results show that at least one class significantly differs from the others at every single technostress creator. Finally, job performance includes significant differences as well.

Subsequently, we use pairwise van der Waerden normal score tests with alpha adjusting by applying the Holm–Bonferroni method (Holm 1979) to determine which types of workplaces differ significantly. We utilized Vargas and Delaney's A (Vargha & Delaney 2000) for investigating the effect sizes. In the following, we focus on reporting significant differences that show at least a moderate effect to meet the call for statistical and practical significance (Mohajeri et al. 2020). For exact values and results from all deducted tests, see Appendix 1 and 2, respectively.

Techno-Uncertainty

For techno-uncertainty as a technostress creator, CIL-LE workplaces significantly distinguish from all other types and show explicitly higher values than all other classes. However, in terms of strain, there are no significant differences between the groups. Although a previously conducted general van der Waerden normal score test detected a significant difference between the workplace types, this difference is no longer identifiable at the level of pairwise comparisons. Thus, there is no significant difference concerning strains due to techno-uncertainty. This phenomenon of a globally significant and non-significant pairwise-test results can be observed when weak significant results (the global test had a p-value of 0.036) are further "penalized" by the correction procedure and are therefore no longer significant.

Techno-Insecurity

Regarding techno-insecurity as a technostress creator , CIL-LE again differs from CIL-SME, although, however, the difference is moderate. In this context, CIL-LE reports higher values. On the other hand, there are several significant differences in strains, e. g., CIL-LE considerably distinguishes from BOE-LE and moderately from CIL-SME, with CIL-LE exhibiting higher values. Likewise, CIL-noISB moderately differs from BOE-LE whereby CIL-noISB reports higher values.

Techno-Overload

Techno-overload as a technostress creator does not report any significant differences between the workplace classes. Interestingly, there is a moderate difference in related strain between CSM-SME and CIL-noISB, with CIL-noISB surpassing the other.

Techno-Invasion

Considering techno-invasion as a technostress creator, there is a significant variance between CIL-LE and BOE-LE, with CIL-LE reporting clearly higher values. In terms of strain and besides the respective significant difference between CIL-LE and BOE-LE, there are also significant distinctions between CIL-LE and CSM-SME as well as CIL-noISB and BOE-LE. In this context, CIL-LE has moderately higher values than CSM-SME and significantly higher values than BOE-LE. In comparison, BOE-LE reports clearly higher values than CIL-noISB.

Techno-Complexity

Although the general van der Waerden normal score test detected a significant deviation between the workplace types in terms of technostress creators, this difference disappears at the level of pairwise comparisons. Thus, there is no significant variance concerning techno-complexity as a technostress creator. In contrast, significant differences regarding strain due to techno-complexity between BOE and both CIL-noISB and CIL-LE were observed, with BOE-LE reporting moderately smaller values.

Job Performance

The differences between the types of data scientists' workplaces regarding job performance show that CIL-LE is distinctly different from both CIL-noISB and BOE-LE holding higher job performances. Furthermore, CIL-noISB also performs significantly worse than workplace CSM-SME and BOE-LE.

To sum up, CIL-LE incumbents highly differ from both CSM-SME incumbents and BOE-LE incumbents reporting higher values for technostress creators and technostress-related strains. At the same time, there are also differences between CIL-noISB incumbents and BOE-LE incumbents in terms of strain due to both techno-invasion and techno-complexity. In contrast, CIL-LE employees report higher values for perceived job performance despite their higher demands in both technostress creators and strain due to ICT use.

5 Discussion

In general, data scientists represent a highly digitalized occupational group that is important for today's companies to create knowledge and, accordingly, competitive advantages out of big data. In this paper, we contribute to the problems of detecting employees who fulfil data scientists' tasks by (i) providing a definition based on data scientists' ICT use which is closer to businesses' reality compared to other definitions in the context of job titles and (ii) detecting classes of data scientists' workplaces which differ regarding job-related and company-related attributes. In doing so, we found four kinds of workplaces: customer service management within SMEs (CSM-SME), customer interaction lead position with low levels of innovativeness, support, and bureaucracy (CIL-noISB), customer interaction lead position within large enterprises (CIL-LE), and back office expertise within large enterprises (BOE-LE), with CIL-LE being the largest class of data scientists' workplaces. This suggests that data scientists more likely hold lead positions within large enterprises and exhibit customer contact. These findings are clearly against associating data scientists' workplaces with in-house tasks. Therefore, data science expertise should be considered when hiring employees for leadership workplaces since these workplaces often require the fulfilment of data scientist tasks. Further, it is quite surprising that data scientists often report high levels of innovativeness and support along with high bureaucracy (and low levels each, respectively), which seems to be contradicting. Moreover, it is worth pointing out that data scientists' lead positions are likely to utilize many ICT technologies but use them quite rarely. In contrast, employees without lead responsibilities tend to use relatively few technologies commonly. Thus, leaders have to gain broader knowledge due to the use of ICT.

Subsequently, we found significant differences between the groups regarding technostress. The groups report different levels of technostress creators as well as related strains and, in particular, vary regarding the composition of technostress' roots (i. e., the technostress creators) and suffering (i. e., the technostress-related strains). The results suggest that data scientists holding leadership positions are higher demanded by ICT developments which may be caused by top-down strategies for launching new technologies. Furthermore, leaders within SMEs seem to be less demanded due to new ICT compared to leaders in large enterprises. Also, it is notable that CIL-LE seem to feel more replaceable than CSM-SME incumbents regarding ICT knowledge, while there is no significant difference compared to BOE-LE incumbents. I. e., the combination of leadership and working within a large enterprise seems to guide data scientists to feel less important for their company in terms of ICT-related knowledge. The results further indicate that the use of many technologies which is highly connected to leadership workplaces generally leads to higher strains in this regard and, moreover, strain due to techno-invasion rather occurs within large companies. Lastly, it is also noteworthy that BOE-LE incumbents report significantly less techno-complexity than both the leadership workplace classes. Hence, the findings lead to the conclusion that data scientists who work as leaders are especially in danger of perceiving technostress creators as well as strain due to the use of ICT and, further, employees within large enterprises are more likely to perceive strain due to techno-invasion.

Overall, CIL-LE incumbents reported the highest levels of both perceived technostress creators and technostress-related strain but, at the same time, assessed themselves with the strongest job performance. Since technostress has been shown to negatively influence job performance (Bakker et al. 2008; Bakker & Demerouti 2017; Taris 2006), CIL-LE incumbents seem to overcome this issue more efficiently compared to the other classes of data scientists. In this context, one factor could be that CIL-LE workplaces are highly associated with innovative and supportive culture within the enterprise which may enhance the feeling of being productive and, further, lead to success in performing active coping strategies like seeking social support (Carver et al. 1989). This suggestion is supported by the fact that CIL-noISB incumbents which represent the other leadership class report the worst job performance: they seem to suffer more from technostress by getting less support in overcoming it.

5.1 Theoretical Contribution

Considering technostress as an important aspect of health at the workplace both employers and employees have to carefully deal with, we contribute to current technostress research by successfully adapting concepts of work stress research regarding workplace attributes to technostress context. More specifically, we provide a job-specific view of technostress considering the highly digitalized and heterogeneous job class of data scientists by comparing the detected groups of data scientists' workplaces concerning technostress creators, technostress-related strains, and job performance.

Comparing our results with prior findings regarding the relationships between workplace attributes and general stress at work, we found both equivalent and contradicting results: while technostress goes along with workplaces exhibiting a leadership function and higher level of bureaucracy which is in line with findings regarding overall work stress (Chan et al. 2000; Ganster 2005; Hambrick et al. 2005), a higher level of education surprisingly appears to be associated with technostress as well, disagreeing with the relationship of work stress and education (Golubic et al. 2009). Moreover, technostress is associated with the use of many ICT at work independent of a rare usage while the frequent use of less technologies does not go along with higher technostress. The results further suggest that customer contact is also related to technostress perception which is in line with the relationship of customer contact and overall stress at work (Hartline & Ferrell 1996). In contrast, there are no clear impacts regarding the presence of large companies as well as high levels of both innovative and supportive culture since these attributes go along with both minor and major technostress issues.

5.2 Practical Implications

Our results provide important practical aspects for employers who aim to protect their data scientists from technostress. The variability of perceived technostress between the four types of data scientist workplaces suggests implementing different strategies for dealing with technostress within each group.

Overall, CIL-LE workplaces are associated with the highest level of both technostress creators and strains due to ICT use, so this class requires the highest support in overcoming technostress. As part of support, employers are recommended to explain both the launch process and the requirements of new ICT developments timely and in more detail for countering techno-uncertainty as well as to establish a single point of contact for employees where they may provide feedback whether a technology use is efficient for monitoring techno-overload. Furthermore, managers are suggested to protect the blurring boundaries between work and leisure by limiting employees' availability to their work time for tackling techno-invasion as well as to periodically communicate with their data scientists, underlining that they are important for the company in order to overcome techno-insecurity.

Regarding CIL-noISB incumbents, employers should concentrate on providing support regarding the use of the numerous ICT which have to be handled at these workplaces. By replacing redundant technologies and providing further tutorials for the remaining ones as well as explaining recent developments regarding the ICT used within the company, data scientists will be able to gain more profound and required know-how and the perceived strains due to techno-uncertainty and techno-complexity may be significantly reduced. Moreover, CIL-noISB incumbents should also be supported in protecting blurred boundaries, e. g., by defining clear rules regarding home office or the private use of ICT provided by the company such as mobile phones and laptops. Finally, since these workplaces are associated with significantly lower job performance than all other classes, appreciating achieved productivity is highly recommended.

Since CSM-SME and BOE-LE incumbents generally report relatively low values in technostress and, at the same time, good performance, we suggest focusing on appreciating these groups of data scientists. Further, general support regarding technostress by providing knowledge about the topic and strategies to overcome technostress is recommended.

5.3 Limitations and Future Research

Even though this paper is able to offer a deeper understanding of the heterogeneous and highly relevant job class of data scientists and, further, the level of technostress within these jobs, our investigations have several limitations that have to be taken into account. First, a self-reporting survey in the context of technostress is generally in danger of social desirability bias. Second, we used eight important workplace attributes for detecting classes of data scientist workplaces, but, at the same time, more indicators could help differentiate workplaces, for example, the possibility of using home office or flex time, which was not part of our study. Third, since we aimed to measure the overall level of strains in the context of technostress creators, we could not provide evidence regarding more fine-grained distinctions of strain, e. g., the various facets of burnout or different health issues. Lastly, we asked participants for their overall job performance which does not exhibit a certain causality to the technostress level.

Nevertheless, we were able to provide a deeper understanding of data scientists' workplaces as a job class which has reached particular importance due to the rapid evolution of digitalization at work. Moreover, we proved that technostress should also be considered in the context of individual job classes in order to effectively deal with it. Therefore, our investigations may be seen as a first step for future examinations of technostress within specific job classes and, further, with respect to other workplace attributes to distinguish the necessary internal and external resources to deal with technostress. In this context, we recommend to particularly focus on other high-digitalized jobs like, e. g., IT specialists or online marketing experts.

Appendix

Appendix 1

	CIL-noISB		*CIL-LE*		*BOE-LE*	
	Creator	*Strain*	*Creator*	*Strain*	*Creator*	*Strain*
Techno-Uncertainty						
CSM-SME	1.000	0.180	**< 0.001**	0.10	1.000	1.000
CIL-noISB	-	-	**< 0.001**	1.000	1.000	0.350
CIL-LE	-	-	-	-	**< 0.001**	0.300
BOE-LE	-	-	-	-	-	-
Techno-Insecurity						
CSM-SME	0.481	0.072	**0.001**	**0.006**	> 0.5	1.000
CIL-noISB	-	-	0.147	1.000	> 0.5	**0.025**
CIL-LE	-	-	-	-	0.067	**< 0.001**
BOE-LE	-	-	-	-	-	-
Techno-Overload						
CSM-SME	1.000	0.080	0.350	**0.037**	1.000	> 0.5
CIL-noISB	-	-	0.390	> 0.5	1.000	0.443
CIL-LE	-	-	-	-	1.000	0.434
BOE-LE	-	-	-	-	-	-
Techno-Invasion						
CSM-SME	> 0.5	0.277	0.134	**0.007**	0.134	0.418
CIL-noISB	-	-	0.066	0.418	0.134	**0.017**
CIL-LE	-	-	-	-	**< 0.001**	**< 0.001**
BOE-LE	-	-	-	-	-	-
Techno-Complexity						
CSM-SME	0.319	> 0.5	0.261	> 0.5	1.000	> 0.5
CIL-noISB	-	-	1.000	> 0.5	0.155	**0.026**
CIL-LE	-	-	-	-	0.075	**0.015**
BOE-LE	-	-	-	-	-	-

Table 7. p-values for the pairwise van der Waerden tests comparing the types of workplaces regarding technostress creators; bold values indicate significant results with α = 5% with Holm–Bonferroni correction (Holm 1979)

	CIL-noISB	*CIL-LE*	*BOE-LE*
CSM-SME	**< 0.001**	0.079	0.079
CIL-noISB	-	**< 0.001**	**< 0.001**
CIL-LE	-	-	**< 0.001**
BOE-LE	-	-	-

Table 8. p-values for the pairwise van der Waerden tests comparing the types of workplaces regarding job performance; bold values indicate significant results with α = 5% with Holm–Bonferroni correction (Holm 1979)

Appendix 2

	CIL-noISB		*CIL-LE*		*BOE-LE*	
	Creator	*Strain*	*Creator*	*Strain*	*Creator*	*Strain*
Techno-Uncertainty						
CSM-SME	0.498	0.397	**0.337**	0.405	0.467	0.474
CIL-noISB	-	-	**0.329**	0.504	0.472	0.578
CIL-LE	-	-	-	-	**0.668**	0.572
BOE-LE	-	-	-	-	-	-
Techno-Insecurity						
CSM-SME	0.421	0.388	0.364	0.382	0.452	0.525
CIL-noISB	-	-	0.423	0.489	0.546	0.630
CIL-LE	-	-	-	-	0.613	**0.640**
BOE-LE	-	-	-	-	-	-
Techno-Overload						
CSM-SME	0.479	0.385	0.423	0.392	0.461	0.459
CIL-noISB	-	-	0.429	0.504	0.481	0.570
CIL-LE	-	-	-	-	0.539	0.564
BOE-LE	-	-	-	-	-	-
Techno-Invasion						
CSM-SME	0.490	0.406	0.416	0.384	0.572	0.552
CIL-noISB	-	-	0.409	0.466	0.600	**0.650**
CIL-LE	-	-	-	-	**0.672**	**0.666**
BOE-LE	-	-	-	-	-	-
Techno-Complexity						
CSM-SME	0.406	0.417	0.435	0.446	0.518	0.552
CIL-noISB	-	-	0.509	0.516	0.624	0.637
CIL-LE	-	-	-	-	0.589	0.603
BOE-LE	-	-	-	-	-	-

Table 9. Vargha and Delaney's A for for the pairwise comparisons of the types of workplaces regarding technostress creators; bold values indicate moderate or strong effects (Tomczak & Tomczak 2014); grey values are not significant

	CIL-noISB	*CIL-LE*	*BOE-LE*
CSM-SME	**0.737**	0.433	0.588
CIL-noISB	-	**0.211**	**0.343**
CIL-LE	-	-	**0.646**
BOE-LE	-	-	-

Table 10. Vargha's and Delayne's A for the pairwise comparison of the types of workplaces regarding job performance; bold values indicate moderate and strong effects (Tomczak & Tomczak 2014); grey values are not significant

Appendix 3

Aspect	*Indicators*	*Characteristic*	*Latent Class* CSO-SME (83)	CIL-noISB (91)	CIL-LE (225)	BOE-LE (87)	*All*
Job	Customer Contact	Yes	74	79	190	42	385
		No	9	12	35	45	101
	Leadership Function	Yes	49	62	208	12	331
		No	34	29	17	75	155
	Requirement Level	Non-academic	71	35	67	43	216
		Academic	12	65	158	44	270
	Degree of Digitization	Few, Rarely	10	18	14	14	56
		Few, Often	57	25	37	61	180
		Many, Rarely	9	32	96	4	141
		Many, Often	7	16	78	8	109
Company	Company Size	less than 250	83	56	74	2	215
		250 or more	0	35	151	85	271
	Innovativeness	low	18	84	11	31	144
		high	65	7	214	56	342
	Support	Low	15	71	19	46	151
		high	68	20	206	41	335
	Bureaucracy	Low	18	91	12	21	142
		high	65	0	213	66	344

Table 11. Number of data scientists exhibiting a certain characteristic within a type of workplace

References

Akaike, H., (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, (19:6): 716–723. doi: 10.1109/TAC.1974.1100705.

Apt, W., Bovenschulte, M., Hartmann, E. A. and Wischmann, S., (2016). Foresight-Studie „Digitale Arbeitswelt“. Forschungsbericht, Berlin (accessible from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-47039-5).

Atanasoff, L. and Venable, M. A., (2017). Technostress: Implications for adults in the workforce. *The Career Development Quarterly*, (65:4): 326–338. doi: 10.1002/cdq.12111.

Ayyagari, Grover and Purvis, (2011). Technostress: Technological antecedents and implications. *MIS Quarterly*, (35:4): 831–858. doi: 10.2307/41409963.

Bakker, A. B. and Demerouti, E., (2017). Job demands-resources theory: Taking stock and looking forward. *Journal of Occupational Health Psychology*, (22:3): 273–285. doi: 10.1037/ocp0000056.

Bakker, A. B., van Emmerik, H. and van Riet, P., (2008). How job demands, resources, and burnout predict objective performance: A constructive replication. *Anxiety, Stress, & Coping*, (21:3): 309–324. doi: 10.1080/10615800801958637.

Barber, L. K., Conlin, A. L. and Santuzzi, A. M., (2019). Workplace telepressure and work-life balance outcomes: The role of work recovery experiences. *Stress and Health*, (35:3): 350–362. doi: 10.1002/smi.2864.

Baškarada, S. and Koronios, A., (2017). Unicorn data scientist: the rarest of breeds. *Program*, (51:1): 65–74. doi: 10.1108/PROG-07-2016-0053.

Botta, D., Werlinger, R., Gagné, A., Beznosov, K., Iverson, L. and Fels, S.: Fisher, B., (2007). Towards understanding IT security professionals and their tools. In: *Proceedings of the 3rd symposium on usable privacy and security (SOUPS)*: 100–111. New York, NY: ACM.

Britto, R., Cruzes, D. S., Smite, D. and Sablis, A., (2018). Onboarding software developers and teams in three globally distributed legacy projects: A multi-case study. *Journal of Software: Evolution and Process*, (30:4): 1-17. doi: 10.1002/smr.1921.

Brod, C., (1984). *Technostress: The human cost of the computer revolution*, Reading, MA: Addison-Wesley.

Carver, C. S., Scheier, M. F. and Weintraub, J. K., (1989). Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology*, (56:2): 267–283. doi: 10.1037//0022-3514.56.2.267.

Chan, K. B., Lai, G., Ko, Y. C. and Boey, K. W., (2000). Work stress among six professional groups: the Singapore experience. *Social Science & Medicine*, (50:10): 1415–1432. doi: 10.1016/S0277-9536(99)00397-4.

Chang, S. J., van Witteloostuijn, A. and Eden, L., (2010). From the editors: Common-method variance in international business research. *Journal of International Business Studies*, (41:2): 178–184. doi: 10.1057/jibs.2009.88.

Chen, A. and Karahanna, E., (2014). Boundaryless technology: Understanding the effects of technology-mediated interruptions across the boundaries between work and personal life. *AIS Transactions on Human-Computer Interaction*, (6:2): 16–36.

Cooper, C. L., Dewe, P. J., and O'Driscoll, M. P., (2001). *Organizational stress: A review and critique of theory, research, and applications*, Thousand Oaks, CA: Sage.

Costa, C. and Santos, M. Y., (2017). The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *International Journal of Information Management*, (37:6): 726–734. doi: 10.1016/j.ijinfomgt.2017.07.010.

Davenport, T., (2020). Beyond unicorns: Educating, classifying, and certifying business data scientists. *Harvard Data Science Review*, (2:2). doi: 10.1162/99608f92.55546b4a.

Davenport, T. H. and Patil, D. J., (2012). Data scientist. *Harvard Business Review*, (90:5): 70–76.

Dekker, I. and Barling, J., (1995). Workforce size and work-related role stress. *Work & Stress*, (9:1): 45–54. doi: 10.1080/02678379508251584.

Doyle, A., (2019). Important job skills for data scientists. https://www.thebalancecareers.com/list-of-data-scientist-skills-2062381, Accessed October 7, 2020.

Federal Statistical Office of Germany, (2018a). Erwerbstätige und Arbeitnehmer nach Wirtschaftsbereichen je 1.000 Personen [Employees by economic sector by unit of 1,000 persons].
https://www.destatis.de/DE/Themen/Arbeit/Arbeitsmarkt/Erwerbstaetigkeit/Tabellen/arbeit, Accessed October 12, 2020.

Federal Statistical Office of Germany, (2018b). Erwerbstätige und Erwerbstätigenquote nach Geschlecht und Alter 2008 und 2018: Ergebnis des Mikrozensus [Employment and employment rate by gender and age in 2008 and 2018: Results of the microzensus]. https://www.destatis.de/DE/Themen/Arbeit/Arbeitsmarkt/Erwerbstaetigkeit/Tabellen/erwerbstaetige-erwerbstaetigenquote.html, Accessed October 12, 2020.

Feir-Walsh, B. J. and Toothaker, L. E., (1974). An empirical comparison of the anova f-test, normal scores test and kruskal-wallis test under violation of assumptions. *Educational and Psychological Measurement*, (34:4): 789–799.

Ferraris, A., Mazzoleni, A., Devalle, A. and Couturier, J., (2019). Big data analytics capabilities and knowledge management: impact on firm performance. *Management Decision*, (57:8): 1923–1936. doi: 10.1108/MD-07-2018-0825.

Fischer, T. and Riedl, R., (2020). On the stress potential of an organisational climate of innovation: A survey study in Germany. *Behaviour & Information Technology*, (67:1), 1–22. https://doi.org/10.1080/0144929X.2020.1836258.

Fitzgerald, M., Kruschwitz, N., Bonnet, D., and Welch, M., (2013). Embracing Digital Technology. A New Strategic Imperative. *MIT Sloan Management Review*, (55:2): 1-16.

Ganster, D. C., (2005). Response: Executive Job Demands: Suggestions from a Stress and Decision-Making Perspective. *Academy of Management Review*, (30:3): 492–502. doi: 10.5465/amr.2005.17293366.

German Federal Employment Agency, (2020). Berufenet. Berufsinformationen einfach finden. https://berufenet.arbeitsagentur.de/berufenet/faces/index?path=null/suchergebnisse/kurzbeschreibung&dkz=129987&such=data+scientist, Accessed October 6, 2020.

Gimpel, H., Lanzl, J., Manner-Romberg, T., and Nüsken, N., (2018). Digitaler Stress in Deutschland. Eine Befragung von Erwerbstätigen zu Belastung und Beanspruchung durch Arbeit mit digitalen Technologien. Düsseldorf, Germany: Hans-Böckler-Stiftung.

Gimpel, H., Lanzl, J., Regal, C., Urbach, N., Wischniewski, S., Tegtmeier, P., Kreilos, M., Kühlmann, T. M., Becker, J., Eimecke, J. and Derra, N. D., (2019). Gesund digital arbeiten?!: Eine Studie zu digitalem Stress in Deutschland. Augsburg, Germany: Projektgruppe Wirtschaftsinformatik des Fraunhofer FIT. doi: 10.24406/fit-n-562039.

Golubic, R., Milosevic, M., Knezevic, B. and Mustajbegovic, J., (2009). Work-related stress, education and work ability among hospital nurses. *Journal of advanced nursing*, (65:10): 2056–2066. doi: 10.1111/j.1365-2648.2009.05057.x.

Grace, M. K. and van Heuvelen, J. S., (2019). Occupational variation in burnout among medical staff: evidence for the stress of higher status. *Social Science & Medicine*, (232): 199–208. doi: 10.1016/j.socscimed.2019.05.007.

Hackman, J. and Oldham, G. R., (1976). Motivation through the design of work: test of a theory. *Organizational Behavior and Human Performance*, (16:2): 250–279. doi: 10.1016/0030-5073(76)90016-7.

Hageman, S. J., (1992). Alternative methods for dealing with nonnormality and heteroscedasticity in paleontological data. *Journal of Paleontology*: 857–867.

Hambrick, D. C., Finkelstein, S. and Mooney, A. C., (2005). Executive job demands: New insights for explaining strategic decisions and leader behaviors. *Academy of Management Review*, (30:3): 472–491. doi: 10.5465/amr.2005.17293355.

Harman, H. H., (1967). *Modern factor analysis*: University of Chicago press.

Hartl, E., (2019). A characterization of culture change in the context of digital transformation. In: *25th Americas Conference on Information Systems (AMCIS)*. Cancún, Mexiko.

Hartline, M. D. and Ferrell, O. C., (1996). The management of customer-contact service employees: An empirical investigation. *Journal of Marketing*, (60:4): 52–70. doi: 10.1177/002224299606000406.

Ho, A., Nguyen, A., Pafford, J. L. and Slater, R., (2019). A data science approach to defining a data scientist. *SMU Data Science Review*, (2:3): 1-21.

Holm, S., (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, (6:2): 65–70.

Ismail, N. A. and Abidin, A. Z., (2016). Data scientist skills. *Journal of Mobile Computing & Application*, (3:4): 52–61.

Janssen, M., van der Voort, H. and Wahyudi, A., (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, (70:1): 338–345. doi: 0.1016/j.jbusres.2016.08.007.

Lansisalmi, H., Peiro, J. M. and Kivimaki, M., (2000). Collective stress and coping in the context of organizational culture. *European Journal of Work and Organizational Psychology*, (9:4): 527–559. doi: 10.1080/13594320050203120.

Lazarus, R. S., and Folkman, S., (1984). *Stress, appraisal, and coping*, New York, NY: Springer publishing company.

Linzer, D. A. and Lewis, J. B., (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of statistical software*, (42:10): 1–29.

Manieri, A., Brewer, S., Riestra, R., Demchenko, Y., Hemmje, M., Wiktorski, T., Ferrari, T. and Frey, J., (2015). Data science professional uncovered: How the EDISON project will contribute to a widely accepted profile for data scientists. In: *Proceedings of the 7th IEEE International Conference on Cloud Computing Technology and Science (CloudCom'15)*: 588–593: IEEE.

Mauro, A. de, Greco, M., Grimaldi, M. and Ritala, P., (2018). Human resources for big data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, (54:5): 807–817. doi: 10.1016/j.ipm.2017.05.004.

Mikalef, P., Giannakos, M. N., Pappas, I. O. and Krogstie, J., (2018). The human side of big data: Understanding the skills of the data scientist in education and industry. In: *Global Engineering Education Conference (EDUCON)*: 503–512: IEEE. doi: 10.1109/EDUCON.2018.8363273.

Mohajeri, K., Mesgari, M. and Lee, A. S., (2020). When statistical significance is not enough: Investigating relevance, practical significance, and statistical significance. *MIS Quarterly*, (44:2): 525–559. doi: 10.25300/MISQ/2020/13932.

Müller, O., Fay, M. and vom Brocke, J., (2018). The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. *Journal of Management Information Systems*, (35:2): 488–509. doi: 10.1080/07421222.2018.1451955.

Murawski, M. and Bick, M., (2017). Digital competences of the workforce – a research topic? *Business Process Management Journal*, (23:3): 721–734. doi: 10.1108/BPMJ-06-2016-0126.

Nylund, K. L., Asparouhov, T. and Muthén, B. O., (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal*, (14:4): 535–569.

Okkonen, J., Vuori, V. and Palvalin, M., (2019). Digitalization changing work: Employees' view on the benefits and hindrances. In: Á. Rocha, C. Ferrás & M. Paredes (Eds.). *Information Technology and Systems. Proceedings of ICITS 2019*: 165–176. Cham, Switzerland: Springer.

Osmundsen, K., Iden, J. and Bygstad, B., (2018). Digital transformation drivers, success factors, and implications. In: *The 12th Mediterranean Conference on Information Systems (MCIS)*. Corfu, Greece.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. and Podsakoff, N. P., (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *The Journal of applied psychology*, (88:5): 879–903. doi: 10.1037/0021-9010.88.5.879.

R Development Core Team, (2019). *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing.

Ragu-Nathan, T. S., Tarafdar, M., Ragu-Nathan, B. S. and Tu, Q., (2008). The consequences of technostress for end users in organizations: conceptual development and empirical validation. *Information Systems Research*, (19:4): 417–433. doi: 10.1287/isre.1070.0165.

Rees, D. and Cooper, C. L., (1992). Occupational stress in health service workers in the UK. *Stress medicine*, (8:2): 79–90. doi: 10.1002/smi.2460080205.

Richards, D. and Marrone, M., (2014). Identifying the education needs of the business analyst: An australian study. *Australasian Journal of Information Systems*, (18:2). doi: 10.3127/ajis.v18i2.803.

RStudio Team, (2019). *RStudio: Integrated delopment environment for R*, Boston, MA: PCB.

Şahin, Y. L. and Çoklar, A. N., (2009). Social networking users' views on technology and the determination of technostress levels. *Procedia - Social and Behavioral Sciences*, (1:1): 1437–1442. doi: 10.1016/j.sbspro.2009.01.253.

Salanova, M., Llorens, S., Cifre, E. and Nogreda, C., (2007). El tecnoestrés: Concepto, medida e intervención psicosocial. [Technostress: Concept, measurement and prevention]. *Notatechia de Prevencion*, (730). Madrid, Spain: INSHT.

Santuzzi, A. M. and Barber, L. K., (2018). Workplace telepressure and worker well-being: The intervening role of psychological detachment. *Occupational Health Science*, (2:4): 337–363. doi: 10.1007/s41542-018-0022-8.

Schwarz, G., (1978). Estimating the dimension of a model. *The Annals of Statistics*, (6:2): 461–464.

Schwemmle, M. & Wedde, P. (Eds.), (2012). *Digitale Arbeit in Deutschland. Potenziale und Problemlagen*, Bonn: Friedrich-Ebert-Stiftung, Medienpolitik.

Srivastava, S. C., Chandra, S. and Shirish, A., (2015). Technostress creators and job outcomes: theorising the moderating influence of personality traits. *Information Systems Journal*, (25:4): 355–401. doi: 10.1111/isj.12067.

Stanton, A. D. and Stanton, W. W., (2016). The relationship between big data, data science, digital analytics and the skills and abilities needed to optimise marketing decisions. *Applied Marketing Analytics*, (2:3): 265–279.

Tarafdar, M., Cooper, C. L. and Stich, J.-F., (2019). The technostress trifecta - techno eustress, techno distress and design: Theoretical directions and an agenda for research. *Information Systems Journal*, (29:1): 6–42. doi: 10.1111/isj.12169.

Tarafdar, M., Tu, Q., Ragu-Nathan, B. S. and Ragu-Nathan, T. S., (2007). The impact of technostress on role stress and productivity. *Journal of Management Information Systems*, (24:1): 301–328. doi: 10.2753/MIS0742-1222240109.

Tarafdar, M., Pullins, E. and Ragu-Nathan, T. S., (2015). Technostress: Negative effect on performance and possible mitigations. *Information Systems Journal,* (25:2), 103–132. https://doi.org/10.1111/isj.12042.

Tarafdar, M., Tu, Q. and Ragu-Nathan, T. S., (2010). Impact of technostress on end-user satisfaction and performance. *Journal of Management Information Systems*, (27:3): 303–334. doi: 10.2753/MIS0742-1222270311.

Tarafdar, M., Tu, Q., Ragu-Nathan, T. S. and Ragu-Nathan, B. S., (2011). Crossing to the dark side. *Communications of the ACM*, (54:9): 113–120. doi: 10.1145/1995376.1995403.

Taris, T. W., (2006). Is there a relationship between burnout and objective performance? A critical review of 16 studies. *Work & Stress*, (20:4): 316–334. doi: 10.1080/02678370601065893.

Tehseen, S., Ramayah, T. and Sajilan, S., (2017). Testing and controlling for common method variance: A review of available methods. *Journal of Management Sciences*, (4:2): 142–168.

Thompson, N., Stradling, S., Murphy, M. and O'neill, P., (1996). Stress and organizational culture. *The British Journal of Social Work*, (26:5): 647–665.

Timonen, H. and Vuori, J., (2018). Visibility of work: How digitalization changes the workplace. In: *Proceedings of the 51st Hawaii International Conference on System Sciences*. Waikoloa Village, HI, USA.

Tomczak, M. and Tomczak, E., (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*, (21:1): 19–25.

Travers, C. J. and Cooper, C. L., (1993). Mental health, job satisfaction and occupational stress among UK teachers. *Work & Stress*, (7:3): 203–219. doi: 10.1080/02678379308257062.

Tucker, J. W., (1994). Generalized van der Waerden identities. *Journal of Physics A: Mathematical and General*, (27:3): 659–662.

van der Waerden, B. L., (1952). Order tests for the two-sample problem and their power. *Indagationes Mathematicae*, (14:1): 453–458.

van Dijkhuizen, N. and Reiche, H. M., (1980). Psychosocial stress in industry: A heartache for middle management? *Psychotherapy and psychosomatics*, (34:2-3): 124–134.

Vargha, A. and Delaney, H. D., (2000). A critique and improvement of the “CL” common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, (25:2): 101–132. doi: 10.2307/1165329.

Wallach, E. J., (1983). Individuals and organizations: The cultural match. *Training and Development Journal*, (37:2): 29–36.

Waller, M. A. and Fawcett, S. E., (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, (34:2): 77–84. doi: 10.1111/jbl.12010.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R. and Childe, S. J., (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, (70:1): 356–365. doi: 10.1016/J.JBUSRES.2016.08.009.

VII Conclusion

Overall, data science has reached outstanding relevance within modern marketing research since the big data revolution has provided countless new opportunities of collecting customer data and, subsequently, gaining valuable knowledge out of such data. For creating competitive advantages out of these circumstances, it is crucial for companies to improve the performance of their marketing data scientists. In this context, new applications of machine learning approaches as well as effectively overcoming technostress as a huge downside of digitalization have been shown to be particularly important aspects of improving the performance of data scientists in contemporary marketing contexts. Despite its importance, machine learning research is still at an early stage within marketing research, with numerous opportunities of further examining new tasks and types of data. Furthermore, technostress research lacks investigations regarding both coping strategies for overcoming technostress and technostress issues within specific occupational groups. For tackling these gaps, the following research questions were proposed within this thesis:

> *RQ1: How can data scientists improve their performance by successfully applying machine learning algorithms in contemporary marketing contexts?*
>
> *RQ2: How can data scientists improve their performance by effectively overcoming technostress at work?*

Referring to the improvement of data scientists' performance by successfully applying machine learning algorithms in contemporary marketing contexts, this thesis delivers the following insights:

i. By applying deep long short term memory neural networks for classifying sentiments of written customer reviews, the intuitive variation of hyperparameters within the network architecture does not necessarily lead to improvements in accuracy. Furthermore, positive impacts on the network classification performance caused by hyperparameter variants cannot be accumulated automatically by just combining the underlying hyperparameter variants. Therefore, both researchers and practitioners have to consider the possibilities of unintuitive hyperparameter impacts including interaction effects when optimising deep learning models.

ii. Machine learning approaches successfully perform call centre arrivals' forecasting tasks, with random forest models yielding the highest prediction accuracy. Thereby, machine learning can outperform conventional time series models. Within the context of time series data, cross-validation containing an expanding rolling window constitutes a powerful instrument for comparing different approaches aiming for an optimised model selection in both research and practical environment.

iii. Machine learning approaches successfully tackle the widespread e-commerce problem of online shopping cart abandonment by predicting such abandoners based on aggregated clickstream data. To achieve this, gradient boosting with regularization delivers the strongest performance. At the same time, standard decision tree and boosted logistic regression provide similar results while exhibiting less model complexity and, therefore, represent serious alternatives. Hence, practitioners are recommended to consider their respective computational capacities and machine learning capabilities when deciding upon a machine learning approach for online shopping cart abandonment prediction.

Regarding the improvement of data scientists' performance by effectively overcoming technostress at work, the following deductions are provided:

iv. Increasing stressors due to the use of ICT are associated with higher levels of both exhaustion and productivity. Therefore, employers should avoid extremely low as well as excessive levels of technology-related job demands in order to optimise their employees' productivity.

v. By applying reactive coping as a personal resource moderating the relationship of stressors due to ICT use and employees' exhaustion within technostress context, active-functional coping is associated with lower exhaustion while dysfunctional strategies are, in turn, related to higher exhaustion. Simultaneously, both ways of coping buffer the relationship between technostress creators and exhaustion as a facet of strain. Furthermore, employees who utilize a broad set of various coping strategies may benefit more from applying coping. Nevertheless, both managers and employees are highly recommended to be aware of the long-term disadvantages dysfunctional strategies imply regarding health and performance outcomes.

vi. Due to their heterogeneity of roles and tasks, defining data scientists due to job titles appears to be inappropriate. Therefore, employers are recommended to use data scientists' specific use of digital technologies for detecting employees who fulfil data scientists' tasks.

vii. Considering general job- and company-specific workplace attributes, data scientists' workplaces can be classified into four different groups. These groups differ in the context of technostress, i. e., regarding technostress creators, strains due to the use of ICT, and overall job performance. In this context, workplaces with customer contact, a leadership function, high education, and a high level of bureaucracy are associated with perceiving more technostress, partially contradicting prior findings in the context of general stress at work. Overall, these findings lead to the necessity of providing different support strategies for effectively over-coming technostress depending on the respective class of data scientists.

Considering these results, this thesis provides new theoretical insights within machine learning and technostress as important branches of research which have reached outstanding relevance due to the developments of the big data revolution. Moreover, the revealed findings deliver practical implica-tions marketing managers are highly recommended to be aware of since marketing may strongly benefit from high-performing data scientists who are able to extract valuable knowledge out of vari-ous types of customer data.

For future research, it is important to note that both machine learning and technostress research still provide numerous issues researchers have to deal with in order to improve the performance of data scientists in marketing contexts. Therefore, further studies regarding the application of new machine learning algorithms, the examination of new kinds of marketing data, or investigations of new coping strategies to effectively overcome technostress are highly recommended.

Appendix

Appendix A: Index of Research Papers

Since this thesis is of cumulative nature, several research papers have served as a basis for it. These papers are either published in or submitted to academic journals. In the following, an overview of these papers as well as the respective journals is provided.

Research Paper #1 (Chapter II): Derra, N. D., and Baier, D. (2020). Working in Detail: How LSTM Hyperparameter Selection Influences Sentiment Analysis Results.
This paper has been published in the Archives of Data Science, Series A, 6(1), 1-22. https://doi.org/10.5445/KSP/1000098011/10.

Research Paper #2 (Chapter III): Albrecht, T., Rausch, T. M., and Derra, N. D. (2021). Call Me Maybe: Methods and Practical Implementation of Artificial Intelligence in Call Center Arrivals' Forecasting.
This paper has been published in the Journal of Business Research, 123, 267-278. https://doi.org/10.1016/j.jbusres.2020.09.033.
(VHB JOURQUAL 3: Category B)

Research Paper #3 (Chapter IV): Rausch, T. M., Derra, N. D., and Wolf, L. (2020). Predicting Online Shopping Cart Abandonment with Machine Learning Approaches.
This paper has been published in the International Journal of Market Research, Online First, 1-24. https://doi.org/10.1177/1470785320972526.
(VHB JOURQUAL 3: Category D)

Research Paper #4 (Chapter V): Becker, J., Derra, N. D., Regal, C., and Kühlmann, T. M. (2020). Mitigating the Negative Consequences of ICT Use: The Moderating Effect of Active-Functional and Dysfunctional Coping.
This paper is currently under review at the Journal of Decision Systems.
(VHB JOURQUAL 3: Category B)

Research Paper #5 (Chapter VI): Derra, N. D., Regal, C., Rath, S. H., and Kühlmann, T. M. (2020). Examining Technostress at Different Types of Data Scientists' Workplaces.
This paper is currently under review at the Scandinavian Journal of Information Systems.
(VHB JOURQUAL 3: Category C)

Appendix B: Individual Contribution to the Included Research Papers

The included research papers of this thesis were composed by several authors within different settings each. Hereafter, the composition frame of the various papers is described as well as my respective individual contribution.

Research paper #1, which is presented in chapter II, was composed by two researchers. As the lead author of this paper, I was mainly responsible for all parts of the study including its conceptualization and motivation, the theoretical background, pre-processing and analyses of the data, drawing up the paper, and editing it during the review process including the correspondence with the journal. Daniel Baier constantly participated by providing important and valuable recommendations for all parts of the paper during the creation process.

Research paper #2, which is presented in chapter III, was worked out by three authors. Based on the conceptual framework of Tobias Albrecht and Theresa Rausch, I substantially contributed within several parts of the study, i. e., supplying practically relevant literature, the exchange of knowledge with practitioners and the implementation of machine learning models (in particular, the trial of LSTM models within the study), establishing the two-step research approach, writing the article, and improving the paper during the review process.

Research paper #3, which is presented in chapter IV, was written by three researchers as well. Together with Theresa Rausch, I was centrally involved in all parts of the study including the development of its conceptualization, the selection and implementation of the machine learning models, the literature review, writing the article, and editing during the review process. Lukas Wolf was mainly responsible for the elaboration of the literature review and collaborated during the review process. As the corresponding author, I was further responsible for correspondence with the journal.

Research paper #4, which is presented in chapter V, was composed by four authors. With Julia Becker as the lead author being mainly responsible for the article, I considerably contributed to the paper at carrying out the literature review and, further, supported the conceptualization of the study as well as data analysis, writing the article, and editing during the review process. Christian Regal collaborated within the study’s conceptualization, data collection, data analysis, writing the article as well as the review process. Torsten M. Kühlmann supported the overall development and the writing process by constant advices and recommendations for improving the study.

Research paper #5, which is presented in chapter VI, was written by four researchers. As the lead author, I was mainly responsible for nearly all parts of the article, i. e., conceptualizing the study, conducting the literature review, the data analysis, the writing process, and the correspondence with the journal. Christian Regal supported the conceptualization of the study, data collection, data analysis, and writing the article. Simon Rath was involved in the literature review and Torsten M. Kühlmann constantly provided important comments and recommendations for improving the quality of the paper.