Artificial Intelligence in Business Analytics: Capturing Value with Machine Learning Applications in Financial Services

Marc Schmitt

THESIS

Submitted for the degree of

Doctor of Philosophy

Department of Computer and Information Science,

University of Strathclyde

Email: marcschmitt@hotmail.de

Date: 11.10.2020

Declaration

This thesis is the result of the author's original research. It has been composed by the author and contains minor material that has been previously submitted for examination leading to the award of a degree at the University of Strathclyde in 2016.

Signed: Marc Schmitt
Data: 11.10.2020

Acknowledgments

I would like to express my sincere gratitude to my advisor Dr. Marc Roper for the continuous support of my Ph.D. study and related research, for his patience, open-minded personality, and motivation in difficult times. I could not have imagined having a better advisor and mentor for my Ph.D. study. I would also like to thank all my friends and family members who helped me through difficult times.

Abstract

This Ph.D. thesis explores the strength and applicability of machine learning-based classifiers within the context of business analytics for data-driven decision making. The focus is on supervised binary classification on structured datasets, which are vastly present in relational databases across all enterprises. Advanced analytics has become indispensable for today's corporate world and it is demonstrated that predictive analytics is one of the major contributors to capture business value across the financial services value chain. To test this hypothesis different models as Generalized Linear Models, Random Forest, Gradient Boosting, and Artificial Neural Networks were tested, compared, and combined to test their predictive strength and robustness in different scenarios and use cases. The results indicate the superiority of Gradient Boosting when it comes to structured datasets compared to all other classifiers. This is a major reason why the diffusion of Deep Learning within business analytics is lacking behind. Also, the ensemble learning method stacking – which uses several base learners to create a more powerful super learner – proved to be a viable tool to consistently improve upon the accuracy of even the most powerful candidate models – including Gradient Boosting. Automated Machine Learning (AutoML) was benchmarked against manually tuned models and proved to be a valuable tool to democratize predictive analytics for small to medium-sized corporations and to tackle the skill shortage for ML experts. AutoML has the potential to completely automate the predictive modeling process, but it is mainly concerned with model tuning and selection while ignoring steps at the beginning and end of the pipeline. Also, an ML pipeline setup is suggested that would – once it is automated – be able to reach human expert-level prediction accuracy for binary classification on structured datasets. All those models were tested and applied in the context of different business analytics use cases – with a focus on financial services – to solve problems in credit risk management, insurance claims prediction, and marketing and sales. All use cases demonstrate improvements in prediction accuracy and hence offer direct value gains. Throughout the thesis, there is a consideration of the advantages and constraints when it comes to the use of ML models in the industry including a translation into managerial implications. Also, general economic and business implications are discussed to understand how the field will evolve in the future.

Table of Contents

1 Introduction

1.1 Digital Transformation

The world is becoming digital. More global, and corporations are increasingly subject to competition in an ultra-fast marketplace. In the last decade, there has been an astonishing increase in connectivity, stored data, and advanced analytics. New businesses found their way into the economy and challenged the status quo of many incumbents. The time we are currently witnessing has been referred to as *Industry 4.0* (Reinhard, Jesper, & Stefan, 2016), *The Fourth Industrial Revolution* (Schwab, 2016), or *The Second Machine Age* (Brynjolfsson & Mcafee, 2016). New entrances in the form of innovative start-ups are often global and the natural barriers to only local competition gradually cease to exist. So-called digital natives (e.g. Amazon, Alphabet, Facebook, etc.) started to dominate the market with new business models (e.g. platform companies) that proved to be superior to the old ways of doing business (Parker, Van Alstyne, & Choudary, 2016). It is a time where many traditional companies realized the value of digital technologies to create strategic business value and are faced with the pressure to transform or die (Siebel, 2019).

Having a look at the leading indices across the world we can observe a non-trivial rebalancing w.r.t. to market capitalizations. The most valuable companies used to be traditional companies like Exxon Mobile, General Electric, Microsoft, Gazprom, and Citigroup, but this has changed over the last years. On 28th January 2020, the most valuable companies of the S&P 500 were all digital including Amazon, Microsoft, Apple, Alphabet, and Facebook (Levi & Konish, 2020). Several fortune 500 companies have become bankrupt, where acquired in M&A activities, or developed themselves into digital enterprises with a changed business scope and focus. It is forecasted that within the next 10 years more than 40% of the existing companies today will cease to exists in its current form (Siebel, 2019). This is a worldwide phenomenon. In Germany, Wirecard, a FinTech firm specializing in global payment solutions has replaced Commerzbank from the DAX 30 in 2018, which has been the second-largest Bank in Germany and one of the founding members of the index (Storbeck, 2018). Similar developments can be observed in China with digital leaders like Alibaba, Tencent, Baidu, and JD.com (Candelon, Yang, & Wu, 2019). All those born-digital companies are also at the forefront of AI research.

Nevertheless, those changes are a normal occurrence that can be observed several times during history. New disruptive forces tend to emerge from time to time and put tremendous pressure on the existing economy until a new equilibrium is reached. This usually occurs due

to scientific breakthroughs that have enough power to fundamentally transform the world economy. Technological transformations and disruptions emerge from the sciences, but there is always a point in time when they start to gradually diffuse into the world economy and get adopted by different businesses. This leads to gradual or abrupt changes due to shifts in the job and labor markets, horizontal or lateral integration of businesses, or the formation of completely new enterprises and products, which facilitate a new wave of innovation (Siebel, 2019).

In light of those market shifts, customer expectations are changing rapidly. As the world becomes fully digital, physical stores lose continuously in value and transactions between most counterparties have moved online and become entirely digital. This can be observed around the globe and across several industries. The best example is retail as consumers continuously move their shopping activities online. Traditional bank branches are shut down and offers are exclusively placed in an online or mobile banking environment. Also, social media enables consumers to be more connected and informed, which makes quality and customer service increasingly important. Those shifts in market dynamics have led to more market power for consumers across the globe and the need to become a customer-centric, flexible, and adaptive enterprise has become mandatory.

Companies increasingly gravitate towards advanced analytics, machine learning, and artificial intelligence to compete in this new environment. It seems that the so-called digital native companies have the upper hand and incumbents struggle to transform their legacy systems into a modern big data/digital infrastructure. The advantage of those digital natives is an enterprise architecture build specifically for the modern environment, which is naturally superior compared to old legacy systems that require a complete transformation (Henke et al., 2016).

The need for corporations to survive led to a huge wave of digital transformation (DT) projects across all major industries as incumbents tried to restructure their enterprises into a modern version of themselves. What do we mean by digital transformation?

Hess, Matt, Benlian, & Wiesböck (2016) describe it in the following way: “Digital transformation is concerned with the changes digital technologies can bring about in a company’s business model, which results in changed products or organizational structures or in the automation of processes”. And it occurs according to Verhoef et al. (2019) “in response to changes in digital technologies, increasing digital competition and resulting digital customer behavior.” It is the current struggle or restructuring of old business models towards a modern digital enterprise.

This happens on a global scale and not all incumbent corporations will “survive this current era of mass extinction” (Siebel, 2019).

What portfolio of technologies constitutes digital transformation seems to be quite fluid. The main drivers commonly associated with it are big data, artificial intelligence (AI), cloud computing, robotics, the Internet of Things (IoT), and more recently blockchain. Overall, DT is an ongoing process that has no end and a strict limitation towards certain technologies seems inadequate. There is no reason not to incorporate further technological advances into the definition of what it means to be digital (Chanias, Myers, & Hess, 2019; Warner & Wäger, 2019).

1.2 Big Data and Artificial Intelligence

Several underlying reasons led to those strong shifts in market dynamics, which introduced new and disruptive business models. Big data was the first buzz word in business that was associated with the importance of data. Comments as “data is the new oil” and “data science is the sexiest job in the 21st century” in major business magazines as Harvard Business Review made everyone realize that the status quo in business and the base for the competition is about to change.

The 21st century started with the new word “Big Data” and moved gradually to the world of Data Science. The era of big data is characterized by the availability of different (old and new) data sources and is the origin of the first wave of digital transformation (Baesens, Bapna, Marsden, Vanthienen, & Zhao, 2016; Henke et al., 2016). By now, we live in a world with an immense deluge of data from different sources which increase exponentially every year (Henke et al., 2016).

For business analytics, data sources can be categorized into (1) traditional databases, (2) web-data, and (3) mobile and sensor-based data (Chen, Chiang, & Storey, 2012). Rapidly advancing information technology, storage capabilities, and general hardware improvements have made it possible to store and process those huge amounts of structured as well as unstructured data across all domains. Corporations became huge silos of information imprisoned in large databases and the surge in data availability naturally led to the need to gain insights from these volumes of data (Henke et al., 2016). Besides, cloud infrastructures and world-wide connectivity gave us direct access to those data pools on a global basis, which gives corporations fast and direct access to information and increasing flexibility in their execution (Gampfer, Jürgens, Müller, & Buchkremer, 2018; Zimmermann et al., 2018).

A second development that can be considered a close follower to the big data revolution was the rise of artificial intelligence, which triggered the second wave of digital transformation (Bughin et al., 2017; Henke et al., 2016). Artificial intelligence (AI) became an active field of research due to the advances of artificial neural networks. The breakthrough that drives the current AI revolution is called Deep Learning. Deep learning did not only help AI to increase its popularity, but also increased the scale and scope of possible applications, and is seen as one of the most disruptive technologies since the inception of the internet itself (Goodfellow, et al., 2016). DL is responsible for many aspects of the world that seem by now familiar and usual and it has the potential to further drive technological advances across all sciences and industries. It improved many tasks and brought breakthroughs in text, speech, image, video and audio processing (LeCun, et al., 2015). AI and DL are currently at the peak of the Gartner hype cycle (Columbus, 2019), but investments are strong and are still flowing towards AI (Bughin et al., 2017). The consensus is that we have moved from fundamental progress to the application of AI across various sciences, businesses, and governments (Stadelmann et al., 2018).

The field of economics has also picked up on the technological development of AI and tries to explain how AI will translate into economic changes. Agrawal, Gans, & Goldfarb (2019) and Brynjolfsson, Rock, & Syverson (2019) have introduced the concept of AI as a GPT due to its widespread application possibilities and its general nature to make predictions across domains and fields. A GPT has the following characteristics (Jovanovic & Rousseau, 2005): (1) Pervasiveness – should be able to have an impact on most sectors. (2) Improvement – should become more capable over time and more affordable. (3) Innovation – should have a positive and accelerating impact on the invention of new products and processes.

The initial fear that AI will lead to job losses across many industries as discussed by Pannu (2015) has by now been countered by arguments that every new GPT technology results in a shift in the labor market (Agrawal et al., 2019). While certain positions will find a fast replacement by AI, others will gradually appear, often requiring a different set of skills than earlier positions. This will lead to a slight disruption of the labor market, but not in vast job losses as this was indicated in earlier studies (Agrawal et al., 2019). We can observe that – while there are still warning voices out there – most papers are optimistic about the current development of AI and point towards a bright future.

Overall, AI has emerged as new General-Purpose Technology (GPT) for decision making and will have huge impacts across all industries and sciences. It will foster data-driven decision

making by enabling completely automated end-to-end decision processes. It will also foster innovation and hence drive growth in general due to its nature to be sector overarching.

1.3 Data-Driven Decisions

The era of big data and artificial intelligence led to a change in corporate decision making. At the heart of those developments lies business analytics, a function that drives data-driven decision making by translating raw data into insights. Management has always used data to generate information and insights. Mainly with the help of enterprise resource planning (ERP) and business information systems (BIS). This is not new. What has changed is how we come up with a decision. The earlier, more intuitive decision-making approach from executives was replaced by evidence-based decision making based on advanced analytics and machine learning (Brynjolfsson & Mcelheran, 2019; Delen & Ram, 2018). This had an impact on information technology and IT Strategy, which increased in importance and gradually moved into boardrooms, employing new executive members as the “Chief Digital Officer”.

The number one reason for this cascade of events is the possibility of increased productivity for profit-driven entities in our economy. Data-driven decisions can capture value for corporations in our competitive environment and will gain importance over time. Brynjolfsson, Hitt, & Kim (2011) analyzed whether a focus on data and business analytics aka data-driven decision making (DDD) has a positive impact on corporate performance. Concrete, the authors analyzed 179 listed companies and concluded that the adoption of business analytics results in a 5-6% productivity gain. Further, it seems that the successful utilization of business analytics also impacts other measures as asset utilization, return on equity, and market value. In a more recent study, Brynjolfsson & Mcelheran (2019) finds that early adopters had the most advantages by adopting a DDD strategy. The type of analytics that has a positive impact on the bottom line shifts according to the authors during the study and the latest key driver was predictive analytics.

Grover, Chiang, Liang, & Zhang (2018) think the value proposition of business and data analytics for corporations is less clear but ultimately concluded that it can create value if certain other factors are in place as skilled labor and sound strategic positioning. The authors further argue that those things are crucial as the ability to capture value through analytics is heavily dependent on the required skill-sets that can leverage those analytics capabilities.

Overall, the realization that data-driven decisions and especially predictive analytics have the potential to drive performance by directly impacting the bottom-line started a new wave of

business analytics research. Business analytics and business intelligence constitute a quite long chain of different analytics, which includes retrospective as well as prospective analysis of relevant data. It usually starts from a point of analysis that is descriptive and moves gradually towards methods that offer deeper insights in the form of predictive and prescriptive analytics (Delen & Ram, 2018). Other names, which describe a similar subset of models and activities are data mining, business intelligence, data science, and operations research. The purpose of those functions is on a fundamental level the same. All combine the goal of converting raw data into actionable business insights by utilizing different analytics methods.

Business Analytics is vital in today's world shaped by digital disruption and global competition. Recognizing patterns based on historical data is one of the most useful skills for managerial decision making. Advanced analytics has become indispensable for today's corporate world and predictive analytics has been one of the major contributors to capture business value across all industries. Machine Learning algorithms help us to analyze massive amounts of data, including unstructured and nontraditional data like text and images. Nevertheless, the most economic value is coming from supervised learning on structured data (Ng, 2018). Corporations hold massive amounts of structured data in relational databases and many industries and/or business functions rely heavily on predictive modeling to derive valuable business insights from those data pools.

1.4 Scope of the Research

The objective of this Ph.D. thesis is to explore the strength and applicability of machine learning and business analytics for data-driven decision making to analyze how AI can capture value within financial services. It will be shown that machine learning can contribute to and enhance the performance of several departments across the financial services value chain. To test this hypothesis the currently best performing classifiers (state-of-the-art) are identified and will be applied in different business analytics settings in finance and insurance. The use cases in this thesis cover credit risk management (FinTech), claim assessment (InsurTech), and marketing and sales (Digital Marketing). The technical scope of the thesis is supervised binary classification on structured datasets. See figure 1 for a graphical illustration of the general context and scope of this thesis.

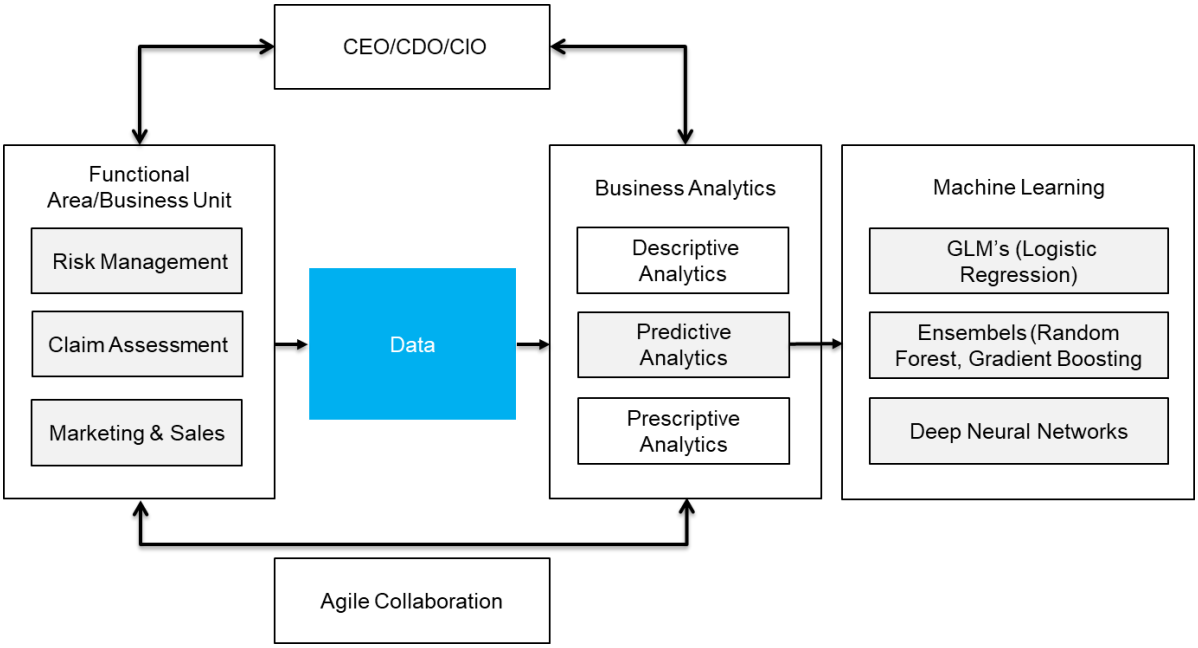

Figure 1. The business analytics function, which utilizes predictive analytics to steer data-driven decision making should be seamlessly integrated into the formal corporate structure using an agile approach to remain flexible to adapt to new demands from the different business functions as well as the management board. The scope of the thesis is supervised binary classification on structured datasets.

The following gaps were identified and will be addressed in this thesis.

Deep Learning vs Gradient Boosting: The realization that there exist different opinions about the performance of tree-based ensembles and deep learning will be further investigated in chapter 3 "Deep Learning vs Gradient Boosting". **Research Question:** Which ML model is the best performing binary classifier for structured data in the context of credit scoring?

Deep Learning in Business Analytics: According to the literature review are AI and DL use cases, especially when it comes to business analytics functions across the value chain still in its infancy. This is surprising given the astonishing media attention and perceived capabilities of deep learning in academia and industry. This topic will be further investigated in chapter 4 "Deep Learning in Business Analytics". The use case focus here is extended and instead of a pure focus on credit risk, two additional uses cases will be taken into account. Insurance claims prediction and marketing and sales. **Research Question:** What is or are the concrete reasons for the slow adoption rate of deep learning in the context of business analytics? Does DL offer advantages over traditional ML models as tree-based ensembles and GLM's for predictive analytics tasks in business analytics? What are the managerial implications of those findings in the context of digital transformation and digital strategy?

Super Learning in FinTech and Credit Risk: FinTech and technology-driven financial markets require strong and accurate predictive analytics. Chapter 5 “Super Learning in FinTech and Credit Risk” will have a look at the current state-of-the-art models for predictive analytics and introduces a way to improve upon those best-performing classifiers. **Research Question:** Is it possible – in the context of credit risk and lending – to consistently improve upon the best binary ML classifiers as identified in the earlier experiments (chapter 3) by utilizing the fusion method stacking? If yes, can a recommendation w.r.t to model configuration be derived that is generally applicable? What about the practical limitations of stacking?

AutoML in Business Analytics: The importance of data-driven / evidenced-based decision making is indispensable in today’s global and competitive market place. The talent gap when it comes to analytics functions is still prevalent and seems to grow over the coming years. A way to solve the skillset shortage and also to help experts with faster prototyping are automated machine learning solutions that promise to completely automate the predictive modeling process. Automated machine learning or AutoML will be further investigated in chapter 6 “AutoML in Business Analytics. **Research Questions:** What are the current possibilities of AutoML and what is the predictive strength of AutoML compared to manually tuned classifiers by a human expert? What are the future implications of AutoML for Business Analytics based on those findings? What further research is necessary to reach a full end-to-end decision engine that can serve as a complete off-the-shelf ML model for business analytics use cases?

Enterprise AI: Towards an End-to-End Data-Driven Decision Engine: The major goal of this chapter is to synthesize the contributions of the earlier chapters into a coherent whole by discussing the status quo of AutoML in light of those earlier findings. **Research Questions:** Is it possible to create an automated end-to-end predictive analytics process that is on par with a manually tuned system? What are the current limitations and gaps regarding such a prediction engine? What are the necessary further research directions required to reach a complete end-to-end decision engine for business analytics?

Stakeholder Implications: This last chapter will draw upon the findings of the earlier content-analysis which spans across all the papers and also on the results of the experimental studies to discuss implications for different stakeholders. The major focus is on the three areas, which were present throughout the thesis – namely credit risk management, insurance claims, and sales and marketing in a financial services context and how machine learning can enhance the performance in those business units. This chapter clearly explains how those models can be applied and where exactly the value contribution can be found, also by leveraging synergies

between different departments. There will also be a discussion on AI as a general-purpose technology (GPT) and how the field will evolve in the future.

A focused introduction, literature review, and gap analysis can be found at the beginning of each chapter. See figure 2 for an overview of the ML methods covered in each part.

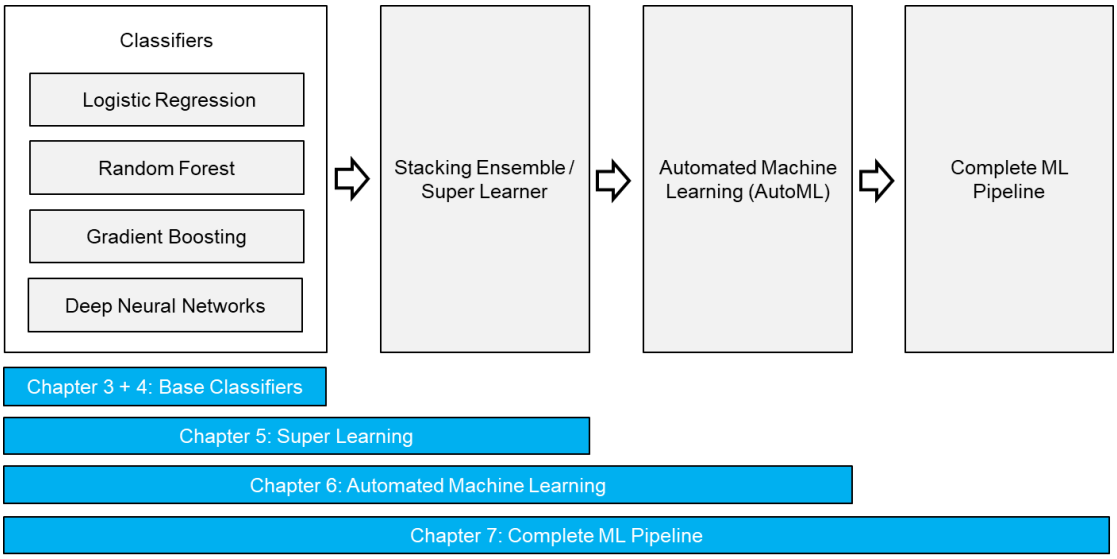

Figure 2. The flow of the thesis gradually evolves using more complex ML methods. Chapters 3 and 4 use exclusively single classifiers and standard ensembles as LR, RF, GBM, and DL. Chapter 5 extends the portfolio with the ensemble method stacking (super-learning), chapter 6 is concerned with the capabilities of AutoML in comparisons to manually tuned models, and chapter 7 combines the findings of the earlier chapter and proposes a fully automated ML pipeline.

The chapters within this thesis are largely individual papers (Chapter 3-6) and can be read as such. The introduction of each chapter contains a focused literature review that is directly relevant to the following analysis and discussion, which increases readability and reduces unnecessary jumping around between chapters. The last two chapters before the conclusion are somewhat unique. Both take the form of synthesis and a slight extension. Chapter 7 focuses more on the technical aspects and discusses the earlier findings in the context of a complete end-to-end ML pipeline for business analytics. Chapter 8 before the conclusion is business focused and will discuss stakeholder implications and how business functions can leverage ML to achieve direct value gains, including a discussion on the economic long-term implications of AI. The conclusion chapter will give a summary of all the findings.

Over the next decade, we have an amazing opportunity to build an AI-powered society and I am glad to contribute slightly to this new technological movement. I genuinely hope you will enjoy reading this thesis.

1.5 Contributions

Research question:

Is it possible to capture value through machine learning applications across the financial services value chain?

This central question was answered through more granular questions as:

- What is the best ML model for credit scoring?
- What are the reasons for the slow adoption of DL in BA?
- Is it possible to consistently improve upon the best ML classifiers by utilizing the fusion method stacking?
- Does AutoML reach the predictive strength of manually tuned classifiers by a human expert?

Contributions:

- Gradient Boosting Machine does outperform Deep Learning in binary classification on structured data in the context of credit scoring.
- In addition to the usual arguments as computational complexity, transparency issues, lacking infrastructure, and skill-shortage, tree-based ensembles as Random Forest and Gradient Boosting Machine beat DL on structured datasets, which offers a logical explanation of why DL adoption is lacking behind expectations.
- It is possible to consistently improve upon the best Machine Learning classifiers by utilizing the fusion method stacking, which trains a so-called super learner by combining several candidate models into one single and more powerful classifier.
- The H2O AutoML framework does not reach the accuracy levels of a manually tuned ML model, but the performance difference is only marginal and AutoML proved to be a powerful method for fast prototyping. It also has the potential to bridge the gap between the skill-set shortage in the field.
- A framework is presented on how to extent AutoML solutions towards a complete end-to-end ML pipeline for business analytics.

The following chapter summaries give a good first impression of what to expect in the following pages.

Chapter 3 - Deep Learning vs. Gradient Boosting:
Benchmarking state-of-the-art classification algorithms for credit scoring

Credit Risk Management is an essential part of financial institutions. In light of the changing lending market structure, advanced analytics has become vital to remain competitive in our fast-paced business world shaped by global competition. The two models currently competing for the pole position are Deep Learning and Gradient Boosting Machines. This paper will benchmark those two algorithms in the context of credit scoring using three distinct datasets with different features to account for the reality that model choice/power is often dependent on the underlying characteristics of the dataset. This study has shown that – for structured datasets – GBM tends to be slightly more powerful than DL and also has the advantage of lower computational requirements. This makes GBM the winner and choice for most problems within credit risk. But it was also shown that the outperformance of GBM is not always guaranteed and ultimately the concrete problem scenario or dataset will determine the final model choice.

Chapter 4 - Deep Learning for Business Analytics:
A Clash of Expectations and Reality

Our fast-paced digital economy shaped by global competition requires increased data-driven decision making based on advanced analytics and Machine Learning. The first wave of digital transformation based on big data and analytics is now gradually replaced by AI, which becomes the driving force behind new digital transformation initiatives. The benefits of Deep Learning (DL) over traditional analytics are manifold, but it comes with limitations that have – so far – interfered with widespread industry adoption. This paper conveys an accurate picture of the current deployment of DL in business analytics. The paper contains three case studies of different business use cases and benchmarks DL against traditional machine learning models. The adoption of Deep Learning is not only affected by computational complexity, lacking big-data architecture, lack of transparency (black-box), and skill shortage, but also by the fact that DL does not outperform traditional ML models in case of structured datasets with fixed-length feature vectors as usually present in relational database systems. It is shown that

DL does not show superior performance for classification problems on structured data across several domains. DL does not achieve higher performance as Gradient Boosting Machine and Random Forest. These results are consistent across all three use cases presented in this study, which offers a logical explanation of why DL adoption is lacking behind expectations. DL should be regarded as a powerful addition to the existing body of ML models instead of a one fits it all solution.

Chapter 5 - Super Learning in FinTech and Credit Risk:
In search of maximum prediction accuracy

Artificial intelligence and machine learning are gradually changing the lending market structure towards full automation. Advanced predictive analytics helped FinTech firms to develop modern lending businesses that foster financial inclusion due to high prediction accuracy, which opens the possibility to disregard collateral as a safety net. This is a big step towards the democratization of debt markets. In search of maximum prediction accuracy, this paper is going to train different configurations of a stacked ensemble model that combine the most powerful baseline models into a so-called super learner. Thereby proving that super learning can improve upon the performance of even the best models currently available. In addition, the observed outcomes were used to derive concrete configuration steps that are generalizable to reach the highest possible prediction accuracy. The four models used as a baseline in this experiment are Logistic Regression, Random Forest, Gradient Boosting Machine, and Deep Learning. The experiment was implemented on three real-world-datasets from the credit risk domain. Also, this experiment is placed in a discussion on financial inclusion and the future of FinTech to convey the importance of ML and AI applications for financial services.

Chapter 6 - AutoML in Business Analytics:
Towards a fully automated predictive analytics process

The realization that data-driven decision-making is indispensable in today's fast-paced and ultra-competitive marketplace has raised interest in industrial machine learning (ML) applications significantly. The current demand for analytics experts vastly exceeds the supply. One solution to this problem is to increase the user-friendliness of ML frameworks to make them more accessible for the non-expert. Automated machine learning (AutoML) is an attempt to solve the problem of expertise by providing fully automated off the shelf solutions for model

choice and hyperparameter tuning. This paper analyzes the potential of AutoML for applications within business analytics, which could help to increase the adoption rate of ML across all business functions. The H2O AutoML framework was benchmarked against a manually tuned model on three real-world datasets to test its performance, robustness, and reliability. The used AutoML framework trains several base learners and combines them via ensemble learning to a stacked super learner. The manually tuned model could reach a performance advantage on all three case studies used in the experiment. Nevertheless, the H2O AutoML package proved to be quite potent. It is fast, easy to use, and delivers reliable results, which come close to a professionally tuned ML model. The experiment proved that the H2O AutoML framework in its current capacity is a valuable tool to support fast prototyping with the potential to shorten development and deployment cycles. It can also bridge the existing gap between supply and demand for ML experts and is a big step towards fully automated decisions for business analytics functions.

Chapter 7 - Enterprise AI:
Towards an end-to-end ML pipeline for Business Analytics

An end-to-end business analytics engine is essentially a comprehensive and automated ML pipeline. The major goal of this chapter is to synthesize the contributions of the preceding chapters into a coherent whole by proposing a complete ML-pipeline that consists of three distinct phases: Phase 1 - Data Preparation; Phase 2 - Model Tuning and Evaluation; and Phase 3 - Model Deployment and Monitoring. AutoML automates the second phase in the pipeline (model tuning and evaluation) and is a vital building block to automate the full pipeline. It is discussed how AutoML can be improved to reach state-of-the-art accuracy levels to fulfill its purpose as the heart of the pipeline. Alternatively, it can be used in its current form. However, to achieve an end-to-end prediction engine for data-driven decision-making extensions towards Phase 1 and 3 are required. Data preparation, which consists of several adjustments as cleaning and feature engineering are not yet automated. Also, there is no consideration of real-world constraints (size, speed, interpretability), and the model choice is purely based on prediction accuracy. Due to the lack of those functionalities, automated monitoring and adjustments are not possible. Those gaps result in clear future research directions which are also discussed in this chapter.

2 Research Methodology

The central idea of the thesis is to analyze the value contribution of modern AI for business analytics and financial services. Is it possible to capture value through machine learning applications across the financial services value chain? The scope of the research is binary classification on structured data. Structured data means tabular data with categorical and numerical variables. To answer this and all the more granular research questions defined in the preceding section a mix of qualitative and quantitative research was chosen as the most suitable approach.

2.1 Qualitative: Content Analysis

The qualitative research approach is based on a literature review and content analysis to identify the current status quo of machine learning in business analytics, including a general understanding of use cases, advantages, problems, and adoption speed of advanced analytics in business and economics. The literature review and content analysis took into account the following materials/sources:

- Research papers from relevant scientific journals across all relevant sciences
- Industry research and white-papers with a focus on world-leading consulting companies as McKinsey & Company, Boston Consulting Group, Deloitte, KPMG, Accenture, etc.
- Reports from official organizations as International Monetary Fund (IMF), European Central Bank (ECB), Bank for International Settlement (BIS), etc.
- Official Announcements and Agendas of Government Entities

Due to the broad background of the topic, it was challenging to limit the relevant search terms utilized during the initial screening phase. The search terms used during the literature search were the following:

- Search Terms for Analytics: Artificial Intelligence, Machine Learning, Digital Transformation, Digital Strategy, Business Analytics, Binary Classification, Predictive Analytics, Black-Box
- Search Terms for Application Areas: Credit Scoring, Credit Risk Management, Lending, FinTech, InsurTech, Insurance, Finance, Digital Marketing, Marketing

Due to the fast development of the field, the focus of the analysis was only on sources going back to 2017. Older papers were included when identified as relevant during the content analysis.

2.2 Quantitative: Experimental Study

Once those state-of-the-art methods were identified an empirical study to verify the strength in comparisons to alternative methods was carried out. This was especially relevant as the conclusions regarding the concrete model choice differ based on the current literature and there is no consensus regarding optimal model choice.

The idea of the quantitative experimental study is to produce generalizable knowledge about the application of state-of-the-art machine learning in business analytics. The study has been carefully designed to be representative of real-world business use cases in finance and insurance. The chosen datasets are all publicly available to facilitate reproducibility by other researchers. Internal validity is given, but external validity (generalization) is not clear. Many papers in this field only used 1 to 3 datasets to justify their findings. While this thesis contains 5 datasets, it might still not be enough to generalize the findings. A description of the datasets can be found in section 3.3.1 and 4.2.2. Overall, the chosen datasets contain different numbers of features and observations and are hence diverse enough to answer the above-formulated research questions.

The following section introduces the necessary building blocks of machine learning and predictive analytics, including the evaluation methods. A description of the datasets, data preparation steps, and the relevant software, can be found in the individual chapters. The same is true for a more detailed description of the individual ML algorithms.

2.3 Quantitative Research Methods

This part introduces the technicalities of machine learning and predictive analytics. Some parts of this chapter contain similarities to my Master's thesis (Schmitt, 2016). However, the content was largely cut, extended and/or re-written, which results in a significant improvement to better serve the purpose of this Ph.D. thesis. For a mathematically exhaustive treatment of ML/DL, it is referred to the deep learning bible from Ian Goodfellow, Yoshua Bengio (2016) and to the statistically motivated exploration of machine learning written by Hastie, Tibshirani, & Friedman (2009). Books that give a good foundation about standard machine learning concepts are Murphy (2012) and Bishop (2006) for general ML and Russel & Norvig (2009) for a more comprehensive view on AI. For applied predictive modeling it is referred to Kuhn & Johnson

(2013). And finally, for a chronological rundown of ML, I would recommend Schmidhuber (2015).

2.3.1 Machine Learning

Machine learning (ML) – whether deep or shallow – can be split into three major categories: Supervised learning; Unsupervised Learning; and Reinforcement Learning. The last one would be semi-supervised learning, which is essentially a hybrid of supervised and unsupervised ML.

Supervised learning (SL) requires labeled data to train a model, which will then be used to make predictions from new (unseen) data. In contrast, the idea of **unsupervised learning** is to gain information out of an unlabeled data-set. Also known as knowledge discovery. Examples are outlier detection and clustering problems (e.g. anomaly detection with autoencoders). **Reinforcement learning** (RL) is the machine-learning method that is closest to real artificial intelligence and mainly responsible (in combination with Deep Learning) for the hype and media attention. RL uses an intelligent agent which tries to optimize its decisions to get the highest reward, which consequently maximizes its value. The idea of a Markov-decision process and dynamic programming is used to optimize a value function at each decision step. RL is still in its infancy and there is still a lot to be done to build a fully autonomous agent (Sutton & Barto, 2017). Standard machine learning algorithms – once understood – are very straightforward in their application. RL instead is a complete system that uses several ML algorithms. RL breaks a complex problem into smaller problems that can be solved by either supervised or unsupervised learning. Only a few applications of RL in business do exist and the field is largely in its infancy.

This thesis is mainly concerned with supervised learning problems where we have preexisting data-sets, that have a sufficient amount of feature sets consisting of input X and output Y pairs. The most common method deployed in the industry is supervised learning. The major reason for this is that supervised ML is the most important method for real-world **predictive analytics** uses cases and is hence the main driver in the industry to capture value for corporations (Gary, 2018; Ng, 2018).

2.3.2 Predictive Analytics

The essential problem in supervised machine learning or predictive analytics is to find a model that predicts an output Y given an input X. This is done by creating an input-output mapping in the form of $f: \mathrm{X} \rightarrow \mathrm{Y}$. The idea is to fit a model that for each observation (feature set) of inputs $\mathrm{X}_i = 1, \ldots, n$ finds a corresponding output Y (Hastie, et al., 2009). Notations differ slightly due to the development in different scientific disciplines as classic statistics, pattern recognition and

computer science. See figure 3 for the different notations w.r.t. input and output variables (Hastie, et al., 2009).

Input (X_1,...,X_n)	Output (Y)
•independent variable •feature •attribute •predictor	•dependent variable •target •response variable •class

Figure 3. Due to the development in different fields as classic statistics, computer science and pattern recognition there exists a variety of names for the input and output variables. This table lists all the names which are frequently used in the machine learning literature.

The type of output depends on the problem type. Within the domain of **supervised learning** regression and classification models can be distinguished. **Classification** problems have discrete outputs, where $\mathrm{Y} \in \{0, \ldots, \mathrm{n}\}$, which are categorical. **Regression** models have real (continuous) outputs, where $\mathrm{Y} \in \mathrm{R}$, which are numerical. Input features could be anything from detailed customer data to predict the default probability of loan applicants (classification) to closing prices of a stock market index to predict future price movements (regression). In general, a predictive model takes the form of

$$\hat{Y} \approx f\,(\mathrm{X}; \mathrm{W}) \tag{1}$$

with the additional parameters W, which are often referred to as weights. The primary goal is to find the model with the best predictive power for a given data set.

Examples of ML models used within business analytics for prediction tasks are GLMs (Bertsimas & King, 2016) as Logistic Regression; Classification and Regression Trees (CART); their ensembles as Random Forest and Gradient Boosting; and more recently (Deep) Neural Networks (Kraus, Feuerriegel, & Oztekin, 2019).

The rationale behind this is the following: **Logistic regression** is widely used in the industry and serves as a good general baseline for binary classification problems. Decision Trees are an all-time favorite, easy and explainable, but lack predictive power compared to their more complex ensembles as Random Forest and Gradient Boosting. **Random Forest** increases upon the accuracy of a single decision tree but remains a sufficiently simple model that is easy to use and delivers good results. **Gradient Boosting** is currently dominating the benchmarks for structured data sets and generally seen as one of the most powerful ML models currently

available. Finally, the relatively new **Deep Learning** offers more flexibility due to its strong performance on unstructured data sets. A more detailed description of the relevant ML methods can be found in the corresponding chapters 3 to 6.

To assess the predictive ability of machine learning models a method to quantify the deviation of the predicted value $\hat{Y} \approx f\,(\mathrm{X};\mathrm{W})$ to the true observation value, Y is required. This is done with a so-called **loss function**. The loss function

$$J(W) = \sum_{i=1}^{n} \mathcal{L}(f(X_i;\mathrm{W});Y_i) \tag{2}$$

is also frequently referred to as cost or objective function, where n is the number of samples and $f(X_i;\mathrm{W})$ the estimation of Y_i. The objective is to choose the parameters W (weights of coefficients) to minimize the loss/error. The loss function is the primary measure of predictive accuracy.

As the predictive model adjusts the weights to determine the best parametrization to reduce the error produced by an objective/loss function, the whole idea of machine learning and predictive modeling, therefore, collapses to a single objective, which results in an **optimization problem** (Bertsimas & Kallus, 2019) in the form of

$$w^* = \underset{w \in W}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i;W)\; y_i) \tag{3}$$

The concrete choice of the loss function is motivated by the problem scenario/prediction task. The most common choices to train neural networks are (1) the **mean squared error** loss for regression problems where the outputs are continuous real numbers.

$$J(W) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i;W)\big)^2 \tag{4}$$

And (2) **cross-entropy** for classification problems where the output is discrete. The cost function for the logistic regression (cross-entropy) is different from the mean squared error for linear regression and represented by

$$J(W) = \frac{1}{n} \sum_{i=1}^{n} C(h_w(x), y), \tag{5}$$

Where

$$C(h_w(x), y) = \begin{cases} -\log\big(h_w(x)\big), & if\ \ y = 1 \\ -\log(1 - h_w(x)), & if\ \ y = 0 \end{cases}. \tag{6}$$

$As\ y\ \in \{0,1\}$, this can also be written as

$$J(W) = -\sum_{i=1}^{N} [y_i \log\big(h_w(x)\big) + (1 - y_i)(\log\big(1 - h_w(x)\big)]. \tag{7}$$

This is the negative log-likelihood function, also referred to as the cross-entropy error function. Maximizing a likelihood function is equivalent to minimizing a loss function which is the reason for using a negative likelihood (Murphy, 2012).

In summary, to fit a predictive model we have to find the optimal weights W (parameterization) to minimize the cost function $J(W)$. In practice, this is done by using an optimization algorithm called **gradient descent** which can be written as

Repeat {

$$W_i = W_i - \alpha \frac{\partial}{\partial W} J(\mathrm{W}) \tag{8}$$

}.

The parameter α is the learning rate. To minimize the cost function $J(W)$ the partial derivative of $J(W)$ with respect to each weight W needs to be calculated. It then moves in the opposite direction to the gradient as the objective is to minimize the error (Raschka, 2015). The algorithm repeats its iterations and updates the parameters until it reaches a local minimum. Choosing α is not an easy task. A small learning rate will result in a slow convergence as the algorithm only takes small steps. On the other hand, a very large learning rate might not converge at all as it will overshoot the minimum (Murphy, 2012). The learning rate is one of the most important hyperparameters for optimizing a machine learning model.

Several ML algorithms as logistic regression, gradient boosting, and neural networks use the gradient descent algorithm to find the parametrization that minimizes the error of the objective function. Note that this method does not work with the random forest as this model is based on discrete instead of continuous outputs.

Once trained, the predictive power or strength of a machine learning model is measured with the help of an evaluation metric. Different types of those metrics exist and are explained in more detail in chapter 2.5.

2.3.3 Modeling Process

The workflow of predictive modeling can be seen in figure 4. It starts with the preparation, exploration, and cleaning of the data set. This includes the elimination of outliers, missing values, and feature scaling. Feature transformation requires a full understanding of the dataset and prediction purpose and usually requires significant domain expertise. If specific features are not particularly relevant for the response or redundant it is possible to merge those features or eliminate them.

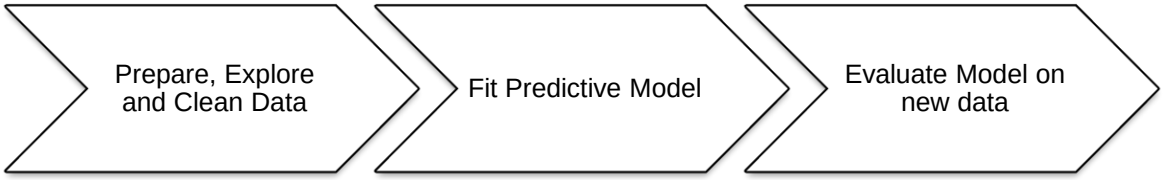

Figure 4. The predictive modeling process usually starts with the preparation, exploration, and cleaning of the data set. In the second step, the prepared input features are used to train the model. After that, the model will be used on new data to evaluate its accuracy.

The second and third step would be the model training and evaluation. Initially, the data set is split into a **training set** and a **test set**. The training set is used to fit the predictive model. The model fitting process contains mainly the optimization of the different hyperparameters to reach the best model configuration for the underlying dataset. The training set might be further separated into different parts (training- and validation sets) with a frequently used method called cross-validation which helps the model to generalize better to unseen data. After the model has been trained the predictive accuracy can be tested on the previously hold-back data set.

The necessity for excessive pre-processing and careful model tuning goes against the notion of an off-the-shelf solution, which is often favored in real-world business analytics and data mining scenarios. An **off-the-shelf method** is characterized by immediate applicability to a problem at hand. Off-the-shelf prediction models can be directly applied to a problem scenario without any significant domain knowledge or pre-processing of the datasets. Automated machine learning solutions (AutoML) are one way to streamline the predictive analytics workflow and will be discussed in chapter 6. Other issues that interfere with the off-the-shelf

property of ML algorithms are lack of interpretability (black-box) and computational complexity (speed), which will further be discussed in chapter 4.

2.3.4 Learning Theory

The problem of **overfitting** is often referred to as a **bias-variance tradeoff**, which means that a model needs to learn the representation of the training set, but just enough to generalize well to new (unseen) data. If a model overfits (high validation error) during the training process and turns out to be a pure representation of the training set (high variance), it will be useless for predictive tasks on data it has not seen before. If the model instead is kept as simple as possible it has low variance but a strong bias. High bias indicates a weak representation of the data, which means the model does not fit the training set very well and is subject to underfitting. It is called tradeoff as the optimum between bias and variance has to be found to produce a model with high prediction accuracy.

See figure 5 for a graphical illustration and a good explanation of the connection between model complexity and prediction error.

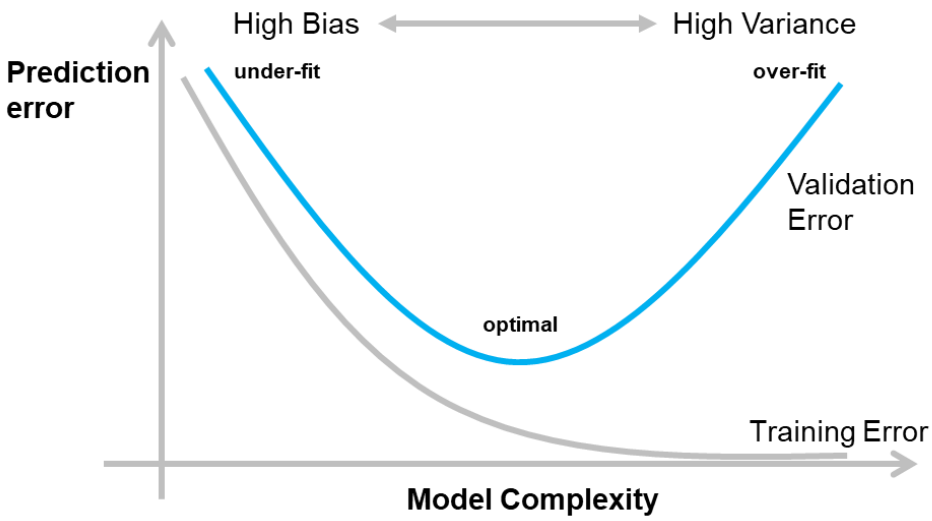

Figure 5 The Bias-Variance trade-off refers to the problem of overfitting. As model complexity increases the training error approaches zero but increasing variance after some optimal point will result in a loss of accuracy for predictions on unseen data sets. Over-fitted models do not generalize well to new data sets. K-fold cross-validation is a good method to tackle the problem of overfitting.

To solve this problem datasets are split into 3 parts. A so-called training set, validation set, and test set. Now the parameters of the learning algorithm (predictive model) can be optimized with the training data and the validation set is used to get the test error. This allows an estimation

of the **generalization ability** of the model on a new currently unknown data, which the algorithm or ML model has not yet seen.

One of the most powerful and widely used methods to prevent model overfitting to improve the generalization ability of the predictive model to unknown data is **K-fold cross-validation**. As can be seen in figure 6 the data set is simply divided into K equally sized parts, one of the K parts are chosen as a validation set, indicated here by the blue parts, and all the other K-1 parts are used as a combined training set to fit the model. This is done for all the parts to receive K different results which are combined in the end to get a better overall predictor that generalizes well (Bishop, 2006).

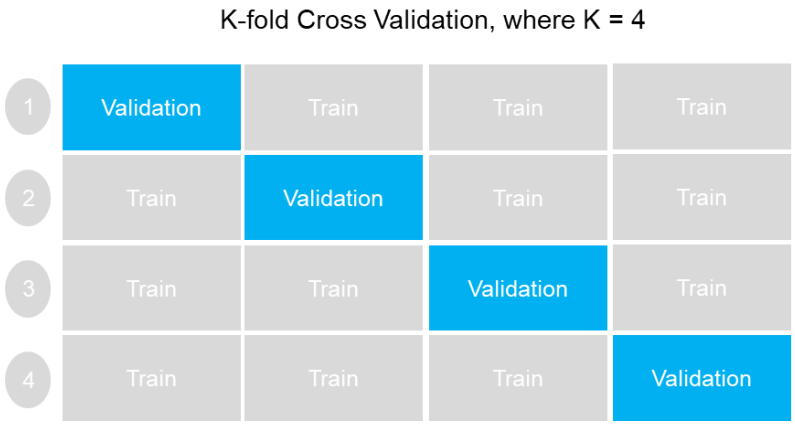

Figure 6 K-fold cross-validation divides the data set into K equally sized parts, chooses one of the K parts (blue) as the test set and uses all the other K-1 (grey) as a combined test set. This is done for all the parts to receive K different results which are combined to get a predictor that generalizes well to unseen data.

2.3.5 Evaluation Methods

To determine the predictive power of an ML model, an evaluation method is required. The performance matrices used in this thesis are Accuracy, AUC, F-score, and LogLoss. Those performance measures (except LogLoss) are based on a concept called confusion matrix which is a contingency table and frequently used in binary classification problems. See figure 7.

		Prediction	
		Good	Bad
Actual	Good	True Positive (TP)	False Negative (FN)
	Bad	False Positive (FP)	True Negative (TN)

Figure 7. A confusion matrix is a basic ingredient for the ROC Curve. It shows the connection between true positives and negatives and false positives and negatives.

True positives (TP) represent good observations that were classified as such, whereas false positives (FP) are observations that were incorrectly classified as good. The same logic applies to true negatives (TN), which represent the predicted values correctly classified as bad, whereas the false negatives (FN) represent observations incorrectly classified as bad.

Accuracy: The model accuracy is the ratio of correctly classified outcomes over the total number of observations and is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

F-score: The F-score is the harmonic average of recall and precision. The range of the F-score lies between 0 and 1. The best performance is reached at 1 indicating perfect precision and recall.

$$Fscore = \frac{2 \times Precision \times \mathrm{Re}call}{Precision + Recall} \tag{10}$$

Where precision is defined as the ratio of true positives over the sum of true positives and false positives.

$$Precision\ (FPR) = \frac{TP}{TP + FP} \tag{11}$$

And recall is defined as the ratio of true positives over the sum of true positives and false negatives.

$$Recall\ (TPR) = \frac{TP}{TP + FN} \tag{12}$$

AUC: In addition, the area under the receiver operating characteristics (ROC) curve (AUC) is used to measure the performance of the classifiers in this study. The closer the AUC comes to 1 the stronger and more accurate is the model. The AUC as a comparison measure is only valid when the underlying distribution is uniform, which means the outcome of each class is equally likely (Flach, Hernández-Orallo, & Ferri, 2011).

As can be seen in figure 8, the FPR represents the x-axis and the TPR represents the y-axis of the ROC plot. A perfect model would have a TPR of 1 and an FPR of 0.

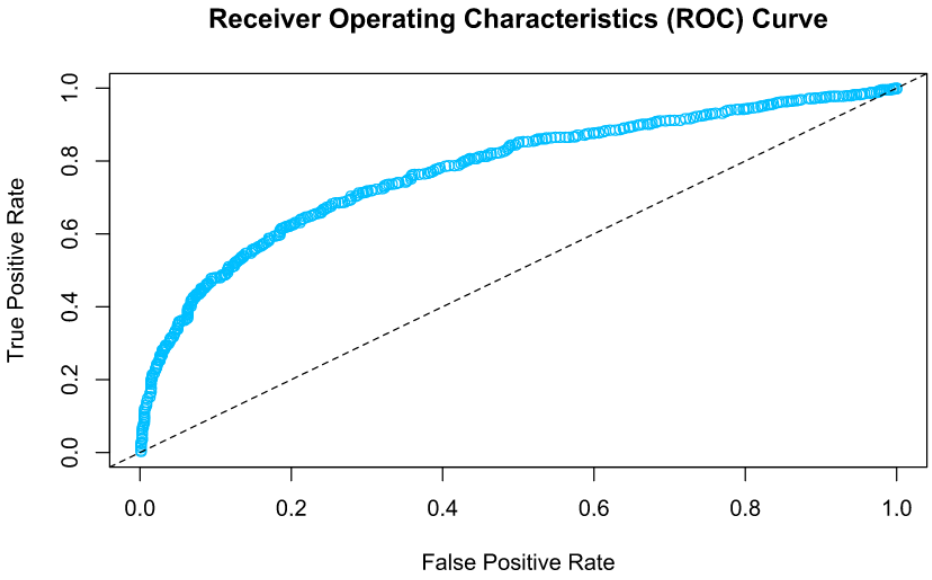

Figure 8. The AUC of the ROC Curve is an accuracy measure for classification problems and will help to assess the predictive power of the classifiers.

The models first predict the probability of default for every observation (customer) and based on a decision boundary assign a classification to the customer of either good (1) or bad (0). The default decision boundary for binary classification problems as in this case is 50% and is represented by a diagonal line from (0,0) to (1,1) of the ROC plot. Hence if we received an AUC of 50% or below our model would be purely random and therefore worthless.

LogLoss: The last metric used is LogLoss. It is also referred to as the cross-entropy error function and is widely used as the objective function when dealing with binary classification problems. But it can also be used as a performance measure. See equation 13.

$$\text{LogLoss}\ (y_i, h_i) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(h_i) + (1 - y_i)(\log(1 - h_i)]. \tag{13}$$

The idea is to penalize bad predictions with a significant loss and reward perfect predictions with a loss of zero. A rejection that is positive by a huge margin difference gets severely punished by this evaluation metrics. The LogLoss is used to determine the reliability of the performance indicated by the other three performance measures. A very high LogLoss is a bad sign and indicates that some predictions are significantly out of line.

2.4 Software

Data preparation and handling are managed in RStudio, which is the integrated development environment (IDE) for the statistical programming language R (R Core Team, 2019). R is one of the go-to languages for Data Science research as well as prototyping in practice. The machine learning models in this paper are developed with H2O, which is an open-source machine learning platform written in Java and supports a wide range of predictive models (LeDell & Gill, 2019). This makes experimentation and research easier. The high abstraction level allows the idea and the data to become the central part of the problem and helps to reduces the effort required to reach a solution. Also, H2O has the advantage of speed as it allows us to move from a desktop- or notebook-based environment to a large-scale environment. This increases performance and makes it easier to handle large data sets. R is connected to H2O by means of a REST API (Aiello, et al., 2016).

3 Deep Learning vs. Gradient Boosting

Abstract

Credit Risk Management is an essential part of financial institutions. In light of the changing lending market structure, advanced analytics has become vital to remain competitive in our fast-paced business world shaped by global competition. The two models currently competing for the pole position are Deep Learning and Gradient Boosting Machines. This paper will benchmark those two algorithms in the context of credit scoring using three distinct datasets with different features to account for the reality that model choice/power is often dependent on the underlying characteristics of the dataset. This study has shown that – for structured datasets – GBM tends to be more powerful than DL and also has the advantage of lower computational requirements. This makes GBM the winner and choice for most problems within credit risk. But it was also shown that the outperformance of GBM is not always guaranteed and ultimately the concrete problem scenario or dataset will determine the final model choice.

Keywords: Credit Scoring, Classification, Deep Learning, Gradient Boosting

3.1 Introduction

The risk management unit is one of the most important business analytics functions within financial institutions. Two important metrics within credit risk management are the probability of default (PD) and the loss given default (LGD). The PD assesses the likelihood of a borrower not willing or able to repay the loan, while the LGD measures the exact loss (outstanding balance – collateral) that would occur in the event of default.

Several developments in today's society have led to a change in the market structure for lending businesses (Claessens, Zhu, Frost, & Turner, 2018), which led to the increased importance of prediction accuracy for credit assessments instead of only relying on post-mortem protection in the form of LGD reduction through collateral. A pure focus on prediction accuracy has the advantage of cash-flow based lending without collateral requirements from borrowers (Frost, Gambacorta, Huang, Shin, & Zbinden, 2019). This can help to foster financial inclusion and is especially important for consumers in developing countries and small to medium-sized corporations, which have no collateral and had traditionally limited access to debt capital (Bazarbash, 2019).

Overall, it would be of significant value for P2P lenders, FinTech's, as well as traditional banks if they could increase the accuracy of the applicant's default probability to assign a correct

credit score during the application process. The better a financial institution/lender can predict the default probability/credit score of certain applicants the better they can shield themselves from potential costly credit losses. Besides, misclassification also results in missed revenue if a potentially good customer is wrongly assumed to be of high credit risk. Machine Learning (Jordan & Mitchell, 2015) plays a fundamental role in achieving this goal as it counters the default problem at the origin which is the decision whether or not to take on a certain applicant.

Statistical and machine learning models have been applied for years in credit risk management. Logistic regression has been the standard method for binary classification and is widely used in financial institutions to assign risk classes to applicants and has served as the main benchmark for years due to its simple application and accurate enough forecasts (Kraus, 2014). ML has been proven to achieve superior performance against generalized linear models very early (Keramati & Yousefi, 2011) and several models as decision trees, random forest, gradient boosting, support vector machines, and neural networks (NN) have gained popularity and are increasingly used in practice (Bazarbash, 2019).

An analysis of the current literature reveals a clear trend showing that the models competing for the highest prediction accuracy are Gradient Boosting Machines (GBM) and Deep Neural Networks (DNN) aka Deep Learning (DL).

Hamori et. al (2018) conducted an interesting empirical study analyzing the performance of ensemble learning in comparison to deep learning. The authors test different NN as well as DL configurations by switching activation functions (ReLU, Tanh) and come after 100 test runs – which are averaged – to the conclusion that GBM does outperform DL as well as NN.

Those findings are in line with other papers as (Addo, Guegan, & Hassani, 2018), but there is also a body of literature that favors DL and comes to the conclusion that DNN or NN are superior in general (Kraus et al., 2019; Lessmann, Baesens, Seow, & Thomas, 2015). The literature seems to be conclusive and we can conclude that there are currently only two models that compete w.r.t to predictive accuracy in credit scoring: Deep Neural Networks and Gradient Boosting Machines.

Surprisingly, there is not a single study exclusively focusing on DL and GBM when it comes to credit scoring. It is usually a mix of several machine learning models and there is a clear gap of a direct comparison of GBM and DL for assessing the correct default category of a loan applicant. Also, most studies within the credit risk domain have relied on only one dataset to benchmark different algorithms w.r.t to prediction accuracy (Addo et al., 2018; Hamori et al.,

2018). Therefore, it might be that the correct model choice is dependent on the underlying dataset.

The goal of this study is a direct comparison of GBM and DL in terms of prediction accuracy to correctly classify the default risk of a customer. Three distinct datasets with different features will be used to account for the possibility that model choice/power is based on the characteristics of the underlying dataset. In doing so this study will shed light on the predictive power and usefulness of both models for credit risk management within the lending market.

The structure of this chapter is as follows. Section 2 “Theory and Methods” introduces the methods used for the following empirical study: Gradient Boosting Machine (GBM) and Deep Learning (DL). Section 3 “Experimental Design” introduces the datasets and the process of model tuning (hyperparameter optimization). Section 4 “Results and Discussion” analyses the findings and discusses their implications. The last section gives a conclusion and a future outlook.

3.2 Theory and methods

The two models compared in this study within the context of default categorization are Gradient Boosting Machines and Deep Learning and are introduced next.

3.2.1 Deep Learning

Recent advances in AI research have improved the capabilities of Artificial Neural Networks and the new paradigm Deep Learning was born (LeCun, Bengio, & Hinton, 2015). The three factors that helped DL to become mainstream are advances in data availability (Big Data), processing power (GPUs), and optimization algorithms (Goodfellow et al., 2016). One of the major advantages of DL is its ability to work with unstructured data-sets, which improved many tasks and brought breakthroughs in text, speech, image, video and audio processing (LeCun, et al., 2015). In 2014 the South Korean Go champion Lee Sedol was defeated by Deep Minds AlphaGo (Silver et al., 2016). This initial success of deep reinforcement learning was soon followed by AlphaGo Zero (Silver et al., 2017) and several other gaming-related multimedia appearances as StarCraft (Pang et al., 2019) and Dota 2 (Katona et al., 2019). Deep learning did not only help AI to increase its popularity, but it also has a wider range of possible applications and is seen as one of the most disruptive technologies since the inception of the internet itself (Goodfellow, et al., 2016). Soon, the business world picked up on those developments and DL was increasingly used to enhance existing business analytics functions, including credit risk management (Bughin et al., 2017; Chui et al., 2018)

Deep Learning comes with many architectures as feed-forward artificial neural networks (ANN), Convolutional neural networks (CNNs), as well as Recurrent Neural Networks (RNNs). The best architecture for transactional (tabular) data, which are not sequential – as in this study – is a multi-layer feedforward artificial neural network. Other, more complex architectures as RNNs do not possess any advantage in those cases (Candel & LeDell, 2019).

The architectural graph of a feed-forward neural network can be seen in figure 9. The first column represents the input features and is called the input layer. The last single neuron represents the output where the final activation function is applied to. The two layers in the middle are called hidden layers. In case the neural network has more than one hidden layer it is called a deep neural network. A deep learning model can consist of several hidden layers and is trained with stochastic gradient descent and backpropagation (Goodfellow et al., 2016).

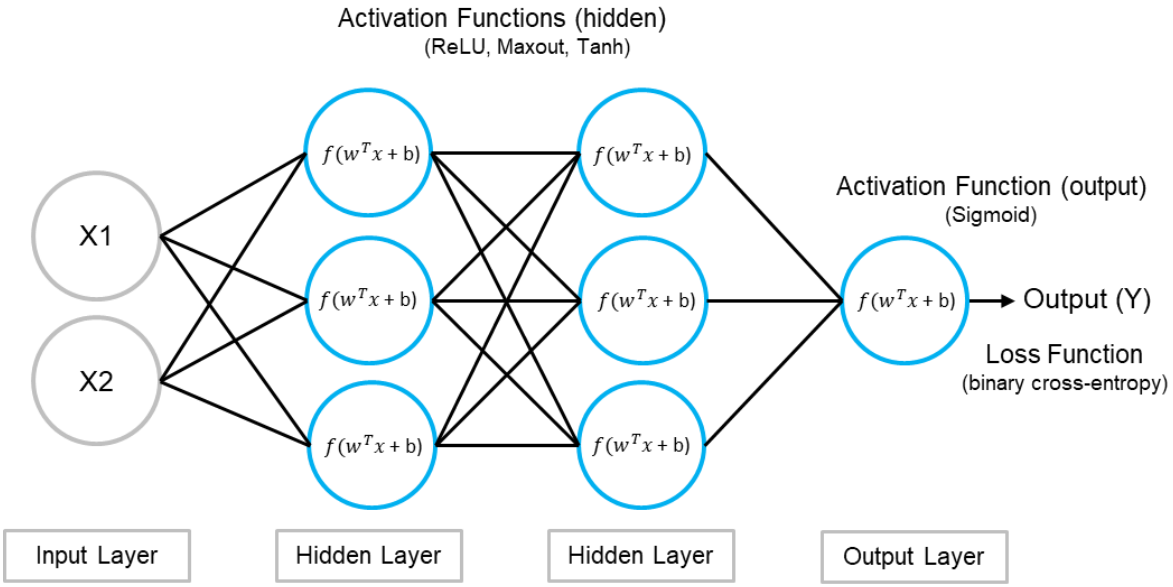

Figure 9. The deep learning model used in this experiment is called a feedforward artificial neural network as the signal flow through the network evolves only in a forward direction. It is the most appropriate choice for problems based on structured datasets as used in this study. It contains one input as well as one output layer and various hidden layers. At each node, a linear combination of input variables and weights are fed into an activation function to calculate a new set of values for the next layer.

A standard neural network operation consists of multiplying the input features by a weight matrix and applying a non-linearity (activation function). Input variables $X_i = (X_1, X_2, \dots, X_n)$ are fed into the neural network, weights $W_i = (W_1, W_2, \dots, W_n)$ are added to each of the inputs and a linear combination of $\sum X_i W_i = w^T x$ is calculated. This linear combination plus the bias term or interceptor serves as input for the activation function to calculate the output Y, which serves

either as input for the next layer or represents the final output/prediction. A neural network is trained with stochastic gradient descent and backpropagation.

Applying a non-linearity in the form of an activation function is essential for neural networks to be able to learn complex (non-linear) representations of the input data-sets. The activation function transforms the output at each node into a nonlinear function.

This study will build three different DL classifiers using the following activation functions for the hidden layers:

- the rectified linear unit (ReLU): $g(z) = \max(0, \mathrm{z}) \in [0, \infty)$,
- the hyperbolic tangent function (Tanh): $g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \in [-1,1]$, and the
- the Maxout function: $g(z) = \max\left(w^k z + b^k\right) \in (-\infty, \infty),\ k \in \{1, \dots, K\}$.

The activation function most widely used (at the time this thesis was written) is the rectified linear unit (ReLU). As developments in DL are quite fast, I recommend checking the best/most common approaches w.r.t architectures as well as the concrete choice of the activation function to solve different problems regularly.

As the scope of the research is binary classification on structured data the output activation function used is the sigmoid function $sigm(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{e^z+1} \in [0,1]$ in line with the binary cross-entropy loss function.

3.2.2 Gradient Boosting

GBM is state-of-the-art when it comes to accuracy, especially for supervised learning problems on structured data-sets (Ng, 2018). The first boosting algorithm – AdaBoost – was introduced by Freund and Schapire (1997). Four years later Friedman (2001) introduced the Gradient Boosting Machine, which is a more general form of the earlier algorithm due to the possibility to switch the loss function, which makes the AdaBoost algorithm essentially just a subset of the GBM introduced by Friedman (2001).

Boosting belongs together with bagging (Breiman, 1996a) and stacking (Caruana, Niculescu-Mizil, Crew, & Ksikes, 2004) to the family of ensemble learning techniques and builds models in sequential order. The goal of ensemble learning is to combine multiple ML algorithms to achieve better predictive performance. The specific idea of boosting is to start with a so-called weak learner – a model only slightly better than random guessing – that gradually improves by correcting the error of the previous model at each step. See figure 10 below.

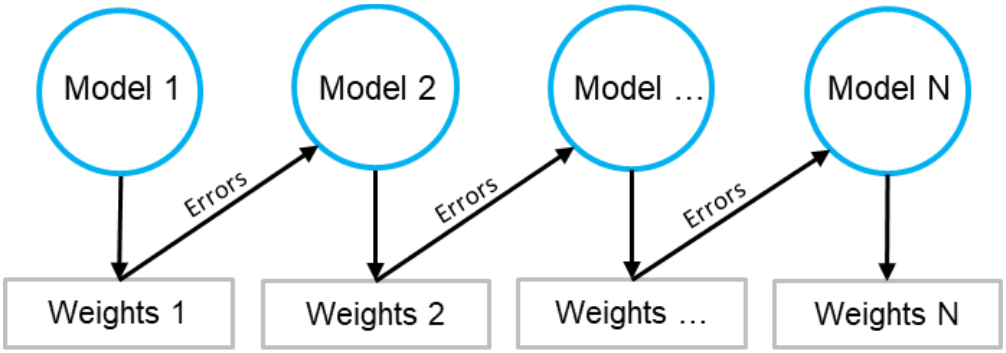

Figure 10. Gradient Boosting starts with a weak learner, typically a decision tree, and improves upon this initial learner iteratively at each step by correcting for the error of the previous model. GBM is one of the best performing ML models currently available.

The most common form of boosting uses decision trees and sequentially ads one tree at a time. This step by step adjustment forces the model to gradually improve the performance and leads to higher accuracy (Hastie, Tibshirani, & Friedman, 2017).

There are several different gradient boosting implementations out there. This study uses the gradient boosting version implemented by Malohlava and Candel (2019) which is based on Hastie et al. (2017).

3.3 Experimental Design

Financial institutions use credit risk models as scoring models to determine a client's default probability. These estimates help to decide whether or not a certain customer should receive a loan or a credit card. The primary objective of this study is to benchmark the above introduced predictive models (Gradient Boosting and Deep Learning) to correctly classify the default category of a customer. Due to the difficulty of determining the best activation function choice three different DL configurations with different activation functions (ReLU, Maxout, Tanh) are run. Only the activations functions in the hidden layers are switched. Due to the research scope - binary classification on structured datasets - the loss function (binary cross-entropy) and output activation function (sigmoid) remain the same. Three distinct datasets containing detailed client data are used.

3.3.1 Data and Preprocessing

The rationale for the used datasets is its relevance for credit risk management. One part of credit risk management is concerned with assessing the likelihood that a counterparty (e.g. loan applicant) will not be able to repay its obligation in part or in full. For this purpose, the

output of credit risk models is either the probability of default or a credit score, which can be binary or multi-class depending on the specific use-case. Credit scoring in this thesis refers to the binary classification of loan applicants, and either a “good” or “bad” label is assigned to the counterparty.

The datasets are from existing retail banks and have been widely used in earlier studies (Guo, He, & Huang, 2019; Hamori et al., 2018; Lessmann et al., 2015; Teng, He, Xiao, & Jiang, 2013; Yeh & Lien, 2009). The contained features resemble the typical information available to a retail bank and are therefore valid as real-world examples. See table 2 for a detailed description of the features contained in dataset 1 and 2.

All three datasets are similarly structured and the prediction goal is the same. In earlier studies was often only one dataset used to test the strength of specific classifiers. More than one dataset allows for easier generalization of the results. Also, the best performing classifier could change based on the underlying dataset itself.

The 3 datasets resemble real-world customer data, are all publicly available, and can be downloaded from the UCI Machine Learning Repositories, which makes the reproducibility of this empirical analysis possible. The datasets contain 23, 20, and 14 features respectively, which are historical client data and will serve as predictor variables to calculate the default category of each observation. All 3 datasets contain a target column that identifies whether or not the client defaulted. See table 1 for the details of each dataset including the resampling information.

Table 1. Description of Datasets

Dataset	Observations	Good	Bad	Balanced (Experimental Setup)	Features
Taiwan	30,000	23,364	6,636	6,636/6,636	23
Germany	1,000	700	300	300/300	20
Australia	690	307	383	307/307	14

The empirical study is based on 3 data sets, each containing several features (predictors) including a target column containing the default information (response). The datasets have been resampled and balanced (equal ratio of good and bad clients) to avoid the tendency for the AUC metric to favour the majority class.

Dataset 1 – Taiwan: The first dataset represents payment information from Taiwanese credit card clients. It was first used by Yeh and Lien (2009) and contains 30,000 observations where 6,636 are flagged as defaults. The dataset contains mainly historical payment information. Each observation (or feature set) contains 23 features including a binary response column for the default information of the credit cardholder.

Dataset 2 – Germany: The second dataset represents detailed customer-level data from a German bank and contains 1,000 observations where 300 are flagged as defaults. Each observation contains 20 features across a diver's range of categories including a binary response column which indicates whether or not a particular client defaulted on their loan payments.

Dataset 3 – Australia: The third dataset contains data for credit card applications for clients based in Australia. The dataset contains 690 observations where 383 are flagged as bad. Each observation contains 14 features including a binary response column indicating whether or not the person defaulted. The attribute names and values in this dataset have been changed to meaningless symbols due to confidentiality reasons.

See table 2 below for a more detailed description of the specific features contained dataset 1 and 2. No such description does exist for the third dataset due to confidentiality reasons.

Table 2. Detailed description of features contained in dataset 1 and 2

Dataset 1 - Taiwan		Dataset 2 - German	
Variable	**Description**	**Variable**	**Description**
X1	Amount of the given credit	X1	Balance of checking account
X2	Gender (1 = male; 2 = female)	X2	Duration in months
X3	Education*	X3	Credit history
X4	Marital status**	X4	For what was the loan taken
X5	Age (year)	X5	Credit amount
X6	Payment history September 2005	X6	Savings account plus bonds
X7	Payment history August 2005	X7	Duration of current employment
X8	Payment history July 2005	X8	Installment rate as % of income
X9	Payment history June 2005	X9	Marital status and gender
X10	Payment history May 2005	X10	Other debtors/guarantors
X11	Payment history April 2005	X11	Present residence since
X12	Amount of bill statement in Sep 2005	X12	Type of owned properties
X13	Amount of bill statement in Aug 2005	X13	Age of applicant
X14	Amount of bill statement in Jul 2005	X14	Housing (rent, own, free)
X15	Amount of bill statement in Jun 2005	X15	Credits at other banks
X16	Amount of bill statement in May 2005	X16	Existing credits at this bank
X17	Amount of bill statement in Apr 2005	X17	Employment/Level of qualification
X18	Amount paid September 2005	X18	The number of dependents
X19	Amount paid August 2005	X19	Registered telephone or none
X20	Amount paid July 2005	X20	Immigrant/foreign worker
X21	Amount paid June 2005		
X22	Amount paid May 2005		
X23	Amount paid April 2005		

* (1 = graduate school; 2 = university; 3 = high school; 4 = others)

** (1 = married; 2 = single; 3 = others)

The datasets were slightly adjusted to better serve the purpose of this study:

Random under-sampling was used to create a balanced data set for this classification study. Imbalanced datasets can result in a bias towards the majority class. The accuracy measures used in this paper – Area under the curve (AUC) – is more reliable when the model is trained with a balanced dataset. Predictive models that strive for maximum AUC tend to gravitate towards a classification that overrepresents the majority category which results automatically in higher prediction accuracy. If the dataset would have an imbalance of 80 to 20, the ML algorithm could always achieve an 80% accuracy without having true predictive power. To address this problem the class distribution (the ration between the two categories good and bad) was adjusted and brought to a state of equilibrium. This is done by under-sampling the majority class and by doing so create an equal ratio of good and bad observations (categories).

The subset of “good” observations (bad observations in case of the Australian dataset) was randomly drawn from the total population.

Another important step during the preprocessing of the data is to replace categorical values with a numerical representation. For example, dataset 1 contains only three numerical values which can directly be used to fit the machine learning models. The other features are categorical and had to be transformed to factor variables to be processed. This is often done by a method called one-hot encoding. One-hot encoding is widely used for classifying categorical data and transforms categorical labels into vectors of zeros and ones. The length of the resulting vectors is equal to the number of categories where each element within the vectors corresponds to one of those categories. This method potentially results in a significant increase in the feature set depending on the number of categories as well as the number of elements within each category. This was done for dataset 2 and 3. Dataset 1 is already clean and consists of only numerical values. Significant preprocessing is not required. One last step – that was done for all data-sets – was to change the response variable from a numeric representation to a binary factor which is necessary for a classification problem.

This study uses a training set, a validation set, and a test set. The first scenario which uses a training set size of 80% (80:10:10) is further referred to as 80:20 split while the second scenario using a training set size of 70% (70:15:15) is referred to as 70:30 split.

3.3.2 Hyperparameter Settings

Machine learning is an empirical process that involves trial, error, and experimentation. Hyperparameter optimization or model tuning describers the process of finding the optimal combination of hyperparameter for a machine learning algorithm. It is a multidimensional optimization problem and becomes more computationally demanding with an increasing number of parameters.

All models were carefully tuned to reach a performance that is adequate for the comparison in this study. In the case of DL, three different models are trained each containing a different activation function in the hidden layers (ReLU, Tanh, and Maxout) while holding all other parameters constant. The H2O framework allows the user to choose the activation functions (ReLU, Maxout, Tanh) for the hidden layers, the appropriate loss function, and the response type.

The H2O deep learning framework uses the above-specified activation function throughout the network (hidden layers) and based on the response column (binary) and loss function choice

(cross-entropy) determines the appropriate activation function (in this case sigmoid) for the final layer.

Dropout has been shown to improve accuracy (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014); hence all 3 DL models use a dropout ratio of 0.50. The concrete hyperparameter settings for each dataset and model in the case of the 80:20 split can be found in table 3.

Table 3. Hyperparameter setting of GBM and DL for the 80:20 split

Dataset	GBM - Parameters	Value	DL - Parameters	Value
Taiwan	ntrees	30	activation*	ReLU, Tanh, Maxout
	max_depth	5	hidden	c(200, 200, 200)
	min_rows	10	epochs	50
	learn_rate	0,2	rate	0,2
Germany	ntrees	400	activation*	ReLU, Tanh, Maxout
	max_depth	30	hidden	c(200, 200)
	min_rows	2	epochs	15
	learn_rate	0,05	rate	0,01
Australia	ntrees	10	activation*	ReLU, Tanh, Maxout
	max_depth	15	hidden	c(200, 200)
	min_rows	10	epochs	15
	learn_rate	0,01	rate	0,003

*These are the activation functions for the hidden layers

The hyperparameter settings of GBM and DL in the case of the 70:30 split can be found in table 4.

Table 4. Hyperparameter setting of GBM and DL for the 70:30 split

Dataset	GBM - Parameters	Value	DL - Parameters	Value
Taiwan	ntrees	30	activation*	ReLU, Tanh, Maxout
	max_depth	5	hidden	c(100, 100)
	min_rows	10	epochs	12
	learn_rate	0,2	rate	0,2
Germany	ntrees	390	activation*	ReLU, Tanh, Maxout
	max_depth	24	hidden	c(200, 200)
	min_rows	2	epochs	15
	learn_rate	0,095	rate	0,01
Australia	ntrees	11	activation*	ReLU, Tanh, Maxout
	max_depth	13	hidden	c(200, 200)
	min_rows	7	epochs	15
	learn_rate	0,01	rate	0,003

*These are the activation functions for the hidden layers

Searching the complete parameter space is computationally demanding. According to Bergstra and Bengio (2012) and Ng (2018)**,** random search results in the best parameter choices compared to grid search and can be completed faster. Hence, an exhaustive grid search is not adequate considering the tradeoff between accuracy and training time. To tune the hyperparameters in this study a combination of random search, grid search, and manual adjustments was used. The random grid search was applied for 15 minutes. Afterward, a grid search was used around a small interval of the parameters determined by random search to further calibrate the models and to improved accuracy. Once this was completed, I have tried micro-adjustments of the hyperparameters to see whether there is a possibility left to enhance the accuracy levels. The manual changes were only done for one parameter at a time while holding all the others constant. Where this was possible it did only impact the accuracy levels slightly. No significant performance improvement could be reached at the final step and mainly the random search plus selective grid search resulted in maximum performance.

In case the hyperparameter value is not mentioned in Table 3 or 4 the default value ascribed by H2O was used during the model training. Concrete advice on parameter choices in ML is subject to further research (Ng, 2018).

3.4 Numerical Results

Three datasets containing detailed customer-level data were used to benchmark Gradient Boosting Machine (GBM) against Deep Learning (DL). Table 5 shows the out-of-sample performance of the trained GBM and DL models and gives a complete summary of the results obtained during this study. It shows the AUC for each of the 3 datasets as well as the two training/test set splits. The AUC as accuracy measures should be diagnostically conclusive as the datasets were resampled and balanced before the model training.

Table 5. **Model results separated by dataset as well as training/test set split**

Dataset	Method	AUC	Method	AUC
	80:20 Split		70:30 Split	
Taiwan	Gradient Boosting	0.773	Gradient Boosting	0.771
	DL + ReLu	0.765	DL + ReLu	0.759
	DL + Tanh	0.744	DL + Tanh	0.741
	DL + Maxout	0.761	DL + Maxout	0.754
Germany	Gradient Boosting	0.885	Gradient Boosting	0.823
	DL + ReLu	0.941	DL + ReLu	0.838
	DL + Tanh	0.919	DL + Tanh	0.816
	DL + Maxout	0.936	DL + Maxout	0.829
Australia	Gradient Boosting	0.988	Gradient Boosting	0.989
	DL + ReLu	0.964	DL + ReLu	0.961
	DL + Tanh	0.961	DL + Tanh	0.943
	DL + Maxout	0.974	DL + Maxout	0.965

3.4.1 Dataset 1: Taiwan

In the case of the 80:20 split GBM achieved an AUC of 0.773 during the out-of-sample test. The best performing DL model used the ReLU activation function and achieved an AUC of 0.765. See figure 11.

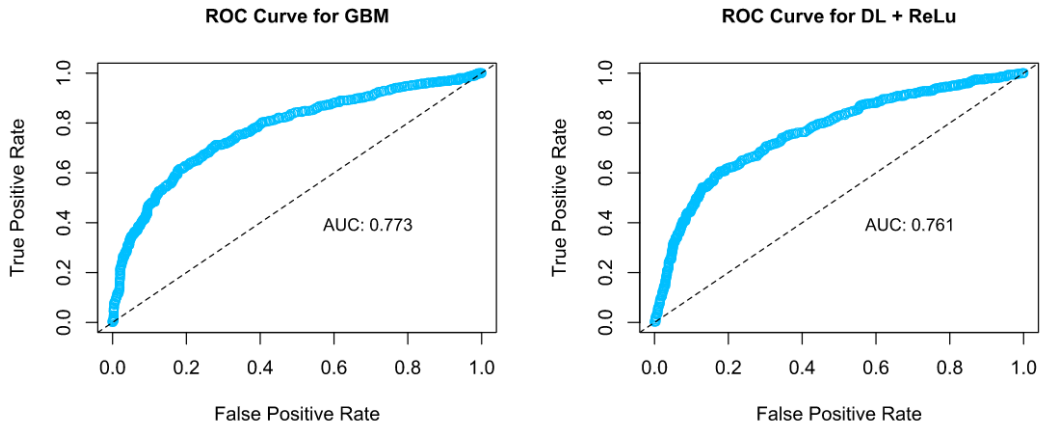

Figure 11. Performance of GBM vs. DL with ReLU function on Taiwanese dataset and 80:20 split

The 70:30 split has a slightly lower classification power in terms of the used metric, which shows an AUC of 0.771 for GBM and an AUC of 0.759 for the DL model, which was also used the ReLU as the activation function. See figure 12.

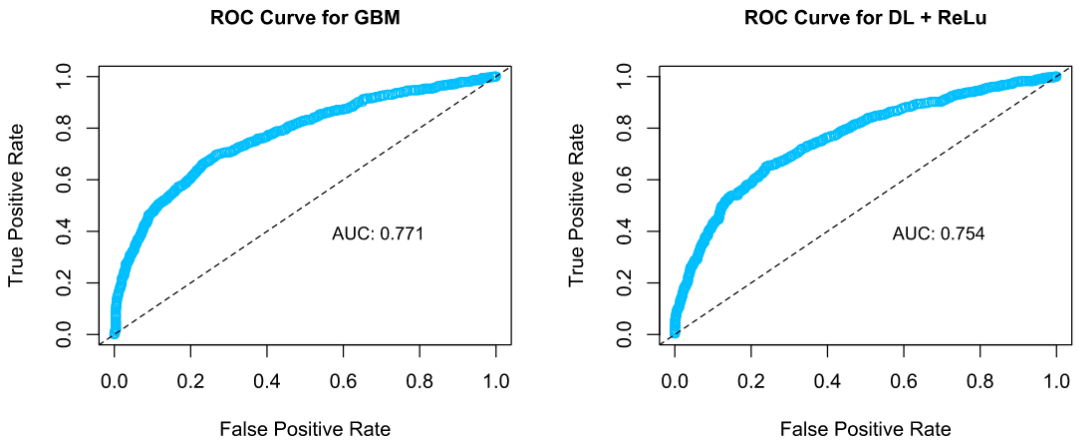

Figure 12. Performance of GBM vs. DL with ReLU function on Taiwanese dataset and 70:30 split

Overall, based on the first dataset GBM remains in terms of accuracy the superior model compared to DL. The DL model utilizing the Tanh activation function was at the lowest performance end at both training/test set splits.

3.4.2 Dataset 2: Germany

The results for Dataset 2 are surprisingly different from Dataset 1. In the case of the 80:20 split GBM achieved an AUC of 0.885 on the test data, which was significantly lower than the best performing DL which achieved an AUC of 0.941. See figure 13.

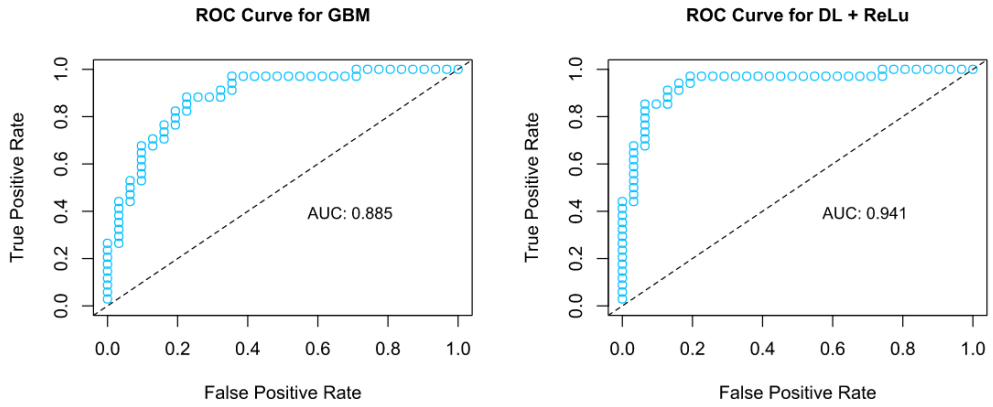

Figure 13. Performance of GBM vs. DL with ReLU function on German dataset and 80:20 split

The 70:30 split for dataset 2 has a slightly different classification accuracy with an AUC of 0.823 for GBM and an AUC of 0.838 for the best DL model that uses again the ReLU activation function. See figure 14.

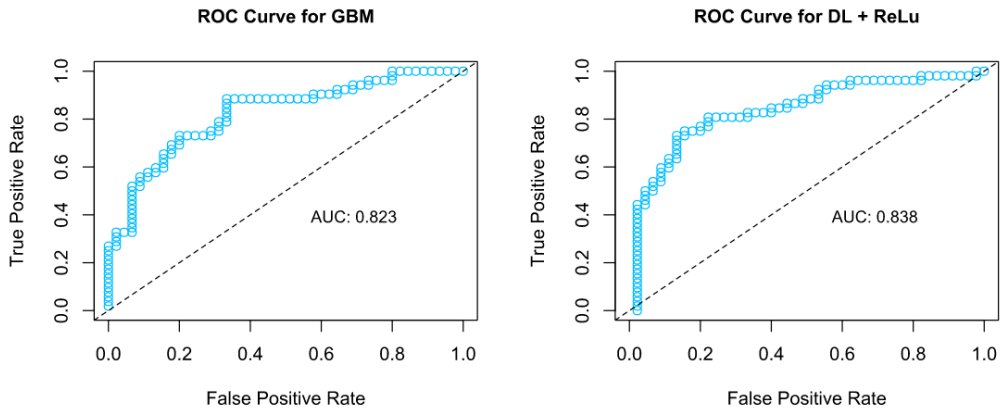

Figure 14. Performance of GBM vs. DL with ReLU function on German dataset and 70:30 split

Overall, the position for the highest classification power is reversed in the case of the second dataset and DL was able to outperform GBM in terms of the accuracy metric AUC. In addition, all 3 DL models could represent the dataset better than GBM, while the DL model with the Tanh activation function takes the lowest spot w.r.t to classification accuracy.

3.4.3 Dataset 3: Australia

The out-of-the sample results obtained for Dataset 3 in case of the 80:20 split show an AUC of 0.988 for GBM, and an AUC of 0.974 for the best DL model, which was this time achieved by the Maxout activation function. See figure 15.

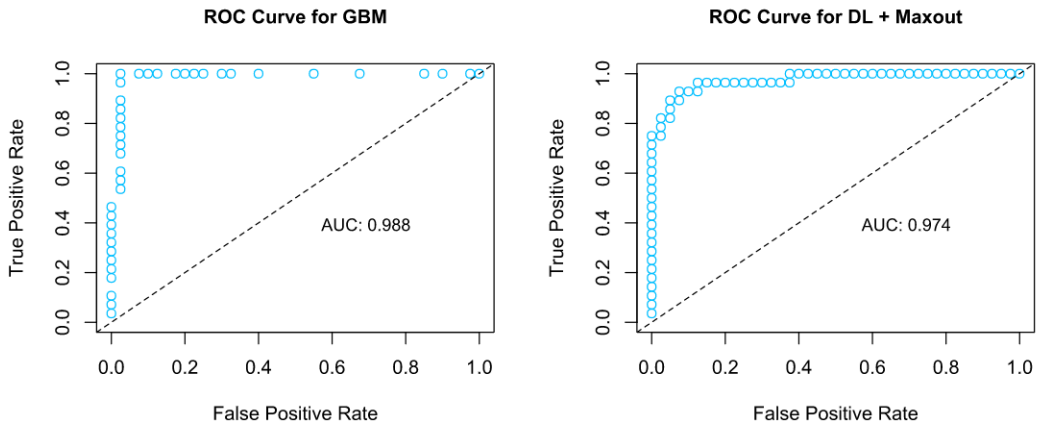

Figure 15. Performance of GBM vs. DL with Maxout function on Australian dataset and 80:20 split

In the case of the 70:30 split GBM achieved a similarly impressive AUC of 0.989 on the test data, while the best performing DL model which also uses the Maxout activation function achieved an AUC of 0.965. See figure 16.

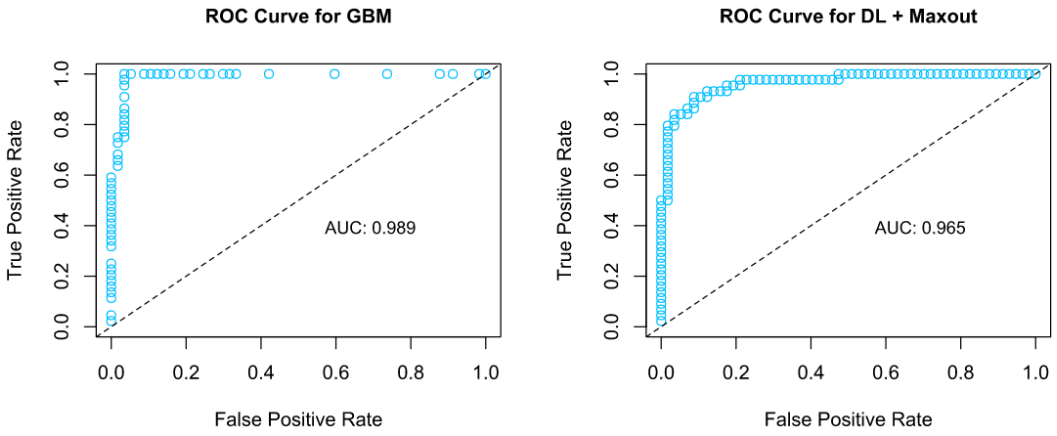

Figure 16. Performance of GBM vs. DL with Maxout function on Australian dataset and 70:30 split

Overall, the performance of GBM is superior to DL in terms of the AUC performance metric for both training/test set splits. Again, the DL model with the Tanh activation function had the lowest AUC among the DL models in both cases.

3.5 Discussion

The results of this study are clearly in favor of Gradient Boosting Machine (GBM) which is based on the results of this experiment superior in terms of accuracy to Deep Learning (DL). This is especially true in the case of dataset 1 and 3, which confirms the findings of Hamori et al. (2018) and Addo et al. (2018). Nevertheless, DL was able to outperform GBM on the second dataset, which suggests that the underlying structure of datasets is important and only slight variations might result in the need for a different model or at least a change in the model configuration.

Given the observation that there exists a tendency to achieve higher prediction accuracy when increasing the number of observations within the training set, it is recommended to use as many observations in the training set as possible. However, the size of the two datasets Germany and Australia with only 1000 and 690 observations respectively are relatively small. Especially in combination with the additional resampling and balancing, which further reduced the number of observations to 600 and 614 made it difficult to go for a 90:10 split. The resulting number of observations in the test set would have been too small, which would have resulted in inadequate performance.

The choice of the activation function for deep learning does indeed have an impact on model performance. Comparing the different DL models and activation function choices, it can be observed that the ReLU activation function performs the best, with the Maxout activation

function as a close follower. The Tanh activation function has shown a consistent underperformance across all the tested scenarios.

The impact of the different training/test set splits (80:20 and 70:30) on model performance was quite trivial and did not alter the essential findings across all three datasets. Overall, the training process GBM is less complex and it is way easier to achieve a satisfactory prediction accuracy with Gradient Boosting Machine then with Deep Learning. Also, the space of hyperparameters for GBM is smaller and fewer variations have to be considered, which results in significantly faster training time for GBM compared to DL.

Decisions have often to be done in real-time, which requires model adjustments to be carried out within minutes or maximum hours. If two or more models deliver more or less the same result the less complex model should be the preferred choice. This becomes particularly important for large datasets. Flexibility and fast reaction times are key in many business functions. Feature engineering and hyperparameter tuning over several days or weeks to arrive at a satisfactory result is not realistic, which is another reason why GBM should be favored over DL.

There is an overall consensus that GBM is superior to DL when it comes to structured (tabular) data sets and DL is dominating tasks based on unstructured data-sets (Ng, 2018). The findings of this study are in line with the current literature but do not suggest a complete switch to GBM. However, papers who propose deep learning as the one-fit all solution seem not to represent reality. Machine learning remains an empirical process. It is therefore recommended to test different models for different datasets to find the one model that can best represent the information contained within the data. It is also advisable to reevaluate parameter settings and model choice after a non-trivial change in the fundamental dataset has occurred as this might result in different requirements w.r.t. model configurations.

Overall, based on this study both algorithms can be considered state-of-the-art for binary classification tasks on structured datasets, while GBM should be the go-to solution for most problem scenarios due to easier use, significantly faster training time and also superior accuracy.

To strengthen the above findings additional datasets can be used, but it is unlikely that the findings will be negatively challenged as results already indicate that there is no guarantee, but a strong tendency that GBM is the preferred choice for structured datasets in the case of binary classification. However, further research could successfully strengthen those findings and confirm that GBM is the best model available for structured datasets.

3.6 Conclusion and Future Research

The global economy has changed significantly over the last years and new entrants in the form of FinTech companies are increasingly disrupting the current lending market structure. The usage of advanced analytics in this increasingly competitive market is essential. Gradient Boosting Machine (GBM) and Deep Learning (DL) are the two dominating forces currently shaping the business analytics landscape when it comes to supporting lending decisions. This study has shown that in the case of structured datasets, GBM tends to be superior in terms of prediction accuracy. It is easier to use and has also the advantage of computational speed. DL was able to beat GBM in one case, which shows that the outperformance of GBM is not always guaranteed. It seems the model choice is also dependent on the concrete problem scenario and underlying characteristics of the dataset and it might be wise to choose a predictive model that is best suited for the problem scenario at hand. Overall, DL and GBM, or in general advanced analytics are powerful models to support businesses operating in the lending market when it comes to the prediction of counterparty defaults.

This study's purpose was to analyze prediction accuracy for binary classification on structured data, but credit scoring could largely move away from using traditional data for the assessment of the default probability. Not only is it possible to utilize a vast array of new data sources that grow consistently in volume, but our ability to harvest and store those data for improved decision making has dramatically improved as well.

Many FinTech companies have already started to use unstructured data (e.g. text mining and social media data) to further enhance the ability of correct default classification of customers, which could further restructure the lending market. Further research could focus on a better understanding of prediction accuracy and how to utilize the special characteristics of DL to use all kinds of unstructured data to support lending decisions. This seems especially relevant due to significant market changes in developing countries that drive forward non-traditional borrowing as peer to peer lending (C. Wang, Han, Liu, & Luo, 2019). Peer to peer lending and other FinTech's might foster financial inclusion, but also tend to cannibalize the market share of incumbent institutions.

Another area in need of further investigation is the current adoption rate of DL in business analytics functions. This seems to be especially relevant for incumbent corporations (Chui et al., 2018), but the adoption and utilization of DL in business are not easy. The necessary skill sets required to develop and deploy advanced prediction models are often not in place. The

questions why DL was not able to find its way into business analytics functions as expected due to the hype and attention it has received recently will be explored in the next chapter.

New emerging models as AutoML could help to close this gap and further democratize ML. A comprehensive analysis of AutoML and its current capabilities in comparisons to manual model tuning can be found in chapter 6.

4 DL in Business Analytics: A Clash of Expectations and Reality

Abstract

Our fast-paced digital economy shaped by global competition requires increased data-driven decision making based on advanced analytics and machine learning. The first wave of digital transformation based on big data and analytics is now gradually replaced by AI, which becomes the driving force behind new digital transformation initiatives. The benefits of Deep Learning (DL) over traditional analytics are manifold, but it comes with limitations that have – so far – interfered with widespread industry adoption. This chapter conveys an accurate picture of the current deployment of DL in business analytics. It contains three case studies of different business use cases and benchmarks DL against traditional machine learning models. It is shown that the adoption of Deep Learning is not only affected by computational complexity, lacking big-data architecture, lack of transparency (black-box), and skill shortage, but also by the fact that DL does not outperform traditional ML models in case of structured datasets with fixed-length feature vectors as usually present in relational database systems. DL does not show superior performance for classification problems on structured data across several domains. DL does not achieve higher performance as Gradient Boosting Machine and Random Forest. These results are consistent across all three use cases presented in this study, which offers a logical explanation of why DL adoption is lacking behind expectations. DL should be regarded as a powerful addition to the existing body of ML models instead of a one fits it all solution.

Keywords: Business Analytics, Predictive Analytics, Digital Transformation, Deep Learning, Ensemble Learning

4.1 Introduction

The last decade was shaped by huge improvements in data storage and analytics capabilities (Baesens et al., 2016; Henke et al., 2016). What started as the Big-Data revolution brought us the age of constant digital change, accelerating globalization, and the consensus that we are moving towards a digital world economy (Davenport, 2018; Warner & Wäger, 2019). Companies operating in today's world have to deal with global competition in an ultra-fast market place (Davenport, 2018).

Amidst all this formed a new paradigm called Deep Learning (LeCun et al., 2015) which emerged out of earlier research of brain-inspired neural networks. DL is part of ML and is one

of the major technologies responsible for driving the current digital revolution (Agrawal et al., 2019; Bughin et al., 2017). DL is capable of learning complex hierarchical representations of data. It was able to outperform traditional methods and has predictive capabilities that come close or surpass human-level intelligence in different areas. The main reasons for the breakthrough of DL stem from developments in three different areas (Goodfellow et al., 2016): (1) Optimization algorithms allow the training of deep neural networks (Hinton, Osindero, & Teh, 2006); (2) The era of "Big Data" increased the amount of large structured, as well as unstructured data-sets, which are now ripe for harvesting (Chen et al., 2012); and (3) hardware improvements, especially GPU's made it possible to train those highly power-hungry models with those huge data-sets. Stadelmann et al. (2018) give a good summary of current applications of DL across different domains. When it comes to image recognition (Krizhevsky, Sutskever, & Hinton, 2012; Szegedy et al., 2015), NLP (Devlin, Chang, Lee, & Toutanova, 2018), and games (Silver et al., 2017; Vinyals et al., 2019), DL is the go-to solution. Accurate performance for unstructured high-dimensional data-sets became only possible due to the advances of DL, which significantly enhances the field of machine learning (Jordan & Mitchell, 2015) to tackle further use cases and take over tasks that were initially only reserved for humans (Agrawal et al., 2019).

We have entered the second wave of digital transformation and the deployment of advanced analytics in the form of machine learning has become a necessity to survive and thrive in this new environment where competitive advantage is mainly based on evidence-based or data-driven decision making (Henke et al., 2016). The function responsible for converting raw data into valuable business insights is called business analytics. It is an interdisciplinary field drawing and combining expertise from machine learning, statistics, information systems, operations research, and management science (Sharda, Delen, & Turban, 2017). Business analytics constitutes a quite long chain of different analytics, which includes descriptive, predictive, and prescriptive analytics (Delen & Ram, 2018). ML operates mainly in the predictive sphere of Business Intelligence but has started to incorporate prescriptive analytics as well (Bertsimas & Kallus, 2019).

Most analytics departments across the corporate value chain have traditionally been using predictive statistics and machine learning models as GLMs, CART, and ensemble learning. Those models are vital tools to help with several analytics tasks that directly impact the bottom-line of firms and organizations (Siebel, 2019). We have moved from fundamental progress to the application of deep learning in various sciences, businesses, and governments (Lee, 2018; Stadelmann et al., 2018). Despite the huge success of DL, a closer investigation of the current

literature reveals that the adoption rate for DL in business functions for analytic purposes is quite low.

Chui et al. (2018) analyzed 100 use cases to demonstrate the current deployment of AI/DL related models across industries and business functions compared to other models referred to as traditional analytics. The result is that while the adoption of DL starts to increase, it seems most units remain working with the older more established analytical models that have been successful already years ago. McKinsey (Chui et al., 2018) also distinguishes departments that have traditionally been using analytics as compared to departments that are foreign to quantitative decision enablers. McKinsey draws a clear picture that shows that the only areas where DL has been utilized so far are traditional analytics arms that have the natural capabilities and skillsets in place to work with modern AI, while technology foreign departments are reluctant to adopt DL models. But even in business units with traditionally strong links to analytics as risk management and insurance remains the utilization of DL quite low and traditional models are still the go-to solution.

The latest paper on the topic confirmed this observation: “While deep learning is on the way to becoming the industry standard for predictive analytics within business analytics and operations research, our discipline is still in its infancy with regard to adopting this technology.” Kraus et al. (2019) has analyzed several papers across the major journals and concluded that DL does not prevail within business analytics functions as perceived due to the current hype and job descriptions.

The main issues why it is not so easy to develop and deploy DL – especially for small to medium-sized corporations – can be partly mapped to the three reasons why DL found its breakthrough in recent years. The “content analysis” of the existing literature identified the following bottlenecks when it comes to the adoption of DL in business analytics functions:

(1) **Computational Complexity:** The hardware necessary to train and validate DL models on large-datasets is tremendous, which makes infrastructure investments quite expensive. This stands in large contrast to the question, whether the development and implementation of those models will materialize and be reflected in a future value increase (Bughin et al., 2017).
(2) **Infrastructure:** Companies need to be able to harvest a continuous flow of unstructured data to capture the value from DL, which is difficult if the necessary “Big Data” infrastructure is not in place (Bughin et al., 2017).

(3) **Transparency:** Another reason is the nature of DL itself. DL is mainly a black-box, which means it can predict correctly, but we lack a causal explanation of why it arrives at a certain decision (Samek & Müller, 2019). This makes it problematic for industries, which are subject to regulatory supervision.
(4) **Skill Shortage:** Talent is required to implement those models as well as subject matter expertise to define use cases (Henke et al., 2016). The current supply and demand gap for ML experts makes it difficult for small- and medium-sized corporations to utilize advanced AI.

Nevertheless, what many studies about the adoption of DL in business analytics seem to ignore is its general value contribution, which should come in the form of improved prediction accuracy. DL has to make a business case for itself to justify its adoption in functional areas, but this is not always given. Several standalone studies comparing the predictive ability of deep learning against traditional machine learning methods on structured data-sets have concluded that DL does not outperform tree-based ensembles (Addo et al., 2018; Hamori et al., 2018). This stands in contrast to the claim that DL offers performance improvements across the board as indicated by Kraus et al. (2019) and also to the general assumption that DL needs to be adopted in every business function (Chui et al., 2018). While the success of DL for unstructured data problems as image recognition and NLP is beyond doubt, the reality about DL for structured data within companies' business analytics functions is less clear and is the main focus of this chapter. Structured data with fixed-length feature vectors are vastly present in many relational databases and standard business uses cases.

Comments as "DL can be a simple replacement of traditional models" are too general and not always true. For structured data, tree-based ensembles as gradient boosting seem to be at least on par with DL across different domains. In support of this claim, an empirical test using three case studies based on real-world data is presented. Concrete, this chapter will contribute to the current body of literature in the following ways: (1) DL is compared to traditional machine learning models as GLMs, Random Forest, and Gradient Boosting based on three real-word use cases within the context of business analytics to verify the assumption that DL does not outperform traditional methods on structured datasets. (2) A comprehensive discussion on the bottlenecks of DL identified during the "Content Analysis" taking into account the findings of the empirical study including managerial recommendations will be given. (3) In the end, a roadmap for future research possibilities for deep learning and business analytics is presented.

This chapter is structured as follows: Section 2 introduces the methods logistic regression and random forest. Second, the experimental design is presented, which includes an explanation

of the dataset, preprocessing steps, and the general setup. In section 3, the numerical results from the three case studies based on real-world data/business problems are presented. All three case studies show that in the case of structured data (tabular data) DL does not have a performance advantage over the tree-based ensembles Random Forest and Gradient Boosting Machine. Section 5 gives a discussion of the technical implications of these results, highlights managerial implications, and suggestions for future research while section 6 concludes with a summary.

4.2 Experimental Design

4.2.1 Methods

The ML models used and compared in this experiment are Logistic Regression (LR), Random Forest (RF), Gradient Boosting Machine (GBM), and Deep Learning (DL).

The **Logistic Regression (LR)** belongs to the big family of generalized linear models (GLMs). GLMs are characterized by taking as input a linear combination of features and link them to the output with the help of a function where the output has an underlying exponential probability distribution like the normal distribution or the binomial distribution (Murphy, 2012). The LR is the standard method for binary classification and widely used in academia and industry. A linear combination of inputs and weights is calculated and applied by feeding $w^T x$ into the logic or sigmoid function represented by

$$sigm(w^T x) = \frac{1}{1 + e^{-w^T x}} = \frac{e^{w^T x}}{e^{w^T x} + 1}. \tag{14}$$

The sigmoid function restricts the range of the output to be in the interval $[0, 1]$.

The recursive partitioning algorithms **Random Forest (RF)** is part of the family of ensemble methods and operates very similar to decision trees with bagging. Bagging (Breiman, 1996a) chooses randomly different M subsets from the training data with replacement and averages these estimates. The random forest creates different decision trees and averages the results in the end to reduce the variance of the prediction model (Murphy, 2012). It is one of the most potent ML algorithms for classification and regression tasks out there.

A description of deep learning and GBM can be found in chapters 3.2.1 and 3.2.2 respectively.

4.2.2 Data and Preprocessing

This experiment is based on three datasets. The idea was to cover several application areas within financial services for supervised classification, hence the chosen datasets cover banking (credit risk), insurance (claims prediction), and marketing and sales for a retail bank. As banks and insurance companies have similar products – non-physical products or services – and therefore use similar distribution channels, the results based on those datasets can be easily generalized for most financial service companies. Insurance companies and banks tend to cooperate during sales activities and often bundle their products – e.g. bank accounts with insurance policies (bancassurance).

All three use cases require the same ML method, which is supervised learning and binary classification and where used in earlier studies, which allows for easy comparison of classifier strength regarding earlier studies. To facilitate reproducibility and comparability the chosen data-sets are all publicly available and can either be downloaded from the UCI machine learning repository or from the public machine learning competition site “Kaggle”, which regularly offers access to high-quality datasets for experimentation. See table 6 for an overview of the case studies/datasets used in this study.

Table 6. Description of Datasets

Business Area	Observations					Description
	Total	y = 0	y = 1	Balanced*	Features	
Credit Risk	30,000	23,364	6,636	6636/6636	23	Prediction whether a customer is going to default on their loan payment
Insurance Claims	595,212	573,518	21,694	21694/21694	57	Prediction whether a policy holder will initiate an auto insurance claim in the next year
Marketing/Sales	45,211	39,922	5,289	5289/5289	16	Prediction whether a targeted customer will open a deposit account after a direct marketing/sales effort

*For the purpose of this study random under-sampling was used to bring the datasets in a balanced state

Credit Risk: The first dataset represents payment information from Taiwanese credit card clients. It consists of 30,000 observations, of which 23,364 are good cases and 6,636 are bad cases (flagged as defaults). Each observation contains 23 features including a binary response column for the default information of the credit cardholder. The features within the dataset

contain mainly historical payment information, but also demographic information as gender, age, marital status, and education. [1]

Insurance Claims: The second dataset represents information about automotive insurance policyholders. It consists of 595,212 observations, of which 573,518 are non-filed and 21,694 are filed claims. Each observation contains 57 features including a binary response column which indicates whether or not a particular policyholder has filed a claim. [2]

Marketing and Sales: The third dataset stems from a retail bank and represents customer information for a direct marketing campaign. It consists of 45,211 observations, of which 39,922 were unsuccessful and 5,289 were successful (resulted in a sale). Each observation contains 16 features including a binary response column indicating whether or not the person ended up opening a deposit account with the bank following the direct marketing effort. [3]

For a more detailed description of the features present in the datasets see section 3.3.1 (credit risk) and the Appendix (insurance and marketing).

The experiment required several adjustments. All three datasets are highly unbalanced. For this study, random under-sampling was used to bring the good as well as the bad cases in a state of equilibrium. This can also be seen in table 5. Example: If highly unbalanced datasets with a ratio of 90:10 are trained it is very easy for the classifier to reach an accuracy of 90% by simply going for the positive observations in all cases. To counter this natural occurring gravitation towards the majority class resampling is used to better gauge the predictive ability of the classifiers. One drawback of under-sampling might be a loss of information, but can be neglected as the major purpose of the dataset is to benchmark the introduced ML classifiers.

Before model construction can take place, several other common preprocessing steps have been performed. A required procedure in ML during preprocessing is to transform categorical values in a numerical representation. Especially the “Case Study 3 – Marketing and Sales” contains predominately categorical strings. Where necessary categorical features were transformed to factor variables with a method called one-hot encoding. H2O has a parameter

[1] The “Credit Risk” dataset can be accessed here: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[2] The “Insurance Claims” dataset can be accessed here: https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data
[3] The „Marketing/Sales dataset can be accessed here: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

setting called one_hot_explicit, which creates N+1 new columns for categorical features with N levels.

For this experimental study, all three datasets are separated into a training set and test set with a proportion of 80:20. To tune the model parameters the training set will be further divided into different training and validation sets using a method called cross-validation during the construction of the classifiers. Cross-validation is used to increase the generalization ability of the classifiers to unknown data and to avoid overfitting. This study uses 5-fold cross-validation.

Model tuning in ML is a highly empirical and interactive process and is essentially based on trial and error. The methods commonly used to help with automating the model tuning process are grid search and random search. Grid search automatically trains several models with different parameter settings over a predefined range of parameters. Overall, this does not change the basic necessity of trying out different combinations of parameters that allow the classifier to adjust adequately to the underlying dataset. This study used a random search, selective grid search, and manual adjustments to arrive at the final parameter settings.

To determine the predictive power of an ML model, an evaluation method is required. The performance metrics used are AUC, Accuracy, F-score, and LogLoss as described in chapter 2.3.5.

4.3 Numerical Results

In this section, three different case studies: Credit Risk, Insurance Claims, and Marketing and Sales are presented to demonstrate that deep learning while being promoted as superior ML solution has difficulties to beat traditional machine learning methods in some cases. Concrete, Logistic Regression, Random Forest, Gradient Boosting Machine, and two different Deep Learning classifiers were trained on each dataset. The first DL model was built with the ReLU activation function whereas the second DL model was built with the Maxout activation function. The ReLU activation function is widely used and has shown to be superior in terms of accuracy and computational speed. The Maxout activation function has been developed to improve classification accuracy in combination with drop out (Goodfellow, Warde-Farley, Mirza, Courville, & Bengio, 2013; Srivastava et al., 2014) and is hence the second choice for this experiment. The Tanh function as used in chapter 3 has been dropped due to consistent underperformance in terms of prediction accuracy and time to completion. Several hyper-parameters were adjusted during the model training process to improve the performance measured by the evaluation metrics AUC, Accuracy, F-score, and LogLoss.

4.3.1 Case Study 1: Credit Risk

Numerical results for the credit risk business case to accurately predict the default category of an applicant. The performance of deep learning is compared to traditional machine learning classifiers via the four evaluation matrices AUC, Accuracy, F-score, and LogLoss. The best performance is highlighted in bold.

Table 7. Numerical results for Case Study 1 - Credit Risk

Method	Out-of-Sample Performance			
	AUC	**Accuracy**	**F-score**	**Logloss**
Logistig Regression	0.712	0.671	0.653	0.623
Random Forest	0.773	0.711	0.688	0.572
Gradient Boosting Machine	**0.774**	0.712	0.691	0.572
Deep Learning + ReLU	0.760	0.700	0.646	0.592
Deep Learning + Maxout	0.762	0.703	0.687	0.599

Table 7 shows clearly that GBM has the best overall performance with the highest AUC, Accuracy, and F-score of 0.774, 0.712, and 0.691 respectively, including a LogLoss of 0.572. RF comes as a close second with an AUC of 0.773 and the same LogLoss as GBM of 0.572. Both ensemble models achieve a better performance in the case of the credit risk dataset than the two DL models with an AUC of 0.760 and 0.762 respectively. The DL + Maxout model has a slightly higher AUC compared to the DL + ReLU, whereas the LogLoss is reversed, which results in a similar performance of the two DL models.

A graphical presentation of the results of each model sorted by the evaluation measure can be found in figure 17. The best performing model GBM is highlighted via a callout text field, which shows the performance of each evaluation metric.

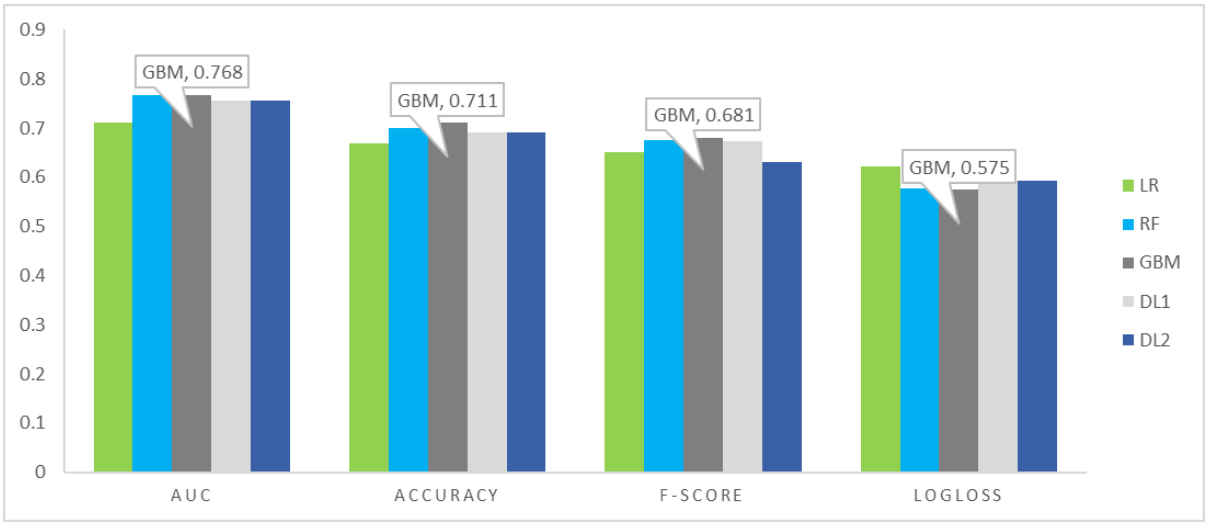

Figure 17. Graphical representation of the performance of each classifier for all 4 performance evaluation metrics for case study 1 - credit risk. Gradient Boosting Machine (GBM) achieves the highest accuracy according to those results.

4.3.2 Case Study 2: Insurance Claims

In table 8 the numerical results for the insurance case study are presented. The goal is to accurately predict whether a policyholder is going to file an insurance claim within the next year. The performance of deep learning is compared to traditional machine learning classifiers via the four evaluation matrices AUC, Accuracy, F-score, and LogLoss. The best performance is highlighted in bold.

Table 8. Numerical results for Case Study 2 - Insurance Claims

Method	**Out-of-Sample Performance**			
	AUC	**Accuracy**	**F-score**	**Logloss**
Logistig Regression	0.629	0.594	0.586	0.667
Random Forest	0.636	0.598	0.584	0.667
Gradient Boosting Machine	**0.640**	0.602	0.588	0.664
Deep Learning + ReLU	0.628	0.597	0.540	0.670
Deep Learning + Maxout	0.633	0.597	0.534	0.669

The results of table 8 are similar to the first case study. GBM is the clear winner in terms of performance with the highest AUC, Accuracy, and F-score of 0.640, 0.602, and 0.588 respectively, including the lowest LogLoss of 0.664. RF takes the second place with an AUC of 0.773 and a LogLoss of 0.664. Both ensemble models achieve a better performance regarding the insurance case study than the two DL models. The DL + Maxout model with an AUC of 0.633 has a slightly higher AUC compared to the DL + ReLU with an AUC of 0.628.

A graphical presentation of the results of each model sorted by the evaluation measure can be found in figure 18. The best performing model (Gradient Boosting) is highlighted via a callout text field.

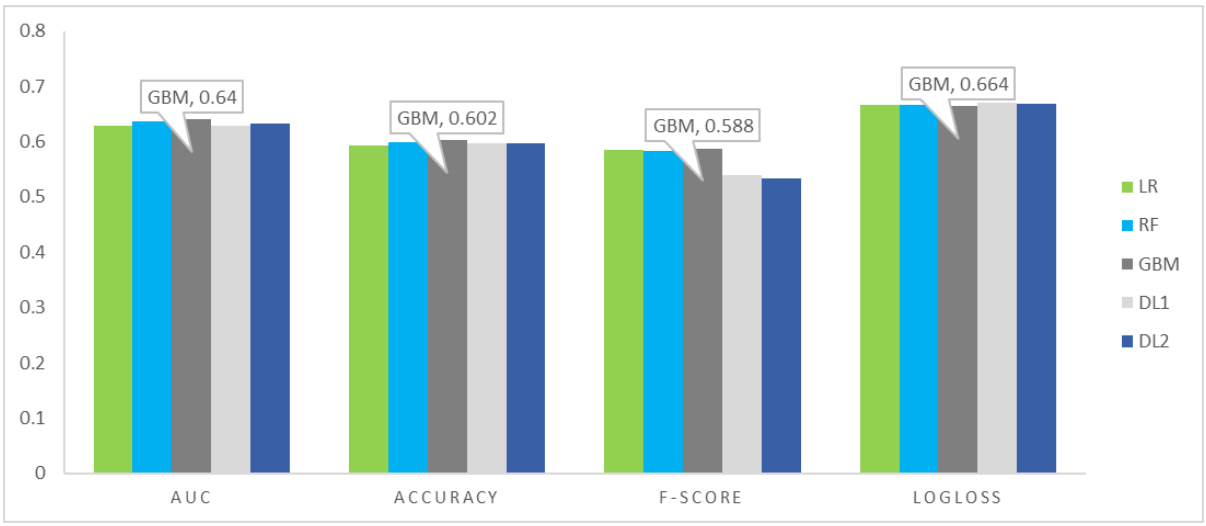

Figure 18. Graphical representation of the performance of each classifier on all 4 performance measures for case study 2 - insurance claims. Also, in the second case study, Gradient Boosting Machine (GBM) achieves the highest prediction accuracy.

4.3.3 Case Study 3: Marketing and Sales

Table 9 shows the numerical results for the marketing and sales case study to accurately predict successful conversions based on a direct marketing effort. The performance of deep learning is compared to traditional machine learning classifiers via the four evaluation metrics AUC, Accuracy, F-score, and LogLoss. The best performance is highlighted in bold.

Table 9. Numerical results for Case Study 3 - Marketing and Sales

Method	**Out-of-Sample Performance**			
	AUC	**Accuracy**	**F-score**	**Logloss**
Logistig Regression	0.918	0.839	0.845	0.377
Random Forest	0.940	0.879	0.888	0.320
Gradient Boosting Machine	**0.940**	0.878	0.886	0.299
Deep Learning + ReLU	0.930	0.861	0.877	0.328
Deep Learning + Maxout	0.930	0.857	0.865	0.336

Based on table 9 the results for the third case study are slightly different from case study one and two. GBM shares the maximum AUC of 0.940 with RF. The RF classifier has also slightly higher Accuracy of 0.879, and also a higher F-score of 0.888 while GBM has still the lowest

LogLoss, which indicates the highest prediction reliability across the models. In line with previous results, both ensemble models achieve a better performance than the two DL models, which have both an AUC of 0.930. LR underperforms all classifiers by a significant amount.

A graphical presentation of the results of each model clustered by the evaluation measure can be found in figure 19. It can be seen that GBM and RF perform better than the two DL models across all performance measures while logistic regression turns out the be the weakest classifier.

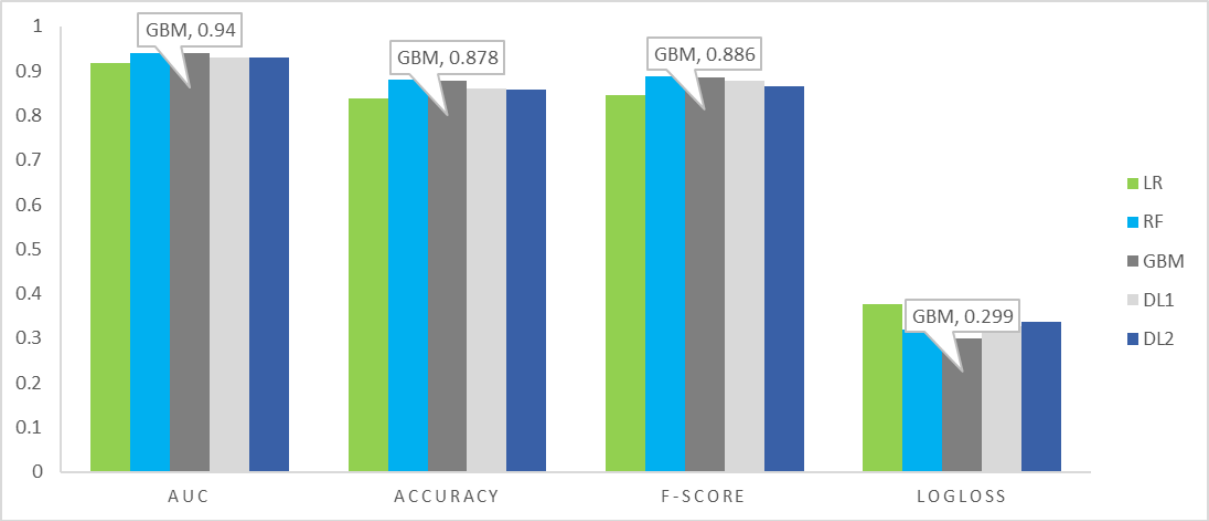

Figure 19. Graphical representation of the performance of each classifier on all 4 performance measures for case study 3 – marketing and sales. Gradient Boosting Machine (GBM) is again the winner, but the results are less significant than before and Random Forest (RF) achieves a very similar performance.

4.4 DL in Business Analytics: A Reality Check

4.4.1 Discussion of Results

To better understand the utility of Deep Learning for Business Analytics it was benchmarked against traditional ML models as GLMs, Random Forest, and Gradient Boosting Machine. Based on the four evaluation measures AUC, Accuracy, F-score, and LogLoss. The empirical results of all three case studies presented (Credit Risk, Insurance Claims, Marketing and Sales) suggest that DL does not have a performance advantage for classification problems based on structured data. Instead, the results are strongly in favor of tree-based ensembles as Random Forest and Gradient Boosting. GBM turns out to be the model with the highest utility for the type of problems analyzed in this study.

Kraus, et al. (2019) benchmarked several baseline models against their proposed embedded DNN model, which resulted in superior performance for DL. The authors recommend fostering the adoption of DL models within the field of Business Analytics and operations research. While the paper of Kraus et al. (2019) is an excellent overview of DL for Business Analytics and very insightful, the analysis does not include GBM as a baseline model in the comparison, which is widely used and known to deliver strong and robust predictions on structured datasets.

Case study two in this chapter uses the same dataset as Kraus et al. (2019) and according to the results of the empirical results is GBM at least on par with the proposed deep architecture by Kraus et al. (2019). Other studies as from Hamori et al. (2018) and Addo et al. (2018) included tree-based ensembles as gradient boosting and came to the same conclusions as the results in this chapter. As the findings of this study are in line with several papers comparing the performance of DL against other ML models there is strong evidence that tree-based methods (GBM as well as Random Forest) do indeed outperform DL models (different configurations have been tested) on most problems containing structured data. Also, DL has several weaknesses as computational complexity, huge data requirements, transparency issues, and needs highly skilled labor, which makes it often difficult to develop and deploy those models at scale. Especially the computational complexity issue results in significantly longer training and validation times compared to all other ML models.

The results strongly suggest that GBM can be seen as the go-to model for most business analytics problems. It is fast, not too complex and delivers for use cases based on structured data the best performance currently available. The results are clear and Business Analytics experts should carefully consider the type, characteristics, and volume of the data at hand to make a final decision about the correct model choice.

4.4.2 Managerial Implications and Digital Strategy

It has been proven that data-driven or evidence-based decisions are superior compared to pure intuitive business decisions and a comprehensive analytics strategy has become necessary for businesses across all industries to capture value at the bottom line. One of the challenges associated with becoming a digital enterprise is how exactly to leverage digital technologies and especially advanced analytics and AI. Current discussions about AI are strongly focused on the applications of DL, but this is not the best way to approach digital transformation. This focus resulted in the problematic assumption that DL adoption in business can be regarded as a benchmark in itself ignoring the question of utility that always needs to be asked before the deployment of any new method or technology.

The main explanation why DL has not found its way into the different business functions as expected is often explained with computational complexity, lacking big-data infrastructure, the non-transparent nature of DL (black-box), and a shortage of skills. But as was demonstrated in this chapter, an additional explanation for the lack of adoption in certain business analytics functions is that DL does not have performance advantages over traditional analytics when it comes to structured data use cases.

For example, many departments that have been utilizing advanced analytics as risk management are perfectly capable of developing and deploying a DL model as the required skillset is identical. Also, the necessary infrastructure to leverage DL in these departments should be in place. The usually described problems cannot be the only reasons. The problem is that DL does not offer any advantage over certain tree-based ensembles for the data present in those departments. Also, the disadvantages as speed and transparency are still present, which makes it, in fact, unreasonable to use DL instead of traditional analytics. DL should be viewed as a valuable addition to the current body of ML that offers the possibility to create new use cases based on its strength instead of forcefully replacing models that are equally powerful and can easily coexist within advanced analytics.

This realization triggers the second argument, which is related to the nature of the underlying dataset. The kind of data present in problems faced within business analytics can largely be divided into three groups (Chen et al., 2012): (1) Structured data from relational database management systems (DBMS), (2) unstructured data, which stem mainly from web-based activities (Social Media Analytics, etc.), and (3) sensor- and mobile-based content, which is largely untouched when it comes to research activities. Many problems in business analytics are indeed based on structured datasets and given that most business functions utilize exactly those kinds of data it should not come as a surprise that DL remains a rather scarce ML algorithm to support their decision making.

The era of Big Data has brought tremendous amounts of data within a single data-set across several domains, which fulfills the requirement of empirical prediction based on deep learning. However, it is important to differentiate and use DL models mainly in line with their strength, which is the usage of vast unstructured datasets, which posed significant problems for traditional analytics. ML overall has been recognized as a General Purpose Technology (GPT) for decision making, which has just started to infuse our economy with the ability to replace mental tasks that were traditionally only reserved for humans (Agrawal et al., 2019). It has also the potential to create completely new business models (Siebel, 2019). Finding use cases that are in line with the strength of DL would help to foster the adoption of DL in business analytics.

And the major strength is unprecedented accuracy on unstructured datasets. Traditional ML models reach a performance plateau quite early and further data are not helpful to increase accuracy. DL has here an advantage as it gains predictive power with every additional data-point (Ng, 2018). This makes DL extremely scalable and future proof, especially since hardware power and the amount of available data will increase continuously over the years. Also, DL eliminates the need for extensive feature engineering as this was usually present in the preprocessing stage of data mining and predictive analytics tasks (LeCun et al., 2015). The time required for preparing data-sets often amounts to 80% to 90% of overall task completion and is one of the major reasons why further advances in DL would indeed be welcoming news for all analytics functions. Overall, practitioners should avoid seeing DL as a simple replacement or enhancement of existing tools for predictive analytics tasks, but more of an opportunity to develop new application areas and use case for business analytics based on the strength of DL, which are predictions based on vast amounts of unstructured data.

4.4.3 Problems and Solutions

DL has several disadvantages. Before developing and deploying DL in a business analytics context the following 5 bottlenecks should be considered:

Computational Complexity & Architecture: DL requires more processing power due to the requirement of large data sets and a complex hierarchical structure, which results in significantly higher model training and validation procedures compared to other ML models. The best solution for small to medium-sized businesses is software as a service (SaaS), which are cloud solutions that commercialize analytics capabilities and sell them to different customers. ML clouds are optimized back-ends that use distributed computing as well as parallel processing to provide maximum processing power. In this way, it is possible to outsource the model training to external platforms, which eliminates the need for a strong infrastructure.

Black-Box: Understanding the underlying logic of predictive models is in certain business areas necessary and industries subject to regulatory scrutiny as finance and insurance tend to have problems with accepting black-box methods like DL. This is one reason why it is often difficult to justify the usage and deployment of "Black-Box" models for certain decisions, especially if the outcome of a decision is nontrivial and could have potential future impacts that face legal issues. The most appropriate solution for those scenarios is to simply avoid black-box models. For an extensive discussion on this topic, it is referred to Rudin (2019). Another way would be to use methods that help to explain the prediction output. For a good overview of model transparency, it is referred to Weller (2019) and Samek & Müller (2019). Given the

broad interest from politics in AI, this seems to be one of the major areas where further research could yield substantial benefits (Miller & Stirling, 2019). Future Research about understanding the DL Black-Box would be of fundamental importance for many business functions to be able to incorporate those new models.

Talent Gap: The McKinsey study about DL adoption in different departments reveals that skillsets are the major problem (Bughin, Seong, Manyika, Chui, & Joshi, 2018). Training DL models do not only require more processing power but are also more complex in terms of hyperparameter optimization during model tuning. Finding the right representations of the underlying dataset is not easy, and since out-of-the-box solutions as tree-based ensembles result often in higher accuracy, a justification to use DL, especially for small datasets is often not given. Besides, many industry professionals and also consulting corporations lack understanding of DL and ML technologies and their differences to fully determine where adoption of DL is necessary and where not (Henke et al., 2016). If an internal search for ML experts and also hiring initiatives are not successful the only option is to hire external consultancies with the right expertise to support the implementation of ML solutions. As ML solutions become more automated (see 4.4.4 future research) domain expertise will become equally important to implement and deploy business analytics solutions (Agrawal et al., 2019). Meaning that – to streamline business processes to capture the value through modern ML solutions – interdisciplinary specialists that speak both languages will be necessary to drive the integration of advanced analytics into the existing enterprise architecture.

Prediction Accuracy: The last argument and the most important one is the fact that DL does not (as widely assumed) outperform traditional ML methods at all tasks. This chapter proved that Gradient Boosting, and also Random Forest outperform DL models on structured classification tasks. Also, traditional methods offer the advantage of easier parameter tuning and faster training time. And in most cases, better transparency.

Overall, and in light of the other problems associated with prediction models based on deep neural networks, decisions w.r.t. to the usage and deployment of DL should only be considered when the concrete business case justifies the development and final deployment of a DL model. And this is mainly the case for business areas utilizing unstructured data sets.

4.4.4 Future Research

The following four key areas could be identified where further research is necessary to increase the utility and hence the adoption of DL in business analytics.

(1) Future research in business analytics could focus on identifying currently non-existing uses which are in line with the strength of DL. Due to its ability to handle huge amounts of unstructured data DL is in terms of future possibilities and new use cases more interesting than traditional analytics. DL possesses the ability to create completely new business models and ways of value generation. (2) Enhancing the prediction accuracy of DL for structured data would be a game-changing development for neural networks. DL has several advantages over traditional methods but has in its current capacity difficulties reaching the performance and accuracy levels of tree-based ensembles as Random Forest and GBM for predictions on structured data. A simple replacement makes hence no sense unless further research in this area realizes performance improvements for DL on structured classification tasks. Developments as dropout (Srivastava et al., 2014) and the Maxout activation function (Goodfellow et al., 2013), which were specifically developed to tackle classification problems are going into this direction, but as shown above, are not enough to reach accuracy levels to justify the replacement of tree-based ensemble models as RF or GBM. Further research could focus on enhancing the ability of DL models to consistently surpass traditional ML models. This would be a significant development, which could result in the extinction of all other ML models. (3) Another issue – especially in light of the skill shortage – is that hyperparameter tuning can be a quite complex undertaking requiring the right talent. A recent development are automated machine learning solutions called “AutoML”, which have started to gain traction and are an interesting field of research that can help to further democratize the use of DL models. AutoML will be further discussed in chapter 6. (4) As this study was in its core only concerned with binary classification it is important to extent it with tests on multiclass classification and regression. Especially regression is relevant for finance and insurance due to the presence of financial times series data in those fields. Several studies have shown that Deep Learning architectures as recurrent neural networks (RNN) and long short-term memory (LSTM) are strong candidates for time series data in finance and offer superior performance (Fischer & Krauss, 2018).

4.5 Conclusion

The progress and breakthroughs achieved by DL are undeniable as can be witnessed by a vast array of new real-world applications all around us. Despite this fact, the adoption rate and hence diffusion across business analytics functions has been lacking behind. This study employed a mix of content analysis and empirical study to explain the current lack of adoption of DL in business analytics functions. The content analysis suggested that the lack of adoption across business functions is based on the four bottlenecks computational complexity, no

existing big-data architecture, lack of transparency/black-box nature of DL, and skill shortage. Also, the empirical study based on three real-world case studies revealed that DL does not as widely assumed offer any performance advantage when it comes to predictions based on structured data sets. This has to be taken into account when using deep learning for data-driven decisions within the context of Business Analytics and helps to answers the question of why analytics departments do not deploy those models consistently. Overall, ML as a GPT for data-driven prediction will further find its way into Business Analytics and keep shaping the field. An important realization is that DL is a valuable additional tool for the ML ecosystem and brought new possibilities to analyze data. But it is not yet possible to replace the other models. Especially tree-based models as Random Forest and Gradient Boosting are powerful classifiers when it comes to structured datasets. Practitioners should concentrate on creating new use cases that leverage the advantage of DL instead of forcing the replacement of traditional models.

5 Super Learning in FinTech: In search of maximum prediction accuracy

Abstract

Artificial intelligence and machine learning are gradually changing the lending market structure towards full automation. Advanced predictive analytics helped FinTech firms to develop modern lending businesses that foster financial inclusion due to high prediction accuracy, which opens the possibility to disregard collateral as a safety net. This is a big step towards the democratization of debt markets. In search of maximum prediction accuracy, this chapter is going to train different configurations of a stacked ensemble model that combine the most powerful baseline models into a so-called super learner. Thereby proving that super learning can improve upon the performance of even the best models currently available. Also, the observed outcomes will be used to derive concrete configuration steps that are generalizable to reach the highest prediction accuracy. The four models used as a baseline in this experiment are Logistic Regression, Random Forest, Gradient Boosting Machine, and Deep Learning. The experiment was implemented on three real-world-datasets from the credit risk domain. Also, this experiment is placed in a discussion on financial inclusion and the future of FinTech to convey the importance of ML and AI applications for financial services.

Keywords: Super Learning, Stacking, Classification, Credit Risk, Lending, FinTech

5.1 Introduction

The digital transformation of the world economy has led to an increased focus on data-driven decision making (Henke et al., 2016). Advanced analytics and machine learning play an increasingly important role in the current business landscape and have found several successful application areas within financial institutions, especially in risk management (Leo, Sharma, & Maddulety, 2019). The lending industry, which requires the assessment of the default probability or credit risk of a counterparty, is one of the most active users of machine learning-based predictive analytics. It was long dominated by universal banks but has recently been disrupted by P2P lending institutions and other FinTech firms (Cooper, 2019). This change in the lending market structure has led to a stronger focus on pure accuracy within lending since it directly affects the core business and hence competitive advantage of FinTech companies (Bazarbash, 2019). One downside is that the increased market penetration from disrupting FinTech's is stripping away revenue from incumbent institutions.

The power of ML solutions seems to be out there and available for everyone. Why is it that FinTech's started to disrupt financial markets and banks had difficulties to compete in the first place? The answer is regulations. The very system that was created to protect banks and act as a safety buffer ultimately made them more vulnerable to new entrances who can leverage advanced technologies like machine learning and artificial intelligence without taking into account regulations as capital requirements, fraud prevention, and other moral hazard reducing safety measures (Stulz, 2019).

Nevertheless, despite the reality that tech startups are shaking up the lending market, FinTech firms do not pose an inherent threat to incumbents. History shows that banks, especially the world-leading organizations are quite adaptable, and usually provide those services quite soon on their own (Stulz, 2019). Or they simply end up buying an emerging FinTech start-up that has made major progress. In contrast, BigTech firms could turn out to be more dangerous than small FinTech startups and severely disrupt the financial systems. Companies that belong to this group are US firms like Amazon, Facebook, Google and Chinese firms like Alibaba and Tencent. They have already a customer base, a modern technology stack, and access to data, which gives them enough leverage to successfully compete with existing financial intuitions (Stulz, 2019).

Yet, looking at the market capitalization of the big financial institutions there is no indication that they will soon be replaced by either FinTech firms nor BigTech firms. There were several predictions over the years that FinTech will lead to the extinction of banks, but none of those turned out to be true (Stulz, 2019). It is more likely that FinTech and BigTech bring numerous benefits and help to foster financial inclusion, and diversification through innovation (FSB, 2019). Another factor is that many FinTech firms do not actively compete with traditional banks, but operate more like a new form of an intermediary (market maker/broker) that matches borrowers and lenders. Banks and hedge funds can act as counterparty here as well (Stulz, 2019). FinTech firms often only leverage a digital infrastructure to connect customers and banks, which makes them a standard financial intermediary (Cooper, 2019).

While those developments in financial markets are exciting. What hasn't changed over the years, is the need for advanced analytics to foster accurate decision making. Several studies over the last years have shown that ML models have advantages, but predictive accuracy depends heavily on the concrete ML algorithm used (Bazarbash, 2019). Incumbent corporations can improve their model accuracy by deploying state-of-the-art ML learning models. This is especially relevant in the light of increasing competition faced by FinTech firms and other technology giants, which increasingly enter the lending market (Frost et al., 2019).

In essence, can all market participants leverage progress in the field of machine learning and artificial intelligence, whether they are traditional banks, FinTech firms, or BigTech. All of them require predictive models to facilitate sound decision making.

This chapter will focus on the predictive part of the lending process that assesses the probability of default and assigns a corresponding credit score, which is binary and either good or bad. A pure focus on prediction accuracy has the advantage of cash-flow based lending without collateral requirements from borrowers (Frost et al., 2019). This can help to foster financial inclusion and is especially important for consumers in developing countries and small to medium-sized corporations, which have no collateral and had traditionally limited access to debt capital (Bazarbash, 2019).

There exists a wide literature when it comes to machine learning in banking risk management. For an exhaustive literature review it is referred to (Leo et al., 2019). One area where advanced analytics has been used for decades is credit scoring and binary classification. The default categorization of a counterparty is an indispensable part of many areas of the financial services industry and several different ML models are continuously used and evaluated due to their direct impact on profitability. Among them are traditional generalized linear models (GLMs) as logistic regression (Kraus, 2014), which has been used in credit risk already for several decades, but also more modern ML approaches as Support Vector Machines, Random Forest, Gradient Boosting Machine, and Deep Learning, which are increasingly used in practice (Addo et al., 2018; Bazarbash, 2019; Hamori et al., 2018; Lessmann et al., 2015).

Given the large size of the lending market and the number of approved loans, Hand & Henley, (1997) argues that only tiny improvements in performance can have a substantial impact on the profitability of financial services provides operating in the lending business. A way to further enhance the predictive ability of classification models would be through stacking, which is also referred to as super learning (Van Der Laan, Polley, & Hubbard, 2007). It has been shown that the fusion of different standalone ML models can further enhance the prediction accuracy in different domains. Using data from healthcare (Kabir & Ludwig, 2019) has shown that super learning can improve the prediction accuracy of several base learners. Other papers proposed multi-stage classifiers, evaluated the predictive power of those ML ensembles within the context of credit scoring and were able to achieve robust performance improvements compared to several earlier studies within the field (Guo et al., 2019; Zhang, He, & Zhang, 2019). Super learning itself can be used with different configurations. A so-called meta-algorithm is required in the process, which is one of the baseline models. Surprisingly, there

exists no assessment of the impact of different choices of meta-learning algorithms on the final performance improvement.

This article will demonstrate that there exists the possibility to further enhance classification accuracy within credit risk by utilizing a fusion strategy as stacking ensembles to create a super learning model. This is done by benchmarking state-of-the-art ML models within the context of credit scoring against three distinct configurations of super learners. Concrete, the chapter will contribute the following: (1) Train the currently dominating classifiers as Logistic Regression, Random Forest, Gradient Boosting Machine, and Deep Learning on three different publicly available datasets containing customer-level data for credit card and loan applications. (2) Use the most powerful base learners (single fully trained models) and combine them via stacking to a model referred to as super learning. Different configurations of the super learner will be assessed by finding the best combination of candidate models and by switching the meta-algorithm used for the training of the super learner. Thereby proving that super learning can improve upon the performance of even the best models currently available. (3) The observed outcomes will be used to derive concrete configuration steps that are generalizable to reach the highest prediction accuracy currently available. (4) In the end, a future outlook for the lending market will be provided.

The structure of this chapter is as follows. Section 2 “Super Learning” will introduce the fusion method stacking. Section 3 “Experimental Design” introduces the primary experimental setup. Section 4 “Numerical Results and Discussion” analyses the findings and discusses their implications. The last section gives a conclusion and a future outlook.

5.2 Super Learning

Super learning requires several pre-trained models, which can be combined via a fusion process to a more powerful super learner. See chapters 3 and 4 for a description of the base learners’ logistic regression, random forest, gradient boosting, and deep learning.

Stacking or super learning is an ensemble method that combines different base learners into a so-called super earner. The goal of stacking is not a gradual improvement over weak learners as is the case for boosting, or the averaging of several outcomes as in bagging, instead, it takes several fully trained prediction models and combines them into a single more powerful learner. Stacking was initially introduced by Wolpert (1992) and later formalized by Breiman (1996b). The theoretical foundation was built by Van Der Laan et al. (2007) who proved that the model created through stacking - which they called super learner ensemble - represents

an asymptotically optimal system for learning The general setup to arrive at the super learner is shown in figure 20.

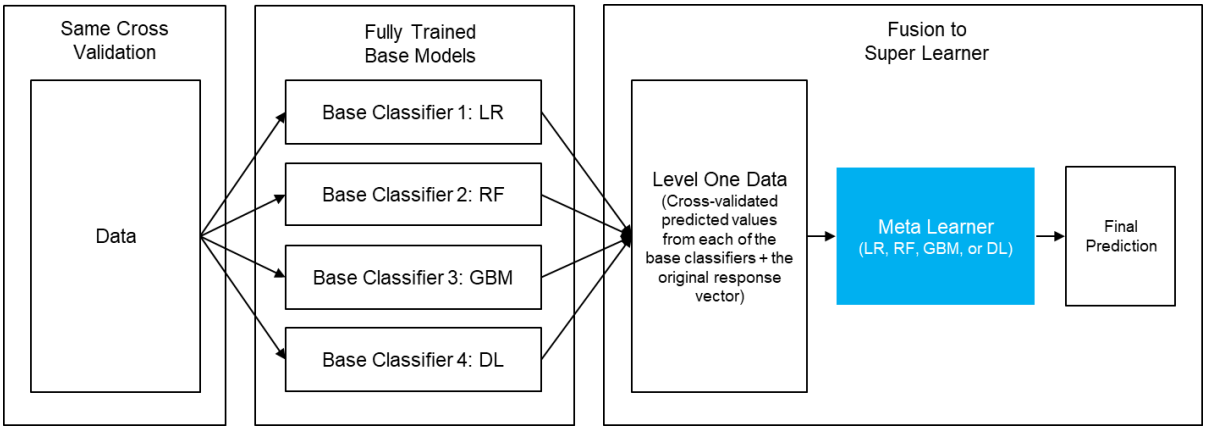

Figure 20. The ensemble method stacking produces a super learner by combining several base classifiers into a single more powerful model. This is done by creating new so-called level one data which is a combination of all the predicted values of the base learners including the original response column. In a final step, the Meta Learner is trained on the new level one data.

The fusion process of the different base classifiers into the super learner requires a similar usage of k-fold cross validation across all the base models. This comes in handy as it is also the best method to tackle the problem of overfitting. Once the base classifiers have been trained the predictions are combined with the original response to create the level one data, which is subsequently used for the training of the meta-learning algorithm.

To generate the final predictions first the predictions of the base classifiers have to be produced. In a subsequent step, those predictions are fed into the meta learner to generate the final ensemble prediction.

Diversification (different ML methods) is usually the better option to get the best results and one of the major advantages of a super learner, but it is possible to stack similar models as well.

5.3 Experimental Design

The primary objective is of this section is to introduce the experimental setup of this study. Three different publicly available datasets are used. The major goal of the experiment in this chapter is to verify the possibility to improve upon the accuracy attainable by the currently dominating ML algorithms identified in chapter 3 and 4 to shed light on the predictive power and usefulness of those ML models for credit scoring within the overall context of credit risk management and ultimately FinTech. This is done by combining several fully trained ML

classifiers to a so-called super learner using a method called stacking. Three different super learners will be created in this study to test the impact of the meta learner on the final accuracy. Once trained the new super learners will be benchmarked against the four base learners' logistic regression, random forest, gradient boosting machine, and deep learning.

Data: The empirical study is based on the same three publicly available data sets as in chapter 3. A split of 80:20 which means 80% of the dataset will be used in the training process and the remaining 20% will be used to test the generalization ability of the trained classifier. The same k-fold cross-validation setup is necessary for the later fusion of the base learners into the meta learner. Hence, during the model training, 80% of the dataset will be split into different training and validation sets, which is done by cross-validation.

Evaluation: The four evaluation methods AUC, Accuracy, F-score, and LogLoss as described in section 2.3.5 will be used to assess the performance of the super learner and the base learners.

Setup: In the first step, the training and validation sets of both datasets were used to fully train the base learners, namely Logistic Regression, Random Forest, Gradient Boosting, and Deep Learning. The prediction results of those base classifiers served as input for the three super learners and also as a baseline to benchmark the performance of the final super learners. The hyperparameter settings for the base learners were chosen by random search and grid search over a predefined range of parameters, including a final manual adjustment. See the Appendix for the concrete hyperparameter values. To enable the following fusion of the models, each classifier was trained with 5-fold cross-validation. In a subsequent step, the predictions of the base learners were used to create the proposed stacked ensemble (as described in section 5.2), which are referred to as SL1 (super learner 1), SL2 (super learner 2), and SL3 (super learner 3). Each super learner uses a different meta-learner algorithm. SL1 uses GLM, SL2 uses DL, and SL3 uses GBM. The mix of models used for each super learner is based on the performance of the candidate models.

5.4 Numerical Results and Discussion

In this section, the numerical results of the experiments are presented to compare the performance of the proposed super learners against the four base learners Logistic Regression, Random Forest, Gradient Boosting Machine, and Deep Learning. This is done by the above introduced four evaluation metrics Accuracy, AUC, F-score, and LogLoss, implemented on 3 datasets. The Accuracy and F-score are reported at a 0.5 threshold level.

A complete overview of numerical test results in terms of the previously defined performance measures for each model and dataset can be found in table 10. The best performing base classifier (candidate model) and the strongest super learner are highlighted in black.

Table 10. Numerical results for each classifier and dataset

Dataset	Classifier	Candidate	AUC	Accuracy	F-score	Logloss
	LR		0.712	0.671	0.653	0.623
	RF		0.769	0.703	0.680	0.577
	GBM		0.775	0.716	0.694	0.570
Taiwan	DL		0.759	0.693	0.716	0.600
	SL1: GLM	DL, GBM	0.779	0.721	0.713	0.566
	SL2: DL	DL, GBM	0.780	0.726	0.708	0.563
	SL3: GBM	DL, GBM, RF	0.776	0.711	0.694	0.569
	LR		0.944	0.877	0.882	0.415
	RF		0.900	0.800	0.800	0.484
	GBM		0.886	0.776	0.776	0.426
Germany	**DL**		0.946	0.815	0.793	0.506
	SL1: GLM	DL, LR	0.946	0.846	0.839	0.470
	SL2: DL	DL, LR	0.947	0.846	0.839	0.360
	SL3: GBM	DL, GBM, RF	0.886	0.754	0.750	0.446
	LR		0.970	0.882	0.871	0.257
	RF		0.981	0.926	0.912	0.247
	GBM		0.987	0.941	0.933	0.246
Australia	DL		0.980	0.897	0.885	0.256
	SL1: GLM	DL, GBM	0.988	0.912	0.900	0.202
	SL2: DL	DL, GBM	0.989	0.853	0.783	0.386
	SL3: GBM	DL, GBM	0.987	0.941	0.926	0.198

The performance of each classifier is depended on the characteristics of the underlying dataset. For the Taiwanese and Australian dataset, GBM has the upper hand in terms of prediction power followed by RF (tree-based methods). In the case of the German dataset, LR and DL reach the highest accuracy (those results are in line with the earlier findings in chapter 3). During model training, all combinations of candidate models were tested. The performance of the proposed super learners varies strongly based on the choice of the underlying classifier. A super learner containing all four classifiers performed weaker on all three datasets compared to a super learner that combines only the two respectively three strongest classifiers available. Only the combinations with the highest accuracy in terms of the performance measures are presented in table 10.

Also, the configuration of the super learner resulting in the highest overall performance is different for each dataset. SL2 for the Taiwanese dataset and SL3 for the Australian dataset were both trained with DL and GBM. SL2 for the German dataset was instead trained with LR and DL. Note that, even though RF is the stronger base classifier in both cases (Australia and

Taiwan) DL turned out to be the better fit for the combination with the GBM classifier. One explanation for why the combination DL and GBM result in a better super learner compared to RF and GBM is the difference of the models itself. GBM and RF are both tree-based algorithms and hence quite similar, while DL can add new information. It seems that the candidate models (base learners) to be trained by the meta-algorithm to become a super learner need to be sufficiently diverse and the findings in this experiment suggest that the different properties of DL and GBM fulfill those requirements.

Table 11 shows the best performing base classifier, the best performing super learner, and the performance differences via the delta based on the evaluation metrics.

Table 11. Comparison of the best baseline model with the best super learner for each dataset

Dataset	Classifier	AUC	Accuracy	F-score	Logloss
Taiwan	GBM	0.775	0.716	0.694	0.570
	SL2 - DL	0.780	0.726	0.708	0.563
	Delta Δ	0.005	0.010	0.014	-0.007
Germany	DL	0.946	0.815	0.793	0.506
	SL2 - DL	0.947	0.846	0.839	0.360
	Delta Δ	0.001	0.031	0.046	-0.146
Australia	GBM	0.987	0.941	0.933	0.246
	SL3 - GBM	0.987	0.941	0.926	0.198
	Delta Δ	0.000	0.000	-0.007	-0.048

The super learner based on DL in the case of the Taiwanese dataset is superior to the best performing base learner GBM. The AUC delta is 0.005, the Accuracy delta is 0.010, the F-score delta is 0.014, and the LogLoss delta is -0.007. A similar outcome can be observed for the German dataset. The super learner could achieve an easily observable and significant edge in performance. The AUC delta is 0.001, the Accuracy delta is 0.031, the F-score delta is 0.046, and the LogLoss delta is -0.146. The performance of the super learners based on table 10 for the Australian dataset is not as obvious. Not a single super learner was able to outperform the best performing base classifier on all evaluation metrics. However, the super learner based on GBM reached the highest prediction accuracy for the Australian dataset with DL and GBM as candidate models and seems to be the best choice. The rationale for this conclusion is the following. There is no delta for AUC and Accuracy, which makes the F-score and the LogLoss the deciding factors. The F-score is with -0.007 slightly lower for the super learner, which is not optimal, but the LogLoss difference of -0.048 is very good and indicates a reliable and robust classifier.

Interesting to see at this point are the different outcomes for all super learners. The meta-learner algorithm had a significant impact on the performance measured by the four evaluation metrics. The super learners using DL as meta-algorithm tends to have the best performance, but this is not guaranteed as the outcome for the Australian dataset shows. So, the concrete choice of the meta-algorithm has to be evaluated for each situation and dataset individually.

What are the implications of those findings w.r.t. model choice? Based on those results the candidate model with the highest prediction accuracy should be chosen first, followed by either the classifier with the second-highest prediction accuracy or by a model that offers the biggest fundamental difference compared to the first choice (e.g. DL and GBM). Other classifiers can be included in the model mix as well to test whether a higher accuracy can be achieved, but this needs to be tested individually. Mixing all existing classifiers, especially models that are weaker than the already included predictors tend to dilute the model mix, which leads to lower performance and does not help to reach higher accuracy levels. Overall, regardless of the outcome of the base learners, it is possible to improve upon the base model performance. This was observed across all three datasets.

Overall, three things need to be considered to reach maximum prediction accuracy in terms of the four evaluation measures AUC, Accuracy, F-score, and LogLoss:

(1) It is necessary to focus on the initial training of the base classifiers as the performance of the super learner is highly dependent on a strong selection of base models. Since the outcome of the final super learner is highly dependent on the base classifiers it is important to push the base learners towards the best possible performance during the initial training process. To reach the highest prediction accuracy the strongest base classifiers need to be taken into account while ignoring weak models as they will only help to dilute the performance of the final classifier ensemble.

(2) The choice of the meta-algorithm has a significant impact on the final performance of the classifier. There is no single configuration of the stacked ensemble learner to reach maximum performance that is the same for the three different datasets. Overall, the usage of DL as a meta-algorithm has shown the best performance in this study.

(3) Using all pre-trained candidate classifiers does not result in higher accuracy. It dilutes the model mix and reduces performance. Instead, it is recommended to choose only a subset of the best performing candidate models available.

5.5 Conclusion and Future Outlook

The current lending market is shaped by technology giants, disruptive startups, and global competition. While FinTech firms are unlikely to lead to the distinction of financial institutions as predicted several years ago, it will lead to significant changes in financial market structure. As we move towards mostly automated digital financial markets, which will foster financial inclusion and sustainable balanced development on a global scale (Hudon, Labie, Szafarz, & Venet, 2019), the deployment of advanced analytics in the form of machine learning has become a necessity to survive in this new environment. Overall, the financial markets shift towards complete digitalization has just begun, but is poised to develop and finally be completed in the coming years. Within this climate, the need for data-driven decision making in the form of predictive analytics has become increasingly important. One area where predictive analytics has been applied excessively is consumer lending and credit risk management in general. The major purpose of predictive analytics here is to calculate the probability of default of a counterparty.

It has been proven that even tiny improvements in prediction accuracy can result in increased business values for lending corporations (Hand & Henley, 1997). Especially when those models are employed at scale on strong platforms. This is even truer today as globalization drives increasingly economies of scale and large portfolios of customers are the norm. This could lead to leaders take-all situations, where a de-facto monopoly has the potential to take large portions of the market in case it can maintain a technological advantage over its peers.

In search of maximum prediction accuracy this study has shown that combining different candidate models to a stronger classifier ensemble – a so-called super learner – is a potent strategy to improve upon the model performance of already existing single classifiers or tree-based ensembles as boosting and bagging. The candidate models used were logistic regression, random forest, gradient boosting machine, and deep learning. Three different real-world credit scoring datasets were employed in this experiment.

Similar to the findings in chapter 3, the results suggest that the performance of the candidate ML classifiers is dependent on the underlying dataset. Also, and this could be shown in all three cases, it was possible to improve upon the performance of those candidate models by combining them via the ensemble method stacking to a so-called super learner. As a general rule the following steps can be regarded as a reasonable guideline to achieve good prediction results: (1) Choose classifiers that have already a high prediction accuracy; (2) Choose classifiers that are sufficiently distinct from each other to provide additional information not

present in the other classifier (e.g. DL and GBM). (3) Avoid weak classifiers as they only tend to dilute the performance of the stacking ensemble.

Overall, it was shown that regardless of the outcome of the base learners, it is possible to improve upon the base model performance. This was observed across all three datasets. Whether the performance improvement justifies the computational complexity that comes with creating a classifier ensemble depends on the degree of improvement and the concrete business case.

ML-based lending will enhance the overall retail banking system by improving efficiency and effectiveness, improving scale and scope of lending as well as its fairness and hence foster financial inclusion of low income / developing country workers and will also shift the customer base towards FinTech utilizing corporations. Overall, FinTech based lending as well as underwriting is already a reality and causes significant shifts in the consumer lending market. Several studies have indicated that the trend described here will continue in the coming years, and further AI-based disruption of the lending market are yet to come. Also, this is not purely lending specifically. These kinds of digital disruptions are currently occurring on a global scale and can be observed in every industry. When it comes to the underlying principles of financial intermediation nothing has changed over the years. Technology just makes things often a bit easier for everyone. The developments will be positive and steer the financial markets towards more inclusion, fairness, and sustainable balanced development across the globe (Hudon et al., 2019).

6 Automated Machine Learning in Business Analytics

Abstract

The realization that data-driven decision-making is indispensable in today's fast-paced and ultra-competitive marketplace has raised interest in industrial machine learning (ML) applications significantly. The current demand for analytics experts vastly exceeds the supply. One solution to this problem is to increase the user-friendliness of ML frameworks to make them more accessible for the non-expert. Automated machine learning (AutoML) is an attempt to solve the problem of expertise by providing fully automated off the shelf solutions for model choice and hyperparameter tuning. This chapter analyzes the potential of AutoML for applications within business analytics, which could help to increase the adoption rate of ML across all business functions. The H2O AutoML framework was benchmarked against a manually tuned model on three real-world datasets to test its performance, robustness, and reliability. The used AutoML framework trains several base learners and combines them via ensemble learning to a stacked super learner. The manually tuned model could reach a performance advantage on all three case studies used in the experiment. Nevertheless, the H2O AutoML package proved to be quite potent. It is fast, easy to use, and delivers reliable results, which come close to a professionally tuned ML model. The H2O AutoML framework in its current capacity is a valuable tool to support fast prototyping with the potential to shorten development and deployment cycles. It can also bridge the existing gap between supply and demand for ML experts and is a big step towards fully automated decisions for business analytics functions.

Keywords: AutoML, Business Analytics, Predictive Analytics, Data-Driven Decision Making, Digital Transformation

6.1 Introduction

The era of struggle towards a modern enterprise has been termed digital transformation. "Digital transformation is concerned with the changes digital technologies can bring about in a company's business model, which results in changed products or organizational structures or in the automation of processes" (Hess et al., 2016). It occurs in response to changes in digital technologies and increasing digital competition, which changes customer behavior and expectations (Verhoef et al., 2019). The major advantage of the last wave of digital transformation was the buildup of a robust infrastructure, which can be built upon to employ new AI-related techniques (Bughin et al., 2017). The new wave of digital transformation will

foster disruptive forces that will exceed the last wave, which was known as "Big Data" revolution and characterized by storing data sets that are larger, more complex, and unstructured in nature, compared to older relational databases systems (Baesens et al., 2016; Chen et al., 2012; Henke et al., 2016).

The increased relevance of information technology and analytics for businesses in every industry makes the separation of IT and business strategy no longer viable and a "fusion" of both into the term "Digital Strategy" has been suggested (Bharadwaj, El Sawy, Pavlou, & Venkatraman, 2013). The realization that data-driven decisions have the potential to drive performance directly impacting the bottom-line of corporations (Brynjolfsson et al., 2011; Brynjolfsson & Mcelheran, 2019) led to the renaissance of business analytics research (Davenport, 2018). Data-driven decision-making is indispensable in today's global, fast-paced, and ultra-competitive market. All major industries and sciences have started to pick up on these developments. Predictive analytics is one of the major dominos to facilitate this new way of decision making. It is part of the complete business analytics chain, which is a complex process involving descriptive, diagnostic, predictive, and prescriptive tools (Delen & Ram, 2018).

The necessity to adopt sophisticated predictive models to make intelligent decisions is without question, but the ability to capture value through analytics is heavily dependent on employees with the required skill-set to leverage those analytics capabilities (Grover et al., 2018). Even though initiatives towards data science education have started to manifest itself (Clayton & Clopton, 2019), the huge demand for talent that makes sense of data and provides useful insights remains tremendous. The use of non-experts when it comes to ML algorithms is problematic as extensive knowledge is required to successfully tune ML models.

Automated machine learning solutions called "AutoML" have started to gain traction, which is a method to automatically tune and compare different algorithms to find the best hyperparameter combination (Feurer et al., 2015). The preceding task of pre-processing and feature engineering of the dataset is only partly supported (Balaji & Allen, 2018), but the end goal of AutoML research is focused on automating the complete predictive modeling process. AutoML could help to fill the existing supply and demand gap when it comes to ML experts. It has also the potential to democratize ML across less quantitative academic disciplines and functional business areas to foster the creation of new research questions and business use cases. Several different AutoML solutions were introduced during the last years. The major goal of the literature review was to choose the best suitable open-source AutoML framework for this study.

Gijsbers et al. (2019) offer an up-to-date comparison of the most mature open-source AutoML frameworks currently available: Auto-WEKA, auto-sklearn, TPOT, and H2O AutoML. The research itself is open-source and accessible online. It receives also regular updates upon the release of a new version. H2O AutoML is one of the top-performing models in this study.

Truong et al. (2019) analyzed the existing body of AutoML frameworks in terms of robustness and reliability taking into account a vast list of open-source and commercialized AutoML solutions. While there is no clear winner across all test cases, H2O managed to outperform all other models for regression and classification tasks.

This chapter will zoom in on the predictive part of business analytics and analyses whether AutoML solutions can enhance the adoption rate of ML across business functions. Based on the literature review is the H2O AutoML framework the best choice for classification tasks and is hence the go-to framework for the following empirical study.

The objective of this study is to test whether the AutoML off the shelve frameworks have a similar performance and/or can beat manually trained ML models. This is important to further drive the adoption of ML solutions across business functions and domains as deep technical knowledge to develop new ML and DL models will require significant theoretical and technical training often not present in corporations. AutoML could speed up the development cycle and counteract the current skill shortage within the area and is the first step towards a full end-to-end decision engine for business analytics.

The H2O AutoML framework is benchmarked against a manually created ML model to compare predictive ability, robustness, and ease of use. Also, these findings will be used to discuss managerial implications for digital strategy. At last, a roadmap for future research will be presented. Overall, the goal is to expand the discussion in the hope to trigger new conversations, and ultimately convince more researchers to think about how to incorporate ML models within business processes.

The rest of the chapter is organized as follows. Section 2 introduces the AutoML framework used. Section 3 describes the general experimental setup. Section 4 describes the outcome of the experiment and presents the performance of the H2O AutoML framework against the manually adjusted ML models. Section 5 discusses the numerical results, managerial implications for digital strategy, and derives future research possibilities based on the findings in this study. Section 6 concludes with a summary.

6.2 AutoML

Automated Machine Learning or AutoML is a method for automating the predictive analytics workflow. Depending on the concrete AutoML solution it might contain preprocessing, feature engineering, as well as model tuning. The current body of AutoML solutions does not handle the pre-processing very well (Truong et al., 2019) and the primary goal of this study is an assessment of the hyperparameter optimization and model choice. Based on the existing literature, is the H2O AutoML framework one of the most mature AutoML solutions currently available and achieves superior performance on classification and regression tasks according to several recent benchmark studies (Gijsbers et al., 2019; Truong et al., 2019). How does it work? See figure 21 below.

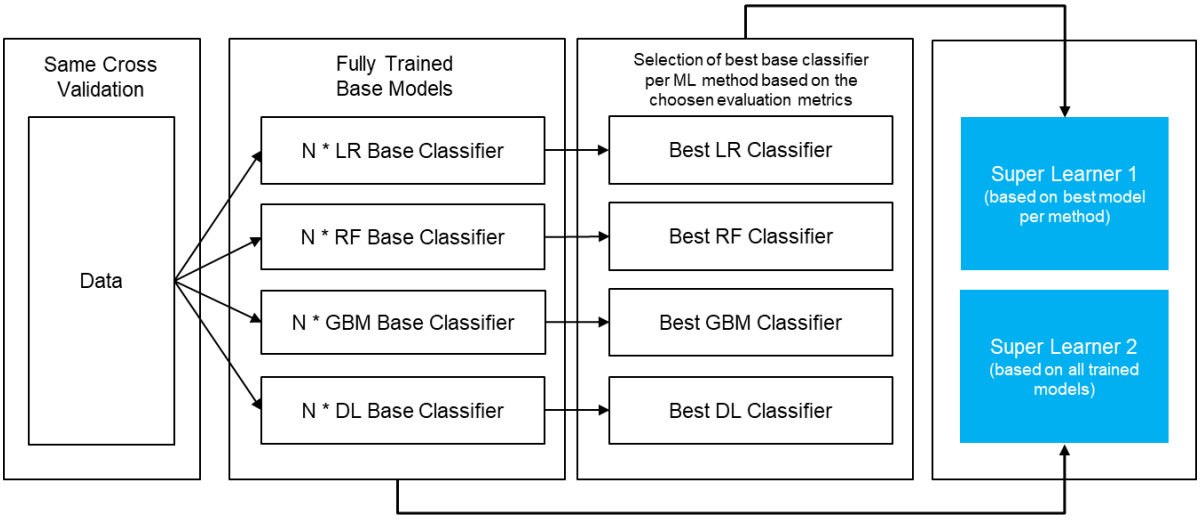

Figure 21. The H2O AutoML framework trains several base learners and in a subsequent step combines those to two different super learners. One super learner is based in all previously trained classifiers, the other takes only into account the best classifier of each ML family (LR, RF, GBM, DL). H2O AutoML automatically ranks the outcomes based on the chosen evaluation metrics.

H2O's AutoML framework (H2O.ai, 2019) creates different candidate models as GLM's, Random Forest, Gradient Boosting, and Deep Learning during an initial training phase and creates via stacking two different super learners. One super learner is based on all the pre-trained candidate models while the other is only an aggregation of the best model out of each family. The major parameters required for the AutoML solution are the *feature columns x*, the *response column y*, the *training_frame*, and the *validation_frame*. Also, the parameters *max_models* and *max_runtime_secs* are used to either specify the maximum number of models trained or the maximum time allowed for the process of model optimization. The H2O AutoML framework uses a random search as the optimization method.

Algorithm 1 Pseudocode for Automated Machine Learning (AutoML)

Input: labeled test dataset Dt, labeled training dataset D1, number of cross-validation sets k, time to completion t**,** choice of meta-learner algorithm M

Step 1: Train Logistic Regression Classifier

Step 2: Train Deep Learning Classifier

Step 3: Train Gradient Boosting Machine Classifier

Step 4: Train Random Forest Classifier

Step 5: Use all pre-trained base classifiers to create super learner 1

Step 6: Use only the best classifier per category to create super learner 2

Step 7: Repeat steps 1-5 until the maximum number of models or time specified has been reached

Output: A list of classifiers build during the run-time in descending order based on their prediction accuracy on the test dataset Dt

6.3 Experimental Design

The primary goal of this study is to benchmark the H2O AutoML framework against a manually trained super-learning ensemble.

Data: The experimental study uses the same three real-world datasets as chapter 4 (credit risk, insurance claims, and marketing). See section 4.2.2 for a description of the datasets and also the preprocessing steps. The chose split is 80:20, which means 80% of the dataset will be used in the training process and the remaining 20% will be used to test the generalization ability of the trained classifier. The same 5-fold cross-validation setup is necessary for the later fusion of the base learners into the meta learner. Hence, during the model training, 80% of the dataset will be split into different training and validation sets, which is done by cross-validation.

Evaluation: The four evaluation methods AUC, Accuracy, F-score, and LogLoss as described in section 2.3.5 will be used to assess the performance.

Setup: The inner workings of the AutoML solution offered by H2O train the 4 base-classifier Generalized Linear Model (LR), Random Forest, Gradient Boosting Machine, and Deep Feedforward Neural Networks. In a subsequent step, it applies the ensemble method stacking to fuse all of those pre-trained candidate models to a super learner to increase the accuracy levels. The best model is automatically selected based on a chosen evaluation measure. To test the strength of this setup I have recreated the inner workings of the H2O AutoML solution by manually training the base-models and combining them via stacking to a super learner.

Overall, the comparisons are between two separately configured super learners. One automatically generated by the H2O AutoML solutions and one manually tuned and configured similarly as already described in the earlier section 5.2 about super-learning/stacking itself.

6.4 Numerical Results

In this section, the experimental results are presented. The manually tuned stacked ensemble learner is compared against the AutoML solution from H2O. Why stacking? This was necessary to recreate the inner workings of the H2O AutoML procedures, which relies on training several different base classifiers, and then combines those pre-trained models for the final ensemble model based on stacking. Three real-world case studies from the business functions credit risk, insurance claims, and marketing were used in this experiment. The four evaluation matrices AUC, Accuracy, F-score, and LogLoss were used to benchmark the H2O AutoML solution against a manually optimized super learner. The Accuracy and F-score are reported at a 0.5 threshold level. The experiment was structured as follows:

In the first step the three baseline models Random Forest, Gradient Boosting Machine, and Deep Learning were carefully trained. To tune the hyperparameter settings of the base models' traditional methods as grid search and random search over a pre-defined range of parameters, as well as manual adjustments, were used during the training process.

Table 12 shows the numerical results for the base classifiers for each dataset. Gradient Boosting obtained the highest overall performance, followed by Random Forest. Deep Learning has the lowest performance scores. This is consistent across all three datasets.

Table 12. Numerical results of optimized base classifiers for all three case studies

Case Study	Method	AUC	Accuracy	F-score	Logloss
Credit Risk	Random Forest	0.769	0.708	0.683	0.574
	Gradient Boosting	0.775	0.716	0.694	0.570
	Deep Learning	0.758	0.703	0.686	0.609
Insurance Claims	Random Forest	0.636	0.598	0.584	0.667
	Gradient Boosting	0.640	0.598	0.586	0.663
	Deep Learning	0.633	0.597	0.534	0.669
Marketing and Sales	Random Forest	0.940	0.877	0.885	0.318
	Gradient Boosting	0.940	0.878	0.886	0.299
	Deep Learning	0.933	0.864	0.871	0.322

In the second step, the candidate models were combined to a so-called super learner via the ensemble method stacking that has been proven to deliver asymptotically optimal

improvements upon a set of base classifiers. For each case study, all three base models (RF, GBM, DL) were used to create the super learner. All three combinations of the baseline models for the stacked ensemble were tested and the best performance could be achieved by using RF, GBM, and DL as input for the super learner for all three case studies. This is not always the case.

In the last step, the stacked super learner created in step two serves as a benchmark for the AutoML solution from H2O to evaluate its performance, robustness, and reliability. Table 13 shows the final comparison of the H2O AutoML solution and the trained super learner.

Table 13. Comparisons of the super learner benchmark model and AutoML for all three case studies

Case Study	Method	AUC	Accuracy	F-score	Logloss
Credit Risk	Stacked Ensemble	0.778	0.717	0.698	0.565
	AutoML	0.776	0.714	0.695	0.567
Insurance Claims	Stacked Ensemble	0.642	0.603	0.592	0.662
	AutoML	0.640	0.599	0.590	0.663
Marketing and Sales	Stacked Ensemble	0.944	0.883	0.889	0.299
	AutoML	0.942	0.884	0.891	0.300

Overall, the results are surprisingly consistent and the stacked Super Learner was able to outperform the AutoML model on all three datasets with an AUC difference of 0.002.

While performance deltas for the other matrices are not identical, the stacked ensemble outperformed the AutoML solution here as well in most cases. For the credit risk case study, the difference is 0.003 for Accuracy, 0.003 for F-score, and 0.002 for LogLoss. The performance difference in the case of the insurance dataset is 0.004 for Accuracy, 0.002 for F-score, and 0.001 for LogLoss. The performance difference for the marketing case study is -0.001 for Accuracy, -0.002 for F-score, and 0.001 for LogLoss. AutoML slightly outperformed the stacked ensemble only on the marketing case study in terms of Accuracy and F-score. Overall, the manually tuned stacked ensemble shows superior performance compared to the AutoML solution for all three case studies.

6.5 Discussion

The purpose of the experimental study presented in this chapter was to test the performance of the H2O AutoML framework compared to a manually tuned ML model in terms of the four evaluation measures AUC, Accuracy, F-score, and LogLoss. This section has three parts: First, the results of the empirical study will be discussed to assess the overall performance of the tested AutoML solution. Second, the findings will be discussed w.r.t. to business analytics

to better understand the managerial implications for digital strategy. And last, a roadmap for future research is provided.

6.5.1 Discussion of Results

Based on the findings of the empirical analysis, which is based on three real-world case studies from credit risk, insurance, and marketing, the H2O AutoML model was not able to outperform the manually tuned classifier.

The AutoML package has difficulties to reach the quality of a manual setup in two ways: (1) The underlying models do not reach the same prediction accuracy as the manually tuned versions. Increasing the running time did not result in a performance improvement either. This was tested on the smaller datasets as credit risk and marketing and higher running time did not have a significant impact on the final output. (2) The H2O AutoML package chooses two stacked ensemble combinations. One based on all the trained models and the other based on the best model for each category. It does not test whether another combination of the candidate models (e.g. smaller subset) results in better performance. This is important as adding weaker models to the pool of models for the stacked ensemble unnecessarily sabotages the performance. Guo et al. (2019) demonstrated that only the best baseline models should be taken into account for the super learner and additional classifiers tend to dilute the performance by adding non-optimal information that results in a reduction of prediction accuracy. The results in chapter 5 are also in line with those findings.

However, the performance delta is not very strong and the AutoML solution provided by H2O is a potent model tuning engine that can significantly speed up prototyping or help practitioners less familiar with ML concepts to set up a powerful model. Nevertheless, for maximum prediction accuracy, careful model tuning and adjustments of hyperparameters done by a data scientist result in the best performance. Based on the small performance improvement it is questionable whether the small edge of manual adjustment as demonstrated by three case studies justifies the time-consuming model creation process when almost the same can be achieved with no knowledge and adjustment efforts. The answer to this question is mainly depended on the use case at hand, and whether a tiny performance improvement justifies the additional time required for manual model tuning. Also, given the strong performance of the AutoML solution created by H2O, it is almost certain that further research will result in prediction accuracy levels that are on par with models adjusted by ML experts.

Overall, AutoML is an important first step towards complete end-to-end decision processes. Due to its relatively strong performance, and consistent results, AutoML has the potential to

become more capable as human engineers over time. This would significantly help to democratize ML for Business Analytics functions, especially for small to medium-sized businesses, which tend to have more difficulties to hire the appropriate talent.

6.5.2 Managerial Implications

Management has always used data to generate information for insights. Mainly in the form of business information systems. This is not new. What has changed is how we come up with a decision. The earlier more intuitive business approach gradually changed towards a more evidence-based or data-driven decision making (Brynjolfsson & Mcelheran, 2019; Delen & Ram, 2018).

This development is also reflected in the current skill demands across all industries (Clayton & Clopton, 2019). This hyperconnected and fast-paced business environment requires employees that are familiar with technology as a business enabler and value generator. The outdated view of IT as a pure cost function needs to be dropped. Older incumbent corporations are often still reluctant to change their mindset when it comes to this new reality.

Business Intelligence and Business Analytics are vital in today's world shaped by digital disruption and global competition. Business Analytics is about converting data to insights to improve management decisions across the complete corporate value chain. It traditionally used different analytic methods to transform data into digestible information to steer corporate decision making.

Companies should strive for digital maturity. This is necessary to remain competitive, but it is not easy for everyone. Not every industry can capture the value associated with superior analytics similar to another. Industries that have traditionally used analytics due to vast amounts of data as banking were able to adapt way faster to the current landscape, but still face significant competition from new startups (Chui et al., 2018). While some of those startups were able to sustain their independence and be listed on the big exchanges, others were swallowed during M&A strikes (Siebel, 2019). Newborn corporations that belong to the categorization of digital natives have a significant advantage as their foundation has been built to facilitate future employment of advanced analytic capabilities and is scalable. They also tend to be more attractive for the young and technology-savvy crowd (Henke et al., 2016).

AutoML might create a new level playing field by democratizing ML solutions across industries and business lines. Even though the findings in this study prove that AutoML does not yet beat careful human engineering when it comes to model tuning, it could help to support the adoption of ML solutions by helping to fill the talent gap. In addition, it is useful to support the skilled

data scientist with fast prototyping and benchmarking. AutoML can be used as a valuable tool to support the predictive modeling process by speeding up prototyping, which helps to accelerate the development cycle and final deployment.

The rapid development of cloud-based open-source and automated ML solutions will democratize the technology space. ML interfaces to TensorFlow like Keras (Falbel & Allaire, 2019) or ML frameworks as H2O (H2O.ai, 2019) helped to increase the level of abstraction and enabled users to better focus on the problem and solution. AutoML is the next step in this process towards higher adoption in the industry by completely automating several key processes in the predictive analytics chain. This will help to foster further adoption of ML in business units and hence accelerate the diffusion of this General-Purpose Technology (GPT) across the economy.

As these developments continue it is very likely that domain knowledge and subject matter expertise will be more important to develop and implement end-to-end AI solutions compared to expertise in machine learning. Agrawal et al. (2019) argue that domain expertise cannot be commoditized, but ML as a GPT can and will be commoditized in different ways. Furthermore, ML will get less cost-intensive over time due to continued innovation within the field itself as well as due to hardware improvements, better software, APIs, and UIs. It can also be perfectly delivered by using cloud solutions to leverage a centralized capable Machine Learning / AI platform (e.g. MS Azure, SAP Hana, etc.).

AutoML is a big first step and might gradually evolve and extend to a fully automated decision engine. Fully automated ML solutions pose the potential to democratize analytics across several industries and business functions but final realizations remain difficult. AutoML is still not able to automatically preprocess complex datasets, which is one of the most time-consuming steps in the data science process. The same is true for the need to move from pure predictive outputs to concrete actionable steps in the form of prescriptive analytics. Until the last steps towards a complete end-to-end process are not solved corporations need to rely on hiring data science experts or external consultants to help them drive the current digital transformation initiatives.

6.5.3 Future Research

Further research is required at both ends of the predictive analytics process. AutoML needs to be able to handle data preprocessing to further automate the ML pipeline. Also, at the end when it comes to deriving actual actions from those predictions there is room for improvement. See figure 22.

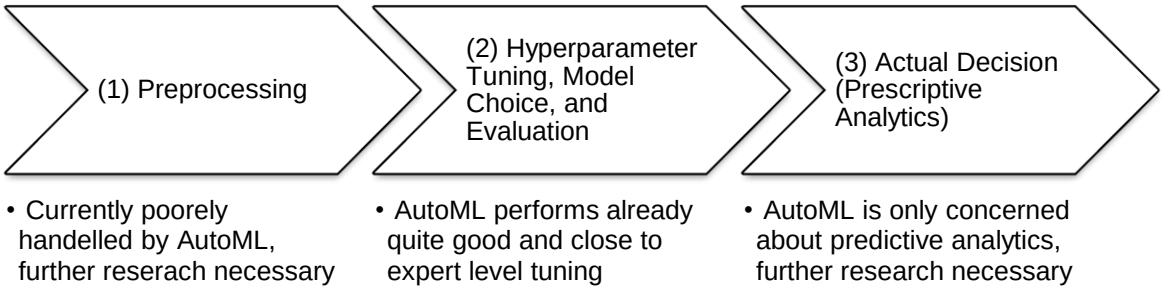

Figure 22. This graphic shows the current capabilities of AutoML and points towards further research necessary to completely automate the predictive analytics workflow to finalize the notion of complete off-the-shelf ML solutions for data-driven decision making.

Current research mainly focuses on predictive tasks and results have to be interpreted by human decision-makers. One of the most interesting questions in Business Analytics is how to move from predictive analytics to a complete end-to-end decision engine that provides managerial decision-makers with concrete actions that can be acted upon. So far, ML and also DL are predominately used for predictive analytics. There are already attempts to combine ML methods with Operations Research/Management science to move from pure prediction to actual decisions, but how to go from a good prediction to a good decision is poorly understood. The major problem is to account for uncertainty in the decision-making process (Bertsimas & Kallus, 2019). Looking at recent studies as Alphastar show that this is possible and that deep reinforcement learning (DRL) is able to reach human-level decision power in an uncertain environment, and this in real-time (Vinyals et al., 2019). Studies on DRL for prescriptive analytics and managerial decision-making in an uncertain environment do not yet exist and would open several new research questions within the field of data and management science. This is a domino that needs to fall to reach full end-to-end decision processes within business analytics.

6.6 Conclusion

The first wave of digital transformation was triggered by big data and is now gradually replaced by Machine Learning and Artificial Intelligence, which has become the new driving force behind the move towards a digital enterprise. This resulted in an increased demand for ML experts and a corresponding skill-shortage, which slowed down the adoption of ML methods for business analytics. AutoML frameworks are expected to be a solution for this current talent gap and could also accelerate the predictive analytic process. It was demonstrated that the H2O AutoML framework in its current capacity does not reach the full prediction accuracy that is possible by careful manual adjustment of the models for two reasons. First, it does not reach

a higher accuracy on the baseline or candidate models than the benchmark model. Second, it has a fixed way to combine the baseline models with ensemble learning. To reach maximum performance an optimal combination of the baseline models has to be determined in each individual case. Despite those findings, this study has shown that AutoML can be a powerful tool. First, it can be used as a baseline during prototyping for ML experts, which can help to accelerate the development and deployment cycles of ML projects; second, it makes ML models more accessible to non-expert users as it further increases user-friendliness by moving the level of abstraction higher; and third, AutoML can be considered as a big step towards a full end-to-end decision engine, which is the ultimate goal of AI in Business Analytics.

7 Enterprise AI: Towards an End-to-End Data-Driven Decision Engine

Abstract

An end-to-end business analytics engine is essentially a comprehensive and automated ML pipeline. The major goal of this chapter is to synthesize the contributions of the preceding chapters into a coherent whole by proposing a complete ML-pipeline that consists of three distinct phases: Phase 1 - Data Preparation; Phase 2 - Model Tuning and Evaluation; and Phase 3 - Model Deployment and Monitoring. AutoML automates the second phase in the pipeline (model tuning and evaluation) and is a vital building block to automate the full pipeline. It is discussed how AutoML can be improved to reach state-of-the-art accuracy levels to fulfill its purpose as the heart of the pipeline. Alternatively, it can be used in its current form. However, to achieve an end-to-end prediction engine for data-driven decision-making extensions towards Phase 1 and 3 are required. Data preparation, which consists of several adjustments as cleaning and feature engineering are not yet automated. Also, there is no consideration of real-world constraints (size, speed, interpretability), and the model choice is purely based on prediction accuracy. Due to the lack of those functionalities, automated monitoring and adjustments are not possible. Those gaps result in clear future research directions which are also discussed in this chapter.

Keywords: Machine Learning Pipeline, Artificial Intelligence, Business Analytics, Data-Driven Decisions, Enterprise AI

7.1 Introduction

The preceding parts have made it clear that business analytics has become mandatory for all industries. The need for data-driven decision making in corporations has developed into a critical necessity to survive in the economy of the 21st century (Siebel, 2019).

The initial chapters 3 and 4 were mainly concerned with finding the strongest prediction models for supervised learning on different structured data sets, while also discussing real-world constraints. Then, chapter 5 introduced and discussed super learning as a reliable method to enhance the performance of those models. And the previous chapter 6 “AutoML in Business Analytics” has introduced the notion of an automated predictive analytics process that does not require any human input during the model tuning and selection process.

This chapter uses the above findings to discuss the status quo in business analytics when it comes to a completely automated decision engine for data-driven decision making. Also, it will be discussed how ML – in theory – can be automated and applied at scale, including ideas on how to translate that silo wise prove of concepts into scalable solutions in an enterprise.

Due to the recent development of AutoML models (Halvari, Nurminen, & Mikkonen, 2020), the idea of a completely automated ML end-to-end solution is within our reach. However, literature regarding a complete end-to-end ML pipeline for business analytics seems not to exist, but several authors are suggesting steps towards this direction (Agrawal et al., 2019; Thomas, 2019). However, Google has recently - March 2020 - announced a beta version of its cloud-based AI Platform Pipeline, which is a very promising project and perfectly captures the idea of a cloud-based scalable end-to-end ML solution. It offers different functionalities as data ingestion, data preparation, feature engineering support, model tuning and evaluation, and deployment (Ramesh & Unruh, 2020).

The remainder of this chapter is structured as follows: The first section briefly introduces the notion of an ML pipeline to kick-start the discussion. Section two discusses AutoML as the heart of the pipeline as it automates the model tuning and evaluation process. This is done by synthesizing the ideas of chapters 3 to 6. Third, necessary AutoML extensions will be discussed that are required to further progress to complete automation. Fourth, we will discuss how one could evaluate such an ML pipeline. Fifth, I will explain how AutoML models can be applied in its current capacity, go over challenges, and limitations as the issue of explain-ability. At last, I will point towards future research to fill the necessary gaps to reach the initial idea of a completely automated prediction engine that does not require human input. Gaps are prediction accuracy (described in earlier findings) for phase two, and also missing research regarding the automation of phase one and phase 3.

7.2 Proposed Pipeline Setup: 3 Phases

The question remains how it will be possible to fully/seamlessly integrate a complex machine learning end-to-end decision system within the enterprise architecture to capture the value of data-driven decision making to its full extent.

This will be a complex and interdisciplinary undertaking requiring the effort of a huge team of experts across business lines and technical specialists responsible for providing domain expertise to provide the functional specification, requirement engineering, prototyping, large-

scale development, and final deployment and maintenance (Agrawal et al., 2019; Bughin et al., 2017).

An end-to-end business analytics engine is essentially a comprehensive and automated ML pipeline, which consists of three distinct phases: Phase 1 - Data Preparation; Phase 2 - Model Tuning and Evaluation; and Phase 3 - Model Deployment and Monitoring. See figure 23.

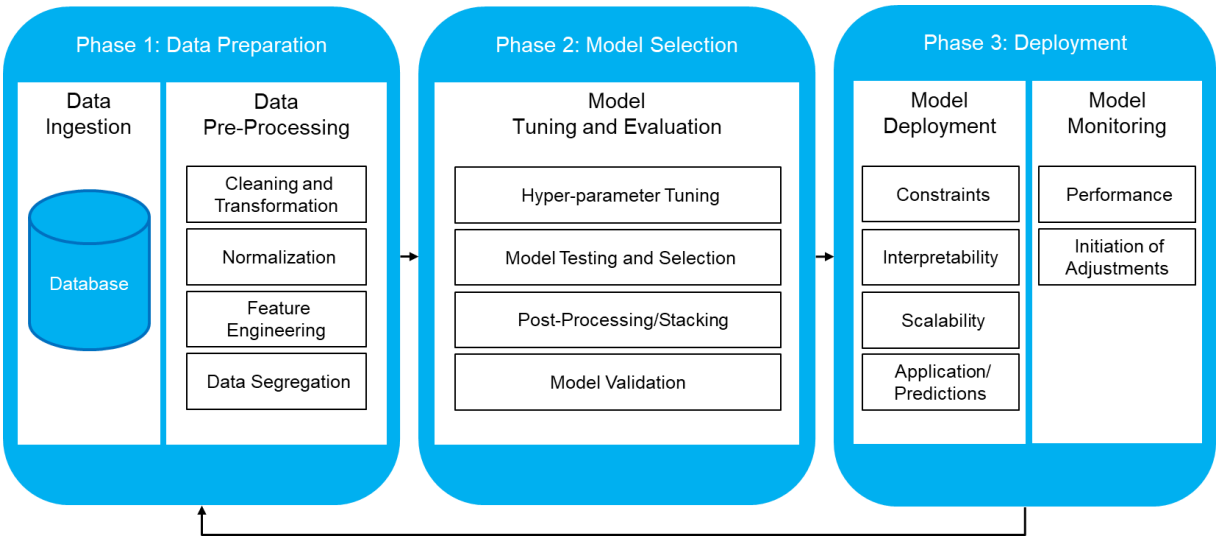

Figure 23. The proposed ML pipeline consists of the three phases Data Preparation, Model Selection, and Model Deployment & Monitoring. AutoML sits at the heart of the ML pipeline and is mainly responsible for model tuning and evaluation. However, if we talk about reaching a fully automated business analytics engine for decision making, AutoML needs to extend its capacities towards phases 1 and 3.

An alternative to complete automation is a human in the loop approach, referred to as augmentation (Davenport, 2018). Due to several current limitations, augmentation is likely to precede the full automation of the ML pipeline (Thomas, 2019). See section 7.5.4 and 7.6 for a discussion on limitations and future research.

7.3 The Heart of the Pipeline - Phase 2

AutoML automates the second phase in the pipeline (model tuning and evaluation) and is a vital building block for a full ML pipeline. It is not enough to just automate the process, but also assure that the performance of the final model choice can be considered state-of-the-art. The following considerations should therefore to be taken into account.

7.3.1 Candidate Models

As the most important part of every ML task is selecting the right candidate model(s), we will start by discussing the findings regarding optimal model choice. In the earlier chapters' different

single classifiers and ensembles thereof were tested for accuracy on several distinct datasets. Based on those earlier findings it is difficult to determine a single candidate model that outshines all the others as model performance tends to depend on the underlying characteristics of the data itself. Overall, the earlier results suggest that Gradient Boosting and Random Forest or more general tree-based ensembles tend to be the best choice for classification problems based on structured data. Deep Learning with a feed-forward architecture turned out to be a powerful model as well and takes the second place. Manual model configuration and parameter tuning were necessary to optimize those ML classifiers to optimally represent the dataset. Also, the model selection process was purely manual after the training phase. This goes against the notion of an off-the-shelf solution for business analytics. See Chapters 3 and 4 for more details.

As shown in chapter 6 AutoML solves this problem by automating the tuning and model selection phase. It has been shown that H2O AutoML (H2O.ai, 2019) is a reliable tool that offers quite good performance, but it does currently only use a random grid search during model optimization, which does not guarantee maximum performance. The only constraint at this point would be the run-time. Given enough time AutoML could test every possible configuration of the setup to assure the selection of the best base-learner parameterization possible to reach maximum prediction accuracy for a given candidate model. However, the time requirement is not very serious and H2O AutoML offers fast convergence of approximately 15 minutes to reach optimal performance (Truong et al., 2019).

Overall, H2O AutoML is fast and can adequately train and identify the best performing ML model for each dataset during the tuning and selection process. However, there is a further possibility to optimize the predictive modeling process in terms of classification accuracy by utilizing the ensemble method stacking.

7.3.2 Post-Processing via Stacking

The findings presented in chapter 5 clearly showed that the different characteristics of GBM and DL can be utilized to improve prediction accuracy beyond the individual algorithms by combining them via the ensemble method stacking (Van Der Laan et al., 2007). It is possible to combine both (or several) of those base models to leverage the fundamental difference of how they represent the dataset, which opens a consistent way to improve upon the base classifiers and guarantees a stronger predictor. The results were consistent across all different test cases, which makes stacking a reliable method to post-process fully trained predictors to improve upon their accuracy levels. Further research taking into account new datasets could either strengthen those findings or weaken it.

The only limitation here was that the concrete selection of the base classifiers cannot be generalized according to the results and is based on the initial training and performance of the base learners, the specific combination of those base learners, and the concrete meta-algorithm applied during the fusion to the super learner. See the discussion in chapter 5 for more details.
Those things naturally require time, which is often a major constraint in a fast-paced business environment and the need for fast prototyping that leads to fast deployment of new solutions or adjustments of already implemented ones can be a major competitive advantage.
Overall, stacking is a reliable performance enhancer, but whether it is a viable tool to use depends ultimately on the computational limitations and concrete business case.

7.3.3 AutoML

In chapter 6 automated machine learning was presented as a solution to streamline the predictive analytics workflow. Concrete, I benchmarked the H2O AutoML framework against a manual setup to test its utility for the ML pipeline. AutoML should be able to select the best candidate model and combine it via stacking into a more powerful super learner. H2O AutoML proved to be a robust and reliable automated ML solution for model tuning and selection that offers strong prediction performance.

However, it was not able to outperform the manual setup of training the base-learners individually, plus the combination with stacking to a super learning ensemble with different configurations. What is missing at this point is a mechanism to select the best stacking setup. To arrive at the maximum possible prediction performance that is possible based on the underlying dataset an optimal selection of the base-learners and meta-algorithm is necessary. As described in section 6.2.2 H2O AutoML trains the candidate models and uses stacking over the best classifiers from all different models to produce the super learner, which tends to produce suboptimal outcomes - see chapter 6 for further details.

The better AutoML can replicate careful manual tuning and final model selection the better. To achieve similar results as the manual setup AutoML needs to optimize the choice of base learners included in the fusion process and drop classifiers that would contribute to a reduction in performance. Guo et al. (2019) introduced an optimization setup for stacking, which chooses the best base models with the help of Bayesian optimization. Using an optimization algorithm as suggested by Guo et al. (2019) would solve this problem and there would be no reasons why AutoML cannot reach similar accuracy levels as a manually tuned version. All possibilities regarding hyperparameter optimization and stacking configuration would be reproduced by the AutoML setup.

Overall, AutoML in its current form is a powerful model that can be used in various business analytics use cases. If the goal is maximum accuracy it is recommended to used stacking in combination with AutoML as a benchmark to reach maximum accuracy while sacrificing transparency and speed. As the H2O AutoML does not select the best base-learning combination it is recommended to build the setup manually and use AutoML as a baseline to improve upon. This will result in superior performance as the current AutoML solution is already quite strong. This setup is simple to execute and results in state-of-the-art prediction results for binary classification problems within business analytics.

The H2O AutoML framework in its current capacity is a valuable tool to support fast prototyping with the potential to shorten development and deployment cycles. It can also bridge the existing gap between supply and demand for ML experts and is a big step towards fully automated decisions for business analytics functions. However, AutoML needs slight improvements to guarantee state-of-the-art performance. Also, it is necessary to extend its capabilities towards phase 1 and phase 3 of the ML pipeline to resemble the complete ML workflow.

7.4 Necessary AutoML Extensions

If the goal of AutoML is complete automation in the form of a fully automated end-to-end ML workflow, several extensions are required. Data acquisition and pre-processing, as well as deployment and monitoring, are important parts of a complete ML pipeline. If we want to fully automate the ML workflow, we need to extend the capabilities of AutoML towards phase 1 and phase 3.

7.4.1 Data Preparation – Phase 1

The preparation of input data, sometimes called data wrangling is an important step (Konstantinou et al., 2019). AutoML sits at the heart of the ML pipeline and is mainly responsible for model tuning and evaluation and – as mentioned in Chapter 6 – data pre-processing is only partially provided by current AutoML solutions. The necessity for excessive pre-processing and careful model tuning of those more complex models goes against the notion of an off-the-shelf solution, which is often favored in real-world business analytics and data mining scenarios. Off-the-shelf prediction models can be directly applied to a problem scenario without any significant domain knowledge or pre-processing of the datasets (Hastie et al., 2017).

Pre-processing steps and data-preparation become even more important when we talk about a centralized Business Analytics core system, which needs to be able to identify data schemas

and possible use cases automatically. Otherwise, a human user has to tell the system via a frontend input the necessary information to configure the ML setup.

Another problem is feature engineering. Feature engineering seems to be often the deciding factor why manual model tuning is largely superior to off-the-self-solutions (Thomas, 2019). This will still be the case, even though all the recommended steps mentioned above are contained in an AutoML solution. If we want to reach a fully automated business analytics engine for decision making (full automation) AutoML needs to extend its capacities towards phase 1, which is a comprehensive and automated preparation of the dataset.

7.4.2 Deployment and Monitoring - Phase 3

First, the primary focus of AutoML is the prediction accuracy of the classifiers. It does not account for any real-world constraints at all. To make a reasonable final model suggestion by itself it is necessary to introduce this functionality in AutoML solutions. AutoML needs to be able to determine an adequate tradeoff between prediction accuracy and real-world constraints given the requirements of a concrete use case/application scenario.

Second, based on the results of chapters 3 and 4 adjustments of the model choice or configuration can become necessary. The findings are clearly in favor of ensemble learning methods as GBM when compared to Deep Learning, but those results could not be generalized, which suggests that the underlying structure of a given dataset is important and only slight variations might result in the need for a different model or at least in a need to change the model configuration. Machine learning remains an empirical process and it is necessary to test various models for different datasets to find the one model that can best represent the information contained within the data. It is also advisable to reevaluate parameter settings and model choice after a non-trivial change in the fundamental dataset has occurred as this might result in different requirements w.r.t. model configurations. This might be the case due to new external input data or due to newly generated data by the ML model itself. See section 7.6.4 for further research directions.

Hence, AutoML is required to build upon its existing capabilities and extend its functionalities towards phase 1 and phase 3 of the above introduced ML pipeline to reach the state of a largely autonomous system.

7.5 Further considerations

7.5.1 Pipeline Evaluation

The evaluation of the proposed automated ML pipeline cannot be done with a single measure. Rather, it should be divided into an objective measure (or measures) for each of the individual phases described in figure 25.

- **Phase 1 – Data Ingestion and Preparation:** The ability of the model to accurately identify the data schema and the appropriateness of the pre-processing applied based on this analysis.

- **Phase 2 – Model Tuning and Selection:** The major part is tuning, validating, and selecting an ML model or ensemble thereof. This step should be based on the usual ML performance metrics as AUC, Accuracy, F-score, and LogLoss. At this point, an array of the best performing models should be given to phase 3 to select them taking into account real-world constraints.
- **Phase 3 – Deployment and Monitoring:** As this phase is mainly concerned with the successful deployment of the model in a real-world scenario, additional constraints have to be taken into account. This is especially important as the preceding model selection was only based on prediction accuracy while ignoring other variables.

As a benchmark one could build a manual setup and compare the quality of the input data after preprocessing, the prediction accuracy of the selected model, and the external requirements and constraints required for deployment.

7.5.2 H2O Driverless AI

Open source tools as H2O AutoML focus mainly on the second step of the ML pipeline, which is hyperparameter optimization and model selection, combined with a combination via stacking to a super learner ensemble. To resemble a full ML workflow, an extension at both ends to phase 1 and 2 is necessary. Given the current importance of evidence-based decision making, it is almost certain that this visible limitation of open-source AutoML solutions will soon be filled. In contrast, commercial tools as H2O Driverless AI extend the pipeline functionality slightly and offer pre-processing steps as recognition of the data scheme as well as feature engineering. Differences also exist regarding hyperparameter search. It seems that the slightly more powerful commercialized version H2O Driverless AI has been optimized with a combination of random search and Bayesian optimization, whereas the open-source version is limited to a random search for parameter tuning. Using the overview of (Gijsbers et al., 2019) there is not

a single open-source AutoML solution currently available that uses a combination of random search and Bayesian optimization. H2O Driverless AI was not part of the evaluation, but it is relevant to mention that steps in this direction are already fruitful. The information was taken out of the official documentation on the website and was not verified in this study (H2O, 2020).

7.5.3 Centralized BA Solutions (Cloud, SaaS)

AutoML is a recent development but vital for further democratization of ML and also necessary for the roll-out and diffusion across business functions and the economy. As a prerequisite though, robust enterprise architecture is necessary to facilitate efforts towards advanced analytic capabilities (Henke et al., 2016). Research is aware of the current challenges and importance to be agile and adaptable in this new digital age and has realized correctly that the cloud is one of the most important areas for the future digital society (Siebel, 2019). Transforming GPT technologies like AI will only prosper at scale when the necessary distribution channels are present (Agrawal et al., 2019). Also, the need for a consistent stream of quality data needs to be assured to make the most of Machine Learning and AI (Bughin et al., 2017; Henke et al., 2016).

Fully automated ML pipelines can be offered as versatile cloud solutions and hence as software as a service (SaaS). Once everything is done (pre-processing, model tuning, and model selection) the best model would be automatically returned for the final deployment. However, the produced model can also live in the cloud, which would completely outsource the ML pipeline to an external provider of analytics or a centralized in-house team. Whether this utility is acquired from an external service provider or built inhouse depends on the strategic importance of AI for the organization and also its current architecture and internal capabilities.

The only sensible approach for long-term business solutions is cloud-based. Not only offers this the possibility to always use a centralized hardware center that offers the best processing power, but it also centralizes maintenance, updates, and monitoring of such a system (Siebel, 2019). A perfect AutoML solution could ultimately replace many data science and ML experts, especially since SaaS cloud solutions could be offered at scale and with a central data science steering committee that maintains this prediction engine.

7.5.4 The Black-box Challenge

There is one constraint that is dominating all the others when it comes to the adoption of ML models in practice. It is the possibility to explain the reasoning behind the predictions. Decision trees or standard linear models do not come close to artificial neural networks or ensembles as bagging and boosting when it comes to prediction accuracy. Nevertheless, they belong still

to the favorite tools in business analytics departments. Model interpretability is often more important than pure predictive ability when it comes to real-world applications and simple models (e.g. decision trees, linear models) fulfill those requirements. In contrast, Deep Learning and Gradient Boosting are more or less black-box models that are not explainable and hence difficult to communicate. As discussed in chapter 4 the black-box nature of especially the most powerful models tends to prevent ML applications to further penetrate specific industries – especially if they are subject to regulatory scrutiny. Risk Management and Insurance – among others – are fields where model interpretability and the need for causal explanations are dominating pure predictive ability.

There is a fundamental difference between causal explanation (understanding) and empirical prediction (practicality). I highly recommend the paper *To explain or to predict?* (Shmueli, 2010). It is somewhat dated, but perfectly captures the essence of those competing philosophies better than any other resource. Both approaches can benefit from each other and a gradual convergence of the two approaches seems natural over time. Empirical prediction is useful to validate the strength of theoretical assumptions, while causal explanation helps us to understand why predictive models arrive at a certain decision.

Nevertheless, there are also reasons why a complete paradigm shift towards explainable ML is questionable. Agrawal et al. (2019) are convinced that the major reason why ML has been able to develop faster and more powerful predictive models than statistics is the pure focus on empirical testing rather than time-consuming causal explanation. Overall, simple empirical prediction without the need for understanding or an explanation is easier than establishing a causal relationship that can be explained. The need for model understanding in fields as economics and business is important, but a shift away from pure prediction will only slowdown further progress. The best would be a gradual improvement of the explain-ability of black-box models, while research about pure predictive power is completely separate. In business, this problem could be solved by fast-paced prototyping and empirical testing in one department to arrive at an immediate decision, including another department responsible for ex-ante explanation and model validation.

Rudin (2019) argues that we should avoid using certain Black-Box models in high-stakes industries and brings as example finance. This is a general statement and while it is true for certain parts of the financial industry, we cannot generalize this observation across the whole sector. We have to distinguish between different areas of finance. While certain areas in financial services are subject to regulatory supervision are highly constraint when it comes to the deployment of black-box models, other fields as trading have a natural inclination to use

anything under the sun to maximize their business objectives. Quantitative Hedge Funds as Bridgewater associates are at the forefront of applying state-of-the-art quantitative models of all kinds with the ultimate goal of improving the ROI of their trading activities.

The asset management industry and all kind of facilitating intermediaries as custodians, brokers, exchanges, etc. traditionally have been very tech-oriented businesses and where always at the front of technological incorporation of the newest models and algorithms to improve efficiency and effectiveness across their functions with the ultimate goal of superior customer satisfaction.

The major issue with interpretability is its nature as a hard constraint. Having the constraint of only explainable models would severely limit the range of model choice of AutoML, which makes the model tuning in phase 2 pointless. Overall, even though a detailed understanding of predictions might not be necessary in all cases it is without question that progress in this area would significantly accelerate the adoption rate of ML models in business, which ultimately would lead to a more widespread diffusion across the economy. See section 7.6.3 for a discussion on further research regarding the explain-ability problematic.

7.6 Future Research Directions

The coming years will utilize the tools that have enabled the global digital transformation of all major industries and will further help to develop the building blocks necessary to reach full automation. Nevertheless, the transition from a research environment to real-world applications poses several challenges. The requirement to deploy ML models at scale with immense amounts of data in a complex corporate setting makes the application of ML for business analytics a complex undertaking. This section briefly discusses the current limitations and constraints of AutoML/Machine Learning and points towards future research directions. Solving those last issues will help AutoML solutions progress towards and finally resemble a complete end-to-end ML pipeline for real-world business analytics applications.

7.6.1 Data Preparation

One of the major problems in supervised predictive analytics is the handling of the initial data. The first step is often regarded as the most important aspect of data science and takes up to 80 or 90 percent of the required time. It consists of different kinds of data cleaning and adjustments, but also of transforming the feature space. Especially the fusion of existing features into more diagnostically conclusive (informative) features seems to be the most important aspect to squeeze out the last bit of performance. Online data science competitions

as Kaggle are often won by superior feature engineering, which indicates that this initial step can be more important than the model tuning process itself, especially when we have already reached the maximum amount of information that can be extracted based on the current characteristics of a given dataset.

Research w.r.t to automated feature engineering exists (Nargesian, Samulowitz, Khurana, Khalil, & Turaga, 2017), however, machines do not possess the expertise of a domain expert and can often not semantically categorize related features. This poses the question whether it is necessary to completely automate this part, or whether human expert input at this stage is something to be desired as it allows for better control and feedback loops, which also helps to better understand ML models and their predictions. Konstantinou et al. (2019) for example, propose a data preparation architecture that automatically transforms the input based on an initial outcome description by the data scientist. Especially since augmentation instead of full automation seems to be the next frontier of digital transformation and AI initiatives in business analytics (Davenport, 2018) integrating domain expert knowledge at this initial step could turn out to be the most reasonable choice.

7.6.2 Real-World Constraints

The primary focus of AutoML is on maximum performance measured by a chosen evaluation metric such as AUC, Accuracy, F-score, or LogLoss. Applying predictive analytics in a real-world setting poses additional challenges and one downside of AutoML in its current form is its exclusive focus on prediction accuracy based on one/several of those evaluation measures. A pure focus on performance and ignoring practical constrains is not optimal (Thomas, 2019). And indeed, most research (including this thesis) sees maximum accuracy as the single most important factor in ML. However, there exist many real-world constraints that AutoML does not account for at the moment and which are crucial for the final deployment in industry. For example, small performance improvements in terms of prediction accuracy do not necessarily justify the selection of a model as this often comes with higher computational resource requirements. This is especially relevant for real-time decision purposes as high-frequency trading and similar applications. However, this needs to be evaluated on a case by case basis taking into account the specific business requirements.

Ultimately, especially if we want to achieve a 100% automated engine when it comes to model training and selection, it is for AutoML to be more flexible regarding hidden constraints as model size, computational complexity (speed), sparseness, and interpretability. Again, the most reasonable choice here could be to include a domain expert in the process to specify the necessary constraints and let the model use an optimization algorithm to balance the trade-off

between pure prediction accuracy (phase 2) and real-world constraints (phase 3). Overall, AutoML needs to become more flexible and take into account the concrete use-case requirements and deployment scenarios during phase 3 of the ML pipeline.

7.6.3 Interpretability – Blackbox

One of the most pressing challenges in data science and business analytics is the problematic of the explain-ability of complex ML models. Without the necessary trust, it becomes difficult to deploy and hence use ML models in practice. As discussed in chapter 4 one of the major reasons why advanced ML (especially Deep Learning) has not found its way into business analytics functions is their lack of transparency and interpretability. The black-box nature of strong algorithms as DL and GBM makes it difficult to communicate results to other stakeholders. In some cases - especially in industries subject to regulatory scrutiny - makes the deployment of black-box models outright illegal and hence impossible.

Breakthroughs w.r.t. to interpretability and transparency issues would significantly help to increase the adoption of ML models for business analytics functions and hence foster faster diffusion across the economy. Research in this area has experienced increased momentum due to the rising interest and therefore funding from different stakeholders.

Samek & Müller (2019) and Montavon, Samek, & Müller (2018) give a good overview of current attempts to solve the black-box issue. Also, research about model interpretability has already translated into software tools such as LIME to solve the black-box issue of certain ML models. LIME, which promises to “explain the predictions of any classifier” is an available solution to tackle the black-box issue of ML (Ribeiro, Singh, & Guestrin, 2016).

Which solution ultimately wins remains to be seen. Further progress in this area is posed to emerge due to the current interest levels from important stakeholders around the world and we will hopefully see further breakthroughs here in the coming years.

7.6.4 Monitoring and Adjustments

The actual deployment of AutoML solutions started to become easier with commercialized tools as H2O Driverless AI, Google’s Cloud AutoML, etc. However, a complete ML pipeline needs to include continuous monitoring of the deployed ML models. As ML models introduce new data to the existing database, which ultimately will be used as input to train the model, the usage of an ML model itself could drive the need for an adjustment and retraining requirements of an already deployed model. ML predictors need to be assessed and changed if the data distribution (fundamental characteristic of the underlying dataset) has been changed. To reach

a completely automated ML pipeline this monitoring and adjustment in phase 3 need to be incorporated in the model. First approaches have already emerged. For example, Wilson et al. (2020) offer the first solution to this problem with a model that offers continuous self-evaluation and adaptations to new data, and the resulting concept drifts over time.

7.7 Conclusion

This chapter introduced the notion of a completely automated ML pipeline as a necessity to reach an end-to-end prediction engine for business analytics and it was shown that we are very close to a complete end-to-end predictive model that can reach state-of-the-art performance without human input. Fully automated ML pipelines could be offered as versatile cloud solutions (SaaS). Whether this utility is acquired from an external service provider or built inhouse depends on the concrete value for the business.

AutoML in its current capacity can sit at the heart of the pipeline and carry out model tuning and evaluation. However, a full ML pipeline for business analytics needs also to be able to pre-process data during phase one and to account for real-world constraints during model deployment. Also, continuous model monitoring would be required as adjustments to model configurations are necessary due to shifting data sets. Current problems are lacking functionality at the beginning and end of the pipeline. Data preparation is largely ignored by open-source AutoML and manual adjustments are required. Also, at the end of the pipeline, there is no appreciation of real-world constraints or steps for further monitoring and model adjustments. Currently, the main focus of model selection is based on prediction performance, usually taking into account only one performance measure. Computational speed, model sparseness, and interpretability might be more important for certain business use-case than a tiny improvement in classification accuracy.

Filling those gaps and combining all those building blocks would result in a complete stand-alone prediction engine that is capable of reaching state-of-the-art accuracy levels without human input. An automated predictive analytics model would be a significant step towards complete end-to-end data-driven decision processes. The ability to automate mental tasks allows us to move from standard robotic process automation based on a chain of if-then decisions to digital robotic process automation, which has the potential to replace or at least augment white color jobs in several industries.

One problem that any business faces, when it comes to the adoption of complex machine learning, especially deep neural networks, is their black-box nature - non-explain-ability - that

makes it difficult to communicate the inner workings of the model. This is especially relevant as many real-world use cases require an understanding of the underlying logic of an applied method. If this prerequisite is not given, despite the predictive superiority - it will be impossible to justify commercial usage and deployment due to regulatory issues. Given the significant importance of AI for the real world the black-box nature has become one of the most important issues when it comes to the deployment of ML models.

However, based on the importance of several stakeholders across domains research regarding the issues of interpretability, transparency, and explainability has gained momentum. Given the current interest of different stakeholders, we hopefully see further progress regarding a better understanding black-box model soon. Further research here would be vital for the democratization of ML and also necessary for the roll-out and diffusion across business functions and the economy.

8 Stakeholder Implications

This part will discuss the relevance and implications of the above findings for different stakeholders. At the 2019 World AI Conference in China Elon Musk compared humanity with a boot-loader. The minimum effort that is required to start a computer, because it cannot start itself. He implied that humanity is the boot-loader for AI and once it's running, we are not required anymore. The founder of Alibaba Jack Ma rolled his eyes, replied, and suggested: "Ok, let's talk about something fun."

The reality is that forecasting the future has always been difficult, especially when it comes to developments as AI in today's complex world. In a way, computers are already more capable of processing certain information and data as humans but are limited to operate in a defined environment – completely incapable of breaking out. Examples of these agents are current gaming-related implementations as AlphaGo Zero (Silver et al., 2017), AlphaStar (Vinyals et al., 2019), or AI bots for Dota 2 (Katona et al., 2019).

That the current state-of-the-art of our predictive models does not resemble true AI is out of the question, but given the speed of technological development over the last years, it might very well be the foundation of the next generation of AI that comes very close to general AI, but for now, all we should be concerned with are the more realistic applications of AI and how we can utilize them to transform our society in the post-digital era towards a smart and automated economy. The following part discusses the managerial as well as economic implications of machine learning and AI.

8.1 Managerial Implications of AI

This part focuses on the managerial implications that can be derived from the earlier findings. It has been shown throughout the thesis that there are several application areas for machine learning in financial services and insurance and it was possible to further increase prediction accuracy by utilizing state-of-the-art machine learning models which indicates the possibility to capture value through machine learning across several business functions.

It was shown that it is still possible to increase the accuracy levels of machine learning for credit risk, insurance, and marketing in the case of structured datasets by using gradient boosting machine (chapters 3 and 4). Also, it is possible to optimize accuracy levels further by creating a super learner (chapter 5). Given the huge volume of those markets, even tiny accuracy improvements can lead to competitive advantage through a more accurate

categorization of bad and good customers or more effective demographic targeting. This chapter will discuss these use-cases in more detail. Due to the increasing volume of available data ML applications are likely to become even more over time.

Example: The ability to translate a stream of customer data into valuable business information will further increase in importance. A new form of banking called open-banking, which uses open APIs and a digital ecosystem to connect different market participants will give everyone increased access to data. Hence, the ability to use that information will be the key to a customer-centric digital business strategy. Customers' expectations are changing rapidly and the old way of doing banking and insurance will soon be obsolete. Currently, only the older generations are still visiting physical bank branches, while internet banking is already standard, and the complete shift to mobile banking is the next step.

Machine learning can help companies/departments to capture business value, but how can this be achieved? This part has three themes which will be discussed. (1) The first part focuses on how exactly machine learning can be leveraged and applied in the three functions presented in this thesis, namely, credit risk management, insurance claims prediction, and sales and marketing, including possible synergies. (2) The second part focuses on the value of a centralized business analytics function and complete automation. (3) Third, the impact of ML on operational task automation and strategic decision making is discussed.

8.1.1 Counterparty Risk

The first and also most detailed use case presented is the assessment of the credit worthiness of a customer. Due to the scope of the thesis, which is binary classification on structured data, the above cases are based on a binary credit score, which was defined as good or bad. However, in practice it is often better to use the probability of default directly. Both concepts have a different outcome when modelled directly, but are strongly related. The credit score or category (classification) can be calculated based on a probability of default (percentage prediction) including a decision threshold.

Counterparty credit risk is one of the largest risk classes and highly relevant for corporations. However, the risk assessment of counterparties is twofold.

The use cases presented are mainly concerned with FinTech lending, which means the credit assessment and direct decision whether a loan will be issued. Machine learning can help us to make the initial decisions whether to engage in a business activity/transaction with a counterparty in the first place.

Example: Many FinTech companies (e.g. Peer-to-Peer Lenders) do assess their customer's creditworthiness with ML models. Whether we call the outcome a credit score (category/classification) or a probability (percentage) is irrelevant for the logic behind this argument. In the end, we have to determine a concrete threshold level which disqualifies a counterparty.

The idea is to only target customers, which are - based on our classification outcomes - not likely to result in a default over the lifetime of the loan. In case of insurance policy, we would assess the likelihood for a policyholder to file an insurance claim during the lifetime of the contract. Both scenarios are fundamentally the same. We do not want a negative event to take place and ML helps us to identify customer segments, which have a high likelihood to cause such an event. Once this segment has been identified we can further focus on a customer demography that is most likely to execute a final purchase (see marketing analytics).

However, PDs are also used to calculate potential losses based on a credit portfolio to determine the capital requirements of a company. Capital requirements can be based on regulatory requirements or self-directed/imposed risk management. This is necessary to shield financial institutions from unexpected market events. The necessity to have a certain capital buffer to avoid bankruptcy during volatile periods is based on regulatory requirements. Banks and insurance companies pool risks in a portfolio and hold specific capital requirements. The need for concentration risk reserves is necessary as larger portfolio defaults are always possible due to systemic risks that cannot be captured by individual risk assessments as shown in the use cases above. Higher debt increases the overall ROI in case the total return on capital is higher than the cost of debt. This so-called leverage effect increases the incentive for companies to increase their debt financing.

Correct assessment of the probability of default is one ingredient to achieve an optimal capital buffer, meaning the tradeoff between equity and debt in a corporation – a so-called capital structure puzzle from a corporate finance perspective. Increasing debt levels increase the risk of default in unexpected credit events, especially if they are subject to concentration risk. An accurate assessment of the PD has the potential to affect the capital requirements calculation for a company positively. Nevertheless, the likelihood of catastrophic events or tail risks in insurance and banking is first of all low, and second, it cannot be avoided by any company. However, these facts do not undermine the value of individual risk assessments for a specific counterparty as quality customers who achieve a high credit score and hence a low probability of default tends to be more resilient in unexpected tail risk events.

While advanced machine learning is in theory the optimal tool to support credit risk management in assessing the default probability of a counterparty it does have limitations in practice. First, while the models presented in this thesis have superior accuracy compared to more traditional assessments certain limitations prevent them to shine in the industry. The biggest barrier are regulatory bodies, which require quantitative models to be explainable. The general black-box nature of Deep Learning and Tree-Based Ensembles as Gradient Boosting makes those models unusable for many problems. Hence, most financial institutions subject to regulatory supervision cannot use these black-box models in many areas and therefore stick to more traditional approaches, which are very well known by the regulators.

Second, the credit score or the concrete probability of default are both parts of credit risk management, but they are not the only important variables. Credit risk is mainly concerned with calculating the expected loss (EL), which is the product of the probability of default (PD), the loss given default (LGD), and the exposure at default (EAD). See equation 9.

$$EL = PD * LGD * EAD \tag{9}$$

In case guaranteed collateral is present during a transactional agreement with a counterparty and this collateral covers a hundred percent of the potential exposure (EAD = 0), the PD in itself is not relevant for the expected loss (EL) calculation and capital requirements.

However, ML can be used for internal risk assessments that are not subject to regulatory supervision, and as we will see in the next section it is the perfect tool for digital marketing.

8.1.2 Marketing Analytics

Another use case for AI, which was also presented earlier is predictive support for marketing and sales teams. The actual distribution channels for companies are of significant importance to bring the product to the market and ML has proven to be a valuable tool to improve actual conversion rates by estimating the probability of how likely a direct marketing effort translates into an actual purchase/business transaction. The AUC and Accuracy scores for the sales and marketing dataset are extremely high and show that ML can significantly support the sales process.

Taking into account demographic data from customers to increase sales seems to be one of the best use cases for ML. Reducing the number of unsuccessful advertisement campaigns can have a significant impact on ads budget especially for social media advertisements as measured by clicks, views, or engagements, which should ultimately translate into a purchase

- high conversion rates. Companies using those models can reduce their advertising and marketing costs significantly as demographics which on average result in a negative return on investment (ROI) can be dropped. It is possible to run a targeted campaign by only showing ads to customer segments that exhibit the highest conversion rates. Machine learning makes it possible to spend the existing advertising budget in a targeted way and hence most effectively. It is also easier to monitor the performance and actual outcome of an advertising campaign or a direct marketing effort. Effective advertising efforts directly translate into bottom-line results for the company and can be monitored in real-time or only with a slight delay. In contrast, the risk is more difficult to determine as default rates are a lacking indicator that only materializes once a customer in fact defaults on a loan or files an insurance claim, which could lie years in the future.

It has been shown that ML can add value for the individual business functions, but the above-introduced cases do not have to live in isolation. Combining the information gained from either credit risk or insurance claims assessment with the marketing and sales data allows determining specific demographics that offer low default rates (low business risk) and high conversions rates and up- and cross-sell possibilities (high business value). Using machine learning/advanced analytics in this way results in a strong competitive advantage due to significant cost reductions. In this case, the earlier discussed issue of the black-box nature is not relevant as sales activities and demographic targeting is not subject to regulations. This is an area where artificial intelligence and machine learning can be applied to its maximum value immediately and only the concrete outcome of a specific campaign is relevant. This can easily be captured by immediate KPIs as initial conversion rates and up- and cross-sell monitoring per customer. However, the adoption of those marketing analytics capabilities is still in its infancy as analyzed by various recent studies (Miklosik, Kuchta, Evans, & Zak, 2019).

8.1.3 Centralized and Automated

Overall, there are two possibilities to use those methods. The first and easier way is to remain at a functional level and develop the necessary tools only for a specific department. If the strategy here is aligned across the company, cross-functional collaboration would still be possible, but the knowledge and data transfer would be more manual and not automated. Increased automation of those advanced analytics capabilities would increase velocity and flexibility, which would make it easier to translate this enhanced prediction and decision power into tangible business value. It would also reduce adjustment times as analytic capabilities are centralized and maintained by an expert team or a hired service provider (see section 7.5.3). Also, an enterprise-wide solution would make it possible to capture interdependencies between

the different results (see figure 1 in section 1.4). An example of interdependencies between different departments is the risk assessment of customers (counterparty risk) and the willingness or likelihood of a customer to buy a certain product (conversion rate). Combining that information is necessary to achieve the maximum results from analytics efforts, which requires cross-functional execution instead of independent silo-wise utilization of those models. The best way to future proof an organization in the long run and to remain competitive is to develop integrated and automated analytics capabilities that operate in a cross-functional manner.

8.1.4 Strategic vs Operational Automation

Earlier general-purpose technologies (GPT) had a huge impact on productivity and tended to change the nature of how we do business. They changed the economic supply chain horizontally as well as vertically and brought rapid innovation and disruption of job markets due to new skill requirements. This is the same for general-purpose machine learning, but there is a slight distinction. Most earlier breakthroughs technologies tended to automate operational tasks using robotic process automation, which consists in its essence of simply adding "if-then" decisions after each other to automate existing and always recurring business processes that are necessary for the core business to be carried out consistently. This time is slightly different. This is the first time in the history of management science that the possibility to achieve automated corporate decision making due to technological advances is within our reach. It has been suggested several times in the past that technology could directly make strategy decisions, but this has never turned out to be true. Peter Ducker wrote several books on the topic of infusing management decisions with technology or completely automating them. While IT and technological progress had a huge impact on operational management and business processes the impact on strategic management (board-room) decisions have so far been limited or non-existent. ML significantly increases the scope of automation possibilities across business functions. So-called white color jobs fall under the umbrella of mental tasks and will gradually be taken over from ML-based systems.

AI systems are only possible because of their constituents, which are ultimately single ML models. Machine learning routines are mainly used for predicting a certain output Y based on a data Input X. The model is trained on past data to predict the future or outcome of future scenarios, but it is only extrapolating the past into the future. The fields of business analytics, business intelligence, and data mining mainly utilize machine learning algorithms for predictive analytics to facilitate correct business decisions, but this is not enough to reach full automation of many complex tasks. It is necessary to combine those standard machine learning models

into a larger system, where each of them is carrying out a standard prediction task. The ML pipeline suggested in chapter 6 is the first step towards further automation for business analytics. Once this is completed, it will be easier to design and develop add-ons that include additional steps as concrete actions that might be able to replace high-level managerial decisions. However, complex AI systems put together based on standalone ML models all handling different aspects of decision-making tasks with a fusion of the different outcomes to a whole similar to what deep mind has done for Go and StarCraft might turn out to be difficult. Game environments are inherently limited and are indeed a very small entity when compared with the realities taking place in the real world.

How far AI-enabled decision making will move up the corporate ladder remains to be seen. So far, it was not possible to help board room decisions with more than standard information as too many variables need to be taken into account, which is difficult to formalize. Dominos will all fall eventually, but every domino needs to fall for another one to continue the chain reaction. How do you eat an Elephant? One bite at a time. Most organizations will be able to slowly adapt by implementing those changes succinctly instead of immediately.

8.2 AI as General Purpose Technology

The field of economics has recently picked up on the strong developments of artificial intelligence. Agrawal et al. (2019) introduced AI as a new general-purpose technology (GPT), which will have a similar effect on the world economy as earlier breakthrough technologies like the steam engine, electricity, or the internet. As mentioned in the introduction – the characteristics of a GPT are according to Jovanovic & Rousseau (2005) pervasiveness, improvement, innovation. Mapping that to the current AI landscape – the major reasons why ML Systems can be considered as a GPT are the following (Agrawal et al., 2019):

- **External improvements:** The breakthrough of AI in recent years can largely be attributed to improvements in data availability, software, and hardware improvements. CPU and other hardware improvements will make it faster as well, it can be delivered through cloud architectures to serve a broader audience across all scientific fields. DL can be improved by new scientific breakthroughs within the field itself, by improved hardware, or by better data-sets, or simply higher amounts of data.
- **Self-improvement:** The very idea of ML is learning from data, which means it can improve upon itself. It can also improve itself by using a competing ML model, which gives AI superior improvement capabilities (Z. Wang, de Freitas, & Lanctot, 2016). A very good example is the current Alpha Go Zero, which keeps getting better by

continuously playing against itself, without the requirement of human knowledge (Silver et al., 2017).

- **Externalities:** Another characteristic of a GPT is the effect it is having on the vertical supply chain within an industry. It will affect all levels within this vertical chain and force suppliers as well as customer-facing units at the end of the supply chain to adapt to this new technology. Examples are Nvidia, Google, and Tesla. All develop graphics processing units (GPUs) specifically for the parallel processing requirements of DL.
- **Geopolitical Power Dynamics:** Leadership in artificial intelligence might lead to a shift in geopolitical power dynamics if the AI arms race keeps moving forward as it currently does. Several government agendas from the USA, China, and Europe indicate huge expectations regarding future AI applications.
- **Investments:** Also the investments from governments all around the world, unprecedented venture capital, and private equity fundings, and complete AI hubs all around the world with the specific goal to drive advancements in this field will make sure that research and progress in AI will remain stable for years to come.

For an exhaustive treatment of this topic, I highly recommend the book *The Economics of Artificial Intelligence: An Agenda* from Agrawal et al. (2019). It is the best currently available discussion on AI as a general-purpose technology and helps to understand the necessity of further research in this field.

The diffusion of AI/ML as a GPT across the economy – even though it is accelerating – has just begun and will take time. First of all, it is normal that it takes time for new cutting-edge research to be translated into business use cases. And second, history shows that the general nature of a new GPT is often not easy to grasp and a lack of understanding of AI as technology makes it difficult for business decision-makers to incorporate those models within the business process (Agrawal et al., 2019). Nevertheless, huge investments from governments, venture capital funds, and corporations across all industries are in place and will facilitate the adoption and integration of AI systems within the corporate architecture. The overall cash injection for this wave of AI is unprecedented and will lead with a high likelihood to further progress in the coming years (Bughin et al., 2019; US Government, 2019).

The next years will be dominated by further automation of business processes, which will keep the demand for highly skilled labor aka white color jobs very high. The complexity of this task, which requires business-understanding, model-understanding and data-understanding requires strong skills in data science and information technology, which has been mentioned

on several fronts to be lacking in most industries. Domain experts with several years' of experience as well as automation experts with advanced degrees in relevant disciplines and/or several years of work experience in a matching or closely related environment or domain will be required.

9 Conclusion

Advanced analytics and Machine Learning have found widespread applications across many business units and functions in all major industries. Within the research scope - binary classification on structured data - the overarching theme of this thesis was to prove that machine learning and business analytics can be used for data-driven decision making to capture value within financial services. Several distinct use-cases with real-world data sets have been utilized to test the hypothesis that ML can have a positive impact on enterprise value and competitive advantage.

Based on the datasets a tendency towards the best suitable models and configurations can be inferred, but a complete generalization cannot be justified. Therefore, further research should - if possible - include additional datasets, datasets with more observations, and datasets containing more divers' features to strengthen the findings of this thesis. Also, further studies could extend the scope and include multiclass classification and regression. Regression is relevant for financial time series data and different neural network architectures as recurrent neural networks and long short-term memory (LSTM) have shown to produce strong results.

Nevertheless, it was shown that ML increases prediction accuracy and has the potential to increase operational performance, which translates to direct value gains in terms of cost reduction or increased conversion rates, which both contribute positively to the value of any corporation.

This central conclusion was reached by answering more granular questions as:

- What is the best ML model for credit scoring?
- What are the reasons for the slow adoption of DL in BA?
- Is it possible to consistently improve upon the best ML classifiers by utilizing the fusion method stacking?
- Does AutoML reach the predictive strength of manually tuned classifiers by a human expert?

The findings to those answers were combined and translated into concrete stakeholder implications.

Concrete, to demonstrate the added value of artificial intelligence in business analytics various ML classifiers as Logistic Regression, Random Forest, Gradient Boosting, and Deep Learning were benchmarked on different business analytics use cases. The focus was on financial

services related applications in credit risk management, insurance claims prediction, and marketing and sales. The following results were reached.

Gradient Boosting and Deep Learning: Gradient boosting is the best performing binary classifier for structured data in the context of business analytics for most use cases and datasets. These first results were important, as the existing academic literature turned out to be quite contradictory when it comes to identifying the best performing classifiers. Many papers are wondering why there is a lack of adoption when it comes to DL. The main reasons usually stated are computational complexity, no existing big-data architecture, lack of transparency/black-box nature of DL, and skill shortage. However, DL does not - as widely assumed - offer any performance advantage when it comes to predictions based on structured data sets. This explains the current reluctance of established business analytics functions to replace working ML solutions with DL models that perform often weaker and exhibit other problems as a lack of speed and transparency. Nevertheless, DL offers proven advantages (e.g. unstructured data) and should be regarded as a valuable addition to the existing pool of ML models as it introduces more flexibility and a wider field of applications.

Super Learning: It is possible to improve upon the best performing binary ML classifiers currently available (also gradient boosting) by utilizing the fusion method stacking. Given that only tiny accuracy improvements can translate into a competitive advantage this method should be taken into account whenever accuracy is the major variable in a business use case.

However, stacking for performance enhancements introduces more complexity in the whole process, which makes it necessary to justifies the increase in accuracy for a specific use case when taking into account other real-world constraints as speed requirements, complexity, and model transparency, etc.

AutoML: AutoML proved to be a valuable tool with the potential to completely automate the predictive modeling process and comes very close to professionally tuned ML models. AutoML has the potential to solve the currently existing skill-shortage. Also, it can help experts with faster prototyping and can be considered a big step towards a full end-to-end decision engine, which is the ultimate goal of AI in business analytics.

ML-Pipeline: It was shown that an end-to-end business analytics engine can be created by integrating existing AutoML solution in a complete ML pipeline, which consists of the three distinct phases: (1) Data preparation, (2) model tuning and evaluation, and (3) model deployment and monitoring. AutoML automates the second phase in the pipeline (model tuning

and evaluation) and is a vital building block, but extensions towards phase 1 and 3 are required to complete the business analytics pipeline.

Business Value: ML can not only increase business value, but it will soon be mandatory if companies want to survive in the current global digital ecosystem. Increasingly fast-paced and global competition is not subject to local borders and can easily enter markets without infrastructure investments. Also, consumer attitude is changing rapidly and especially the younger generation expects financial services to offer similar products as they are used to from Big Tech companies as Apple, Amazon, and Facebook. This is only possible by leveraging AI-based analytics across the whole value chain in an automated a targeted way. The significant value of ML becomes clear during the discussion of the stakeholder implications and, surprisingly, diffusion does not occur at a faster pace.

The use cases demonstrated the vast array of application areas of machine learning in financial services. It can be used to assess the default probability of a counterparty either for the sales decision itself (mainly FinTech and InsurTech) or for the calculation of capital requirements. Using state-of-the-art ML is an absolute necessity to compete in the lending market among other FinTech companies. Also, for the capital requirements calculations is maximum accuracy helpful to find the best risk-return tradeoff possible and hence be able to achieve the highest ROI. Whether the capital requirements are based on regulatory supervision or self-imposed risk steering is here secondary.

Another application area is digital marketing. It is possible to run a targeted campaign by only showing ads to customer segments that exhibit the highest conversion rates. Machine learning makes it possible to spend the existing advertising budget in a targeted way and hence most effectively. It is also easier to monitor the performance and actual outcome of an advertising campaign or a direct marketing effort. Effective advertising efforts directly translate into bottom-line results for the company and can be monitored in real-time or only with a slight delay. Most marketing and sales activities, which focus on maximizing shareholder value allow the usage of AI to its maximum extent and the black-box nature of ML is mostly irrelevant.

Combining analytics capabilities across functions in a centralized unit would be best to leverage cross-functional synergies. Machine learning can reduce the overall risk of a company by avoiding unnecessary dangerous customer demographics, which would further optimize the advertising budget. For example, combining the risk assessment with the assessment of maximum conversion rates makes ML the perfect tool to maximize companies advertising budget in a targeted way by only focusing on low-risk demographics that are also highly likely

to engage with the product. It also offers the possibility to target customers, which are known to engage in additional buying activities (upsell and cross-sell), which can further increase revenue and hence ROI.

All those findings helped to draw a clear picture of the way business analytics will develop in the future. The new position of the Chief Digital Officer is becoming more relevant every year indicating the increasing importance of becoming a digital enterprise on a global scale. The coming years will utilize the correct tools that have enabled the global digital transformation of all major industries, governments, and societies, and will further help to develop the building blocks necessary to reach full automation. The evidence is conclusive and machine learning will be lined up behind other technologies like computers, electricity, the internet, etc. and will have a continuous effect on many aspects of our lives for years to come. ML as GPT has become a reality and will slowly diffuse across the economy by providing for the first time in human history the possibility to automate mental tasks. ML as a GPT for data-driven prediction will further find its way into business analytics and keep shaping the field. Improvements in parallel processing, network infrastructure, and distributed systems, and research to improve the underlying principles of ML models itself will only help to improve the performance and capabilities further over time. The concrete scale and scope of the impact of AI in business analytics for data-driven decision making depend on the concrete industry, but it will cause significant changes in the horizontal as well as vertical supply chains across the economy. The diffusion of AI as a GPT will take time as the history of GPT's shows, but the fire of digital transformation has been ignited and new technological disruptions along the way will keep the flame burning for years to come. The infrastructure has been built and in the post-digital era, it will not be a question of being digital anymore. Businesses reluctant to adopt a digital infrastructure will be wiped out of the current market and replaced by an agile and modern business that can continuously integrate advanced analytics and future technological progress into the existing infrastructure. Distributed ledgers are already on the rise and started to integrate with the existing DT technologies like big data, cloud, IoT, and AI. The current existing technologies are enough to completely restructure our world economy and society into a fully digital version of itself. The starting gun has been fired. Now, it is upon us to make the best out of it.

Tables and Figures

References

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, *6*(2), 1–20. https://doi.org/10.3390/risks6020038

Agrawal, A., Gans, J., & Goldfarb, A. (2019). *The Economics of Artificial Intelligence: An Agenda*. (A. Agrawal, J. Gans, & A. Goldfarb, Eds.). London: National Bureau of Economic Research.

Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., & Zhao, J. L. (2016). Transformational Issues of Big Data and Analytics in Networked Business. *Mis Quartely*, *40*(4), 807–818.

Balaji, A., & Allen, A. (2018). Benchmarking Automatic Machine Learning Frameworks. Retrieved from http://arxiv.org/abs/1808.06492

Bazarbash, M. (2019). *FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk*. *IMF Working Papers*. https://doi.org/10.5089/9781498314428.001

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Bertsimas, D., & Kallus, N. (2019). From Predictive to Prescriptive Analytics. *Management Science*. https://doi.org/10.1287/mnsc.2018.3253

Bertsimas, D., & King, A. (2016). OR Forum—An Algorithmic Approach to Linear Regression. *Operations Research*, *64*(1), 2–16. https://doi.org/10.1287/opre.2015.1436

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., & Venkatraman, N. (2013). Digital business strategy: Toward a next generation of insights. *MIS Quarterly*, *37*(2), 471–482. https://doi.org/10.25300/MISQ/2013/37

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Singapore: Springer Science+Business Media, LLC. https://doi.org/10.1117/1.2819119

Breiman, L. (1996a). Bagging predictors. *Machine Learning*. https://doi.org/10.1007/bf00058655

Breiman, L. (1996b). Stacked regressions. *Machine Learning*. https://doi.org/10.1007/bf00117832

Brynjolfsson, E., Hitt, L., & Kim, H. (2011). Strength in numbers: How does data-driven decision-making affect firm performance? *International Conference on Information Systems 2011, ICIS 2011*, *1*, 541–558. https://doi.org/10.2139/ssrn.1819486

Brynjolfsson, E., & Mcafee, A. (2016). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York, London: W. W. Norton & Company Ltd.

Brynjolfsson, E., & Mcelheran, K. (2019). *Data in Action: Data-Driven Decision Making and Predictive Analytics in U.S. Manufacturing (Working Paper)*. Retrieved from https://ssrn.com/abstract=3422397

Brynjolfsson, E., Rock, D., & Syverson, C. (2019). Artificial Intelligence and the Modern Productivity Paradox. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The Economics of Artificial Intelligence* (pp. 23–60). Chicago: National Bureau of Economic Research | The University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.003.0001

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., … Trench, M. (2017). *ARTIFICIAL INTELLIGENCE: THE NEXT DIGITAL FRONTIER?* McKinsey&Company - McKinsey Global Institute. https://doi.org/10.1016/S1353-4858(17)30039-9

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). *Notes From the AI Frontier: Modeling the Global Economic Impact of AI*. *McKinsey*. McKinsey&Company. Retrieved from https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

Bughin, J., Seong, J., Manyika, J., Hämäläinen, L., Windhagen, E., & Haran, E. (2019). *Notes From the AI Frontier: Tackling Europe's Gap in Digital and AI*. *McKinsey Global Institute*.

Candel, A., & LeDell, E. (2019). *Deep learning with H2O*. (A. Bartz, Ed.), *H2O. ai* (6th ed.). Retrieved from http://h2o.ai/resources/

Candelon, F., Yang, F., & Wu, D. (2019). Are China's Digital Companies Ready to Go Global? Retrieved 18 February 2020, from https://www.bcg.com/publications/2019/china-digital-companies-ready-go-global.aspx

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*. https://doi.org/10.1145/1015330.1015432

Chanias, S., Myers, M. D., & Hess, T. (2019). Digital transformation strategy making in pre-digital organizations: The case of a financial services provider. *Journal of Strategic Information Systems*, *28*(1), 17–33. https://doi.org/10.1016/j.jsis.2018.11.003

Chen, H., Chiang, R. H. L. ., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly: Management Information Systems*, *36*(4), 1165–1188. https://doi.org/10.5121/ijdps.2017.8101

Chui, M., Manyka, J., Mehdi, M., Henke, N., Chung, R., Nel, P., & Malhotra, S. (2018). *Notes from Hundrets of Insights From the AI Frontier Use Cases*. *McKinsey Global Institute*. McKinsey&Company.

Claessens, S., Zhu, F., Frost, J., & Turner, G. (2018). Fintech Credit Markets around the World: Size, Drivers and Policy Issues. *BIS Quarteley Review*, (September), 29–49. Retrieved from https://www.bis.org/publ/qtrpdf/r_qt1809e.pdf

Clayton, P. R., & Clopton, J. (2019). Business curriculum redesign: Integrating data analytics. *Journal of Education for Business*, *94*(1), 57–63. https://doi.org/10.1080/08832323.2018.1502142

Columbus, L. (2019). What's New In Gartner's Hype Cycle For AI, 2019. Retrieved 19 February 2020, from https://www.forbes.com/sites/louiscolumbus/2019/09/25/whats-new-in-gartners-hype-cycle-for-ai-2019/#6dbac2e6547b

Cooper. (2019). *Fintech and Banking: What do we know?*

Davenport, T. H. (2018). From analytics to artificial intelligence. *Journal of Business Analytics*, *1*(2), 73–80. https://doi.org/10.1080/2573234x.2018.1543535

Delen, D., & Ram, S. (2018). Research challenges and opportunities in business analytics. *Journal of Business Analytics*, *1*(1), 2–12. https://doi.org/10.1080/2573234x.2018.1507324

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Retrieved from http://arxiv.org/abs/1810.04805

Falbel, D., & Allaire, J. (2019). Keras: R Interface to 'Keras'. R Package. Retrieved 22 November 2019, from https://cran.r-project.org/web/packages/keras/index.html

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. *Advances in Neural Information Processing Systems*, *2015-Janua*, 2962–2970.

https://doi.org/10.1007/978-3-030-05318-5_6

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Flach, P., Hernández-Orallo, J., & Ferri, C. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, 657–664.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. https://doi.org/10.2307/2699986

Frost, J., Gambacorta, L., Huang, Y., Shin, H. S., & Zbinden, P. (2019). *BigTech and the changing structure of financial intermediation*. *BIS Working Papers*. Retrieved from https://www.bis.org/publ/work779.htm

FSB. (2019). *BigTech in finance: Market developments and potential financial stability implications*.

Gampfer, F., Jürgens, A., Müller, M., & Buchkremer, R. (2018). Past, current and future trends in enterprise architecture—A view beyond the horizon. *Computers in Industry*, *100*, 70–84. https://doi.org/10.1016/j.compind.2018.03.006

Gary, M. (2018). *Deep Learning: A Critical Appraisal*. Retrieved from https://arxiv.org/abs/1801.00631

Gijsbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B., & Vanschoren, J. (2019). An Open Source AutoML Benchmark, 1–8. Retrieved from http://arxiv.org/abs/1907.00909

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. *Deep Learning*. https://doi.org/10.1016/B978-0-12-391420-0.09987-X

Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). Maxout networks. *30th International Conference on Machine Learning, ICML 2013*, (PART 3), 2356–2364.

Grover, V., Chiang, R. H. L., Liang, T. P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems*, *35*(2), 388–423.

https://doi.org/10.1080/07421222.2018.1451951

Guo, S., He, H., & Huang, X. (2019). A Multi-Stage Self-Adaptive Classifier Ensemble Model With Application in Credit Scoring. *IEEE Access*, *7*, 78549–78559. https://doi.org/10.1109/ACCESS.2019.2922676

H2O.ai. (2019). H2O AutoML. Retrieved 13 January 2020, from http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html

H2O. (2020). H2O Driverless AI. Retrieved 12 March 2020, from https://www.h2o.ai/products/h2o-driverless-ai/

Halvari, T., Nurminen, J. K., & Mikkonen, T. (2020). Testing the Robustness of AutoML Systems. Retrieved from http://arxiv.org/abs/2005.02649

Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, C. (2018). Ensemble Learning or Deep Learning? Application to Default Risk Analysis. *Journal of Risk and Financial Management*, *11*(1), 12. https://doi.org/10.3390/jrfm11010012

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society. Series A: Statistics in Society*. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Stanford, California: Springer. https://doi.org/10.1007/b94608

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning Second Edition*. *Springer*. Springer. https://doi.org/111

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (2016). *THE AGE OF ANALYTICS: COMPETING IN A DATA-DRIVEN WORLD*. Retrieved from https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

Hess, T., Matt, C., Benlian, A., & Wiesböck, F. (2016). Options for Formulating a Digital Transformation Strategy. *MIS Quarterly Executive*, *15*(2), 123–125. https://doi.org/10.1108/10878571211209314

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Comp.*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hudon, M., Labie, M., Szafarz, A., & Venet, B. (2019). Sustainability, FinTech and financial Inclusion, 162–172. https://doi.org/10.4337/9781788114226.00024

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jovanovic, B., & Rousseau, P. L. (2005). General Purpose Technologies. In *Handbook of Economic Growth (Chapter 18)* (Vol. 1, pp. 1181–1224). https://doi.org/10.1016/S1574-0684(05)01018-X

Kabir, M. F., & Ludwig, S. A. (2019). Enhancing the Performance of Classification Using Super Learning. *Data-Enabled Discovery and Applications*, *3*(1). https://doi.org/10.1007/s41688-019-0030-0

Katona, A., Spick, R., Hodge, V. J., Demediuk, S., Block, F., Drachen, A., & Walker, J. A. (2019). Time to Die: Death Prediction in Dota 2 using Deep Learning, 1–8. https://doi.org/10.1109/cig.2019.8847997

Konstantinou, N., Abel, E., Bellomarini, L., Bogatu, A., Civili, C., Irfanie, E., … Paton, N. W. (2019). VADA: an architecture for end user informed data preparation. *Journal of Big Data*. https://doi.org/10.1186/s40537-019-0237-9

Kraus, M., Feuerriegel, S., & Oztekin, A. (2019). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*, 1–14. https://doi.org/10.1016/j.ejor.2019.09.018

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, 1–9. https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. *Springer*. https://doi.org/10.1007/978-1-4614-6849-3

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeDell, E., & Gill, N. (2019). H2O: R Interface for 'H2O'. R Package. Retrieved 22 November 2019, from https://cran.r-project.org/web/packages/h2o/index.html

Lee, K.-F. (2018). AI Superpowers by Kai-Fu Lee.

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks*, *7*(1). https://doi.org/10.3390/risks7010029

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2015.05.030

Levi, A., & Konish, L. (2020). The five biggest tech companies now make up 17.5% of the S&P 500 — here's how to protect yourself. Retrieved 29 January 2020, from https://www.cnbc.com/2020/01/28/sp-500-dominated-by-apple-microsoft-alphabet-amazon-facebook.html

Malohlava, M., & Candel, A. (2019). *Gradient Boosting Machine with H2O*. *H2O. ai* (7th ed.). H2O. Retrieved from http://h2o.ai/resources/%0Ahttp://h2o-release.s3.amazonaws.com/h2o/master/3805/docs-website/h2o-docs/booklets/GBMBooklet.pdf

Miklosik, A., Kuchta, M., Evans, N., & Zak, S. (2019). Towards the Adoption of Machine Learning-Based Analytical Tools in Digital Marketing. *IEEE Access*, *7*(Ml), 85705–85718. https://doi.org/10.1109/ACCESS.2019.2924425

Miller, H., & Stirling, R. (2019). *Government Artificial Intelligence Readiness Index*. Retrieved from https://www.statista.com/study/50485/artificial-intelligence/

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, Massachusetts; London, England: The MIT Press. https://doi.org/10.1038/217994a0

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., & Turaga, D. (2017). Learning feature engineering for classification. In *IJCAI International Joint Conference on Artificial Intelligence*. https://doi.org/10.24963/ijcai.2017/352

Ng, A. (2018). *Machine Learning Yearning*. https://doi.org/10.1007/978-981-10-1509-0_9

Pang, Z.-J., Liu, R.-Z., Meng, Z.-Y., Zhang, Y., Yu, Y., & Lu, T. (2019). On Reinforcement Learning for Full-Length Game of StarCraft. *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v33i01.33014691

Pannu, A. (2015). Artificial Intelligence and its Application in Different Areas. *Certified International Journal of Engineering and Innovative Technology*.

Parker, G., Van Alstyne, M., & Choudary, S. (2016). Platform revolution: How networked markets are transforming the economy and how to make them work for you. *W.W. Norton & Company*.

R Core Team. (2019). R: A language and environment for statistical computing. Retrieved 22 November 2019, from https://www.r-project.org/

Ramesh, A., & Unruh, A. (2020). Google: Cloud AI Platform Pipelines. Retrieved from https://cloud.google.com/blog/products/ai-machine-learning/introducing-cloud-ai-platform-pipelines

Reinhard, G., Jesper, V., & Stefan, S. (2016). Industry 4.0: Building the digital enterprise. *2016 Global Industry 4.0 Survey*. https://doi.org/10.1080/01969722.2015.1007734

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should i trust you?' Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Russel, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall - Pearson Education, Inc. https://doi.org/10.1017/S0269888900007724

Samek, W., & Müller, K.-R. (2019). *Towards Explainable Artificial Intelligence*. https://doi.org/10.1007/978-3-030-28954-6_1

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schmitt, M. (2016). *Artificial Intelligence in Finance: Machine Learning Applications for Risk Management in Financial Institutions*. University of Strathclyde, Glasgow. Retrieved from www.marcschmitt.com

Schwab, K. (2016). The Fourth Industrial Revolution: what it means and how to respond. *World Economic Forum*.

Sharda, R., Delen, D., & Turban, E. (2017). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*. Pearson Education Limited. Retrieved from www.pearsonglobaleditions.com

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Siebel, T. M. (2019). *Digital Transformation: Survive and Thrive in an Era of Mass Extinction* (1st ed.). New York: Rosetta Books.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., … Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*. https://doi.org/10.1038/nature24270

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*.

Stadelmann, T., Amirian, M., Arabaci, I., Arnold, M., Duivesteijn, G. F., Elezi, I., … Tuggener, L. (2018). Deep learning in the wild. *ArXiv*, *11081 LNAI*, 17–38. https://doi.org/10.1007/978-3-319-99978-4_2

Storbeck, O. (2018). Commerzbank to be replaced by Wirecard in Dax index. Retrieved 29 January 2020, from https://www.ft.com/content/2d32b806-b150-11e8-8d14-6f049d06439c

Stulz, R. M. (2019). *FinTech, BigTech, and the future of Banks*.

Sutton, R. S., & Barto, A. G. (2017). *Reinforcement learning: an introduction 2018 complete draft*. *UCL,Computer Science Department, Reinforcement Learning Lectures*. https://doi.org/10.1109/TNN.1998.712192

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Teng, G. E., He, C. Z., Xiao, J., & Jiang, X. Y. (2013). Customer credit scoring based on HMM/GMDH hybrid model. *Knowledge and Information Systems*. https://doi.org/10.1007/s10115-012-0572-z

Thomas, J. (2019). *Gradient Boosting in Automatic Machine Learning: Feature Selection and Hyperparameter Optimization*. Ludwig Maximilians Universit München.

Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, C. B., & Farivar, R. (2019). Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. Retrieved from http://arxiv.org/abs/1908.05557

US Government. (2019). *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update*.

Van Der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*. https://doi.org/10.2202/1544-6115.1309

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Qi Dong, J., Fabian, N., & Haenlein, M. (2019). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, (July 2018). https://doi.org/10.1016/j.jbusres.2019.09.022

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., … Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*. https://doi.org/10.1038/s41586-019-1724-z

Wang, C., Han, D., Liu, Q., & Luo, S. (2019). A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM. *IEEE Access*, *7*, 2161–2168. https://doi.org/10.1109/ACCESS.2018.2887138

Wang, Z., de Freitas, N., & Lanctot, M. (2016). Dueling Network Architectures for Deep Reinforcement Learning. *ArXiv*, (9), 1–16. https://doi.org/10.1109/MCOM.2016.7378425

Warner, K. S. R., & Wäger, M. (2019). Building dynamic capabilities for digital transformation: An ongoing process of strategic renewal. *Long Range Planning*, *52*(3), 326–349. https://doi.org/10.1016/j.lrp.2018.12.001

Weller, A. (2019). *Transparency: Motivations and Challenges* (Vol. 2). https://doi.org/10.1007/978-3-030-28954-6_2

Wilson, J., Meher, A. K., Bindu, B. V., Chaudhury, S., Lall, B., Sharma, M., & Pareek, V. (2020). Automatically Optimized Gradient Boosting Trees for Classifying Large Volume High Cardinality Data Streams Under Concept Drift. https://doi.org/10.1007/978-3-030-29135-8_13

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*. https://doi.org/10.1016/S0893-6080(05)80023-1

Yeh, I. C., & Lien, C. hui. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems*

with Applications, *36*(2 PART 1), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Zhang, W., He, H., & Zhang, S. (2019). A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2018.12.020

Zimmermann, A., Schmidt, R., Sandkuhl, K., Jugel, D., Bogner, J., & Möhring, M. (2018). Evolution of Enterprise Architecture for Digital Transformation. *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOCW*, *2018-Octob*(October), 87–96. https://doi.org/10.1109/EDOCW.2018.00023

Appendix

Table 14. Feature Description for the Marketing/Sales dataset

Variable	Description
X1	age (numeric)
X2	job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management', 'retired','self-employed','services','student','technician','unemployed','unknown')
X3	marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
X4	education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate', 'professional.course','university.degree','unknown')
X5	default: has credit in default? (categorical: 'no','yes','unknown')
X6	housing: has housing loan? (categorical: 'no','yes','unknown')
X7	loan: has personal loan? (categorical: 'no','yes','unknown')
X8	contact: contact communication type (categorical: 'cellular','telephone')
X9	month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
X10	day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
X11	duration: last contact duration, in seconds (numeric).
X12	campaign: number of contacts performed during this campaign and for this client (numeric)
X13	pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 in case of no contact)
X14	previous: number of contacts performed before this campaign and for this client (numeric)
X15	poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

While more features in the case of marketing analytics might be helpful to sharpen the predictive model and increase accuracy levels, the existing features are sufficient to answer the research question.

Feature Description Insurance dataset

The dataset does not contain a detailed description similar to the marketing or credit risk datasets. The features are shortcuts (e.g., ind, reg, car, calc). The car shortcut indicates that the information contained here is related to the insured car: Type of car, age, etc.

Also, the features have a postfix to indicate the type.

- Feature with ‘bin’ represents a binary feature
- Feature with ‘cat’ represents a categorical feature
- Feature with ‘calc’ represents an extra calculated feature
- All others are continuous or ordinal
- -1 values represent missing values which have been already informed in the data overview.

The fact that the information is not detailed does not make it less relevant. The fact that it contains 57 different features makes it the most feature-rich dataset in this thesis and one that does simulate a real-world scenario.

Table 15. Hyperparameter settings - Deep Learning in Business Analytics

Dataset	Parameters					
	RF	**Value**	**GBM**	**Value**	**DL**	**Value**
Credit Risk	ntrees	50	ntrees	40	activation*	ReLU, Maxout
	max_depth	20	max_depth	7	hidden	c(50, 50)
	min_rows	3	min_rows	10	epochs	10
	-		learn_rate	0.2	rate	0.2
Insurance Claims	ntrees	30	ntrees	60	activation*	ReLU, Maxout
	max_depth	10	max_depth	5	hidden	c(200, 200)
	min_rows	8	min_rows	10	epochs	5
	-		learn_rate	0.01	rate	0.005
Marketing/Sales	ntrees	50	ntrees	60	activation*	ReLU, Maxout
	max_depth	20	max_depth	7	hidden	c(100, 100)
	min_rows	3	min_rows	10	epochs	9
	-		learn_rate	0.15	rate	0.005

*These are the activation functions for the hidden layers

Table 16. Hyperparameter settings - Super Learning in FinTech (candidate models)

Dataset	Parameters					
	RF	**Value**	**GBM**	**Value**	**DL**	**Value**
Taiwan	ntrees	50	ntrees	40	activation*	ReLU
	max_depth	20	max_depth	7	hidden	c(50, 50)
	min_rows	3	min_rows	10	epochs	10
	-		learn_rate	0.2	rate	0.2
Germany	ntrees	40	ntrees	39	activation*	ReLU
	max_depth	15	max_depth	7	hidden	c(200, 200, 200)
	min_rows	4	min_rows	10	epochs	15
	-		learn_rate	0.2	rate	0.01
Australia	ntrees	50	ntrees	29	activation*	ReLU
	max_depth	18	max_depth	15	hidden	c(50, 50)
	min_rows	5	min_rows	10	epochs	10
	-		learn_rate	0.01	rate	0.01

*This is the activation function for the hidden layers

Table 17. Hyperparameter settings - AutoML in BA (candidate models for super learning)

Dataset	Parameters					
	RF	**Value**	**GBM**	**Value**	**DL**	**Value**
Credit Risk	ntrees	50	ntrees	40	activation*	ReLU
	max_depth	20	max_depth	7	hidden	c(50, 50)
	min_rows	3	min_rows	10	epochs	10
	-		learn_rate	0.2	rate	0.2
Insurance Claims	ntrees	30	ntrees	60	activation*	ReLU
	max_depth	10	max_depth	5	hidden	c(200, 200)
	min_rows	8	min_rows	10	epochs	5
	-		learn_rate	0.01	rate	0.005
Marketing/Sales	ntrees	50	ntrees	60	activation*	ReLU
	max_depth	20	max_depth	7	hidden	c(100, 100)
	min_rows	3	min_rows	10	epochs	9
	-		learn_rate	0.15	rate	0.005

*This is the activation function for the hidden layers