Degree Project in Industrial Management

Second cycle, 30 credits

Machine Learning Implementation for Prediction of Probability of Default in Credit Risk

THERESA DÖÖS

ANNIE HOLGERSSON

TRITA – ITM-EX 2024:285
Stockholm, Sweden 2024

www.kth.se

Machine Learning Implementation for Prediction of Probability of Default in Credit Risk

by

Theresa Döös
Annie Holgersson

Master of Science Thesis TRITA-ITM-EX 2024:285
KTH Industrial Engineering and Management
Industrial Economics and Management
SE-100 44 STOCKHOLM

Implementation av Maskininlärning för Prediktion av Sannolikhet för Fallissemang inom Kreditrisk

av

Theresa Döös
Annie Holgersson

Examensarbete TRITA-ITM-EX 2024:285
KTH Industriell teknik och management
Industriell ekonomi och organisation
SE-100 44 STOCKHOLM

KTH Industrial Engineering
and Management

Master of Science Thesis TRITA-ITM-EX 2024:285

Machine Learning Implementation for Prediction of Probability of Default in Credit Risk

Theresa Döös
Annie Holgersson

Approved 2024-06-07	Examiner Jannis Angelis	Supervisor Christian Thomann
	Commissioner Länsförsäkringar Bank AB	Contact person Gustav Rosquist

Abstract

Probability of Default (PD) models can be used for various purposes within the banking and insurance sector. The models classify an observation as the probability of that observation defaulting within a future time span, often a year. Within banks, this type of model is used to calculate risk grades and capital requirements and needs to be accepted for implementation by Finansinspektionen. Due to requirements from Finansinspektionen regarding transparency and explainability, logistic regression is primarily used today when constructing these models. However, there is an interest in exploring how more advanced machine learning models would perform in this field.

On behalf of Länsförsäkringar, this study focuses on how three models built using three different machine learning algorithms perform when trained on the same data as Länsförsäkringar's current model. The algorithms used are Random Forest, XGBoost, and Artificial Neural Networks, and the dataset used consists of private customers holding loans between the years 2007 and 2019. In addition, the study also covers current literature in the field, feature analysis, variable selection, and training of hyperparameters for model optimization. The model that performs the best according to the selected performance measures AUC, Brier score and log loss is the XGBoost model, which is in accordance with findings from several previous studies. The transparency and explainability of this model are found to be inferior to that of logistic regression, but the model does not lack transparency altogether. The study suggests further analysis of how these models could be implemented in the field of PD modelling and how the requirements from Finansinspektionen and EU could be interpreted and changed in order to make reality of the implementation machine learning in risk management.

Key-words
Risk management, credit risk, probability of default, machine learning, Random Forest, XGBoost, Artificial neural network

KTH Industriell teknik och management

Examensarbete TRITA-ITM-EX 2024:285

Implementation av Maskininlärning för Prediktion av Sannolikhet för Fallissemang inom Kreditrisk

Theresa Döös

Annie Holgersson

Godkänt 2024-06-07	Examinator Jannis Angelis	Handledare Christian Thomann
	Uppdragsgivare Länsförsäkringar Bank AB	Kontaktperson Gustav Rosquist

Sammanfattning

Modeller för beräkning av sannolikheten för fallissemang kan användas för olika ändamål inom bank- och försäkringssektorn. De klassificerar en observation som sannolikheten för att den observationen fallerar inom en framtida tidsperiod, ofta ett år. Hos vissa banker används denna typ av modell för att beräkna riskklasser och kapitalkrav och behöver genomgå en process för godkännande hos Finansinspektionen. På grund av krav av Finansinspektionen gällande transparens och förklarbarhet används idag främst logistisk regression vid konstruktion av dessa modeller. Det finns dock ett intresse för att utforska hur mer avancerade maskininlärningsmodeller skulle prestera inom detta område.

På uppdrag av Länsförsäkringar fokuserar denna studie på hur tre modeller, byggda på tre olika maskininlärningsalgoritmer, presterar när de tränas på samma data som Länsförsäkringars nuvarande modell. Algoritmerna som används är Random Forest, XGBoost och Artificial Neural Networks, och datsetet som används består av privatkunder med lån mellan åren 2007 och 2019. Dessutom innehåller studien även en litteraturstudie av området, variabelanalys, variabelval och träning av hyperparametrar för att optimera modellprestationen. Den modell som presterar bäst enligt de utvalda prestationsmåtten AUC, Brier score och log loss är XGBoost, vilket stämmer överens med resultat från flera tidigare studier. Transparensen och förklarbarheten hos denna modell har visat sig vara lägre än för logistisk regression, men möjligheten till transparens är inte obefintlig. Studien föreslår ytterligare utredning av hur dessa modeller skulle kunna införlivas inom PD-modellering och hur kraven från Finansinspektionen och EU skulle kunna tolkas och behöva förändras för att användningen av maskininlärning skulle bli verklighet inom riskhantering.

Nyckelord

Riskhantering, kreditrisk, sannolikhet för fallissemang, maskininlärning, Random Forest, XGBoost, artificiellt neuralt nätverk

Contents

List of Figures

List of Tables

Acknowledgements

We would like to express our sincere gratitude to Marie Frentz, Gustav Rosquist and Johan Wenelius at Länsförsäkringar for their invaluable guidance, support and encouragement throughout the duration of this thesis. Their expertise and insights have been instrumental in shaping this last chapter of our studies. We would also like to thank our supervisor Christian Thomann at the School of Industrial Engineering and Management at KTH for his insightful feedback, constructive criticism and valuable suggestions, which have greatly contributed to the refinement and enhancement of this thesis.

Theresa Döös & Annie Holgersson

Acronyms

ANN	Artificial neural networks
AR	Accuracy ratio
AUC	Area under the curve
BS	Brier score
CV	Cross validation
FN	False negative
FP	False positive
FPR	False positive rate
IRB	Internal ratings-based
IRK	Intern riskklassificering (Swedish)
IV	Information value
LL	Log loss
PD	Probability of default
PFI	Permutation feature importance
RFE	Recursive feature elimination
ROC	Receiver operating characteristic
SMOTE	Synthetic minority oversampling technique
FP	False positive
TP	True positive
TPR	True positive rate
WoE	Weight of evidence

1 Introduction

1.1 Background

The use of machine learning and AI has become immensely popular over the last couple of years and even more so following the boom of generative AI. The possible fields of implementation seem to be endless as organizations become more data driven and digital and it has been shown that high performing organizations invest great amounts to develop and implement more AI driven solutions. This compels companies to make similar investments to maintain their market share [Chui et al., 2023]. Financial institutions, including insurance companies, banks, and other credit institutions, will be undergoing a significant transformation in their approach to risk management as AI and machine learning techniques have become increasingly effective tools for consistent decision making. Today, many credit institutions use so called IRB-models to aid in managing their assessment of credit risks. IRB is an abbreviation of ”internal ratings-based” and the IRB-models consist of different submodels. An important part of the IRB-modelling is the probability of default model, which is used to calculate the risk of a customer defaulting. IRB-models are directly connected to the capital adequacy requirement at the bank and can vary a lot in design between companies and area of implementation. The alternative to using the IRB approach for credit risk is to use the standardized approach. The standardized approach for credit risk is mainly used by small and medium-sized banks and as the name indicates, the method means that a bank uses standardized risk weights to calculate how much capital they should hold [Swedish Bankers' Association, 2020]. However, in Sweden, the IRB-banks account for more than 90 percent of the banking sector's assets and what all IRB models in Sweden have in common is that they must be approved by Finansinspektionen, the Swedish Financial Supervisory Authority. Credit institutions must follow the regulations given by Finansinspektionen and one of the most important aspects to consider when creating IRB-models is that the decisions taken need to be explainable to the customers [Finansinspektionen, 2018].

Due to the regulations and the approval process by Finansinspektionen, Swedish credit institutions are unable to implement any advanced machine learning models in their IRB-processes. However, there is a wish to explore the field of machine learning within risk management in order to be ready for future shifts in legislation. Länsförsäkringar Bank, a Swedish bank active in credit risk assessment and lending, is one of the Swedish financial institutions wanting to explore the potential of such a model. Länsförsäkringar present some successful projects where more advanced machine learning models have been put into professional use in the UK and express a desire to investigate the feasibility and accuracy achieved when using such tools in Sweden [Rosquist, 2024].

The downside to using more advanced machine learning models in decision making is the risk of loosing transparency and explainability [Li and Han, 2023]. Most machine learning models use some type of black-box feature, where data is processed by an algorithm in which the operations and decisions taken are not directly visible to the user. This means that the decisions taken by a model using black-box methodology will not be presentable to the customer to the extent required. Swedish law states that all decisions taken by insurance companies and banks need to be presented and motivated to the customer, meaning all customers have the right to know exactly on what grounds their motion was dismissed, that being a variable in a linear regression model or answer in a questionnaire [Finansinspektionen, 2018]. This means the use of machine learning in risk assessment will require some changes to the concept of transparency, as insurance companies and banks will need to change their routines to be able to still communicate decisions clearly to customers. In addition to this, the European Parliament passed the European Union AI Act in March 2024 which outlines regulations regarding the use of artificial intelligence and machine learning [European Union, 2024]. This is the first comprehensive legal framework outlining the responsibilities, risks and safety measures a market actor must take in order to use artificial intelligence in a public setting. The framework states that any system using artificial intelligence must not do harm and only be used under strict risk assessment and regulation if the implementation is classified as having limited to high risk. For the use of artificial intelligence in risk management one can interpret the AI Act such that as long as the customer is made aware that the decisions made is a result of artificial intelligence, it can be used. However, time will tell how the new AI Act will be implemented, interpreted and used.

1.2 Historical background

Insurance companies, credit institutions and banks in Sweden do not only follow the guidelines by Finansinspektionen but by doing so they also follow the Basel accords as implemented in European Union legislation. The reason behind banks having special regulations is connected to the government and a country's macro economy as well as to the individual's role in society. Having stable and safe banks is a requirement for a healthy society as banks offer financial aid and stability to both individuals and companies. If banks start facing difficulties covering their risks or liabilities the instability can spread to other parts of the market. Private customers can loose their confidence in the banking sector and request to withdraw their deposited funds, resulting in a bank run. At the same time the customers are hesitant to borrow money from the bank if unrest has spread, resulting in a shrinking interest income for the banks. As the banks loose the deposits and see their income decrease, they can shift their business to save and try to cover their losses, resulting in further distress on the market. The same applies for companies wanting to invest in or receive funds for growing from the bank. When banks face financial distress or uncertainty and this starts to affect the companies usually dependent on support from the banks the distress spreads to the job market, the country's export and production and so starts the downward facing spiral of economic distress in society. In the 1990's the Swedish market faced hardship when, in brief, regulations regarding banks' credit risk was eased at the same time as the Swedish currency was weak and inflation was high. The result was, among other things, the default of Nordbanken and the government takeover that followed when the bank needed new capital to fulfill their capital requirement after the market shift. After the difficulties of Nordbanken several banks followed, including Första Sparbanken and Göta Bank. When a Swedish bank is on the verge of collapsing the Swedish government steps in to prevent a further deepening of the crisis, resulting in big public spending which can lead to further public hardship going forward. The financial crisis in the 1990's shows that secure banks are a necessity for having a stable economy, and how regulations and guidelines can act as a necessary tool to prevent financial hardship [Englund, 2015].

In 2007 and 2008 the world was hit by a financial crisis partly caused by issues on the US housing and fixed-income market and by banks facing difficulties covering risk. When the investment bank Lehman Brothers filed for bankruptcy in the fall of 2008 the crisis was a fact and several countries saw businesses and banks default. The bankruptcy of Lehman Brothers came for many investors and observers as a shock and became a wake up call to the risks of the financial market and the need for new global regulations [Riksbanken, 2023]. However, not only banks and credit institutions can cause financial hardship and bank defaults, but similar issues occur when customers start to default on their loans. Because of this, it is important for a credit institution such as a bank to have IRB-models in place to monitor the risk of customer defaults and to be aware of the bank's ability to cover potential losses. This is where the Basel accords play a significant role.

1.3 The Basel accords

The Basel accords are frameworks covering capital adequacy created by the Basel Committee. The Committee was founded to enhance financial stability through improvement of banking supervision and global collaboration. The Basel accords were introduced to enhance the stability and resilience of the global financial system, and the accords have undergone iterative enhancements in response to evolving financial markets since introduction [Basel Committee, 2024]. While the Basel Committee has no actual mandate to influence or regulate domestic and international legislation, the accords are implemented by the European Union making it binding for Swedish banks. The first accord introduced in 1988, Basel I, can be seen as laying the groundwork for international banking regulation and collaboration. The primary focus on the first accord is on credit risk as Basel I mandated a minimum capital requirement of 8% based on a bank's portfolio's risk-weighted assets. The capital requirement refers to the minimum level of capital a financial institution must have available and is divided into *Tier 1* and *Tier 2* capital. *Tier 1* capital is a bank's primary source of funding and consists of shareholder's equity and retained earnings while *Tier 2*, on the other hand, includes revaluation reserves, hybrid capital instruments and subordinated term debt, general loan-loss reserves, and undisclosed reserves. After the implementation of Basel I several flaws were identified with the framework, including the fact that Basel I only covered credit risk, leading to the introduction of Basel II.

The second Basel accord, Basel II, was introduced in 2004 to further cover risk sensitivity and align capital requirements more closely with the risk profile of banks. The idea behind Basel II was to expand the accords to cover more than only credit risk, and the focus of the new accord was to introduce an international standard for banking regulation. Basel II consists of three pillars: minimum capital requirement, supervisory review and market discipline [Basel Committee, 2024]. The first pillar aims to ensure that banks maintain sufficient capital to cover risks, particularly credit, operational, and market risks. This includes monitoring the probability of default to be able to cover potential customer defaults. The focus of the second pillar is the supervisory review of banks' risk management practices and adequacy of a bank's capital in relation to its risk. This means banks must conduct internal risk assessments in order to determine whether additional capital is needed beyond the minimum requirements to cover the bank's risks. It also means that supervisory authorities, such as Finansinspektionen in Sweden, review and evaluate the banks' risk management processes, internal controls and capital adequacy to ensure a common standard for risk management on the market. Lastly, the third pillar of Basel II covers market discipline through encouraging transparency and disclosure of relevant information for all actors on the market. This means banks are required to disclose information regarding their risk management practices, capital adequacy and risk exposures, giving market participants, such as shareholders, the chance to assess a bank's risk management and financial health and hence make informed investment decisions.

After the financial crisis in 2008 the Basel committee identified flaws and weaknesses in the systems and frameworks for financial regulation, and presented the Basel III framework to address these. Built upon the earlier Basel frameworks the aim of Basel III is to strengthen the banking sector's resilience and improve the stability of the global financial system. The new framework was introduced in 2010 and was implemented in stages and finalized in 2019. The Basel III framework strengthened capital and liquidity requirements and introduced measures including leverage ratio, liquidity coverage ratio and net stable funding ratio for more accessible monitoring of banks' health. Basel III puts greater emphasis on the need for high-quality capital to ensure banks' resilience, meaning *Tier 1* capital is considered more loss-absorbing and important. The introduction of the different ratios of Basel III aims to prevent excessive leverage in the banking sector and ensure sufficient liquidity to withstand market shifts by promoting more stable funding structures over time and the building of buffers for times of economic hardship and stress. Basel III also emphasises enhanced disclosure requirements to improve transparency, requiring banks to present more information about their risk exposures, risk management practices, and capital adequacy than before as a tool to prevent a future financial crisis [Basel Committee, 2024].

Following the Basel accords' implementation in European Union legislation, mainly through the Capital Requirement Regulation and Capital Requirement Directive, banks and financial institutions are heavily regulated to limit the risk of financial crisis and market instability [European Banking Authority, 2024]. All banks active in Sweden have to obey the regulations presented by Finansinspektionen which consist of local and European Union wide laws, incorporating the Basel accords. This means banks must have practices in place for managing risk, including liquidity, market and credit risk, which includes the use of IRB-models.

1.4 Risk management at Länsföräkringar Bank AB

Länsförsäkringar Bank's risk management department is responsible for the risk modelling of the bank which includes monitoring the risk of default among credit customers. In order to model the risk of customer default and monitor the bank's credit risk Länsförsäkringar hence uses a probability of default model. The customers covered in the model investigated in this report are private, non-corporate customers that have all received a loan from Länsförsäkringar Bank, including secured and unsecured loans. The probability of default model predicts the risk of a customer defaulting, that is loosing their ability to make required payments to the bank. In order to monitor the risk of their lending operations in accordance with the regulations previously discussed, Länsförsäkringar's risk management department uses a variety of statistical models and measurements but as Länsförsäkringar is subject to the frameworks discussed, limitations to what models to use exists. This gives Länsförsäkringar restrictions when developing new ways to model credit risk, including not being able to use more advanced machine learning models. However, as times change regulations need to be updated, and one can argue that the rise of machine learning and artificial intelligence will result in an update of the current regulations.

1.5 Research Question

The purpose of this thesis is to explore the potential and accuracy of machine learning models applied to assess probability of default within credit risk modelling for private customers. Following the frameworks discussed above, the use of machine learning in probability of default modelling will study how one can strengthen the prediction power of the IRB-models and limit the risk for Länsförsäkringar in accordance with regulations by Finansinspektionen. The purpose will be fulfilled through answering the following two research questions:

- Is there an added value in using more advanced machine learning methods when model performance is weighed against being able to explain the model outcome?
- Is it possible to use machine learning to develop new decision models that, when compared to the performance of linear models, make more accurate decisions and performs stable now and over time?

1.5.1 Scope

The study focuses on building a more advanced model for calculation of probability of default in risk classification and compare it to the performance of linear models. With regards to risk, the thesis deals with Länsförsäkringar Banks' capital coverage for credit risk through the use of probability of default models. The clients covered by the model are private, non-corporate customers holding secured and unsecured loans with Länsförsäkringar Bank.

1.6 Key findings

In addressing these research questions, the thesis found that machine learning models, particularly XGBoost, demonstrated a strong predictive ability, achieving an AUC score exceeding 0.95. The implementation of these models also streamlined the process by reducing the need for extensive manual data management and simplifying updates over time, improving operational efficiency. While transparency remains a challenge compared to logistic regression, variable importance measures like Permutation Feature Importance (PFI) offer some insight. Overall, the findings underscore the potential benefits and applicability of integrating machine learning into IRB modeling, highlighting the need for updated regulatory frameworks and advanced expertise within financial institutions to fully leverage these technologies.

2 Literature review

The use of machine learning in financial risk management is believed to be the next step towards making consistent decisions and reducing risk [Azzone et al., 2022]. As the climate is changing and the acceptance for usage of more advanced algorithms and data management is increasing, insurance companies, credit institutions and banks strive towards finding the right balance between traditional models and machine learning based algorithms. The discussion regarding the use of machine learning in risk management is centered around the need for transparency since all decisions regarding credit and insurance limits must be clearly stated and easily presented to the customer meaning the use of black box techniques is limited. However, this seem to be changing [Rosquist, 2024].

Azzone et al. [2022] discussed the topic in their article in which they evaluated how machine learning methods such as random forest could be used to predict lapse in life insurance. The researchers found that the use of models based on more advanced machine learning greatly outperformed traditional models such as logistic regression models. In their study, Azzone et al. found that linear models, such as logistic models, are keen to miss the factors related to heterogeneity, and tend to not be able to capture more advanced relationships between factors. Their research concludes that traditional linear models tend to be favorable to the idea that the economic condition of the insured have greater impact than it in reality has. This means that in many cases the traditional linear models in short result in wrong decisions, also known as lapse decisions. This research hence shows the importance of implementing machine learning in decision making to limit the number of lapse decisions and hence increase profitability and accuracy [Azzone et al., 2022].

Many researchers have implemented machine learning methods as a way to effectively manage large amounts of data, and limit the risk of errors in data management. In the last decade, most industries have transitioned towards a more data driven approach and in their article ”Machine learning applications in nonlife insurance” Grize et al. [2020] discuss the use of machine learning in a changing insurance industry. They state that since the insurance industry has switched to a more data driven approach many companies need to challenge their traditional methods. In their article the authors compare the use of more traditional methods such as generalized linear models and multivariate regression with more advanced machine learning methods, including neural networks and random forest, in non-life insurance decisions. Grize et al. [2020] find that the implementation of machine learning in the insurance industry is long overdue and vital for the industry to adapt as the greater amount of data requires a more foolproof and data driven way for companies to make decisions. The risk of making the wrong decision is too high as traditional models, such as logistic regression, are unfit to manage big amounts of data. At the same time, consumers and companies across the globe have low tolerance for mistakes, putting pressure on credit institutions and insurance companies to limit the risk of mistakes. Risk management and limiting of mistakes has never been more important as the market and business climate change more rapidly than ever before. The authors hence argue that in order to survive insurance companies and credit issuers must implement machine learning methods which are superior when managing big and more complex data sets [Grize et al., 2020].

Based on the reasoning presented above one can argue that a switch to machine learning methods in risk management, both in insurance and credit risk situations, is vital for the survival of the company. However, is the change to a more advanced data driven approach an easy step to take? In their summary of previous studies made in the field of machine learning application in financial risk management, Mashrur et al. [2020] presents challenges and opportunities for the implementation of machine learning to help researchers navigate the complex world of machine learning implementations. The authors state that as the need for more data driven decision making is putting new pressure on financial institutes to evolve, the implementation of machine learning is challenging. Studies have shown that the implementation is most difficult in cases when the quality and size of the data is limited. The success of most machine learning methods is deeply dependent on the quality of the data set, and since most industries are still in the early stages of data driven decision making and big data management, most companies lack the ability to gather and manage the required data. The authors suggest that in order to succeed with the implementation of machine learning methods the first step is to implement ways to gather, prepare and manage the big amount of data required, in contrast to what many seem to do today. Another factor that the authors have identified to be challenging for many actors is the level of explainability of the machine learning methods, something that is vital for insurance and credit risk companies. Mashrur et al. state that the areas where machine learning has been successfully implemented and previous research been exhaustively focused include credit rating and fraud detection, proving that the use of machine learning in advanced decision making is possible and successful [Mashrur et al., 2020].

Due to the major effect machine learning has had on industry and academia the past decade the field of applying more advanced models to finance and credit risk assessment is not undiscovered. Shi et al. [2022] made a systematic review of 76 major research contributions to investigate the accuracy and efficiency of different models applied to the field. When reviewing the machine learning driven credit risk algorithms they find, similarly to the earlier mentioned study by Azzone et al. that the machine learning models in general perform better than the traditional linear models. Especially the deep learning algorithms, such as neural network models, show a noticeably better result when looking at the statistical accuracy measure AUC and ACC. They also mention the challenges with the models, both discussed in previous paragraph; transparency and limited and imbalanced data sets. When training a model it is desirable to have a balanced data set, but within the credit risk data imbalance is usually severe, making the model building dependent on data sampling methods which in turn affects the accuracy and efficiency of the model [Shi et al., 2022]. When applied to larger credit scoring data sets, Kruppa et al. [2013] also show that machine learning models give a better result than classical credit risk models. In their study they compared the accuracy of a classical general linear regression to Random Jungle, a fast random forest implementation when applied to credit risk. The random forest method using probability estimation trees received an AUC score of 0.959 which is outstanding, especially compared to the different versions of logistic regression used in the study that received scores of ~ 0.75 [Kruppa et al., 2013].

The choice of what kind of machine learning models to apply when assessing credit risk is a difficult one since more complex models have not been widely used. Chen et al. [2017] made a study applying Support Vector Machines when assessing credit scores, achieving an overall accuracy of 77.8% when applied to real world data set from a bank in Bulgaria. They also achieve a 64% Gini coefficient, another measure of model accuracy, but the score is not remarkably good [Chen et al., 2017]. Another study of credit risk analysis using machine and deep learning models by Addo et al. [2018] argue, in contradiction to the study of Shi et al. [2022], that the algorithms based on artificial neural networks do not necessarily provide the best performance. The result of their study shows, similarly to the study of Kruppa et al. [2013], that the tree based methods are the best and most stable models [Addo et al., 2018].

Similar to the findings of Kruppa et al. and Addo et al. a study conducted at Cornell University found that while deep learning algorithms has proven to be of great use in text and image processing, tree algorithms, including Random Forest and XGBoost, have a stronger track record when dealing with tabular datasets. The study by Grinsztajn et al. [2022] was conducted using 45 tabular datasets from various domains only including i.i.d. and real-world data and used in building both classification and regression models. The authors found that not only do models using artificial neural networks appear to be more time consuming and less accurate, but that tuning the hyperparameters does not result in making the neural networks better, coming to the conclusion that the use of tree algorithms are far superior. Grinsztajn et al. explains the lack of accuracy of the artificial neural networks by discussing the difference in structure of a tree-based algorithm and a neural network, coming to the conclusion that since hyperparameter tuning did not seem to result in better model performance the inherent properties of the neural networks are to be blamed. Firstly, the authors found that neural networks are biased to overly smooth solutions preventing the models from learning irregular patterns. Secondly, Grinsztajn et al. found that neural networks require a bigger dataset of informative features as they lack the ability to effectively sort out the uninformative features. This results in less accurate model performance and weaker models Grinsztajn et al. [2022].

As machine learning is a very current topic and the wider use of it is fairly new, many researchers are testing different ways of implementing it in various parts of the market. The ethical part of using machine learning within the financial market is discussed by Rizinski et al. [2022]. They state that the automation achieved by machine learning algorithms will lead to significant cost savings and that it is, if used correctly, more explainable than one might think. They apply a XGBoost model, a model considered a "black box"-model, to asses binary classification of credit approvals in their research. They then show how one can make an explainability analysis with the results by plotting how the different means of the explanatory variables affect the decision, similarly to how one can plot the coefficients in regression. By plotting the results of the decisions taking by humans at the bank and the decisions taking by the model, they come to the conclusion that the decisions taken by the model is less likely to be biased. They argue that both ways to make decisions, by humans or by machine learning models, can potentially make mistakes, but that the decisions made by a machine algorithm is more transparent than the ones taken by humans due to better decisions and reduced bias and mistakes [Rizinski et al., 2022].

The articles and studies discussed above presents a small part of the topics being covered in current research. While the topic of machine learning implementation is commonly studied, few articles discuss the implementation by Swedish insurance companies and banks and what the Swedish market would look like if machine learning methods were to be used. The current supervisory authorities in the European Union, the European Banking Association and the European Central Bank, have no guidelines or practises in place to support the use of machine learning based models in accordance with current regulations for credit risk management. This means that Finansinspektionen currently has no guidelines on how to support and evaluate the use of more advanced machine learning models by Swedish banks, making the implementation infeasible. As a result, the methods used in this analysis will not be implemented by Länsförsäkringar, but will be used to assess whether or not the use of machine learning is preferable when legislation changes.

The use of linear models in credit risk management is however common and established. After the financial crisis in 2008 and the implementation of the Basel III framework that followed, many studies have been made on risk factors associated with credit risk. Most studies evaluate how credit institutions and banks manage the risks associated with lending credit to corporate customers, and identify which risk factors seem to be of higher significance related to the probability of default.

With the 2007-2008 financial crisis in the US being partly created by defaults in American households' mortgage loans, the Basel III framework was created to put pressure on banks to stress test their systems as well as keep their liquidity and funding stable in order to avoid the risk of bank runs. As a results, banks strengthened their requirements on households wanting to borrow money. In their study *LAPS: Computing Loan Default Risk from User Activity, Profile, and Recommendations*, Uriawan et al. [2022] analyze the variables used when reviewing an individual's loan application. Factors discussed in this study include traditional variables such as credit score and collateral but also features containing personal data, job information, assets and salary. When analyzing the relationship between these factors and the chances of receiving a loan, the researchers argue that information regarding the potential borrower's financial health, personal life and behaviour is of importance. Uriawan et al. complete their study with the creation of a *trustworthiness score* in which factors such as age, salary and assets give a potential lender an overall idea of the reliability and integrity of the borrower. The idea behind using a scoring such as this is to see which factors other than credit score and collateral impact the probability of default, giving the reader an idea of what factors to consider upon creating probability of default models.

When analyzing the credit risk of corporate clients banks tend to look at risk factors such as total assets and liabilities, turnover and financial measures including EBITDA and net debt to paint a picture of the company's overall health. One can also include macroeconomic factors in the analysis to incorporate systematic risks associated with lending. This is fairly common to use when analyzing private customers as well, and in their article *Macroeconomic Factors of Consumer Loan Credit Risk in Central and Eastern European Countries*, Kanapickienė et al. [2023] look at systematic and unsystematic credit risks banks face in Central and Eastern Europe when lending to consumers. The authors look at general macroeconomic factors including GDP, inflation and labour market characteristics, and conclude that economy growth indicators, such as GDP and inflation, which have a direct impact on the consumers' financial health, impact the banks' credit risk by making the risk easier to manage. The study by Kanapickienė et al. show that when analyzing the credit risk of consumer banks the macroeconomic factors impacting systemic risk is relevant to take into consideration. This can include looking at the applicants employment status, spending habits and housing situation Kanapickienė et al. [2023].

3 Economic theory

3.1 Risk management

For banks and insurance companies risk management is, as mentioned previously, a vital part of operations. Following the financial crisis in 2007-2008 and the bankruptcy of Lehman Brothers both credit institutions and governments had to restructure and strengthen their risk management. This together with the implementation of the Basel III framework that followed the crisis, has resulted in countries and unions, including the EU, applying regulations to minimize the risk of banks going under. Both in Sweden and abroad, banks are an important pillar in society and the well being of banks are therefore vital for a nation's economic growth, stability and prosperity.

Under EU legislation, credit institutions must hold enough capital to be able to cover credit, operative and systematic risks, in order to prevent bank failure and bank runs. In an operative setting, this means banks must be cautious of who they choose to lend capital to. Risk management in a credit risk setting hence is a challenge of balancing the risk of lending and the potential profit from interest. In order to manage this challenge banks use the measure probability of default to assess the risk of a customer not being able to repay their debt. A credit institution which manages to balance this risk can make great profit and be a stable and trustworthy partner for private and corporate customers, making probability of default models a vital part of a bank's strategy.

3.2 Probability of default

A probability of default (PD) model estimates the probability that an obligor defaults on their credit obligation within twelve months. It is typically a binary classifier returning a probability of an observation/customer defaulting in the near future (usually within the following year). It's predictions are used together with *Exposure At Default* and *Loss Given Default* to calculate *Expected Loss* which in return is used when allocating capital for risk mitigation. The *Exposure At Default* is the total value a lender is exposed to when a loan defaults while the *Loss Given Default* is the amount of money a lender loses when a borrower defaults. Under the IRB approach, credit institutions apply their own internal PD models, which under the Basel accords and the capital requirements regulations must be approved by the supervisory authority. In Sweden Finansinspektionen acts as the supervisory authority for IRB modelling. The PD models commonly used by Swedish banks use historical data and logistic regression to build their models which estimate the probability of default. In their regression models, banks use default as their dependent variable and a variety of financial, behavioral and macroeconomic factors as independent variables to create estimations. Independent variables often include ratios and measures of assets and liabilities, personal information such as co-borrower and income, and factors covering spending behaviour, such as whether the client is paying their bills on time. As the models are trained, tested and evaluated on historical data, they tend to loose accuracy over time, resulting in time consuming processes to update the models. Models tend to loose accuracy over time as they are trained on historical data becoming more obsolete with time.

3.3 Risk grades

In order to easily classify customers and monitor how they behave over time banks and insurance companies use risk grades which splits the customers into buckets based on predefined limits with each grade consisting of customers with similar characteristics [Riksbanken, 2004]. These limits often correspond to different intervals of the probability of default. A bank can for example choose to create 12 risk grades were the model is supposed to classify low risk customers in risk grade 1 and high risk customers in higher grades. The probability of default model used by the bank hence outputs a value between 0 and 1 which is the probability of default, which the bank then can use to classify the customers in buckets. Each grade has a lower and upper limit and some identify the ideal spread of customers would put most customers in the lower grades and very few in the high risk grades, resulting in assumed lower risk for the bank. However this distribution does not only depend on the model and its prediction but indeed on the costumer portfolio as well. Banks can use migration matrices to visualize how customers move from year to year between their risk grades, and either spot trends and changes or evaluate the predictive power of the model over time. The risk grades can also be used when pricing insurances or mortgages since riskier customers should receive a higher premium to cover the increased risk.

4 Technical theory

4.1 Machine learning algorithms

Machine learning has increased in popularity over the past couple of years and the fields of use only seem to grow. A machine learning model can, when compared to a linear statistical model, capture non-linear relationships more efficiently and hence be used to solve more advanced problems. There are a broad variety of models belonging to the machine learning class of algorithms, ranging from generative and non-generative models and unsupervised to supervised machine learning algorithms. IBM defines machine learning as *"through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics"* [IBM, 2024b]. Arguments for using machine learning methods are that these algorithms tend to be customizable and able to identify advanced relationships and patterns in big data sets that simpler models and human analysts fail to identify, while a downside to these models is that they require big and unbiased datasets, as the accuracy increases with the size of the dataset (within limitation). If a machine learning model is trained on a smaller dataset this can result in errors that make the model's predictions correct while the model is actually completely wrong [IBM, 2024b].

One aspect worth discussing in regards to machine learning models is the risk of overfitting the model. Overfitting occurs when the model predicts accurately for training data but not for new data, i.e. test data. This can occur for multiple reasons including training the model on a dataset that is too small or too noisy, that is contains large amounts of irrelevant data, or if the model is trained for too long on one sample set of the data.

The machine learning algorithms presented below has been selected as relevant for the purpose of this study, based on the findings of the literature review in Chapter 2. The tree models have proven to be suitable for similar purposes, while neural networks have shown varying results in the same areas.

4.2 Random forest

Random forest is a machine learning algorithm that can, by using a combination of outputs of multiple decision trees, solve both classification and regression problems. The random forest algorithm consists of multiple decision trees, which all consist of questions and branches with decision nodes stating e.g. yes/no or true/false. In using a decision tree one can split the data into subsets and use information gain, Gini coefficient or mean squared error to evaluate the strength of the split. The random forest algorithm creates an uncorrelated collection of decision trees, where each tree is comprised of a random data sample and number of variables from the training set, with each decision tree resulting in a prediction, with the prediction in a probability of default model being either *default* or *no default*. The final result from the random forest algorithm is then the average or most popular result from the predictions in the decision trees.

Random forest algorithms have several hyperparameters, which are set before the model is trained, including node size, number of trees and number of variables sampled. Because of the speed and flexibility of models built on the random forest algorithm the possible areas for application include finance, economics and healthcare, and as the models have a reduced risk of overfitting the use of random forest has become very popular [Schonlau and Zou, 2020].

4.3 Extreme gradient boosting

Gradient boosting is a class of ensemble machine learning algorithms used in prediction problems, with extreme gradient boosting (XGBoost) being an implementation of the gradient boosting ensemble algorithm. Similar to random forest models, ensemble algorithms are based on decision trees which are added one at a time to correct prediction errors made by previous models. Gradient boosting models are trained to minimize a loss function by adding new models correcting previously made errors and consist of three components: a loss function to be optimized, decision trees to make predictions and an additive model to which the decision trees are added to minimize the loss function. The optimization stage involves identifying the parameters for each model that minimize the loss function and the trees are structured so that at each step the algorithm chooses the variable providing the most information gain. The loss function measures the difference between the predicted value and true value with XGBoost using a tailored loss function for each problem. For classification problems, logistic regression is commonly used as loss function while mean squared error is used in regression problems [Chen and Guestrin, 2016].

As XGBoost uses regularization terms at each step which penalise complex models, this algorithm tends to result in more generalized and simpler models. The algorithm also uses tree pruning which removes branches in the decision trees that do not significantly contribute to model improvement. Because of this, XGBoost is known for being fast and able to make accurate predictions with limited risk of overfitting. Similar to other tree algorithms, XGBoost uses various hyperparameters for regularization, learning rate, tree depth and subsampling [Chen and Guestrin, 2016].

4.4 Artificial neural network

Artificial neural networks (ANNs) get their name from the structure of the human brain and is commonly used in pattern recognition, classification and regression. An ANN consists of nodes (also known as neurons) organized in layers, with an input layer, one or more hidden layers and an output layer. Each node has its own weight and threshold and connects to other nodes. The input layer receives raw input data and sends it into each node, where each node represents a variable in the input data. The hidden layers consist of nodes which can be seen as individual linear regression models consisting of input data, weights, a threshold and an output. When data reaches a node the input data is multiplied by its weight and summed, and the output is passed through a so called activation function which determines the output of the node. Commonly used functions are ReLu, sigmoid and tanh. If the value of the output is greater than some known threshold the node is activated and sends the output forward to the next layer. One can describe the linear regression of each node as follows:

$$\hat{y} = x_1 \times w_1 + x_2 \times w_2 + ... + x_n \times w_n + b \tag{1}$$

where $\hat{y}$ is the predicted outcome, x_i is a variable, w_i the variable's weight and b the predetermined threshold. If $\hat{y}$ is greater than 0 the output would be one while a $\hat{y}$ smaller than 0 results in an output of 0. To evaluate the accuracy of an ANN one uses a cost function, which is to be minimized to ensure a strong predictive ability. Commonly used cost functions are the mean squared error and log loss. When training the ANN, the model adjusts the weights and thresholds used to find the values resulting in the smallest cost function. This procedure is called backpropagation, since the goal is to backpropagate the errors from the output nodes to the input nodes [Mahbobi et al., 2023].

There are many different types of neural networks and they are hence used in various fields. E.g. convolutionary neural networks are greatly beneficial in pattern recognition such as image processing as they consist of convolutional layers, pooling layers, and fully connected layers while feed forward networks work as great classifiers of text topics or sentiment [IBM, 2024a]. Other commonly known applications of ANNs include Google's search engine and Apple's Siri virtual assistance. This thesis will be using a multi perceptron classifier, which is a type of feed forward network. In this type of network the hyperparameters to be set are for example the number and size of hidden layers, activation functions of each layer, the optimizer of the backpropagation and learning rate.

4.5 Parameter tuning

Machine learning algorithms uses hyperparameters to manage the model training. The hyperparameters differ between models and can be seen as external configuration variables which lets the user tailor the model further. Which values to use as hyperparameters are set manually before training the model and one can use parameter tuning to find the optimal values for a classification problem. It is important to note that hyperparameters can not be estimated from the dataset but must be set before model training [Zainab et al., 2020]. Hyperparameter optimization can however be used to tune the parameters to find the best possible values for each model. This is done by comparing the accuracy of the model for a variety of values of each hyperparameter, and choosing the values that corresponds to the best model accuracy. Optimization of hyperparameters can therefore be seen as a vital tool for creating accurate models.

There exists several techniques for hyperparameter training. One such is randomized search. It is a more effective method that, instead of trying all possible combinations of hyperparameters instead samples a defined number of combinations from the hyperparameter space. Giving the method a list of possible values for each hyperparameter, the method terminates by evaluating the performance of each combination using cross validation and selecting the combination that yields the best performance metric [Zainab et al., 2020].

5 Mathematical theory

This chapter presents the mathematical theory behind the current model used for predicting probability of default. This is presented to get a deeper understanding of how it works and how it can be compared to the algorithms utilized in this study. Additionally, the chapter covers techniques used for variable selection, imbalanced dataset and evaluation of model performance. The model performance measures are to be used when comparing the different models to each other over time and for checking if the models meets the performance requirements.

5.1 Logistic regression

Logistic regression is used to predict a binary outcome based on historical observations, and is commonly used in prediction of credit risk and probability of default. The model predicts the value of a binary dependent variable by analyzing the linear relationship between the dependent variable and one or more continuous or binary independent variables. Logistic regression is part of the group of models known as *Generalized linear models (GLM)* and uses a decision boundary value to classify or predict an outcome, such as default or no default. The boundary value varies from 0 to 1 making logistic regression a good choice of model when one wants to predict the outcome as a probability [Montgomery et al., 2012].

A logistic regression model can be expressed as follows:

$$Logit(\pi) = \frac{1}{1 + exp(-\pi)} \tag{2}$$

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \times x_1 + ... + \beta_k \times x_k \tag{3}$$

where $Logit(\pi)$ is the dependent variable, x is the independent variable and β_i is the coefficient often estimated using maximum likelihood estimation [Montgomery et al., 2012]. The model will predict a probability between 0 and 1 which will be translated to a binary value of 0 or 1. A probability of less than 0.5 will predict a 0, while a probability of 0.5 and above will predict a 1.

5.2 Monotonic binning

Data binning is a data processing technique used to minimize the impact of small observation errors without loosing important information. The method organizes the data values for each variable into small intervals, so called bins, which are then replaced by a representative value, e.g. the mean or median of the interval. Monotonic binning requires the binning process to be ordinal, and is used to transform continuous variables into categorical ones. One can also use monotonic binning to manage missing values, as the method puts the data points which have missing values into a separate bin. Monotonic binning algorithms often use some type of decision tree to create the bins and split the data.

5.3 Correlation

Correlation between variables describes the association between two or more variables, and illustrates how when one variable changes the other changes as well. This works as an indicator of the relationships between variables in a dataset and is an important factor to consider when selecting variables. Two variables are considered perfectly correlated if the correlation coefficient takes a value of -/+1, while a value of 0 means the variables are independent. There are multiple ways of measuring correlation with Pearson's correlation coefficient, Spearman's correlation coefficient and Kendall's Tau being three popular methods [Montgomery et al., 2012].

Pearson's correlation coefficient is most commonly used when measuring linear correlation. Pearson's takes on a number between -1 and 1 measuring strength between two variables with a value of -1 indicating a perfect negative relationship and 1 indicating a perfect positive relationship. The formula for calculating Pearson's correlation coefficient for variables x and y looks as follows:

$$r_{x_i x_j} = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} \tag{4}$$

[Montgomery et al., 2012]

Spearman's rank correlation coefficient measures the rank correlation between two variables which is the statistical strength between the rankings of the variables. Spearman's measures the monotonic relationship between two variables and is equal to the Pearson correlation of the ranks of the variables. Spearman's rank correlation takes on values in the range of $[-1, 1]$ is defined as below:

$$\rho = 1 - \frac{6 \times \sum d_i^2}{n \times (n^2 - 1)} \tag{5}$$

where d is the difference in rank between the two variables and n is the number of observations [Montgomery et al., 2012].

Kendall's Tau, also known as Kendall rank correlation coefficient, measures the ordinal association between two variables and is also a measure of rank correlation. It is similar to Spearman's correlation coefficient and also ranges from $[-1, 1]$, but is preferable when using a smaller dataset. Kendall's Tau is defined below:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j)sgn(y_i - y_j) \tag{6}$$

where n is the number of observations, x and y are the two variables and

$$\sum_{i<j} sgn(x_i - x_j)sgn(y_i - y_j) \tag{7}$$

is the difference between the number of concordant pairs and discordant pairs [Montgomery et al., 2012].

5.4 Multicollinearity

Multicollinearity occurs when two or more predictor variables have a strong linear dependency, which can result in overfitting and loss of model accuracy. One way to measure multicollinearity is by the use of a correlation matrix and a cutoff value. The correlation matrix measures the correlation between all pairs of independent variables, often using Pearson's correlation coefficient, which can then be used to remove all variables which have correlation too strong to fit in the model. For example, one can remove the variable which have the smallest significance with the dependent variable in all pairs which have a correlation coefficient above the absolute value of 0.8 [Montgomery et al., 2012]. Causes for multicollinearity include using a dataset with variables that are too similar and using too many variables in comparison to the number of observations.

5.5 Overfitting

When training a classification or regression model the risk of overfitting is to be acknowledged. Overfitting occurs when a fitted model can not generalize to new data resulting in the model giving accurate predictions for the training data but not for the testing data. This can occur when the model is trained for too long on the training data, when the dataset is too small or when the model is too complex resulting in it learning from the irrelevant information in the dataset decreasing the ability to derive patterns in new data. Overfitting can be discovered by comparing the accuracy when using the training and validation datasets and by analyzing the variance. If the model has a low error rate when using the training data but a high error rate on other datasets the model is probably overfitted. To prevent overfitting one can use a variety of methods including early stopping, feature selection and expansion of the training set [European Banking Authority, 2023].

5.6 Managing an imbalanced dataset

Using an imbalanced dataset can cause major problems when building models. An imbalanced dataset is characterized by a small class ratio, meaning the size of the minority class is significantly smaller than the size of the majority class. In Länsförsäkringar's case the minority class is default and the majority class is no default. To manage this one can oversample the minority class or undersample the majority class. When oversampling the majority class one randomly duplicates observations of the minority class to balance the dataset. This can however result in overfitting and a common method to manage this risk is the synthetic minority oversampling technique (SMOTE). The SMOTE algorithm synthesize new data points from the minority class instead of duplicating them, resulting in new "fictional" data points to balance the dataset. The SMOTE algorithm selects a random point in the minority class and finds k number of nearest neighbors to that point, often a value of k=5 is used. Following this, the algorithm chooses one of the selected neighboring points and creates a synthesized point between the neighbor and the initially selected point. When doing this, one limits the risk of overfitting the dataset while managing the issues related to imbalanced datasets [Bowyer et al., 2011].

5.7 Recursive feature elimination using k-fold cross validation

Recursive feature elimination (RFE) is a method used for variable selection that fits a model and removes the least significant variables until a predetermined number of variables is reached. The method recursively removes small number of variables per loop and in that removes dependencies and collinearity from the model. The RFE model uses a predetermined number of variables to keep in the model, and in order to find the optimal number of variables to keep one can use k-fold cross validation. Cross validation is a statistical method used to compare multiple models' ability to make predictions on new data. k-fold cross validation splits the dataset into k subsets and uses, for each group, one subset as a test set and the rest as training set. One then fits a model on the training set and evaluates it on the test set, and summarize the overall skill of the full model by observing the scores of the different submodels created. Most commonly used is a 10-fold cross validation where one divides the dataset into 10 groups. Cross validation is hence used in RFE to choose the number of variables to keep in the model [Jemai and Zarrad, 2023].

5.8 Permutation feature importance

Permutation feature importance (PFI) provides insight into how much each variable contributes to the overall performance of the model and can hence be used in variable selection. The idea behind PFI is to evaluate how the model's predictive performance changes when values of a particular variable are shuffled randomly while the other features stay the same. If the overall performance decreases greatly after shuffling this indicates that the variable is more important for the model's predictive ability, while a low impact on the performance indicates that the variable is not of importance to the model. PFI can use for example AUC or R^2 as measures of impact [Kaneko, 2022].

5.9 Hyperparameter tuning using k-fold cross validation

Hyperparameter tuning is used to find the optimal hyperparameters for each model to achieve the best possible model performance. Such parameters are set before training the model and hold important properties for the model performance such as learning rate and number of nodes or trees. A common way to find the best possible hyperparameters for a model is to use k-fold cross validation described above. To find the optimal set of hyperparameters using cross validation one first defines the hyperparameter space and hence limits the range of values for each parameter. One then uses cross validation to compare all possible subsets of hyperparameters to find the combination resulting in the best possible model performance [Zainab et al., 2020].

The use of hyperparameter tuning can increase the risk of overfitting and is often time consuming with high computational cost. However, it increases the model performance and generalizability which makes it highly beneficial.

5.10 Model performance

5.10.1 Receiver operating characteristic curve

The receiver operating characteristic (ROC) curve is used to evaluate the performance of a classification model and plots the parameters *True Positive Rate (TPR)* and *False Positive Rate* [Bradley, 1997]. A *TPR* can also be defined as a recall and is therefore defined as below:

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

Consequently, *FPR* is defined as follows:

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

where *TP* is the number of true positive predictions, *FP* is the number of false positive predictions and *TN* and *FN* are the numbers of *True Negatives* and *False Negatives* respectively. The ROC curve plots the *TPR* against the *FPR* at different classification thresholds, where lowering the threshold results in more cases predicted as positive and hence an increase in both true and false positive predictions.

The ROC curve can hence be used to organize, assess and illustrate a model's performance and depict the relative trade off between benefits of the model and the costs associated with wrong predictions. To further analyze the performance of the model one tends to use the AUC ROC curve, which is the area under the curve of the ROC [Bradley, 1997].

5.10.2 Area under the curve

The area under the curve (AUC) is used to evaluate how well a model predicts true values. When evaluating the performance of a PD model the AUC ROC curve is commonly used, which maps the area under the ROC curve. This means the AUC ROC curve measures a model's ability to correctly classify a binary classifier, such as default or no default, where an AUC ROC curve of close to or equal to 1 indicates the that the model can correctly differentiate between all positive and negative points. A low score of the AUC ROC curve indicates that the model is not suitable for classifying data points. Overall, a value of the AUC ROC curve of above 0.8 indicates a good model performance, 0.7-0.8 is considered fair and 0.5 is equivalent to random guessing [Bradley, 1997].

5.10.3 Brier score

The Brier score (BS) is a metric which is widely used for assessing the accuracy of probabilistic forecasts, frequently utilized within areas such as weather forecasts and risk prediction models [Ala'raj and Abbod, 2016, Li and Chen, 2020]. It measures the mean square difference between the predicted probabilities given by the model and the actual outcomes. Mathematically it can be defined as follows:

$$\text{BS} = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2 \tag{10}$$

where p_i represents the predicted probability, o_i denotes the observed outcome (either 0 or 1) and N represents the total number of observations. Hence, a lower Brier score indicates better accuracy and calibration of the model's probabilistic predictions. A Brier score of 0 indicates perfect accuracy, meanwhile a Brier score of 1 indicates complete inaccuracy.

5.10.4 Log loss

Log loss (LL), often referred to as cross entropy, is a commonly used metric for evaluating the performance of classification models in scenarios where the outputs are probabilities for each class. Similarly to Brier score, it measures the dissimilarity between the predicted probability distributions and the actual observed class labels. It can be defined as below:

$$\mathrm{LL} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)\right) \tag{11}$$

where p_i represents the predicted probability for the positive class, y_i denotes the true class label(0 or 1) and N represents the total number of observations. A lower log loss indicates better alignment of predicted probabilities and actual outcomes. Hence, log loss serves as another measure in assessing the performance of binary classification models [Roberts, 2023].

6 Methodology

6.1 Initial data management

The dataset used in this study was collected by Länsförsäkringar's risk management department and is the same dataset used when building Länsförsäkringar's current PD model. The model covers private, non-corporate customers with loans at Länsförsäkringar, both unsecured and secured loans with the majority of the customers holding mortgage loans. The dataset contains a variety of variables covering behavioral, financial and qualitative features collected both from Länsförsäkringar's own databases and from external sources covering credit debt and income. The structure is such that for each year the current customers are listed with the features presented above, as well as information about whether or not they defaulted as a customer over the course of the following year. The data was provided as files of yearly data from 2007-2019 and it was decided, similarly to the building of Länsförsäkringar's current model, to build the challenger models on the years 2014-2017. Data from 2018 and 2019 was saved to be used as test data.

The variables in the dataset are continuous, binary and categorical. The categorical variables are transformed to binary using dummy variables. The response variable is binary and takes the value of 1 if a customer defaulted in the following year and 0 otherwise. Since the observations of the dataset corresponds to accepted customers of Länsförsäkringar, the number of defaults is very low, less than 1%. This results in a highly imbalanced dataset.

6.1.1 Statistical analysis

The initial step was to conduct a statistical analysis of the dataset to get a picture of the overall statistics of the variables, including how many points were missing and how balanced the dataset was. The statistical analysis showed that most data points had missing values and that the dataset was highly imbalanced, i.e. the number of defaults was a small fraction of the observations. To manage this it was decided that the SMOTE method would be used to oversample the minority class to 50% of the size of the majority class and that further analysis of the variables was needed to combat the issue of missing values.

Moreover, the dataset contained an extensive amount of variables, some being duplicates and identification variables, e.g. dates of events for customers. An initial manual selection of what variables to include was therefore conducted. The selection was based on the factors discovered important for credit risk evaluation in the literature review as well as on discussions with Länsförsäkringar. The first variable selection hence resulted in 57 variables covering behavior, financial and qualitative features of the customer. Remaining data management followed the flow visualized in Figure 1 and will be discussed in more detail below.

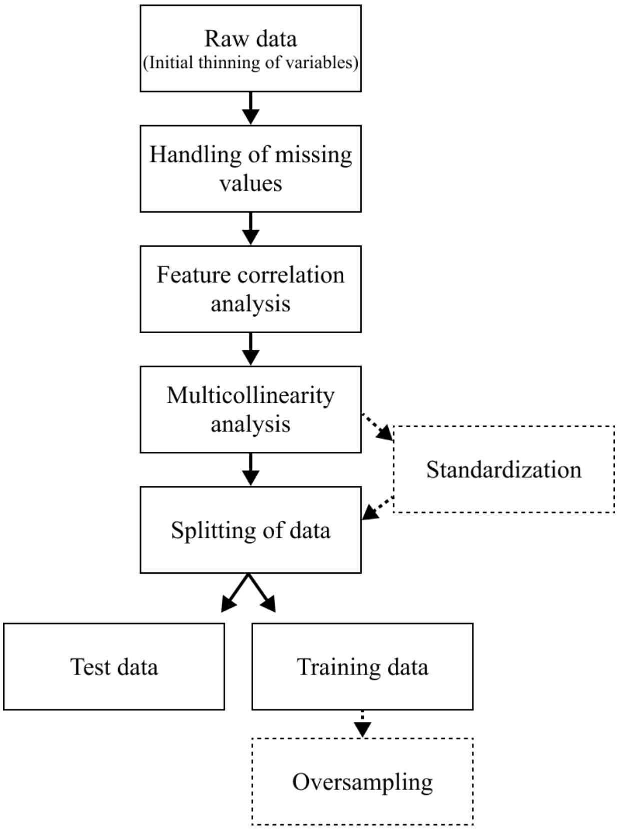

Raw data
(Initial thinning of variables)
Handling of missing values
Feature correlation analysis
Multicollinearity analysis
Standardization
Splitting of data
Test data
Training data
Oversampling

Figure 1: Flow of data handling

6.1.2 Missing values

As mentioned the statistical analysis showed that the dataset had issues with missing values. First, a *complete case analysis* was considered, but as this would result in a reduction of observations by 92% it was regarded as infeasible. As the variables were of different types and ranges it was decided that a study of what a missing value actually meant for each variable had to be conducted. For example, a missing value for a financial variable such as assets could mean that there were no information available about the customer's assets but one would assume that the customer had such not registered at Länsförsäkringar. On the other hand, a missing value for a qualitative or behavioural variable indicates that since no information is known, the value is probably 0, i.e. the feature is not true for that specific customer. Such a variable might be if the customer is deceased. Under this study it became clear that it was not always possible to know why there were missing values, whether it was due to incomplete data collection or human factors. It was hence decided that since no additional customer information could be collected, one way of handling all missing values would be by replacing them with 0 as this was believed to be the most accurate interpretation of a missing value. However, as the reason behind every missing value was not identified, it was decided to create a second dataset where median-imputation was applied. This was done in order to identify if there exist any variables which were more affected by the choice of imputation method. The implementation of the two imputation methods could be used to decide which of the two methods to be most suitable for each variable. This ultimately resulted in the creation of two datasets and the parallel development of models, applying all algorithms to each dataset.

6.1.3 Monotonic binning

Monotonic binning was initially used to transform the continuous variables into optimal and monotonic categorical ones. The implementation however did not prove to be beneficial to the dataset and had issues transforming the data into even buckets. After discussions with Länsförsäkringar and analysis of the impact of the method it was hence decided to not use monotonic binning for the model development.

6.1.4 Feature correlation analysis

As the first thinning of variables decreased the amount to 57 variables, it was discussed how to reduce the amount further to facilitate future work. It was hence decided that a bivariate analysis were to be performed, where each predictor variable was analysed together with the response variable. The measures analyzed using hypothesis testing were Pearson's, Spearman's and Kendall's. If any of these measures got a p-value of less than the significance level of 5% the variable would be kept.

6.1.5 Multicollinearity analysis

In order to reduce redundancy in the model pairs of predictor variables were analyzed for correlation. A common used cut-off value for bivariate correlation is ±0.8 [Chan et al., 2022]. Hence, if the correlation between a pair of predictor variables was greater than or equal to a threshold of ±0.8, the variable with least correlation with the response variable was dropped from the dataset.

6.1.6 Standardization

The last step of the data preprocessing was to standardize the continuous variables in order to ensure all features were on the same scale limiting the risk of bias and improving model performance. Standardization, or z-score, was calculated for each x using the following formula:

$$z_{ki} = \frac{x_{ki} - \bar{x}_k}{s_k} \tag{12}$$

with k being the variable, x_{ki} being the original observation, $\bar{x}_k$ the mean of variable k and s_k the standard deviation. The standard deviation of each sample was calculated as follows:

$$s_k = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ki} - \bar{x}_k)^2} \tag{13}$$

and the mean as

$$\bar{x}_k = \frac{1}{n}\left(\sum_{i=1}^{n} x_{ki}\right) \tag{14}$$

where n is the sample size [Montgomery et al., 2012].

This method was only applied when building and training the ANN, as tree algorithms are argued not to be sensitive to the scale of the input features.

6.1.7 Splitting of data

When all above was applied, the data was split into two sets; test data and training data. The test data consisted of all observations from 2018 (and later on 2019), while the other group consisted of the data from 2014-2017. Lastly, the minority class in the training data was oversampled to 50% of the majority class before training the models.

6.2 Model building

Initially each model was created using some default parameters. Then, for each model, Random Forest, XGBoost and ANNs, hyperparameter tuning and RFE was used to find the best possible model fit. Hyperparameter tuning was used to find the hyperparameters resulting in the best possible model for each of the three algorithms and RFE was used to select the optimal subset of variables giving the highest accuracy.

As the problem with the imbalanced datasets is greater for ANNs it was decided that two different ANN models would be built, one that would be trained using oversampled data and one using the original dataset after variable selection and removal of correlated variables.

6.2.1 Variable selection - RFE

RFE was performed for each model with 5-fold cross validation. The measure to optimize was chosen to be AUC and the method resulted in using 40 features for the Random Forest model, 13 features for the XGBoost model and 18 features for the ANN model, when applied on the zero-imputed dataset. Corresponding numbers for the models built on the median-imputed dataset were 42, 13 and 34.

6.2.2 Hyperparameter tuning

For each model, different types of hyperparameters were tuned. For the tree based models this included the dept of the trees, number of trees, learning rate and maximum number of terminal nodes. For ANN this included number of neurons in each layer and what kind of activation function to be use.

6.3 Model evaluation and comparison

The models' predictive abilities were evaluated using AUC ROC as a measure of accuracy. Log loss and Brier score were calculated to compare the accurate and well-calibrated probability estimates of the models. Migration matrices were used to track and compare movements between risk grades year to year. The AUC was also used to analyze overfitting, as a high model accuracy on the training data with a low precision on test data could indicate overfitting. Each model, Random Forest, XGBoost and ANN (with and without oversampling), was trained and tested on both zero-imputed and median-imputed datasets to analyze the impact of missing value management of the dataset.

7 Results

7.1 Preprocessing of data

The initial manual variable selection resulted in 57 variables being retained. When applying the two different imputation methods for managing missing values, *zero imputation* and *median imputation*, both yielded the same result in terms of features correlation analysis and multicollinearity analysis. Five variables did not pass the criteria of having a significance level of 5% and were hence removed. This was a minor partition of the variables, giving the indication that the chosen variables were relevant for building the model.

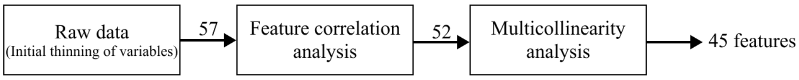

Raw data
(Initial thinning of variables)
57
Feature correlation analysis
52
Multicollinearity analysis
45 features

Figure 2: Initial selection of variables

Furthermore, the multicollinearity analysis using Pearson's correlation coefficient and a cutoff value set to ±0.8 resulted in the removal of 7 variables. As can be seen in Figure 2 these initial selections yielded 45 variables. The multicollinearity analysis is visualized using the heatmap in Figure 3, where a darker shade indicates stronger correlation and a lighter shade a weaker correlation. The values and heatmaps for the different imputation methods were almost identical, resulting in the same deletion of variables. However, the presented visualization is from the analysis where *median imputation* was used.

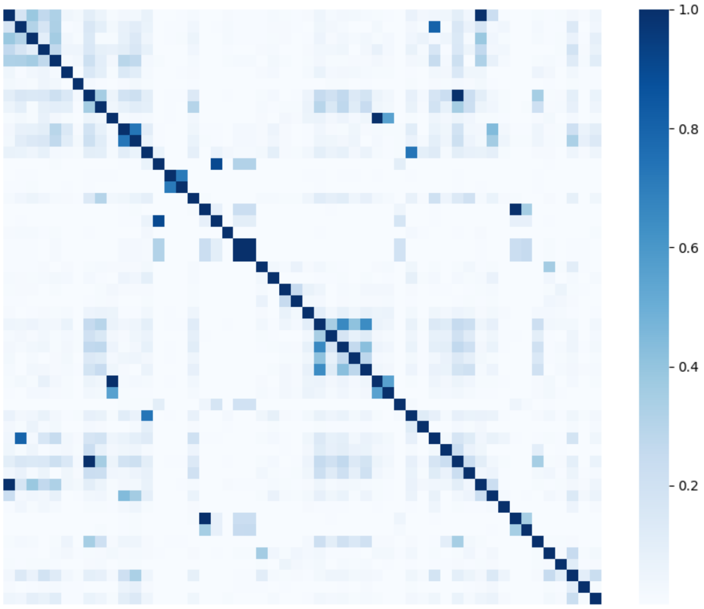
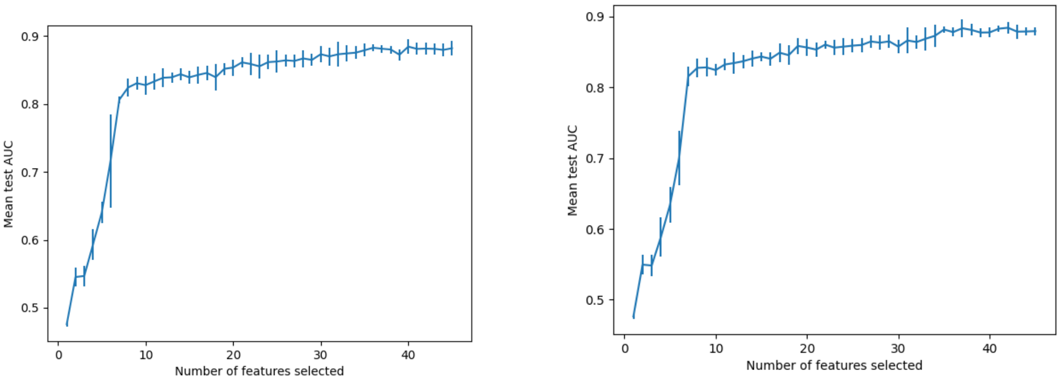


Figure 3: Heatmap of features correlations

7.2 Variable selection

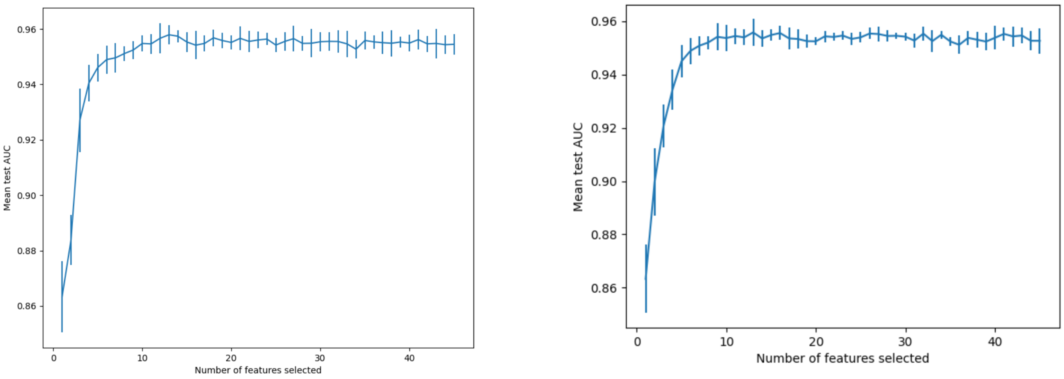
For every model and method for managing missing values, variable selection was carried out. Hereafter the two methods used to manage missing values will be called *zero imputation* and *median imputation* and each plot will be labeled with the associated method. The following plots show how the mean value of AUC obtained by cross validation with 5 folds changes with the number of variables as well as the corresponding error bars representing the standard deviation for every number of selected features by RFE. Python's *StratifiedKfold()* function was applied to ensure that the positive class was evenly distributed between the folds.



(a) With *zero imputation*

(b) With *median imputation*

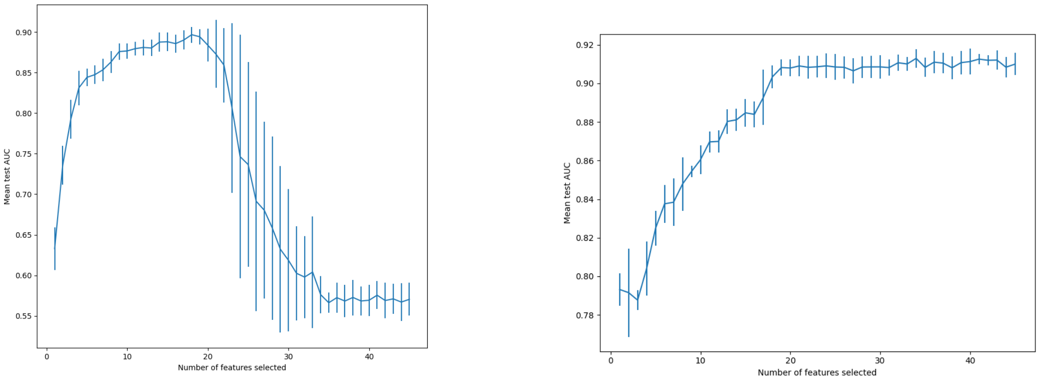
Figure 4: RFECV for Random Forest



(a) With *zero imputation*

(b) With *median imputation*

Figure 5: RFECV for XGBoost



(a) With *zero imputation*

(b) With *median imputation*

Figure 6: RFECV for ANN (based on logistic regression)

The optimal number of features chosen by application of RFE using cross validation with AUC as score were given by the algorithm for each model. However, these numbers were not always the numbers of features chosen. In addition, a visual interpretation of how the AUC score varied with number of features was conducted. When the increase in AUC started to decrease and an increase in number of features did not contribute significantly to the model's performance, a model with a smaller number of features was preferred for simplicity. The number of optimal features obtained by RFECV and chosen can be found in Table 1.

Method	**Optimal number of features by RFECV**	**Chosen number of features**
Random Forest$_{zero}$	40	**21**
Random Forest$_{median}$	42	**19**
XGBoost$_{zero}$	13	**13**
XGBoost$_{median}$	13	**13**
ANN$_{zero}$	18	**18**
ANN$_{median}$	34	**19**

Table 1: Optimal number of variables by RFECV and the number of variables chose

When only observing the graphs of the RFECV for **Random Forest** in Figure 4 the graph is steep in the beginning, showing it is possible to build a model only consisting of 7 features that achieve an AUC score of > 0.80. When adding more than 7 variables the graph still seems to increase but this time with a smaller linear effect on AUC. The algorithm determined the optimal number of features to be 40 and 42 respectively for the two imputation methods. However, a discussion about the number of variables was held where it was concluded to only keep 21 and 19 variables. This decision was taken on the ground of wanting to avoid overfitting by keeping the complexity of the model in check. The effect of adding the remaining ~ 20 features resulted in an improvement in the performance metric AUC of $\sim$ 0.01-0.02, which was considered unjustifiable. The variables selected by RFE can be found in Table 2 and 3. These were almost exactly the same, with all 19 variables selected for the model built on median-imputed data being selected for the model built on zero-imputed data, with two additional variables.

As can be seen in Figure 5 the AUC using **XGBoost** and only including one feature was $\sim$0.86, which shows that it is possible to build a fairly good model with only one feature. This feature was *Beteende 3*. With 2-3 features the model received an AUC score of > 0.90. When adding additional features the performance measure stabilized and increased with slower pace. It could be argued that to use less than 13 features is preferred, as the increase in AUC when adding the last 8 features is only $\sim$0.01. The peak in AUC value was however reached with 13 features included and that was also the chosen number of variables for both imputation methods. The names of the features can be found in Table 2 and 3. The imputation methods gave very similar sets of features, with only two features setting them apart, meaning the two imputation methods had 11 variables in common.

Similar to XGBoost, the **ANN** using zero imputation receives a high AUC score when including only 4 features. Thereafter the increase is not as steep. As seen in Figure 6, after adding 18 features the AUC declines for each added feature. Worth noticing here is that the error bars span over a greater area as well, indicating that there is more uncertainty observed when using these variables. The graph of the AUC for median imputation does not decrease as quickly as when using zero imputation, but starts instead with a greater value in AUC for the first feature and stabilizes around 19 features. The algorithm returned 34 as the optimal number of features when using median imputation, as this resulted in the highest AUC. However, as the graph shows, the performance metric is quite steady and differs a maximum of 0.01 in AUC between 19 and 34 features. It was hence decided to move forward using 19 variables to limit the risk of creating a too complex model. Out of the chosen variables 12 were the same for the different imputation methods.

The different variables chosen by RFE and included in the different models can be found in Table 2 and 3. In both tables, the variables encoded in blue represents those selected by RFE for two different models, while the variables encoded in pink were selected by RFE for all three models.

Chosen variables - *zero imputation*

Random Forest$_{zero}$	**XGBoost**$_{zero}$	**ANN**$_{zero}$
Beteende 3	Beteende 3	Beteende 3
Beteende 4	Beteende 4	Beteende 4
Beteende 11	Beteende 5	Beteende 12
Beteende 12	Beteende 9	Beteende 13
Beteende 13	Beteende 14	Beteende 14
Beteende 14	Beteende 15	Beteende 16
Beteende 15	Beteende 17	Beteende 17
Beteende 16	Finansiell 4	Finansiell 1
Finansiell 1	Finansiell 6	Finansiell 4
Finansiell 2	Finansiell 14	Finansiell 6
Finansiell 3	Kundattribut 1	Finansiell 10
Finansiell 4	Kvalitativ 2	Kundattribut 1
Finansiell 6	Kvalitativ 4	Kundattribut 2
Finansiell 7		Kundattribut 3
Finansiell 8		Kundattribut 4
Finansiell 10		Kundattribut 10
Finansiell 14		Kvalitativ 1
Finansiell 15		Kvalitativ 2
Kundattribut 8		
Kundattribut 12		
Kvalitativ 3		

Table 2: Features chosen by RFE with set number of features. Blue color indicates that the feature was selected for two models, while pink color indicates that the feature was selected for all models.

Chosen variables - *median imputation*

Random Forest$_{median}$	**XGBoost**$_{median}$	**ANN**$_{median}$
Beteende 3	Beteende 3	Beteende 3
Beteende 4	Beteende 4	Beteende 4
Beteende 11	Beteende 5	Beteende 9
Beteende 12	Beteende 9	Beteende 12
Beteende 13	Beteende 15	Beteende 13
Beteende 15	Finansiell 4	Beteende 16
Finansiell 1	Finansiell 6	Finansiell 3
Finansiell 2	Finansiell 14	Finansiell 4
Finansiell 3	Kundattribut 1	Finansiell 6
Finansiell 4	Kundattribut 3	Kundattribut 2
Finansiell 6	Kvalitativ 1	Kundattribut 3
Finansiell 7	Kvalitativ 2	Kundattribut 8
Finansiell 8	Kvalitativ 4	Kundattribut 10
Finansiell 10		Kundattribut 12
Finansiell 14		Kundattribut 13
Finansiell 15		Kvalitativ 1
Kundattribut 8		Kvalitativ 2
Kundattribut 12		Kvalitativ 3
Kvalitativ 3		Kvalitativ 4

Table 3: Features chosen by RFE with set number of features. Blue color indicates that the feature was selected for two models, while pink color indicates that the feature was selected for all models.

7.3 Tuning of hyperparameters

Below the result from applying randomized search for each model and their chosen hyperparameters is presented, using both zero imputation and median imputation. The values in the "Tested values" column represent the values that were tested while the values that were considered to be optimal are presented in bold font. An important point to make is that the AUC value remained relatively stable regardless of the hyperparameters used.

Random Forest

Hyperparameter	**Tested values**	**Optimal$_{zero}$**	**Optimal$_{median}$**
Number of trees	100, 200, 300	**200**	**300**
Maximum depth	None, 10, 20, 30	**10**	**10**
Minimum samples split	2, 5, 10	**2**	**5**
Minimum samples leaf	1, 2, 4	**1**	**4**
Max features	auto, sqrt	**sqrt**	**sqrt**
Bootstrap	True, False	**True**	**True**

Table 4: Hyperparameters for Random Forest

XGBoost

Hyperparameter	**Tested values**	**Optimal$_{zero}$**	**Optimal$_{median}$**
Number of trees	100, 200, 300	**200**	**200**
Maximum depth	3, 5, 7, 10	**3**	**3**
Minimum child weight	1, 3, 5	**3**	**3**
Subsample	0.5, 0.7, 0.9	**0.7**	**0.7**
Colsample by tree	0.5, 0.7, 0.9	**0.9**	**0.9**
Gamma	0, 0.1, 0.2	**0.1**	**0.1**
Learning rate	0.01, 0.1, 0.2	**0.1**	**0.1**

Table 5: Hyperparameters for XGBoost

ANN

Hyperparameter	**Tested values**	**Optimal**$_{zero}$	**Optimal**$_{median}$
Number of neurons	32, 64, 128	**32**	**32**
Activation function	relu, tanh, sigmoid	**tanh**	**sigmoid**

Table 6: Hyperparameters for ANN

Hyperparameters in Table 4 for Random Forest are somewhat similar between the imputation methods, but the optimal number of trees, minimum sample splits and minimum number of sample leafs differs. The change in AUC when these parameters were changed was not remarkably large but rather negligible. The same holds for the result of XGBoost. The algorithm for randomized search resulted in the same set of optimal hyperparameters for both methods of imputation, found in Table 5. This is not unexpected considering that the two models had 11 variables in common.

For ANN the number of neurons used was the same for both imputation methods favoring the lowest tested number; 32. A smaller number of neurons often results in a model which has a weaker ability to capture complex patterns, but it also keeps the model less computationally complex making it less likely to be overfitted. In terms of activation functions both tanh and sigmoid are non-linear functions following an s-shape able to capture more complex relations, compared to the linear relu function. Tanh was beneficial to be used in combination with zero imputation while the use of median imputation imply the sigmoid function to be the preferred activation function.

7.4 Model performance

Table 7 contains our selected performance measures for model comparison. As can be seen in the table below XGBoost is the preferred model as it performs the best on all measures.

Model	**AUC**	**Brier score**	**Log loss**
Random Forest$_{zero}$	0.9330	0.0013	0.0072
Random Forest$_{median}$	0.9331	0.0013	0.0071
XGBoost$_{zero}$	0.9523	0.00128	0.00651
XGBoost$_{median}$	0.9511	0.00129	0.00656
ANN$_{zero} - SMOTE$	0.9128	0.0316	0.1130
ANN$_{median} - SMOTE$	0.9182	0.0466	0.1646
ANN$_{zero}$	0.9305	0.0013	0.0071
ANN$_{median}$	0.9412	0.0013	0.0069

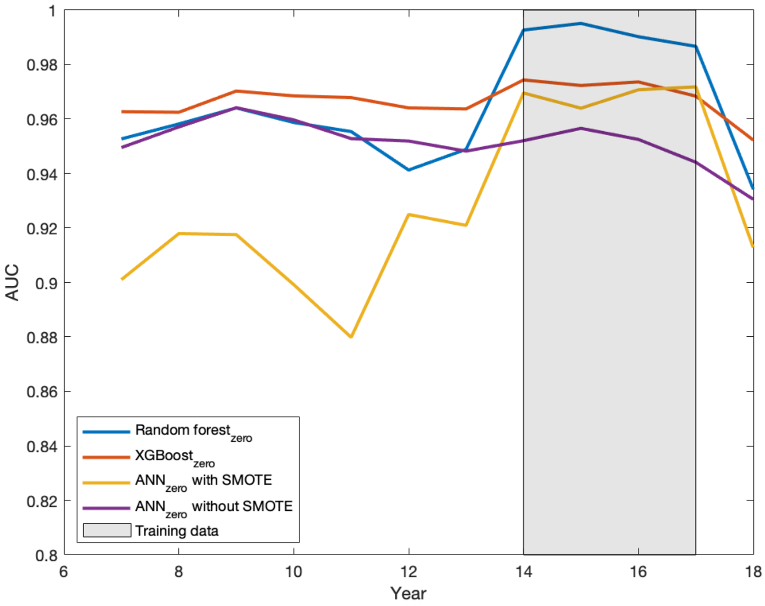
Table 7: Performance Metrics of different Models, based on year 2018

In Table 7 it is clear that models built using Random Forest algorithm were able to receive a high AUC on the test data. Similar to the results by Kruppa et al. [2013], the performance of the Random Forest model is excellent. It receives a score which would be regarded as acceptable for an actual implementation of the model. When looking at the graphs in Figure 7 it can on the other hand be seen that the models seem to be considerably overfitted to the years of the training data. During these years the models provide an AUC score of close to 1, and drops to $\sim$ 0.93-0.95 on the out-of-sample years. This indicates that the trees in the models most likely have some splits in them that only hold for the small amount of defaulting observations in the training data, which are not applicable to the observations in the test data. This can also be seen in the subgroup plots in appendix, Figure 14.

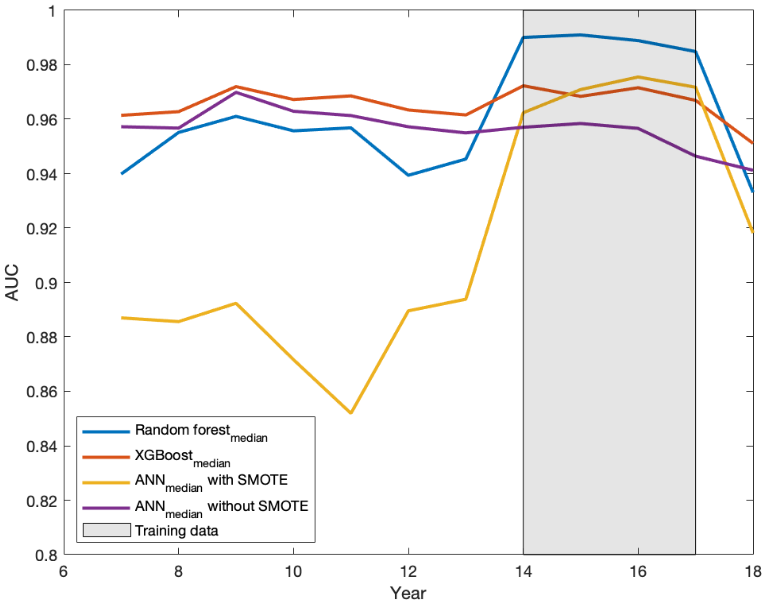
When looking at the model performance scores of XGBoost it is possible to conclude that it is a stable and well performing algorithm for the purpose and data of this thesis. Both models of different data imputation methods receive an AUC score of > 0.95 and a brier score close to zero, indicating good model performance for probabilistic classification. Looking at the plot of AUC over the years 2007-2018 it could be argued that the models are slightly overtrained on the specific years of the training data, but it is not to a great extent as this results in a fluctuation of around 0.01 in AUC.

As can be observed in Table 7 the models trained on oversampled data using SMOTE received a lower AUC score on the test data. This is also clear when looking at the graphs in Figure 7. Here, the AUC curve of ANN with SMOTE is superior to ANN without SMOTE for the years of the training data, but performs worse on out-of-sample data. This is the case for both imputation methods. It is possible that the very low default rate causes this problem. Since the default rate, i.e. observations of the positive class, is lower than 1% of the data it is not surprising that the models are showing signs of overtraining when this small part of the data is oversampled to being such a big part of the dataset. Here, it could be argued that SMOTE should have been used with a smaller amount of oversampled data, e.g. 10% instead of 50%. If one looks at the ANN models without oversampling the AUC curve is much more stable over time, even though it is decreasing a little in the year 2018. The brier score is significantly greater for the models trained on oversampled data, indicating once again that the models trained on data without oversampling are superior.

Figure 7a and 7b show graphs of the AUC score for each model over the years. The grey area marks the years of training data.



(a) With *zero imputation*

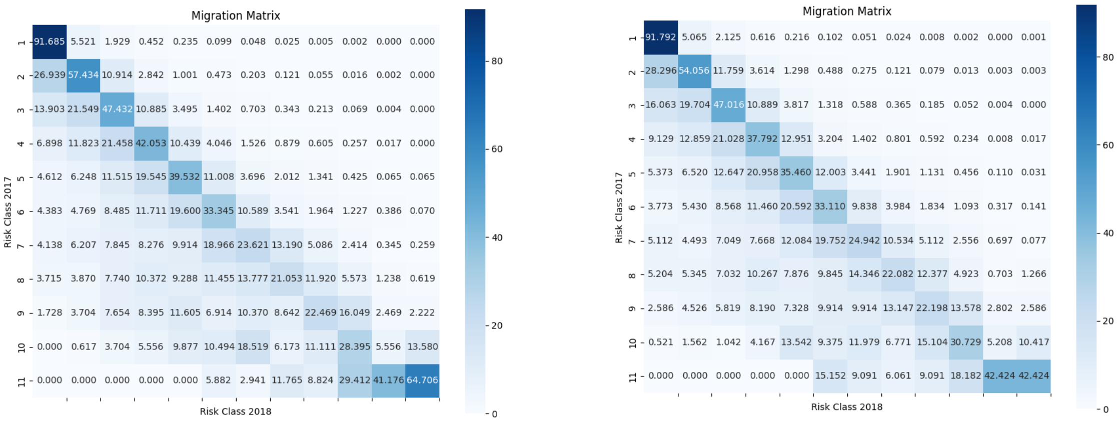


(b) With *median imputation*

Figure 7: AUC over the years with different models

7.4.1 Migration matrix

The number within a cell in a matrix, at position (i,j) indicates the percentage of customers moving from risk class i year 2017 to risk class j year 2018. On the right is an additional column indicating the actual default rate for risk class i.



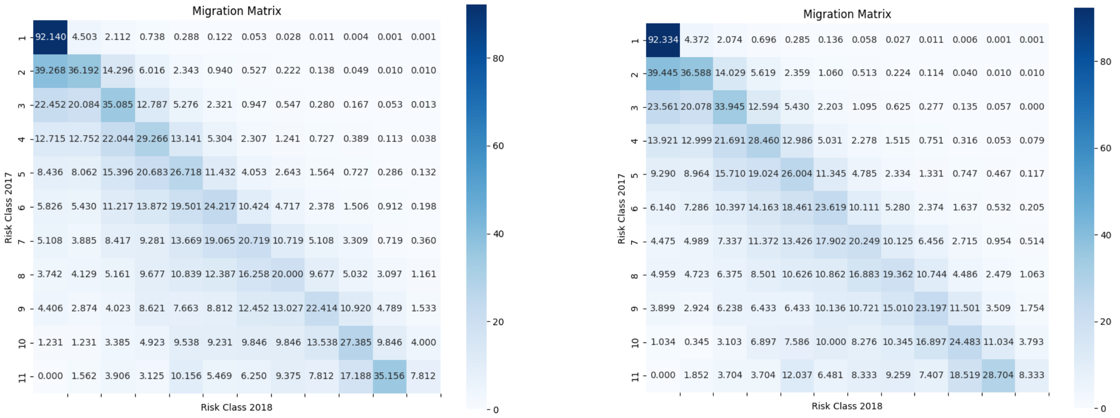
1	91.685	5.521	1.929	0.452	0.235	0.099	0.048	0.025	0.005	0.002	0.000	0.000
2	26.939	57.434	10.914	2.842	1.001	0.473	0.203	0.121	0.055	0.016	0.002	0.000
3	13.903	21.549	47.432	10.885	3.495	1.402	0.703	0.343	0.213	0.069	0.004	0.000
4	6.898	11.823	21.458	42.053	10.439	4.046	1.526	0.879	0.605	0.257	0.017	0.000
5	4.612	6.248	11.515	19.545	39.532	11.008	3.696	2.012	1.341	0.425	0.065	0.065
6	4.383	4.769	8.485	11.711	19.600	33.345	10.589	3.541	1.964	1.227	0.386	0.070
7	4.138	6.207	7.845	8.276	9.914	18.966	23.621	13.190	5.086	2.414	0.345	0.259
8	3.715	3.870	7.740	10.372	9.288	11.455	13.777	21.053	11.920	5.573	1.238	0.619
9	1.728	3.704	7.654	8.395	11.605	6.914	10.370	8.642	22.469	16.049	2.469	2.222
10	0.000	0.617	3.704	5.556	9.877	10.494	18.519	6.173	11.111	28.395	5.556	13.580
11	0.000	0.000	0.000	0.000	0.000	5.882	2.941	11.765	8.824	29.412	41.176	64.706

(a) With *zero imputation*

1	91.792	5.065	2.125	0.616	0.216	0.102	0.051	0.024	0.008	0.002	0.000	0.001
2	28.296	54.056	11.759	3.614	1.298	0.488	0.275	0.121	0.079	0.013	0.003	0.003
3	16.063	19.704	47.016	10.889	3.817	1.318	0.588	0.365	0.185	0.052	0.004	0.000
4	9.129	12.859	21.028	37.792	12.951	3.204	1.402	0.801	0.592	0.234	0.008	0.017
5	5.373	6.520	12.647	20.958	35.460	12.003	3.441	1.901	1.131	0.456	0.110	0.031
6	3.773	5.430	8.568	11.460	20.592	33.110	9.838	3.984	1.834	1.093	0.317	0.141
7	5.112	4.493	7.049	7.668	12.084	19.752	24.942	10.534	5.112	2.556	0.697	0.077
8	5.204	5.345	7.032	10.267	7.876	9.845	14.346	22.082	12.377	4.923	0.703	1.266
9	2.586	4.526	5.819	8.190	7.328	9.914	9.914	13.147	22.198	13.578	2.802	2.586
10	0.521	1.562	1.042	4.167	13.542	9.375	11.979	6.771	15.104	30.729	5.208	10.417
11	0.000	0.000	0.000	0.000	0.000	15.152	9.091	6.061	9.091	18.182	42.424	42.424

(b) With *median imputation*

Figure 8: Migration matrices for Random Forest



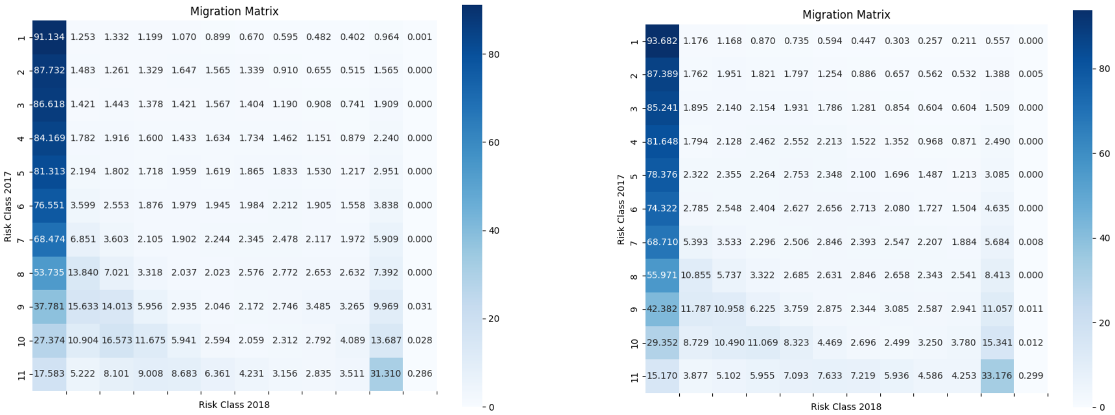
1	92.140	4.503	2.112	0.738	0.288	0.122	0.053	0.028	0.011	0.004	0.001	0.001
2	39.268	36.192	14.296	6.016	2.343	0.940	0.527	0.222	0.138	0.049	0.010	0.010
3	22.452	20.084	35.085	12.787	5.276	2.321	0.947	0.547	0.280	0.167	0.053	0.013
4	12.715	12.752	22.044	29.266	13.141	5.304	2.307	1.241	0.727	0.389	0.113	0.038
5	8.436	8.062	15.396	20.683	26.718	11.432	4.053	2.643	1.564	0.727	0.286	0.132
6	5.826	5.430	11.217	13.872	19.501	24.217	10.424	4.717	2.378	1.506	0.912	0.198
7	5.108	3.885	8.417	9.281	13.669	19.065	20.719	10.719	5.108	3.309	0.719	0.360
8	3.742	4.129	5.161	9.677	10.839	12.387	16.258	20.000	9.677	5.032	3.097	1.161
9	4.406	2.874	4.023	8.621	7.663	8.812	12.452	13.027	22.414	10.920	4.789	1.533
10	1.231	1.231	3.385	4.923	9.538	9.231	9.846	9.846	13.538	27.385	9.846	4.000
11	0.000	1.562	3.906	3.125	10.156	5.469	6.250	9.375	7.812	17.188	35.156	7.812

(a) With *zero imputation*

1	92.334	4.372	2.074	0.696	0.285	0.136	0.058	0.027	0.011	0.006	0.001	0.001
2	39.445	36.588	14.029	5.619	2.359	1.060	0.513	0.224	0.114	0.040	0.010	0.010
3	23.561	20.078	33.945	12.594	5.430	2.203	1.095	0.625	0.277	0.135	0.057	0.000
4	13.921	12.999	21.691	28.460	12.986	5.031	2.278	1.515	0.751	0.316	0.053	0.079
5	9.290	8.964	15.710	19.024	26.004	11.345	4.785	2.334	1.331	0.747	0.467	0.117
6	6.140	7.286	10.397	14.163	18.461	23.619	10.111	5.280	2.374	1.637	0.532	0.205
7	4.475	4.989	7.337	11.372	13.426	17.902	20.249	10.125	6.456	2.715	0.954	0.514
8	4.959	4.723	6.375	8.501	10.626	10.862	16.883	19.362	10.744	4.486	2.479	1.063
9	3.899	2.924	6.238	6.433	6.433	10.136	10.721	15.010	23.197	11.501	3.509	1.754
10	1.034	0.345	3.103	6.897	7.586	10.000	8.276	10.345	16.897	24.483	11.034	3.793
11	0.000	1.852	3.704	3.704	12.037	6.481	8.333	9.259	7.407	18.519	28.704	8.333

(b) With *median imputation*

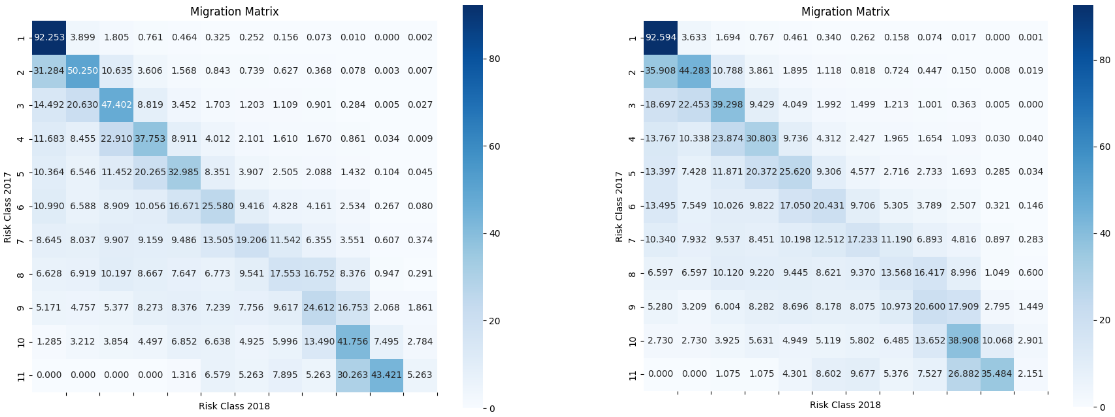
Figure 9: Migration matrices for XGBoost



(a) With *zero imputation*

(b) With *median imputation*

Figure 10: Migration matrices for ANN, with SMOTE



(a) With *zero imputation*

(b) With *median imputation*

Figure 11: Migration matrices for ANN, without SMOTE

The migration matrices in Figure 8 look quite stable for Random Forest, with the observations of the lower risk grades being steady with bell shaped migrations. However there are some questionable migration rates in the higher risk grades. In the far right column one can see that the default rate within each grade seems to be decreasing for the model built on zero imputed data, but has some irregularities in the median imputed data.

For XGBoost the migration matrices show reasonably steady populations within the risk grades. Even though the diagonal is not as steady as preferred, the distribution of the grades in the matrices have similar bell curves centralized around the diagonals which is favourable, although some deviations from the bell shape can be observed. The same holds for the migration matrices for ANN without SMOTE, seen in Figure 18. They look quite steady overall, with the majority of each customer staying within the same risk grade from one year to the next. The risk grades do not display perfect bell shapes, but the diagonals are quite clear, indicating that the migrations from one grade to another is usually done to grades nearby. The same can not be said for the matrices for ANN with SMOTE. For this model most of the customers from year 2017 moves to risk grade 1 in year 2018. Looking at the distribution of risk grades in 2018 in Figure 20 risk grade 11 is the second greatest grade after risk grade 1 and the rest of the grades are quite evenly sized.

7.4.2 Distribution of risk classes and performance of subgroups

To get an overview of what the distribution of the risk classes look like histograms of the predicted risk classes were plotted for each model. These can be found in Figure 18a,19, 20 and 21 in Appendix. The histogram bars are declining with increased risk classes for all models except for the ANN using SMOTE.

Plots of the performance of subgroups for each model can be found in the appendix. Similar graphs as in Figure 7 were created for three different subgroups for each model. The chosen subgroups were observations holding unsecured loans, mortgage loans and observations not in possession of any savings. They were all fairly stable and alike, except for the ANN using SMOTE where the subgroup of mortgage loans showed a noticeable deterioration in performance for several years. This is not a problem for ANN without SMOTE and indicates once again that using oversampling of this size leads to overfitting. The plots can be found in Figure 15, 16 and 17.

7.5 Final model

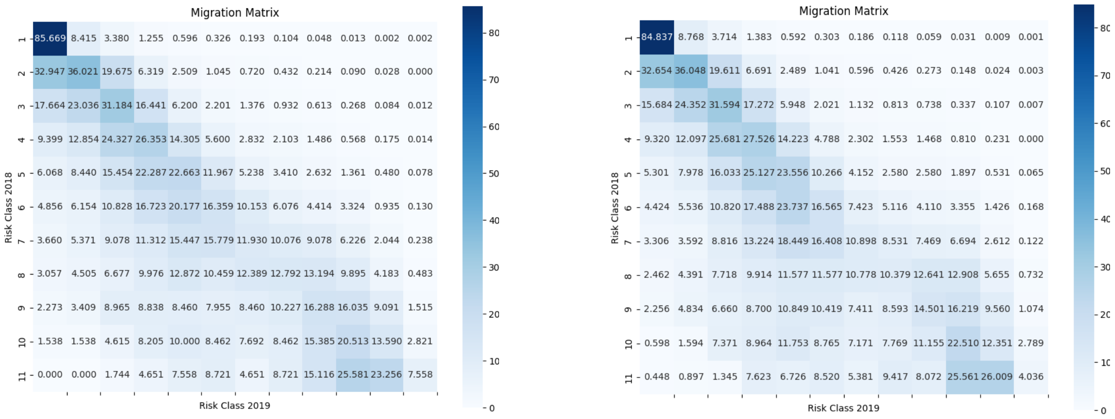
Looking at the different models' performances in Table 7 it is clear that XGBoost, whether it is using zero imputation or median imputation, performs better than the other models with regards to all measures used to assess model performance. The results in brier score and the log loss measures are fairly comparable for ANN without SMOTE, Random Forest and XGBoost, as there are only minor differences between them. One can argue that the same holds for AUC, as there is at a maximum only two percentage points differing between the models. Still, the AUC for the models built using XGBoost is greater and when looking at the plots of AUC in Figure 7 XGBoost is superior overall (with Random Forest and ANN with SMOTE on the training years being the only exception). Indeed, a higher AUC could have been achieved with a greater amount of variables in the ANN models and the Random Forest models. The XGBoost models did not however require as many variables as ANN to reach a high AUC score and they had the most steady performances over the years as well. Hence, XGBoost was considered to be the best algorithm and the models built using it were thus also chosen as the final models. These models were tested on new validation data from the year 2019 with results presented below.

Final model	**AUC**	**Brier score**	**Log loss**
XGBoost$_{zero}$	0.9534	0.0012	0.0066
XGBoost$_{median}$	0.9527	0.0013	0.0068

Table 8: Performance Metrics of final model, based on year 2019

AUC				
Imputation	**All**	**Blank loan**	**Mortgage loan**	**Non saving**
zero	0.9534	0.9392	0.9496	0.9233
median	0.9527	0.9364	0.9511	0.9255

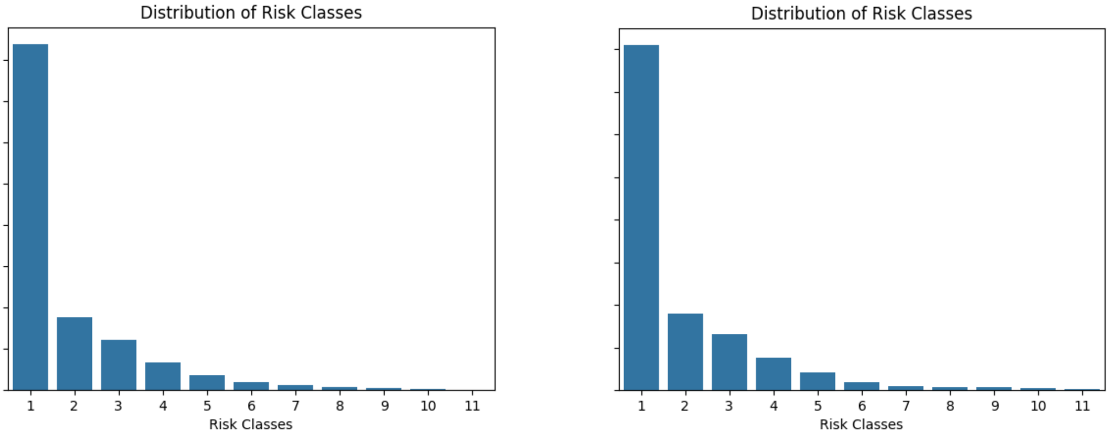
Table 9: AUC of subgroups, based on year 2019



(a) With *zero imputation*

(b) With *median imputation*

Figure 12: Migration matrices for final model



(a) With *zero imputation*

(b) With *median imputation*

Figure 13: Distribution of risk classes based on year 2019

The results of the final models are as expected after observing the previous performances of the models. The AUC is similar to that received when using the data from 2018, although being a little greater in 2019 while the brier score and log loss are about the same (the sets of observations for each year were around the same size). Similarly to data from 2018, the performance of different subgroups are overall steady with the non-saving subgroup being the one that differs the most from the overall portfolio. This can be seen as an indication of limited overtraining. The migration matrices are not looking as favourable as desired, as the diagonals are not well-defined, meaning there are many customers moving between risk grades. This was a problem when looking at the matrices for 2017-2018 as well. In most risk grades the distribution of the migrations look fairly bell shaped, but there are some exception in the grades > 7. However, the migration matrices are not a measure of model performance but rather a tool for analyzing the behaviour of the customers as the matrices give a clear representation of the risk grades and the movement of the portfolio of customers. It is desired to have customers staying in the same or nearby grade from one year to the next, as it affects the capital adequacy requirement [Rosquist, 2024]. The capital adequacy requirement in return is wanted by credit institutions to not increase unpredictably from one year to the next. Nonetheless, behavior from customers can change from one year to another changing their probability of default and hence their risk grades. The risk grades correspond to the same risk intervals currently used in Länsförsäkringar's risk grade. It is however possible to receive different migration matrices and risk grade distributions with other intervals.

Lastly, there was an interest in investigating the feature importances of the chosen features for the final model. Here two different measures were applied on the two different sets of features (each set of the two imputation methods). The measures used were *Permutation feature importance* and XGBoost's implementations own measure, *gain*. Gain can be explained as being the reduction in the model's loss function achieved by adding a split based on the feature. A higher gain indicates that a feature has a larger impact on improving the model's predictive accuracy. The results of this measure can be found in Table 10, where "X" indicates that the respective feature was not a part of this model's set of features.

Variable	$\mathbf{PFI}_{zero}$	$\mathbf{Gain}_{zero}$	$\mathbf{PFI}_{median}$	$\mathbf{Gain}_{median}$
Beteende 3	0.0459	80.5	0.0463	83.8
Finansiell 4	0.0100	22.6	0.0114	20.9
Finansiell 6	0.0085	15.8	0.0068	12.6
Beteende 17	0.0078	22.3	X	X
Kvalitativ 4	0.0074	24.5	0.0083	24.6
Kvalitativ 1	X	X	0.0055	17.2
Kvalitativ 2	0.0045	11.0	0.0043	10.4
Beteende 4	0.0031	22.3	0.0034	20.8
Beteende 14	0.0015	10.6	X	X
Finansiell 14	0.0009	8.5	0.0007	9.5
Beteende 15	0.0008	9.3	0.0010	8.7
Beteende 9	0.0007	13.5	0.0008	11.4
Beteende 5	0.0005	13.9	0.0008	15.0
Kundattribut 1	-0.0004	8.0	-0.0007	8.6
Kundattribut 3	X	X	0.0009	18.4

Table 10: Feature importances for final model

8 Discussion

8.1 Performance of models

The conclusion that the tree based methods are the best in terms of chosen performance metrics are not unexpected, since as discussed in the literature review so was the conclusion of other similar studies made [Addo et al., 2018, Kruppa et al., 2013, Grinsztajn et al., 2022]. Similarly to the discovery by Grinsztajn, XGBoost performs superior to ANN on the tabular data used in this study. This goes against the findings in the studies of Shi et al. [2022] and Azzone et al. [2022] which both found neural networks models to be superior in regards of the AUC measure. The neural network in this thesis does not seem to perform as bad as it did in Grinsztajn's study, which could be explained by the size of the dataset. It is mentioned in Grinsztajn's study that it is crucial to use a big dataset when training neural networks, which can be argued to be the case when training the ANN models in this thesis. Nonetheless the data was highly unbalanced with the minority class being less than 1% of the dataset and when this minority class was oversampled with SMOTE it resulted in a model overtrained on the positive class of the training data. This resulted in an almost perfect AUC score for the years of the training data and significantly lower AUC for the out-of-sample years. It is possible that the results might be better with a smaller ratio of the number of samples in the minority class in relation to the number of samples in the majority class. This study was conducted only using this ratio set to 0.5, but it is a possibility the resulting model would not be as overtrained if the ratio was smaller.

Furthermore, the aforementioned small changes in AUC when tuning the hyperparameters are consistent with the findings of Grinsztajn, confirming that the use of hyperparameter training does not make the model significantly better. This was not the case for the tree based methods either, as they did not get a significantly better AUC when the hyperparameters were tuned. There could be several reasons to this. The initial parameters set, corresponding to each algorithm implementation's default settings, could have been close to optimal values. It is also possible that the search spaces for the parameters were not sufficiently explored. Even though they were selected to represent a bigger span of variables, the span could have been chosen to be greater.

Another thing that can be observed in the results is the difference in performance depending on the imputation method used. As mentioned earlier, it is arguable that a missing value in many of the variables correspond to a zero, as this imply that if there is no information regarding the customer having an unsecured loan with Länsförsäkringar, the customer most likely has no unsecured loan with the bank. However, this is not an assumption that can be argued to be true for all variables in the dataset. This was the reason why the two imputation methods were applied and investigated. For the tree based models, XGBoost and Random Forest, both zero imputation and median imputation yielded very similar sets of variables, having only minor differences between their performances. This result suggests that the choice of imputation method might not affect the models' predictive capabilities significantly. Moreover, the hyperparameter training resulted in identification of the fairly similar set of optimal hyperparameters which did not improve the AUC remarkably. For ANN the methods resulted in two sets of variables that differed a bit more than for the tree based models. Notably though, the AUC did not improve remarkably with hyperparameter tuning as the performance metrics remained fairly stable across the different imputation methods. This holds for all algorithms, which indicates that the models are well-suited for the given data set and task.

If a more advanced machine learning model would be used in this setting it might be relevant to check how to minimize the missing values in the dataset. As mentioned in the study of Mashrur et al. [2020] in the literature review, one critical step in order to success with the implementation of machine learning methods is to implement ways to gather, prepare and manage the big amount of data required. If there was a routine of how to treat missing values along with information of what exactly a missing value of that variable meant, the above challenge of finding suitable imputation method would not be a problem.

As mentioned in the literature review all of the algorithms have some difficulties with transparency and explainability, with neural networks being the most extreme. However, there are methods and scores that can still be used to explain parts of the predictions. For Random Forest and XGBoost there are techniques such as feature importance that can help in giving insights into the model's behavior and the relationship between features and predictions. For the ANN models it is more difficult, but scores like permutation feature importance can still be used. Nevertheless it is impossible to neglect that transparency and interpretability to some extent are sacrificed in exchange for superior predictive performance. The requirements from Finansinspektionen concerning the building and use of IRB models are open for interpretation when describing the requirements for explainability and hence an industry practise is often referenced [Cervenka, 2016]. It could hence be argued that as the models used in this project are performing just as good and in some ways better than logistic regression, with regards to stability of the AUC, the use could be beneficial and plausible.

As can be seen in Table 2 and 3, there are some features which are chosen by all or most of the models. For example, *Beteende 3, Beteende 4, Finansiell 4* and *Finansiell 6* are included in all models while *Beteende 15* is part of all tree models and *Beteende 14* is part of all models built on zero imputed data. The latter is interesting in particular since that discrepancy in variable selection depends on data imputation. It could be explained by the missing values of the variable being clustered around zero or the median, or that the absence/presence of another variable affected the choice. Further analysis of what a missing value of that variable actually mean would be of importance when further expanding the model for implementation.

8.2 Variable importance

The results in Table 10 give insights into the predictive behavior of the final models. By comparing the importance scores of the different models it is possible to get some understanding of how imputation strategies affect the importance rankings. Two features that emerge as significant contributors for both models are *Beteende 3* and *Finansiell 4*. This indicates that they have robust predictive power regardless of imputation method. Additionally, *Kvalitativ 4* is also ranked high in all measures for both models. Looking at the individual models and the different measures it can be identified that the rankings are fairly similar, having the highest scored features in the upper half and lowest in the lower half. This means that the top ∼6 features of Table 10 are important when building an XGBoost model.

It is clear that these kind of measures give insight into which variables are important and affect the predictive ability of the model. However they do not offer direct interpretation of the relationships' direction or magnitude. Compared to the coefficients received when using logistic regression this might seem non-transparent. From the coefficient of logistic regression it is possible to analyze the strength of the association from the magnitude of the coefficient and the direction from the sign of it. Hence, logistic regression offers a direct interpretation of feature importance which is not accessible when using an XGBoost model. The importance scores for the final models can be used to gain insight into a model's decision-making process, but they are not as transparent as the coefficients of the models used today. Consequently it is interesting to discuss in which scenario the limit for the need of transparency is crossed. Since Finansinspektionen's requirements are stated somewhat vague when discussing the need for transparency it could be argued that these measures would hold and that hence the more advanced models, such as XGBoost, could be used.

Furthermore, after discussions with Länsförsäkringar it was clear that the chosen set of variables used in the final XGBoost model did not diverge greatly from the expected result. Most of the features had already been investigated during Länsförsäkringar's internal model development. This again shows that the variables chosen as important in our models are in line with what the market considers important and further strengthens the use of models and methods such as these. The difference between current models and more advanced models is the more advanced ones ability to capture nonlinear relationship.

8.3 Machine learning in risk management

8.3.1 Feasibility of use

As previously discussed, the use of machine learning is beneficial and has many fields of implementation across multiple industries. The benefits of using machine learning are focused on topics such as accuracy, limitation of bias, performance over time and the overall benefits of automation. The analysis of the use of machine learning in risk management in this thesis has shown that the use of such models is connected to all these benefits. The resulting models of this study are shown to have strong prediction power over time, limit the need for manual labor as well as have strong AUC. However, as argued in earlier sections, the use of machine learning may limit transparency and result in models which outputs are not easily communicated to the customer when used in IRB modelling. The models presented above can, in ways described earlier in this thesis, provide different rankings of how important different features are for the total performance of a model. They do however not present any information about exactly how much each feature impacts the individual outcome, nor do they give the user any indication of why the model predicted a certain customer as default or no default. As Swedish regulations state that all decisions made by a bank must be transparent enough to be communicated to a customer it is valuable to discuss whether the increase in accuracy now and over time is worth the decrease in explanability.

As there are some different interpretations to how a decision must be communicated to a customer the use of machine learning models at Länsförsäkringar will be subject to the structures in place at the bank. The main argument regarding transparency Länsförsäkringar communicates is the need for insurance officers working at local offices across Sweden to understand the decision process in such a way that they can communicate it clearly to the customer. This puts pressure on the clarity required when communicating the decisions made by the models and requires both easily interpreted models as well as well formulated model documentation. This is however one of the downsides to using machine learning models since there is no clear motivation given by the model compared to when using a linear regression model in which one can use the variable coefficients to argue for feature importance. One can hence discuss how much there is to gain from limiting transparency to increase accuracy. Since the models are performing well over time and XGBoost in particular is shown to be a valid contender to current models in place, one need to take a step back and study the main idea behind the need for PD models.

When PD models are created the aim is to model and monitor how customers' behaviour with regards to repaying debt affects the credit risk taken by the bank in accordance with the capital requirement regulated by Swedish law. This means that banks both need models such as these to monitor current customers' behaviour to hedge for shortfalls in debt payments, as well as to guarantee that the bank fulfills the capital requirements stated by Swedish law. In the situation of fulfilling capital requirements to cover credit risk a tree based machine learning model might be the best way forward since our models show that these types of models tend to be accurate without loosing prediction power over time. This would limit the need for regularly updates to the model structure. The use of PD modelling for monitoring capital requirement might therefore be benefited by implementation of machine learning algorithms as this field of use does not require the transparency in decision making that customer interactions do. In accordance with this, the European Banking Authority and the EU's AI Act limit the use of machine learning when interacting with customers but have no clear regulation in place for managing the use of machine learning in internally used models [European Banking Authority, 2023]. Since regulations might change, the implementation of for example XGBoost at banks is connected to some uncertainty, but if the use is purely internal the implementation might be approved. However, if a bank uses a model such as ours to categorize their customers in risk grades which are then used in further analysis and mapping of customers this will not be allowed given the current regulations. The model might have unknown biases based on the training data which in addition to the overall issue with transparency does not create optimal conditions for getting the approval by Finansinspektionen.

One can hence argue that the use of machine learning in PD modelling is beneficial if the aim is to monitor the capital requirement but because of the lack of explanability it is not beneficial nor approved to be used when communication with customers. As the datasets used when building the models in this thesis consist of current customers who have already received loans from Länsförsäkringar, an argument is that this model can be used to monitor movements between risk grades and flag customers which behaviour changes drastically. If the model were to be used like this the benefits of machine learning implementation are great and as this requires no customer interaction it might be approved to be used.

As the results show that machine learning models are performing on a comparable level as regression based ones there seem to be benefits connected to using more advanced models. Since the models of this thesis were created using limited manual data management, human intervention and time these types of models might have more advantages than just the accuracy. The time and resources saved when limiting manual data management create opportunities for employees to use their time more efficiently, creating opportunities for further operational improvements. At the same time, the implementation of automated and standardized feature selection methods limit the risk of bias and give the bank a chance to identify variables previously not believed to be of importance. This can both result in more fair and data-driven decision making as well as in more substantiate decisions, which is beneficial when measuring credit risk and monitoring a bank's capital requirement.

8.3.2 Challenges linked to implementation

Connecting this deliberation to the initial research question, *Is there an added value in using more advanced methods applying machine learning when model performance is weighed against the possibility of explaining the model outcome to the client?*, it can be concluded that the use of machine learning models in risk management has some benefits with regards to accuracy over time. The switch to machine learning, and the use of tree-based algorithms in particular, is in fact connected to added value, however only when used in internal modelling. As concluded above, the limit in transparency and explanability that follows the use of ANNs and decision trees will require some updates to the current regulations and is more dependent on the statistical skills and knowledge possessed by the insurance officer than the use of regression models. As stated in a 2023 report by the European Banking Authority credit institutions which have implemented machine learning in their IRB models express the need for additional know-how skills in order to succeed with the implementation. These skills include theoretical knowledge of machine learning models, mathematical and statistical understanding of model structure and machine learning techniques used [European Banking Authority, 2023]. This would require organizations to invest in training of current employees, hiring of new competencies and expanding of knowledge sharing across organizational divisions. The success of implementing machine learning in IRB modelling is hence based on the organization's ability to assemble the skills necessary and share knowledge with divisions previously not required to possess deep technical skills. This includes training insurance officers across the country to interpret and communicate decisions taken by the machine learning model as well as ensuring testing and validation operations have the essential qualities. If the model was only to be used in internal IRB modelling the training of insurance officers would not be as vital, however the overall literacy in the organization would still need to be high.

The implementation of machine learning would not only be linked to the need for human skills in the bank but also within Finansinspektionen which is responsible for evaluating and approving IRB models in Sweden. As the current regulations require skills associated with regression analysis in order to assess model feasibility, Finansinspektionen would need to invest in training and knowledge gathering in order to fully evaluate the models proposed by banks if machine learning based ones were to be accepted. As Finansinspektionen is a public office this would require investments from the Swedish government, dependent on the government's attitude to the use of machine learning. The successful use of machine learning in credit risk management is hence deeply dependent on multiple stakeholders and their attitude towards the implementation.

9 Conclusion

The objective of this thesis was to analyze whether the use of machine learning in probability of default modelling is beneficial when explainability and transparency is weighted with model performance. The research has been focused on answering the stated research questions:

- Is there an added value in using more advanced machine learning methods when model performance is weighed against being able to explain the model outcome?
- Is it possible to use machine learning to develop new decision models that, when compared to the performance of regression models, make more accurate decisions and performs stable now and over time?

With regards to the first research question it was found that the use of machine learning models do in fact add value to the organisation and are transparent to the user, depending on the definition of transparency. All three machine learning algorithms performed well with regards to AUC score and limit the need for manual data management, creating an opportunity for organisations to restrain the time required for preprocessing of data. Regarding transparency it is possible to measure the importance of the variables in the models using for example PFI, however it is not as easily interpreted as analyzing the beta coefficients in linear regression. Secondly, the use of machine learning algorithms such as these proved to be easily implemented. The time required for implementing these models limited the need for manual data management making it easy to update models with time. With regards to this, machine learning models can be argued to be of benefit when building models to be used over time, as the use limits time consuming data management making it easier to update the model when accuracy declines.

When implementing and testing the three machine learning models Random Forest, XGBoost and ANNs, different hyperparameters, variables and imputation methods were used to find the best possible model with regards to the AUC score. The implementation of automated hyperparameter tuning and variable selection using RFE proved to efficiently find well performing models while limiting the need for manual labor. The two different imputation methods, zero and median imputation, proved to have similar performances for the tree based models while the ANNs were discovered to be more sensitive to the method used.

With regards to the AUC score and Brier score XGBoost was found to be the best performing model, no matter the imputation method, in accordance with findings of previous studies. The use of XGBoost outperformed the other models and received an AUC score of over 0.95 on the validation data from 2019, while having somewhat stable migration matrices. The use of variable importance measures showed that it is possible to get information regarding variable importance for an XGBoost model, even though the scores are not as tangible as the beta coefficients received using logistic regression. The use of XGBoost can hence be argued to be beneficial in IRB modelling, opening up the discussion regarding the need for transparency and easily interpreted model outputs.

The high AUC score received when using XGBoost indicates the benefits of implementing machine learning in risk management. The limiting of manual data processing and increase in standardization that follows the use of models such as XGBoost is argued to be of great use to the banking sector as it increases efficiency. Despite the benefits, the balance between transparency and model performance is harder to reach when using machine learning compared to logistic regression models and would require some changes to the legislation covering transparency in IRB modelling. The implementation of models such as XGBoost would in addition require changes to the competencies and knowledge sharing of both banks and Finansinspektionen as the use of machine learning requires deep understanding of the more advanced models' structures and theory. The use of machine learning in PD modelling would hence both require updates to the current regulations regarding IRB models and to the knowledge base of banks, legislators and customers. However, as stated in this study, the use of machine learning in risk management seem to have some advantages, why further studies on machine learning legislation is needed to truly tell the potential of the implementation of these algorithms. However, this study proves the usefulness and applicability of machine learning in PD modelling.

9.1 Future research

The use of machine learning in risk management is still relatively new and undocumented. In particular, the implementation of machine learning models on the Swedish credit risk market has not yet been widely studied with research covering risk management of private, non-corporate customers being even more rare. However, given the topicality of artificial intelligence and machine learning the field is only going to grow in popularity and application moving forward. It is therefore of interest to study the application of machine learning in further settings connected to risk management and to analyze the differences between various subclasses of machine learning algorithms. This study only looks at the results of implementing decision trees and artificial neural networks why it is of relevance to study the use of other algorithms, both generative and non-generative. Machine learning as a concept covers a broad spectrum of algorithms of different structures and levels of complexity meaning more research need to be done conducted to evaluate what algorithms are of best use in risk management.

This study covers private customers at Länsförsäkringar Bank AB in Sweden, meaning the study is limited both in geography and in type of customers. Further research is needed to evaluate the broader impact of using machine learning in risk management for different markets and groups of customers. The Swedish market is unique in the sense that Swedish households generally tend to have quite high mortgages with floating rates in comparison to borrowers in other parts of Europe [Helgesson, 2021]. In addition, the Swedish property rental market is smaller than in other countries meaning most Swedes own their homes resulting in big dependencies on the possibility to loan money from banks [Herold et al., 2021]. To evaluate how machine learning can be used by risk departments to monitor their capital requirement across the globe further analysis will therefore need to be conducted on additional markets outside of Sweden.

Another limitation to this study is that it only concerns the use of machine learning in PD modelling, while this only covers one part IRB modelling. Further studies will therefore benefit by implementation of machine learning in additional IRB models to create a broader picture of the general use of such algorithms in risk management.

As mentioned, the use of machine learning and artificial intelligence is yet a topic to be further expanded and as so new regulations and limitations of use are to be expected to be introduced over the coming years. Initiated by the EU's AI Act [European Union, 2024] further regulations for the use of machine learning and artificial intelligence are to be expected and as so the use of machine learning in risk management is subject to changes moving forward. This can result in both an expansion and limitations of the use of machine learning in risk management why future research will give a more clear picture of what the use of machine learning will look like moving forward.

References

P. M. Addo, D. Guegan, and B. Hassani. Credit risk analysis using machine and deep learning models. *RISKS*, 2018. URL https://www.mdpi.com/2227-9091/6/2/38.

M. Ala'raj and M. F. Abbod. Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, 104:89–105, 2016. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2016.04.013. URL https://www.sciencedirect.com/science/article/pii/S0950705116300569.

M. Azzone, E. Barucci, G. G. Moncayo, and D. Marazzina. A machine learning model for lapse prediction in life insurance contracts. *Expert Systems with Applications*, 2022. https://www.sciencedirect.com/science/article/pii/S0957417421015700?via%3Dihub.

Basel Committee. History of the basel committee. 2024. URL https://www.bis.org/bcbs/history.htm.

K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011. URL http://arxiv.org/abs/1106.1813.

A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(96)00142-2. URL https://www.sciencedirect.com/science/article/pii/S0031320396001422.

A. Cervenka. Bankerna borde stoppa miljardregnet. *Svenska Dagbladet*, 2016.

J. Y.-L. Chan, S. M. H. Leow, K. T. Bea, W. K. Cheng, S. W. Phoong, Z.-W. Hong, and Y.-L. Chen. Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics*, 10(8), 2022. ISSN 2227-7390. doi: 10.3390/math10081283. URL https://www.mdpi.com/2227-7390/10/8/1283.

M. Chen, Y. Dautais, L. Huang, and J. Ge. Data driven credit risk management process: A machine learning approach. In *ICSSP'17: PROCEEDINGS OF THE 2017 INTERNATIONAL CONFERENCE ON SOFTWARE AND SYSTEM PROCESS*, pages 109–113, 2017. URL https://dl-acm-org.focus.lib.kth.se/doi/abs/10.1145/3084100.3084113.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, Aug. 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

M. Chui, L. Yee, B. Hall, A. Singla, and A. Sukharevsky. The state of ai in 2023: Generative ai's breakout year. 2023. URL https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year#/.

P. Englund. The swedish 1990s banking crisis: A revisit in the light of recent experience. *Riksbanken*, 2015. URL https://archive.riksbank.se/Documents/Avdelningar/AFS/2015/Session%201%20-%20Englund.pdf.

European Banking Authority. Machine learning for irb models: Follow-up report from the consultation on the discussion paper on machine learning for irb models. 2023. URL https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf.

European Banking Authority. The basel framework: the global regulatory standards for banks. 2024. URL https://www.eba.europa.eu/activities/basel-framework-global-regulatory-standards-banks.

European Union. Ai act. 2024. URL https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Finansinspektionen. Stora förändringar av ramverket för bankernas kreditriskmodeller, 2018. URL https://www.fi.se/sv/publicerat/nyheter/2018/stora-forandringar-av-ramverket-for-bankernas-kreditriskmodeller/#dela.

L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on tabular data? *Cornell University*, 2022. URL https://arxiv.org/abs/2207.08815.

Y.-L. Grize, W. Fischer, and C. Lützelschwab. Machine learning applications in nonlife insurance. *Applied Stochastic Models in Business and Industry*, 2020. URL https://onlinelibrary-wiley-com.focus.lib.kth.se/doi/10.1002/asmb.2543.

G. Helgesson. Sweden: Wealthy but vulnerable households. *Nordea Economic Outlook*, 2021.

T. Herold, S. Westerberg, and A. Lodén. Den låsta dörren till hyresmarknaden. *Stockholms Handelskammare*, 2021.

IBM. What is a neural network? Technical report, 2024a.

IBM. What is machine learning?, 2024b. URL https://www.ibm.com/topics/machine-learning.

J. Jemai and A. Zarrad. Feature selection engineering for credit risk assessment in retail banking. *Information*, 14(3), 2023. ISSN 2078-2489. doi: 10.3390/info14030200. URL https://www.mdpi.com/2078-2489/14/3/200.

R. Kanapickienė, G. Keliuotytė-Staniulėnienė, D. Vasiliauskaitė, R. Špicas, A. Neifaltas, and M. Valukonis. Macroeconomic factors of consumer loan credit risk in central and eastern european countries. *Economies*, 2023. URL https://doi.org/10.3390/economies11040102.

H. Kaneko. Cross-validated permutation feature importance considering correlation between features. *Analytical Science Advances*, 3(9-10):278–287, 2022. doi: 10.1002/ansa.202200018.

J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler. Consumer credit risk: Individual probability estimates using machine learning. *EXPERT SYSTEMS WITH APPLICATIONS*, 2013. URL https://www-sciencedirect-com.focus.lib.kth.se/science/article/pii/S0957417413001693.

T. Li and L. Han. Dealing with explainability requirements for machine learning systems. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1203–1208, 2023. doi: 10.1109/COMPSAC57700.2023.00182.

Y. Li and W. Chen. A comparative performance assessment of ensemble learning for credit scoring. *Mathematics*, 8(10), 2020. ISSN 2227-7390. doi: 10.3390/math8101756. URL https://www.mdpi.com/2227-7390/8/10/1756.

M. Mahbobi, S. Kimiagari, and M. Vasudevan. Credit risk classification: an integrated predictive accuracy algorithm using artificial and deep neural networks. *ANNALS OF OPERATIONS RESEARCH*, 2023.

A. Mashrur, W. Luo, N. Zaidi, and A. Robles-Kelly. Machine learning for financial risk management: A survey. *Institute of Electrical and Electronics Engineers*, 2020. URL https://ieeexplore-ieee-org.focus.lib.kth.se/document/9249416/references#references.

D. Montgomery, E. Peck, and G. Vining. *Introduction to Linear Regression Analysis.* John Wiley & Sons, Inc., 5th edition edition, 2012.

Riksbanken. Internal risk classification systems and risk-sensitive capital requirements. *Riksbank Financial Stability Report*, pages 71–84, 2004. URL https://archive.riksbank.se/Upload/Dokument_riksbank/Kat_publicerat/Artiklar_FS/finstab04_1_artikel2.pdf.

Riksbanken. The financial crisis 2007-2010. 2023. URL https://www.riksbank.se/en-gb/markets/measures-in-response-to-financial-turmoil/the-financial-crisis-2007-2010/.

M. Rizinski, H. Peshov, K. Mishev, L. T. Chitkushev, I. Vodenska, and D. Trajanov. Ethically responsible machine learning in fintech. *IEEE ACCESS*, 2022. URL https://ieeexplore.ieee.org/document/9869843.

A. Roberts. Binary cross entropy: Where to use log loss in model monitoring, 2023. URL https://arize.com/blog-course/binary-cross-entropy-log-loss/.

G. Rosquist. Länsförsäkringar, 2024.

M. Schonlau and R. Y. Zou. The random forest algorithm for statistical learning. *Stata Journal*, 20, 2020. doi: 10.1177/1536867X20909688.

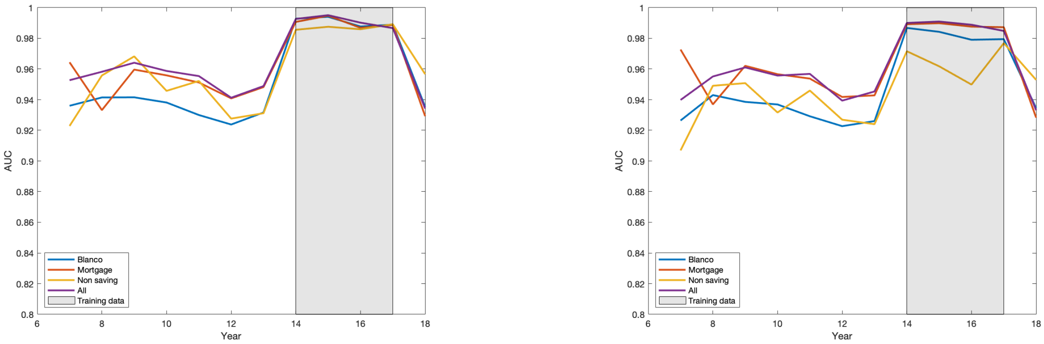
S. Shi, R. Tse, W. Luo, S. D'Addona, and G. Pau. Machine learning-driven credit risk: a systemic review. *NEURAL COMPUTING & APPLICATIONS*, 2022. URL https://link-springer-com.focus.lib.kth.se/article/10.1007/s00521-022-07472-2.

Swedish Bankers' Association. Det här är basel 4, 2020. URL https://www.swedishbankers.se/repository/bankfokus/bankfokus-nr-1-2020/det-haer-aer-basel-4/.

W. Uriawan, O. Hasan, Y. Badr, and L. Brunie. Laps: Computing loan default risk from user activity, profile, and recommendations. In *2022 FOURTH INTERNATIONAL CONFERENCE ON BLOCKCHAIN COMPUTING AND APPLICATIONS (BCCA)*, 2022. URL https://ieeexplore-ieee-org.focus.lib.kth.se/document/9922034.

A. Zainab, A. Ghrayeb, M. Houchati, S. S. Refaat, and H. Abu-Rub. Performance evaluation of tree-based models for big data load forecasting using randomized hyperparameter tuning. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 5332–5339, 2020. doi: 10.1109/BigData50022.2020.9378423.

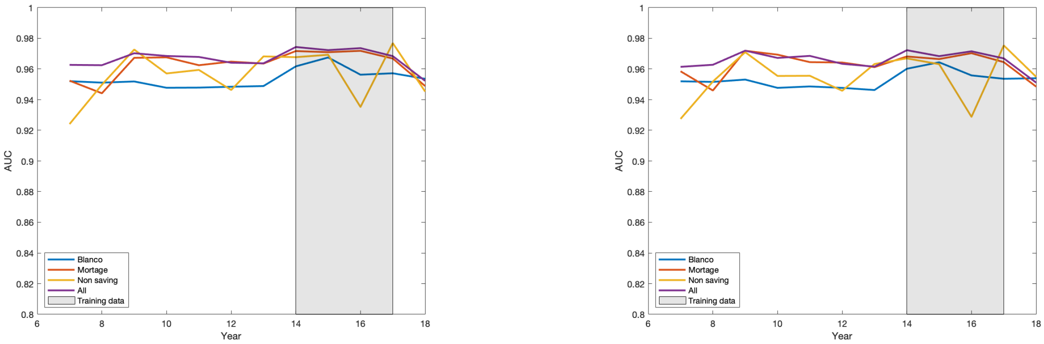
10 Appendix



(a) With *zero imputation*

(b) With *median imputation*

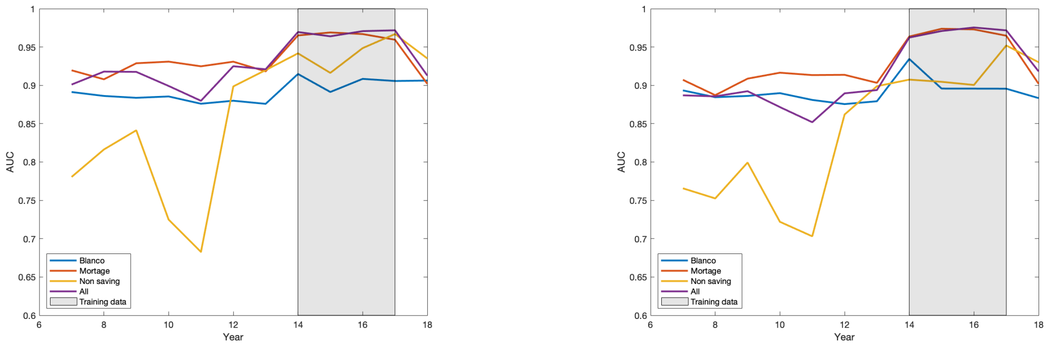
Figure 14: AUC of subgroups by Random Forest over the years



(a) With *zero imputation*

(b) With *median imputation*

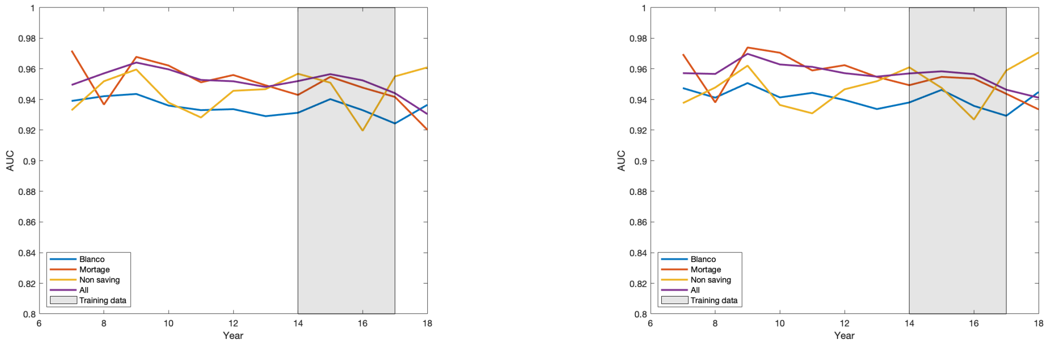
Figure 15: AUC of subgroups by XGBoost over the years



(a) With *zero imputation*

(b) With *median imputation*

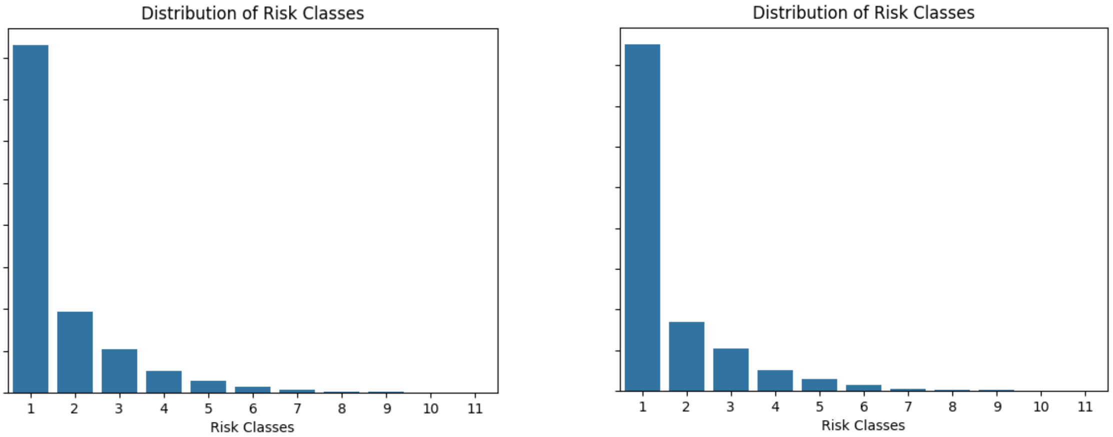
Figure 16: AUC of subgroups by ANN with SMOTE over the years



(a) With *zero imputation*

(b) With *median imputation*

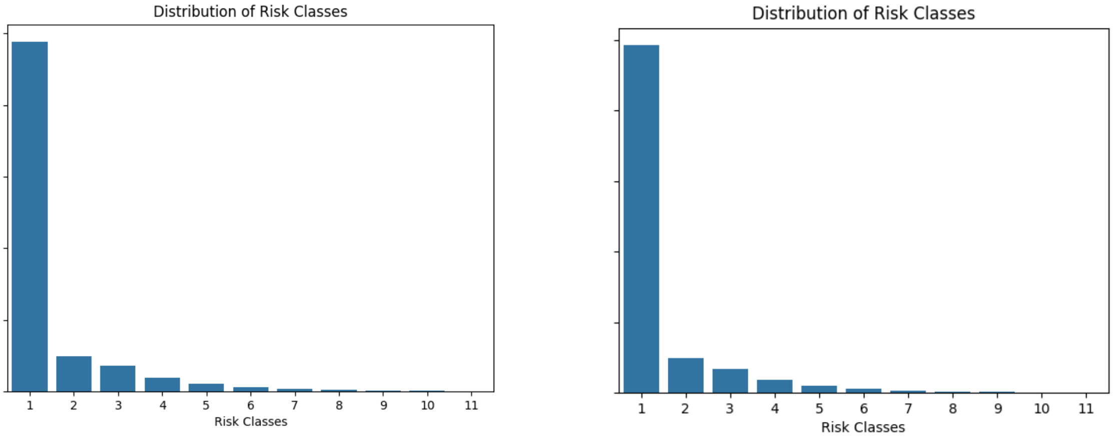
Figure 17: AUC of subgroups by ANN without SMOTE over the years



(a) With *zero imputation*

(b) With *median imputation*

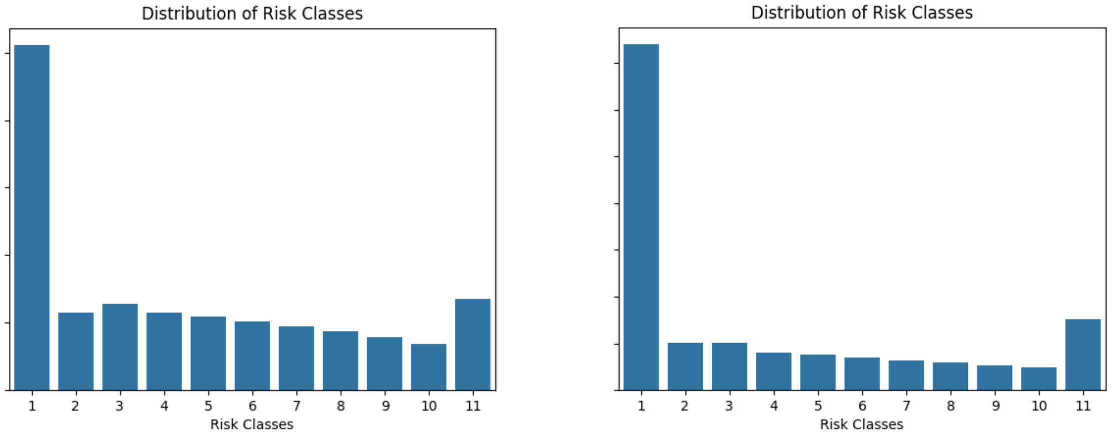
Figure 18: Risk classes with Random Forest for 2018



(a) With *zero imputation*

(b) With *median imputation*

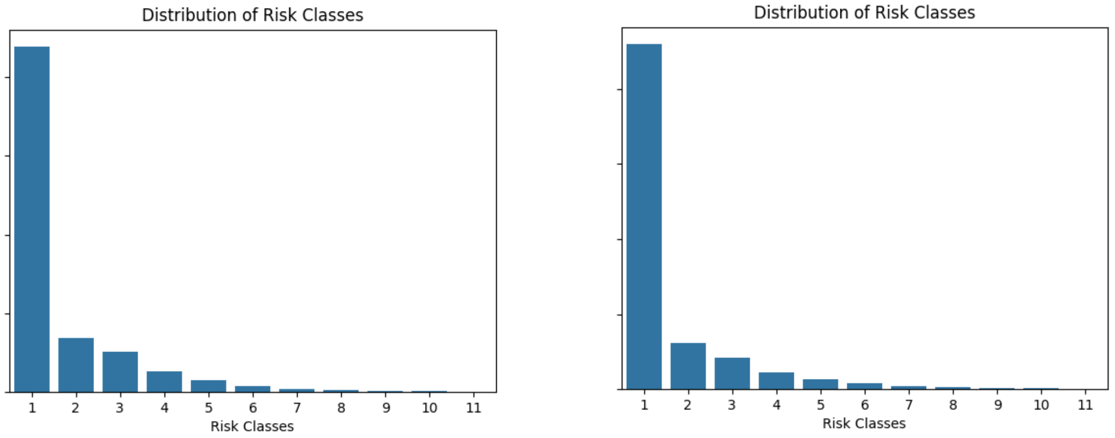
Figure 19: Risk classes with XGBoost for 2018



(a) With *zero imputation*

(b) With *median imputation*

Figure 20: Risk classes with ANN with SMOTE for 2018



(a) With *zero imputation*

(b) With *median imputation*

Figure 21: Risk classes with ANN without SMOTE for 2018

TRITA – ITM-EX 2024:285
Stockholm, Sweden 2024

www.kth.se