HELSINGIN YLIOPISTO

Interpretable Machine Learning for Credit Risk Predictions

Master’s Programme in Economics
General track
Master’s Thesis

Rami Erkkilä
21.02.2025
Helsinki

Faculty: Faculty of Social Sciences

Degree programme: Master’s Programme in Economics

Study track: General track

Author: Rami Erkkilä

Title: Interpretable Machine Learning for Credit Risk Predictions

Level: Master’s Thesis

Month and Year: February 2025

Number of pages: 50

Keywords: Credit risk, Probability of default, Machine learning, Random Forest, Logistic regression, SHAP

Supervisors: Jani Luoto, Mika Meitz

Abstract:

Credit risk models are a vital part of risk management in banks and other credit institutions. Appropriate risk models help mitigate issues caused by uncontrolled rise in default rates. Therefore, credit institutions are required to estimate the probability of default (PD) for each customer to decide whether a loan is granted and to calculate their capital requirements. This estimation is traditionally conducted using statistical methods such as logistic regression, but the recent rise of artificial intelligence (AI) and machine learning (ML) has introduced other alternative models.

Research has shown that the machine learning models generally outperform regression models, but the lack of interpretability limits their adoption in the field of credit risk. In this thesis, a comparison between a random forest model and logistic regression is performed to verify whether the random forest model would yield performance improvements in estimating the probability of default. The main research goal is then to evaluate whether the interpretability of the random forest model could be enhanced to meet regulatory requirements such as General Data Protection Regulation (GDPR) and guidelines set by the European Banking Authority.

In this thesis the random forest model outperforms the traditional logistic regression model, achieving a higher AUC score. Furthermore, with the help of the SHAP framework this thesis shows that the model estimates can be made interpretable both on global and local levels. Institutions adopting these interpretable models could therefore provide transparent explanations behind their decisions.

While the findings in this thesis highlight the potential of interpretable ML models in credit risk management, constantly changing regulatory environment remains a challenge. Therefore, further research is needed to assess the viability of ML models in credit risk management.

Tiedekunta: Valtiotieteellinen tiedekunta

Koulutusohjelma: Taloustieteen maisteriohjelma

Opintosuunta: Yleinen opintosuunta

Tekijä: Rami Erkkilä

Työn nimi: Tulkittavat koneoppimismallit luottoriskien ennustamisessa

Työn laji: Maisterintutkielma

Kuukausi ja vuosi: Helmikuu 2025

Sivumäärä: 50

Avainsanat: Luottoriski, Maksukyvyttömyysriski, Koneoppiminen, Satunnaismetsä, Logistinen regressio, SHAP

Ohjaajat: Jani Luoto, Mika Meitz

Tiivistelmä:

Luottoriskimallit ovat tärkeä osa pankkien ja muiden rahoituslaitosten riskienhallintaa. Tehokkaat luottoriskimallit ehkäisevät nousevien maksukyvyttömyysosuuksien aiheuttamia ongelmia. Sääntelyn mukaan rahoituslaitosten täytyy mallintaa maksulaiminlyöntien todennäköisyys asiakkailleen luotonmyöntöpäätösten tueksi sekä riittävän oman pääoman vaatimusten määrittämiseksi. Mallinnus on yleensä suoritettu käyttämällä perinteisiä tilastollisia menetelmiä kuten logistista regressiomallia, mutta tekoäly- ja koneoppimismallien kehittyminen on mahdollistanut vaihtoehtoisten mallien hyödyntämisen.

Aiemmat tutkimukset ovat osoittaneet, että koneoppimismallit yleisesti suoriutuvat perinteistä logistista regressiomallia paremmin, mutta heikko tulkittavuus rajoittaa niiden käyttöä luottoriskimallinnuksessa. Tässä maisteritutkielmassa verrataan satunnaismetsäalgoritmin sekä logistisen regressiomallin suorituskykyä maksulaiminlyöntien todennäköisyyden ennustamisessa. Tutkielman päätavoitteena on arvioida, voiko satunnaismetsän tulkittavuutta parantaa, jotta sääntelyn kuten yleisen tietosuoja-asetuksen (GDPR) sekä Euroopan pankkiviranomaisen (EPV) asettamien ohjeiden vaatimukset täyttyisivät.

Tämän tutkielman tulokset osoittavat satunnaismetsäalgoritmin suoriutuvan perinteistä logistista regressiomallia paremmin tuottaen korkeamman käyränalaisen pinta-alan (AUC). Lisäksi SHAP-kehikko mahdollistaa koneoppimismallien, mukaan lukien satunnaismetsän, tulkittavuuden sekä globaalilla että lokaalilla tasolla. Näitä tulkittavia malleja hyödyntävät rahoituslaitokset pystyisivät näin ollen antamaan läpinäkyviä perusteluja mallien antamille päätöksille.

Vaikka tämän tutkielman tulokset osoittavat tulkittavien koneoppimismallien mahdollisuudet luottorikien hallinnassa, jatkuvasti muuttuva sääntely-ympäristö luo haasteita niiden laajamittaiselle käyttöönotolle. Tämän takia tarvitaan lisätutkimusta, joka pyrkii arvioimaan koneoppimismallien toimivuutta luottoriskien hallinnassa pidemmällä aikavälillä tulevaisuudessa.

Contents

1 Introduction

Credit risk prediction is a crucial part of the lending process in the financial sector. Banks generally use traditional statistical models like linear or logistic regression for credit scoring. Usually, credit scores are expected to be interpretable and should have statistical properties. General data protection regulation by the EU (2016/679) requires that a data subject must be provided with the logic of automated decision-making. Additionally, the European Banking Authority (2017) mandates that banks ensure transparency in credit risk models. Therefore, due to regulatory restrictions, statistical methods are widely used. There could, however, be advantages to using more sophisticated techniques like machine learning.

Machine learning methods are generally known to have higher predictive powers compared to traditional statistical models. Lessmann et al. (2015) compared over 40 ML methods and concluded that tree-based models like Random Forest outperform standard logistic regression in credit scoring. These models are often able to utilize non-linear relationships between variables to enhance predictions. More accurate credit scoring would benefit commercial banks in multiple ways. More precise credit scoring would allow banks to make better credit decisions leading to fewer defaults. This would also likely increase the profitability of commercial banks since banks could decrease their capital requirements. However, machine learning models have characteristics that are referred to as "black boxes” since their predictions are usually hard to explain, as models contain complex mathematical functions (Loyola-Gonzalez, 2019).

The main goal of this study is to examine if there are benefits of using interpretable machine learning algorithms instead of traditional statistical models when predicting credit risk. Additionally, compliance with banking regulations is investigated to find out if these models would be applicable in practice. My thesis contributes to the existing literature by analysing the performance of an interpretable random forest model on a publicly available credit risk dataset. While previous research (e.g., De Lange et al., 2022) has studied interpretable ML in credit risk, there is little research on whether SHAP framework or other interpretability methods could be used to meet regulatory standards. This thesis also helps to answer whether these models meet the regulatory restrictions that guide the lending process of commercial banks.

2 Literature Review

Due to the significance of the banking sector in the economy, it is easy to argue that it plays a crucial role in keeping the economy functioning. Banks grant credit and allow institutions to save and invest, all of which are crucial for economic growth. Without banks, the economy would not be able to function effectively. (Naili and Lahrichi, 2022) Economic growth depends on the reallocation of resources. This reallocation requires some risk-taking. A functioning competitive economy allows projects or individual firms to fail. (OECD, 20211) Banks are traditionally viewed to issue short-term deposits to finance their longer-term loans. Banks are responsible for analysing the riskiness of these long-term loans before and after lending. Since banks hold these loans to maturity, they are vulnerable to credit risk since borrower's repayment capacity may deteriorate. (Saunders and Allen, 2010). Given the banking sector's substantial effects on the economy, it is important to mitigate risks originating from this sector. For this reason, effective risk assessment plays a critical role in the financial sector.

Interest in risk assessment is often justified by financial crises in the 1980's, 1990's, and more recently 2007-2008 financial crisis. Additionally, there are multiple smaller cases where more accurate risk modelling could have prevented the crisis and enabled more efficient resource allocation (Galindo and Tamayo, 2000). The rapid expansion of the credit base contributed to these crises creating a regulatory incentive to avoid new crises through stricter risk assessments. Credit expansion continues and the impact of a new financial crisis could be more severe than those in the past. In the US, the total amount of outstanding consumer credit was over 13.63 trillion dollars in 2008 (Khandani et al., 2010). In the second quarter of 2024 this amount was already 17.80 trillion dollars (Federal Reserve Bank of New York, 2024). Household credit growth is an important predictor for banking crises. While enterprise credit expansions are also a factor behind these crises, but their contributions are smaller. (Büyükkarabacak and Valev, 2010) Therefore, strict risk assessment when expanding credit bases is desirable.

There have been multiple regulatory responses that have been aimed at mitigating issues related to insufficient risk management. EU banking rules require banks to maintain a set capital level to increase their resilience to unexpected credit losses (ECB Banking Supervision, 2021). Basel III is an international regulatory framework for banks. It raised the minimum capital requirement for common equity capital from 2 % to 4.5 % of risk-weighted assets (RWA) (Slovik and Cournède 2011). Increased capital buffers increase bank's resilience to unexpected credit losses. However, there is a trade-off between reduced risks and the bank's profitability. Reducing a bank's risk level increases its capital costs (Baker and Wurgler, 2015). However, better risk management inside banks may enhance profitability since banks have more tools to analyze the risk profiles of customers. Banks have two different options for risk assessment: standardized approach (SA) and internal ratings-based approach (IRB). Under the SA approach banks use standardized weights for their risk-weighted assets. Internal models allow banks to utilize their own internal data for risk evaluation. IRB approach results in more accurate risk profiling and therefore usually leads to decreased capital requirement. (ECB Banking Supervision, 2021) However, it is not clear which method maximizes profitability, as the IRB approach often leads to increased staff costs as models require development and maintenance. In this thesis, analysis is concentrated on IRB modelling since the interest is on credit score models, e.g. probability of default (PD) estimation.

2.1 Regulation

Credit risk modeling conducted in EU banks follows the regulatory framework of the EU and EBA. One of the main legislative acts is Regulation (EU) 2016/679 (GDPR) which is an important part of the privacy laws and human rights laws in the EU. It grants individuals the right to access and limit the use of their personal data. There are several articles that are important for PD estimation.

Article 5 of the Regulation (EU) 2016/679 limits the usage of personal data. Interpretation is that personal data must be limited to what is required for PD estimation. Therefore, the data must not be used for other unrelated purposes. Additionally, article 12 provides the right for transparency and interpretability. This in addition to Article 13 requires institutions to provide information about the logic behind automated decision-making. This interpretability concern is a crucial part of

PD estimation since the usage of black-box ML models is prohibited in automated decision making. This is one of the main drivers for interpretable machine learning algorithms in credit risk management as techniques like SHAP could be used to address this issue.

The guideline from the European Banking Authority (2017) states that an institution should carefully analyze and choose meaningful risk drivers for PD modeling. This in addition to the requirements of GDPR limits the number of variables that can be used as explanatory variables in the ML model. However, ML models are good at finding complex interactions between many variables (Mullainathan and Spiess, 2017). Therefore, the analysis of risk drivers is not so simple as there can be a lot more relevant risk drivers through these complex interactions. It can be argued that the usage of ML models is in line with regulations for this part since it allows the usage of all relevant risk drivers. In traditional methods like logistic regression, the number of variables is usually limited and therefore ML could provide improvements. It is however important to note that the institutions are required to consult the relevant business experts when selecting risk drivers (EBA, 2017). In the case of ML algorithms, this may be difficult if the number of explanatory variables is high. Therefore, institutions would likely be limited to a lower number of candidate risk drivers which could result in decreased performance of machine learning models.

The guideline from the European Banking Authority (2017) also states that institutions are required to verify that the underlying theoretical assumptions of a model used for PD estimations are met in practice. It should be noted that it may be difficult to verify the assumptions behind the standard models like logistic regression. For example, there are usually more complex relationships between the variables, and the linearity assumption may not be met. ML algorithms usually are less restrictive and are not subjected to different assumptions (Barboza et al., 2017). However, this comes with a significant cost: fewer assumptions mean that models are not as interpretable as traditional statistical models. This motivates the interest in interpretable ML algorithms.

Another important regulatory issue is statistical uncertainty. Institutions are required to use a margin of conservatism (MoC) that covers possible and likely estimation errors. MoC consists of three different parts:

1. Category A: MoC related to data and methodological deficiencies.
2. Category B: MoC related to relevant changes to underwriting standards, risk appetite, collection and recovery policies, and any other source of additional uncertainty.
3. Category C: MoC related to the general estimation error.

Since the interest in this thesis is on PD estimation relevant MoC is the category C MoC. For statistical methods like logistic regression these MoC:s can be estimated using properties of the model. Since the confidence intervals for each risk factor are specified, these intervals can be used to find appropriate levels of MoC. For the ML algorithm, the process is not that simple. Alonso Robisco and Carbó Martínez (2022) propose a framework to partially address this issue. They rank different models based on their model riskiness. However, it is unclear whether this approach would fully meet regulatory requirements since they do not specify MoC levels directly on the estimation results. The margin of conservatism (MoC) is often overlooked in studies that evaluate the performance of machine learning models in PD estimation. Measurement of MoC for ML algorithms is beyond the scope of this thesis but is something that would be an interesting topic for future research.

2.2 Credit Risk Prediction and Scoring Models

Banks face a difficult decision when determining whether to lend or not. They must evaluate each customer’s repayment ability with the information they possess. Historically profitable customers can still default in the future. Therefore, banks incorporate risk premiums for every loan. These risk premiums are stored in an internal account called the expected loss reserve. These reserves can be used to cover losses from defaulted loans. (Bluhm, Overbeck and Wagner, 2010)

A bank can estimate the probability of default (PD) for each customer. They can also estimate the loss-given default (LGD), which is the fraction of exposure expected to be lost when a customer defaults. Additionally, a bank can calculate exposure at default (EAD), which is the total value to which a bank is exposed at the time of default. Given these estimates, a bank can then estimate its expected losses using the formula

$$EL = EAD \times LGD \times PD.$$

In this thesis, the main interest is in the probability of default (PD) estimation. Often the PD estimation process is straightforward but there are situations where it can be challenging (Bluhm, Overbeck and Wagner, 2010). PD is traditionally estimated using a statistical model. Linear or logistic regression are popular choices, though multiple other options exist. Modern statistics including algorithms such as k-Nearest-Neighbours or tree-based systems can also be applied. (Bussmann et al., 2021)

Credit scoring can be defined as the assessment of the risk that is associated with a bank's lending process. It is used for a variety of different loan types including mortgages and credit cards. (Costa e Silva et al., 2020) It is a well-known problem in economics since it was one of the first fields where machine learning methods were applied (Dumitrescu et al., 2022). Dumitrescu et al. argue that the performance of machine learning (ML) methods has been improved since the early 2000's and therefore they are increasingly used by banks for credit scoring. However, due to regulatory concerns regarding the usage of ML in credit risk scoring, traditional statistical methods are still widely used. ML models generally lack explainability and interpretability and are often considered black-boxes (Dumitrescu et al., 2022). Black-box models are models that contain complex mathematical functions and are therefore hard to explain and understand by experts in applications (Loyola-Gonzalez, 2019). It is argued by Bussmann et al. (2021) that black-box ML is not suitable for regulated financial services. Due to these regulatory reasons PD estimation is still usually done using a logistic regression model that is interpretable and explainable.

Credit scoring models are based on historical data on the bank's existing clients. The idea is to assess the customer’s likelihood of being a good or bad payer. This assessment is continuously conducted during the loan period. Initially, a screening is conducted before a loan transaction to identify whether the customer is likely to meet repayment obligations. A customer is then monitored throughout the lifetime of the loan until it is written off. (Costa e Silva et al., 2020) It is important to note that there are several social issues related to credit scoring. Since the performance of the scoring model is highly dependent on the data available, there is an incentive to use all relevant socio-economic variables in these models. However, it is illegal to use some characteristics including race and sex (Costa e Silva et al., 2020; Regulation (EU) 2016/679, 2016). Although, it is still debatable whether institutions should be allowed to use these characteristics for modelling since there is evidence that these variables influence customer's likelihood of default. For example, Lin et al. (2017) argue that gender has a statistically significant effect on default risk. The dataset used in thesis does not include variables that would be ethically questionable so these considerations are not necessary in this case but are still something that should be kept in mind when building models.

2.3 Machine Learning Models in Credit Risk

Machine learning (ML) is traditionally defined as a field of study that allows computers to learn from the given data without being explicitly programmed. The idea is to apply algorithms to make machines learn from given data. There is a large variety of different ML algorithms available and usually, there is no single algorithm that is the best to solve a given problem. The preferred algorithm depends on the problem one is trying to solve. Machine learning can be divided into multiple subcategories. The most applied ones are supervised learning, unsupervised learning, semi-supervised learning, and neural networks. Supervised learning is a subfield of ML where the task is to find a function that maps given input-output pairs. These methods require a training dataset that has an output variable that is predicted or classified by the algorithm. All different supervised learning algorithms learn patterns from the training dataset and then apply those patterns to previously unseen test dataset. (Mahesh, 2020) In this thesis, the applied ML algorithm is the random forest algorithm which belongs to the supervised learning subcategory of ML. The algorithm is described in greater detail in its dedicated section.

Why are machine learning algorithms used in credit risk? Realistic and accurate risk models require accurate predictors of individual risk (Tamayo and Galindo, 2000). The accuracy of machine learning algorithms has improved since the early 2000s, largely due to the development of ensemble methods such as bagging and boosting, which combine multiple models to enhance predictive performance (Dumitrescu et al., 2022). The random forest is an example of a bagging model where multiple decision trees are combined to form a single model. Lessmann et al. (2015) compared over 40 different ML methods and confirmed that the random forest outperforms the traditional logistic regression model in credit scoring. It is however important to note that it is unlikely that a single model outperforms others in every problem given. Therefore usually, multiple different models should be considered as a candidate model. Additionally, laboratory experiments often can overestimate the model's prediction accuracies. External validity should be tested before models are applied to real-world problems. Dumitrescu et al. (2022). argue that over the last decade, banks have been widely adopting ML models as challenger models. They also mention that wider adoption has been restricted by regulatory issues since these models are not generally interpretable.

The performance of machine learning methods can be measured using multiple different methods. One of the most used methods is called the area under the ROC curve (AUC). ROC curve allows a visual representation of model performance. It is drawn by first calculating the true positive rate (TPR) and false positive rate (FPR), and then graphing TPR over FPR. In a perfect model, ROC curve is a horizontal line and TPR is 1.0 and FPR is 0.0. The AUC measure can then be calculated based on this ROC curve. The area under ROC curve indicates how well a model can differentiate between two randomly given points, one negative and one positive. AUC measures the probability that a positive point is ranked higher than a negative one. In a perfect model, AUC is therefore 1. (Google developer, 2024)

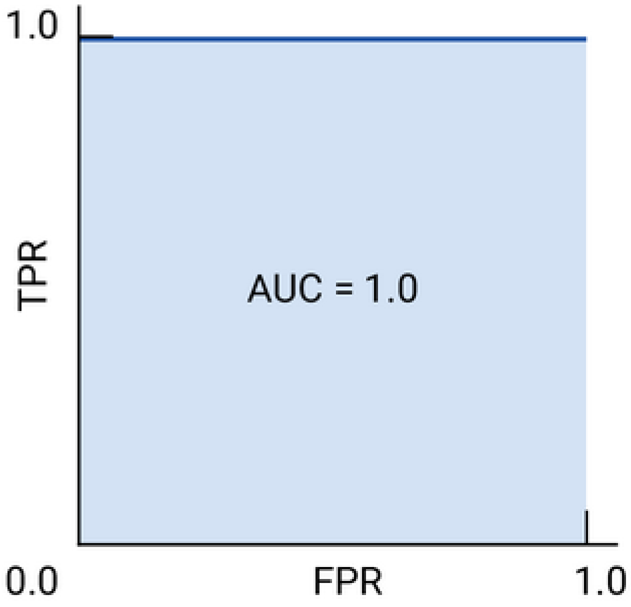


Figure 1. Example of ROC and AUC in a perfect model (Google developer, 2024)

AUC has been widely used as a performance metric for machine learning methods. Bradley (1997) discussed the properties of AUC and its favourable characteristics. It is argued that AUC has multiple desirable properties that make it a more accurate performance metric than a single accuracy measure. AUC for example has an increased sensitivity in the analysis of variance compared to overall accuracy. AUC also gives a low score for random or one-class-only classifiers. (Bradley, 1997) Especially the latter one is important for a good classifier in machine learning. The one-class classifier is a classifier that achieves high accuracy by predicting only a single class. This can happen when the target variable has a large class imbalance, meaning that the other class is far more common compared to the other one. When predicting the probability of default these imbalances are often a major factor since the number of non-performing loans (defaults) is often small compared to performing loans.

It is important to note that there are also drawbacks of using AUC score to compare different models. AUC ignores the predicted probability values and the goodness-of-fit of the model. For example, with class imbalances the mean probabilities are biased towards the more common event. (Lobo et al., 2008) Since this thesis is dealing with highly imbalanced dataset this could raise concerns about the validity of AUC as the selected performance measure. However, since the AUC has been widely used and proven to be an effective performance measure it will be used as a performance metric in this thesis also, even though there are some possible issues. Previous studies that have used AUC to compare different classification algorithms include Lessmann et al., 2015; Fitzpatrick and Mues, 2016; Moscato et al., 2021; Dumitrescu et al., 2022; Bücker et al., 2022. To mitigate possible issues other performance metrics introduced later in this thesis are used as well.

Other commonly used metrics to measure the performance of classifiers are precision and recall. Since the problem in hand is a binary classification task and a classifier labels observation as negative (non-delinquent) or positive (delinquent) a confusion matrix can be used to represent classifications. The confusion matrix is constructed from four different measures. There are true positives (TP) which are observations that are correctly labelled as positive and true negatives (TN) which are observations that are correctly labelled as negative. Additionally, false positives (FP) are observations that are incorrectly labelled as positive even though the correct class would be negative. Finally, false negatives (FN) are observation that are incorrectly labelled as negative. One example of a confusion matrix is shown in figure 8. When the confusion matrix is built one can then then calculate precision and recall scores. Recall is defined as the number of TP values divided by the sum of TP and FN:

$$Recall = \frac{TP}{TP + FN}$$

Similarly, the precision score can be defined as the number of TP values divided by the sum of TP and FP:

$$Precision = \frac{TP}{TP + FP}$$

Intuitively recall score describes the proportion of all actual positives that were classified correctly as positives and precision is the proportion of all the model's positive classifications that are positive. (Davis and Goadrich, 2006) In this thesis precision and recall will be used as an additional performance metrics to validate model performance. However, to ensure comparability with other studies the AUC score will be the primary performance metric.

2.4 Interpretability and Explainability in Machine Learning

Concepts of significance, relevance, and explainability help to understand the problem with machine learning algorithms. Statistical and societal sides are two different perspectives that can be studied to understand this problem. Statistical significance and relevance should be familiar concepts for all economists. They describe whether a coefficient is meaningful for the outcome of the dependent variable. Economic significance and relevance on the other hand describe how meaningful a one-unit change in the independent variable is in terms of its impact on the dependent variable. (Hoepner et al., 2021) The important topic for this thesis is explainability. Explainability ensures that from the statistical perspective, a model is transparent and can be replicated. It also ensures from the economic perspective that decision-makers can identify the steps in the model to rationalize their decisions. (Hoepner et al., 2021)

Interpretability has been widely researched. Loyola-Gonzalez (2019) discusses interpretability and explainability in machine learning and the differences between black-box and white-box models. They argue that white-box models such as decision trees offer higher interpretability since they produce results that are easy to understand for humans. These models provide representations that are easy for even non-technical users to follow. They also highlight newer approaches that aim to combine the benefits of black-box and white-box models. These so-called hybrid models retain the interpretability of white-box models and higher accuracies of black-box models. Often used techniques include rule extractions and visual explanations.

Bussmann et al. (2021) propose an explainable model that is argued to be usable in credit risk management. Their approach includes Shapley values (SHAP) where the predictions are grouped using the similarity in the explanations. This approach has its background in game theory, and Shapley values were originally introduced by Shapley (1953). Bussmann et al. (2021) argue that the advantage of using SHAP methods is that it allows the presentation of variable contributions to the prediction in a machine learning model. This method is also applicable to any machine learning model. They utilize the method proposed by Lundberg and Lee (2017) where predictions of a model are expressed as linear combinations of binary variables that describe if a variable is included in the model. Their main result is that the XGBoost algorithm outperforms the logistic regression model in credit risk prediction. Additionally, they show that XGBoost model can be explained using Shapley values.

Gramegna and Giudici (2021) apply the same SHAP framework proposed by Lundberg and Lee (2017) to real Small and Medium-sized enterprises (SME) data. The data used is SME data from the Italian Chamber of Commerce. They used a similar XGBoost algorithm as Bussmann et al. (2021) to predict the probability of default of Italian SMEs. Their main result is that the SHAP approach outperforms other interpretable models called Locally Interpretable Model Agnostic Explanations (LIME). Naturally, it is hard to extend to these results and conclude that the SHAP approach would always be preferred but these results indicate that the SHAP approach should be a well-suited approach for model interpretability.

Bücker et al. (2022) compare the performance of several different machine learning models to a traditional scorecard-based logistic regression. They observe that the performance of their logistic regression model is surprisingly good and more complex machine learning methods do not necessarily improve the accuracy of a credit risk model. This result is somewhat surprising since almost all other studies have shown that more complex ML algorithms outperform traditional logistic regression models. These results are likely achieved since they apply excessive data preparation for a logistic regression but use no further data preprocessing for ML models. It is also possible that for this particular dataset, a logistic regression is the better-performing model. These results could be achieved if the underlying problem is highly linear. Bücker et al. (2022) also apply the SHAP method to make ML models interpretable and argue that model exploration processes including the SHAP method would become inevitable in the future since the models are becoming more complex over time. They also argue that to meet regulatory requirements these interpretability methods should be applied if ML methods are used.

De Lange et al. (2022) also use a similar approach for interpretable ML in credit risk modeling. They apply the LightGBM model to a Norwegian credit risk dataset provided by a medium-tier bank in Norway. The TreeSHAP method proposed by Lundberg et al. (2019) is applied to a LightGBM model that is a tree-based gradient boosting framework. TreeSHAP is a method that allows for faster Shapley value calculation when using tree-based ML algorithms. They compare the performance of the LightGBM model to a classical LR model that is currently applied in the bank. It is determined that the LightGBM model outperforms the LR model by 17 % when measured by the change in the AUC score. However, some of the increased performance could be explained by the inclusion of additional explainable variables in the LightGBM model. This performance increase was reduced when the analysis was performed using the same variables as in the LR model, but the increase was still significant (9 % in AUC score). They also conclude that the SHAP method could be used to enhance the interpretability of a ML model. There is however one minor problem with this study. Like other studies mentioned in this section, they do not give a clear justification for why the interpretable model meets the regulatory requirements. However, De Lange et al. (2022) argue that the local explanations of SHAP solve the problem of explainability for each individual prediction.

Additional issues concerning interpretability arise from class imbalance. Since credit risk data is usually highly imbalanced as the number of defaults is low compared to performing loans, there is a concern that this imbalance could cause issues with interpretability. Chen, Calabrese, and Martin-Barragan (2024) showed that the class imbalance influences the interpretive performance of SHAP. They showed that SHAP is more stable when the ML model is trained using a balanced dataset. Large class imbalance also causes greater variance in Shapley values. Both are unwanted qualities since banks are required to give an explanation of a decision to a customer so they can improve their features to obtain a loan. If customers are given misleading or wrong explanations, they are not able to improve their features and are therefore more likely to be declined a loan also in the future. Chen, Calabrese, and Martin-Barragan (2024) argue that this kind of misinformation could lead to financial losses for both the institution and the customer. Additionally, the declined confidence on financial institutions could lead to a confidence crisis. It is important to note that almost always a resampling is used when dealing with imbalanced datasets. Resampling is a method where oversampling (increasing the number of minority class instances) or under sampling (reducing the majority class instances) are applied. Other more advanced method like Synthetic Minority Over-sampling Technique (SMOTE) can also be applied to improve model balance (Thomas et al., 2017). Cost-sensitive or algorithm-based methods should therefore be used to achieve unbiased interpretation results (Chen, Calabrese, and Martin-Barragan, 2024).

Even though the SHAP method is chosen for this thesis, there are other candidate methods that try to achieve similar results. Popular alternatives are partial dependency plots (Friedman, 2001; Zhao and Hastie, 2021) and already mentioned Locally Interpretable Model Agnostic Explanation (LIME). Other widely used alternatives include global surrogate modeling where a simpler model is trained to mimic the behavior of a complex black-box model (Lualdi et al. 2022). The SHAP method has achieved better classification results in credit risk than LIME (Gramegna and Giudici, 2021). Additionally, the SHAP method is a relatively simple method to implement and would therefore be suitable for real word tasks like PD estimation. For these reasons, the SHAP is the selected method for this thesis. More details about and theoretical background of SHAP follow in the section 3.3.

3 Methods

In this section we build the theoretical background for all models used in this thesis. At first, we look at the logistic regression model that will be used as a benchmark model in this thesis. Some alternatives for the logistic regression model in credit risk are also briefly described. Then we built the necessary theoretical background behind the random forest model (RF). To understand the structure of RF we describe the building blocks behind RF called decision trees. Finally, theory behind SHAP framework, which is used to enhance the interpretability of random forest model, is presented.

3.1 Logistic Regression

Logistic regression (LR) is a well-known classifier that is widely used on a large variety of different classification tasks. It performs also reasonably well on credit scoring (Fitzpatrick and Mues, 2016). The logistic regression model has a clear advantage over other machine learning methods. It allows statistical testing of the importance of input variables (Costa e Silva et al., 2020). Logistic regression models are also highly interpretable. Even though there are more advanced methods most international banks are still using logistic regression for credit scoring. (Dumitrescu et al., 2022)

A logistic regression model can be used to model the conditional probabilities of default using the logistic cumulative distribution function $F(.)$. The conditional probability can then be calculated using

$$P(y_i = 1|x_i) = F\big(\eta(x_i;\beta)\big) = \frac{1}{1 + \exp\left(-\eta(x_i;\beta)\right)}$$

where $\eta(x_i;\beta)$ is an index function defined as

$$\eta(x_i;\beta) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$$

where

$$\beta = \big(\beta_0, \beta_1, \dots, \beta_p\big) \in R^{p+1}$$

is the vector of unknown parameters. The estimator $\hat{\beta}$ for these unknown parameters can be estimated by maximizing the log-likelihood function

$$L(y_i, \beta) = \sum_{i=1}^{n} \left\{ y_i log\{F(\eta(x_i; \beta))\} + (1 - y_i) log\{1 - F(\eta(x_i; \beta))\} \right\}.$$

There is an important assumption behind the model. The index $\eta(x_1; \beta)$ must be linearly related to the predictive variables. With this assumption and model relative contribution of each predictor for the probability of default can be achieved using marginal effects

$$\frac{\partial P(y_i = 1|x_i)}{\partial x_{i,j}} = \beta_j \frac{\exp(\eta(x_i; \beta))}{[1 + exp(\eta(x_i; \beta))]^2}$$

Therefore, a positive (negative) predictive variable with a statistically significant coefficient has a positive (negative) effect on the probability of default. Due to the linearity assumption of the index $\eta(x_i; \beta)$, the model ignores non-linear relationships between y_i and predictive variables x_i. (Dumitrescu et al., 2022) This linearity assumption is the main limitation behind the logistic regression and therefore there is incentive to use other methods that can consider non-linear relationships between predictive variables and a target variable. Dumitrescu et al. (2022) propose the usage of a penalized logistic tree regression (PLTR) model, which combines decision trees with a logistic regression model. Fitzpatrick and Mues (2016) conclude that the Boosted Regression Trees (BRT) and Random Forests (RF) outperform a logistic regression model in mortgage default prediction. They argue that these models can capture non-linear relationships and therefore have greater performance compared to a logistic regression model. More recent Gradient Boosting or ensemble methods are also proposed to be used since they are expected to reach higher accuracies when predicting default risks. However, these models lack the interpretability of a logistic regression and therefore require other explainable models to be developed. (Gramegna and Giudici, 2021)

There are also other advantages of using a logistic regression model. Since banks are often required to create scorecards that describe how each predictive variable affects a customer's credit score, LR models are highly valued. This transparency meets regulatory requirements and gives an incentive to use relatively simple models. Since logistic regression models are still widely used, are highly interpretable, and are known to have reasonably good classification performance, they will be used as a benchmark model in this thesis.

3.2 Random Forest

Random forest is a classification algorithm that combines multiple decision tree predictors into a single model (Breiman, 2001). Therefore, it is first necessary to understand the process of decision tree modeling. Decision trees are interpretable ML models that can be used for various classification tasks. A decision tree is a hierarchical model where functions or decision rules are applied for the explanatory variables recursively to achieve discrimination between target classes. Decision trees are constructed by two distinct parts: branch nodes and leaf nodes. In a decision tree, a feature space containing all the explanatory variables is split into smaller subspaces using decision rules at each node. (Myles et al., 2004) Figure 2. shows the general structure of a decision tree algorithm.

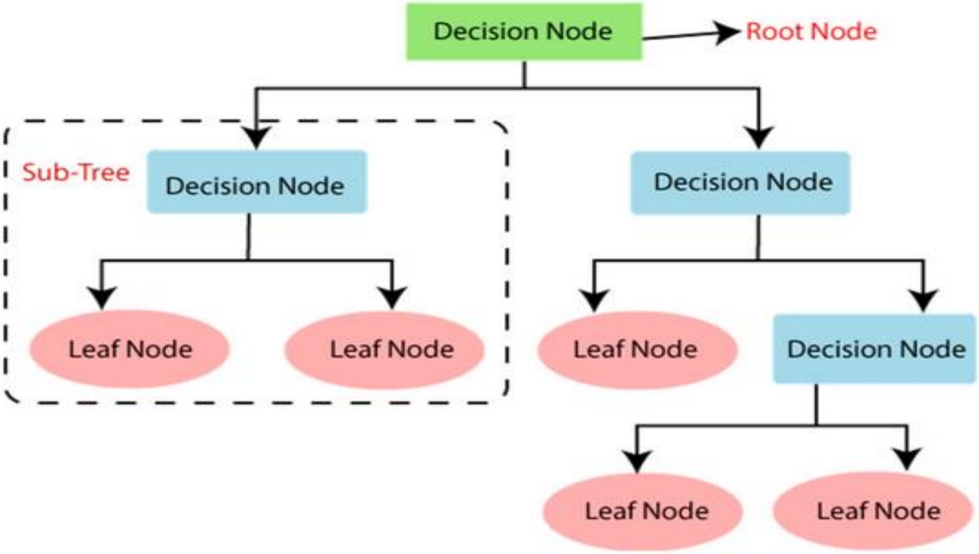


Figure 2. Decision tree (Alnemari and Alshammari, 2023)

The objective of a decision tree is to find a set of decision rules that split data at points that maximize class distinction. A set of partition rules is determined by a scoring criterion. There are two commonly used scoring criteria: information gain and the Gini index. Information gain can be calculated using the concept of Info. Info is described by equation

$$Info = -\sum_j (\frac{N_j(t)}{N(t)}) log_2 (\frac{N_j(t)}{N(t)}),$$

where N_j is the number of classes in sample j, $N(t)$ is the number of samples in node t and $N_j(t)$ is the number of class j samples in node t. A partition that maximizes the change in this Info (InfoGain) is then selected. InfoGain can be calculated using

$$\text{InfoGain} = \text{Info}(\text{Parent}) - \sum_{k} (p_k) Info(Child_k)$$

where $info(q)$ is the information from subspace q and p_k is the proportion of samples that pass into subclass k. (Myles et al., 2004)

The other widely used scoring criteria is the Gini index which measures the reduction in class impurity from portioning the feature space. This impurity can be calculated using

$$impurity = 1 - \sum_j | \left| p(j) \frac{N_j(t)}{N_j} \right| |^2).$$

The Gini index can then be calculated as

$$Gini = impurity(Parent) - \sum_k (p_k) Impurity(Child_k).$$

The selection of scoring criteria should not have a great effect on the model's prediction accuracy. Selection can affect which features are selected for a tree. (Myles et al., 2004) The main advantage of decision trees is their simple interpretability. However, simplicity comes with a cost on prediction accuracy. Trees also tend to overfit. Overfitting is a process where the model has more complexity than necessary. Overfitting can happen when the overly flexible model is fit on a dataset. An example of this would be to fit a neural network for data that conforms to the linear model. This would result in increased complexity without any additional performance improvement or even decreased performance. Overfitting can also happen when a model with irrelevant components is used: for example, using a linear regression model with excess irrelevant predictors. (Hawkins 2004)

A random forest model consists of multiple decision trees. For each tree a random subsample of the training dataset is used. These trees then vote for the most popular class in a classification problem and the most popular class is then selected as an output of a model. This bagging process solves the problem of overfitting and reduces the variance of unstable classifier such as individual decision trees (Thomas et al., 2017). Random forest classifier is usually constructed from the individual decision trees by utilizing random feature selection. For each tree, a random subset of features is selected, and a tree is grown on this new dataset. (Breiman, 2001). Construction of random forest requires a process called hyperparameter tuning. Random forest is also known to provide a good performance with default settings, but the performance can in some cases vary significantly depending on the selected parameters. Hyperparameters for a random forest are for example the number of trees, splitting (selection) criteria, and the sample size. (Probst et al., 2019) Hyperparameter tuning is beyond the scope of this thesis but for good practice, a grid search is used to find the best-performing model for the default prediction task. Since the grid search method is used to find out which hyperparameter combination provides the best classification results from the given set of possible hyperparameters, it may fail to find the absolutely best-performing subset of parameters. However, the main interest is on the interpretability of the selected model and not the absolute accuracy, and therefore this method is expected to be adequate for the purpose of this thesis.

The goal is to utilize interpretable tree-based machine learning methods e.g., random forest and compare their performance against a commonly used logistic regression model. Performance would be measured by AUC score as stated before. Additionally, other metrics like accuracy, precision, and recall. Instead of tree-based methods, potentially even more sophisticated methods like gradient boosting (LightGBM) or artificial neural networks (ANN) could be used. However, these could increase the complexity of models with uncertain performance improvement. Therefore, the objective is to utilize relatively simple tree-based methods e.g. random forest.

Implementation process of the methods is conducted using Python since I am the most familiar with it. The code and outputs should be understandable to people who have experience with R. R could be employed for the regression model due to its advanced statistical packages relative to Python. However, since the logistic regression model should be a baseline model, it will also be implemented in Python. Statsmodels for reference logistic regression and scikit-learn for other models, are the required Python packages for the implementation. Pandas and Numpy are used for data preparation and SHAP will be implemented using the SHAP package for Python.

3.3 SHAP

As already stated in the literature review part in the section 2.4, Lundberg and Lee (2017) propose a method that has its background on game theory. The idea is to measure feature importance i.e. how each feature impacts the classification result of a single data point. To understand this process, let's first consider the most fundamental case, namely, linear regression. The prediction of a linear regression model can be presented as

$$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \,.$$

For each instance x_j we can find the contribution ϕ_j on the prediction $\hat{f}(x)$ by using

$$\phi_j(\hat{f}) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j),$$

where $E(\beta_j X_j)$ is the mean effect estimate for feature j. Then we can calculate the sum of all feature contributions for one instance using

$$\sum_{j=1}^{p} \phi_j(\hat{f}) = \sum_{j=1}^{p} (\beta_j x_j - E(\beta_j X_j)) = (\beta_0 + \sum_{j=1}^{p} \beta_j x_j\,) - (\beta_0 + \sum_{j=1}^{p} E(\beta_j X_j)\,)$$
$$= \hat{f}(x) - E(\hat{f}(X))$$

meaning that the sum of feature contributions equals the difference between the predicted value and the average predicted value. (Molnar, 2018)

To find the feature contributions for any model we can used so called Shapley value. The Shapley value is defined as

$$\phi_j(\hat{f}, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M\,!} [f_x(z') - f_x(z' \setminus i)]$$

where $|z'|$ is the number of non-zero entries in z', and $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x'. M is a subset of the features used in the model and notation $z' \setminus i$ is equal to setting $z'_i = 0$. (Lundberg and Lee, 2017) Calculating SHAP values is a difficult task in practise. Estimating $f_x(z' \setminus i) = E[f\ (x) \mid x_S]$ where S is the set of non-zero indexes in z′ and

$E[f\ (x)\mid x_S]$ is the expected value of the function conditioned on a subset S of the input features is challenging. However, a method called *Tree SHAP* can be used to efficiently calculate SHAP values for a tree by estimating $E[f\ (x)\mid x_S]$. (Lundberg et al., 2019) These SHAP values can be presented using different SHAP plots. Commonly used plots are SHAP summary plots and SHAP dependence plots (Lundberg et al., 2019). Let's look at some details of these plots to understand how these can be used to display the feature contributions of a model in a way that is interpretable.

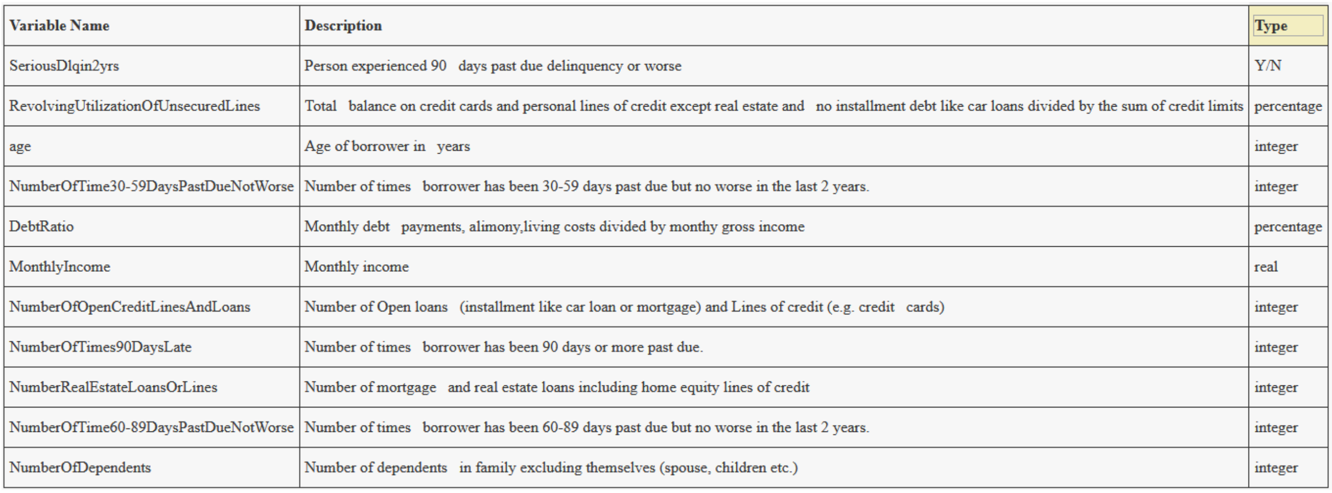
The feature importances of variables in a ML model including random forest models are traditionally presented using bar charts. However, these plots do not represent the range and distribution of impacts of a feature, and they do not display how these feature values are linked to feature's impact. SHAP summary plots solve this issue by leveraging individualized feature attributions while retaining interpretability. The other commonly used plot is called SHAP dependence plot. They replace the traditionally used partial dependence plots. The main advantage of using SHAP dependence plots instead of partial dependence plots is that they can display vertical dispersion by considering the interaction effects in the model. More details and examples of these plots are shown in the section 5.3 where these plots are used to explain the feature importances of a random forest model on a global level. In the section 5.3 local effects are also shown which are also an interesting part for the goals of this thesis.

4 Data and Data Preparation

In this chapter data used for the probability of default estimation is described with greater details. I also present some limitations and challenges related to the dataset used for this task. Additionally, data preprocessing is described. Special emphasis is on the preprocessing of the training data used for the logistic regression model since the usage of advanced binning algorithm is required.

4.1 Data Collection

Since credit risk data usually contains sensitive personal data, banks cannot share datasets that could be used for this kind of research. Fortunately, several publicly available datasets exist. The dataset intended to be used for this study is the Give Me Some Credit dataset available on Kaggle, which is the world's largest data science community with many publicly available resources. The dataset contains historical data for 150 000 borrowers and should be sufficient for this kind of study.
The data contains 10 different variables that can be used to predict whether a customer is likely to default during the next 2 years. A detailed description of the dataset is shown in table 1.



Variable Name	Description	Type
SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

Table 1. Description of the dataset (Cukierski, 2011)

The objective is to use Give me some credit dataset available publicly on Kaggle. The dataset was originally used for a credit risk prediction competition held in Kaggle which is the world’s largest data science community online. The dataset includes data for 150 000 borrowers and the variables included in the dataset are shown in table 1.

Unfortunately, the dataset lacks more detailed descriptions. It is unclear where the dataset was originally collected from. Most likely data is pseudonymized customer data provided by a bank. Lacking descriptions for the dataset may raise some concerns about the validity of this study. However, I argue that this dataset is sufficient for the task at hand. Since the main task is to compare classification performance of two different classification algorithms and not to draw any causal inference, the lack of the descriptions should not be a major issue. Additionally, the dataset is widely used to measure the performance of different classification algorithms, and it has been used in other research articles as well for example Fitzpatrick and Mues (2016).

The main advantage of the dataset is that enables relatively simple classification task. The idea is to predict whether a customer has a serious delinquency during the next 2 years (variable SeriousDlqin2yrs). Dataset also has predefined set of variables so detailed feature selection process is not required. This can be also a challenge since it is possible that some meaningful risk drivers are missing from the dataset. However, since the objective is to compare the performance between two models, lacking variables should not present a threat for the validity of the results. It is important to emphasize that the goal is not to assess whether variables have causal effects on the probability of default. If one wants to estimate causal effects, then variable selection process and the properties of variables should be very carefully studied

4.2 Data Preprocessing

To understand the relationships between variables correlation plot was drawn. The plot shows the linear correlation between variables using the Pearson correlation coefficient. This plot is used to determine whether the feature selection is adequate for the classification task.

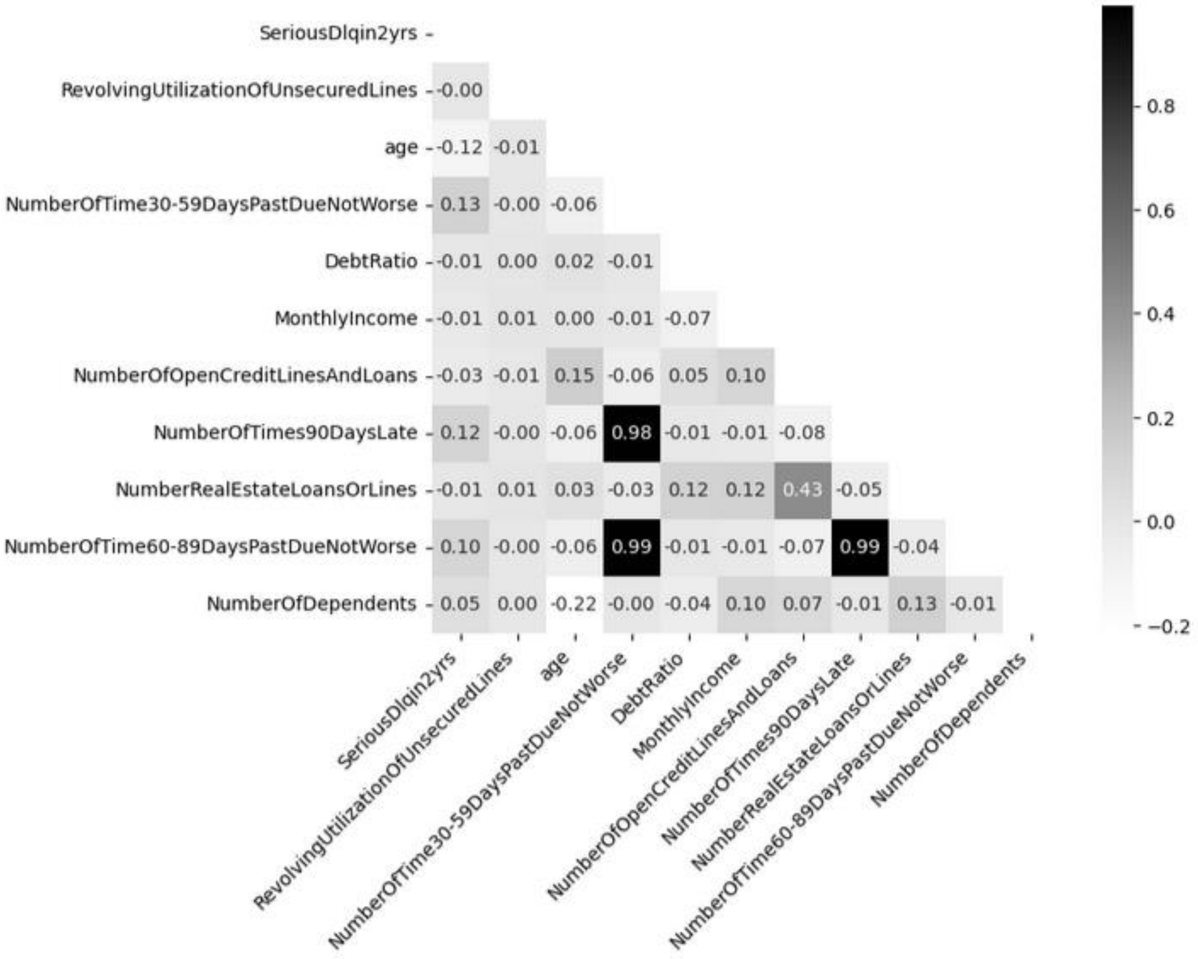


Figure 3. Lower triangle of the correlation matrix.

The correlation matrix shows a few issues within the variables. Variables describing short term delinquencies are highly correlated. This conclusion is backed by the variation inflation factor (VIF). VIF can be defined using the concept of tolerance. Tolerance is measured using equation

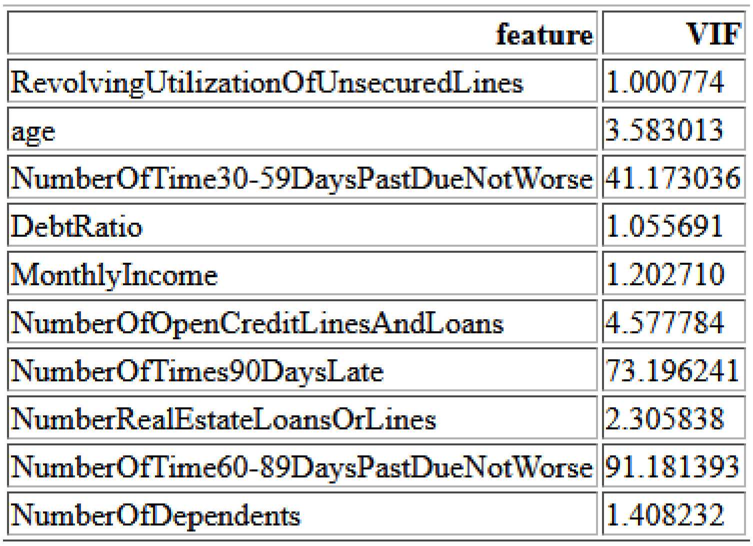
$$Tolerance\ =\ 1 - R^2\ ,$$

where R^2 is the coefficient of determination for the regression where a specific variable is regressed on all other remaining independent variables.

From this VIF can be defined as the reciprocal of tolerance:

$$VIF = \frac{1}{Tolerance}.$$

VIF value indicates how much variance of the estimates is being inflated by multicollinearity. There does not exist a formal cutoff point that can be used to detect a significant multicollinearity. Generally, VIF values exceeding 10 are considered to indicate multicollinearity. Even smaller values can be significant in some cases and should be considered as well. (Senaviratna and A Cooray, 2019).



feature	VIF
RevolvingUtilizationOfUnsecuredLines	1.000774
age	3.583013
NumberOfTime30-59DaysPastDueNotWorse	41.173036
DebtRatio	1.055691
MonthlyIncome	1.202710
NumberOfOpenCreditLinesAndLoans	4.577784
NumberOfTimes90DaysLate	73.196241
NumberRealEstateLoansOrLines	2.305838
NumberOfTime60-89DaysPastDueNotWorse	91.181393
NumberOfDependents	1.408232

Table 2. Variance Inflation Factor (VIF) for Independent Variables

VIF values for the variables in the *Give me some credit* dataset are shown in table 2. High VIF values for delinquency variables verify the conclusion that there is significant multicollinearity present within the delinquency variables. To avoid issues in the estimation process of the logistic regression model variables “NumberOfTime60-89DaysPastDueNotWorse” and “NumberOfTimes90DaysLate” are omitted from the dataset. Those variables will be omitted also from the random forest model since comparability between models should be maintained.

The second important consideration is the handling of missing values. For “MonthlyIncome” there are 29 731 observations with missing values. Additionally, for “NumberOfDependents” there are 3 924 missing values present in the dataset. For the simplicity of implementation these missing values are replaced by the value of 0. This makes the implementation of a binning algorithm, that is introduced later in this thesis, more straightforward. Replacement can raise concerns if some information is lost. Indeed, it is possible that a better approach would be to create a dedicated category for missing values to preserve the possible information value that missing values could still possess. However, since the idea is to compare the performance between two different models and not optimize the model performance, this approach is argued to be sufficient.

One of the assumptions for logistic regression is that data should not have outliers. To investigate this box plots (or box-and-whisker plots) are plotted for each variable and are shown in figure 4. From figure 4, it is evident that the data has outliers. These outliers are marked as points that can be seen outside of whiskers in box plots. The whiskers in these plots extend to points that lie within 1.5 times the interquartile range (IQR) of the lower and upper quartile. This IQR is the set of datapoints contained within the second and third quartiles (25 % and 75 %) of the dataset. Therefore, the box in a box plot describes this IQR and the black line in the middle of the box is the median value of the given data.

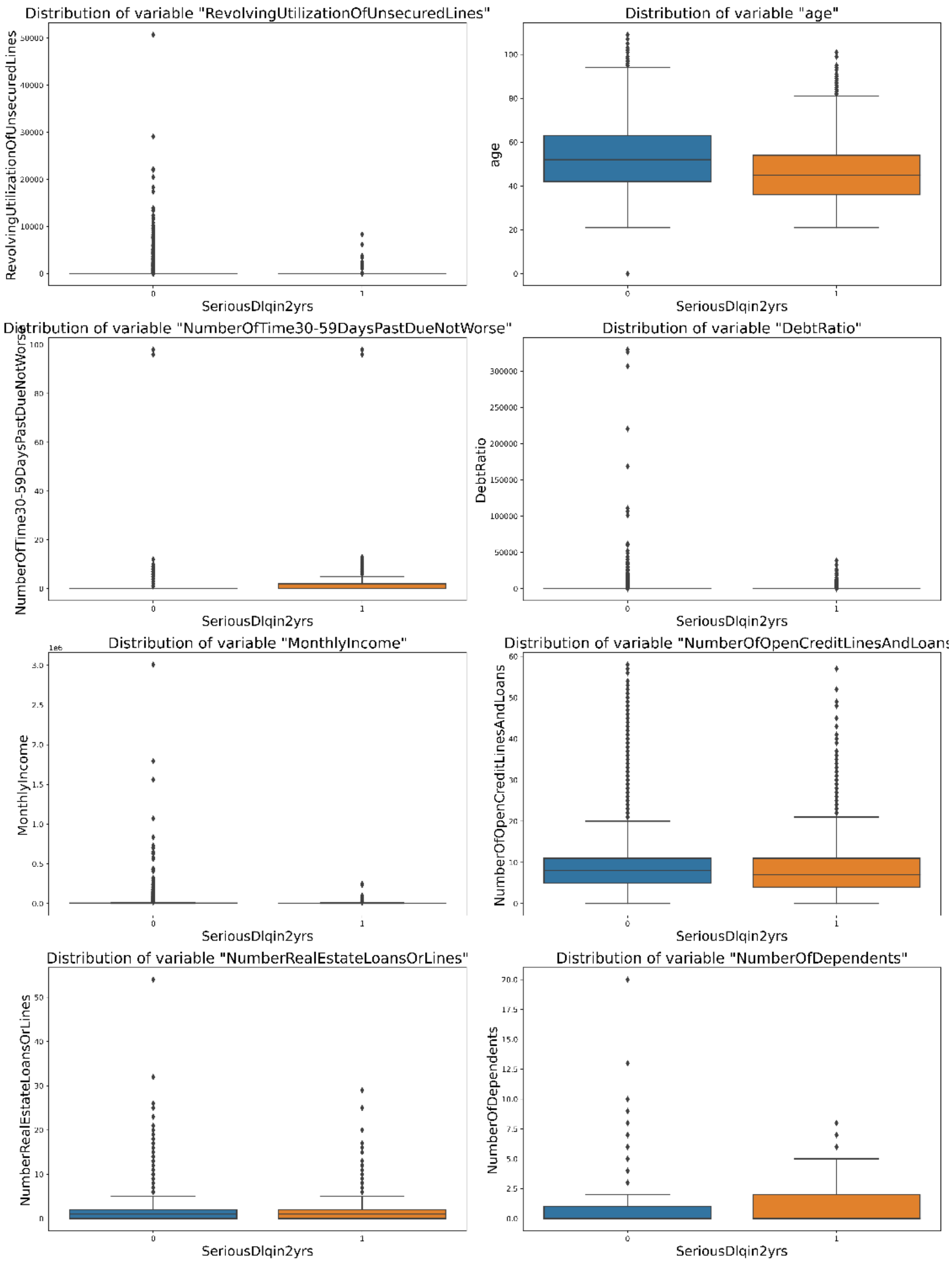


Figure 4. Variable distributions

Since the variables clearly have outliers that could significantly affect logistic regression estimates, these outliers should be omitted from the dataset. To avoid information loss instead of omitting observations with outliers, a binning algorithm is used. There are numerous different options for a binning algorithm that could be used to categorize continuous variables. Some commonly used methods include for example equal-width and equal-size binning algorithms that are the most obvious and straightforward approaches as well. In equal-width binning the number of bins is predefined. Then for each variable the values are binned in the bins with equal widths based on the range of the values. In the equal-size binning the number of bins is also predefined. With this approach the goal is to have the same number of observations in each bin. Therefore, the width of a bin varies based on the density of a variable. (Mironchyk, P. and Tchistiakov, V., 2017) These simple binning algorithms fail to consider the relationship with the target variable (dependent variable). Additionally equal-width binning may be affected by the outliers since the range of values is affected by these outliers. Equal-size binning can deal with outliers, but the number of bins can vary widely, which could complicate the interpretation of the results.

To avoid these issues more advanced binning algorithms are proposed. These include for example Chi-Merge and Optimal binning. (Mironchyk, P. and Tchistiakov, V., 2017) To obtain all the field specific requirements a binning algorithm called monotone optimal binning proposed by Mironchyk and Tchistiakov is selected for this thesis. The algorithm meets all the requirements for a good binning algorithm. It allows for monotonicity meaning that when moving from one bin to another, there is a monotonic change in risk. Additionally, the algorithm ensures that there is maximum correlation between the risk indicator (independent variable). (Mironchyk, P. and Tchistiakov, V., 2017)

It is possible that the choice of a binning algorithm affects the performance of a logistic regression significantly. Usage of monotone optimal binning proposed by Mironchyk and Tchistiakov may raise also raise some concerns if the binning algorithm is indeed the best available algorithm for this task. Since the article is a working paper and shows some clear indications of weaknesses it is possible that some other binning algorithm would perform better. For example, the algorithm is not tested comprehensively and Python implementation of the algorithm using

MOBPY package is not completely error free. It is possible that a bin has zero accounts for defaults which violates the criteria for a good binning algorithm. The choice of a binning algorithm is an interesting topic that has not been widely studied and would be a possible topic for further research. However, for the purpose of this thesis the monotone optimal binning is deemed adequate.

Categorized variables can be presented using weight of evidence (WoE) encoding that is used as a measure of bin's predictive ability separating defaulted and non-defaulted customers. WoE values can also be used to compare the effects of bins on to the target variable. The WoE is calculated using the following equation:

$$WoE_i = \left[\ln\left(\frac{Relative\ frequency\ of\ Goods}{Relative\ frequency\ of\ Bads}\right)\right] * 100 \ ,$$

where the relative frequency of goods is the ratio of non-defaulted customers in bin_i to the total number of non-defaulted customers and the relative frequency of bads is the ratio of non-defaulted customers in bin_i to the total number of defaulted customers. (Davis et al., 2023) Categories and Woe values are highlighted in the figure 5. The left y-axis (blue bars) indicates the WoE value for each category and the proportion of defaulted loans for each category is shown on the right y-axis (red dots). Categories are monotonic for each variable as the WoE is strictly increasing or decreasing when moving to the right on categories indicating that the binning is working as expected.

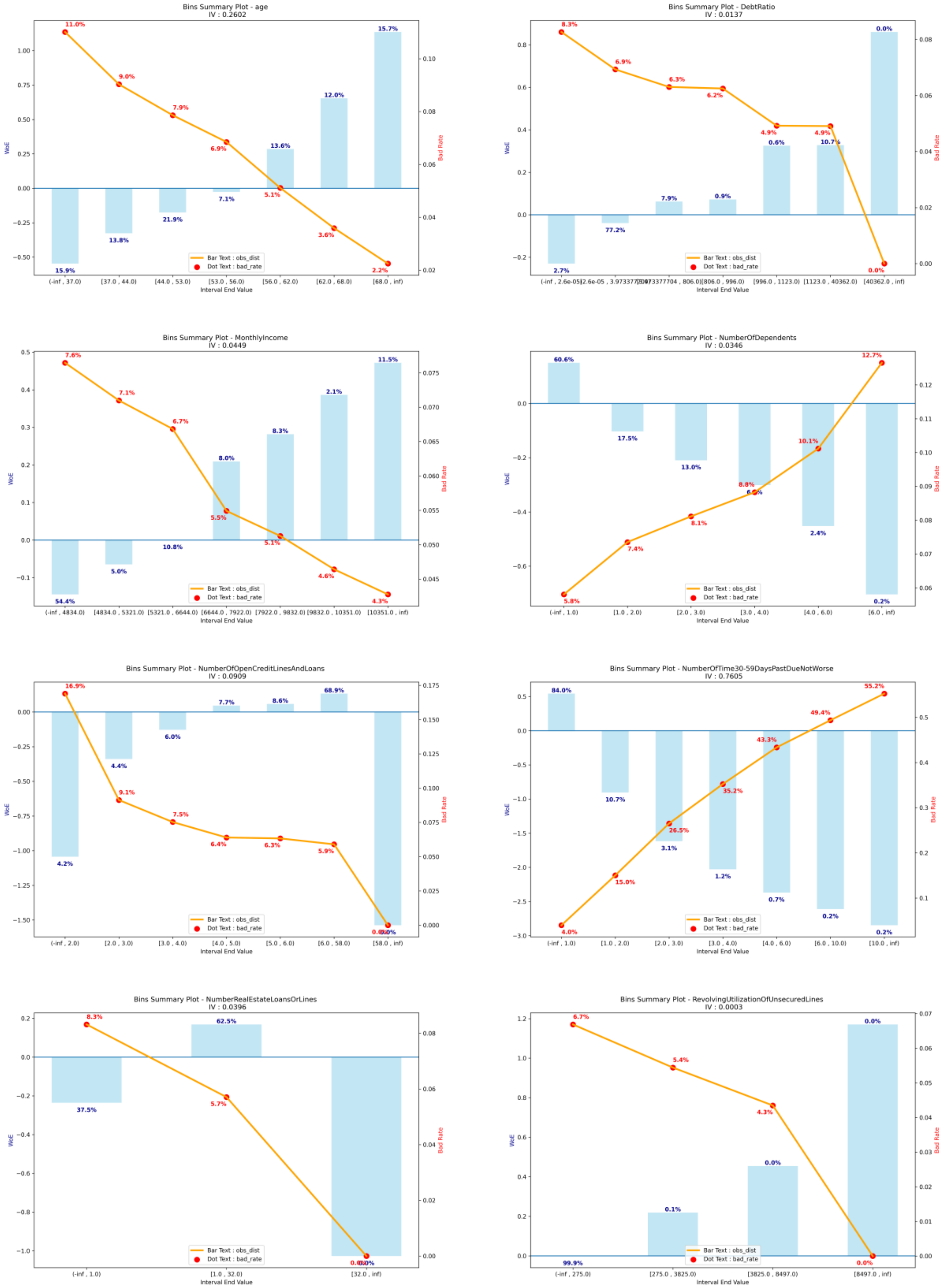
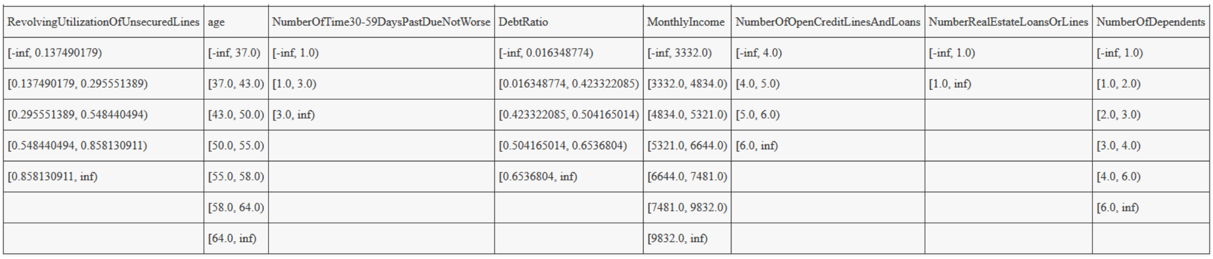


Figure 5. Binning summary results

There is still a slight issue since some categories have only non-defaulted observations. This violates the requirements of a good binning algorithm. This issue rises from the MOBPY implementation of the algorithm since there seems to be a bug which allows for these categories to exist even though “min_bads” parameter has been used. This parameter should have ensured that there is adequate number of defaults in each category but for unknown reason different parameter values do not influence bins.

Exact reason behind this issue was hard to determine but it was likely identified to be the incorrect handling of large outlier values. To attend this issue outliers were removed from the dataset for affected variables including RevolvingUtilizationOfUnsecuredLines, DebtRatio, MonthlyIncome and NumberRealEstateLoansOrLines. It was determined that the extreme outlier values for these variables were causing issues with the binning algorithm. While removing these outliers may cause some information loss since the number of observations is reduced. However, to meet assumptions behind the logistic regression model and the selected binning algorithm this trade-off is necessary. After these modifications the categories for each variable are presented in table 3.



RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30-59DaysPastDueNotWorse	DebtRatio	MonthlyIncome	NumberOfOpenCreditLinesAndLoans	NumberRealEstateLoansOrLines	NumberOfDependents
[-inf, 0.137490179)	[-inf, 37.0)	[-inf, 1.0)	[-inf, 0.016348774)	[-inf, 3332.0)	[-inf, 4.0)	[-inf, 1.0)	[-inf, 1.0)
[0.137490179, 0.295551389)	[37.0, 43.0)	[1.0, 3.0)	[0.016348774, 0.423322085)	[3332.0, 4834.0)	[4.0, 5.0)	[1.0, inf)	[1.0, 2.0)
[0.295551389, 0.548440494)	[43.0, 50.0)	[3.0, inf)	[0.423322085, 0.504165014)	[4834.0, 5321.0)	[5.0, 6.0)		[2.0, 3.0)
[0.548440494, 0.858130911)	[50.0, 55.0)		[0.504165014, 0.6536804)	[5321.0, 6644.0)	[6.0, inf)		[3.0, 4.0)
[0.858130911, inf)	[55.0, 58.0)		[0.6536804, inf)	[6644.0, 7481.0)			[4.0, 6.0)
	[58.0, 64.0)			[7481.0, 9832.0)			[6.0, inf)
	[64.0, inf)			[9832.0, inf)			

Table 3. Variable categories from monotone optimal binning

5 Estimation

In This section model estimation process for logistic regression as well as for a random forest is described. The logistic regression model is built on the prepossessed dataset described in the previous section. Model uses WoE transformed categorical variables as independent variables to predict the target variable (default probability). The dataset is divided into a training dataset used for model training and validation dataset that provides unbiased evaluation for the model's performance. Training and validation datasets consist of 70 % and 30 % of total observation respectively. The model is described with greater details and the model's performance is measured using metrics that were described in the section 2.3. The main metric is the AUC score but other metrics like precision and recall are reported as well to fully understand the performance of a model.

After the logistic regression model is trained and evaluated, a random forest model is built. The theoretical background on the model is described with greater details in section 3.2. The model is built separately on two different datasets. At first, the same pre-processed dataset is used as was used for the logistic regression. This approach allows for strict performance comparison where the only performance gains are explained by the properties of a model and not the differences between datasets. As the dataset was built to meet the assumptions behind the logistic regression model it is also a good idea to test whether a random forest model would gain performance improvements if these assumptions were removed. Since the data for random forest model is not required to meet these assumptions, it is sensible to measure the performance of the model using the whole dataset available as well. Additionally, a grid search is performed to find optimal parameter values for the model. Finally, the performance between the logistic regression model and the random forest is compared to identify whether it possible to achieve performance gains from more complex machine learning models.

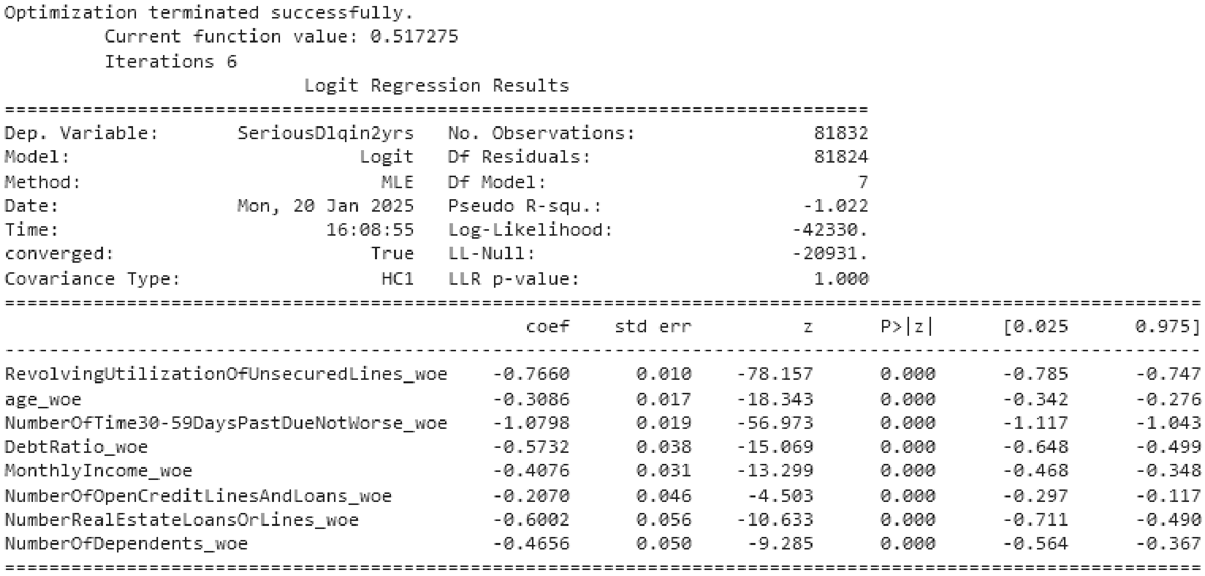
After the modelling process the main goal is to analyse the explainability of the model is using SHAP framework. Since even a small performance increase could yield substantial economic benefits for banks, there are high incentives for using the best available models. However, to meet regulatory requirements described in the section

2.1 the usage of black-box models is prohibited. This provides clear motivation for this analysis.

5.1 Logistic Regression Model

As already mentioned, the goal is to estimate a logistic model that is used as a benchmark model in this thesis. The model is estimated using pre-processed dataset that contains WoE values for each binned variable. Bins are shown in table 3. and the WoE values are presented in the figure 5.

To understand the estimation results the idea is to at first investigate the estimation summary from another Python package called statsmodels. Logistic regression model was estimated using the preprocessed dataset and estimation results are presented in figure 6. Coefficients for each variable are negative as can be seen from the figure 6 and all coefficients are statistically significant. These negative coefficients are explained by the nature of the used model. Negative coefficients indicate that increase in an independent variable leads to lower odds of defaulting. Since WoE values are used as independent variables these coefficients are expected to be negative by definition. Larger WoE values indicate that distribution of defaults > distribution of non-defaults for a variable. Therefore, increase in WoE would result in lower odds of defaulting which is expected. Note that there is no causal interpretation here, WoE values only describe correlation and cannot be used for causal estimates per se.



```
Optimization terminated successfully.
         Current function value: 0.517275
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:       SeriousDlqin2yrs   No. Observations:                81832
Model:                          Logit   Df Residuals:                    81824
Method:                           MLE   Df Model:                            7
Date:                Mon, 20 Jan 2025   Pseudo R-squ.:                  -1.022
Time:                        16:08:55   Log-Likelihood:                -42330.
converged:                       True   LL-Null:                       -20931.
Covariance Type:                  HC1   LLR p-value:                     1.000
============================================================================================================
                                               coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
RevolvingUtilizationOfUnsecuredLines_woe    -0.7660      0.010    -78.157      0.000      -0.785      -0.747
age_woe                                     -0.3086      0.017    -18.343      0.000      -0.342      -0.276
NumberOfTime30-59DaysPastDueNotWorse_woe    -1.0798      0.019    -56.973      0.000      -1.117      -1.043
DebtRatio_woe                               -0.5732      0.038    -15.069      0.000      -0.648      -0.499
MonthlyIncome_woe                           -0.4076      0.031    -13.299      0.000      -0.468      -0.348
NumberOfOpenCreditLinesAndLoans_woe         -0.2070      0.046     -4.503      0.000      -0.297      -0.117
NumberRealEstateLoansOrLines_woe            -0.6002      0.056    -10.633      0.000      -0.711      -0.490
NumberOfDependents_woe                      -0.4656      0.050     -9.285      0.000      -0.564      -0.367
============================================================================================================
```

Figure 6. Estimation summary of Statsmodels Logistic regression.

Since the goal is to build a classification model LogisticRegression classifier from Scikit-learn is used to build the final model. It is important to note that therefore the coefficients in figure 6 do not present the coefficients of the final model and are used only for the reference and to develop understanding of a logistic regression models. Th final logistic regression model is estimated using Scikit-learn machine learning package in Python. Model is fitted using LogisticRegression classifier that implements a regularized logistic regression. There are multiple different solvers available that can be used for the estimation process. Solver is a method used to solve the unconstrained optimization problem, that is in this case the problem of maximizing the log-likelihood function described in the section 3.1. The default solver LogisticRegression classifier is called L-BFGS which stands for a limited memory BFGS or limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm (Xiao et al., 2008). Selection of the solver should not be important for the estimation results but is more related to the computational efficiency. Unregularized logistic regression problem is an unconstrained convex optimization problem that is usually solved efficiently using standard convex optimization methods like Newton's method. However, when regularization is added the problem becomes more complex and may require other methods. (Lee et al., 2006) Since regularization is advisable to avoid overfitting in the modelling process L2 penalty (Ridge Regression) is used to avoid overfitting issues. For the logistic regression all parameter values were therefore left as default values. It should be noted that the parameter optimization could yield performance increments also for the logistic regression but since the goal is to build a benchmark model this process was omitted.

The estimated logistic regression model yields the ROC curve shown in figure 7. The corresponding AUC score for the model calculated from the ROC curve is 0.82. Generally, the AUC score between 0.8 and 0.9 can be considered as “excellent“ (Mandrekar, 2010). This would indicate that the logistic regression model has a good discriminatory power between non-defaulted and defaulted observations. Compared to other studies the AUC is relatively high. For example, Fitzpatrick and Mues (2016) achieved the AUC score of approximately 0.77 for their logistic regression model. Naturally direct comparison between two studies is not sensible since they used different dataset. Dumitrescu et al. (2022) estimate a logistic regression model using the same Give me some credit dataset that was used in this thesis. Their linear logistic regression model only achieves the AUC score of approximately 0.70 which is significantly lower that was achieved in this thesis. There are couple of reasons that could explain this difference. The binning method used in this thesis to deal with nonlinear relationships and outliers was not used by Dumitrescu et al. In fact, their data preprocessing was relatively simple, and they only filled missing values with the mean value of corresponding variable. Other explanation could be that Dumitrescu et al. used cross fold validation where a model is trained using different sub-samples of dataset and evaluated multiple times on different validation sets. Therefore, it is possible that the high AUC score achieved in this thesis is only by chance. To investigate the second explanation, a 5-fold cross validation AUC scores were calculated for the logistic regression model estimated in this thesis. The mean AUC from these five cross folds was approximately 0.819. Therefore, the difference between model performances must come from differences in data preprocessing or the model setting. The monotonic optimal binning likely improves the model performance and therefore the selected approach was justified.

Precision score is quite high 94 % of non-defaulted observations were classified as non-defaulted. However, the recall score is low, only approximately 8 %. The low recall score indicates that only 182 defaulted observations out of 2384 defaults in the validation dataset were correctly classified as defaults. This low recall is expected since the classifier classifies observations with log odds of 0.5 (cut-off point) or more as defaults. Since the dataset is imbalanced, probabilities are biased towards the more common event. Therefore, to achieve higher recall different cut-off point would be required (for example 0.2). In a real word application, a bank would never agree to loan for a customer with the default probability of 50 % and therefore lower cut-off values would be used. Since the main performance metric in this thesis is the AUC score, finding the optimal cut-off point is not required.

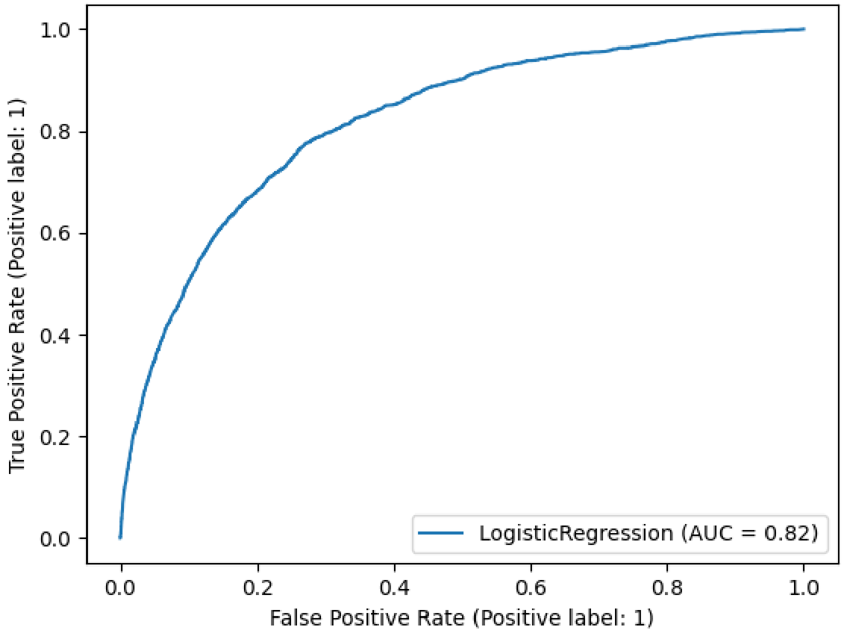


Figure 7. Logistic regression model, ROC curve and AUC score

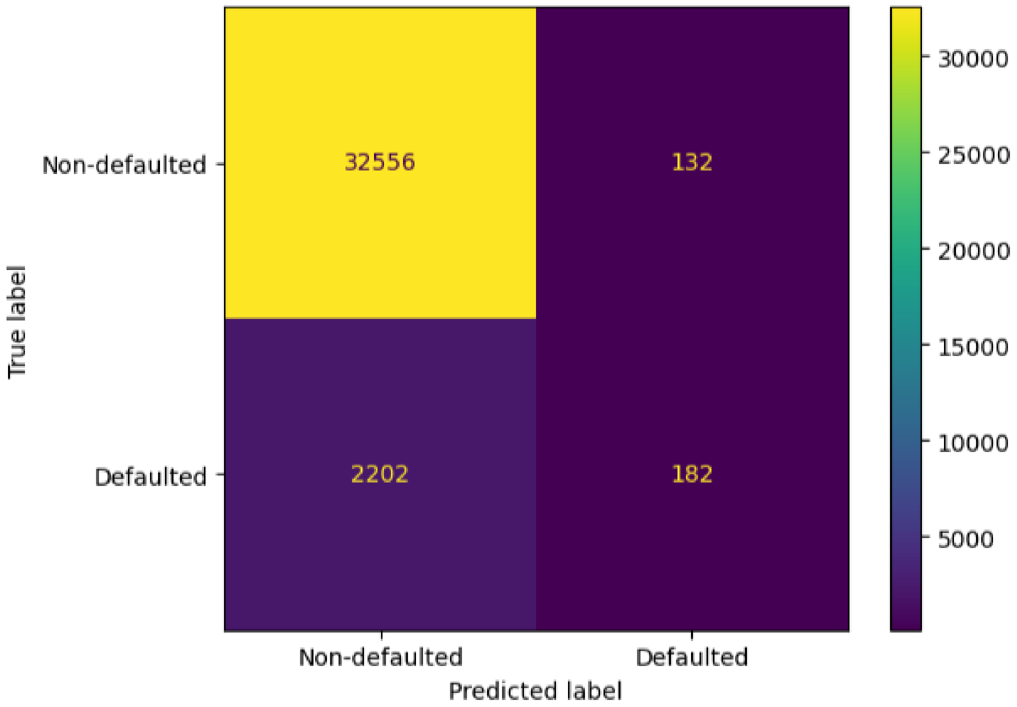
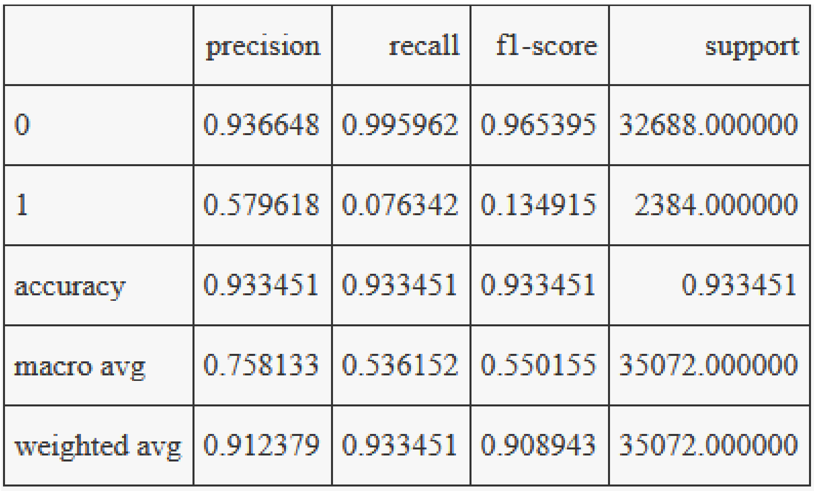


Figure 8. Logistic regression model, confusion matrix



	precision	recall	f1-score	support
0	0.936648	0.995962	0.965395	32688.000000
1	0.579618	0.076342	0.134915	2384.000000
accuracy	0.933451	0.933451	0.933451	0.933451
macro avg	0.758133	0.536152	0.550155	35072.000000
weighted avg	0.912379	0.933451	0.908943	35072.000000

Figure 9. Logistic regression model, classification report

5.2 Random Forest Model

The baseline random forest was trained on the same training dataset as the logistic regression model. This selection leads to an information loss as binning omits outliers from the dataset and enables linearity. The random forest algorithm is able to deal with outliers and non-linear relationships. Therefore, it is expected that this baseline model will not perform optimally. However, to understand where the possible performance improvements are coming from, comparison between this base model and the logistic regression model is performed.

The initial random forest was trained using default class weights meaning that the training sample was not underweighted or overweighted based on the distribution of the target variable. This will significantly reduce the recall score of the model since will be oversensitive for the majority class (non-defaulted) as discussed in the section 2.3. The number of estimators (individual trees in the forest) was set as 30. To avoid overfitting the max depth was set as 9. This controls the maximum depth of a tree. If this is left as 0, nodes in each tree would be expanded until all leaves are pure or until all leaves contain a single sample. This would result in overfitting and is therefore controlled by setting the maximum depth for each tree.

The figure 10. displays the ROC curves as well as the AUC scores from the initial random forest and the logistic regression models. Since the random forest was trained using the same binned dataset, the performance increase is insignificant. The random forest discriminates well between defaults and non-defaults but as expected, we can see from figures 11. and 12. that the recall score for this model is low. Only 72 out of 2384 serious delinquencies were correctly classified. This would be easily fixed by using a sampling method but since we are mainly interested about the discriminatory power of the model (AUC score), this sampling process is omitted for now. The justification for this it that the AUC score already considers all possible decision thresholds. However, to evaluate true performance benefits of a random forest model, we need to build a model that utilises non-linear dependencies within the data. Therefore, we train an additional model using the full available dataset.

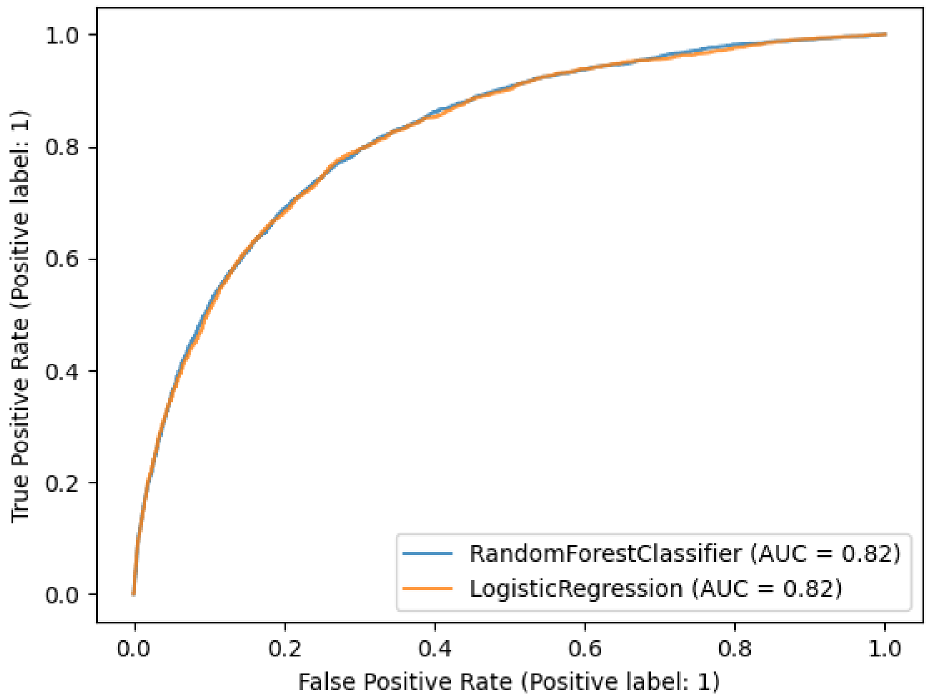


Figure 10. ROC curve and AUC scores for the initial random forest and logistic regression models.

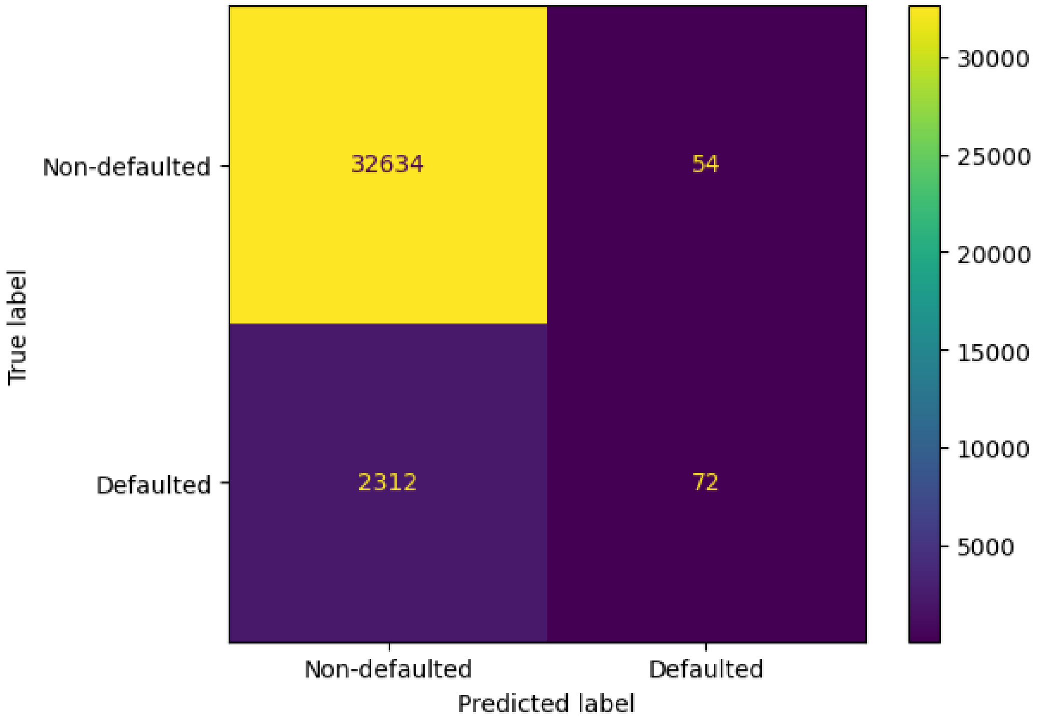
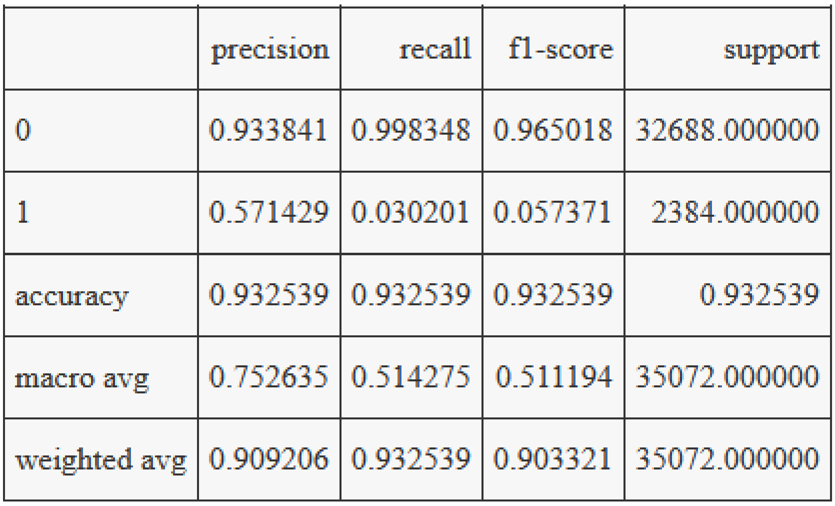


Figure 11. Initial random forest model, confusion matrix



	precision	recall	f1-score	support
0	0.933841	0.998348	0.965018	32688.000000
1	0.571429	0.030201	0.057371	2384.000000
accuracy	0.932539	0.932539	0.932539	0.932539
macro avg	0.752635	0.514275	0.511194	35072.000000
weighted avg	0.909206	0.932539	0.903321	35072.000000

Figure 12. Initial random forest model, classification report

The second model was trained using the full available dataset. Since the random forest algorithm is not limited by the linearity assumption or the existence of outliers within the data, this approach is feasible. For this model the same train/validation split was used as 70 % of data was used for the training process and 30 % of the data was used to validate the performance of the model. A grid search was performed to find the optimal parameter values for this final model. The grid search is a method where optimal hyper-parameter values are exhaustively generated from a grid of parameter values. In practice a model is given a grid of possible hyper-parameter values, and a model is trained using all possible combinations of these values. This is usually combined with a cross-fold validation to ensure that each parameter combination is evaluated exhaustively. To find the optimal set of hyper-parameters a grid search was performed where possible values for maximum depth of individual trees was a set (6,9,12) and the number of estimators (individual trees) was a set (20, 40, 100). The optimal set of parameters that resulted in the highest AUC score was a maximum depth of 9 and 100 estimators. The optimization process was limited to these two hyper-parameters only for performance reasons since the possible parameter combinations grow fast if more parameters are included in the grid search process. It should be noted that since this grid search was narrow, it is likely that an even better model exists. However, even with this limited optimization the random forest outperforms the logistic regression model.

For this optimal model the ROC curve as well as the AUC score are displayed in figure 13. This time the performance of the random forest model is significantly improved. It achieved the AUC score of 0.86 (0.862 for 5-fold cross validation). There is a significant improvement with the discrimination power of the model. The results are in line with the Dumitrescu et al. (2022) as the AUC for the random forest model in their study was approximately 0.85. A slight difference in the AUC scores between this thesis and their paper likely results from different hyper-parameters used for model estimation. Additionally, the recall score is significantly improved, and the model was able to classify 17 percent of serious delinquencies correctly. As already mentioned, this could be further improved using different sampling methods but since we are interested in the discriminatory power of models this is not required.

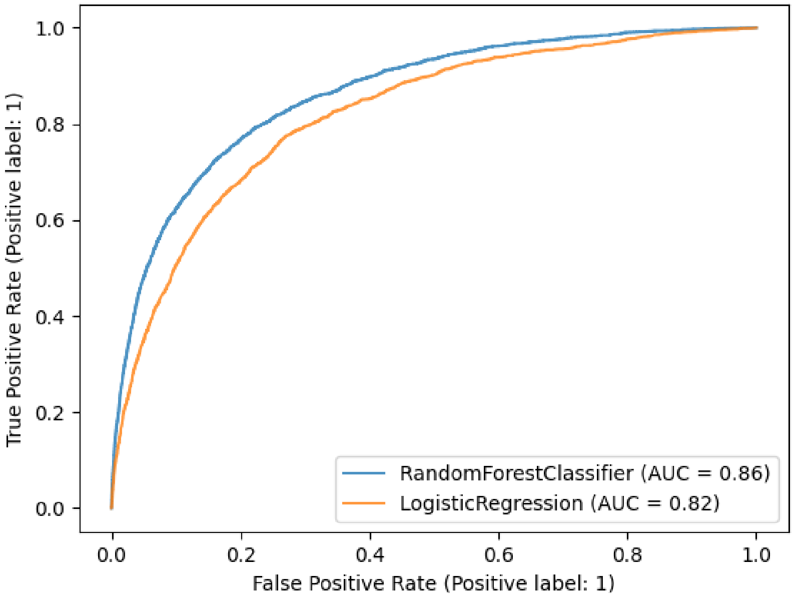


Figure 13. ROC curve and AUC scores for the optimized random forest and the benchmark logistic regression models.

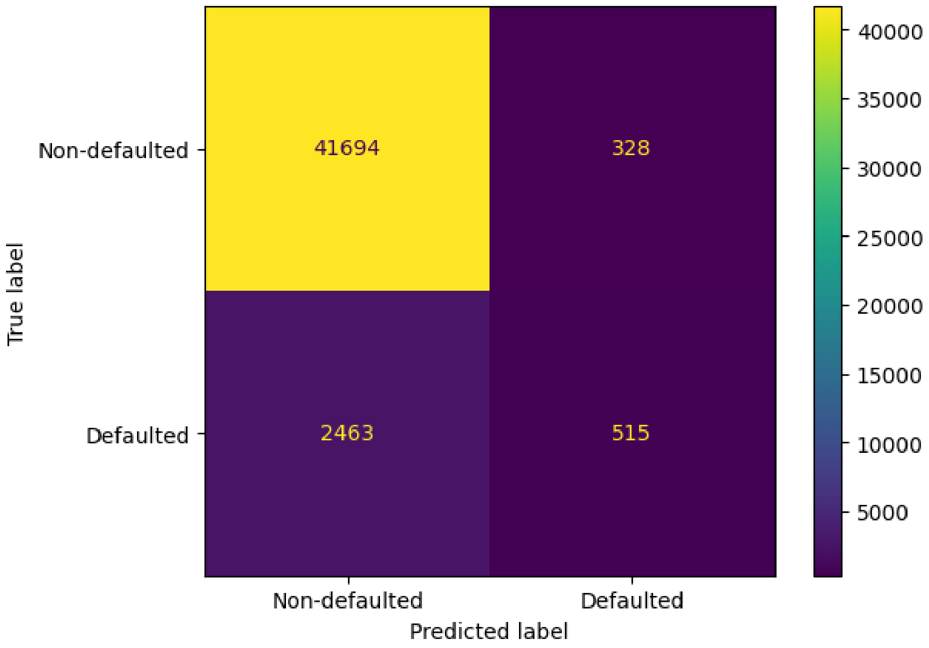
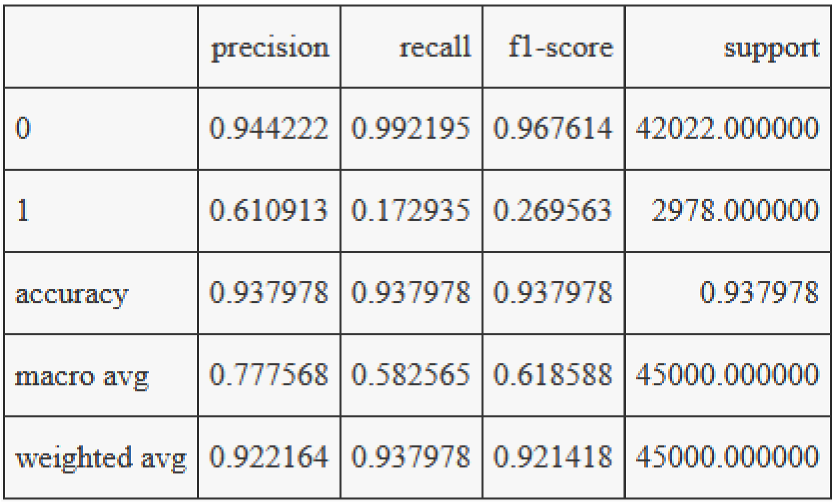


Figure 14. Optimized random forest model, confusion matrix



	precision	recall	f1-score	support
0	0.944222	0.992195	0.967614	42022.000000
1	0.610913	0.172935	0.269563	2978.000000
accuracy	0.937978	0.937978	0.937978	0.937978
macro avg	0.777568	0.582565	0.618588	45000.000000
weighted avg	0.922164	0.937978	0.921418	45000.000000

Figure 15. Optimized random forest model, classification report

5.3 Interpreting the Random Forest Model

In this section the built optimized random forest model is interpreted using SHAP framework. Interpretation is done by visualizing SHAP values using SHAP plots described in the section 3.3. The idea is to see whether the feature importances of the model can be presented in a way that is easy to understand could be used to explain these black-box models in a way that would meet regulatory requirements. SHAP values are calculated in Python using TreeExplainer from SHAP package. This TreeExplainer implements algorithm proposed by Lundberg et al. (2019) to efficiently calculate SHAP values for a tree-based model.

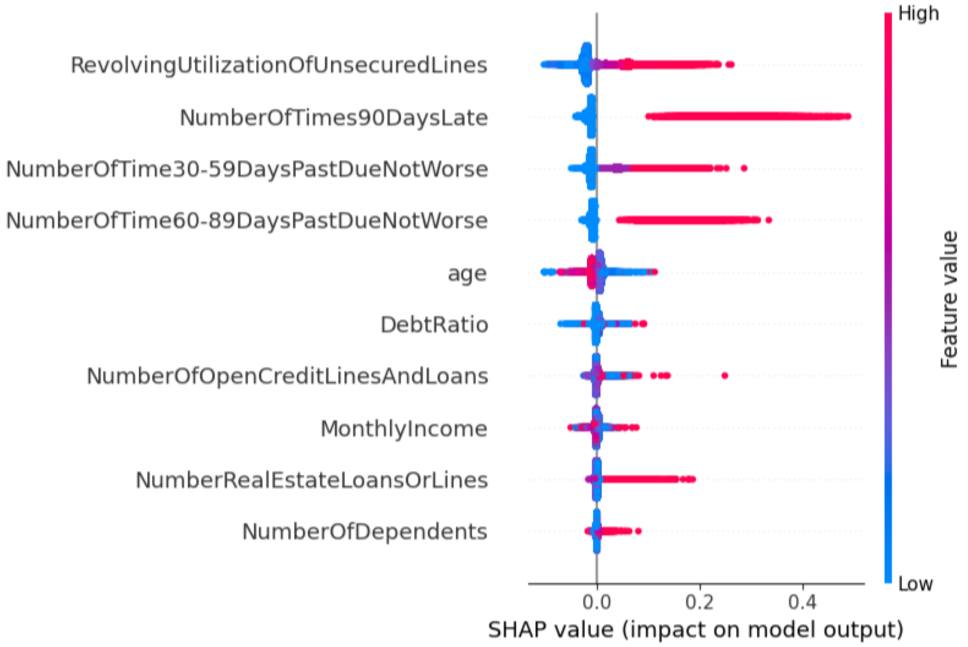


Figure 16. SHAP summary plot of the optimized random forest model.

To plot global feature importances, features are first ordered by their global impact and then SHAP value for each observation is plotted horizontally. If there are more observations with the same SHAP value these observations are stacked vertically to show density. (Lundberg et al., 2019) The SHAP summary plot displayed in figure 16 shows the global importance of each feature in the random forest model. The most important feature is in this case RevolvingUtilizationOfUnsecuredLines that describes total balance on credit cards and personal lines of credit except real estate and no instalment debt like car loans divided by the sum of credit limits. This is followed by the features that describe previous delinquencies. In this plot higher SHAP values indicate higher log-odds of serious delinquency. For example, higher values of RevolvingUtilizationOfUnsecuredLines generally lead to higher log-odds of default. It is important to note that the effect is not necessarily linear as can be easily seen for example for age variable where higher values generally lead to lower log-odds of default but can also lead to high log-odds depending on the observation. This is possible since the random forest model is able to capture non-linear interactions between the independent variables.

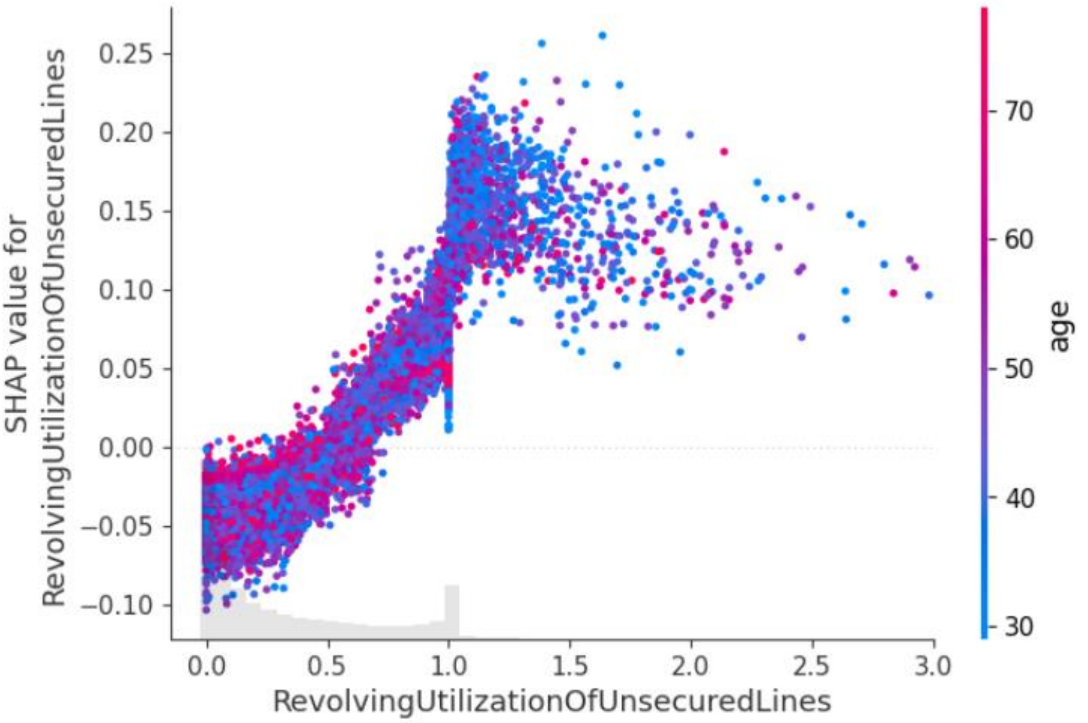


Figure 17. SHAP dependence plot for RevolvingUtilizationOfUnsecuredLines showing interaction with age.

Other SHAP plot called SHAP dependence plot can be used to visualize these interactions between variables. These plots display how the feature's attributed importance changes based on its value. They can show interactions between variables since each dot representing value of the variable is coloured by the value of interacting variable. If there are clear dependencies between variables these dependence plots would reveal these. Unfortunately for the credit risk dataset used for this thesis, there are no clear interactions between variables that could be easily visually presented. For example, in the figure 17 the interaction between

RevolvingUtilizationOfUnsecuredLines and Age is displayed. As can be seen, there is no clear indication that the impact of RevolvingUtilizationOfUnsecuredLines depends on Age, at least not linearly. This is not a problem for interpretability, but clear dependency would have helped to present the benefits of SHAP dependence plot more clearly. For interested readers, more intuitive example with clear interactions between variables can be found for example in Lundberg et al. (2019).

Next the focus is sifted to local explanations. The idea is to present the features importances for a single observation and therefore interpret the classification results for an individual borrower. The SHAP waterfall plots can be used for this task. These plots show the contribution of each feature value for the classification result (De Lange et al., 2022). In waterfall plot red colouring describes positive contribution, i.e. increase in the log-odds of serious delinquency. In contrast blue bars indicate that the feature value leads to decreased log-odds of serious delinquency. For example, in the figure 18. contribution for each feature value displayed for a randomly selected individual that does not have serious delinquency. The figure describes how the model reaches its prediction (f(x)) for a single observation. The most important feature for the prediction is displayed at top and the actual feature values are shown on the left.

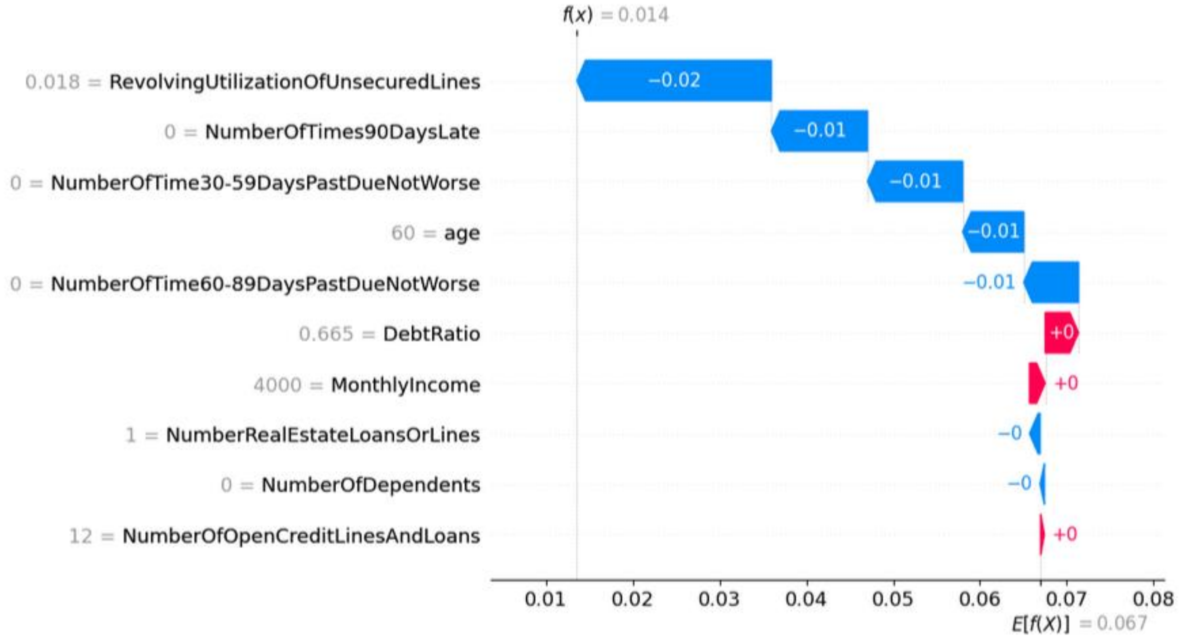


Figure 17. SHAP waterfall plot for single non-defaulted observation in the training dataset.

Similarly for a defaulted observation, local feature contribution can be shown using a waterfall plot. From the figure 18, we an see that the top 4 features show significant increases in the log-odds for serious delinquency. We can easily see that the most important factor why the individual was classified as delinquent was the number of time that the person had 60 to 90 days delinquent payments. The person has also had previous over 90 day delinquent payment which significantly increases the log-odds of being seriously delinquent in the future.

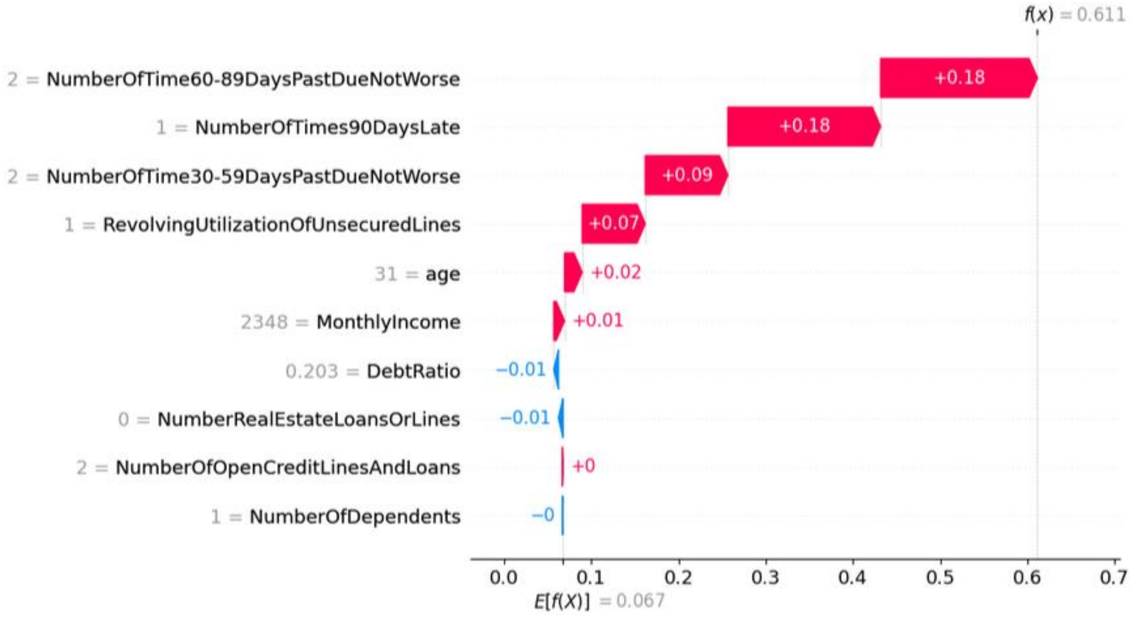


Figure 18. SHAP waterfall plot for single defaulted observation in the training dataset.

As we have shown in this section SHAP values and SHAP plots can be used to make the predictions of machine learning model significantly more interpretable. In the next section we will investigate the regulatory requirements and see whether this interpretability method is something that could be used in practise and would satisfy the regulatory requirements described in the section 2.1. It is important to note that one of the main limitations of this study is that we do not build a scoring card. For example, parameters acquired from the logistic regression are usually used to build a scoring card where each featured value is assigned a score using linearly transformed parameters from the LR. It should be possible to use SHAP values to build similar scoring cards. This is still relatively new subject and has not been extensively studied but for example, Hlongwane et al. (2024) show that it is possible to use SHAP for this purpose. However, since the score cards are beyond the scope of this thesis, further research would be required to confirm this approach.

6 Interpretability and Regulatory Requirements

Since the banking sector is a highly regulated field, there are multiple different regulatory requirements for credit scoring. The most important ones are described with details in section 2.1. In this section the idea is to look at these requirements and see whether the interpretable machine learning model built in this thesis would meet these requirements.

According to article 13 of the Regulation (EU) 2016/679 institution is required to provide meaningful information behind automated decision-making. Interpretable machine learning model that utilizes SHAP values meets this requirement since the logic behind the classification can be explained for each individual. These explanations would help a borrower to understand why they received a given risk classification and how their behaviour could improve their creditworthiness.

The guideline from the European Banking Authority (2017) states that an institution should carefully analyse and choose meaningful risk drivers for PD modelling. Based on the results shown in this thesis it is possible to argue that an interpretable ML model would be preferred over the logistic regression model since the model can built using more meaningful variables as there are fewer assumptions behind ML models. This is one of the major benefits of using ML methods over traditional statistical methods in addition with increased risk classification accuracies. The guideline does not mandate the usage of a specific model and therefore the usage of interpretable ML would be possible.

As mentioned in the section 2.1 one of the major drawbacks of using ML models in PD estimation is that the margin of conservatism (MoC) estimation becomes significantly harder. The EBA guideline requires institutions to cover the possible estimation errors using MoC, particularly in cases where model assumptions are not clearly specified. Since for the traditional statistical models this uncertainty in estimates can assessed using confidence intervals, applying MoC is relatively straight forward process. For ML models this process is more complex. SHAP values offers a partial solution by identifying feature contributions for each variable but does not completely solve the problem. Therefore, additional research is required to determine whether the statistical uncertainty in these interpretable ML models could be properly assessed.

The European Banking Authority (2017) guideline also mandates that the estimates of a credit risk model should remain stable over time and that all default observations are included in the model calibration. To meet the assumptions behind logistic regression model and to achieve adequate classification performance, some observation including default observation were omitted from the training dataset as described in the section 4.2. However, for the random forest model all the observations were usable. This would favour the usage of ML algorithms where there are less requirements for the training data. Stability of estimates over time could be ensured using different class imbalance handling techniques that were briefly mentioned in the section 2.4.

One of the main issues concerning the usage of ML models in credit risk is that since the usage of AI is becoming more popular regulation is still adapting to this situation. For example, the European parliament resonantly published new rules on artificial intelligence. In this regulation AI is classified according to its risk. Credit risk models are classified as high-risk AI systems since they may determine access to financial resources and essential services (housing). The regulation also mentions concerns over possible discrimination as these AI models (including ML) could reinforce biases related to age, gender or race. Therefore, these models will be regulated, and providers of these high-risk systems must ensure accuracy and robustness of these models. Additionally, providers must design their high-risk AI systems to allow deployers to implement human oversight. (EU, 2024/168)

7 Conclusions

In this thesis we compared the performance between logistic regression model and random forest model in default risk predictions. It was shown that the random forest model outperforms LR model when predicting for a serious delinquency in Give Me Some Credit dataset. The random forest was able to achieve the AUC score of 0.86 compared to 0.82 of logistic regression. The reason behind this improvement is that the ML models are able to capture complex, non-linear feature interactions unlike regression models that assume linear relationships between variables. It is also argued that based on other research the performance of ML models in general is usually superior compared to traditional statistical models when predicting credit risk. However, there are still some limitations that would require further research. Benefits of these interpretable machine learning models are clear on technical level but the economic benefits of using these models are not studied in this thesis. It is possible that institutions would still prefer the interpretability of less complex models if the benefits of using ML models are not significant enough on economical level.

This study demonstrated that the SHAP framework can be used to enhance the interpretability of random forest models to meet regulatory requirements. Models could be explained on both global and local levels meaning that the institutions using these interpretable ML models could provide their customers reasoning behind their credit decisions. As the regulation of AI models is currently changing fast due to rise of generative AI models, it is difficult to draw definite conclusion whether these ML models can be used in the field of credit risk in the future. In this thesis it is argued that these models fill the current regulatory requirements but since the regulation is in constant state of change it is difficult to say whether these models would meet future requirements.

Since the data availability of public credit risk datasets is limited, there are some issues concerning the data. These issues were mitigated using data preprocessing techniques including variable binning algorithm. It was shown that with adequate preprocessing techniques currently available public datasets can be used to measure performance of different models. However, for further research it is recommended that these comparisons are extended into different datasets to validate these results.

The future research should also extend the SHAP framework into other ML models such as gradient boosting algorithms and neural networks. Other interpretability methods like LIME and counterfactual explanations should also be studied since there are multiple alternatives for the SHAP framework. Additionally, these interpretable machine learning models could be used to create scoring cards to further improve interpretability of these models.

In conclusion, while interpretable machine learning models like random forest with SHAP framework show promise in improving credit risk predictions and fulfilling regulatory requirements, their adoption in banking sector highly depends on the future regulations and the development of these interpretable models. As the regulatory landscape continues to evolve, it is necessary that these models continue to evolve as well to remain viable options for credit risk management.

References

Alnemari, S. & Alshammari, M., 2023, Detecting Phishing Domains Using Machine Learning. Applied sciences. [Online] 13 (8), 4649-.

Alonso Robisco, A. & Carbó Martínez, J. M., 2022, Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial innovation (Heidelberg)*. [Online] 8 (1), 1–35.

Baker, M. and Wurgler, J., 2015, Do strict capital requirements raise the cost of capital? Bank regulation, capital structure, and the low-risk anomaly. American Economic Review, 105(5), pp.315-320.

Barboza, F. et al., 2017, Machine learning models and bankruptcy prediction. Expert systems with applications. [Online] 83405–417.

Bluhm, C., Overbeck, L., & Wagner, C, 2010, Introduction to Credit Risk Modeling (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781584889939

Bradley, A. P., 1997, The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, [Online] 30 (7), 1145–1159.

Breiman, L., 2001, Random forests. Machine Learning, 45, pp.5-32.

Büyükkarabacak, B. & Valev, N. T., 2010. The role of household and business credit in banking crises. Journal of Banking & Finance, [Online] 34 (6), 1247–1256.

Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J., 2021. Explainable machine learning in credit risk management. Computational Economics, 57(1), pp.203-216.

Bücker, M., Szepannek, G., Gosiewska, A. and Biecek, P., 2022. Transparency, auditability, and explainability of machine learning models in credit scoring. Journal of the Operational Research Society, 73(1), pp.70-90.

Chen, Y., Calabrese, R. and Martin-Barragan, B., 2024. Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 312(1), pp.357-372.

Classification: ROC and AUC, Google Developer, 2024, accessed 08 November 2024, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Costa e Silva, E. et al., 2020. A logistic regression model for consumer default risk. Journal of Applied Statistics, [Online] 47 (13–15), 2879–2894.

Cukierski, W., 2011, Credit Fusion, Give Me Some Credit, Kaggle, https://kaggle.com/competitions/GiveMeSomeCredit

Davis, J. & Goadrich, M. (2006) 'The relationship between Precision-Recall and ROC curves', in *ICML 2006 : proceedings, twenty-third International Conference on Machine Learning*. [Online]. 2006 New York, NY, USA: ACM. pp. 233–240.

Davis, R., Lo, A.W., Mishra, S., Nourian, A., Singh, M., Wu, N. and Zhang, R., 2023. Explainable Machine Learning Models of Consumer Credit Risk. *Journal of Financial Data Science*, *5*(4).

de Lange, P. E. et al., 2022. Explainable AI for credit assessment in banks. Journal of risk and financial management. [Online] 15 (12), 1–23.

Dumitrescu, E. et al., 2022. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, [Online] 297 (3), 1178–1192.

European Banking Authority, 2017. Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures. https://eba.europa.eu/regulation-and-policy/model-validation/guidelines-on-pd-lgd-estimation-and-treatment-of-defaulted-assets

Fitzpatrick, T. & Mues, C., 2016. An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. European Journal of Operational Research, [Online] 249 (2), 427–439.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189-1232.

Gramegna, A. and Giudici, P., 2021. SHAP and LIME: an evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence, 4, p.752558.
Hawkins, D. M., 2004. The Problem of Overfitting. Journal of Chemical Information and Computer Sciences, [Online] 44 (1), 1–12.

Hlongwane, R., Ramabao, K. and Mongwe, W., 2024. A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. Plos one, 19(8), p.e0308718.

Hoepner, A. G. F. et al., 2021. Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective. The European Journal of Finance, [Online] 27 (1–2), 1–7.

Household debt and credit report Q2 2024, Federal Reserve Bank of New York, 2024, accessed 06 November 2024, https://www.newyorkfed.org/microeconomics/hhdc

Khandani, A.E., Kim, A.J. and Lo, A.W., 2010. Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), pp.2767-2787.

Lee, S.I., Lee, H., Abbeel, P. and Ng, A.Y., 2006, July. Efficient l~ 1 regularized logistic regression. In Aaai (Vol. 6, pp. 401-408).

Lessmann, S. et al., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, [Online] 247 (1), 124–136.

Lin, X. et al., 2017. Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. Applied Economics, [Online] 49 (35), 3538–3545.

Lobo, J.M., Jiménez-Valverde, A. and Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography, 17(2), pp.145-151.

Loyola-Gonzalez, O., 2019. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. IEEE Access, [Online] 7154096–154113.

Lualdi, P. et al., 2022. Exploration-oriented sampling strategies for global surrogate modeling: A comparison between one-stage and adaptive methods. *Journal of computational science*. [Online] 60101603-.

Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. 2019. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv:1802.03888.

Lundberg, S., & Lee, S.-I., 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 30 (pp. 4765–4774). Curran Associates, Inc.

Mahesh, B., 2020. Machine learning algorithms-a review. International Journal of Science and Research (IJSR), [Internet], 9(1), pp.381-386.

Mandrekar, J.N., 2010. Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology, 5(9), pp.1315-1316.

Mironchyk, P. and Tchistiakov, V., 2017. Monotone optimal binning algorithm for credit risk modeling. Utr. Work. Pap.

Molnar, C., 2018. A guide for making black box models explainable. URL: https://christophm. github. io/interpretable-ml-book, 2(3), p.10.

Moscato, V., Picariello, A. and Sperlí, G., 2021. A benchmark of machine learning approaches for credit score prediction. Expert Systems with Applications, 165, p.113986.

Mullainathan, S. & Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. The Journal of economic perspectives. [Online] 31 (2), 87–106.
Myles, A. J. et al., 2004. An introduction to decision tree modeling. Journal of Chemometrics, [Online] 18 (6), 275–285.

Naili, M. and Lahrichi, Y., 2022. The determinants of banks' credit risk: Review of the literature and future research agenda. International Journal of Finance & Economics, 27(1), pp.334-360.

Organisation for Economic Co-operation and Development. Secretary-General. et al., 2011. Bank competition and financial stability. Paris: OECD.

Pedregosa et al., 2011, JMLR 12, pp. 2825-2830. Scikit-learn: Machine Learning in Python

Probst, P. et al., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, [Online] 9 (3), e1301-n/a.

Regulation (EU) 2016/679—General Data Protection Regulation (GDPR), EU, 2016, Official Journal of the European Union.

Regulation (EU) 2024/168 of the European Parliament and of the Council of 13 June 2024 harmonized rules on artificial intelligence and amending certain Union legislative acts. Official Journal of the European Union.

Saunders, A. & Allen, L. , 2010. Credit risk measurement in and out of the financial crisis: new approaches to value at risk and other paradigms. [Online].
Senaviratna, N.A.M.R. and A Cooray, T.M.J., 2019. Diagnosing multicollinearity of logistic regression model. Asian Journal of Probability and Statistics, 5(2), pp.1-9.

Shapley, L., 1953. A value for n-person games. Contributions to the Theory of Games, 28(2), 307–317.

Slovik, P. and Cournède, B., 2011. “Macroeconomic Impact of Basel III,” OECD Economics Department Working Papers, No. 844, OECD Publishing. http://dx.doi.org/10.1787/5kghwnhkkjs8-en

Tamayo, P. & Galindo, J., 2000. Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. Computational Economics, [Online] 15 (1), 107–143.

Thomas, L., Crook, J. and Edelman, D., 2017. Credit scoring and its applications. Society for industrial and Applied Mathematics.

What are internal models? ECB Banking Supervision, 2021, accessed 04 November 2024, https://www.bankingsupervision.europa.eu/about/ssmexplained/html/internal_models.en.html

Xiao, Y., Wei, Z. and Wang, Z., 2008. A limited memory BFGS-type method for large-scale unconstrained optimization. Computers & Mathematics with Applications, 56(4), pp.1001-1009.

Zhao, Q. & Hastie, T., 2021. Causal Interpretations of Black-Box Models. Journal of business & economic statistics. [Online] 39 (1), 272–281.