Developing an Early Warning System for Retail Customer Deterioration: A Data-Driven Approach

Jaap Beltman
MSc. Industrial Engineering & Management

UNIVERSITY OF TWENTE.

University of Twente
July 21, 2023

UNIVERSITY OF TWENTE.

This report is written as part of the educational program Industrial Engineering & Management. The assignment has been executed in the context of the specialization Financial Engineering & Management at ING Bank in Amsterdam.

Author

J.F. Beltman

Research institute: University of Twente
Faculty: Behavioural, Management and Social Sciences
Educational program: Industrial Engineering & Management
Specialization: Financial Engineering & Management

Supervisors University of Twente

Dr. J.R.O. Osterrieder (Jörg)
Dr. M.R. Machado (Marcos)

Supervisors ING Bank

Leon Dusée, MSc
Dr. Markus Haverkamp

Colloquium date: 28 July 2023
Pages: 91

Preface

First of all, I would like to thank my supervisors. This graduation journey started like a roller coaster and 11.000+ km away in Indonesia, where I followed courses at the time. Without the necessary preparations I had to start with my graduation assignment. Eventually, everything ended really well and I'm sure I couldn't have achieved this without all of my supervisors. Also, I want to thank the End-to-End team, where I performed my graduation assignment, for the good times. It was pleasant to work with all members of the team. Next to that, of course, I have to thank my family and friends for their support. I'm genuinely happy how these last months worked out and I'm keen on what lies ahead.

Abstract

This thesis addresses the existing gap in literature concerning the prevention of retail customers from becoming overdue. The literature gap in the area of retail EWS is the lack of a comprehensive approach that incorporates multiple data sources instead of relying on a single data source, along with the need for alternative methods, such as behavioral characteristics, for identifying deteriorating clients. The purpose of this thesis is to present the design and evaluation of an EWS for retail loans. Considering that financial institutions are often subject to regulations, this thesis avoids the use of algorithms that are considered poorly interpretable. The algorithms employed in this study include logistic regression, XGBoost, and Random Forest. Additionally, a fourth model, so called meta-model, is created by using the output of the other models, as input for a new model, which is trained by a logistic regression model. Regarding overall performance, the random forest algorithm achieves an average AUC of 0.775, while XGBoost and logistic regression achieve an AUC of 0.75. The meta-model outperforms the individual models with an AUC of 0.80. Compared to the random forest model, the XGBoost and logistic regression model did fit more on their top features. The meta-model consistently achieves the best fit when using the random forest predictions, as it is the top-performing individual algorithm. XGBoost serves as the second algorithm, while the logistic regression model provides the least significant contribution to the meta-model in terms of importance. The top-performing meta-model is utilized to gain practical insights into the timeliness of the EWS signals. The model reveals that a higher threshold for warning signals results in alerts closer to the overdue date, indicating increased sensitivity to emerging client deterioration. Conversely, lower thresholds focus more on the client's overall status. Furthermore, using the top ten features for training yields satisfactory overall results, but incorporating features beyond the top ten provides valuable supplementary information.

Contents

1 Introduction **7**
1.1 Research context . . . 7
1.1.1 Background . . . 7
1.1.2 Problem description . . . 8
1.2 Research methodology . . . 9
1.2.1 Research objective . . . 9
1.2.2 Research questions . . . 9
1.2.3 Research methodology . . . 10
1.3 Research design . . . 11
1.4 Scope . . . 13

2 Context analysis **14**
2.1 Credit risk . . . 14
2.2 Financial Early Warning Systems . . . 14
2.2.1 Purpose of EWS . . . 15
2.3 Deterioration detection approach . . . 17
2.3.1 Expert judgement . . . 17
2.3.2 Analogous approach . . . 18
2.3.3 Parametric approach . . . 19
2.3.4 Comparison of methods for EWS . . . 19
2.3.5 Conclusion . . . 20
2.4 Added value of data-driven EWS . . . 20
2.5 Design and implementation challenges . . . 22

3 Literature review **24**
3.1 Data quality and preprocessing . . . 24
3.1.1 Data cleaning . . . 24
3.1.2 Data reduction . . . 25
3.1.3 Imbalanced datasets . . . 25
3.2 Introduction to Machine Learning . . . 26
3.3 Types of ML algorithms . . . 27
3.3.1 ML types . . . 27
3.3.2 Application areas . . . 28
3.3.3 ML in EWS . . . 29
3.4 ML models . . . 29
3.4.1 Model interpretability and explainability . . . 29
3.4.2 Model description . . . 30
3.5 Model ensembling and optimization . . . 33
3.6 Model validation . . . 34
3.6.1 Performance evaluation metrics . . . 34
3.6.2 Cross validation . . . 36
3.6.3 Train-test split . . . 37

4 Model development **38**
4.1 Data pre-processing . . . 38
4.1.1 Data description . . . 38
4.1.2 Data analysis . . . 39
4.1.3 Data preparation . . . 40
4.1.4 Target variable . . . 42
4.2 Model training . . . 42
4.2.1 Trade-off decision . . . 42
4.2.2 Methodology . . . 43

5 Results & Discussion **45**
5.1 Model fitting description . . . 45
5.1.1 Random Forest . . . 45
5.1.2 Logistic Regression . . . 46
5.1.3 XGBoost . . . 47
5.1.4 Meta-model . . . 49
5.2 Model validation . . . 50
5.2.1 Predictive power . . . 50
5.2.2 Warning signals . . . 51
5.2.3 Feature importance . . . 52
5.2.4 Random dataset validation . . . 52
5.2.5 Cross-validation . . . 53
5.3 Discussion . . . 54

6 Conclusion **56**
6.1 Final considerations . . . 56
6.2 Limitations & further research . . . 57

References **60**

Appendix **65**

A Description datasets **65**

B Cleaning datasets **71**

C Data reduction **82**

D Parameter tuning **85**

E Performance metrics **86**

List of Figures

1 DSRM process model by Peffers et al. (2007) . . . 11
2 Research design . . . 12
3 Lending stages . . . 16
4 Deterioration detection framework . . . 17
5 From Lawson et al. (2021), Subsets of Artificial Intelligence . . . 27
6 From Jo (2021), Supervised learning . . . 28
7 From Jo (2021), Unsupervised learning . . . 28
8 From Nandi & Pal (2022), trade-off explainability - accuracy . . . 30
9 From Brett Lantz (2015), Entropy curve . . . 31
10 From Nandi & Pal (2022), Random forest . . . 32
11 From Larner (2022), Confusion Matrix . . . 34
12 From Trifonova et al. (2014), Area under the Curve (AUC) . . . 36
13 From Ren et al. (2019), K-fold cross validation . . . 36
14 Clients per lending category . . . 38
15 Stacking methodology . . . 43
16 Random Forest variable importance . . . 46
17 Logistic Regression variable importance . . . 47
18 XGBoost frequency importance . . . 48
19 XGBoost gain importance . . . 49
20 Meta-model variable importance (4 iterations) . . . 50
21 Histogram of the mortgage features . . . 75
22 Histogram of the current account features . . . 79
23 Histogram of the credit card features . . . 81
24 Correlation matrix of mortgage features . . . 82
25 Correlation matrix of current account features . . . 83
26 Correlation matrix of credit card features . . . 84
27 All feature performance metrics . . . 86
28 Meta-model 90th percentile signal distribution . . . 87
29 Meta-model 95th percentile signal distribution . . . 87
30 Meta-model 99th percentile signal distribution . . . 87
31 Meta-model 90th percentile last signal . . . 88
32 Meta-model 95th percentile last signal . . . 88
33 Meta-model 99th percentile last signal . . . 88

List of Tables

1 Benefits and drawbacks of expert judgement 18
2 Benefits and drawbacks of the analogous approach 18
3 Benefits and drawbacks of the parametric approach 19
4 Summary of several data-driven EWS approaches 23
5 Overview of binary classification metrics 35
6 Merging data sets 41
7 AUC with top ten features 52
8 AUC with random dataset 53
9 Cross validation iterations 53
10 Input variables mortgage dataset 65
11 Input variables current account dataset 68
12 Input variables credit card dataset 70
13 Percentage zero's mortages 71
14 Percentage zero's current account 72
15 Percentage zero's credit card 73
16 Parameter tuning 85

Abbreviations

AI - Artificial Intelligence
AUC - Area Under the Curve
CNN - Convolutional Neural Networks
DSR - Differentiated Sampling Rates
DSRM - Design Science Research Methodology
DT - Decision Tree
DTE SBD - Decision Tree Ensemble based on SMOTE, Bagging and DSR
EWS - Early Warning System(s)
FDP - Financial Distress Prediction
IDE - Integrated Development Environment
IFRS9 - International Financial Reporting Standard 9
LIME - Local Interpretable Model-agnostic Explanations
ML - Machine Learning
ML LightGBM - Gradient Boosted Decision Trees with Light Gradient Boosting
NN - Neural Network
OVO SVM - One-Versus-One Support Vector Machine
PCA - Principal Component Analysis
P2P - Peer-to-Peer
ROC - Receiver Operating Characteristic
ROS - Random OverSampling
RUS - Random UnderSampling
SHAP - SHapely Additive Explanations
SMOTE - Synthetic Minority Over-sampling Technique
SVM - Support Vector Machine
XGBoost - eXtreme Gradient Boost

1 Introduction

1.1 Research context

1.1.1 Background

Prevention is often considered a better approach than cure in many contexts, including credit risk management. In the realm of credit risk, early detection of payment problems can make a significant impact, as the cure may come too late if symptoms are not identified in a timely manner. To address this, banks employ Early Warning Systems (EWS) within their credit risk departments to detect and prevent loan payment arrears. An efficient EWS can be the determining factor between successful loan restructuring, enabling timely actions to be taken to assist the client in returning to normalcy, and loan recovery, marking the client as in default and initiating debt collection procedures. In cases where the financial situation of the client cannot be restored, the credit agreement must be liquidated. This latter outcome is not ideal, as it can harm both the customer and the lender. Timing is crucial in this scenario, as prompt action can prevent a situation from deteriorating to the point of loan recovery. Therefore, the implementation of robust EWS is of utmost importance in the banking industry, as it allows for early identification and mitigation of potential credit risks, ensuring the financial stability of both the bank and its clients.

Despite the demonstrated effectiveness of EWS, the literature on these systems is limited. Existing literature on EWS primarily focuses on corporate loans and is limited on retail loans (Allen, DeLong, & Saunders, 2004). Although corporate loan defaults can significantly impact financial institutions, the volume of retail loans, including mortgages, credit card loans, personal loans, and current accounts, is much higher. While it is challenging to implement a real-time, non-lagging EWS for corporate loans (Boonman, Jacobs, Kuper, & Romero, 2017; Gunther & Moore, 2003), this is less of an issue for retail customers as their expenses, balances, and other financial information can be tracked any point in time. Currently, EWS rely heavily on historical data (Du, Liu, & Lu, 2021). For corporate EWS, quarterly or annual reports and financial statements are the primary data sources. Financial ratios, such as liquidity, solvability, and debt ratio, are examined to obtain an understanding of the financial well-being of a company. The provision of these reports yields significant information; however, if the information is received tardily, its worthiness may be questioned. The success of EWS depends on timely and accurate data, as any delay in data collection and dissemination can lead to missed opportunities for early intervention and risk mitigation. In the context of EWS, time is a critical factor, and any lag in the system can compromise its usefulness. Given the volume of retail loans, the possibility to track the real-time status of the retail borrower, and the wholesale banking focus in literature on EWS, there is a promising opportunity for the research on retail EWS.

For the retail side, Kocenda & Vojtek (2009) created an default prediction model to distinct low risk and high risk clients. Nevertheless, this research only uses socio-demographic variables, such as marital status and level of education. Kocenda & Vojtek (2009) admit that, within the retail domain, research needs to be conducted on models that solely focuses on behavioral characteristics (the behavior of the client). Leow & Crook(2014) predicted changes of clients regarding their credit card loan payment status. Again, this study only

uses socio-demographic variables, such as age, job and address details. Also, Leow & Crook (2014) emphasize the importance of further research with a different method of predicting these transitions. Kwon & Park (2023) built an EWS that warns when household debt gets too high with macro-economic indicators, such as GDP, CPI and interest rates. Nevertheless, this EWS does not possess the ability to judge individual clients and thus is not suitable as retail EWS.

In the limited research conducted in the area of retail EWS, a comprehensive perspective that incorporates various input variables to identify deteriorating clients is noticeably lacking. The existing studies primarily rely on a single data source, without considering specific client behaviors, instead focusing on social backgrounds or macro-economic trends. The studies themselves acknowledge the need for further research to concentrate on client behavior and explore alternative methods for identifying deteriorating clients. It can even be questioned if socio-demographic variables are suitable variables for retail EWS, since variables, such as level of education barely or do not change and are primarily fixed indicators, which can not detect emerging risks. Besides, these studies possess elements that can be used for an EWS, but the purpose of the studies itself is not to create a retail EWS for individual client cases.

The use of real-time multivariate data can also make current EWS more dynamic. With the exponential growth of data availability and accessibility, it is imperative for EWS to adapt to a more data-driven approach. As per Cisco's report [1], all forms of data traffic are increasing over the coming years, making it imperative for EWS to be capable of adapting quickly to changing situations. Although it may be questionable how much of historical data is still relevant today, a significant amount of historical data is available for use in the EWS. Besides, the increase in data traffic might favor data-driven EWS in the coming years.

Furthermore, it is commonplace for a retail borrower to possess multiple types of lending products or financial instruments that offer insight into their lending behavior. For instance, a retail client with a low balance on their current account may indicate a potential risk in meeting their mortgage payments. This presents an opportunity to establish a comprehensive EWS that is fortified by diverse data sources. Also, the missed payments for one lending product can lead to arrears on other types of lending products. A comprehensive EWS can prevent the financial institution from further losses by creating a warning signal on the client.

1.1.2 Problem description

Current literature on EWS is primarily focused on corporate loans, neglecting other loan types, such as retail loans (Allen et al., 2004). However, retail loans, including mortgages, credit card loans, personal loans, and current accounts, constitute a significant volume of lending activity. Implementing real-time EWS for retail customers can be efficient, as their financial information can be tracked in real-time and warning signals can be send out quickly. Also, due to increasing data traffic a data-driven EWS might be favorable.

[1] https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

The existing literature in the area of retail EWS lacks a comprehensive approach where multiple data sources are used instead of one single data source. Also, existing literature reports that further research is needed about behavioral characteristics (Kočenda & Vojtek, 2009) and about alternative methods (Leow & Crook, 2014) for identifying deteriorating clients. Besides, the current literature is containing elements that can be used for a retail EWS, but literature lacks a study that is conducted with the purpose to develop a retail EWS.

1.2 Research methodology

1.2.1 Research objective

Following the shortcomings in literature, the aim of this thesis is to explore the feasibility and potential of developing a robust and efficient EWS specifically designed for retail loans. The objective of this thesis is therefore to use data from multiple sources. Especially, the EWS should more focus on the behaviour and the changing status of the client, rather than generic input variables, such as socio-demographic and macro-economic variables.

To achieve this, the research concentrates on collecting high-quality data from retail loan clients, such as expenses, balances, spending categories and income. The data is analyzed to understand the characteristics of the dataset and can be processed if necessary. The next step is to create a retail EWS model that is data-driven. Once the model is developed, it can be validated and its performance can be measured, with respect to timeliness of the warnings, specificity and sensitivity of the model. If the performance of the retail loan EWS is satisfactory, there is an opportunity to implement the model for daily use. However, implementing the model is not the objective of this research.

1.2.2 Research questions

Given the stated research objective, the primary research question of this study is as follows:

How does the integration of various lending products into a comprehensive Early Warning System for retail loans contribute to the early detection of client deterioration?

The central inquiry encompasses a broad scope, encompassing a managerial perspective. To effectively examine the path towards the development of a real-time and dynamic system, it is imperative to identify specific sub-questions that target distinct segments of the main question. Hence, the formulation of the following subsidiary questions will prove beneficial in advancing towards the overarching inquiry. Particularly, due to the lack of knowledge in the literature on certain topics, a comprehensive investigation is needed to lay a solid foundation for the development of a retail loan EWS. Each subsidiary question is accompanied by a concise description and a defined objective.

1. What are the current practices and methodologies that can be employed to early detect deterioration?

Initially, it is necessary to chart the prevailing circumstances. This step is of paramount importance as it serves as a reference point for future assessment of progress and helps to establish clear objectives for potential solutions. Furthermore, a comprehensive understanding of the conventional deterioration detection approach is necessary to prevent redundant efforts and facilitate the advancement of the latest developments in this field.

1a. How can the deterioration process be modelled?
This sub-question is focused on identifying the different states or conditions that a client can be in within the context of a deterioration process. Understanding the various states is essential for detecting deterioration at the earliest possible stage. It is crucial to have a clear mapping of the deterioration process, and the identification of states can help with developing a model to predict and track the deterioration process. By defining the different states a client can be in, the deterioration process can be more accurately modeled and monitored, allowing for early detection of deterioration and timely intervention to prevent further damage.

1b. How can a framework be created to categorize methods involved in deterioration detection?
To enhance current practices, it is necessary to construct a theoretical framework comprising various approaches. While each category has its own set of advantages and disadvantages, the creation of a theoretical framework enables a deliberate selection of the appropriate approach to be employed.

2. How can client data be modelled to create a retail EWS?
After mapping the deterioration process and determining the appropriate approach for retail client EWS, the subsequent step is to model the data into a retail EWS. The theoretical frameworks in sub-question 1b outlines a suitable approach to categorize clients into stages. The purpose of the EWS is to detect clients that are transferring to deteriorating stages early. The objective is to create a retail EWS that can be validated and is suitable for implementation.

3. How can the retail client EWS be validated?
To evaluate the performance of the model, the last step is to use suitable validation metrics to measure this performance. The retail EWS should be validated both from a theoretical side, as well as practical side. This means that, besides the validation metrics, also the model should be assessed on what this implies for the warning signals provided, to gain a practical perspective.

1.2.3 Research methodology

In order to establish a well-defined and structured approach throughout the research process, it is important to incorporate a comprehensive research methodology. The choice of methodology is relevant, as it should align with the purpose and goals of the study. Any discordance between the methodology and the objectives of the research may lead to confusion and hinder the efficacy of the study. Hence, it is crucial to have a clear understanding of the research objectives in order to select an appropriate methodology. The

research objective, as outlined in Section 1.2.1, is to develop a retail client EWS. Given that the ultimate aim is to create a tangible product, it is necessary to consider a methodology that places emphasis on the design phase of product development. Furthermore, the methodology should also provide space for conducting a thorough literature review, in order to address questions 1 and 2.

In view of the aforementioned criteria, the Design Science Research Methodology (DSRM) has been chosen as the methodology for this research. This method was developed by Peffers et al. (2007) and provides a scientific framework for product development, making it highly useful for this study. Additionally, this methodology allows for process iteration, which is particularly beneficial in the design phase.

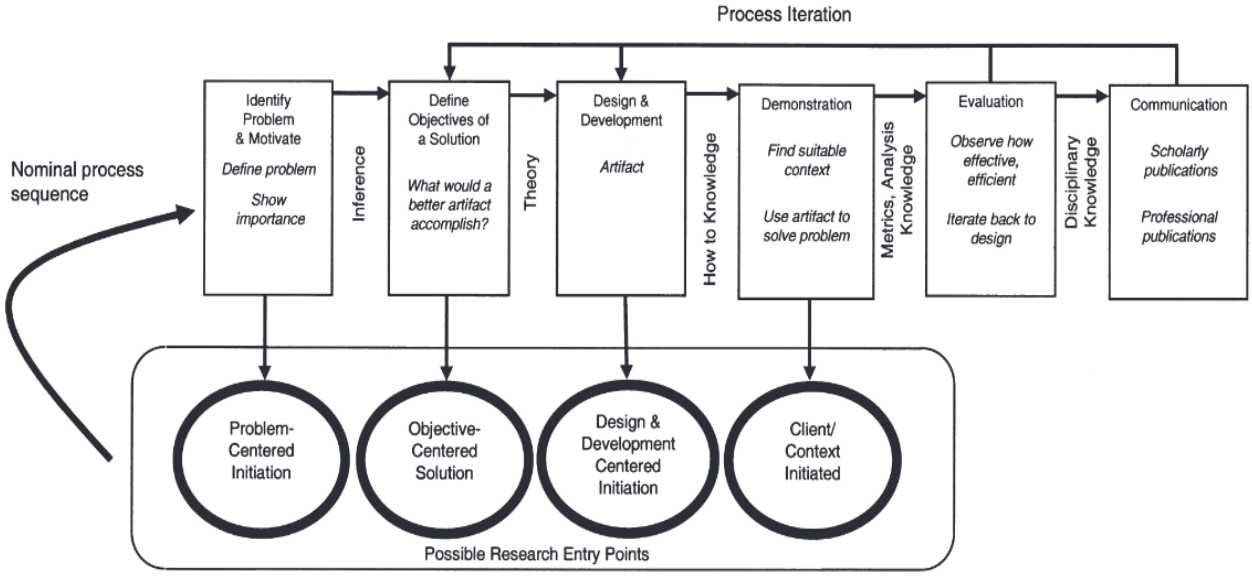


Figure 1: DSRM process model by Peffers et al. (2007)

Figure 1 illustrates the stages of the DSRM methodology, which consists of six distinct phases. The first stage involves defining the problem, which has been discussed in the early introduction. The subsequent stages include defining the solution objectives, design and development, demonstration, evaluation, and communication. As demonstrated in the figure, this methodology places a strong focus on the design aspect while also taking into account relevant literature.

1.3 Research design

Thus far, several topics have been mentioned, such as the research questions and research methodology. However, it is preferable to align these different topics in order to get a coherent overview of activities that connect to both the research question and the research methodology. The research design will merge the research methodology, research question and activities into. Figure 2 shows the overview of the sequential steps taken in this thesis. For each stage of the DSRM activities are connected to research questions.

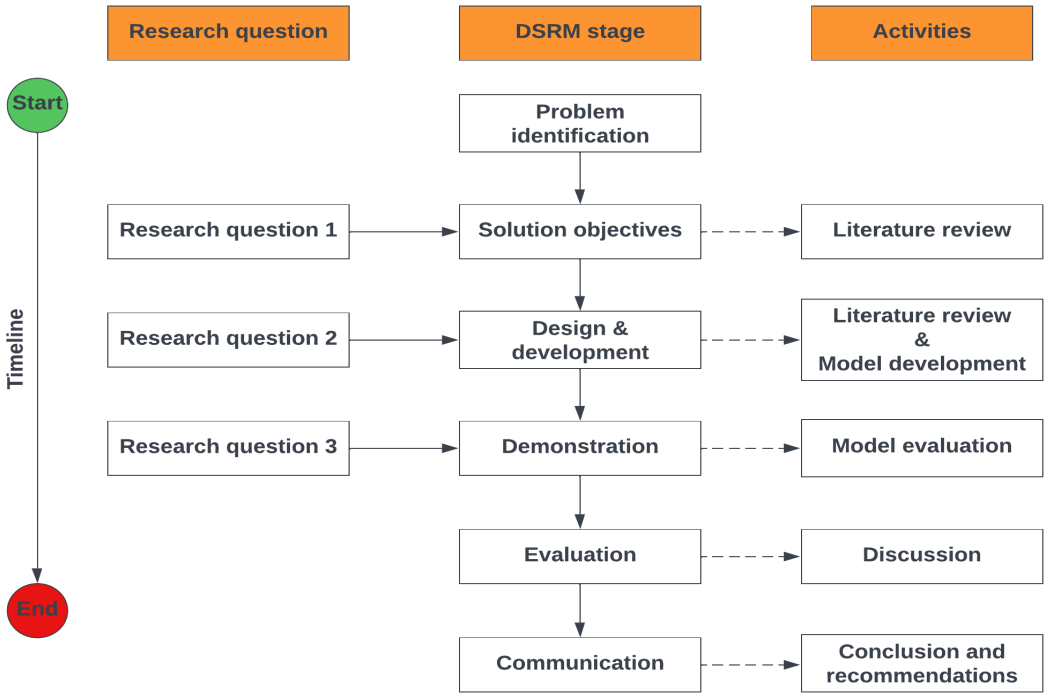


Figure 2: Research design

The DSRM stages summarized:

- **Problem identification**: The first step in DSRM entails defining the specific research problem and justifying the value of a solution. The present phase can be regarded as the opening stage of the thesis, and it is expounded upon in the introduction section.
- **Solution objectives**: The phase of solution objectives is essential for establishing a comprehensive understanding of the state-of-the-art EWS. The mapping of a norm or the framework for a norm serves as a foundation to assess the impact of a retail client EWS in deterioration detection. Moreover, it is crucial to keep pace with the latest developments in the field. In this stage both theoretical and practical perspectives are considered, thereby bridging the gap between theory and practice.
- **Design & development**: The design and development segment of this study provides an overview of the development process, following the standard established in the previous section. Due to the scientific nature of Design Science Research Methodology DSRM, a comprehensive analysis of the existing literature can be conducted before embarking on the design phase. This section concentrates on exploring the methodologies for translating retail client data into an EWS model that is appropriate for retail clients. This activity entails conducting an extensive review of the literature and based on this framework to construct a retail EWS. The primary aim of this phase is to present a product that can be appraised to assess its applicability in resolving the research issue.
- **Demonstration**: The demonstration phase of DSRM tests and validates the model. The primary aim of this phase is to critically examine the demonstrated model. As

DSRM allows multiple iterations, the evaluation stage serves as an opportunity to quickly identify any areas for improvement and implement necessary changes to the model. This feedback loop facilitates the refinement of the model and ensures that the final product aligns with the research objectives.

- **Evaluation**: The evaluation is the stage where the results from the demonstration are interpreted. This stage gives the meaning behind the validation metrics and what this implies for the retail EWS.
- **Communication**: In the communication phase of the research, the study's conclusions are drawn, and practical and research recommendations are provided.

1.4 Scope

The research is considered exploratory due to the nature of the topic, and as such, it is important to establish the boundaries of the study. The primary focus is to develop an EWS that suits the purpose of retail clients. This thesis can be seen as the roadmap towards the development of a comprehensive retail client EWS with multiple lending products as input variables. Nevertheless, it is important to mention that this thesis will limit its activities to the development and validation of the results and therefore it is not created for implementation. For the implementation of the EWS model, the model has to be integrated in current systems and an user interface needs to be established in order to make the model workable. However, this implementation is beyond the scope of this research and therefore no practical efforts will be made to implement the model in any way.

Moreover, during this thesis choices have to be made regarding the model development, such as a threshold to determine the sensitivity of the model. Some choices can be made from reasonable perspective, however, other choices involve organisational decisions. The preference for a certain option can be a managerial decision. This thesis will outline the options, but not make a decision as the decision might pose unfavorable consequences for the organisation.

2 Context analysis

2.1 Credit risk

The credit risk department of a financial institution bears the responsibility of managing the potential financial losses that may arise due to borrowers' inability to repay their loans or fulfill their financial obligations. The importance of credit risk management cannot be overstated as the absence of proper credit risk assessment may lead to a loss of market share and adversely affect financial stability (Markov, Seleznyova, & Lapshin, 2022). Effective credit risk management plays a critical role in ensuring financial stability, and credit risk is subject to external evaluation by central banks and auditors, as noted by Markov et al. (2022). These evaluations are designed to ensure that banks are compliant with key regulatory requirements, such as Basel III and International Financial Reporting Standard 9 (IFRS9) (Markov et al., 2022; Padhan & Prabheesh, 2019).

In addition to the upfront checks performed during credit origination, the monitoring of credit risk is also a crucial aspect of credit risk management. This involves the ongoing evaluation of borrowers' creditworthiness and their ability to meet their financial obligations over time. Monitoring credit risk allows financial institutions to identify potential risks early on and take appropriate actions to mitigate them. After a borrower fails to make a payment on a loan, the recovery process starts with issuing repayment notices in steps - leading to litigation suit as the final step (Uddin, Akter, Mollah, & Al Mahi, 2022). From a logical standpoint, the current situation can be considered far from optimal, as it requires significant effort on the part of the financial institution and causes disruption to the client's activities.

Identifying non-performing or deteriorating loans at the earliest opportunity is of importance, considering the substantial losses inflicted on various market participants, including investors, financial institutions, shareholders, and the broader economy, due to loan defaults (Z. Zhang, Wu, Qu, & Chen, 2022). Accordingly, this thesis focuses specifically on the monitoring phase of the lending process.

2.2 Financial Early Warning Systems

The definition of an "Early Warning System" is broad and can vary depending on the context in which it is used. Therefore, it is important to specify the definition of EWS and the purpose for which it is being used. EWS are not limited to the financial sector and are applied in various fields (Klopotan, Zoroja, & Meško, 2018). Even within the financial sector, there are multiple purposes for which EWS can be implemented.

Samitas et al. (2020) use EWS to predict financial crises, whereas An et al. (2022) use EWS to indicate liquidity shocks. In financial literature, EWS are often referred to as Financial Distress Prediction (FDP) models that provide early warning signals (Z. Zhang et al., 2022; Bräuning, Malikkidou, Scalone, & Scricco, 2019). It is important to note that FDP and EWS are not exactly the same, as FDP models have a more narrow scope and focus on the stability of the institution to maintain the stability within the financial market, whereas EWS are a more generic instrument. Nevertheless, the overlap between EWS and FDP in terms of warning signals creates a reason to incorporate FDP literature

in this thesis.

As mentioned earlier, this thesis is written in the context of credit risk. Therefore, in this thesis, EWS is defined as a tool that detects the deterioration of the credit status of a client before an event happens and should be seen in the context of credit risk.

2.2.1 Purpose of EWS

Banking crises have become a frequent occurrence in recent decades, necessitating the development of EWS (Ionela, 2014; Percic, Apostoaie, & Cocriş, 2019). The development of such systems can be geared towards a range of objectives. While the existing literature predominantly emphasizes EWS applications in the context of banking crisis events, this thesis will concentrate exclusively on the credit risk aspect of EWS, as mentioned earlier.

Iustina (2012) classifies EWS into two categories. The former refers to EWS that are subordinate to micro surveillance activities, while the latter refers to EWS that provide support for macro-prudential supervision. Though EWS are not new within credit risk, they were not initially developed for detecting deterioration of clients as mentioned earlier. They were developed for macro-prudential supervision. However, nowadays, EWS are suitable for micro surveillance activities as well. Macro-prudential supervision focuses on the stability of the financial system as a whole and the identification of systemic risks. EWS that support macro-prudential supervision are designed to identify risks that may affect the entire financial system. On the other hand, micro-surveillance activities focus on the monitoring of individual institutions and the identification of potential risks to their financial health, such as non-performing loans. The purpose of the EWS in this thesis is therefore dedicated to the micro-surveillance activities, with a focus on providing early warning signals of potential credit default by individual borrowers or a group of borrowers.

Within literature the benefits of efficient EWS are indisputable. EWS give a chance to management to take advantage of opportunities to avoid or mitigate potential problems before they occur (Koyuncugil & Ozgulbas, 2012; Davis & Karim, 2008). However, there is a need to emphasize that this is the case for accurate EWS only, as there is a danger of false alarms leading to inappropriate policy action (Davis & Karim, 2008). Timely intervention actions based on early warning signals can prevent further deterioration of a financial institution (Bräuning et al., 2019). The actions that can be taken do depend on the nature of the loan as well as the circumstances. Within the context of wholesale banking, this early warning signals can cause commercial lenders to be restructured instead of recovered. This is both beneficial for the client, as well as the financial institution, since recovering the loan is a time-expensive and non-preferable job, which disrupts the daily activities of the client. Also, in smaller cases early warning signals can be of great value, as customers can be reminded and/or warned for the consequences in absence of their repayment. Nevertheless, the exact actions undertaken remain a managerial decision and are out of the scope of this thesis, but these examples highlight the benefits of early warning signals.

Loans are available in various types and sizes, including small personal loans and large commercial loans, each displaying distinct characteristics. In addition to personal loans,

other types of loans include credit card loans, overdrafts on current accounts, and mortgages. Despite their differences, all loan types share a common characteristic: the risk of non-repayment. Song et al. (2023) have classified loans into two categories, namely "good" or "default," based on the stages at which customers can be present. However, banks do not immediately classify customers who are past due as "default" customers (European Banking Authority, 2016). Moreover, it would be erroneous to consider a customer who is past due as a "good" customer. Consequently, there is a need to incorporate a third category, namely the "overdue" state, into Song et al.'s (2023) classification. In the "overdue" state, the customer is behind on payments and may either progress to the "default" state if they cannot fulfill their obligations or return to the "good" state if they can catch up on their payment schedule. Figure 3 points out the stages in which a customer can be present during the lending process.

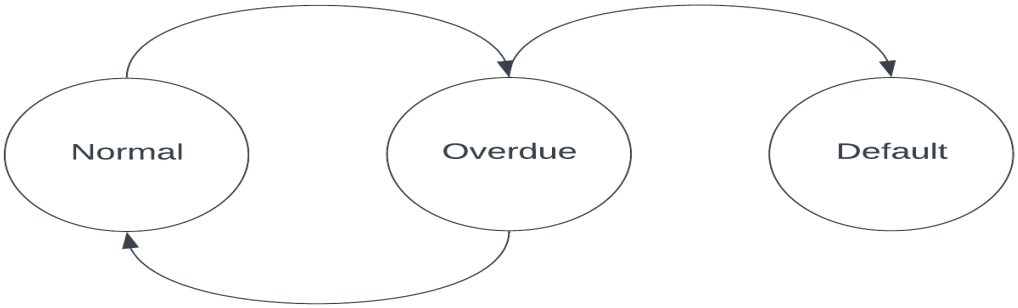


Figure 3: Lending stages

With an understanding of the stages of a lending customer, it is now possible to explain where EWS come into play. Figure 4 illustrates the typical progression of events for a defaulting customer. It is important to note that while all defaulting customers become overdue, not all overdue customers become defaulters. Therefore, this figure does not account for all possible scenarios. Compared to the normal and default states, the overdue state is temporary, and clients cannot remain in this state for longer than a predetermined number of days. However, the maximum number of days in the overdue state is not fixed, as it depends on the type of loan (European Banking Authority, 2016). Eventually, if a lender reaches the maximum number of days, it is either classified as default or is set back to normal.

Recapitulating the definition of EWS in this thesis:*a tool that early detects deterioration of the credit status of a client before a event happened.* The definition highlights that the warning signal must be issued before the occurrence of any event, including becoming overdue. Therefore, the EWS operates during the period in which the client's credit status is still considered to be 'good'. In contrast, during the overdue and default stage, non-repayment events have already occurred, rendering any warning signals in this stage unsuitable for early detection.

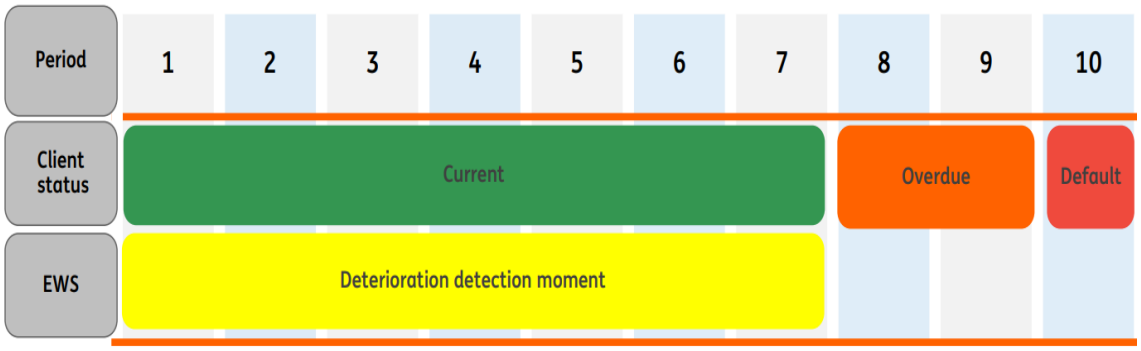


Figure 4: Deterioration detection framework

2.3 Deterioration detection approach

Even though, the approach may differ, early deterioration detection is not something new. It has always been in the interest of financial institutions to detect deterioration as early as possible. However, the approach used to accomplish this can differ per financial institution. The approaches that are used for deterioration detection should be feasible for the specific loan that is granted. As example, big loans are less frequent in the portfolio's of banks than creditcard loans. Nevertheless, the detail of information provided to the financial institution is much higher in wholesale banking (Allen et al., 2004). The deterioration detection method for wholesale banking can use financial statement analysis and ratio analysis. Compared to big commercial loans, for creditcards there is no annual financial statement present, the absence of this leads to the fact that the approach used for wholesale banking to detect early deterioration is not a feasible approach for smaller clients. Nevertheless, the volume of creditcard customers is much higher than big commercial loans. Therefore, the approach used to detect early deterioration is completely dependent on the specifications of the loan and the information present.

Given that there is not one way to detect deterioration, it urges the need to consult literature to systematic categorize several approaches. General types often used within literature for estimation include expert judgement, analogous approach and parametric approach (Budiono, Kiswanto, & Soemardi, 2014). The approaches are not strict and a combination of several approaches can be used, nonetheless it is important to understand these approaches in essence. The following subsections several approaches are analysed and their advantages and disadvantages are exhibited.

2.3.1 Expert judgement

Expert judgement is an approach that utilizes the experience of an expert to arrive at a decision (Shih & Mabon, 2021). It is presumed that the individual possesses sufficient knowledge and expertise in the field to enable sound judgement. In the context of credit risk, expert judgement can be applied to determine whether a loan is deteriorating and likely to default or become overdue. Expert judgement is a useful tool for making swift decisions since knowledge can be readily applied to the specific topic. Additionally, expert judgement can facilitate decision-making even when limited information is available (Hanea, Nane, Bedford, & French, 2021). However, there is a counterargument that

expert judgement may potentially be biased and overlook the full range of available evidence (Granger Morgan, 2013). Moreover, expert judgement can be subjective and lack transparency in opinions (Shih & Mabon, 2021). Furthermore, evaluating an entire loan portfolio can be a time-consuming process if an expert is required to make judgements on the status of each individual loan. Table 1 presents a summary of the aforementioned advantages and disadvantages of the expert judgement approach.

Benefits	**Drawbacks**
High level of knowledge	Potential for bias
Fast decision-making	Lack of transparency
More impactful with limited information	Time consuming on larger scale

Table 1: Benefits and drawbacks of expert judgement

2.3.2 Analogous approach

In literature, the analogous approach is often employed as an comparative method. This approach involves examining historical events to evaluate similar cases and assuming that the findings can be applied to the specific case at hand (Budiono et al., 2014). Regarding credit risk, this could entail considering a company to be deteriorating if its debt-ratio surpasses a certain threshold. Experts' opinions may also be included in the analogous approach, as there are no strict algorithms governing the process (Parviz, 2022). This results in the incorporation of benefits such as advanced knowledge and enhanced decision-making, while concurrently reducing biases and promoting transparency through the use of decision rules. This approach can be particularly useful when data is limited, as more generalized rules can be applied instead of a detailed approach, and it can also expedite decision-making through the use of specific decision rules (Parviz, 2022). However, the analogous approach is somewhat generic and lacks attention to detail, as it focuses on generic decision-making rules. When dealing with debt-ratios, it is important to take into account whether the company is a start-up or a corporation, as start-ups are likely to have higher debt. Furthermore, the initial investment of time required to develop the approach may not necessarily yield a return (Parviz, 2022). Finally, the analogous approach assumes that the circumstances of prior events are applicable to the present situation, which may not always be the case. Table 2 presents an overview of the advantages and disadvantages of the analogous approach.

Benefits	**Drawbacks**
Includes expert opinion	Generic decision rule
Useful with limited data	The use might not offset the development time
Works better on large scale than expert judgement	Assumes there are similarities

Table 2: Benefits and drawbacks of the analogous approach

2.3.3 Parametric approach

The parametric approach is the most data-driven of all methods, relying on quantitative analysis and statistical modeling to predict credit risk events based on historical data and established risk factors (Parviz, 2022). As this approach is grounded in statistics, it is more objective and transparent when compared to expert judgement and the analogous approach. Furthermore, the parametric approach can be more accurate since it makes decisions that fit the specific situation and do not rely on generic rules (Parviz, 2022). It can be applied on a large scale, such as to evaluate the risk associated with a financial institution's portfolio. However, the efficiency of this approach is dependent on the availability of sufficient data, and in the absence of such data, the models may not be useful (Parviz, 2022). Additionally, the parametric approach can only take into account well-understood risks and may not be suitable for evaluating new risks or changing conditions (Parviz, 2022). Furthermore, the development of the model can be extremely time-consuming. Finally, the parametric approach may not be as flexible as the expert judgement and analogous methods. Table 3 presents an overview of the benefits and drawbacks of the parametric approach.

Benefits	**Drawbacks**
Objective and transparent	Development is time consuming
High accuracy and tailoring to the situation	Dependent on the availability of data
Works better on large scale than expert judgement & analogous approach	Cannot deal with changing environment and/or unknown risks
	Less flexible than the expert judgement & analogous approach

Table 3: Benefits and drawbacks of the parametric approach

2.3.4 Comparison of methods for EWS

When considering various approaches, each has its own advantages and disadvantages. However, when developing a retail Early Warning System (EWS), certain approaches are better suited than others. It is important to acknowledge that the number of retail clients is extremely high in the context of an EWS. The parametric approach excels in handling this large volume by efficiently processing retail customers on a large scale. On the other hand, expert judgement becomes impractical at this scale as evaluating all clients within a reasonable timeframe is not feasible. The analogous estimation method, however, is less affected by this limitation as it compares clients and sets a threshold instead of individually assessing each client. Nevertheless, it assumes similarities among clients, which may not be applicable to retail clients without proper evidence. Additionally, while it performs well with limited data, the abundance of retail clients in the development of the retail EWS renders this limitation less significant. Despite the drawbacks of the parametric estimation method impacting the retail EWS, its benefits outweigh these limitations. Given the heavy regulation of financial institutions, objectivity and transparency are required. Furthermore, a tailored approach ensures that a generic decision rule does not apply to all clients, which is advantageous considering the high volume of retail clients.

2.3.5 Conclusion

As previous subsections have demonstrated, there is no one approach that can be applied universally, and the approach used should be tailored to the situation at hand. It is crucial to categorize these approaches fundamentally to comprehend the reason for selecting a specific approach. Additionally, the approaches are often contrary to each other, where the parametric approach is time-efficient once it has been developed, expert judgement is time-consuming for large-scale work. However, in a changing environment, the reliability of the parametric approach may be questionable. Nevertheless, for the purpose of a retail EWS the parametric approach is the most suitable due to its ability to process large volumes of clients, the transparency and its tailored approach.

2.4 Added value of data-driven EWS

To make the most of a data-driven EWS, it's important to consider relevant studies from existing research when making design decisions. In this section, we focus on the EWS designed for credit risk purposes. There is limited information on retail EWS, but that doesn't mean there is a lack of knowledge about EWS in general. To incorporate existing knowledge about EWS development and related fields, we review literature on other types of EWS, often within a corporate setting.

Fu et al. (2020) conducted a study that focused on the peer-to-peer (P2P) lending market in China, which has experienced rapid expansion in the past decade. Numerous P2P lending platforms failed to manage financial risks effectively, resulting in a significant number of defaults, losses for investors, and business closures. In an effort to enhance the process, Fu et al. (2020) employed artificial neural networks. They used deep learning models to mine indicative clues from highly colloquial comment texts about P2P lending platforms and predict whether or not a platform would produce a default event. The results of this study were promising for this field.

Sun et al. (2021) explore the use of a support vector machine (SVM) for multiclass financial distress prediction in companies. The SVM method is integrated with three decomposition and fusion methods to build three models, which are tested on four financial status categories: financial soundness, financial pseudosoundness, moderate financial distress, and serious financial distress. The study finds that One-Versus-One Support Vector Machine (OVO-SVM) outperforms the other two models and is preferred for multiclass FDP. The study also shows that when the class distributions are balanced in the training dataset, the performance of the models greatly improves. Overall, the SVM method is effective in predicting the financial status of companies, especially for categories that are difficult to predict using human expertise.

Sun et al. (2018) discuss the importance of enterprise credit evaluation for risk management, especially in times of financial crises. It proposes a new decision tree (DT) ensemble model, named Decision Tree Ensemble based on SMOTE, Bagging and DSR (DTE-SBD), for imbalanced enterprise credit evaluation. DTE-SBD uses a combination of the synthetic minority over-sampling technique (SMOTE) and the Bagging ensemble learning algorithm with differentiated sampling rates (DSR) to address the problem of class imbalance. The proposed model is compared to five other models in an empirical

experiment using financial data from 552 Chinese listed companies, and the results show that DTE-SBD outperforms the other models and is effective for imbalanced enterprise credit evaluation.

Barboza et al. (2017) examined the prediction of bankruptcy and default events in credit risk management. The authors compared traditional statistical methods and artificial intelligence models with machine learning techniques. Using data from 1985 to 2013 on North American firms, over 10,000 firm-year observations are analyzed. The study reveals that machine learning models, particularly when incorporating additional financial indicators, significantly enhance prediction accuracy compared to traditional models. New variables such as operating margin, change in return-on-equity, change in price-to-book, and growth measures are used as predictive variables. On average, machine learning models exhibit approximately 10% higher accuracy than traditional models. Among the tested models, the random forest technique achieves the highest accuracy of 87%, while logistic regression and linear discriminant analysis attain 69% and 50% accuracy, respectively. Bagging, boosting, and random forest models outperform other techniques, and the inclusion of additional variables consistently improves prediction accuracy. According to the author, their research contributes to the ongoing debate about the superiority of computational methods over statistical techniques in credit risk prediction.

Jones (2017) mentions that parametric models such as multiple discriminant analysis and logit were found to have constraints in handling predictors. To address this, the author turned to the gradient boosting model, a statistical learning method renowned for its capacity to handle a vast number of predictors. With a sample of 1115 US bankruptcy filings and 91 predictor variables, the study aimed to identify the predictors' strengths and overall performance in bankruptcy prediction. Surprisingly, non-traditional variables like ownership structure/concentration and CEO compensation emerged as powerful predictors, often overlooked in prior research. Unscaled market and accounting variables, serving as proxies for size effects, followed closely in predictive strength, while market-price measures and financial ratios played a moderate role. However, macro-economic variables, analyst recommendations/forecasts, and industry variables exhibited weaker overall predictive power. This research wants to highlight the limitations of traditional models and emphasize the significance of including non-traditional variables for more accurate bankruptcy predictions.

Hosaka (2019) conducted a study to apply convolutional neural networks (CNNs) in predicting corporate bankruptcy. Financial ratios were transformed into grayscale images for training and testing the CNN model. The dataset included financial statements from 102 Japanese bankrupt companies and 2062 currently listed Japense companies. Synthetic data points were generated to increase the dataset size. The CNN model based on GoogLeNet architecture was trained using 7520 images. The performance of the CNN model was compared to traditional methods such as decision trees, linear discriminant analysis, and support vector machines. The CNN-based bankruptcy predictions outperformed these conventional techniques. This research demonstrates the potential of CNNs in financial analysis and highlights their superiority in bankruptcy prediction.

Korol's (2013) research included data from 185 companies listed on the Warsaw Stock Ex-

change and 60 companies listed on stock exchanges in Mexico, Argentina, Peru, Brazil, and Chile. This author focused on forecasting the bankruptcy risk of enterprises in Latin America and Central Europe The companies were divided into learning and testing datasets, and financial ratios and their dynamics were analyzed. The author's developed models demonstrated high efficiency and were among the first to compare forecasting differences between the two regions. The research demonstrated the value of developing early warning models, as all models presented in the article exhibited high forecasting effectiveness ranging from 74.07% to as high as 96.66%.

Li et al. (2021) focus on the use of credit scoring tools to identify bad borrowers, particularly those who may be fraudsters on online lending platforms. The authors employ a multi-layer structured Gradient Boosted Decision Trees with Light Gradient Boosting Machines (ML-LightGBM) approach to capture early defaulted borrowers. Considering the imbalanced sample distribution and the costs associated with misclassification, they incorporate a cost-sensitive framework into the classification models' loss function to enhance predictive accuracy. The empirical results, based on a dataset of 1.6 million online loans, demonstrate that the proposed cost-sensitive ML-LightGBM algorithm outperforms other predictive models. This suggests that the cost-sensitive ML-LightGBM technique holds promise for fraud detection and credit scoring in this context.

In conclusion, there are multiple approaches to utilizing data for EWS. Table 4 provides a summary of the referenced papers. An intriguing observation from the literature is that defaults or deterioration's often receive inadequate representation, resulting in imbalanced datasets. Furthermore, the input variables used to detect deterioration yield interesting and diversified information. Research literature reveals that various sources, such as text mining/sentiment analysis, financial ratios, non-traditional indicators, and individual creditworthiness, are employed for this purpose. Moreover, this literature review highlights a predominant focus on defaults rather than overdue situations. Additionally, the majority of the literature concentrates on companies, with only a small portion addressing personal lending, specifically P2P default prediction. As far as my knowledge extends, a comprehensive EWS for retail clients has not yet been explored in the literature.

2.5 Design and implementation challenges

Previous section outlined the importance of data-driven EWS and the progress that has been made in recent years in developing more robust EWS. Hewamalage et al. (2023) clearly elaborates on the fact that machine learning has become more accessible due to the enormous growth of data, which imposes the quality challenges nowadays. Therefore, this section needs to identify known pitfalls or challenges within the development of EWS.

Data quality is key within this field and a well known understanding is that *"the model is as good as the data"*. Common problems are data bias, where the data is not representing the population. Furthermore, data incompleteness, data entry errors and data inconsistency pose also problems in the design of EWS. This raises the importance of the use of high quality data in combination with proper pre-processing.

Once the data is gathered and preprocessed, the model selection is important. The rise

Author	Year	Data-driven method	Input variables	Loan type
Fu et al.	2020	Artificial Neural Network	Text mining/ sentiment analysis	P2P lending
Sun et al.	2021	Support Vector Machine	Financial ratios	Chinese Company Lending
Sun et al.	2018	Decision Tree	Financial ratios	Chinese Company Lending
Barboza et al.	2017	Various Machine Learning Algorithms	Financial ratios	US Company lending
Jones	2017	Gradient Boosting	Financial Ratio's & non-traditional variables	US Company lending
Hosaka	2019	Convolutional Neural Networks	Financial ratios	Japanese Company Lending
Korol	2013	Discriminant analysis, decisional trees, artificial neural networks	Financial ratios	Latin/European Company Lending
Li et al.	2021	Light Gradient Boosting	Borrower's & financial status variables	P2P lending

Table 4: Summary of several data-driven EWS approaches

in number of machine learning algorithms is high, but not all algorithms are suitable for every purpose. Section 3 will explain this in more detail, but the choice of a non-suitable algorithm can decrease the results. Furthermore, within machine learning some models are known as 'blackbox' models. Petch et al. (2022) describes the term "black box" as models that are sufficiently complex that they are not straightforwardly interpretable to humans. Regulators have become aware of the ethical risk within artificial intelligence (Kaplan & Haenlein, 2020). Even though, EWS are not mandatory credit risk models, there is a need for explainability and interpretability to successfully implement an EWS.

3 Literature review

3.1 Data quality and preprocessing

The data quality and preprocessing are fundamental steps in ML practices. If the data is successfully preprocessed, the dataset can be considered a reliable and suitable source to apply ML algorithms (Luengo, García-Gil, Ramírez-Gallego, García, & Herrera, 2020). The first step is to ensure the acquisition of high-quality data. Preprocessing is akin to a cleaning process, but if the dataset is of low quality, preprocessing alone may not suffice. Benhar et al. (2020) differentiate between two processes in preprocessing: data cleaning and data reduction. This section addresses common issues in data cleaning and discusses techniques for data reduction.

3.1.1 Data cleaning

Data cleaning is the first step in the process after the raw data has been gathered. If the data is clean, this step can be seen as redundant, however, data is rarely clean (Luengo et al., 2020). Data cleaning is defined as the process of dealing with missing data, detecting and eliminating outliers,identifying and removing noise and correcting inconsistencies (Benhar et al., 2020). For each of these steps, fitting solutions will be proposed.

- **Missing data**: A missing value is a value that has not been stored or gathered due to a faulty sampling process, cost restrictions or limitations in the acquisition process (Luengo et al., 2020). Missing values cannot be ignored as they can cause problems and difficulties. Improperly handling missing values can lead to the extraction of poor knowledge and inaccurate conclusions (Wang & Wang, 2010).introduction section.

- **Outlier detection**: Outliers are extreme values that are far away from the bulk of the data points. They are often caused by measurement errors, exceptional true values, misreporting and sampling errors (Smiti, 2020). Outliers are identified by their distance from the center of the distribution or their deviation from the expected pattern of data. Outlying data that inhibits data analyt- ics should better be removed while those carrying important information should be kept (Smiti, 2020).

- **Noise elimination**: Noise in data refers to random or irrelevant variations that lack any meaningful pattern or structure. Inaccuracies in data formatting, such as using an incorrect data type (e.g., using an integer instead of a boolean), can introduce noise to the data (Smiti, 2020). To mitigate the effects of noise, Luengo et al. (2020) suggest the use of noise polishing techniques and filters in some instances.

- **Inconsistencies**: Inconsistencies in data refer to data points that deviate from a standardized format or convention. For instance, using "NL" and "The Netherlands" interchangeably may represent the same entity; however, such differences in notation may lead ML algorithms to perceive them as separate entities. As a result, identifying and resolving inconsistencies in data is crucial for ensuring the accuracy and effectiveness of ML algorithms.

3.1.2 Data reduction

Once the data is cleaned data reduction comes in to play. Types of data reduction are feature selection, feature extraction and transformation/discretization.

- **Feature selection**: Feature selection is the process of identifying a set of relevant features based on specific criteria, while discarding non-informative ones that are either redundant or irrelevant (Stańczyk, Zielosko, & Jain, 2018). Feature selection can eliminate irrelevant and redundant features that could create accidental correlations in learning algorithms. These correlations can ultimately diminish the generalization ability of the model (Luengo et al., 2020). Besides, feature selection can make the model run faster. Models and visualizations created using data with fewer features will be more straightforward to comprehend and interpret (Luengo et al., 2020).

- **Feature extraction**: Feature extraction techniques are an alternative approach that involves creating a new set of features by combining the original ones based on specific criteria (Luengo et al., 2020). A Principal Component Analysis (PCA) is a well known method. PCA is a linear feature extraction technique that aims to transform a dataset into a lower-dimensional space, while retaining as much of the original information as possible.

- **Transformation**: To further prepare the data for analysis, standardization can be used to transform the values of a variable to have a mean of 0 and a standard deviation of 1. This is particularly useful when dealing with features that have vastly different scales. Normalization is another technique used in data preprocessing that scales the values of a variable to a specific range, typically between 0 and 1. Normalization is particularly useful for features that have a wide range of values and is often used as a substitute for standardization when dealing with features that have similar scales. Log transformation is yet another technique that is often used to transform skewed data by taking the logarithm of the values. This helps to reduce the impact of outliers and make the data more normally distributed, which can make it easier to analyze. Binning involves dividing a continuous variable into a set of discrete intervals or bins. This can be useful when dealing with numerical data that has a large range of values, making it easier to analyze and compare.

3.1.3 Imbalanced datasets

As revealed by the context analysis, credit risk often involves imbalanced datasets. In binary classification problems, imbalanced datasets refer to datasets that are predominantly dominated by one category. In credit risk management, this means that there is a large number of clients who do not default and only a smaller number who do. Consequently, the prediction models, particularly for the minority class, are inadequate (Fosić, Žagar, Grgić, & Križanović, 2023). However, this is not desirable, as the purpose of the model is to detect defaults early. Model developers often miss this due to the use of inappropriate validation metrics (Fosić et al., 2023). Metrics such as accuracy can show positive numbers when everything is predicted in favor of the dominant category, which means the model does not differentiate between the two categories, which of course causes the whole model to be useless. Therefore, it is necessary to employ methods that improve the imbalanced

dataset problem and use appropriate metrics. Sidumo et al. (2022) proposed several solutions to the imbalanced dataset problem, such as random over-sampling (ROS), random under-sampling (RUS) and synthetic minority over-sampling techniques (SMOTE).

- **Random over-sampling**: The ROS technique replicate samples from the minority class (Sidumo et al., 2022). This way, the minority class becomes more balanced with the majority class and the performance of the ML model improves. Compared to other methods, ROS is relatively simple, since it just replicates sames that are already present. Critics often point out that this method fails to bring in any novel insights to the data and has the potential to cause overfitting (Santos, Soares, Abreu, Araujo, & Santos, 2018).
- **Random under-sampling**: As per Sidumo et al. (2022), the technique of RUS involves removing instances from the majority class to balance the dataset, making it the antithesis of ROS which balances the data by adding instances. While RUS is an easily implemented method, it has a drawback - discarding samples from the majority class can result in the loss of important information, particularly when the data is already limited in quantity.
- **Synthetic minority over-sampling techniques**: SMOTE is a statistical technique that increases the number of minority samples in the dataset by generating new instances (Sidumo et al., 2022). SMOTE generates comparable instances to the existing minority points, resulting in broader and less distinct decision boundaries that boost the generalization ability of classifiers. This, in turn, enhances their performance (Santos et al., 2018). However, also SMOTE can induce overfitting in the ML model (Luengo et al., 2020).

3.2 Introduction to Machine Learning

Machine learning (ML) is the study of computer algorithms that seek to improve automatically through experience (Lawson et al., 2021). ML techniques are popular for prediction of bankruptcy and default risks in credit risk management, as Barboza et al. (2017) report that these techniques have a 10% higher accuracy than traditional models. ML can be used to detect patterns in data that can predict future values. These predictions are based on one or multiple input variables that could have a correlation. Therefore, a low correlation between variables lowers the predictive power of ML models. These ML based predictions can be used to support informed decision-making and in the context of credit risk to manage risks and actively mitigate them.

ML is a branch of Artificial Intelligence that is especially useful within banking and finance (Nazareth & Ramana Reddy, 2023). It is important to know where exactly ML positions within the field of AI. Lawson (2021) defines ML as a subset of AI, but also within ML several subsets are present. These subsets are neural networks and deep learning. It is known that methods such as deep learning outperform conventional ML models (Oliveira & Bollen, 2023). Nevertheless, neural networks and deep learning models are typical examples of black box models (Guidotti et al., 2018). As elaborated on in section 2.5, black box models fierce strong resistance and are not widely accepted due to their

lack of transparency. Since this thesis is limited to implementation, but still wants to make it possible to implement the model, it should clearly distinct black box models from explainable models. It is acknowledged that ML can learn by analyzing existing datasets and identifying patterns in data that are missed by humans (Liu, Fu, Yang, Xu, & Bauchy, 2019). However, the division made by (Lawson et al., 2021) shows that there should be great conscious for the way the model is trained, since the way that the model is trained eventually determines if a model is explainable to humans.

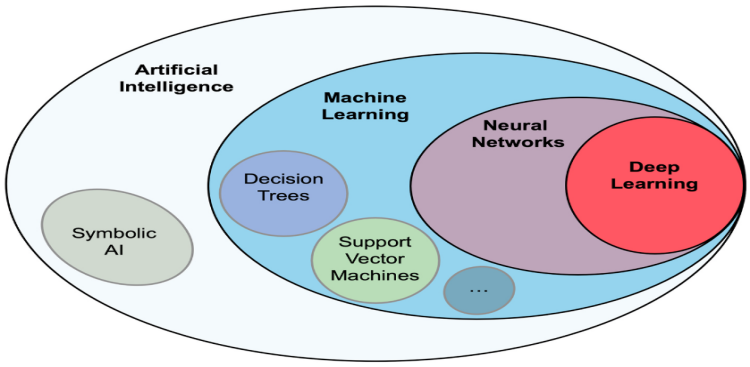


Figure 5: From Lawson et al. (2021), Subsets of Artificial Intelligence

3.3 Types of ML algorithms

In previous section already a great distinct in several types of models has been made. This section highlights the types of ML algorithms and the purpose of these algorithms. It is important to understand the different types of algorithms ML to eventually select the right type of ML to develop the EWS model. For this purpose, the book *'Machine learning foundations'* from Jo (2021) is consulted.

3.3.1 ML types

The first fundamental distinct within ML is made between supervised and unsupervised learning. This distinct can be seen as the main division in ML.

Supervised learning

Supervised ML is referred to as labeled ML. In this both the input and the output variables are provided. The goal of the algorithm is to learn the relationship between the input data and the correct output, so that it can make accurate predictions on new unseen input data. The supervised learning is the type which is mentioned the most frequent in literature (Jo, 2021). Figure 6 visualizes the core concept behind supervised learning.

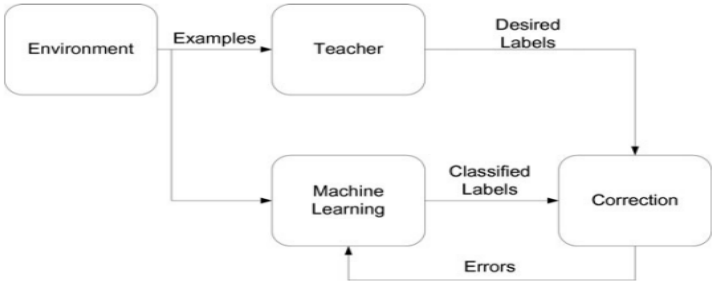


Figure 6: From Jo (2021), Supervised learning

Unsupervised learning

The second type of ML is unsupervised learning. Whereas, supervised learning is labelled, unsupervised learning is unlabelled. This means that there are input variables, but no output variables. The concept behind this is that within unsupervised learning the algorithm learns the similarities between the training examples and define the similar metric between them. Figure 7 expresses the idea behind unsupervised learning.

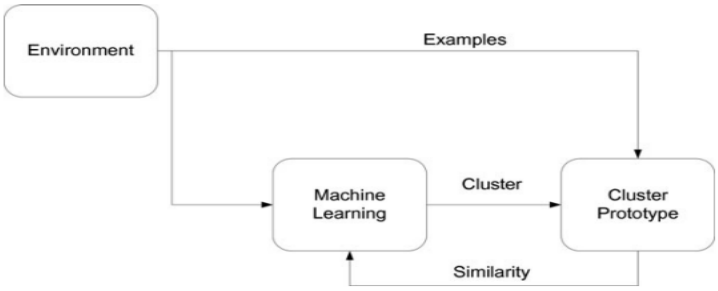


Figure 7: From Jo (2021), Unsupervised learning

3.3.2 Application areas

Previous section elaborated on the main distinction within ML is supervised and unsupervised learning. This section will translate these broad distinctions to specific areas in which the types can be applied. The four area's that are mentioned by Jo (2021) within the aforementioned literature are classification, regression, clustering and hybrid tasks. Following explanations are based on 'Machine Learning Foundations' from Jo (2021).

- **Classification**: Classification is a type of supervised ML which involves assigning input variables to a predefined category. Supervised learning produces discrete output variables, but the input variables may be continuous. Binary classification is the most widely used type of classification, where items are assigned to either a positive or negative category. Multiple classification, on the other hand, allows for more categories and is more specific in its classification of items.

- **Regression**: Regression, like classification, is a supervised ML method where the output variable is continuous. The key distinction between regression and classification lies in the nature of the output variable. Regression models produce continuous output variables, whereas classification models produce discrete output variables. There are several types of regression, including univariate regression, which generates a single output variable based on the input, and multivariate regression, which

generates multiple output variables. Time series regression, which predicts a sequence of values over time, is also a type of regression.

- **Clustering**: In contrast to classification and regression, clustering is a form of unsupervised ML. Clustering also has as output variable a discrete value. Clustering is in some ways comparable to classification. Based on the similar metrics the clustering algorithm tries to distinguish the item in different categories.
- **Hybrid tasks**: Hybrid ML tasks involve the combination of multiple ML techniques to achieve the desired outcome. This approach involves using a sequence of different algorithms to address the complexity of the problem. A typical example of hybrid ML is to first apply a classification algorithm to categorize data into predefined classes, and then apply a regression algorithm to predict the numerical value of an output variable within each category.

3.3.3 ML in EWS

It is evident that EWS is predominantly geared towards classification, as the model must provide an answer regarding whether a client's condition is deteriorating or not. This can be viewed as a binary classification scenario as described in the previous section. Clustering is not suitable for this purpose, as the algorithm needs to classify the client into a predefined category. The findings of this section have positive implications for the remaining sections, as they provide a clear indication that the focus should be on classification tasks. Therefore, the subsequent sections will explore the various techniques and algorithms related to classification algorithms.

3.4 ML models

3.4.1 Model interpretability and explainability

According to Adilkhanova et al. (2022), white-box models have been favored by researchers due to their simplicity and transparency, but black-box models typically offer higher prediction accuracy. This presents a trade-off between model performance and interpretability. The demand for explainability is not only coming from financial institutions, but also from other parties such as regulators (H. Zhang, Zhang, Zhang, & Zhu, 2023). One prevalent post-hoc method is feature relevance, which computes relevance scores for input variables and attributes the prediction to the set of features (Barredo Arrieta et al., 2020). Feature relevance can be divided in feature importance, feature correlation and feature interaction (H. Zhang et al., 2023). Methods such as Shapely additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) focus on feature importance.

The aim of this thesis is to achieve high model performance while maintaining sufficient explainability. Figure 8 from Nandi & Pal (2022) shows a spectrum of classification algorithms where interpretability is offset against accuracy. In all cases, higher accuracy leads to lower interpretability. Nandi & Pal report that Support Vector Machines (SVM) and Neural Networks (NN) are low in interpretability. Therefore, in this thesis, we set a cutoff at SVM. This does not imply that SVM and NN are completely unexplainable algorithms, but their high degree of complexity raises questions about their added value.

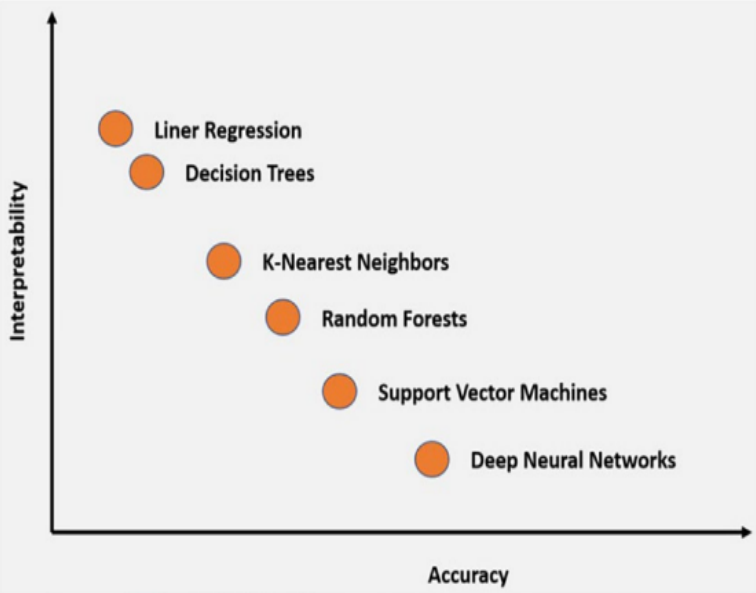


Figure 8: From Nandi & Pal (2022), trade-off explainability - accuracy

3.4.2 Model description

For the binary classification problem a couple of ML algorithms are useful. This thesis discusses decision trees, logistic regression, K-nearest neighbors and random forest as prediction algorithms for the development of a retail EWS.

Decision trees

Decision trees are possibly the most used method in ML (Brett Lantz, 2015). Decision tree learners use a tree structure to model the relationships between features and potential outcomes and act as a type of classifier, but can be used as regression algorithm as well. A tree with a wide trunk that splits into narrower branches as you move upward. In a decision tree classifier, branching decisions guide examples to a final predicted class value. Once the model is generated, several decision tree algorithms produce the resulting structure in a format that is legally compliant and easily understandable by humans (Brett Lantz, 2015).

In constructing a decision tree, the main challenge is to identify the best feature to split the data, with the goal of maximizing purity, where a subset containing only one class is considered pure. Entropy, a measure of randomness, is used to determine the best splits. The decision tree aims to minimize entropy, thereby increasing homogeneity within the groups. Entropy is measured in bits, with the minimum denoting complete homogeneity and the maximum representing maximum diversity, with no group holding even a small plurality. The notation for entropy can be found in equation 1. in the context of a binary classification problem, "Pi" represents the probabilities associated with the positive (i = 1) and negative (i = 2) outcomes.

$$Entropy = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{1}$$

As can be derived from the formula, a lower entropy is better because it indicates a higher degree of homogeneity within the resulting groups. As can be seen in Figure 9 and derived from the formula, in case of a binary classification a 50/50 split results in maximum entropy and the concerning indicator is not useful. On the other hand, a 100/0 or 0/100 split results in minimum entropy and the corresponding indicator is very useful. Other commonly used criteria are Gini index, Chi-Squared statistic, and gain ratio (Brett Lantz, 2015).

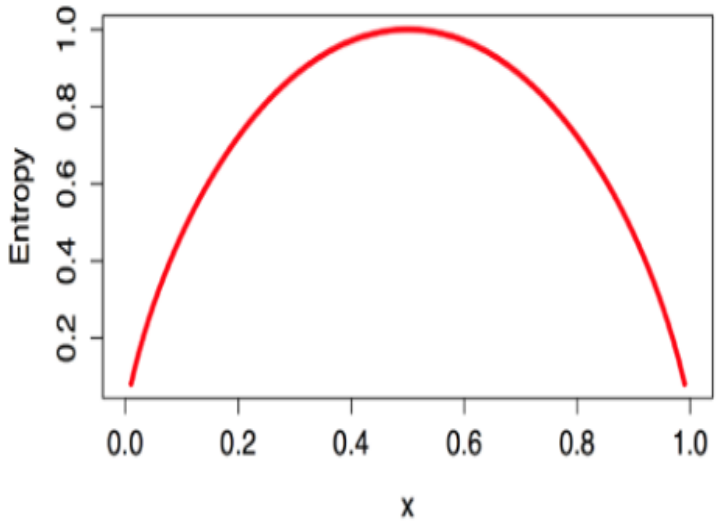


Figure 9: From Brett Lantz (2015), Entropy curve

Logistic regression

Logistic regression is a linear model that uses probabilities to predict a categorical response (Luengo et al., 2020). This method can be applied to both binary problems and multiclass problems. Logistic regression is used to predict the likelihood of an event occurring by fitting a sigmoidal curve to the data. The algorithm of logistic regression has a transparent characteristic (Barredo Arrieta et al., 2020). However, for multiple predictor variables the algorithm can become less intuitive (Nandi & Pal, 2022). The logistic function transforms the linear combination of input features into a probability value between 0 and 1. Based on the threshold applied the instance can be categorized in one of the two categories in case of binary classification. The formula is the following:

$$P(Y_i = 1|X_i) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_i X_i)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_i X_i)} \quad (2)$$

In this formula the beta can be explained as the weights. The weights are assigned in the learning phase. Whereas the X's are the input values for the particular instance. The output value is a number between 0 and 1 and as mentioned, based on the threshold, the value is categorized.

K-nearest neighbors

The K-nearest neighbors algorithm predicts the class of a test sample by taking a vote among its K nearest neighbors, where the neighborhood relationship is defined by a distance measure between samples (Barredo Arrieta et al., 2020). In terms of model explainability, it is crucial to note that the predictions generated by KNN models rely on the concept of distance and similarity between examples. This approach to prediction resembles the experience-based decision-making process of humans, which makes decisions based on past similar cases (Barredo Arrieta et al., 2020). The euclidean formula can be used for K-nearest neighbors, where Pi-Qi is the distance between two training instances:

$$Euclidean = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{3}$$

In the context of k-nearest neighbors, the Euclidean distance formula is used to determine the distance between an observation and all other observations in a dataset, in order to identify the k-nearest neighbors to the observation of interest. The k-nearest neighbors are then used to make a prediction or classification for the observation of interest, based on their similarity to the k-nearest neighbors. Also, K-nearest neighbor is a typical explainable classification algorithm (Barredo Arrieta et al., 2020).

Random forest

A Random Forest is an ensemble of decision trees, where each tree is trained independently using a randomly selected subset of the data (Luengo et al., 2020). At every node of the decision tree, a random subset of features is used to compute the output (Nandi & Pal, 2022). Nandi & Pal (2022) visualize the random forest as can be seen in Figure 10. A random forest trains multiple trees and, in case of the binary classification problem, with majority voting the decision trees decide in which category the sample belongs.

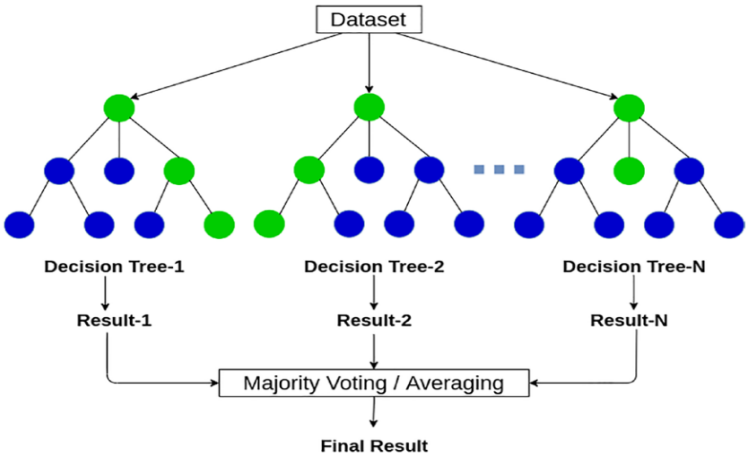


Figure 10: From Nandi & Pal (2022), Random forest

Due the fact that the random forest is constructed out of multiple decision trees the technique explanation is redundant. However, in literature it is generally acknowledged that random forests can improve decision trees (Nandi & Pal, 2022). Compared to random forests, decision trees are very sensitive to overfitting (Prajwala, 2015). Besides, random forests can effectively handle high-dimensional problems and accommodate non-linear relationships between predictors (Darst, Malecki, & Engelman, 2018).

XGBoost
eXtreme Gradient Boosting (XGBoost), an algorithm that works with trees, uses gradient boosting. During this process, each subsequent tree is constructed sequentially to minimize the error of the previous tree (W. Zhang et al., 2020). The trees in the sequence learn from their predecessors by updating the residual errors. As the boosting progresses, the subsequent trees learn from an updated version of the residuals. To expedite the learning process, distributed computing plays a vital role by parallelizing the computation across multiple machines or processors. This parallelization facilitates faster training of the ensemble of trees, resulting in accelerated learning (W. Zhang et al., 2020). Nevertheless, XGBoost is known for overfitting issues (W. Zhang et al., 2020).

3.5 Model ensembling and optimization

Model ensembling and optimization are two techniques employed in ML to enhance model accuracy and performance. Ensembling is a method that combines multiple models to make more accurate predictions. This typically involves training various models on the same data and then merging their predictions in some way. Optimization tunes hyperparameters to improve model performance. Ensemble learning techniques, including bagging, stacking, and boosting, are approaches that can be used (Khelfa, Ba, & Tordeux, 2023).

Bagging, an abbreviation for Bootstrap Aggregating, is a popular ensemble learning technique used in ML to improve the accuracy and stability of models. The basic idea behind bagging is to train multiple models on different subsets of the training data and combine their predictions through averaging or voting (Vynokurova & Peleshko, 2018). To implement bagging, we first divide the training data into several random subsets, with replacement. Then, we train a separate model on each subset using the same learning algorithm. During testing, the predictions of all models are combined in some way, such as taking the average or majority vote. This can help to reduce overfitting and improve the generalization of the model. As can be concluded, random forest already incorporates this in the algorithm itself, but it can be applied on other algorithms as well.

Stacking is an ensemble learning technique that trains multiple models and using their predictions as input features to train a meta-model that combines them (Vynokurova & Peleshko, 2018). The basic idea behind stacking is to learn a model that can effectively combine the strengths of individual models and improve the overall performance. This way the benefits that every ML algorithms individually has to offer, such as non-linearity detection, can be used to reinforce the outcome. Nevertheless, as might become clear, this ensembling method can be time-consuming.

Boosting is an ensemble learning technique used in ML to improve the accuracy of models by sequentially training weak models and combining them into a strong model (Forsyth, 2019). The concept behind boosting is to train a weak model on the original dataset. Then, we modify the dataset by increasing the weights of the misclassified examples and retrain the model on the modified dataset. This process is repeated several times, with each iteration focusing on the examples that are still misclassified. This can help to reduce

bias and improve the generalization of the model (Forsyth, 2019).

Lastly, a well known optimization technique is hyperparameter tuning. Hyperparameter tuning is the process of selecting the optimal values for the hyperparameters of a ML model, such as the number of branches in a decision tree . Hyperparameters are settings that are determined before training the model, and they can significantly impact the performance and generalization of the model. By iterating over a predefined set of values, the most optimal values can be extracted to use in the model.

3.6 Model validation

3.6.1 Performance evaluation metrics

To evaluate how a model performs it is necessary to establish performance evaluation metrics that can judge the model. As found, EWS models can be modelled with classification ML algorithms. This section will review the methods possible for to judge upon the performance of the model in the context of binary classification. In literature often used methods are sensitivity/recall/true positive rate, specificity/true negative rate, precision, accuracy, F-score and AUC (Area under Curve) (Kaur & Rattan, 2023; Valero-Carreras, Alcaraz, & Landete, 2023; Assy, Mostafa, El-khaleq, & Mashaly, 2023). Each of these methods will be described in this section and drawbacks of the several methods will be mentioned. The first four metrics (sensitivity,specificity, precision, accuracy, F-score) can be explained best in one Figure. Figure 11 highlights the confusion matrix from Larner (2022).

		True Status	
		Condition present	**Condition absent**
Test Outcome	**Positive**	True positive [TP]	False positive [FP]
	Negative	False negative [FN]	True negative [TN]

Figure 11: From Larner (2022), Confusion Matrix

Accuracy is the best evaluation metric to choose for balanced datasets (Assy et al., 2023). It is important to recap that the aforementioned EWS literature often dealt with inbalanced datasets and thus this is an important note. Precision is especially useful to see which proportion of the positive predicted instances was also truly positive (Valero-Carreras et al., 2023). In contrast, sensitivity measures the proportion of true positive instances identified by the model among all positive instances, whereas specificity measures the proportion of true negative instances identified by the model among all negative instances (Valero-Carreras et al., 2023). Based on this explanation, it can be argued that utilizing a combination of multiple evaluation metrics is advantageous, as a high score on

one metric does not necessarily indicate a well-performing model. The F1-score incorporates both precision and sensitivity, as evidenced by its formula. However, it does not reveal the relative contribution of precision and sensitivity to the overall score. Figure 5 provides an overview of the various metrics.

Metric	**Equation**	**Remarks**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Biased for unbalanced datasets
Precision	$\frac{TP}{TP+FP}$	Only accounts for the model's capability to detect relevant instances
Sensitivity	$\frac{TP}{TP+FN}$	Only accounts for the model's capability of positive predictions
Specificity	$\frac{TN}{TN+FP}$	Only accounts for the model's capability of negative predictions
F1-score	$\frac{Precision*Sensitivity}{Precision+Sensitivity}$	Does not show the relative contribution
AUC	$\int TPR\, d(FPR)$	Does not specify sensitivity & specificity at specific thresholds

Table 5: Overview of binary classification metrics

The last metric mentioned is the AUC. For a binary classification problem, the Receiver Operating Characteristic (ROC) curve is created by plotting the sensitivity, on the y-axis, against the specificity, on the x-axis, at various thresholds (Zhou, 2023). For each classification algorithm a separate ROC curve can be plotted. Verbakel et al. (2020) suggest that ROC is a suitable method to determine the risk threshold for the desired specificity and sensitivity level. For some instances it is crucial to detect every positive instance and therefore a low threshold should be applied, on the other hand, if intervention is expensive, the choice for a higher threshold might be more convenient (Verbakel et al., 2020). As the name suggests, the AUC equals the area below the ROC curve. The higher the value the better the quality of the model as the line converges to the upper left corner (Trifonova, Lokhov, & Archakov, 2014). If the AUC value is higher than 0.5, it indicates that the algorithm is better than pure chance in a binary setting (Zhou, 2023; Verbakel et al., 2020).

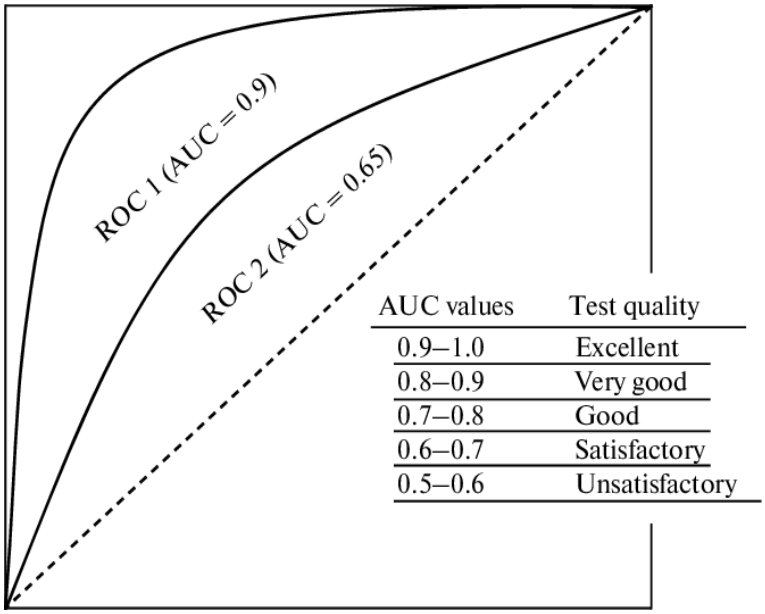


Figure 12: From Trifonova et al. (2014), Area under the Curve (AUC)

3.6.2 Cross validation

A well known validation method for ML within literature is K-fold cross validation (Tsamardinos, 2022). This method ensures that the model is trained multiple times with a different set of data. K-fold cross-validation equally divides the dataset into K parts, trains the classifier on K - 1 parts and evaluates the trained classifier on the left-out part (Xue, Dobbs, Bonvin, & Honavar, 2015). Figure 13, from (Ren, Li, & Han, 2019), visualizes the K-fold cross validation process. Tsamardinos (2022) mentions that this method can reduce the model's variance, but that this method is more urgent for smaller datasets.

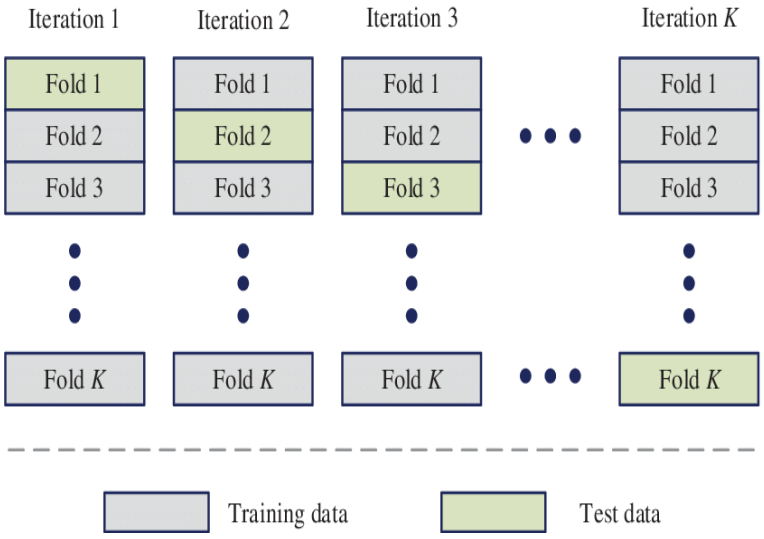


Figure 13: From Ren et al. (2019), K-fold cross validation

Even though, K-fold cross validation might be the most popular type of cross validation, there are more types. If the variance of the model is really important, then stratified k-fold cross validation reduces this further (Tsamardinos, Greasidou, & Borboudakis, 2018). The

data is partitioned into folds with an additional constraint that requires the distribution of the outcome in each fold to closely match the distribution of the outcome across all samples (Tsamardinos, 2022). Also, monte carlo cross validation, or shuffle-split, is a well known method in literature. Each iteration this randomly takes a part of the data as training data and the remaining data as test data (Xu & Liang, 2000).

3.6.3 Train-test split

Train-test split is a widely used method for evaluating the performance of ML models. This technique involves dividing the available data into two subsets - a training set and a testing set. The training set is used to train the model, while the testing set is used to evaluate its performance on unseen data. It is crucial to maintain a balance between the sample sizes of both the training and testing sets. However, it is practically impossible to have all samples present in both sets. This can result in the model learning the validation samples, making the validation process invalid. To overcome this challenge, a predefined ratio should be established for the train-test split.

Tran et al. (2022) opted for a 70-30% ratio for the train-test split, while Tsamardinos (2022) used a 90-10% ratio. There is no consensus on the ideal ratio for the train-test split in the literature. However, a study conducted by Veganzones & Séverin (2018) on bankruptcy prediction with imbalanced datasets found that an imbalanced distribution with the minority class representing only 20% significantly impacts prediction performance.

4 Model development

4.1 Data pre-processing

4.1.1 Data description

The goal of this thesis is to develop a comprehensive EWS model with multiple sources of data from retail client. Therefore, the data used in this thesis is originating from clients that posses either one or multiple lending products. The lending products in this thesis involve mortgages, credit cards and current accounts. A description from the raw input variables that are used in this thesis can be found in appendix A. For confidentiality reasons, the input variables will only receive a brief description and will be referred to as a number throughout this thesis. The data reaches from december 2015 until december 2019, due to the availability of data and the features are measured per month. Most customers have 49 months of data, since the the number of months between december 2015 and december 2019 equal 49. Not from all customers all 49 months are present for several reasons. Reasons include that the person became client, left as client or defaulted within the aforementioned time window. Within the dataset there are 589.700 unique clients. Figure 14 shows a venn diagram with the number of clients per lending category.

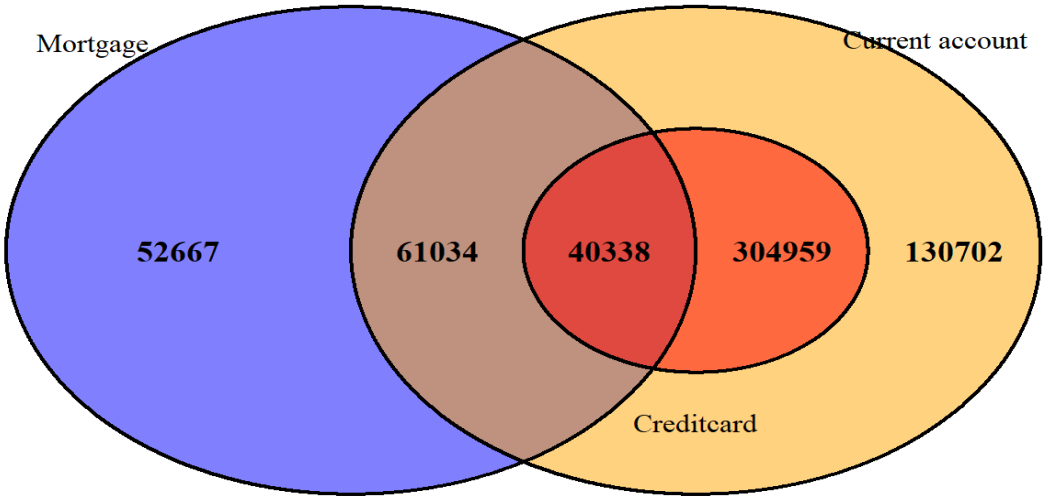


Figure 14: Clients per lending category

From this Figure it becomes clear that the vast majority of the clients has a current account. In total 537.033 clients have a current account, 154.039 clients have a mortgage and 345.297 have a credit card. It is not surprising that all clients who poses a credit card also have a current account, since the credit card is monthly deducted from the current account. Only a small proportion of the clients has all three lending products. Due to the sheer number of data points in these three datasets, which together represent almost three billion observations, computational issues arise. In combination with the scope of this

thesis to develop a comprehensive model it is justified to only include the 40.338 clients that have all three lending products. Therefore, client numbers that are not represented in all three datasets are dropped. This causes the number of data points to shift from almost three billion to approximately 240 million data points.

4.1.2 Data analysis

This section undertakes an analysis and cleans the data in order to have a high-quality dataset for the model. The process of data cleaning encompasses the identification and processing of missing values, detection of outliers, reduction of noise, and resolution of inconsistencies. For all three datasets this process is conducted.

- **Mortgage**
  In the dataset, missing values are represented by the value of zero. However, for some variables, zero is also a valid value. This presents a challenge as we want to treat missing values, but not valid values. To address this issue, Table 13 in Appendix B1 provides an overview of the percentage of zeros per feature. If a feature has more than one percent zero values, it is inspected to determine if the high percentage of zeros is significant. However, after inspection, it was found that none of the variables had a significant percentage of zeros. Since it is not possible to distinguish a missing value from a true zero, and the percentage of missing values in the remaining features is typically less than 0.1%, no further action will be taken.

  Furthermore, a statistical analysis is performed on the dataset and histogram plots are created, which can be found in Figure 21. Indicators used are the mean, standard deviation, maximum, minimum and range. Due to confidentially reasons, this statistical analysis is not published and axis's have been removed. From this analysis it turns out that feature 26 consists out of one value. Also, feature 1,6,18-25,27 and 28 are predominately existing out of one value, which makes distinction based on these features redundant. Furthermore, noise and inconsistencies are not found in this dataset.

- **Current account**
  Following the equal methodology, the current account dataset has 102 features, which is far more than the mortgage dataset. Table 14 shows that there are relatively much zero's, while the number of numerical variables is high, which are shown in Table 11. Logically, an integer is more frequent zero than a numeric variable. Also, this dataset is bothered with noise, since some data types are wrongly noted. From 80% the variable is dropped, due to the lack of information. Feature 16, 30-37, 49, 50, 57, 58, 64 and 77 are excluded from this research. From the plots in Figure 22, it turns out that feature 38 only has one value and therefore is also excluded. Also, feature 4,7,9,11,12,20-23,27-29,42,43,51,52,55,60,69,72-74,78,81-84,89-92 and 102 are again dominated by one value and are therefore excluded.

- **Credit card**
  The credit card dataset contains 59 features. Table 15 again shows the percentages zero's per feature. Besides the high number of features with a high level of zero's, this dataset contains features with a 100% zero score in it. After inspection it turns out that a vast amount of features is not useful due to their extreme percentage zero's.

Feature 11-17, 23-31, 39, 42, 44, 49, 52-56 is excluded. Based on the histograms shown in Figure 23, it can be observed that feature 57 contains only a single value. Consequently, feature 57 is excluded from further analysis. Also, 5,8,10,36-38,58,59 are merely consisting out of one value and are therefore deleted as well.

4.1.3 Data preparation

This section focuses on the process of data preparation for ML algorithms through data reduction. Data reduction plays a vital role in enhancing computational efficiency, improving model performance, mitigating overfitting risks, and increasing interpretability. Specifically, this section employs a correlation technique to achieve data reduction. Additionally, the datasets are merged, scaled, and subsequently split into distinct test and train sets.

Correlation

Correlation analysis is a valuable preprocessing step for ML. It helps identify the relationships between variables, enabling a better understanding of the dataset. This understanding aids in feature selection by identifying highly correlated features that can be potentially redundant or offer similar information. It also helps in identifying variables that have a strong association with the target variable, allowing for the prioritization of influential features. By creating a correlation matrix a quick overview can be provided to show which features are highly correlated among each other. Appendix C shows the correlation matrices in Figure 24, 26 and 25. Variables with a correlation of 1 are removed, since they have no distinctive characteristics and after scaling these variables would have the same values. In particular the mortgage and current account features showed variables duplicated features and consequently had the correlation value of 1.

Merging the datasets

The objective of this thesis, as stated in its scope, is to develop a comprehensive retail EWS. This involves merging diverse sources of data into a single dataset in order to predict whether a client will become overdue. Since the datasets only include clients with all three lending products, all clients are present in all three datasets. Thus, the primary key column used for merging the datasets is the client number. Additionally, the lending product statuses are recorded on a monthly basis, so the secondary key column for merging is the date (month/year). Unlike the primary key column, not all clients have lending records for every month. This is understandable because clients may already possess a credit card and a current account, but may not have a mortgage or apply for one at a particular time. Table 6 indicates the key columns for the merger of the datasets.

Client number	Month/ year	Feature 1	Feature 2	...
1	09/2020			
1	10/2020			
2	09/2020			
...	...			

Table 6: Merging data sets

Scaling

Scaling helps to balance feature magnitudes, ensuring fair consideration of each feature during the learning process. By facilitating faster and more efficient optimization, scaling contributes to quicker convergence of algorithms. It also improves model performance by preventing bias and inaccurate predictions, particularly in algorithms sensitive to feature scales. Overall, scaling data is promoting fairness, accuracy, efficiency, and convergence of models.

Min-max scaling, is one method for scaling numerical data in ML. It rescales the values of a feature to a specific range, typically between 0 and 1. The process involves subtracting the minimum value of the feature from each data point and then dividing it by the range (the difference between the maximum and minimum values). This transformation ensures that the minimum value of the feature is mapped to 0, and the maximum value is mapped to 1, while maintaining the relative relationships between the other values. Equation 4 presents the mathematical expression for min-max scaling. In this equation the maximum taken of the whole sample, max(x), and the minimum is taken, min(x). Then each individual data point (Xi) is scaled based on the formula.

$$Min - maxValue = \frac{x_i - min(x)}{max(x) - min(x)} \tag{4}$$

Besides, numerical data also a vast amount of data is categorical data. For this purpose label encoding is used. Label encoding is another technique used to represent categorical variables with numerical labels. Each category is assigned an unique numerical label, often starting from 0 or 1. Label encoding assumes an ordinal relationship between the categories.

Train test split

Train test split involves dividing the available data into two sets: the training set and the testing set. The training set is used to train the model, while the testing set is used to assess its performance on unseen data. This split helps in estimating how well the model generalizes to new, unseen examples and helps detect overfitting or underfitting issues.

Caution must be exercised when dealing with this dataset due to the presence of multiple months for each client. A standard split cannot be performed in such cases, as it would result in different months of the same client being present in the training and testing sets. Although customer identification numbers and dates are not provided, the model may indirectly infer associations between instances due to consistent parameters through-

out the dataset, such as credit card limits. This can lead to the model inadvertently learning from validation or test instances, rendering the validation results unreliable. To address this, the train-test split is conducted based on customer identification numbers rather than individual data points. The split ratio is set at 80:20, where 80% of the clients are allocated to the training set and 20% to the test set.

4.1.4 Target variable

In 100% of the cases that the credit card went overdue the current account also went overdue. This is explainable, since the credit card debt is deducted from the current account. In the mortgage dataset, in 98% of the time if the mortgage went overdue, also the current account went overdue. It is unknown why this was not the case in the remaining 2% and these overdue situation were removed from the dataset. The target variable for the development of the retail EWS is therefore the status on the current account.

The target variable for the ML model is the window 3 months before the client went overdue. This time is selected, since a shorter time frame might result in not enough time to detect the deterioration of the client. A longer time results in that the model is trained on non-deteriorating months and that the model is learning to detect financially healthy clients.

Given that clients can become overdue multiple times, it is important to ensure that the model does not have multiple opportunities to recognize an overdue client. This would introduce a bias during the validation stage if the model misses the initial overdue occurrence and only identifies the situation as overdue on the second instance, consequently labeling the client as overdue. Conversely, considering that only a small fraction of clients become overdue more than once, it is not worthwhile, in terms of effort, to examine each situation individually rather than focusing on each client. Consequently, only the initial overdue occurrence per client is utilized in the test set. It is worth noting that this is relevant solely for the test set since it is employed to validate the model. Conversely, the opposite holds true for the training set, where it is advantageous for the model to encounter all overdue situations per client, particularly considering the slight imbalance in this dataset.

4.2 Model training

4.2.1 Trade-off decision

After preprocessing, the dataset was reduced to a total of 1.8 million rows. Each row represents one month of one client, whereas most clients have 49 months in total and thus 49 rows. The cleaned dataset includes 40.338 clients, out of which 6.617 clients have been overdue at least once. Despite selecting specific variables, a dilemma arises. Providing a large amount of data to the model enhances its learning ability. However, if the model is overwhelmed with data and cannot effectively process samples, the abundance of data becomes useless. For instance, in a random forest algorithm, increasing the number of trees generally improves results but also increases running time. Also, if cross validation is applied then the models have to be trained multiple times.

To ensure a balanced approach, a total of 10.000 random clients were chosen for training and testing the model. Following an 80:20 train-test split, 8000 clients were used for training, while 2000 clients were reserved for validation. This selection methodology aims to provide an adequate number of client months for the retail EWS, while maximizing the algorithm's potential.

For the cross-validation metric, Monte Carlo cross-validation was chosen. This approach involves randomly partitioning the dataset into training and validation sets multiple times. Each partitioning is treated as an iteration, and this process is repeated for a specified number of iterations. The Monte Carlo method introduces variability in the training and validation sets across iterations. A total of 4 iterations were performed for each algorithm. This means that the algorithms underwent training 4 times, utilizing the 8000 clients for training and validating the results on the 2000 clients.

4.2.2 Methodology

This thesis incorporates three different algorithms: logistic regression, XGBoost and random forest. Additionally, a stacking meta-model will be constructed using a majority voting approach. The purpose of this meta-model is to predict whether a client will become overdue based on the individual predictions of the three aforementioned algorithms. The meta-model is a new ML algorithm that learns from the outputs of the three ML algorithms. This way it can apply weights to the different algorithms and use individual characteristics of the ML models to reinforce the models performance. As ML algorithm for the meta-model logistic regression is selected, since it can deal well with limited input variables. The stacking process is visualized in Figure 15.

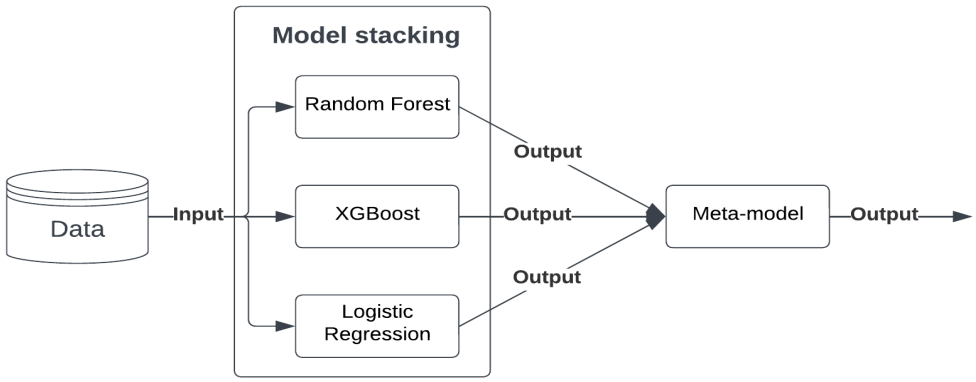


Figure 15: Stacking methodology

Hyperparameter tuning is a crucial step in ML model development. Hyperparameters are parameters that are not learned from the data but are set manually before training the model. These parameters control the behavior and performance of the model, making their selection a critical task. The process of hyperparameter tuning involves searching for the optimal combination of hyperparameter values that results in the best model performance. This was merely selected based on systematic experimentation and evaluation of different parameter settings. Figure 16 in appendix D, shows the list hyperparameters tested.

Moreover, this thesis includes an iteration specifically focused on analyzing the relevance of features beyond the top ten. By conducting this iteration on an earlier version that included all features, it is possible to directly compare the performance and assess the impact brought by the top ten features. This comparison allows to determine if the dataset can be effectively reduced to only these influential drivers while still achieving comparable results.

R was chosen as the programming language for the ML model in this study, primarily because of its exceptional capabilities in handling large dimensions and samples. The preprocessing tasks were efficiently performed using R, benefiting from its proficiency with extensive datasets and its suitability for statistical analysis. Moreover, R offers a comprehensive set of ML packages that encompass the specific algorithms employed in this thesis. To facilitate the development process, R-Studio was selected as the integrated development environment (IDE) for this project.

5 Results & Discussion

This section aims to provide a detailed exposition on the process of model fitting, focusing on individual algorithms employed. Additionally, the obtained model outcomes will be analyzed, encompassing both the model that incorporates all versions as well as the version incorporating only the most significant features. Finally, a comprehensive conclusion will be drawn regarding the findings.

5.1 Model fitting description

To comprehend the rationale behind the predictions made by the models, it is needed to elucidate the fitting process of each individual model. By doing so, we can facilitate a comparative analysis of the various models in subsequent sections and potentially provide explanations for discrepancies observed within the models.

Due to the importance of what the EWS functionality actually is, there is a need for a short recapitulation. The model is trained to recognize the three months before a client is going overdue. The three months window strikes a balance between including deteriorating months, but excluding still financially healthy months of the client. Nevertheless, as could be seen earlier in Figure 4, the deterioration detection window is in the time before the overdue situation occurs. Consequently, if a warning signal is provided somewhere in this time window and the client goes overdue eventually, the client is a True Positive. Vice versa, if even a single month of a non-overdue client is remarked as overdue, then the client is a False Positive. The results will also show the practical implications of the model by showing the timeliness of the warning signals.

5.1.1 Random Forest

In order to judge each feature, the variable importance is extracted. For each predictor variable, the Variable Importance function in R is used to perturb the values of that variable while leaving the other variables unchanged. The model's predictions are then re-evaluated using the permuted variable, and the decrease in performance metric is recorded. The outcome is a ranking of multiple variables, enabling the identification of important features based on their respective levels of importance. Figure 16 shows the sorted output of an iteration of a random forest model.

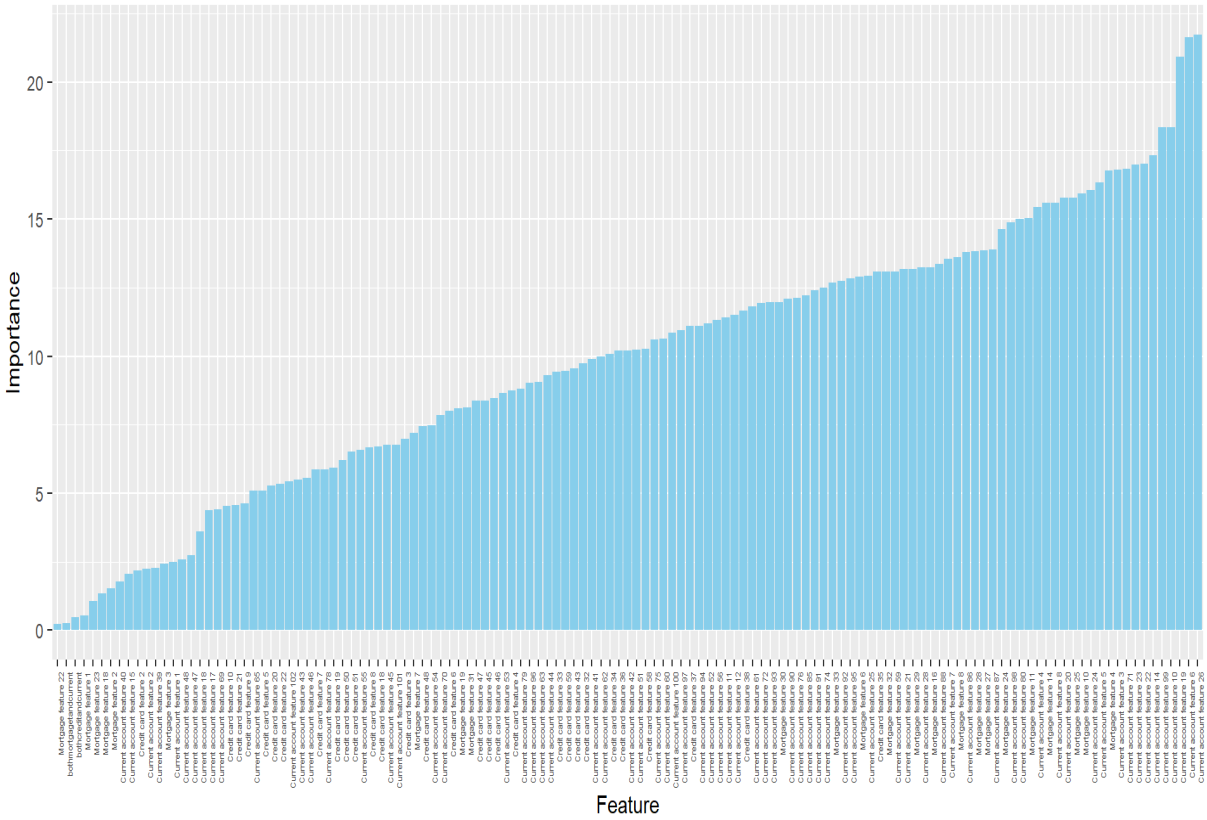


Figure 16: Random Forest variable importance

The plot reveals that while there are a few prominent features, the model does not rely solely on a small set of variables for its fitting. On the contrary, the model utilizes a substantial number of features to make predictions, indicating a diverse selection of variables. Conversely, the lower end of the graph highlights a few features that are almost negligible, as the model barely utilizes them in its prediction process. The plots of the other iterations exhibit a similar pattern, with minor variations observed primarily in the ranking of features rather than the levels of importance. The overall importance levels of the features remain consistent across these iterations. The analysis reveals that the top-ranking features primarily belong to the mortgage and current account datasets, indicating their significant influence on the model's predictions. However, features from the credit card dataset are noticeably sparse in the distribution and hold relatively less importance when they do appear. In comparison to the mortgage and current account features, the credit card features exhibit lower levels of importance in contributing to the model's predictive performance.

5.1.2 Logistic Regression

Following the same methodology as the Random forest fitting description, the most important features are exhibited in this section. In Figure 17 the feature importance of a logistic regression iteration can be found.

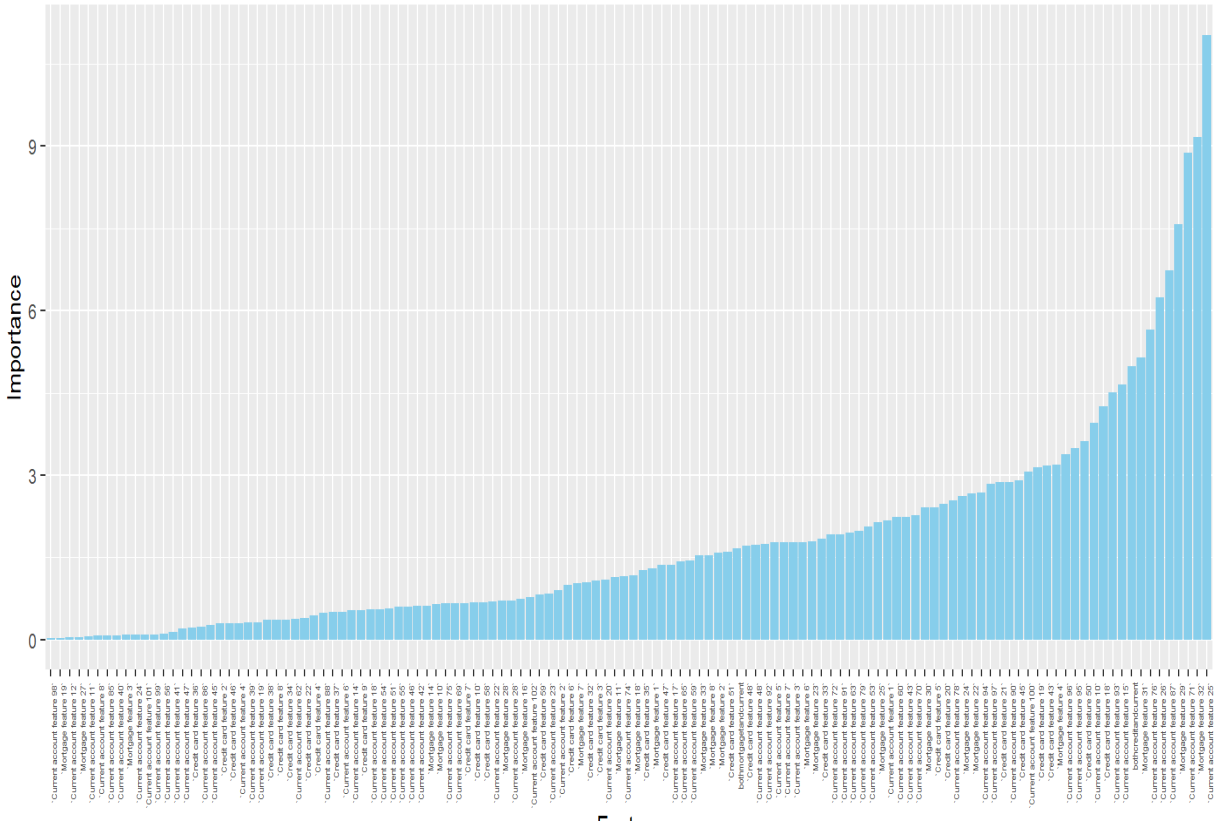


Figure 17: Logistic Regression variable importance

There are notable differences in the importance rankings between the logistic regression and Random Forest models. The distribution of the most important features in the logistic regression model exhibits a slightly more curved pattern, suggesting that it is fitted on a smaller subset of features compared to the Random Forest model. The logistic regression model relies heavily on the first 10-15 features, which are the primary drivers of its predictions.

Furthermore, the logistic regression model displays a larger number of (almost) redundant features compared to the Random Forest model. This is evident from the lower end of the distribution, where features demonstrate little to no significant importance in the logistic regression model. Similar to the Random Forest model, the credit card features in the logistic regression model exhibit reduced importance when compared to the mortgage and current account features. Other logistic regression iterations, followed a similar distribution.

5.1.3 XGBoost

In contrast to the methodologies used for Logistic Regression and Random Forest, the XGBoost algorithm is judged on a combination of gain and frequency. The gain metric measures the improvement in the model's loss function attributed to a specific feature when it is used for splitting in the trees. The frequency metric counts the number of times a feature is used for splitting across all the trees in the ensemble. Figure 18 shows the frequency and Figure 19 shows the gain.

The frequency plot reveals that a substantial number of features are employed in the model, indicating their usage during the modeling process. However, upon examining the gain plot, it becomes apparent that only a select few features significantly contribute to the model's improvement. While one feature appears to be dominant, there are other features that also play a role in enhancing the model's performance. Interestingly, the XGBoost algorithm primarily gains from a limited number of features, despite utilizing a broader set of features. This observation aligns with the findings from the logistic regression and random forest models, where the credit card dataset exhibits comparatively less relevance in influencing the model's predictions.

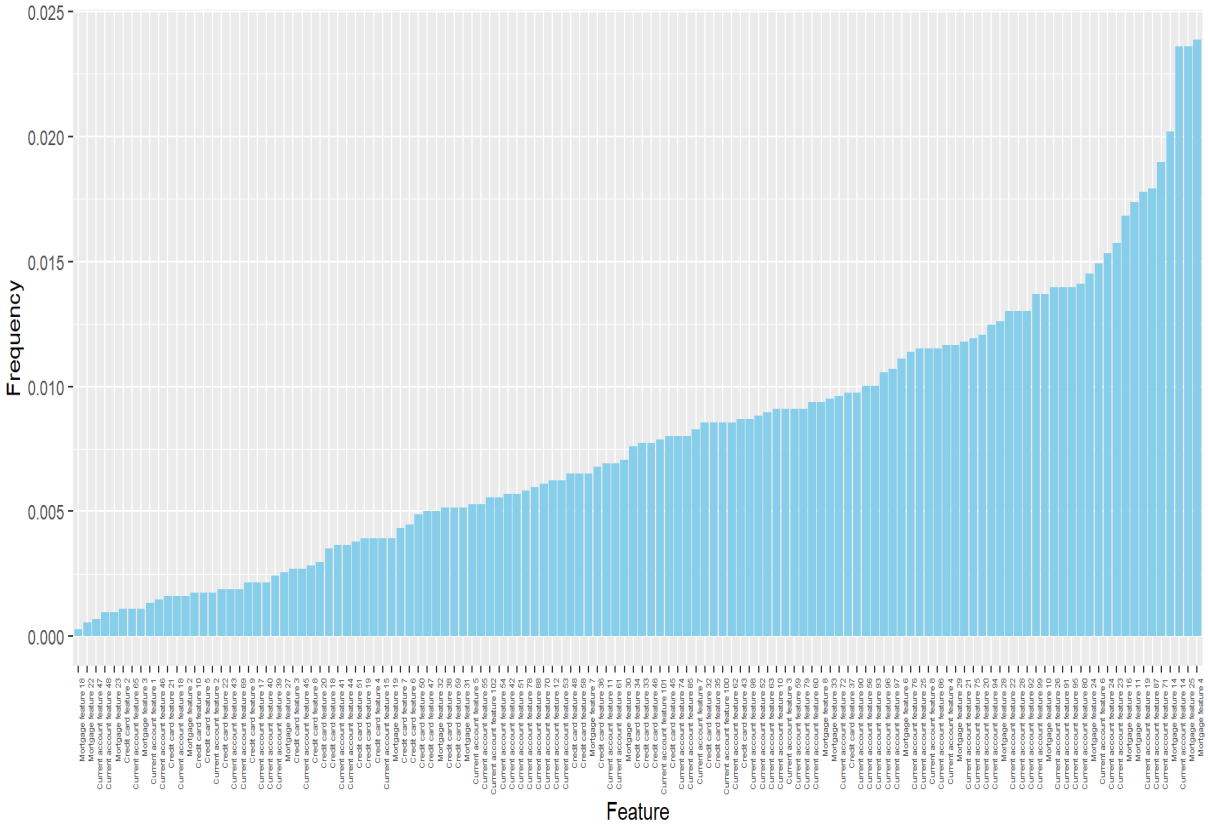


Figure 18: XGBoost frequency importance

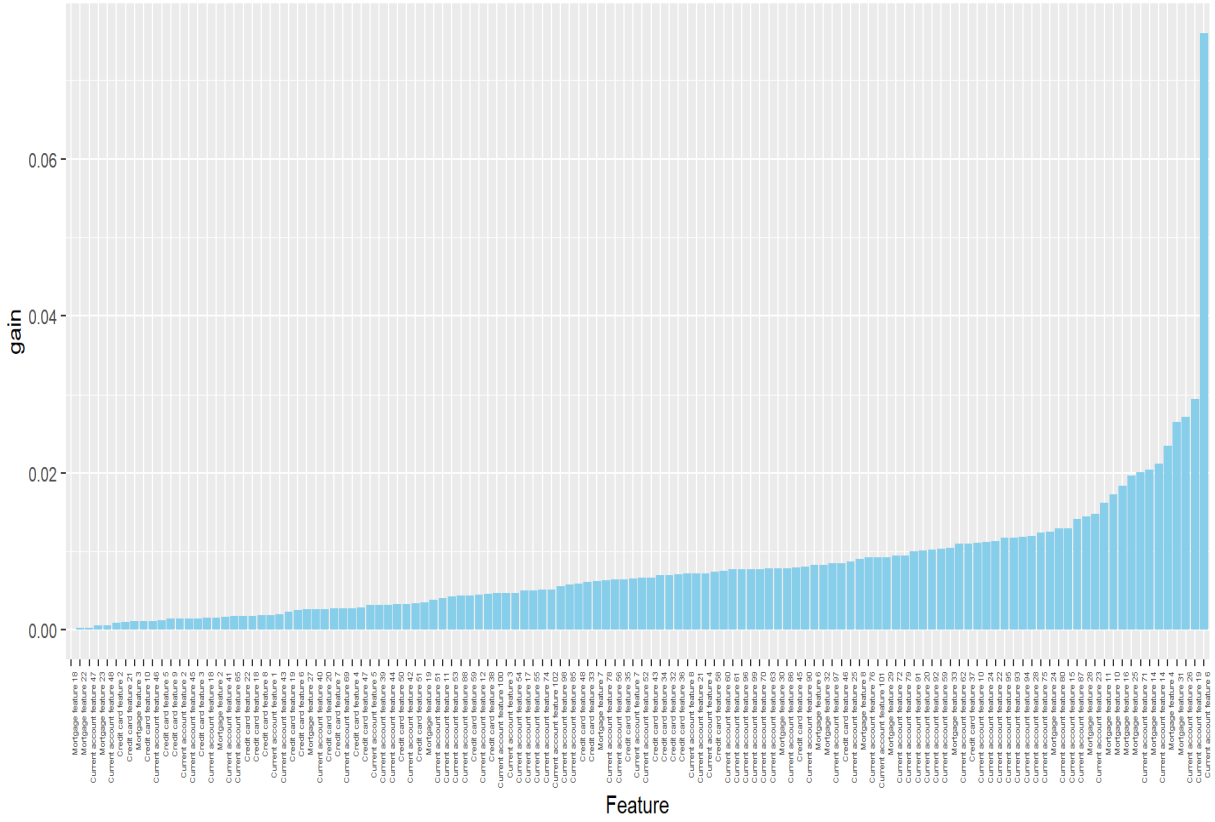


Figure 19: XGBoost gain importance

5.1.4 Meta-model

The meta-model is a special case, since its input variables are not features, but the predictions values of the random forest, logistic regression and XGBoost. Therefore, it has only three features. The model was trained with a logistic regression and therefore also the same performance measure methodology as the individual logistic regression model will be used. In Figure 20, it can be seen that the meta model is mostly fitted on the random forest. However, it also includes the predictions of the logistic regressoin and XGBoost well. The logistic regression is in all cases the weakest input algorithm, whereas the XGBoost algorithm is supportive, but its importance fluctuates throughout the iterations.

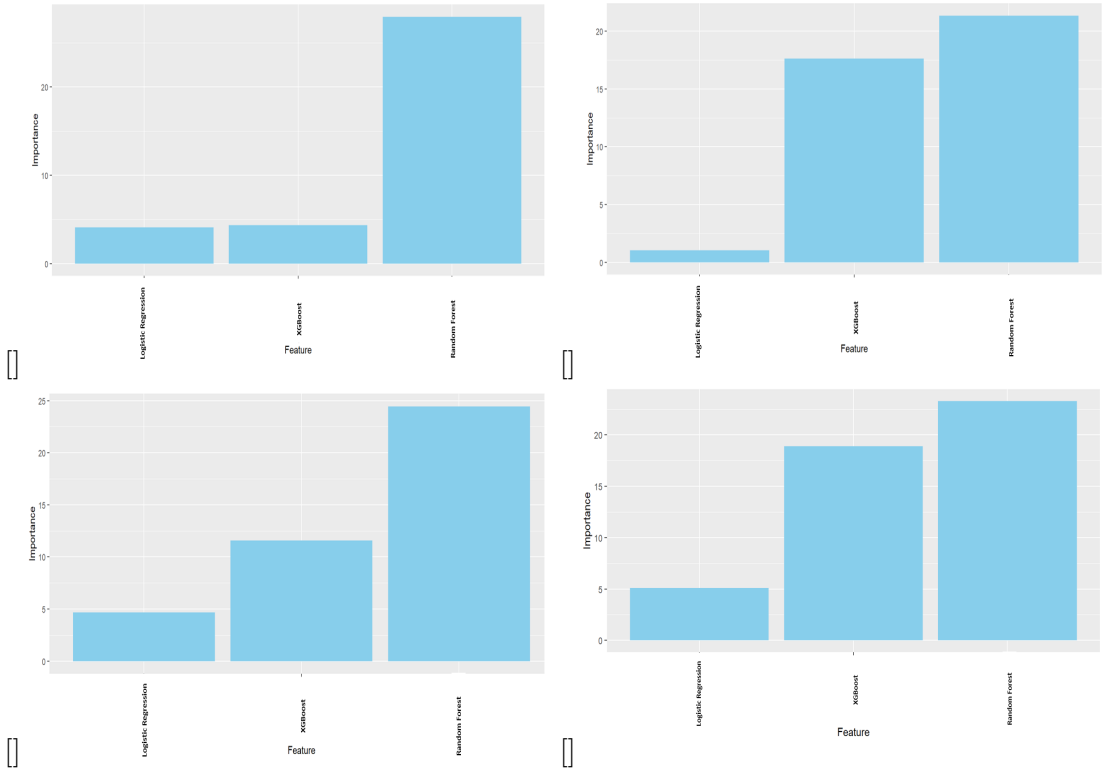


Figure 20: Meta-model variable importance (4 iterations)

5.2 Model validation

5.2.1 Predictive power

The performance metrics which are used here are the sensitivity, specificity, negative predictive value and AUC. Whereas the first four performance metrics require a certain threshold, the AUC does not require this threshold and gives the performance in one eye sight. Nevertheless, AUC gives a direct score without knowing the sensitivity/specificity ratio. Therefore, it might be the case that the model only performs really well at a certain threshold, but not at others. Therefore, the sensitivity, specificity and negative predictive value are all measured at the first quarter, median and third quarter. It is not intended to strictly judge the model on the several thresholds, but it gives a good impression about the status at several thresholds.

Figure 27 in Appendix E1 gives the overview of the metrics in all four iterations. On average the random forest has a AUC of 0.775, logistic regression has a AUC of of 0.747 and XGBoost 0.746. Overall the random forest slightly outperforms the other two algorithms when looked at the AUC. Even though it has to be said that all algorithms perform really constant and no big outliers are detected in the iterations, the random forest is even more constant than the logistic regression or XGBoost. The standard deviation for the four random forest iterations is below 0.01, while the logistic regression and XGBoost are almost 0.02.

In addition, when considering the average AUC, the stacking meta-model outperforms all the individual algorithms with an average AUC of 0.802. Although the standard deviation is slightly higher compared to the random forest, it is still lower than that of

logistic regression and XGBoost. Throughout all iterations, the meta-model consistently demonstrates the best performance in terms of AUC. This result is expected because the random forest serves as the most important input variable for the meta-model, suggesting that the meta-model is consistently able to outperform the best individual model.

Upon examining the sensitivity, specificity, and negative predictive value at various thresholds, it can be deduced that the models exhibit a logical progression. When the threshold is set low (1st quarter), it is consistently observed that the specificity is high, indicating accurate and timely warnings. However, the sensitivity reveals that this is primarily due to the abundance of warning signals rather than the model's accuracy. Conversely, as the threshold is increased, the specificity decreases, but there is typically a reduction in false warnings. It is also noteworthy that the overall performance of all four models is relatively similar at higher thresholds, with the random forest and meta-model showing superior performance around the median. While the sensitivity remains relatively consistent across all models, the random forest and meta-model outperform in terms of specificity. Additionally, it is important to highlight that the negative predictive value remains consistent across different thresholds. This indicates that regardless of the threshold used to classify instances as negative, the models consistently excel in accurately identifying instances that are genuinely negative.

5.2.2 Warning signals

In addition to identifying warning signals for deteriorating customers, it is crucial to analyze the timing of when these signals are provided. This information is significant as it demonstrates the effectiveness of the "early" warning system in delivering timely warnings. Previous findings indicated that the model performed well when using a higher threshold. However, when the threshold is lowered, the sensitivity and specificity reveal that true positive clients are only identified because more customers are marked as overdue. In fact, if the threshold is set low enough, all customers receive a warning signal. Consequently, evaluating warning signals at a lower level skews the model's performance regarding timeliness.

To address this concern, the analysis focuses on the 3rd quarter of all models. This approach allows us to determine the number of months before the occurrence of an overdue situation that warning signals are provided. Since a customer can receive a warning signal each month, a histogram is generated to illustrate the total count of warning signals provided.

The distribution of warning signals before a client becomes overdue is depicted in Appendix E2 in Figure 28, 29, and 30 for the 90th, 95th, and 99th percentiles, respectively. These Figures illustrate how the signals are distributed among the top 10%, 5%, and 1% of clients, respectively, who are marked as overdue. For this purpose, the meta-model was used since it was the top performer among all models. It is important to note that false positives, which do not result in reaching an overdue date, are logically excluded from these distributions.

The Figures indicate a notable pattern when examining different percentiles. At the 99th percentile, a significant majority of warning signals are observed to be generated shortly

before the client becomes overdue. In contrast, when analyzing the 95th and 90th percentiles, it is apparent that a relatively higher proportion of warning signals are provided further in advance before the overdue situation occurs, compared to the 99th percentile.

It is possible that the observed difference in warning signal timing across percentiles could be influenced by several factors. One factor could be that customers already warned in the 99th percentile continue to receive subsequent months marked as risky, while new overdue customers may not be identified as early as those in the 99th percentile.

To gain further insights into this matter, Figure 31, 32, and 33 in Appendix E2 provide information about the timing of the last warning signal before a customer becomes overdue. Upon examining these figures, a noteworthy observation emerges: in the 99th percentile, on average, the last warning signal is received shortly before the overdue situation compared to the 90th percentile. This finding suggests that customers in the 99th percentile tend to receive their final warning signal shorter in relation to their overdue date than customers in the 90th percentile.

5.2.3 Feature importance

The last iteration of model that included all features is runned again, but only with the top ten features of all individual models to evaluate the influence of the main predictors. Since exactly the same train and test set are used, this yields a fair comparison. The AUC results are shown in Table 7.

Model	AUC with all features	AUC with top 10	Difference
Random Forest	0.7809	0.7443	-0.0366
XGBoost	0.759	0.7169	-0.0421
Logistic Regression	0.7593	0.724	-0.0353
Meta-model	0.8035	0.7796	-0.0239

Table 7: AUC with top ten features

An interesting observation is that when only a limited number of features are used, the performance of all models is consistently weaker compared to when the models are trained with all features. Notably, the individual models are particularly affected by the reduced feature set, experiencing a significant decline in performance. Although the meta-model also exhibits a decrease in performance, it is less pronounced than that of the individual models.

5.2.4 Random dataset validation

It is important to guarantee that input variables actually lead to the decision to generate a warning signal for a retail client. Therefore, a test is needed to ensure this. With random dataset validation a new dataset is generated with the purpose to ensure that random parameters do not lead to correct predictions and to validate that the model works based on conditional probability. The approach used was to take the maximum and minimum of each input feature and generate a random number in between. Since the data was already normalized between 0 and 1, this effectively meant that a value between 0 and 1 was

selected for each input parameter. The target variable had the same overdue/non-overdue ratio as the original dataset. The results from the random dataset usage can be found in Table 8. In total 10.000 random customers are created.

Model	AUC original dataset	AUC with random dataset	Difference
Random Forest	0.7809	0.5022	-0.2787
XGBoost	0.759	0.5166	-0.2424
Logistic Regression	0.7593	0.5112	-0.2481
Meta-model	0.8035	0.5090	-0.2945

Table 8: AUC with random dataset

According to Verbakel et al.(2020) and Zhou (2023) the AUC results of the model using random data should not exceed 0.5 in a binary classification context, since the random data follows pure chance. The table shows that this theory is approached, with values slightly above 0.5. The fact that it is not exactly 0.5 in the context of a binary classification is not surprising, since there can always be a variation. More validation samples could help reduce this error, however, in general it can be seen that predictions are actually dependent on input variables.

5.2.5 Cross-validation

Cross validation is applied to the model. In more specific, Monte Carlo cross-validation was the selected method for the several iterations performed. Thus far, only an average AUC has been mentioned. This already partly reduces the bias, since the end result is not based on a single iteration. Nevertheless, it is also important to shortly discuss the deviation between the several iterations. For this purpose, we take a t-test due to the low number of iterations. The different values of the t-test can be found in Table 9.

Model	Mean	Standard deviation	Standard Error	Margin Error	Lower bound (95% CI)	Upper bound (95% CI)
Random Forest	0.7750	0.0091	0.0046	0.0145	0.7605	0.7895
Linear	0.7465	0.0186	0.0093	0.0295	0.7170	0.7761
XGB	0.7463	0.0188	0.0094	0.0300	0.7163	0.7763
Meta	0.8022	0.0168	0.0084	0.0267	0.7755	0.8289

Table 9: Cross validation iterations

We typically see that the 95% confidence interval of all models does not covers a long range, even though only four iterations are performed. From the table it can be seen that the Random Forest has the lowest margin error, whereas XGBoost has the highest margin error. According to Trifonova et al. (2014), we can conclude that it is 95% sure that all models can be classified as good or higher, since the lower bound is in all cases higher than 0.7 AUC.

5.3 Discussion

The aim of this thesis was to create a retail EWS. Overall the results were better than expected. All models showed an AUC far above 0.5, which means that according to Zhou (2023) and Verbakel et al. (2020) the algorithm is better than pure chance in a binary setting. According to Trifonova et al. (Trifonova et al., 2014) all individual models can be remarked as 'good' and the meta-model, which outperformed all individual models, can even be remarked as 'very good'.

Also among the models the performance is as expected. Nandi & Pal (2022) described that methods, such as linear and logistic regression are transparant algorithms, but their accuracy is limited. Whereas, algorithms, such as random forest, have reduced interpretability, but higher performance. The logistic regression and XGBoost algorithm scored almost evenly with an AUC of 0.75, whereas for the random forest this was slightly with an AUC of 0.775. Therefore, the results confirm this statement from Nandi & Pal (2022). Also, the model ensembling technique 'stacking', proposed by Khelfa et al. (2023), increased the results and the meta-model yielded an average AUC of above the 0.8.

Also, when we look at the note of Verbakel et al. (2020) to look at the models performance at several thresholds, we find differences. If the threshold is lowered the model is getting more specific, but less sensitive. This was also found in this thesis, as a lower threshold always led to a less sensitive model. It was suspected that particularly at a higher threshold the model performed better and that for lower thresholds it was rather a case of 'luck' due to the relaxation of the threshold.

Hence, in addition to the aforementioned statement, it was crucial to assess the timing of warning signals provided prior to the occurrence of overdue situations. The findings distinctly demonstrated that when the threshold was set higher, the signals were distributed shortly before the situations became overdue. Conversely, when the threshold was lowered, the warnings were issued significantly earlier before the situations became overdue. However, it was important to consider the possibility that this discrepancy could be attributed to two factors: first, earlier months of clients already marked as overdue in the 99th percentile could have been classified as overdue, and second, newly identified overdue clients may have influenced the results. Therefore, an examination was conducted to determine the last signal received by clients before they became overdue. This analysis revealed that the distribution of the last signal followed a similar pattern to that of all signals. Consequently, it can be argued that the shift towards earlier warning signals was primarily caused by newly identified clients marked as overdue. This shift is likely due to the model's decreased sensitivity towards sudden problems that trigger a signal and increased sensitivity towards the overall financial status of clients. For instance, individuals with lower balances are targeted earlier than those with higher balances. Consequently, the retail EWS is more effective at higher thresholds compared to lower ones.

Upon examining the top ten features of the models, it becomes apparent that their performance is subpar compared to when the models are trained with all features. It was initially unexpected that utilizing the top features per model would result in a significant decrease in performance, which did not occur. However, there was also no noticeable improvement,

likely because the original models were not affected by overfitting or collinearity issues, and the reduced features did not contribute to enhanced decision-making capabilities for the models.

6 Conclusion

6.1 Final considerations

The motivation for this thesis was the absence of focus in literature on prevention of retail customers going overdue. Whereas, the focus of PD models is not to prevent overdue and default situations, EWS do include this focus. Especially, in the retail side in credit risk is underrepresented in literature. Therefore, the development of a retail EWS could yield an innovate way to detect customer deterioration early. Therefore, the main question in this research was:

How does the integration of various lending products into a comprehensive Early Warning System for retail loans contribute to the early detection of client deterioration?

A retail client can be present in one of the three states: normal, overdue and default. The purpose of a EWS is to detect deterioration before a non-payment situation occurs. Therefore, this thesis focused on the transition of normal to overdue clients, since before a client goes default, first a client has to be overdue. Since literature lacks knowledge about EWS several approaches have been investigated. A possible distinction is expert judgement, analogous approach and parametric approach. Each approach has their individual advantages, however, due to the fact that the parametric approach works on a large scale compared to analogous approach and expert judgement, can be tailored to the situation and is objective and transparent this method is selected for the development of an EWS.

Within the parametric domain, ML is an effective approach to create an EWS. This technique is well-suited for predicting bankruptcy and default risks in credit risk management. Within literature the retail EWS can be seen as a binary classification problem. Also, for the assessment of the binary classification model multiple evaluation metrics are necessary, since basing results on one indicator might give a distorted image. Some ML models are known for higher accuracy, however, in general this comes with reduces interpretability. Due to the fact that financial institutions are heavily regulated, to some extent explainability is required. Therefore, the choice is on more explainable algorithms, which are reinforced by model ensembling and optimization methods.

The EWS was designed in four ways, with a random forest, logistic regression, XGBoost and a meta-model created by stacking all previous models. It was found that the random forest model was using more features to construct a model compared to the xgboost and logistic regression. In all iterations the random forest was the most important feature for the meta-model, whereas the XGBoost was the second feature and logistic regression the least important feature. Regarding the overall performance the random forest scored on average 0.775 AUC, the XGBoost and logistic regression 0.75 AUC and the meta-model 0.80 AUC. The top performing meta-model showed that a higher threshold for warning signals results in signals closer to the overdue date than at a lower threshold. This suggested that at a higher threshold the model was more sensitive to emerging client deterioration, whereas for lower thresholds looked more at the clients overall status. Using the top ten features to train the model shows overall good results, but the features outside the top

ten yield valuable extra information.

In conclusion, on the question how the development of a comprehensive retail loan EWS can contribute to early deterioration detection, this thesis proved that deterioration of clients can be detected early. The thesis showed that an effective EWS can be developed with an AUC of >0.80, which can be classified as very good. In addition, a more sensitive model can detect upcoming deterioration early, whereas a more specific model goes more into the overall status of the client.

6.2 Limitations & further research

One of the primary challenges encountered during the course of this research was the limited computational resources available for data processing and analysis. Due to these constraints, it became necessary to utilize a small subset of the available data and reduce the performance parameters. Consequently, this restriction may have influenced the precision and accuracy of the model's predictions. By employing only a fraction of the available data, there is a possibility that important patterns, trends, or outliers present in the excluded portion were not captured in the analysis. As a result, the generalizability and external validity of the research findings might be compromised, as the conclusions drawn from the limited dataset may not fully reflect the characteristics of the entire population or phenomenon under investigation.

Moreover, the reduction of parameters within the models may have limited the model's ability to capture complex relationships and interactions present in the data. Random forests are known for their ensemble nature, leveraging multiple decision trees to improve prediction accuracy and generalization. However, due to computational constraints, a smaller number of trees were used, potentially reducing the model's precision and robustness. It is important to acknowledge that the limitations arising from data reduction and the reduced number of trees were not arbitrary choices but were necessary compromises given the computational constraints. However, it is crucial to recognize that these limitations introduce potential biases and uncertainties that must be considered when interpreting the results.

The utilization of only a fraction of the available data and the reduction in the number of trees within the random forest model due to computational constraints represent important limitations in this thesis. These limitations may impact the generalizability, precision, and accuracy of the findings. Future research should consider employing a larger sample size and increasing the number of trees to validate and enhance the results presented here.

Secondly, the effectiveness of EWS in identifying and mitigating financial risks is well-established in various industries. However, the credit risk sector presents unique challenges and dynamics that are distinct from other sectors. The factors influencing overdue situations in retail require specific attention and understanding. Unfortunately, the existing literature on retail EWS is relatively scarce and lacks comprehensive coverage.

The predominant emphasis on defaults rather than overdue situations in the literature

poses a limitation to this research. While defaults are undoubtedly crucial indicators of financial distress, timely identification of overdue situations can enable proactive interventions and prevent the escalation of financial difficulties for retailers. The lack of dedicated research on this aspect limits the availability of established frameworks, methodologies, and empirical evidence specific to retail EWS, hindering the development of comprehensive models and best practices.

Furthermore, the existing literature on EWS predominantly focuses on businesses. Although some studies touch upon retail within a broader context, there is a clear dearth of research dedicated exclusively to retail EWS. The retail sector possesses unique characteristics, including customer behaviour. These factors necessitate a tailored approach to EWS, which requires a deeper understanding of the specific challenges and risk factors faced by retailers.

The scarcity of dedicated research restricts the availability of well-established frameworks and empirical evidence specific to retail EWS. Literature should aim to bridge this gap by exploring and developing methodologies tailored to the unique dynamics of the retail sector, thereby enhancing the understanding and applicability of EWS in this domain.
The recommendations for further research are divided in technical recommendations and recommendations in the context of EWS.

- **Technical recommendations**
  While the primary focus of this thesis was not on creating completely explainable algorithms, there were valid reasons for not favoring unexplainable models. However, it is well-established that black box models possess significant strength and potential. Therefore, further research can delve into exploring the possibilities offered by black box models. Especially, the development of hybrid models that combine the strengths of black box models with interpretable models. The interpretability of models is essential in domains where understanding the underlying decision-making process is crucial for regulatory compliance or gaining stakeholders' trust, such as the credit risk sector. Investigating methods to strike a balance between interpretability and predictive performance by incorporating interpretability measures or leveraging model-agnostic interpretation techniques can pave the way for more transparent and reliable decision-making in complex domains.

  Also, exploring the utilization of specialized hardware for ML purposes can significantly enhance results. While regular computers are suitable for everyday tasks, they may fall short in handling intensive computations like training ML models. This highlights the need for specialized hardware to overcome limitations stemming from limited computational resources. Overall, specialized hardware could significantly improve results due to their ability to handle complexer datasets and run computational heavier models.

- **Retail EWS recommendations**
  Implementing smaller time steps in EWS can bring significant benefits. The primary objective of an EWS is to detect threats early, and smaller time steps contribute to

achieving this goal. By adopting smaller intervals, such as weekly or daily monitoring, EWS can facilitate rapid response and mitigation strategies, thereby minimizing the impact of threats. Additionally, the limitations of monthly reporting become apparent when considering scenarios where a client's condition deteriorates early in a month, potentially leading to overdue situations before any action is taken. By relying solely on monthly information, the effectiveness of an EWS can be compromised. Hence, integrating smaller time steps into the EWS framework becomes crucial to ensure prompt and effective risk detection and response.

Besides, the initial development of the EWS relied on data exclusively sourced from a single country, without incorporating any personal information. However, it is crucial to consider the applicability of the research conclusions to other countries, given the distinct economic, social, and regulatory environments that each country possesses. To enhance the effectiveness of the retail EWS, it would be advisable to augment the dataset with socio-demographic details like age, profession, nationality, and geographic information. By including such information, we can potentially enhance the performance and accuracy of the system, as it takes into account a broader range of factors that influence retail dynamics. This expansion would allow for a more comprehensive and tailored analysis across different countries and contribute to a more robust and globally applicable EWS.

Furthermore, it is needed to emphasize the importance of striking a balance between false positives and true positives. While this aspect has been partially addressed, it warrants further consideration. Although the scope of this thesis intentionally excluded managerial decisions like establishing a threshold for identifying deteriorating clients, it is good to acknowledge that in real-world scenarios, such exclusions are not feasible. Every instance of failing to identify a deteriorating client or falsely marking a client as positive incurs associated costs. Therefore, it is interesting further research to delve into the preferences of credit risk professionals regarding the appropriate threshold. Conducting research on the preferred threshold could lead to the development of a more practical and effective implementation model.

References

Adilkhanova, I., Ngarambe, J., & Yun, G. Y. (2022, 9). *Recent advances in black box and white-box models for urban heat island prediction: Implications of fusing the two methods* (Vol. 165). Elsevier Ltd. doi: 10.1016/j.rser.2022.112520

Allen, L., DeLong, G., & Saunders, A. (2004, 4). Issues in the credit risk modeling of retail markets. *Journal of Banking & Finance*, *28*(4), 727–752. doi: 10.1016/j.jbankfin.2003.10.004

An, H., Wang, H., Delpachitra, S., Cottrell, S., & Yu, X. (2022, 6). Early warning system for risk of external liquidity shock in BRICS countries. *Emerging Markets Review*, *51*. doi: 10.1016/j.ememar.2021.100878

Assy, A. T., Mostafa, Y., El-khaleq, A. A., & Mashaly, M. (2023). Anomaly-Based Intrusion Detection System using One-Dimensional Convolutional Neural Network. *Procedia Computer Science*, *220*, 78–85. Retrieved from `https://linkinghub.elsevier.com/retrieve/pii/S1877050923005483` doi: 10.1016/j.procs.2023.03.013

Barboza, F., Kimura, H., & Altman, E. (2017, 10). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417. doi: 10.1016/j.eswa.2017.04.006

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020, 6). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. doi: 10.1016/j.inffus.2019.12.012

Benhar, H., Idri, A., & L Fernández-Alemán, J. (2020, 10). *Data preprocessing for heart disease classification: A systematic literature review.* (Vol. 195). Elsevier Ireland Ltd. doi: 10.1016/j.cmpb.2020.105635

Boonman, T. M., Jacobs, J. P. A. M., Kuper, G. H., & Romero, A. (2017, 11). Early Warning Systems with Real-Time Data. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3067757

Bräuning, M., Malikkidou, D., Scalone, S., & Scricco, G. (2019). A new approach to Early Warning Systems for small European banks. doi: 10.2866/128724

Brett Lantz. (2015). Divide and Conquer-Classification Using Decision Trees and Rules..

Budiono, H. D., Kiswanto, G., & Soemardi, T. P. (2014). Method and model development for manufacturing cost estimation during the early design phase related to the complexity of the machining processes. *International Journal of Technology*, *5*(2), 183–192. doi: 10.14716/ijtech.v5i2.402

Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018, 9). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics*, *19*. doi: 10.1186/s12863-018-0633-8

Davis, E. P., & Karim, D. (2008, 6). Comparing early warning systems for banking crises. *Journal of Financial Stability*, *4*(2), 89–120. doi: 10.1016/j.jfs.2007.12.004

Du, G., Liu, Z., & Lu, H. (2021, 4). Application of innovative risk early warning mode under big data technology in Internet credit financial risk assessment. *Journal of Computational and Applied Mathematics*, *386*. doi: 10.1016/j.cam.2020.113260

European Banking Authority. (2016). *Final Report Guidelines on the application of the definition of default under Article 178 of Regulation (EU) No 575/2013.*

Forsyth, D. (2019). *Applied Machine Learning*. Springer International Publishing. doi: 10.1007/978-3-030-18114-7

Fosić, I., Žagar, D., Grgić, K., & Križanović, V. (2023, 6). Anomaly detection in NetFlow network traffic using supervised machine learning algorithms. *Journal of Industrial Information Integration*, *33*, 100466. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S2452414X23000390 doi: 10.1016/j.jii.2023.100466

Fu, X., Ouyang, T., Chen, J., & Luo, X. (2020, 7). Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks. *Information Processing and Management*, *57*(4). doi: 10.1016/j.ipm.2020.102236

Granger Morgan, M. (2013). Use (and abuse) of expert elicitation in support of decision making for public policy. Retrieved from www.pnas.org/cgi/doi/10.1073/pnas.1319946111 doi: 10.1073/pnas.1319946111/-/DCSupplemental

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018, 8). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5). doi: 10.1145/3236009

Gunther, J. W., & Moore, R. R. (2003, 10). Early warning models in real time. *Journal of Banking and Finance*, *27*(10), 1979–2001. doi: 10.1016/S0378-4266(02)00314-X

Hanea, A. M., Nane, G. F., Bedford, T., & French, S. (2021). *Expert Judgement in Risk and Decision Analysis*. Retrieved from http://www.springer.com/series/6161

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023, 3). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, *37*(2), 788–832. doi: 10.1007/s10618-022-00894-5

Hosaka, T. (2019, 3). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications*, *117*, 287–299. doi: 10.1016/j.eswa.2018.09.039

Ionela, S. A. (2014). Early Warning Systems – Anticipations Factors of Banking Crises. *Procedia Economics and Finance*, *10*, 158–166. doi: 10.1016/s2212-5671(14)00289-5

Iustina, B. (2012). Development of an early warning system for evaluating the credit portfolio's quality. A case study on Romania. *Prague Economic Papers*(3), 347–362. doi: 10.18267/j.pep.428

Jo, T. (2021). *Machine Learning Foundations*. Springer International Publishing. doi: 10.1007/978-3-030-65900-4

Jones, S. (2017, 9). Corporate bankruptcy prediction: a high dimensional analysis. *Review of Accounting Studies*, *22*(3), 1366–1422. doi: 10.1007/s11142-017-9407-1

Kaplan, A., & Haenlein, M. (2020, 1). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*, *63*(1), 37–50. doi: 10.1016/j.bushor.2019.09.003

Kaur, M., & Rattan, D. (2023, 2). *A systematic literature review on the use of machine learning in code clone research* (Vol. 47). Elsevier Ireland Ltd. doi: 10.1016/j.cosrev.2022.100528

Khelfa, B., Ba, I., & Tordeux, A. (2023, 2). *Predicting highway lane-changing maneuvers: A benchmark analysis of machine and ensemble learning algorithms* (Vol. 612). Elsevier B.V. doi: 10.1016/j.physa.2023.128471

Klopotan, I., Zoroja, J., & Meško, M. (2018, 8). Early warning system in business, finance, and economics: Bibliometric and topic analysis. *International Journal of*

Engineering Business Management, *10*. doi: 10.1177/1847979018797013

Kočenda, E., & Vojtek, M. (2009). *Default Predictors and Credit Scoring Models for Retail Banking* (Tech. Rep.).

Korol, T. (2013, 3). Early warning models against bankruptcy risk for Central European and Latin American enterprises. *Economic Modelling*, *31*(1), 22–30. doi: 10.1016/j.econmod.2012.11.017

Koyuncugil, A. S., & Ozgulbas, N. (2012, 5). Financial early warning system model and data mining application for risk detection. *Expert Systems with Applications*, *39*(6), 6238–6253. doi: 10.1016/j.eswa.2011.12.021

Kwon, Y., & Park, S. Y. (2023, 2). Modeling an early warning system for household debt risk in Korea: A simple deep learning approach. *Journal of Asian Economics*, *84*. doi: 10.1016/j.asieco.2022.101574

Larner, A. J. (2022). *The 2x2 Matrix Contingency, Confusion and the Metrics of Binary Classification* (Tech. Rep.).

Lawson, C. E., Martí, J. M., Radivojevic, T., Jonnalagadda, S. V. R., Gentz, R., Hillson, N. J., ... Garcia Martin, H. (2021, 1). *Machine learning for metabolic engineering: A review* (Vol. 63). Academic Press Inc. doi: 10.1016/j.ymben.2020.10.005

Leow, M., & Crook, J. (2014, 7). Intensity models and transition probabilities for credit card loan delinquencies. *European Journal of Operational Research*, *236*(2), 685–694. doi: 10.1016/j.ejor.2013.12.026

Li, Z., Zhang, J., Yao, X., & Kou, G. (2021, 6). How to identify early defaults in online lending: A cost-sensitive multi-layer learning framework. *Knowledge-Based Systems*, *221*. doi: 10.1016/j.knosys.2021.106963

Liu, H., Fu, Z., Yang, K., Xu, X., & Bauchy, M. (2019, 12). Machine learning for glass science and engineering: A review. *Journal of Non-Crystalline Solids: X*, *4*. doi: 10.1016/j.nocx.2019.100036

Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2020). *Big Data Preprocessing Enabling Smart Data* (Tech. Rep.).

Markov, A., Seleznyova, Z., & Lapshin, V. (2022, 11). Credit scoring methods: Latest trends and points to consider. *Journal of Finance and Data Science*, *8*, 180–201. doi: 10.1016/J.JFDS.2022.07.002

Nandi, A., & Pal, A. K. (2022). *Interpreting Machine Learning Models*. Apress. doi: 10.1007/978-1-4842-7802-4

Nazareth, N., & Ramana Reddy, Y. V. (2023, 6). *Financial applications of machine learning: A literature review* (Vol. 219). Elsevier Ltd. doi: 10.1016/j.eswa.2023.119640

Oliveira, R. A. d., & Bollen, M. H. (2023, 1). *Deep learning for power quality* (Vol. 214). Elsevier Ltd. doi: 10.1016/j.epsr.2022.108887

Padhan, R., & Prabheesh, K. P. (2019, 12). Effectiveness of Early Warning Models: A critical review and new agenda for future direction. *Buletin Ekonomi Moneter dan Perbankan*, *22*(4), 457–484. doi: 10.21098/bemp.v22i4.1188

Parviz, R. (2022). Project estimating and cost management..

Percic, S., Apostoaie, C.-M., & Cocriş, V. (2019). *Early Warning Systems For Financial Crisis - A Critical Approach* (Tech. Rep.).

Petch, J., Di, S., & Nelson, W. (2022, 2). *Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology* (Vol. 38) (No. 2). Elsevier Inc. doi: 10.1016/j.cjca.2021.09.004

Prajwala, T. (2015, 1). A Comparative Study on Decision Tree and Random Forest Using R Tool. *IJARCCE*, 196–199. doi: 10.17148/ijarcce.2015.4142

Ren, Q., Li, M., & Han, S. (2019, 1). Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. *Big Earth Data*, *3*(1), 8–25. doi: 10.1080/20964471.2019.1572452

Samitas, A., Kampouris, E., & Kenourgios, D. (2020, 10). Machine learning as an early warning system to predict financial crisis. *International Review of Financial Analysis*, *71*. doi: 10.1016/j.irfa.2020.101507

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018, 11). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches. *IEEE Computational Intelligence Magazine*, *13*(4), 59–76. doi: 10.1109/MCI.2018.2866730

Shih, W. Y., & Mabon, L. (2021, 9). Understanding heat vulnerability in the subtropics: Insights from expert judgements. *International Journal of Disaster Risk Reduction*, *63*. doi: 10.1016/j.ijdrr.2021.102463

Sidumo, B., Sonono, E., & Takaidza, I. (2022, 11). *An approach to multi-class imbalanced problem in ecology using machine learning* (Vol. 71). Elsevier B.V. doi: 10.1016/j.ecoinf.2022.101822

Smiti, A. (2020, 11). *A critical overview of outlier detection methods* (Vol. 38). Elsevier Ireland Ltd. doi: 10.1016/j.cosrev.2020.100306

Song, Y., Wang, Y., Ye, X., Zaretzki, R., & Liu, C. (2023, 6). Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme. *Information Sciences*, *629*, 599–617. doi: 10.1016/j.ins.2023.02.014

Stańczyk, U., Zielosko, B., & Jain, L. C. (2018). *Advances in Feature Selection for Data and Pattern Recognition*. Retrieved from `http://www.kesinternational.org/organisation.php`

Sun, J., Fujita, H., Zheng, Y., & Ai, W. (2021, 6). Multi-class financial distress prediction based on support vector machines integrated with the decomposition and fusion methods. *Information Sciences*, *559*, 153–170. doi: 10.1016/j.ins.2021.01.059

Sun, J., Lang, J., Fujita, H., & Li, H. (2018, 1). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, *425*, 76–91. doi: 10.1016/j.ins.2017.10.017

Tran, K. L., Le, H. A., Nguyen, T. H., & Nguyen, D. T. (2022, 11). Explainable Machine Learning for Financial Distress Prediction: Evidence from Vietnam. *Data*, *7*(11). doi: 10.3390/data7110160

Trifonova, O. P., Lokhov, P. G., & Archakov, A. I. (2014). *Metabolic profiling of human blood* (Vol. 60) (No. 3). doi: 10.18097/pbmc20146003281

Tsamardinos, I. (2022, 12). *Don't lose samples to estimation* (Vol. 3) (No. 12). Cell Press. doi: 10.1016/j.patter.2022.100612

Tsamardinos, I., Greasidou, E., & Borboudakis, G. (2018, 12). Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine Learning*, *107*(12), 1895–1922. doi: 10.1007/s10994-018-5714-4

Uddin, M. H., Akter, S., Mollah, S., & Al Mahi, M. (2022, 9). Differences in bank and microfinance business models: An analysis of the loan monitoring systems and funding sources. *Journal of International Financial Markets, Institutions and Money*, *80*. doi: 10.1016/j.intfin.2022.101644

Valero-Carreras, D., Alcaraz, J., & Landete, M. (2023, 4). Comparing two SVM models through different metrics based on the confusion matrix. *Computers and Operations Research*, *152*. doi: 10.1016/j.cor.2022.106131

Veganzones, D., & Séverin, E. (2018, 8). An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, *112*, 111–124. doi: 10.1016/j.dss.2018.06.011

Verbakel, J. Y., Steyerberg, E. W., Uno, H., De Cock, B., Wynants, L., Collins, G. S., & Van Calster, B. (2020, 10). ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *Journal of Clinical Epidemiology*, *126*, 207–216. doi: 10.1016/j.jclinepi.2020.01.028

Vynokurova, O., & Peleshko, D. (2018). Using Stacking Approaches for Machine Learning Models. *IEEE Second International Conference on Data Stream Mining & Processing*.

Wang, H., & Wang, S. (2010). Mining incomplete survey data through classification. *Knowledge and Information Systems*, *24*(2), 221–233. doi: 10.1007/s10115-009-0245-8

Xu, Q.-S., & Liang, Y.-Z. (2000). *Monte Carlo cross validation* (Vol. 56; Tech. Rep.). Retrieved from `www.elsevier.comrlocaterchemometrics`

Xue, L. C., Dobbs, D., Bonvin, A. M., & Honavar, V. (2015, 11). *Computational prediction of protein interfaces: A review of data driven methods* (Vol. 589) (No. 23). Elsevier B.V. doi: 10.1016/j.febslet.2015.10.003

Zhang, H., Zhang, X., Zhang, T., & Zhu, J. (2023, 7). Capturing the form of feature interactions in black-box models. *Information Processing and Management*, *60*(4). doi: 10.1016/j.ipm.2023.103373

Zhang, W., Zhang, R., Wu, C., Goh, A. T. C., Lacasse, S., Liu, Z., & Liu, H. (2020, 7). State-of-the-art review of soft computing applications in underground excavations. *Geoscience Frontiers*, *11*(4), 1095–1106. doi: 10.1016/j.gsf.2019.12.003

Zhang, Z., Wu, C., Qu, S., & Chen, X. (2022, 7). An explainable artificial intelligence approach for financial distress prediction. *Information Processing and Management*, *59*(4). doi: 10.1016/j.ipm.2022.102988

Zhou, T. (2023, 4). Discriminating abilities of threshold-free evaluation metrics in link prediction. *Physica A: Statistical Mechanics and its Applications*, *615*. doi: 10.1016/j.physa.2023.128529

Appendix

A Description datasets

A1. Description mortgage dataset

Feature	Description	Data type
1	General information about the mortgage	Integer
2	General information about the mortgage	Integer
3	General information about the mortgage	Integer
4	Information about the size of the mortgage	Numeric
5	Information about the size of the mortgage	Numeric
6	Information about the size of the mortgage	Numeric
7	Information about the mortgage payments	Integer
8	Information about the mortgage payments	Integer
9	Information about the mortgage payments	Integer
10	Information about the mortgage payments	Numeric
11	Information about the mortgage payments	Numeric
12	Information about the mortgage payments	Numeric
13	Information about the mortgage payments	Numeric
14	Information about the size of the mortgage	Numeric
15	Information about the size of the mortgage	Numeric
16	Information about the size of the mortgage	Numeric
17	Information about the size of the mortgage	Numeric
18	Information about the mortgage payments	Integer
19	Information about mortgage defaults	Numeric
20	Information about mortgage defaults	Numeric
21	Information about mortgage defaults	Numeric
22	General information about the mortgage	Integer
23	General information about the mortgage	Integer
24	Information about the mortgage payments	Numeric
25	Information about the size of the mortgage	Numeric
26	General information about the mortgage	Integer
27	Information about mortgage defaults	Numeric
28	Information about mortgage defaults	Numeric
29	Information about mortgage defaults	Numeric
30	Information about mortgage defaults	Numeric
31	Information about the mortgage payments	Integer
32	Information about the mortgage payments	Integer
33	General information about the mortgage	Integer

Table 10: Input variables mortgage dataset

A2. Description current account dataset

Feature	Description	Data type
1	General information about the current account	Integer
2	General information about the current account	Integer
3	Information about the balance	Numeric
4	Information about the balance	Numeric
5	Information about the balance	Numeric
6	Information about the balance	Numeric
7	Information about the balance	Numeric
8	Information about the balance	Numeric
9	Information about the balance	Numeric
10	Information about the balance	Numeric
11	Information about the balance	Numeric
12	Information about the balance	Numeric
13	Information about the balance	Numeric
14	General information about the current account	Integer
15	Information about the balance	Integer
16	Information about the balance	Integer
17	Information about the balance	Integer
18	Information about the balance	Integer
19	Information about the balance	Numeric
20	Information about the balance	Numeric
21	Information about the balance	Numeric
22	Information about the balance	Numeric
23	Information about the balance	Numeric
24	Information about the balance	Numeric
25	Information about the balance	Numeric
26	Information about the balance	Numeric
27	Information about the balance	Numeric
28	Information about the balance	Numeric
29	Information about the balance	Integer
30	Information about spendings	Numeric
31	Information about spendings	Numeric
32	Information about spendings	Numeric
33	Information about spendings	Numeric
34	General information about the current account	Integer
35	General information about the current account	Integer
36	General information about the current account	Integer
37	General information about the current account	Integer
38	General information about the current account	Integer
39	Information about spendings	Integer
40	Information about spendings	Integer
41	Information about spendings	Integer
42	Information about the balance	Numeric
43	Information about the balance	Integer

44	Information about the balance	Integer
45	Information about spendings	Integer
46	Information about spendings	Integer
47	Information about the balance	Integer
48	Information about spendings	Integer
49	Information about the balance	Integer
50	Information about the balance	Integer
51	Information about the balance	Numeric
52	Information about the balance	Numeric
53	Information about spendings	Numeric
54	Information about spendings	Numeric
55	Information about the balance	Numeric
56	Information about spendings	Numeric
57	Information about the balance	Integer
58	Information about the balance	Numeric
59	Information about spendings	Integer
60	Information about spendings	Integer
61	Information about spendings	Integer
62	Information about spendings	Integer
63	Information about spendings	Integer
64	Information about the balance	Integer
65	Information about spendings	Integer
66	Information about spendings	Numeric
67	Information about spendings	Numeric
68	Information about spendings	Numeric
69	Information about spendings	Integer
70	Information about spendings	Integer
71	Information about spendings	Numeric
72	Information about spendings	Numeric
73	Information about spendings	Numeric
74	Information about spendings	Numeric
75	Information about spendings	Numeric
76	Information about spendings	Numeric
77	Information about the balance	Integer
78	Information about spendings	Numeric
79	Information about the balance	Numeric
80	Information about spendings	Numeric
81	Information about the balance	Numeric
82	Information about spendings	Numeric
83	Information about spendings	Numeric
84	Information about spendings	Numeric
85	Information about spendings	Numeric
86	Information about spendings	Numeric
87	Information about spendings	Numeric
88	Information about spendings	Numeric
89	Information about spendings	Numeric

90	Information about spendings	Numeric
91	Information about spendings	Numeric
92	Information about spendings	Numeric
93	Information about spendings	Numeric
94	Information about spendings	Numeric
95	Information about spendings	Numeric
96	Information about the balance	Numeric
97	Information about spendings	Numeric
98	Information about the balance	Numeric
99	Information about spendings	Numeric
100	Information about spendings	Numeric
101	Information about the balance	Numeric
102	Information about the balance	Numeric

Table 11: Input variables current account dataset

A3. Description credit card dataset

Feature	Description	Data type
1	Information about the credit card payments	integer
2	Information about the credit card payments	integer
3	Information about the credit card payments	integer
4	Information about the credit card payments	numeric
5	Information about the credit card payments	numeric
6	Information about the credit card payments	numeric
7	Information about the credit card payments	numeric
8	Information about the credit card payments	numeric
9	Information about the credit card payments	numeric
10	Information about the credit card payments	numeric
11	Information about the credit card payments	numeric
12	Information about the credit card payments	numeric
13	Information about the credit card payments	numeric
14	Information about the credit card payments	numeric
15	Information about the credit card payments	numeric
16	Information about the credit card payments	numeric
17	Information about the credit card payments	numeric
18	Information about the credit card payments	numeric
19	Information about the credit card payments	numeric
20	Information about the credit card payments	numeric
21	Information about the credit card payments	numeric
22	Information about the credit card payments	numeric
23	Information about the credit card payments	numeric
24	Information about the credit card payments	numeric
25	Information about the credit card payments	numeric
26	Information about the credit card payments	numeric
27	Information about the credit card payments	numeric
28	Information about the credit card payments	numeric
29	Information about the credit card payments	numeric
30	Information about the credit card payments	numeric
31	Information about the credit card payments	numeric
32	Information about the credit card payments	numeric
33	Information about the credit card payments	numeric
34	Information about the credit card payments	numeric
35	Information about the credit card payments	numeric
36	Information about the credit card payments	numeric
37	Information about the credit card payments	numeric
38	Information about the credit card payments	numeric
39	Information about the credit card payments	integer
40	Information about the credit card payments	integer
41	Information about the credit card payments	integer
42	Information about the credit card payments	integer
43	Information about the credit card payments	numeric

44	Information about the credit card payments	numeric
45	Information about the credit card payments	numeric
46	Information about the credit card payments	numeric
47	Information about the credit card payments	numeric
48	Information about the credit card payments	numeric
49	Information about the credit card payments	numeric
50	Information about the credit card payments	numeric
51	Information about the credit card payments	numeric
52	Information about the credit card payments	numeric
53	Information about the credit card payments	numeric
54	Information about the credit card payments	numeric
55	Information about the credit card payments	numeric
56	Information about the credit card payments	numeric
57	Information about the credit card payments	integer
58	Information about the credit card payments	numeric
59	Information about the credit card payments	numeric

Table 12: Input variables credit card dataset

B Cleaning datasets

B1. Percentage zero's mortgage dataset

Feature	Percentage zero's
1	0.00000000
2	44.68021768
3	84.74203434
4	0.09259985
5	0.09259985
6	7.98189627
7	0.00000000
8	0.05598142
9	0.05598142
10	0.02384150
11	0.09259985
12	0.02384150
13	0.09259985
14	0.09259985
15	0.09259985
16	0.00000000
17	0.00000000

Feature	Percentage zero's
18	99.53798866
19	0.00000000
20	0.00000000
21	0.00006586
22	97.39804961
23	99.66160880
24	0.45799385
25	0.47347107
26	0.00000000
27	0.00000000
28	0.00000000
29	52.47355537
30	41.89866573
31	96.81340603
32	96.81340603
33	0.00454437

Table 13: Percentage zero's mortages

B2. Percentage zero's current account dataset

Feature	Percentage zero's	Feature	Percentage zero's
1	0.00000	52	0.76922
2	48.20908	53	22.13630
3	0.03141	54	24.50566
4	0.04186	55	24.15735
5	0.15559	56	16.19089
6	0.10167	57	98.06684
7	0.85571	58	93.80762
8	0.85571	59	0.61674
9	0.05181	60	1.01180
10	0.72581	61	0.66521
11	0.85571	62	12.22790
12	0.85571	63	12.43557
13	0.93693	64	93.73118
14	0.00000	65	77.95280
15	62.93480	66	0.61674
16	81.83775	67	1.01180
17	62.93480	68	0.66521
18	62.93480	69	10.51169
19	0.22135	70	1.01180
20	0.01934	71	1.00374
21	0.18600	72	0.61674
22	0.01412	73	1.01180
23	0.18600	74	0.66521
24	0.18600	75	12.22790
25	3.61374	76	12.43557
26	0.91658	77	93.73118
27	0.02107	78	77.95280
28	0.19900	79	20.62827
29	62.85398	80	0.64848
30	90.35831	81	5.54423
31	96.50383	82	0.61674
32	84.48655	83	1.01180
33	93.78961	84	0.66521
34	98.00936	85	0.77650
35	98.32532	86	1.45361
36	99.70250	87	0.87211
37	99.79127	88	0.65043
38	0.00000	89	0.61674
39	0.61674	90	0.61674
40	49.43557	91	1.01180
41	0.61674	92	0.66521
42	0.61674	93	12.43557
43	4.48385	94	12.22790
44	0.76922	95	7.42854
45	22.13630	96	20.62827
46	24.50566	97	0.77650
47	24.15735	98	1.47523
48	16.19089	99	0.87422
49	98.06684	100	0.78728
50	93.80762	101	40.39988
51	4.48385	102	37.41556

Table 14: Percentage zero's current account

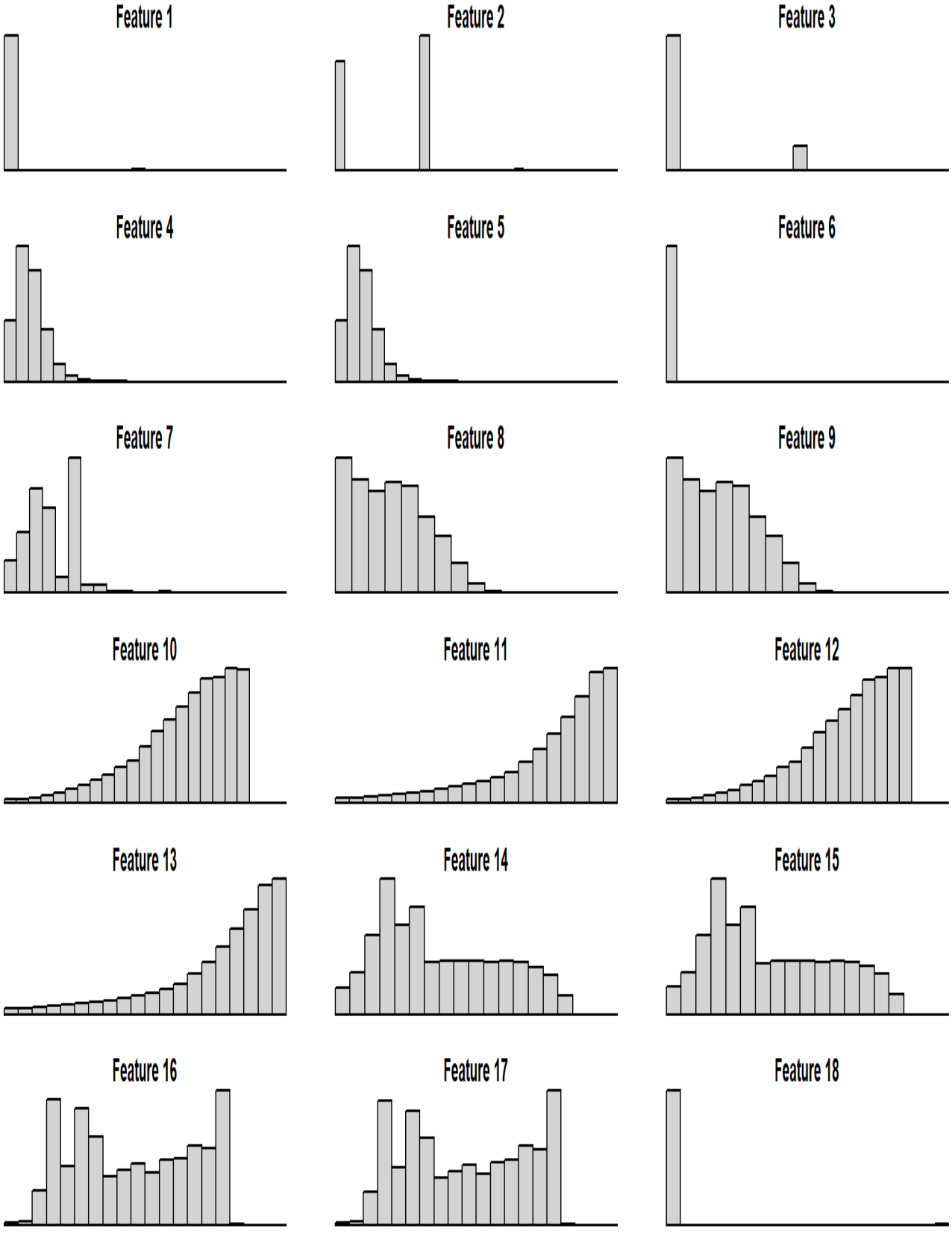
B3. Percentage zero's creditcard dataset

Feature	Percentage zero's
Feature 1	0.000509059563362644
Feature 2	0
Feature 3	0
Feature 4	0
Feature 5	73.1214853679313
Feature 6	14.5467163961298
Feature 7	41.9681430525248
Feature 8	48.9771296506833
Feature 9	62.425210665816
Feature 10	72.5519325597891
Feature 11	82.509986051768
Feature 12	93.5042302849715
Feature 13	94.6301851958692
Feature 14	98.9649122211626
Feature 15	98.2528227352788
Feature 16	99.275608241335
Feature 17	99.6747109390113
Feature 18	14.5467163961298
Feature 19	41.9681430525248
Feature 20	48.9771296506833
Feature 21	62.425210665816
Feature 22	72.5519325597891
Feature 23	95.34940151564
Feature 24	82.509986051768
Feature 25	93.5042302849715
Feature 26	94.6301851958692
Feature 27	90.8448182827045
Feature 28	98.9649122211626
Feature 29	98.2528227352788
Feature 30	99.275608241335

Feature	Percentage zero's
Feature 31	99.6747109390113
Feature 32	0
Feature 33	0
Feature 34	0.139821693403606
Feature 35	0.119119937826859
Feature 36	0.0058541849786704
Feature 37	0
Feature 38	0
Feature 39	100
Feature 40	100
Feature 41	100
Feature 42	100
Feature 43	1.59471392549404
Feature 44	87.4510030170263
Feature 45	14.1716243411921
Feature 46	27.4240567974724
Feature 47	23.9034008572563
Feature 48	38.349577989622
Feature 49	81.9277915978022
Feature 50	48.7031707623337
Feature 51	76.579866354896
Feature 52	80.8611421260364
Feature 53	95.3264938352887
Feature 54	93.2560637478323
Feature 55	96.5377162230495
Feature 56	98.4423625793709
Feature 57	0
Feature 58	0
Feature 59	0

Table 15: Percentage zero's credit card

B4. Histograms of mortgage features



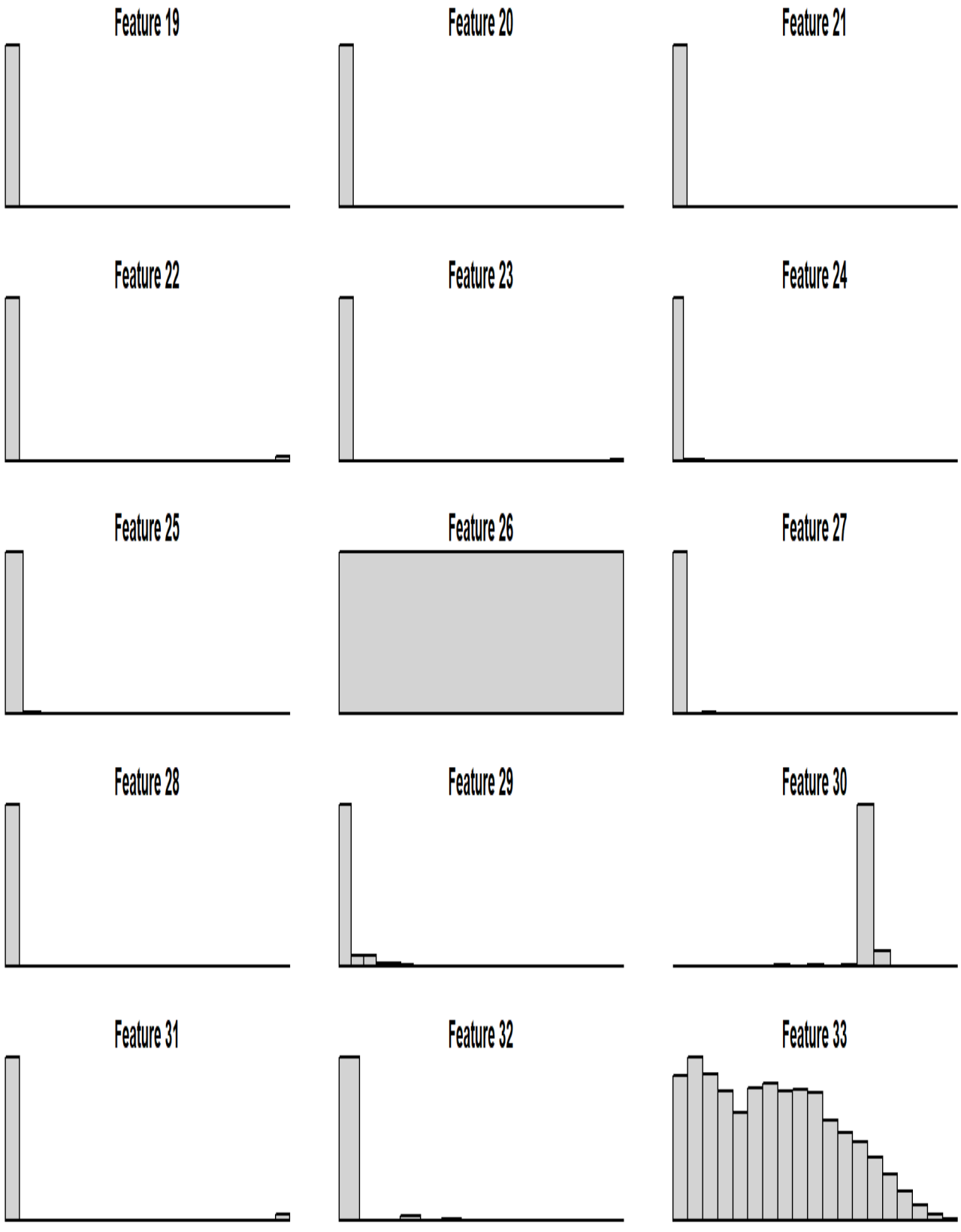
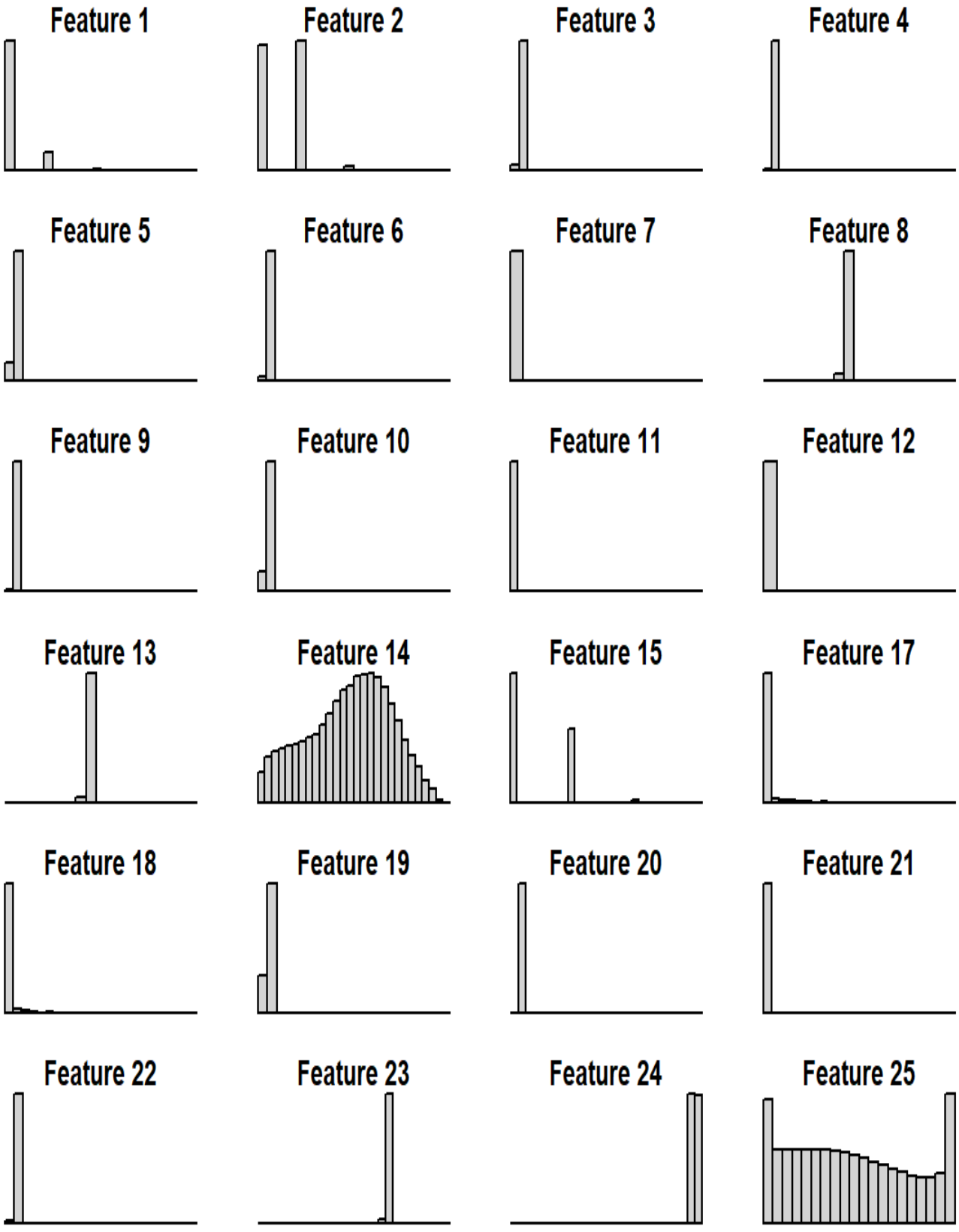
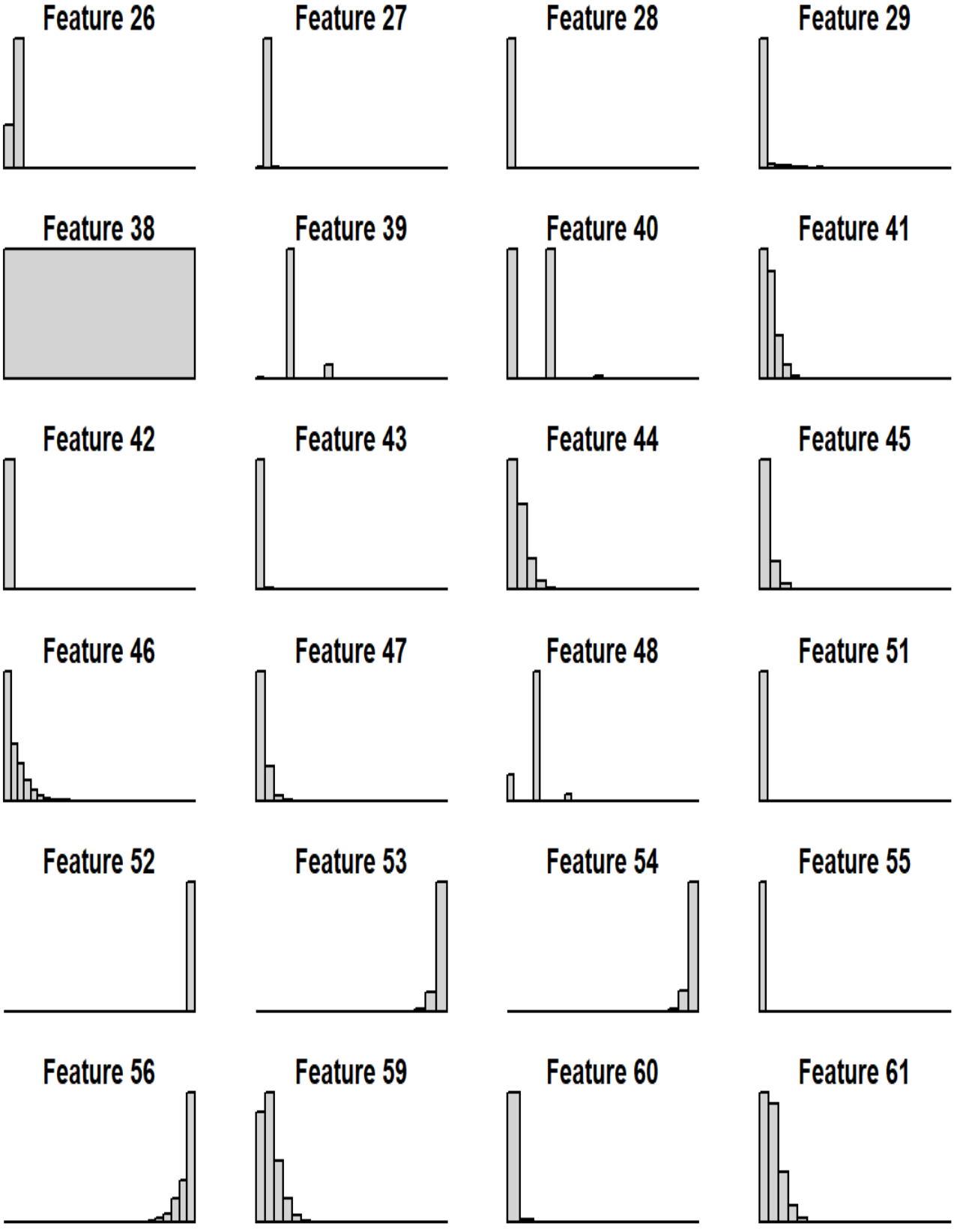
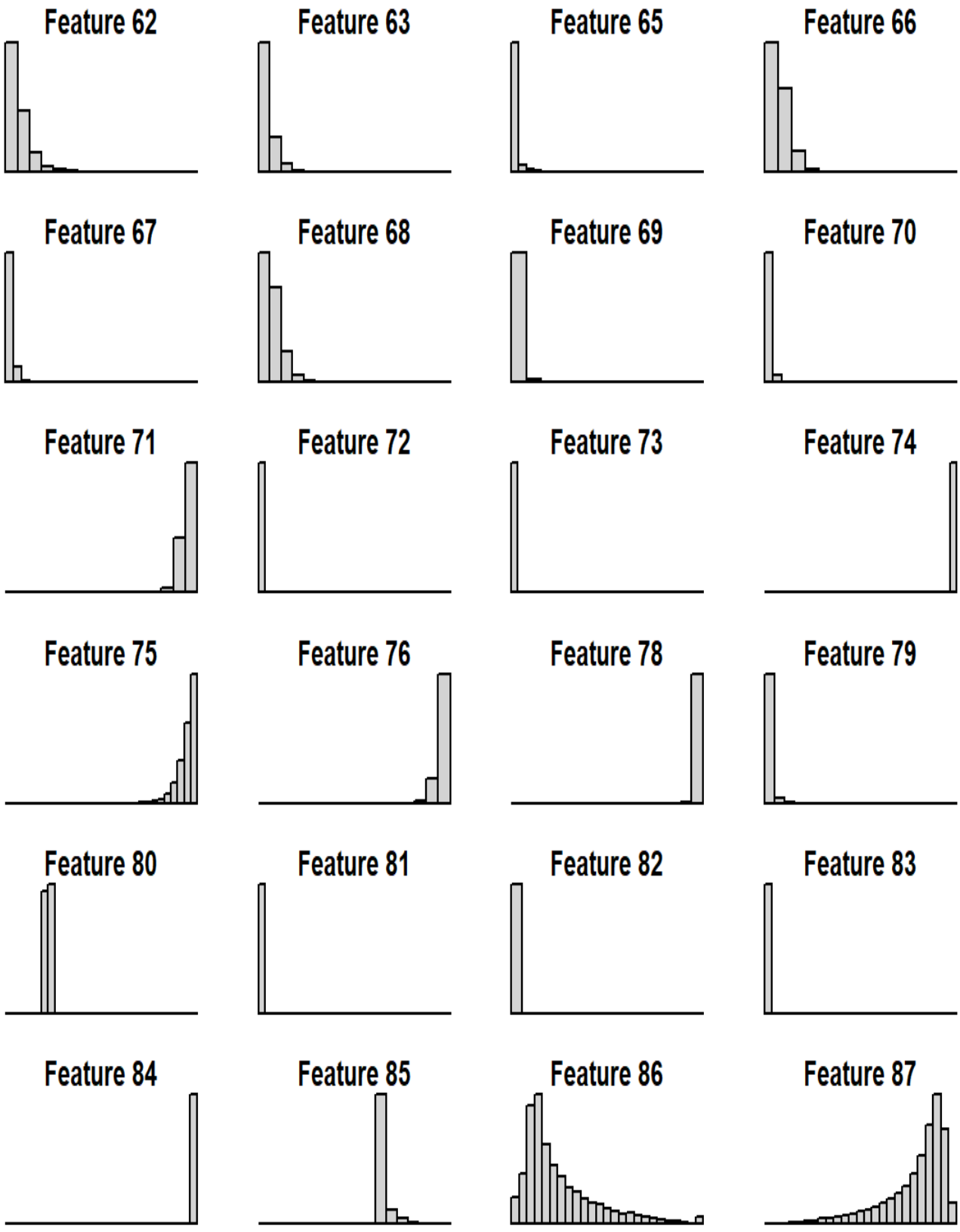


Figure 21: Histogram of the mortgage features

B5. Histograms of current account features




Feature 26
Feature 27
Feature 28
Feature 29
Feature 38
Feature 39
Feature 40
Feature 41
Feature 42
Feature 43
Feature 44
Feature 45
Feature 46
Feature 47
Feature 48
Feature 51
Feature 52
Feature 53
Feature 54
Feature 55
Feature 56
Feature 59
Feature 60
Feature 61


Feature 62
Feature 63
Feature 65
Feature 66
Feature 67
Feature 68
Feature 69
Feature 70
Feature 71
Feature 72
Feature 73
Feature 74
Feature 75
Feature 76
Feature 78
Feature 79
Feature 80
Feature 81
Feature 82
Feature 83
Feature 84
Feature 85
Feature 86
Feature 87

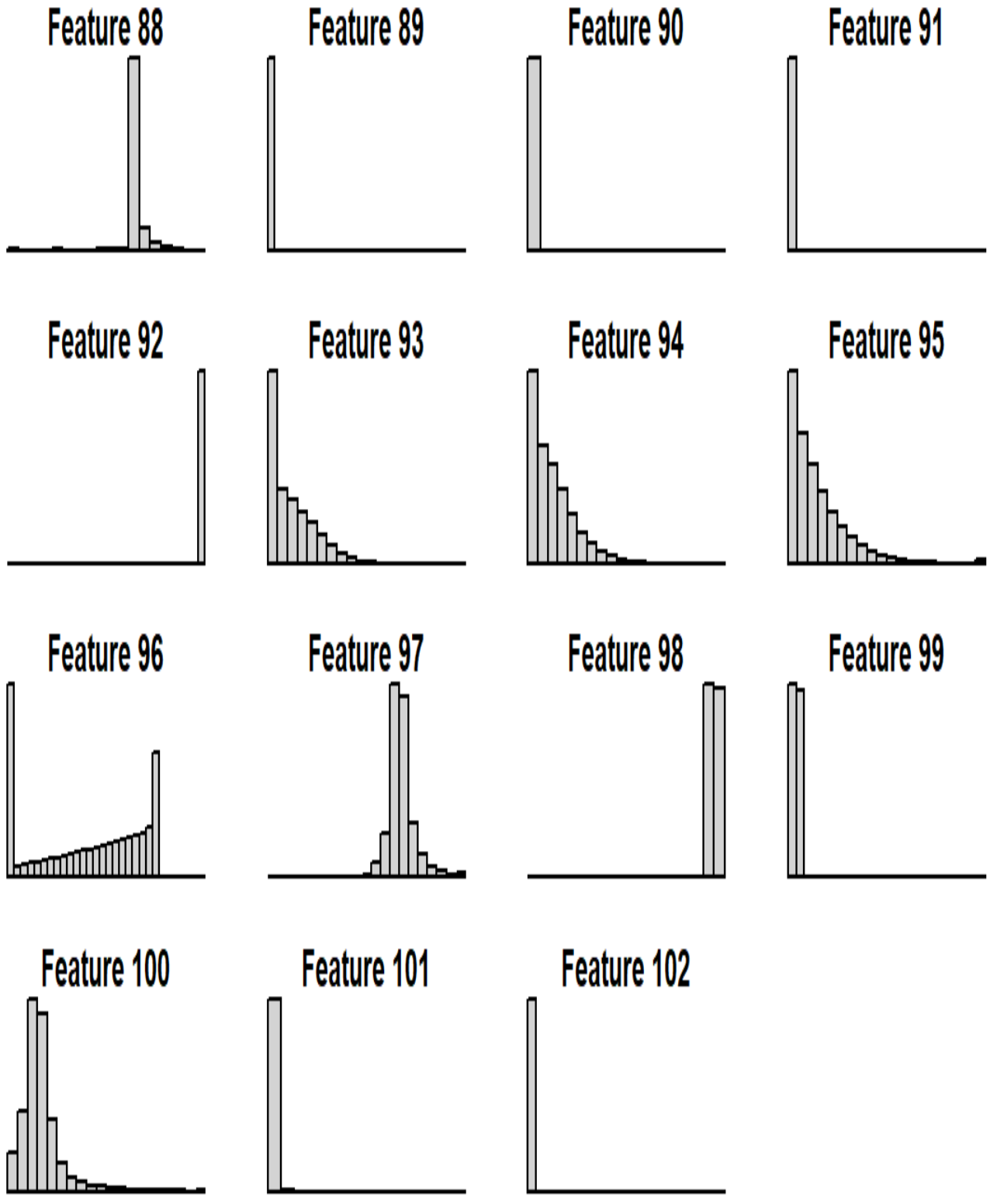
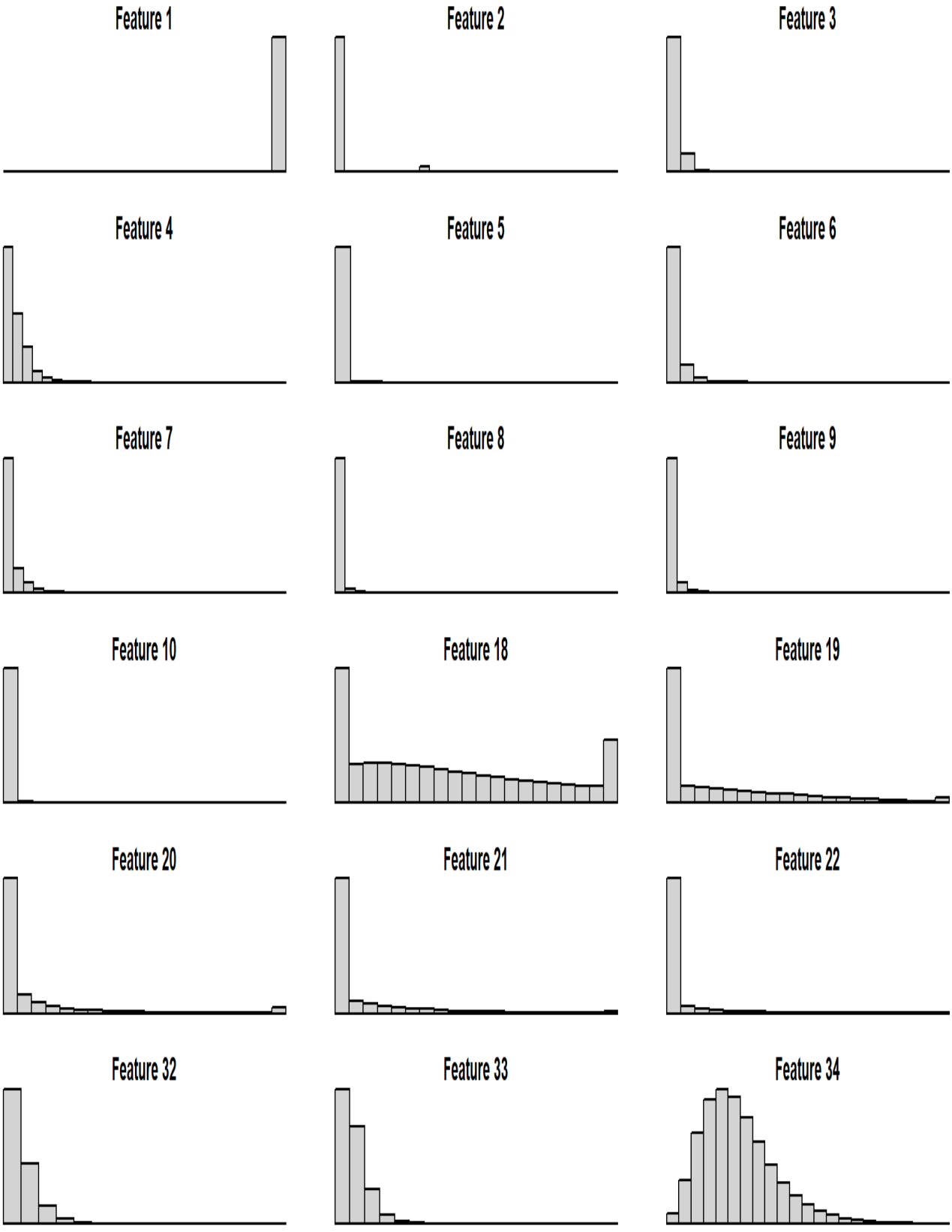


Figure 22: Histogram of the current account features

B6. Histograms of credit card features



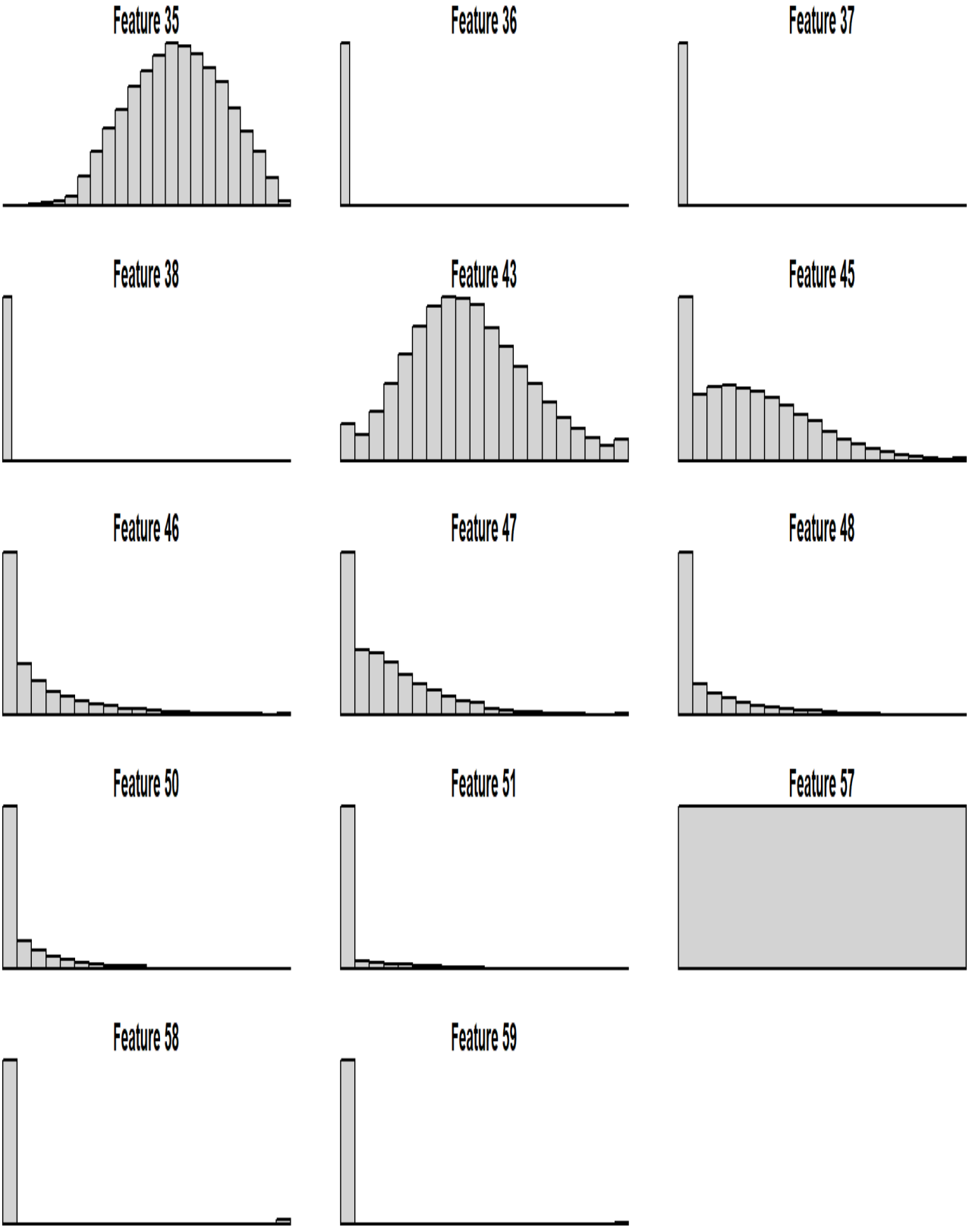


Figure 23: Histogram of the credit card features

C Data reduction

C1. Correlation matrix mortgage features

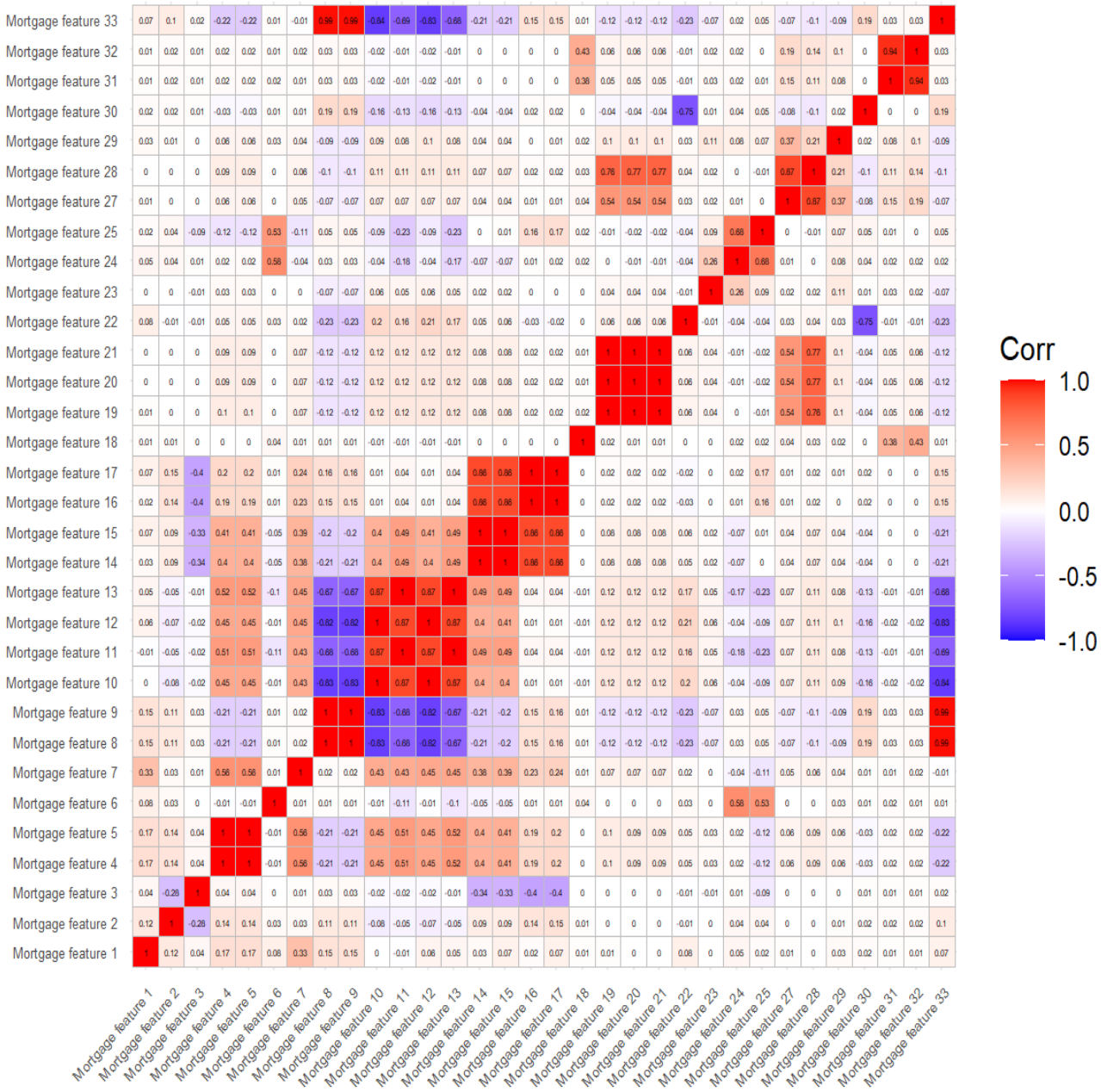


Figure 24: Correlation matrix of mortgage features

C2. Correlation matrix current account features

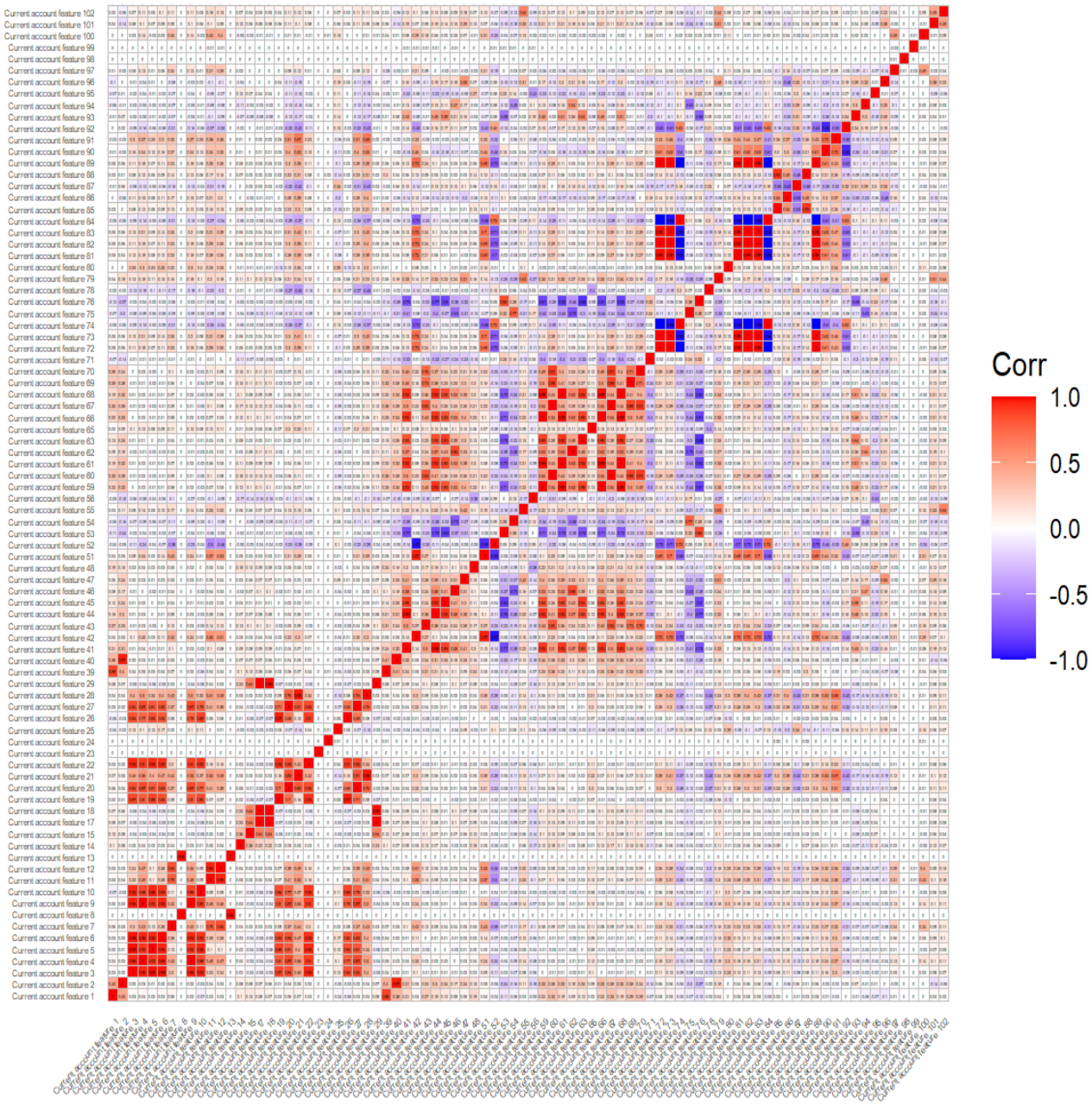


Figure 25: Correlation matrix of current account features

C3. Correlation matrix credit card features

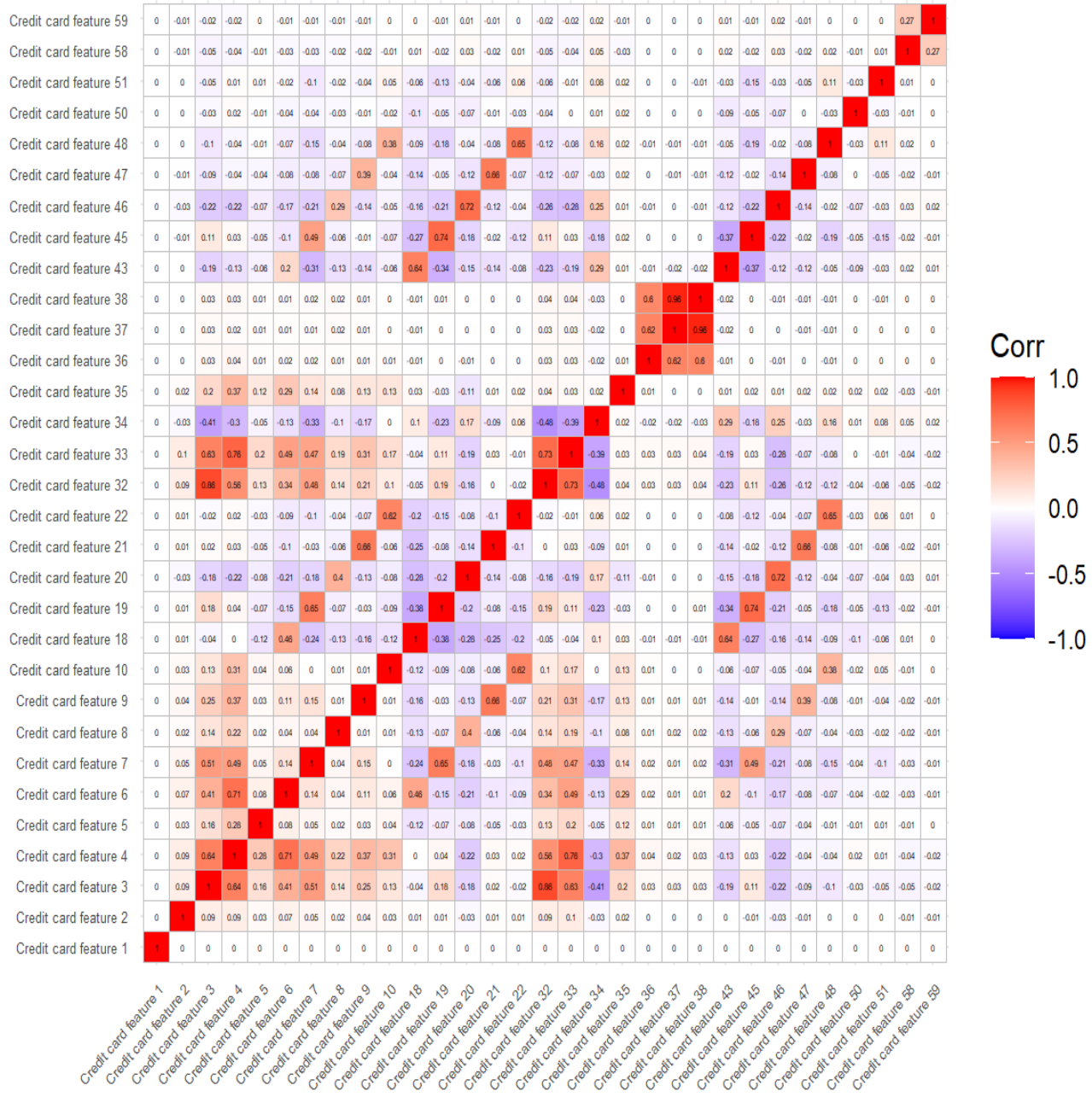


Figure 26: Correlation matrix of credit card features

D Parameter tuning

Model	*Parameters*
Random Forest	"n_estimators" = 100, "mtry" = sqrt(number of features) - 1 "weights"=NULL, "replace"=TRUE, "nodesize" = 1, "maxdepth" = NULL, "mincut" = 1
XGBoost	"n_round"= 200, "objective" = binary:logistic "eta" = 0.3, "max_depth" = 6, "subsample" = 1, "colsample_bytree" = 1, "min_child_weight" = 1, "gamma" = 0, "lambda" = 1, "alpha" = 0, "nthread" = 1, "verbosity" = 1.
Logistic regression	"family" = binomial, "link" = logit, "weights" = NULL, "offset" = NULL, "control" = list(), "model" = TRUE,

Table 16: Parameter tuning

E Performance metrics

E1. All features performance metrics

			1st quarter				Median				3rd			
Model	Iteration	AUC	Accuracy	Sensitivity	Specificity	Negative predictive value	Accuracy	Sensitivity	Specificity	Negative predictive value	Accuracy	Sensitivity	Specificity	Negative predictive value
Random Forest	1	0.7841	0.3402	0.31179	0.91343	0.98644	0.5586	0.54530	0.82686	0.98453	0.7662	0.7709	0.6696	0.9792
Random Forest	2	0.7704	0.3358	0.3034	0.9051	0.98209	0.5538	0.54070	0.79869	0.98042	0.7682	0.7750	0.6416	0.9758
Random Forest	3	0.7645	0.3622	0.33570	0.88428	0.98282	0.4811	0.46362	0.82556	0.98128	0.7624	0.7678	0.6563	0.9778
Random Forest	4	0.7809	0.2976	0.26288	0.93248	0.98615	0.5308	0.51318	0.85370	0.98465	0.7681	0.7738	0.6640	0.9768
Linear Regression	1	0.7597	0.2877	0.25634	0.92049	0.98488	0.5267	0.51207	0.82155	0.98304	0.7614	0.7683	0.6219	0.9762
Linear Regression	2	0.7474	0.29	0.25639	0.91653	0.98283	0.5321	0.5182	0.7905	0.9788	0.7611	0.7695	0.6039	0.9731
Linear Regression	3	0.72	0.2892	0.25908	0.88256	0.97754	0.528	0.5155	0.7737	0.9782	0.754	0.7638	0.5613	0.9717
Linear Regression	4	0.759	0.2897	0.25470	0.92926	0.98504	0.5282	0.51274	0.81190	0.98034	0.7588	0.7665	0.6174	0.9734
XGBoost	1	0.7568	0.2877	0.25739	0.89929	0.98100	0.529	0.51522	0.80742	0.98183	0.763	0.7692	0.6378	0.9772
XGBoost	2	0.7506	0.293	0.25929	0.92144	0.98401	0.5301	0.5158	0.7954	0.9792	0.7621	0.7698	0.6187	0.9741
XGBoost	3	0.7186	0.2877	0.25707	0.89119	0.97899	0.5251	0.51318	0.75993	0.97683	0.7579	0.7668	0.5820	0.9731
XGBoost	4	0.7593	0.2945	0.25989	0.92765	0.98501	0.5318	0.51679	0.80707	0.98000	0.7643	0.7712	0.6383	0.9750
Meta-model	1	0.8226	0.2983	0.26263	0.95935	0.99171	0.5371	0.5187	0.8780	0.9875	0.7717	0.7751	0.7073	0.9800
Meta-model	2	0.801	0.2979	0.26047	0.94656	0.98829	0.5379	0.51961	0.85496	0.98414	0.7688	0.7739	0.6794	0.9766
Meta-model	3	0.7816	0.295	0.2599	0.9286	0.9850	0.5312	0.51583	0.80952	0.97995	0.7683	0.7731	0.6825	0.9778
Meta-model	4	0.8035	0.2954	0.26140	0.94167	0.98839	0.5388	0.52237	0.85000	0.98511	0.7717	0.7750	0.7083	0.9806

Figure 27: All feature performance metrics

E2. Warning signal distribution

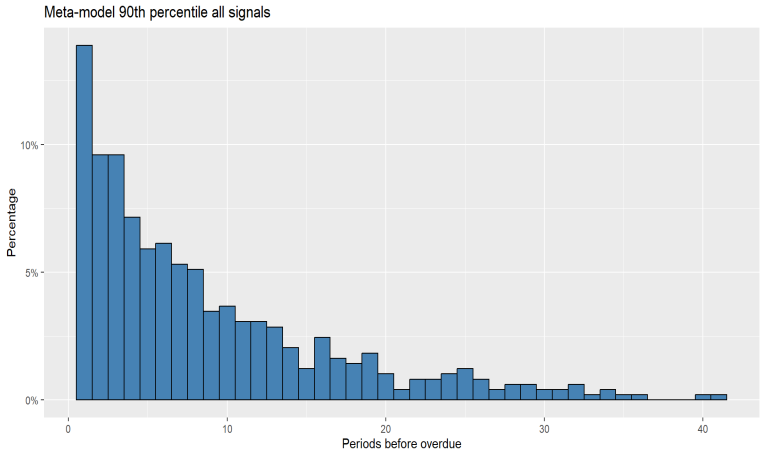


Figure 28: Meta-model 90th percentile signal distribution

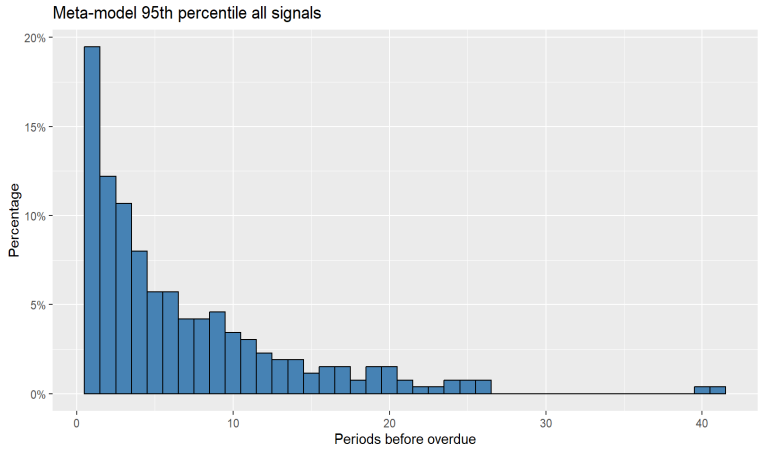


Figure 29: Meta-model 95th percentile signal distribution

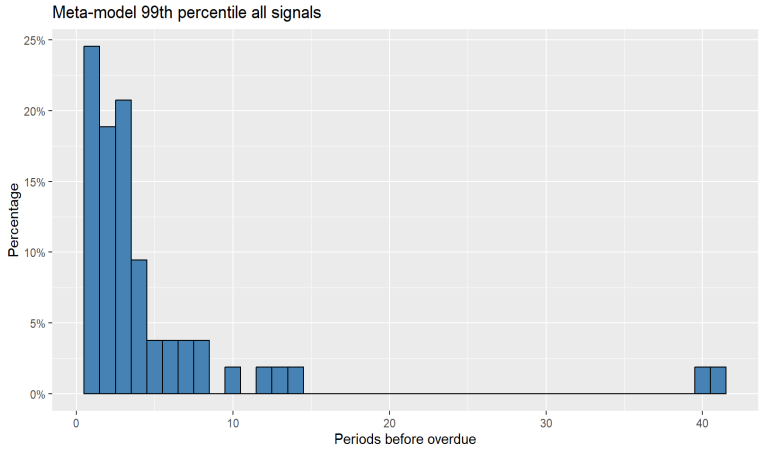


Figure 30: Meta-model 99th percentile signal distribution

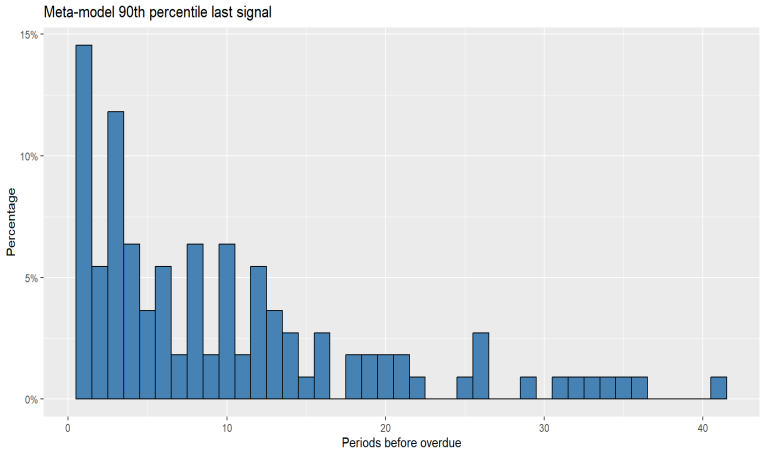


Figure 31: Meta-model 90th percentile last signal

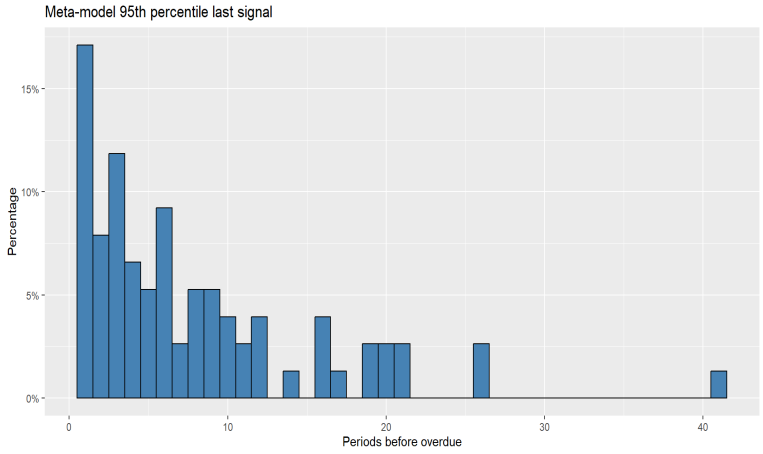


Figure 32: Meta-model 95th percentile last signal

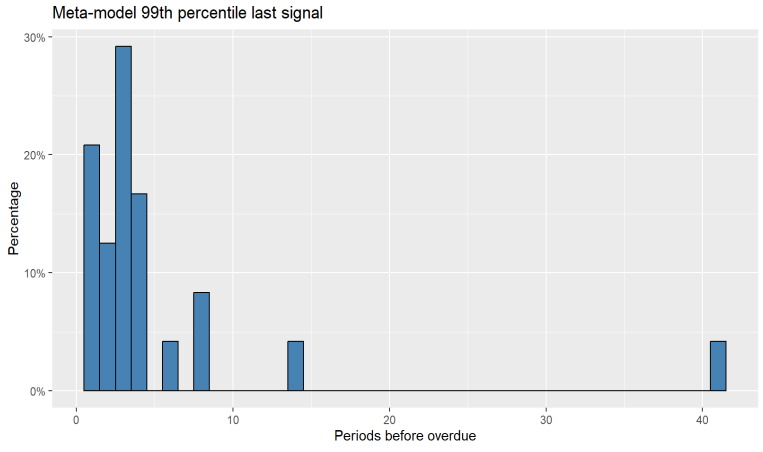


Figure 33: Meta-model 99th percentile last signal