TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Artjom Pahhomov 211744IAPM

INTRODUCING AN EXPLAINABLE AI MODEL FOR LOAN ELIGIBILITY PREDICTION: A CASE STUDY OF AS LHV PANK

Master's Thesis

Supervisor: Sven Nõmm
PhD

Co-supervisor: Indrek Arandi
MSc

Tallinn 2023

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Artjom Pahhomov 211744IAPM

LAENUKÕLBLIKKUST HINDAVA XAI MUDELI JUURUTAMINE AS-I LHV PANK NÄITEL

Magistritöö

Juhendaja: Sven Nõmm
PhD

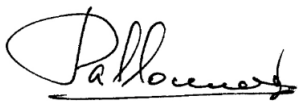
Kaasjuhendaja: Indrek Arandi
MSc

Tallinn 2023

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature, and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Artjom Pahhomov



08.05.2023

Abstract

This thesis addresses the development and integration of a customer loan eligibility prediction model based on machine learning and artificial intelligence principles to improve efficiency, reliability, and accuracy in the decision-making process, which is a crucial topic for AS LHV Pank and other financial institutions. The novelty of this research lies in the analysis of financial behaviour through the examination of account statements, providing a unique perspective on loan applicants' financial habits and patterns. Acknowledging the financial industry's need for transparency to maintain trust and compliance, this research innovatively incorporates eXplainable Artificial Intelligence (XAI) techniques to enhance the model's interpretability and ensure transparency.

The primary research questions focus on the feasibility of loan decision-making based on account statements, the selection of suitable algorithms for financial behaviour analysis, and ensuring the transparency and interpretability of the AI model. The proposed methodology follows a standard machine learning framework, from data preprocessing and feature selection to validation and XAI technique implementation. A minimum viable product is developed to demonstrate the viability of incorporating the XAI model into the existing business process, while considering extensibility, efficiency, and other relevant metrics. The obtained results are validated by quantitative evaluation of the model's performance and XAI technique effectiveness, end-user feedback, and adherence to initial requirements and regulations.

The thesis is written in English and is 73 pages long, including 6 chapters, 28 figures and 7 tables.

Annotatsioon

Laenukõlblikkust hindava XAI mudeli juurutamine AS-i LHV Pank näitel

Käesolev lõputöö käsitleb laenukõlblikkust hindava prognoosimudeli väljatöötamist masinõppe ja tehisintellekti abil ning selle integreerimist olemasolevasse äriprotessi, eesmärgil tagada efektiivsem, usaldusväärsem ning täpsem otsustusprotsess, mis on aktuaalne teema nii AS LHV Panga kui ka muude finantsasutuste jaoks. Uuring keskendub laenutaotlejate finantskäitumise, tarbimisharjumuste ja -mustrite analüüsimisele pangakonto väljavõtete abil, mida peetakse käesoleva töö peamiseks uudsuse komponendiks. Arvestades finantsvaldkonna vajadust läbipaistvate ning usaldusväärsete lahenduste järele, kaasatakse uuenduslikult käesolevas töös seletusliku intellektitehnika (XAI) põhimõtteid, parendamaks mudeli tulemuste tõlgendatavust.

Magistritöö peamisteks eesmärkideks on teostada laenuotsuste ennustamine pangakonto väljavõtete põhjal, tuvastada sobiv algoritm probleemi lahendamiseks ning tagada mudeli läbipaistvus ja tõlgendatavus. Töös kasutatav metoodika järgib standardset masinõppe raamistikku, alates andmetöötlusest ja tähtsamate tunnuste valikust kuni valideerimise ja XAI-tehnikate rakendamiseni. Tõestamaks mudeli rakendatavust reaalsetes äriprotsessides, arendatakse välja minimaalne elujõuline toode (MVP) ning seejärel integreeritakse see olemasolevasse infosüsteemi, arvestades lahenduse laiendatavuse ja efektiivsusega. Saadud tulemusi valideeritakse kasutades mudeli jõudlust ja XAI-tehnikate efektiivsust kirjeldavaid kvantitatiivseid mõõdikuid ning nõuetele vastavuse ja lõppkasutaja tagasiside abil.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 73 leheküljel, 6 peatükki, 28 joonist, 7 tabelit.

List of Abbreviations and Terms

AI	Artificial Intelligence
ADASYN	Adaptive Synthetic Sampling
API	Application Programming Interface
AUC	Area Under Receiver Operating Characteristics Curve
CDS	Credit Decision System
CNN	Convolutional Neural Networks
DL	Deep Learning
DT	Decision Tree
DWH	Data Warehouse
HTTP	Hypertext Transfer Protocol
LIME	Local Interpretable Model-Agnostic Explanations
k-NN	k-Nearest Neighbors
ML	Machine Learning
NN	Neural Networks
PCC	Percentage Correctly Classified
REST	REpresentational State Transfer
RF	Random Forest
ROSE	Random Over-Sampling Examples
SMOTE	Synthetic Minority Over-sampling Technique
SHAP	SHapley Additive exPlanations
SQL	Structured Query Language
SVM	Support Vector Machines
US	Undersampling
XAI	eXplainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Context

Improving internal processes and mitigating potential risks is a relevant issue for AS LHV Pank and many other companies operating in the financial field. As loans are typically one of the main sources of revenue for the banking industry, it is essential to improve the decision-making process by making it more efficient, reliable, and accurate from the company's point of view. Numerous types of research have shown that certain improvements could be achieved by applying solutions based on the principles of machine learning and artificial intelligence [1, 2, 3].

When approached without due consideration, critical concerns arise: the powerful models frequently operate as "black boxes", where even the researchers who create them face difficulties in fully grasping the mechanisms behind their predictions [4]. Considering the unique characteristics of the financial sector, this problem is of significant magnitude and must be addressed during the development of machine learning-based models. In response to this challenge, Explainable Artificial Intelligence (XAI) techniques have garnered considerable academic interest, as they aim to improve the explainability of machine learning models and facilitate a more comprehensive understanding of their output [4, 5].

1.2 Problem Statement

In this thesis, the main objective is to investigate the development of a customer loan eligibility prediction model and integrate it into an existing business process. The proposed model should possess the capability to analyse the financial behaviour of a loan applicant based on their account statement, and provide an explainable prediction of the decision regarding a given application. Initially, the model is intended to serve an advisory role in the loan decision-making process, implying that its verdict would not be decisive. Instead, it would offer supportive recommendations to help decision makers. The need is dictated

by the strategic roadmap of the company: machine learning is viewed as a tool that might expand the application throughput and speed up the decision-making phase, further automate the process, and simplify it for the employees, with an explainable output being a key criterion, particularly in the financing industry.

Furthermore, the model must be used by other internal applications automatically, and, according to the regulations and specifics of the domain, its usage must also be traceable subsequently. Therefore, integration is another crucial topic that should be addressed in this master's thesis. The model needs to be provided as a microservice to ensure that other applications can access it whenever an account statement is sent as a request. The financial industry regulations must also be accounted for: all data sources that were utilised to make the decision, including all iterations of the aforementioned model, must be retained. The ability to automatically perform the analysis of the account statement using a machine learning-based model would improve the speed, accuracy, and efficiency of decision-making for all parties involved in the process.

The use of explainable AI principles is the primary element of originality; the model must allow users to understand how the model achieved the particular result. Given that the primary function of the model is that of an advisor, it must identify the factors to which a bank employee should pay attention when considering a particular loan applicant.

In summary, the primary objective of this thesis is to address several key questions that will contribute to developing an efficient solution to the stated problem.

- Investigate the feasibility of making loan decisions based solely on account statements, and assess the need of incorporating additional customer data features.
- Identify the most appropriate algorithm or ensemble of algorithms for financial behaviour analysis.
- Explore the integration of the AI model into the existing business workflow.
- Examine strategies to enhance the interpretability and transparency of the AI model for its users.

1.3 Methodology

The approach used to create a predictive model is rather standard in the field of machine learning. Research conducted by the University of Witwatersrand School of Computer Science and Applied Mathematics [6] concludes with a proposal of a specific framework (or workflow) that could be used as a guideline for developing machinery in the field of financing (see Figure 1). The proposed workflow is substantiated by the statistical evidence and closely aligns with classical methodology, and therefore shall be referred to while approaching the problem. First, the data must be preprocessed and then analysed to detect and extract the features that influence the model outcomes the most; these steps are known as feature selection and feature engineering. As a next step, based on the available materials and literature, the algorithms will have been decided that will further be used to experiment with model building. Then, the validation is conducted by evaluating and comparing algorithms' performances using statistical model goodness metrics. Finally, various XAI techniques will be analysed, compared and applied to increase model transparency.

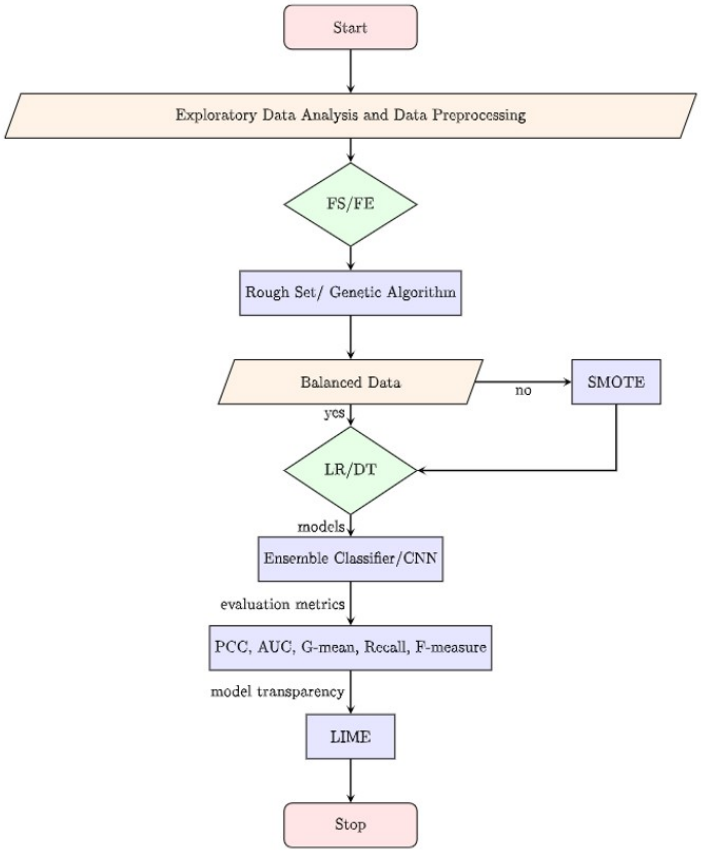


Figure 1. Guiding machine learning framework for credit scoring [6].

To solve the subproblem of integrating the resulting model into the current process, a minimum viable product (MVP) will be developed to provide a tangible proof of the model's feasibility within the existing business workflow. As the solution is being designed for use in a large company, the current infrastructure state, used technologies, and other metrics frequently applied in software development (such as extensibility, efficiency, speed, etc.) shall be taken into consideration. The resulting product is expected to be ready for a seamless incorporation with an internal Credit Decision System.

1.4 Work Structure

The background section of present thesis seeks to provide details required to understand the current situation: the state of the industry in general and the company in particular; the intent and starting position of the current project. An analysis section consists of a concise overview of the machine learning process, its necessary steps, and the explanation of the algorithms that are being considered in the thesis; furthermore, the most important aspects and peculiarities of financial data under analysis are described. The section explaining the proposed solution for the problem contains the description of three main subprocesses: building the model using machine learning techniques, integrating the XAI techniques, and introducing the solution into the loan decision-making process. The emphasis of the final section is on assessing the outcomes attained, specifically:

- Evaluating the model's performance by applying chosen metrics and comparing the values to the established standards;
- Examining the effectiveness of the implemented XAI techniques through both quantitative and qualitative evaluation approaches;
- Conducting a social evaluation by collecting feedback from end users to understand the model's trustworthiness and efficacy;
- Ensuring the developed solution adheres to the initial requirements and complies with relevant regulations.

2 Background

2.1 Loan Eligibility Assessment

Loan eligibility assessment is a process conducted by financial institutions with the goal of assessing the creditworthiness of a given client and determining their suitability. Assessing various factors and analysing data, such as credit score or other characteristics of financial behaviour, allows lenders to make informed and rational decisions on whether to approve or reject the application. This process is crucial for the institutions, as it facilitates risk mitigation, enables sustainable profitability, and safeguards against potential default.

A comprehensive overview of loan eligibility processes can be challenging to obtain, as the specific methods used are often considered proprietary information by financial institutions. Based on the author's personal experience in the financial sector, lending criteria and assessment methods may vary significantly across different companies. This diversity is also reflected in credit scoring, one common approach to evaluating borrower creditworthiness. The multitude of credit scoring models and techniques highlights the complexity and adaptability of credit assessment in response to changing customer profiles, economic conditions, and industry standards [7]. Therefore, it is essential for lending institutions to recognise the variability of loan eligibility evaluation approaches and develop approaches to improve credit risk assessment and decision-making that are well suited to achieve the goals set by a specific company.

2.2 Related Work

The use of machine learning in banking and finance is not a new concept; therefore, multiple literature and research reviews have been conducted to provide the reader with an overview of various techniques and approaches. The author believes that a review of the literature done at the University of Witwatersrand School of Computer Science and Applied Mathematics in Johannesburg is particularly valuable because it provides an incredibly

thorough and well-written overview of the 74 studies published in the 2010s [6]. The selection of the best method for model creation is a significant component of all AI-related activities; thus, it is vital to analyse the tools commonly applied in a certain industry. Compared to a 2016 literature survey that included approximately 150 articles over a period of almost two decades [8], one can see a clear shift towards more sophisticated machine learning algorithms – such as convolutional neural networks (CNN), and deep learning (DL) – and/or ensemble classifiers, as they seem to outperform more traditional single classifiers [6, 8]. The ability of CNN and DL to identify characteristics that are more discriminative between borrowers explains their high performance. However, it is believed that the lack of transparency in DL models is both their primary issue and the main reason why they are used less frequently.

In general, the most popular techniques used in credit scoring and loan approval are support vector machines (SVM), neural networks (such as CNN or DL), and various ensemble classifiers, with the popularity of deep learning already increasing [6, 8, 9]. The quality of AI credit scoring models is often validated using German and Australian credit datasets; and referring to the performance of the algorithms, by comparing the PCC (Percentage Correctly Classified) and AUC (Area Under Receiver Operating Characteristics Curve) of the results of the reports that were validated with those credit datasets, it can be observed that the quality of the model is approximately the same in all cases (see Figure 2). Boosting algorithms, such as Gradient Boosting, have shown slightly better results, since the logic behind such classifiers is to combine several AI models into a more powerful machinery.

The study by the University of Witwatersrand concludes that there are two main problems regarding the current state: the problem of imbalanced datasets and the inability of major fraction of applied machine learning models to explain their predictions [6].

2.3 Company Context and Processes

LHV provides a wide range of loan products to meet the diverse needs of its customers. Within the scope of this thesis, however, the main focus is on loans offered by AS LHV Finance, a subsidiary that specialises in smaller loans. The primary loan types within this category are hire-purchase and consumer loans. Hire-purchase is a financing solution

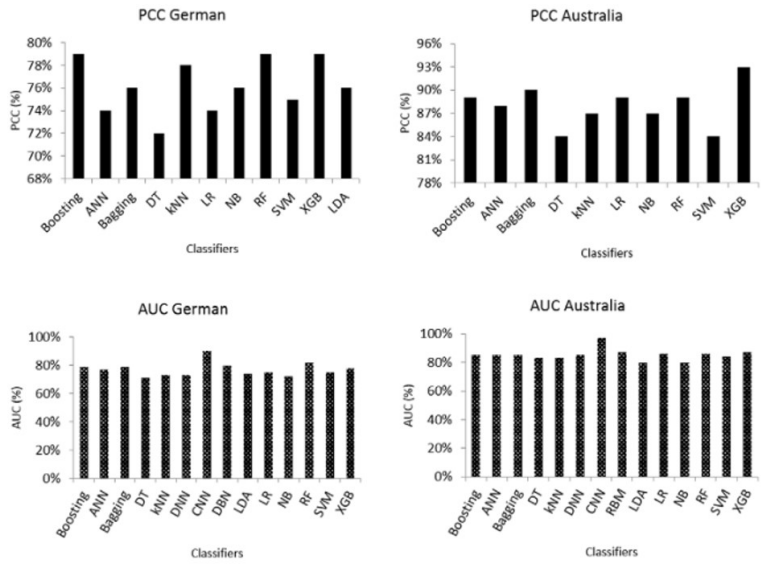


Figure 2. Average performance of algorithms on German and Australian credit datasets [6].

that enables individuals to obtain goods immediately while spreading the total cost over a predetermined period, resulting in manageable monthly payments; this product is accessible to customers through collaborations with various vendors, including e-commerce platforms and on-site stores. On the contrary, consumer loans are designed for significant purchases in one time, such as vehicles and appliances, or to secure funds for home improvement projects.

Individuals seeking a loan must navigate a detailed application and eligibility assessment process. Upon submitting the loan application, it is directed to the Credit Decision System (CDS), where it undergoes evaluation through an automated decision tree. This tree comprises a series of step-by-step controls, commonly referred to as checks, that assess the viability of the application. The automated process in CDS can yield two potential outcomes: an automatic positive or negative decision (approval or rejection), or a redirection to manual control. In the event of manual control, loan managers effectively evaluate the application by hand. They may contact the applicant if further clarification is required or if additional steps must be taken to satisfy the eligibility criteria. If all necessary conditions are met, the loan manager presents an offer to the applicant. The overarching goal of CDS and loan managers is to conduct an efficient decision-making process, delivering an initial offer or the final decision to the client as quickly as possible.

2.4 Available Data Description

Initially, applicants are required to provide extensive personal information, including the number of national identification documents, educational background, employment status, and the number of dependent individuals. In addition, financial data, such as monthly income and liabilities, must be submitted for consideration. The financial data provided, however, is also validated and compared with actual account statements. Depending on business rules, the client may be asked to provide an account statement to further substantiate their financial position. In instances where the client's primary bank is not LHV, manual submission of an account statement is required; otherwise, the account statement is fetched automatically from other internal systems. In cases where multiple accounts are held, statements from each account may be requested to ensure a comprehensive understanding of the applicant's financial circumstances.

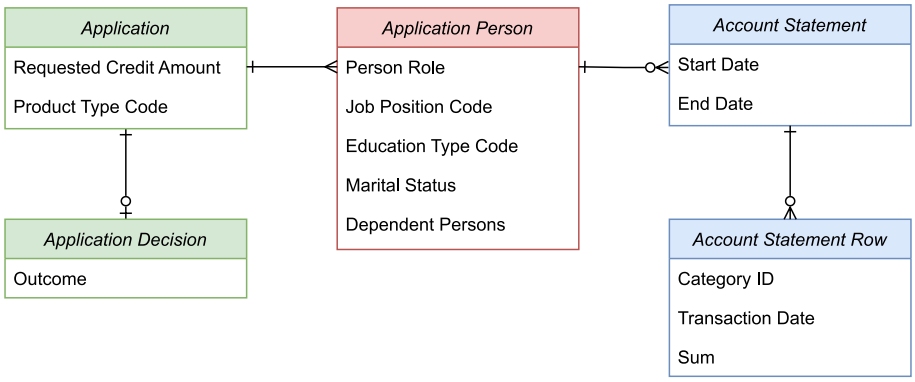


Figure 3. Simplified Entity Relationship Diagram visualising the structure of data under analysis.

Transaction data derived from account statements is systematically organised into a hierarchical structure of categories and subcategories. The categorisation process, conducted by a partner service specialised in account statement analysis, involves assigning each transaction in the account statement to a specific category, represented by a category identification number, and subsequently linking it to a broader parent category. Although the account statement data are divided into more than 500 unique transaction categories, parent categories offer a higher level of abstraction by consolidating related expenses. For example, individual expenses such as heating, security and electricity are subsumed into the overarching parent category of utility services expenses. This categorisation framework not only enables efficient data management and organisation, but also facilitates a more

meaningful interpretation of spending behaviour and trends across all the financial aspects while evaluating loan application.

2.5 Regulatory and Ethical Considerations

Adherence to ethical standards is crucial in the world of financial technology because it promotes trust and guarantees adherence to legal requirements. According to an article on ethically responsible machine learning in financial technology, the main ethical principles in finance are:

- integrity and objectivity,
- competence and fairness,
- confidentiality,
- professionalism and diligence [10].

Those principles can be kept by applying explainable artificial intelligence (XAI) techniques [10]. Furthermore, several regulations and agreements, such as Basel III agreements [11], require or promote credit scoring algorithms to be transparent and explain their decisions. However, as pointed out in the review, only a small fraction of studies (around 8%) have considered incorporating transparency and XAI techniques into the workflow [6]. With the observation of this serious drawback in numerous studies, this issue should certainly be addressed and thoroughly investigated in this particular thesis.

In the financial industry, privacy is of utmost importance, especially when working with private consumer information, such as account statements. Thus, the use of anonymised data is essential in the creation and use of credit risk models. In the context of given problem, anonymousity of the data refers to the removal of personally identifiable information, such as customer names, bank account numbers and transaction descriptions, as well as identification numbers. This approach will protect sensitive data and ensure that the AI model does not acquire unnecessary bias.

In addition to ethical concerns, there are also multiple regulatory aspects to consider. For instance, Estonain Creditors and Credit Intermediaries Act obligates financial institutions

to provide adequate and sufficient decision explanations to the borrowers [12]. Another Estonian law, the Money Laundering and Terrorist Financing Prevention Act, requires financial institutions to maintain comprehensive records and cooperate with law enforcement agencies or supervisory authorities [13]; therefore, all information used to evaluate loan applications should be retained and provided when requested by authorities. When applying for a loan, the customer grants permission for processing according to the LHV Principles of Processing Customer Data [14]. Thus, the analysis of customer data and implementation of the AI model in this study shall be conducted in strict compliance with aforementioned guidelines and regulations.

3 Analysis and Methodology

This chapter aims to offer an analysis of the technologies and methodologies integral to the addressed problem. The objective is to establish a solid foundation for consecutive research and a thorough theoretical understanding of the topic.

3.1 Machine Learning

Machine learning is a type of artificial intelligence that involves the use of algorithms and statistical models to allow computer systems to learn from data and make predictions or decisions based on this learning, without being explicitly programmed [15]. Real-world data can be too complex to comprehend; therefore, large datasets can be used to train machine learning algorithms that use mathematical and statistical methods to find patterns and relationships in the data. This makes situations where the data instances have a lot of hidden correlations that are not always obvious very well-suited for machine learning. Because of its capacity to analyse and make sense of large datasets, especially now that the amount of data is increasing exponentially, this approach has attracted a lot of interest in recent years. Depending on the approach, machine learning can be classified into many groups, the following being the most relevant to solving the problem.

- *Supervised Learning*: This method uses labelled data, where the desired output is known, to train the machine learning model. Based on the examples presented in the training data, the model learns to map inputs to outputs [15].
- *Unsupervised Learning* With this method, a machine learning model is trained on unlabelled data, as the desired output is not known. Without any explicit instruction, the model discovers links and patterns in the data on its own [15].
- *Reinforcement Learning*: The machine learning model shall learn by interaction with the environment and collecting feedback in the form of rewards or penalties, with an objective to maximise the overall reward over time [16]. In a given context, the goal could be to minimise the number of defaults or to maximise the profit from loans.

Each of the techniques discussed could potentially instruct a model to identify loan applications that should be rejected; however, the decision was made to employ supervised learning as the main vector of methodology. This approach allows for the utilisation of historical data, which is advantageous due to the method's capacity for a structured training process and its use of labeled data to investigate a precise correlation between inputs and outputs.

3.1.1 Workflow

A general consensus on the machine learning framework has been achieved in most publications, articles, and researches, despite some slight differences in the way the workflow should be divided into key steps. After familiarising with corresponding literature and articles [17, 18, 19] author believes that the workflow can generally be divided into following steps:

1. Data collection: gathering the required information from numerous sources to train the model, which involves selecting the appropriate data sources, obtaining the data, and organising it in a usable form for further steps.
2. Data exploration and preprocessing: analysing and comprehending gathered data, cleaning, and transforming it into a format suitable for analysis, e.g. handling missing values, dealing with outliers, etc.
3. Feature engineering: extracting relevant features from the data so that model can perform better, e.g. choosing the most significant features, producing new features through transformations, and lowering the dimensionality of the data.
4. Model selection: selecting the model architecture or algorithm that best addresses the issue at hand while considering the nature of the problem, the available data, and the intended performance standards.
5. Model training: acquiting a model with the patterns in training data so that, given the validation data, it could make correct predictions or return the expected outputs as frequently as possible.
6. Evaluation: measuring how well the model generalises to new data by assessing the trained model's performance on a set of validation data by using evaluation metrics.

7. Model deployment: incorporating the trained model into a program or system to make predictions on fresh data, which entails packaging the model in a way that other software elements can use it, and providing an API for users to communicate with the model.

The academic work structure described above is generally suitable for achieving the goals of this thesis. However, all of the aforementioned sources do not explicitly include the additional step of explainability analysis. Explainability analysis uses XAI methodologies to explain how the model works and the causes of its predictions: e.g. feature importance analysis, various model-agnostic techniques, or the creation of understandable models. This is essential for solving the problem stated in this thesis, and thus, will be included in the workflow and precede the model deployment step.

3.2 Shortcomings & Challenges

In data analysis, every domain of origin comes with its own characteristics and peculiarities, and according to some studies [20, 21], financial data often pose several challenges for machine learning algorithms. The author therefore believes that the most severe shortcomings and bottlenecks of financial data include:

- Weak predictability: economic growth, and hence financial market behavior, is difficult to predict; financial data has low "signal-to-noise ratio", and there are a lot of variables that impact the performance [20].
- Data volume and quality: the availability of data in finance is limited, compared to other fields – financial data cannot be synthesised, and, furthermore, the level of noise can make modelling complicated and imprecise [21].
- Evolving markets: financial markets and trends are adaptive by nature, they rapidly evolve and overall have a dynamic character [20].
- High dimensionality: complex financial data, including extensive financial statements, have many dimensions, which results in more complicated feature extraction and more complex and resource-intensive machine learning models.
- Interpretability requirement: since financial data is frequently used to impact critical business decisions, it must frequently be comprehensible and transparent [20, 21].

3.2.1 Dataset Balancing

When significant disproportions exist between the classes in a classification problem, this is known as an imbalanced dataset issue [22]. For instance, in a scenario that involves the identification of credit card fraud, the proportion of fraudulent transactions to non-fraudulent transactions is substantially smaller – this may result in a biased model that does not adequately represent the minority class. Although not as severe an issue as in the provided example above, the imbalanced dataset problem is also present in the data under analysis, where the rejected applications make up only about 30-40% of total number of applications, and thus this issue needs to be accounted with.

Traditionally, the methodologies for addressing imbalanced data are categorised into three predominant groups [23]:

- Data-level methods: modifying the collection of examples to balance distributions and/or remove certain samples.
- Algorithm-level methods: directly modifying existing learning algorithms to alleviate the bias towards majority objects and adapt them to mining data with skewed distributions.
- Hybrid methods: combine the advantages of two previous groups.

Data-level techniques are a common approach to solve this issue, which involve altering the dataset to produce a more even distribution of classes. Techniques like oversampling, undersampling, and data augmentation can be used to accomplish this [24, 25]. Oversampling implies generating new synthetic instances using methods – e.g., SMOTE, ADASYN, and ROSE – or duplicating the existing data points to provide fresh examples for the minority class; one particularly popular group of algorithms are SMOTE-based (Synthetic Minority Over-sampling Technique) approaches that generate synthetic examples by interpolating between existing minority class examples [26]. Contrarily, undersampling seeks to limit the number of instances from the majority class by discarding some samples at random or choosing the samples that are most representative [24]. Data augmentation involves modifying the existing examples in the dataset by applying transformations or adding noise to create new variations of the same example, which can help to increase the diversity of

the dataset under analysis [25].

The most popular algorithm-based technique is cost-sensitive learning: this method involves assigning various costs (weights) to different classes or samples, resulting in a higher penalty for misclassifying the minority class than the majority class [24]. The main advantage of this method is that no data are lost during the process, nor are synthetic data introduced as well. Some machine learning algorithms, such as decision trees and ensemble methods, have options to assign greater importance to minority class samples during training [22].

For imbalanced financial data, cost-sensitive techniques that assign different misclassification costs to different classes can be useful and have remarkable performance results [27]. However, the effectiveness of cost-sensitive methods depends on the particular issue and dataset at hand, and other strategies such as ensemble methods and data-level strategies may also be useful. Regarding the oversampling approach, in one study, it was noted that, although the methods perform fairly well with unbalanced financial data sets, the effect of noise examples could also be amplified [28].

3.3 Machine Learning Algorithms

There are numerous different algorithms from which to select in order to address the problem using machine learning. Classification and regression are the two categories into which all methods can be split: regression techniques forecast continuous numerical values, whereas classification algorithms categorise data into distinct categories or classes depending on the attributes that are present. This section will concentrate on exploring several classification approaches that may be used to categorise data into different groups or classes based on their properties because the current issue requires the labelling of data.

3.3.1 Overview

According to numerous sources that seek to provide a systematic overview of supervised machine learning methodologies [29, 30, 31], classification approaches can be broadly categorised into five major groups, based on the logic behind.

- Instance-based learning: predicting the result by the similarity between new and stored training data, without grasping a generalised model of the data [30]. The most well-known instance-based algorithm is *k*-Nearest Neighbor classifier.
- Logic-based learning: solving problems sequentially or incrementally by applying logical (or symbolic) functions [31]. Decision trees and rule-based classifiers are prominent examples in this category.
- Probabilistic learning: utilising distributive statistics and functional analysis to determine a probabilistic relationship [29, 31]. This category includes such examples as Naive Bayes Classifier, and Logistic Regression [29].
- Support Vector Machine: identifying and determining the best decision boundary or hyperplane to optimise the separation of different classes maximally [29].
- Neural Networks and Deep Learning: mimicing the functioning of the human nervous system by constructing a network of connected cells (or neurons) to recognise complex patterns in data [29].

Due to its capacity to increase the predictability and stability of machine learning models, ensemble classifiers have emerged as a quite well-liked methodology. To produce a more reliable prediction, ensemble classifiers aggregate multiple models, frequently with varying strengths and weaknesses.

The author has chosen to proceed with a select few algorithms based on their demonstrated ability to perform tasks of a similar nature and their general popularity, as revealed by the extensive research carried out in the School of Computer Science and Applied Mathematics of the University of Witwatersrand [6]. The following section will provide a concise summary and technical description of each supervised learning algorithm that is being considered to solve the problem stated in the thesis.

K-Nearest Neighbors

K-Nearest Neigbor (k-NN) is a straightforward algorithm that is fairly easy to comprehend. It is used when there is little prior knowledge of the distribution of the data [31]. Given a new data point, k-NN locates the k closest data points in the training set and predicts the class based on the average value or majority class of the k-nearest neighbors [29].

The neighbors are detected using a provided distance metric (e.g., Euclidean, Manhattan, Mahalanobis distance). Despite the simplicity of the approach, it has been shown to achieve astonishing results in complicated data, as found in the study [31].

Decision Tree

Decision trees (DT) are one of the most important approaches in data science, because of their ability to produce intensive results, organised structure, and natural interpretability [30]. The algorithm generates a tree-structured model with decision nodes that are arranged in a hierarchy according to the properties of the input data. Each leaf (or dead end) node is a predicted outcome, whereas each internal node, known as the split criterion, has a condition on one or more feature variables that splits the data into two or more branches; the goal of each split conducted is to maximise the separation of the different classes among the children nodes [29]. The decision tree works quite well with the datasets where determining the split boundary is more straightforward, e.g. datasets with discrete categorical features.

Support Vector Machine

A techique called Support Vector Machine (SVM), which is naturally defined for binary classification, seeks to solve a problem by finding an optimal hyperplane separating the data points from different classes and maximising the region between the boundary and data (i.e. margin) [29]. Despite most datasets being initially non-separable, SVM solves the separation problem by projecting the data points onto a higher dimensional space where the dataset becomes separable [30]. This is achieved by employing a variety of optimisation approaches, such as *kernel trick* [29]. SVM is therefore a strong and efficient technique to deal with high-dimensional data and is capable of addressing non-linear decision boundaries [31].

Random Forest

Random Forest (RF) is an ensemble learning algorithm, and is a well-known example of bagging (or bootstrap aggregating), i.e. creating multiple weak classifiers models on various subsets of the data in parallel, and then combining their predictions [29, 32]. In this approach, the training data is randomly divided into multiple subsets of randomly chosen

features – abovementioned two methods of including randomness enable it to handle a wide range of datasets and surpass a single decision tree in terms of performance [32]. A model is then constructed for each subset using decision tree algorithm, and the final prediction for a new data point is given by combining the outputs from each base model. Due to its randomness, the Random Forest algorithm is also claimed to be resistant to noise and outliers [29].

XGBoost

Extreme Gradient Boosting, or XGBoost, is another example of ensemble learning algorithms that has gathered significant attention in recent years. XGBoost is, similarly to Random Forest, based on the decision trees, however instead of bagging this algorithm utilises boosting principle [33], i.e. builds weak models sequentially by assigning more weight to previously incorrectly categorised samples to increase overall accuracy [29]. When combining weak classifiers, this algorithm uses a gradient descent technique with the objective of minimising the loss, hence the name [33]. XGBoost has been proven to be successful in solving complicated problems in finance, including credit risk assessment [6] and banking [32].

Neural Networks

Neural networks, inspired by the human nervous system, aim to mimic the learning process observed in biological neurons by adjusting connection weights among artificial neurons in response to external stimuli [29]. In order to learn patterns and make predictions or judgements based on the input data, neural networks process the data through interconnected layers of artificial neurons with configurable weights, and then provide an output. The strongest property of this approach is the diverse and flexible architecture: there is a wide variety of architectures, ranging from a simple single-layer perceptron to complex multilayer networks [29]; deep learning, a subset of neural networks, has gained more popularity in recent years, as it leverages multiple hidden layers to enable improved learning capabilities [30]. Due to neural networks' adaptability, capability to handle large-scale data, and handling both linear and non-linear models, they are highly efficient in solving complex problems, including those in finance, such as fraud detection and credit scoring [6, 32].

3.3.2 Comparison

The following table (see Table 1) aims to provide a comprehensive comparison of various machine learning algorithms, analysing their individual strengths and limitations.

Table 1. Comparison of machine learning algorithms.

Method	Strengths	Limitations
k-NN	■ The simple and intuitive implementation ■ Similarity-based approach can perform well on imbalanced data [29, 34]	■ Distance-based approaches can be ineffective for high-dimensional data ("*curse of dimensionality*") [29] ■ Relatively high computational expense for large datasets [34] ■ Challenging to choose k, as small values risk noise interference and large values reduce sensitivity to data locality [29]
DT	■ Easily comprehensible and interpretable by design ■ Ability to handle both numerical and categorical features [29, 34] ■ Application requires low computational power [35]	■ Known to be relatively sensitive to minor changes in data ■ May be a subject to over-fitting in case of noisy data or when the dataset is not representative enough [35] ■ Often yields a locally optimal solution, not globally optimal one [34]
RF	■ Reduces the decision tree over-fitting problem [34] ■ More robust towards noise and outliers in the dataset [29] ■ Commendable performance in high-dimensional space [29]	■ Computationally and temporally more expensive than a decision tree [31, 34] ■ Harder to interpret compared to a decision tree
SVM	■ Potentially effective in high-dimensional space ■ Capable of solving nonlinear classification problems [35] ■ Less vulnerable to overfitting and aims to achieve globally optimal solution [34]	■ Computationally expensive for large datasets [34] ■ Difficult to understand results due to input transformation [35] ■ Noise in the data potentially degrades performance [31, 34]
XGBoost	■ Effectiveness, high performance and accuracy across various domains [6, 31] ■ Manages to handle the imbalance of data by employing regularisation [33]	■ Being an ensemble algorithm, requires high computational power and longer training time [31, 34] ■ Less interpretable than a single decision tree

Method	Strengths	Limitations
NN	■ Ability to efficiently solve complex non-linear problems with non-contiguous distributions [29] ■ Performs well on a vast amount of data in high-dimensional space ■ Due to flexibility is considered to be a universal approximator [29]	■ Requires a large amount of data to perform the generalisation ■ Relative complexity of optimising the parameters and network structure [29] ■ Requires high computational power for training and operation [35] ■ Potential for over-fitting if data is not representative enough or contains too much noise [29, 35]

Among the discussed algorithms, Random Forest (RF), Support Vector Machines (SVM), and XGBoost demonstrate promising potential for financial data analysis, particularly in the context of account statement with numerous features. The ensemble methods of RF and XGBoost are adept at mitigating the overfitting issue prevalent in decision trees, while SVM excels in high-dimensional feature spaces while also maintaining a lower risk of overfitting. Additionally, XGBoost's ability to handle imbalanced data through regularisation makes it well-suited for financial applications where certain classes might be under-represented. Although Neural Networks (NN) have demonstrated success in a variety of domains, their implementation in this context may be inappropriate due to their inherent complexity and demanding computational requirements; therefore, this study will not delve into the more advanced Deep Learning (DL) approaches.

3.4 Explainable AI

The term *eXplainable Artificial Intelligence* (XAI) refers to the creation of AI systems that can give human users clear and concise justifications for the decisions these models make. The ability to explain AI-based judgments is critical for regulatory compliance and user acceptance in many application areas where it has substantial effects on specific people and society at large: for example, in precision medicine, where far more information is required than a simple binary prediction; other examples include autonomous vehicles in transportation, security and finance [5]. In other words, interpretability of a model by using XAI is way to become trustful for an end-user [36]. To increase the understandability and transparency of AI models, researchers are constantly coming up with new methods and strategies in the fast developing subject of XAI.

The literature clearly distinguishes between models that are interpretable by design and those that can be understood using external XAI approaches: models that can be understood by themselves are called *transparent*, while the latter approach is referred to as *post-hoc explainability* [5]. The usage of a certain subset of algorithms capable of creating fully transparent models is the easiest and most confident way to achieve complete interpretability [37]. Algorithms that create those models are, for example, decision trees, *k*-Nearest Neighbors, Naive Bayes, rule-based algorithms, and linear/logistic regression. Because they are simple to understand and can be used to provide explicit rules for decision making, these models can be used in applications where interpretability is a crucial need. Nevertheless, they might not always perform as well as more complex, black-box models.

3.4.1 Overview of Methods

Based on the range of applicability, the XAI techniques are divided into model-agnostic and model-specific categories [5]. Model-specific approaches are created for specific models or algorithms and may offer more accurate and exact explanations but are restricted to a certain class of models. On the other hand, model-agnostic techniques work with any machine learning model, typically by examining feature input and output pairs; these methods tend to be more broad and interpretable but may not take advantage of the specific qualities of a model, such as structural data or weights [37].

An additional criterion for classifying XAI methods is the scope of interpretability [37]. Global interpretability in XAI involves understanding the behavior of a machine learning model as a whole, based on a comprehensive view of its features and each of the learned components. In contrast, local interpretability involves understanding its behaviour at the level of individual predictions. Local interpretability techniques aim to identify the features or inputs that contributed the most to a particular prediction. Local explanations emphasise making individual predictions more interpretable and therefore can be more accurate than global explanations [37].

When developing an interpretable AI model, a wide range of XAI techniques can be considered. However, the context of the current thesis enforces certain constraints on the considered approaches.

- Need for a model-independent approach: the algorithm used for model building is not fixed and might be subject to change in the future.
- Local interpretability requirement: it is expected of the service to offer an explanation for every single prediction.
- Technique reliability: the bank is a rather conservative organisation, and the use of methods that have undergone extensive study and testing is essential.

As a result, two popular local model-agnostic methods, LIME and SHAP, were selected for further comparison.

LIME

Local Interpretable Model-Agnostic Explanations (LIME) is a post-hoc and model-agnostic technique for explaining AI model predictions, which underlying principle is to create a locally faithful and interpretable surrogate model for explaining the decisions of the black-box model [36]. The algorithm to generate the explanation is as follows:

1. create a dataset of instances that are similar to the instance to be explained by introducing random perturbations to its input features;
2. obtain the predictions for the perturbed dataset using the black-box model, and weight the new samples according to their proximity to the point under analysis;
3. train the weighted and interpretable model (e.g., by using Linear Regression) that approximates the behaviour of black-box on the generated dataset;
4. extract the feature importance scores from the interpretable model and generate a comprehensive explanation [36, 37].

Due to the versatility in handling diverse data types, such as tabular data, text, and images, LIME has found successful applications in various domains, including financials [6].

SHAP

SHapley Additive exPlanations (SHAP) is a model-agnostic method for explaining the output of a machine learning black-box model that uses the cooperative game theory approach to quantify the contribution of each feature to the prediction. In general, SHAP

calculates the Shapley values by creating a power set of features, the set that contains all possible subsets of the attributes provided, and evaluates the model predictions for each subset [38].

While using Shapley values is the primary idea behind SHAP, the inventors of the idea have also suggested various approaches, such as KernelSHAP (kernel-based estimation approach inspired by local surrogate models [38]), or TreeSHAP (an efficient estimation approach for tree-based models [39]). In addition, SHAP provides both local and global interpretations: by combining the Shapley values obtained from every local explanation, the global explanation can be acquired, allowing parties to assess the overall behaviour of the model, and thus improve model performance [37].

3.4.2 Comparison

To select the most appropriate XAI technique for a specific problem, it is necessary to conduct a comparative analysis of the methods based on the most relevant criteria. The following section summarises the main strengths and weaknesses (or limitations) of LIME and SHAP in terms of various aspects.

- *Range of applicability*. Both LIME and SHAP are designed to be model-agnostic. However, LIME can be used with any black-box model, and the integration is seamless. Alternatively, due to the slow and complex nature of the default implementation of KernelSHAP [37], specialised techniques have been developed to optimise performance for specific algorithms [39]. However, this can result in a slightly more model-specific approach.
- *Scope*. LIME and SHAP are used to generate explanations at the local level by observing the alterations in model behaviour in a given neighbourhood. Furthermore, SHAP is also able to provide global interpretability.
- *Performance*. LIME traditionally uses a linear model to approximate the local behaviour of the model [36], which might lead to less accurate results. On the other hand, SHAP is able to generate more reliable and robust explanations due to its totally different mathematical approach. Furthermore, the explanations obtained from LIME can have a high variance even for two neighbouring instances, while

SHAP is more stable, as noted in one study [40].

- *Interpretability*. The feature importances calculated by LIME derive from the interpretable model (e.g., linear model), and therefore are more easily understandable and intuitive for non-experts, compared to the concept of Shapley values, utilised by SHAP [37].
- *Complexity*. Compared to LIME, which is generally a simpler approach and thus requires fewer computational resources to calculate, SHAP is a more sophisticated technique. Thus, LIME is a preffered choice when primary considerations are simplicity, efficiency, and speed.

On the basis of the comparison, it can be concluded that both LIME and SHAP techniques have certain advantages as well as drawbacks in terms of providing explainability. Therefore, both approaches are believed to be considered and experimented with in the implementation stage of the project.

4 Proposed Solution

This section presents the details of the implementation of a proposed solution to the stated problem. The proposed solution follows a typical machine learning pipeline, including exploratory data analysis, data preparation, model building, and enhanced by applying explainable artificial intelligence (XAI) methods and deployment techniques. The entire process is carried out using Python – one of the most popular programming languages among data scientists, that provides a rich set of required libraries and tools.

4.1 Data Acquisition

As stated in Section 2.5, the company's area of business requires that all the information used to evaluate loan applications and make decisions is maintained and backed up. Furthermore, besides the system-specific databases, the company additionally utilises a central repository of information principle known as data warehouse (DWH). The abovementioned system, among others, is actively used by data analysts. All the information is stored in the relational database and is accessible by SQL querying. Thus, there are no obstacles to acquire the raw data needed for a subsequent analysis.

It has been determined to only use financial data no older than from the year 2022. This limiting approach was chosen because it offers the most recent data for analysis, allowing insight into the latest patterns and currently active policies in the organisation. Furthermore, by using this approach, it will be ensured that the data are not distorted by the impact of external factors (such as economic swings and inflation rate), leading to up-to-date results.

The decision was made to separate the 2023 data into a standalone test dataset. To accurately and impartially assess a machine learning model's performance, validation and test datasets must be held divided. Validation data are used to fine-tune hyperparameters and choose the best performing model. But, since hyperparameter optimisation might lead to overfitting to the validation data, which will in turn overstate the model's performance, it is sensible to perform a final evaluation on the test dataset, that has never been seen before.

4.2 Exploratory Data Analysis

The data used for analysis and model building in this project is strictly limited to the data provided by the loan applicant, and can be divided into two categories: personal data from the application, and multiple account statements, each containing numerous transactions. The main subject of the current analysis is the data from the account statement with categorised transactions. However, as demonstrated by the graph depicting the number of transactions (see Figure 4), due to the company's business rules and processes, an account statement is not necessarily expected to have a large number of transactions; furthermore, the amount of transactions does not drastically influence the application decision.

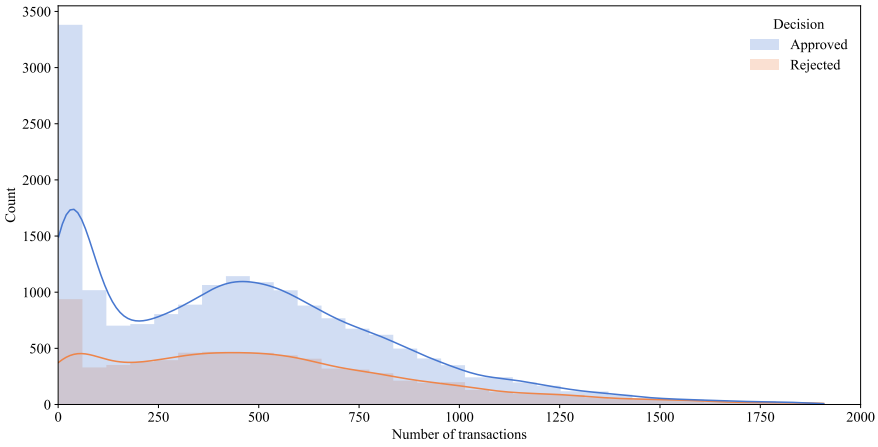


Figure 4. Number of transactions in account statement grouped by the outcome of application.

The applicant's income is one of the most crucial factors to take into account when evaluating a loan application. As an example, it is possible to extract income that has been divided into several categories from the account statement. As a result, the relationship between the applicant's salary and the outcome of their loan request can be observed (see Figure 5). There are also additional categories of income/expense that indicate customer's good creditworthiness or solvency, such as moderate overall expenses, low debt-to-income ratio, investment income, etc.

Despite the limited and difficult-to-process account statement data, it is nevertheless possible to spot the tendencies that must cause a rejection of the loan application. These aspects may include, but not restricted to, excessive credit card use, debt collection costs, excessive gambling, or usage of payday loans (see Figure 6). All the aforementioned may

signal customer's irresponsible financial behaviour. When assessing a loan application, it is critical to look for any indications of irresponsible financial behavior on the applicant's account statement because this could indicate a higher chance of loan default.

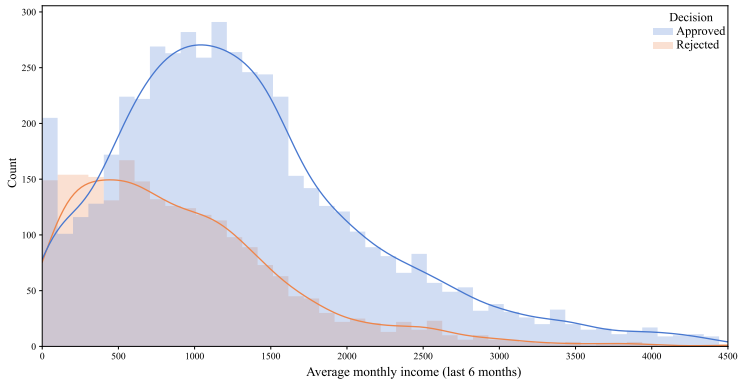


Figure 5. Applicants' average salaries grouped by the application outcome.

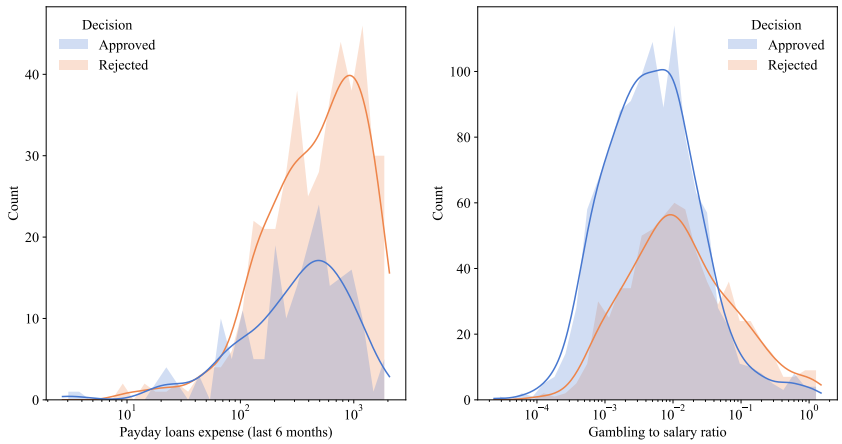


Figure 6. Amount of payday loan expenses and gambling to salary ratio grouped by the application outcome.

To sum up, even though there were no features that were consistently effective for classifying data, the analysis revealed important details about the nature of data and demonstrated that the stated problem could potentially be resolved. Additionally, few data-quality issues (e.g., outliers, product-specific abnormalities) have also surfaced, which will be addressed in the following chapter.

4.3 Data Preparation

In the following section, the attention is directed towards the preparation of the dataset. This step encompasses crucial steps, such as data cleaning, outlier detection, feature engineering, and reduction of data dimensionality, with the overarching objective of ensuring the appropriateness of the data for model building.

4.3.1 Cleaning

As a first step, it was deemed necessary to exclude the refinancing loan product from the dataset, the main reason being its distinctive features, which are not shared by other loan products. Since the refinancing loan is mostly used by the customer to pay off the current debts, the patterns of one's financial behaviour might differ drastically. Filtering out the refinancing loan applications from the dataset has resulted in ca 11% dataset shrinkage.

Secondly, it was decided to eliminate from the dataset the account statements with the small number of transactions, as it was decided that such statements do not provide adequate data for analysis and do not accurately reflect the customer's financial activity. Analysing empty accounts can result in assumptions about the customer's creditworthiness or solvency that are unreliable or deceptive. Following discussions with the business side, it was decided to eliminate account statements with fewer than 50 transactions, resulting in another ca 14% dataset volume reduction.

During financial statements analysis, it is also essential to identify outliers so that unexpected data points do not skew the results or interpretations. It has been chosen to apply the Z-score method for outlier detection with the following formula (given the feature with a mean μ and a standard deviation σ):

$$Z = \frac{x - \mu}{\sigma}$$

Applying the Z-score and filtering out the points with $|Z| > 3$ has resulted in the elimination of an additional approximately 0.2% of the total volume of the data set.

4.3.2 Feature Engineering

Whilst using the applicant-provided personal data in analysis requires minor initial pre-processing, such as encoding categorical features, adapting account statement data is less straightforward and rather challenging. It has been decided to approach the problem using the following strategies.

- *Aggregation*. Combining the many-side of one-to-many relationship using functions (e.g., sum, count, average).
- *Discretisation*. Converting continuous features, such as transaction timestamps, into discrete bins or categories.
- *Domain knowledge*. Creating features that capture essential patterns within the given domain based on the knowledge of current processes.

Each transaction category has a broader parent category, as indicated in Section 2.4, therefore, a more general category was chosen as the main criterion for grouping, leading to the creation of approximately 60 primary categories. However, there exist some important exclusions and certain categories must be regrouped manually: for instance, although credit card charges, payday loans, are also generally considered to be loans, there is a significant distinction from the bank's point of view; on the other hand, bank classifies a variety of income categories more generally as a salary. To depict the customer's financial behaviour over time, the transaction timestamps were discretised by the number of months elapsed, and subsequently, the transaction data were additionally aggregated by applying the cumulative sum function. Only the groupings corresponding to the most recent one, three, and six months were retained, as the current business process focusses primarily on the evaluation of transactional data within the timeframes of 30, 90, and 180 days.

The main advantage of grouping transactions by both category and month is that it makes it possible to analyse the transaction history more thoroughly and concisely, which might uncover broad behavioral trends or anomalies (see Figure 7). Although using a relatively small number of features and being a less complex approach, it offers a fairly accurate representation of behaviour over time. Yet, the main drawback of the described approach is the loss of data granularity: firstly, it is assumed that spending behavior is consistent

over given timeframe; additional trade-off is that outlier transactions, such as irregular or one-time large transactions, can potentially cause the representation to be less accurate.

Account statement (as of 2023-03-31)		
2022-10-01	Salary	3000
2022-10-10	Gambling	-60
2022-11-01	Salary	3000
2022-11-25	Loans	-250
2022-12-01	Salary	3500
2023-01-01	Salary	3500
2023-01-15	Gambling	-35
2023-01-20	Gambling	75
2023-02-01	Salary	4000
2023-02-25	Loans	-500
2023-03-01	Salary	4000
2023-03-15	Gambling	-25

	last 1 mo	last 3 mos	last 6 mos
Salary income	4000	11500	21000
Loans expense	0	500	750
Gambling income	0	75	75
Gambling expense	25	60	120

Figure 7. Transformation of example transaction data using aggregation and discretisation.

In addition to transactional data, multiple domain-specific metrics must be calculated, since they provide valuable information regarding financial stability and responsibility. Assembled metrics are mostly ratios in nature, for instance, loans to salary ratio (loan expenses divided by salary income), gambling to salary ratio (gambling expenses to salary income), etc.

4.3.3 Feature Selection

The feature engineering conducted in the previous section resulted in the dataset with over 200 features, which is rather inefficient to use for model building. Therefore, the objective of applying feature selection algorithms is to select the most informative features with respect to the class label [29]. Although there are many different methods, since the account statement data is overwhelmingly numerical, it has been decided to use mathematical filter models as a starting point for the selection.

Mathematical filter models were chosen for feature selection, in contrast to exhaustive feature selection or recursive feature elimination, for a number of reasons. One important consideration is that the chosen features must have logical interpretation from a business standpoint: whereas very accurate models may be produced through exhaustive feature selection and recursive feature reduction, the chosen features might not be truly significant

or understandable. The use of mathematical filter models made it possible to choose features in a well-thought-out manner.

Various metrics were used in the analysis, such as the Fisher score, information gain, correlation coefficient, chi-squared metric, etc. Each of the aforementioned metrics assesses the importance of given features, but the approach is different. Although almost all metrics were available in the Python scikit-learn[1] package, Fisher score required custom implementation. Given mean μ and standard deviation σ, while j denoting the subset of data belonging to the class j and p_j being the fraction of data belonging to class j, the Fisher score is calculated as follows [29]:

$$F = \frac{\sum_{j=1}^{k} p_j(\mu_j - \mu)^2}{\sum_{j=1}^{k} p_j \sigma_j^2}$$

After the manual review, all the results have proven to be rather predictable and logical (see Figure 8, features are sorted by the average ranks of each metric), with information gain being the closest to the prior domain-based expectations.

As a result of the feature selection stage, it was decided that 40 features with the highest discriminatory power are the optimal number for further analysis. However, this outcome might later be subject to change during the model selection step or by applying XAI techniques.

4.4 Model Selection

In this section, multiple experiments were carried out and evaluated to determine the most suitable algorithm to solve the problem and to decide on the optimal set of conditions to build a model. The experiment in question is a binary classification problem, where the positive class indicates that the loan application should be rejected.

The metric values presented and compared in all the experiments in the following section represent average values of the outcomes obtained through cross-validation. Furthermore,

[1] https://scikit-learn.org/

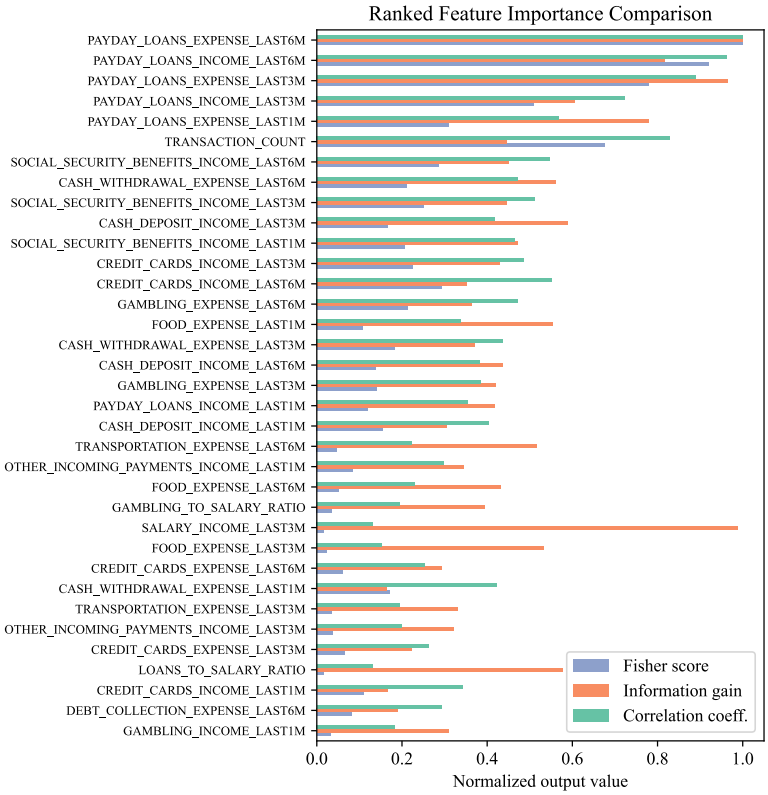


Figure 8. Most important features by several feature selection metrics

all experiments were conducted and listed below using the nested cross-validation approach. Nested cross-validation is a technique for selecting models and tuning its hyperparameters that aims to solve the issue of overfitting the training dataset. Nested cross-validation divides the dataset into training, validation, and testing sets and then uses two loops to assess model performance (see Figure 9): the inner loop optimises the hyperparameters, while the outer loop is used to evaluate the performance of the model [41].

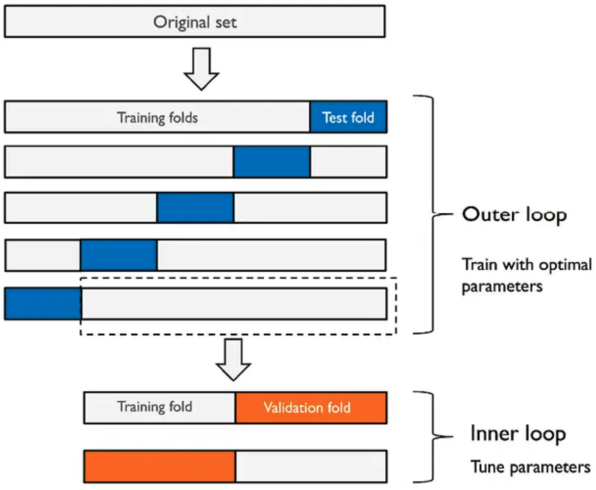


Figure 9. Nested cross-validation approach visualised [42].

4.4.1 Evaluation

Evaluation of machine learning models is a critical step in determining their ability to solve problems, and to properly assess their efficiency, the right evaluation criteria must be chosen. The typical approach to define the performance of a binary classification algorithm is to use confusion matrix (see Table 2).

Table 2. Confusion matrix in the context of current problem.

		Actual outcome	
		Rejected	**Approved**
Predicted outcome	**Rejected**	True Positive (TP)	False Positive (FP)
	Approved	False Negative (FN)	True negative (TN)

A variety of metrics can be derived from the confusion matrix to evaluate the performance of a machine learning model, and in this research the following metrics were chosen for describing model performance.

- Accuracy – the ratio of the number of correct predictions to the total number of predictions, describing the overall performance of the model.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

- Precision – the ratio of true positives to the total number of positive predictions, representing the ability of the model to avoid false positives.

$$\text{precision} = \frac{TP}{TP + FP}$$

- Sensitivity (also known as recall, or true positive rate) – the proportion of actual positive cases that are correctly identified as positive by the model.

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

- Specificity – the proportion of actual negative cases that are correctly identified as negative by the model.

$$\text{specificity} = \frac{TN}{FP + TN}$$

- F_1-score – the harmonic mean of precision and recall. This metric is a good measure of model performance when classes are imbalanced, since it also takes into account the number of false positives and false negatives.

$$F_1 = \frac{2}{\text{precision}^{-1} \cdot \text{recall}^{-1}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

- Area Under ROC Curve (ROC AUC) – the area under the Receiver Operating Characteristics (ROC) curve, which is a graph showing the true positive rate versus the false positive rate ($\frac{FP}{FP+TN}$) at different classification thresholds. ROC curve visualises the model's performance across different thresholds and may be helpful in choosing the most appropriate decision border, while the area under this curve shows how well the model distinguishes the classes.

The selected metrics are considered to be the preferred method for evaluating models in this domain, as indicated by a systematic literature survey [6]. To maintain brevity, only accuracy, F_1-score, and ROC AUC will be presented in subsequent sections. Although the main drawback of aforementioned metrics is the assumption of equal costs of false positives and false negatives, the business side finds it acceptable, since the model serves as an advisor and either type of misclassification implies additional work.

4.4.2 Determining Baseline

The goal of the initial step was to grasp a preliminary estimation of the algorithm performance and create a baseline to be referenced in the future experiments. For this experiment, only the features of the account statement were provided to the algorithms and then the results were evaluated (see Table 3).

Table 3. Initial model performance comparison.

Algorithm	Accuracy	ROC AUC	F_1-score
k-NN	$72.6 \pm 0.5\%$	$75.1 \pm 0.8\%$	$46.1 \pm 1.7\%$
Decision Tree	$75.6 \pm 0.5\%$	$76.1 \pm 2.3\%$	$56.5 \pm 3.6\%$
Random Forest	$78.0 \pm 0.9\%$	$82.4 \pm 0.9\%$	$58.0 \pm 1.5\%$
XGBoost	$78.9 \pm 0.5\%$	$83.1 \pm 0.5\%$	$64.0 \pm 0.1\%$
SVM	$48.9 \pm 0.3\%$	$51.7 \pm 0.1\%$	$18.0 \pm 5.5\%$
Neural Network	$69.5 \pm 0.7\%$	$70.1 \pm 1.1\%$	$43.9 \pm 1.8\%$

The result has pointed out that k-NN and Neural Network algorithms performed reasonably well overall, but it is clear from the experiment's findings that they had trouble identifying the minority class, as seen by their lower F_1-scores. Conversely, Decision Tree based algorithms surpassed other competitors, showing rather high accuracy overall, and they were able to identify the minority class, as seen by their superior F_1-scores. Expectedly, the ensemble algorithms turned out to be the most successful. SVM, on the other hand, performed rather poorly, with low accuracy and severe difficulty recognising the minority class. However, the results have provided an initial baseline and ideas for possible optimisation.

4.4.3 Applying Balancing Techniques

The following step focused on addressing the problem of an unbalanced dataset to improve the quality of the models. Two various strategies were utilised to approach the issue: data-level and algorithm-level methods. In terms of the k -NN and SVM algorithms, the data was also preliminarily scaled.

The first experiment was set out to find the best data-level technique for balancing data. Various techniques of oversampling and undersampling were applied, including random methods and synthetic techniques (such as SMOTE and ADASYN), and the resulting models were evaluated to choose the most suitable technique for each algorithm.

The application of data-level balancing techniques has resulted in an overall performance improvement, especially in the ability to detect the minority class, as indicated by the F_1-score (see Table 4). Despite the minimal decrease in average accuracy and the ROC AUC in terms of k-NN and the Decision Tree, their sensitivity has actually increased. For k-NN and Neural Network, the random technique was more efficient, while other algorithms performed better with synthetic data creation approaches such as SMOTE.

Regarding algorithm-level techniques, the experiment involved assigning weight to each sample, with a higher value for minority class, and providing it to the models that support this feature. As a result, the model was encouraged to give more significance to the points with higher weights and improve its performance in the minority class.

Table 4. Model performance comparison – data-level balancing methods.

Algorithm	Best Technique	Accuracy	ROC AUC	F_1-score
k-NN	Random US	$70.7 \pm 0.6\%$	$74.9 \pm 0.8\%$	$58.5 \pm 1.2\%$
Decision Tree	SMOTE	$75.1 \pm 1.0\%$	$75.9 \pm 1.8\%$	$59.4 \pm 2.7\%$
Random Forest	ADASYN	$77.2 \pm 0.8\%$	$81.4 \pm 0.7\%$	$64.3 \pm 1.3\%$
XGBoost	SMOTE	$78.0 \pm 0.4\%$	$82.7 \pm 0.3\%$	$65.7 \pm 0.1\%$
SVM	SMOTE	$55.9 \pm 2.8\%$	$56.8 \pm 1.5\%$	$33.8 \pm 2.6\%$
Neural Network	Random US	$70.3 \pm 2.5\%$	$69.3 \pm 2.1\%$	$48.2 \pm 2.7\%$

According to the experiment's findings (see Table 5), the algorithm-level sample weight strategy was more optimal compared to the baseline and the results of applying data-level balancing techniques. The resulting models showed better overall performance and the ability to identify the minority class, as pointed out by F_1-score. The effectiveness of the sample weight approach is believed to be explained by the fact that the imbalance is approached without removing any data or adding artificial data.

Table 5. Model performance comparison – weighted sample approach.

Algorithm	Accuracy	ROC AUC	F_1-score
Decision Tree	$74.9 \pm 0.5\%$	$77.4 \pm 2.1\%$	$62.4 \pm 2.9\%$
Random Forest	$78.3 \pm 0.4\%$	$82.3 \pm 1.0\%$	$58.1 \pm 1.5\%$
XGBoost	$79.4 \pm 0.4\%$	$83.7 \pm 1.1\%$	$67.3 \pm 0.2\%$
SVM	$58.2 \pm 3.7\%$	$60.6 \pm 2.6\%$	$40.8 \pm 4.5\%$

The results of the experiments conducted showed that DT-based algorithms, including Decision Tree itself, Random Forest, and XGBoost, have noticeably outperformed other approaches: these algorithms demonstrated higher accuracy, better ability to distinguish the minority class, and efficiency when working with a high-dimensional dataset. Thus, only stated algorithms will be considered in further attempts to improve prediction quality.

4.4.4 Further Performance Improvements

As mentioned in Chapter 2.3, the financing products of the company are divided into two main categories: hire-purchase and consumer loans. Since the rules and conditions for evaluating loan applications slightly vary depending on the product, it was hypothesised that developing separate models could lead to more accurate predictions. Therefore, it was decided to conduct an experiment of creating and applying two distinct models for each product type. There are approximately the same number of applications in both categories,

with around 52% of applications being consumer loans, and the approval ratio is slightly higher for hire-purchase (see Figure 10).

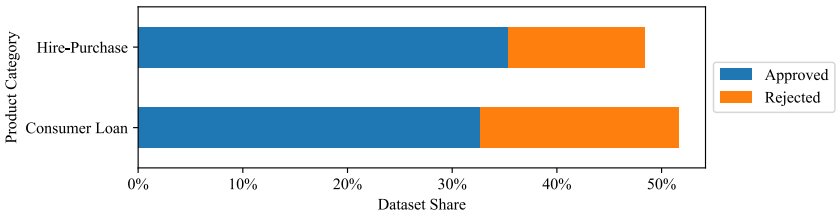


Figure 10. Composition of the dataset by product category and decision.

The results (see Table 6) indicate that training dedicated models on the separate dataset can lead to improved performance. In both cases, the F_1-score and the ROC AUC have improved, indicating better overall performance, especially highlighting good results of a hire-purchase model built with the XGBoost algorithm. Although the consumer loans model seemingly is worse in terms of accuracy compared to the previous experiment, the weighted average accuracy of both models combined did not drop for any algorithm: e.g., XGBoost models have weighted average accuracy of 80.4%.

Table 6. Product-based model performance comparison.

Product Model	Algorithm	Accuracy	ROC AUC	F_1-score
Hire-Purchase	Decision Tree	$78.0 \pm 0.5\%$	$74.4 \pm 1.1\%$	$59.6 \pm 2.7\%$
	Random Forest	$80.8 \pm 0.8\%$	$85.0 \pm 0.6\%$	$64.9 \pm 1.0\%$
	XGBoost	$82.7 \pm 0.8\%$	$86.0 \pm 0.6\%$	$67.9 \pm 0.9\%$
Consumer Loans	Decision Tree	$72.5 \pm 1.8\%$	$76.5 \pm 1.9\%$	$64.7 \pm 1.9\%$
	Random Forest	$76.2 \pm 1.0\%$	$82.7 \pm 1.1\%$	$68.6 \pm 0.8\%$
	XGBoost	$78.2 \pm 0.8\%$	$84.2 \pm 0.8\%$	$69.2 \pm 1.1\%$

All previous experiments focused solely on the data from the account statements. However, as stated in Section 4.2, it is also possible to utilise the personal data provided by the applicant in the application, which has the potential to enhance the results. To test this hypothesis, another experiment was conducted where the dataset was augmented with aforementioned personal data features.

The experiment confirmed the assumption, showing that the models using all three algorithms significantly improved the performance for both types of products (see Table 7). The performance of the XGBoost algorithm has once again surpassed that of other algorithms. All models have seen rather serious increase of the ROC AUC metric and

F_1-score values, with the consumer loans model excibiting a better F_1-score due to its more evenly balanced classes. The stability of the models' results has also remained unchanged for all algorithms.

Table 7. Product-based model performance comparison on data containing applicant provided personal information.

Product Model	Algorithm	Accuracy	ROC AUC	F_1-score
Hire-Purchase	Decision Tree	$77.9 \pm 2.3\%$	$81.6 \pm 1.4\%$	$63.4 \pm 2.4\%$
	Random Forest	$82.7 \pm 0.5\%$	$88.2 \pm 0.5\%$	$69.2 \pm 1.1\%$
	XGBoost	$85.2 \pm 0.8\%$	$89.5 \pm 0.7\%$	$71.2 \pm 1.2\%$
Consumer Loans	Decision Tree	$75.9 \pm 1.2\%$	$81.4 \pm 0.6\%$	$68.6 \pm 0.7\%$
	Random Forest	$80.6 \pm 0.8\%$	$87.3 \pm 0.7\%$	$73.4 \pm 0.8\%$
	XGBoost	$82.0 \pm 0.6\%$	$88.5 \pm 0.7\%$	$74.5 \pm 0.9\%$

A common approach to enhance the performance of the model is to reduce the dimensionality of the dataset by eliminating redundant or irrelevant features. Another experiment was therefore conducted with the aim of examining the impact of lowering the dimensionality to n most important account statement features on the evaluation metrics of the model, and to identify the value of n that can maximise the model's performance. The output of the feature selection stage (Section 4.3.3) defines the importance order in which the features are included in the dataset.

According to the results, adding more features generally improves the performance of both hire-purchase (see Figure 11) and consumer loan (see Figure 12) models in terms of accuracy. Regarding the detection of the minority class, the results of the Decision Tree tend to decrease, while the performance scores of the Random Forest and XGBoost algorithms show a significant improvement. In lower dimensions, the RF algorithm outperforms XGBoost in terms of the F_1-score and the ROC AUC metric, while in higher dimensions, XGBoost surpasses RF and takes the lead. The above-mentioned findings indicate that the overall performance of the models improves in higher dimensions, and since no problems with prediction speed nor data transformation were detected, the dataset's dimensionality does not need to be decreased.

Based on the acquired statistics, the ensemble algorithms outperform others in terms of solving the loan evaluation problem. Despite the simplicity and natural interpretability, the Decision Tree algorithm produces on average less accurate results along with lower

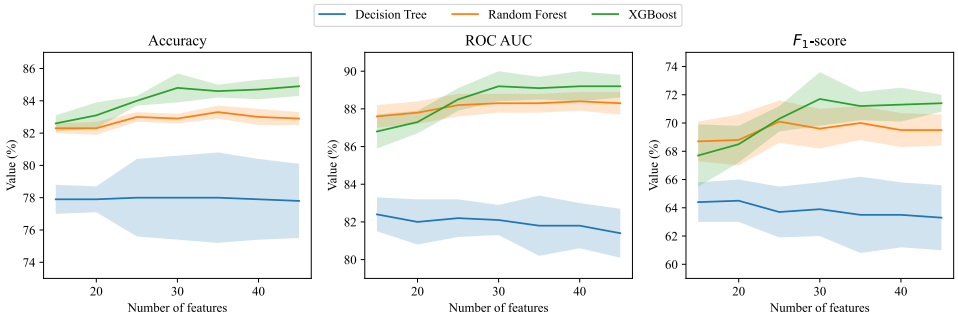


Figure 11. Performance curve of models for hire-purchasing.

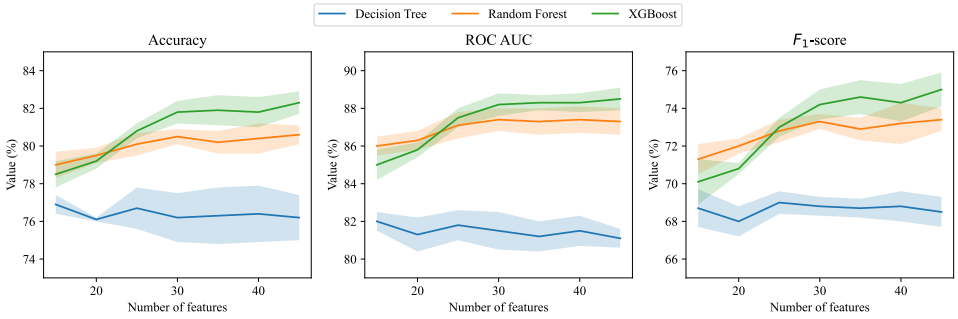


Figure 12. Performance curve of models for consumer loan products.

stability, as seen by higher standard deviation values. Furthermore, higher values of the ROC AUC metric and the F_1-score demonstrate that both RF and XGBoost have consistently good abilities to detect and distinguish the minority class. However, XGBoost has exhibited slightly better overall efficiency by all evaluation metrics and has produced the most stable and reliable results in all experiments.

To summarise the experiment section, the results have shown that the most optimal approach is to use distinct models for each product types, and utilising both the data provided in the application by the customer and approximately 40 most essential features obtained from the account statement. Both models have achieved the highest efficiency by using XGBoost algorithm.

4.5 Integration of XAI Techniques

The next crucial step in achieving the stated objective is the integration of XAI techniques. Due to the availability of both LIME[1] and SHAP[2] as Python packages, both techniques

[1]https://github.com/marcotcr/lime
[2]https://github.com/slundberg/shap

are seamlessly integrated into the current workflow. While the flexibility and simplicity of LIME have made it more straightforward to integrate, the visualisation of SHAP requires some additional steps to perform correctly. Furthermore, an optimised variant of TreeSHAP was applied to the XGBoost model to improve SHAP performance by providing an efficient approximation of the default SHAP algorithm [39]; however, this led to a more model-specific approach.

Upon examining the two techniques, LIME has been found to provide a more concise and straightforward explanation that is easier to comprehend. However, the default output generated by LIME lacks aesthetic appeal and therefore must undergo some alterations to increase its user-friendliness (see Figure 13). Conversely, the explanations of SHAP can be noted for their visual appearence, that might be very informative for a data scientist, but is believed to be overly complex for non-expert users, such as loan managers (see Figure 14).

Figure 13. Example LIME explanation plot for single random data instance.

Regarding the precision of two techniques, both LIME and SHAP were able to generate explanations that were logically correlated with the observed outcome of the loan application. Nevertheless, LIME was observed to have a slightly higher variance in its outputs, while SHAP yields more accurate and stable explanations, due to its non-linear and more sophisticated approach. In terms of performance, however, even the optimised variants of SHAP were found to be slower than LIME, with their speed decreasing with the high number of features.

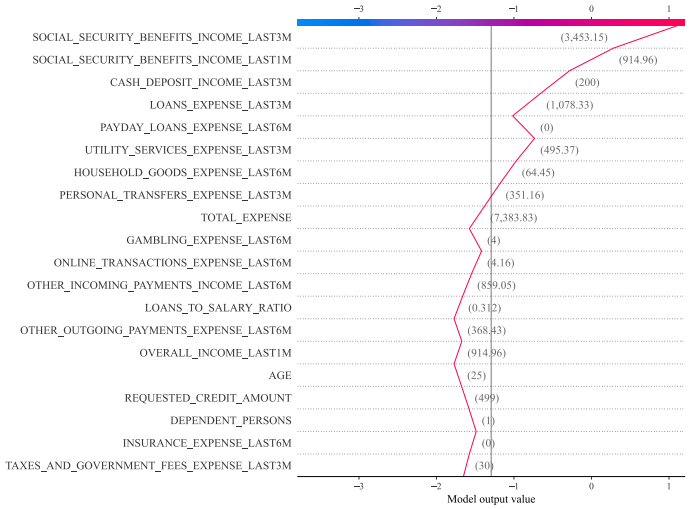


Figure 14. Example SHAP dependency plot for single random data instance.

The SHAP framework provides a variety of explanation and visualisation tools that are not limited to the local interpretability scope and also offer global explanations, enabling a clear and concise comprehension of the behaviour of the model. The global explanation summaries generated by SHAP visualisation tools (e.g., see Figure 15) offer significant benefits to the data analysts of the company by allowing them to examine the decision-process of a given model, inspect the underlying data patterns and detect potential areas for improvement; however, it is rather unnecessary to display thorough and complex explanations to the end-user, as those might be beyond the context and provide superfluous information.

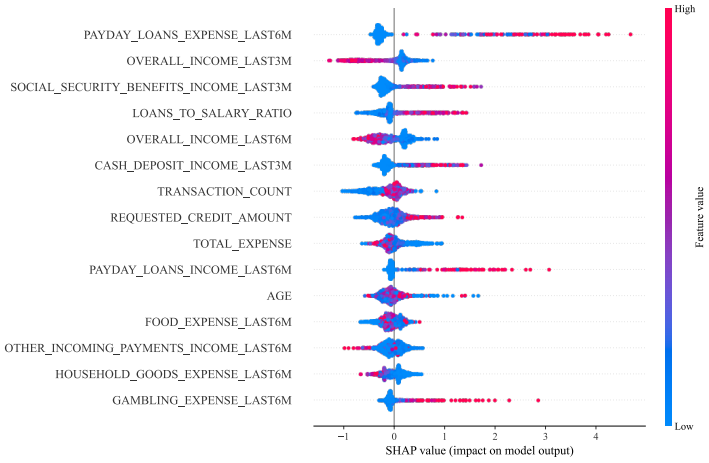


Figure 15. SHAP summary plot for hire-purchase model.

After careful consideration, LIME was selected as the preferred technique for locally explaning a single prediction to the loan manager, with key factors being an intuitive

and comprehensible logic behind the approach, as well as the ability to offer clear and understandable explanation, with fairly concise visualisation. The relative simplicity of LIME positively impacts computational resource utilisation and processing speed, which is particularly essential in systems that prioritise maintaining rapid data throughput and performance; the approach of LIME remains effective in high-dimensional data.

4.6 Model Deployment

Upon the development of an AI model, the ultimate stage in the machine workflow entails model deployment, with its primary objective being the model's ability to be applied within a production environment. There are multiple methods and tools to achieve the goal of exposing the model via an accessible API endpoint. However, regardless of the approach chosen, it is crucial to consider several factors important to the company, including scalability, efficiency, maintainability and security.

The initial phase involves developing an API service that is accessible by the HTTP protocol and conforms to the REST guidelines, as this is one of the common practises at the company in the context of microservices. The decision was made to adopt OpenAPI[1], a modern framework for creating APIs in Python, that is known for its high performance, a lightweight design, and excellent scalability. The primary objective of the API wrapper is to encapsulate the machine learning models and LIME explainer objects, which were previously stored in memory through the use of the joblib[2] module. The API service was designed to receive the HTTP request with a JSON-formatted object consisting of the data required for analysis, as well as supplementary options, e.g., the preferred number of features in the explanation or whether the explaination graph should be generated (see Figure 16).

Upon receiving a request, the primary responsibility of the API service is to perform a data transformation to a format supported by the model, execute the analysis and obtain the prediction by invoking the appropriate model, and use a LIME explainer to generate an explanation (see Figure 17).

[1]https://fastapi.tiangolo.com
[2]https://pypi.org/project/joblib

```
POST /evaluate HTTP/1.1
Host: *****
Content-Type: application/json

{
  "data": {
    "productType": "CF",
    "requestedCreditAmount": 10000,
    "applicantData": {
      "educationCode": 3,
      "jobPositionCode": 5,
      "maritalStatusCode": "V",
      "gender": "M",
      "age": 35,
      "dependents": 2,
      "residency": "EE"
    },
    "accountStatements": [
      {
        "periodStartDate": "2022-01-05",
        "periodEndDate": "2022-07-05",
        "rows": [
          {"categoryId": 8, "date": "2022-01-05T13:24:00Z", "sum": -50.00},
          {"categoryId": 24, "date": "2022-01-05T10:12:00Z", "sum": -200.00},
          {"categoryId": 85, "date": "2022-01-06T08:15:00Z", "sum": 1500.00},
          ...
        ]
      },
      ...
    ]
  },
  "options": {
    "explanationLength": 5,
    "generateGraph": true
  }
}
```

Figure 16. Example HTTP request for account statement analysis.

```
{
  "model": "account-statement_CF_v1"
  "prediction": 0.7235,
  "explanation": {
    "features": [
      {
        "name": "Debt collection expense (last 6 mos)",
        "value": 30.25,
        "importance": -0.29
      },
      {
        "name": "Overall income (last 1 mo)",
        "value": 3084.80,
        "importance": 0.21
      },
      {
        "name": "Social security benefits income (last 3 mos)",
        "value": 8.80,
        "importance": -0.17
      },
      ...
    ],
    "graph": "PHN2ZyB4bWxucz0iaHR0cDov..."
  }
}
```

Figure 17. Example JSON-formatted response of the API service.

Following the integration of the API layer, the subsequent step involves deploying the model to a server for accessibility. The modern approach is to perform a containerisation of the service, which enables efficient and platform-independent deployment by assembling the service along with its dependencies and libraries into a single package. Containerisation of a given application using Docker[1] has proven to be an optimal approach due to robust management of Python dependencies, an isolated runtime environment, easier maintainability, and horizontal scalability. Following containerisation, the service is effortlessly deployable onto a server, becoming accessible to other systems (see Figure 18).

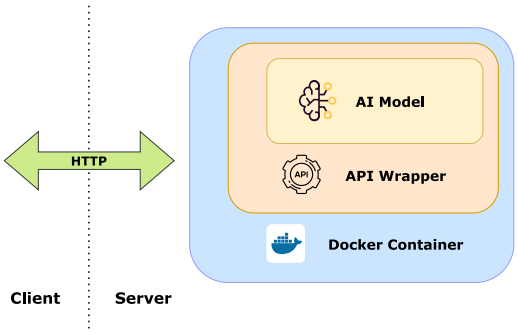


Figure 18. System architecture diagram of the developed API service.

As the main achievement of the given section, the developed AI models were efficiently encapsulated within an API service. Various security measures were also implemented to the application, according to the internal security guidelines of the company. The resulting application exhibits a lightweight microservice architecture that is simple to maintain and an optimal solution in terms of scalability. Consequently, the application is readily available for integration with the Credit Decision System, allowing the loan eligibility evaluation process enhancement for the organisation.

[1] https://www.docker.com

5 Results

The present thesis focused on the problem of loan eligibility prediction through the examination of the applicant's financial data. In the study, various aspects of the issue were thoroughly analysed and multiple experiments were conducted. Consequently, all the goals set in the thesis were effectively accomplished: interpretable machine learning models were developed, optimised, and integrated into the current application decision-making process in the role of an advisory tool.

The findings of this thesis have demonstrated that the development of a machine learning model that focusses exclusively on the analysis of account statement data can yield a satisfactory level of performance. By using aggregation and grouping techniques, it was possible to engineer numerous features describing account statement and use them in subsequent steps. However, augmenting the analysed data with personal information from the loan application has been shown to further enhance the model's performance and accuracy, surpassing the results obtained solely from the account statement analysis.

One of the main goals of the current thesis was to identify the optimal algorithm for the analysis of the account statement when processing loan applications. In order to address the issue, multiple cutting-edge algorithms were evaluated. The results have shown that the Decision Tree-based algorithms have demonstrated the highest performance and efficiency in accurately identifying the minority class (rejected applications, in this context). Overall, ensemble algorithms exhibited favorable performance on average, with XGBoost proving to be the most suitable algorithm for the stated classification problem.

Incorporating the developed AI model into the existing business process required that it be easily accessible to users with minimal effort. A comprehensive Proof of Concept was conducted to ensure successful deployment of the model. The solution has been designed to be accessible via HTTP and as a result, the model is now available for integration into the current workflow, with providing predictions automatically by the request.

In addition to addressing the main problem of loan application classification through account statement analysis, this study also placed an emphasis on the interpretability of the developed AI model – given the nature of the field, it was deemed crucial to incorporate explainability techniques to ensure transparency for end users. As such, two state-of-the-art model-independent XAI techniques, LIME and SHAP, were carefully analysed and successfully integrated into the process, offering a higher degree of transparency in the decision-making process.

5.1 Quantitative Validation

5.1.1 Prediction Quality

As previously indicated in Section 4.1, data originating from the year 2023 was isolated into a distinct dataset for testing purposes. This methodology mimics a production environment by assessing the performance on newly received data from the model's perspective, and thus allows more rigorous and unbiased evaluation. The approach also enables the evaluation of the model's resilience to the concept drift phenomenon.

To assess the performance of the developed machine learning models, the confusion matrix and receiver operating characteristic (ROC) curve were used. As illustrated in Figure 19 and Figure 20, both the hire-purchase model and the consumer loan model demonstrate performance that is comparable to the average goodness measures obtained in Section 4.4. The accuracy of the hire-purchase model is 84.7%, and that of the consumer loan model is 81%; in terms of F_1-score, the results are 70.8% and 75.6%, correspondingly, the main reason behind the difference being the greater imbalance of the hire-purchase loan application data. Although both models exhibit admirable specificity rates of 93.8% and 91.5%, respectively, their sensitivity rates of 62.8% and 67.5% show relative weakness.

In general, the performance remains rather commendable, even when evaluated on prospective data, as perceived from the models' perspective. However, the results suggest that the models demonstrate stronger proficiency in identifying approved loan applications, or negatives, and are less effective at detecting rejected applications, or positives. Poor sensitivity rates may be due to the nature of rejected applications that may be rejected for a

multitude of reasons beyond account statement, such as prior debt or poor credit history; as for now, there is no efficient way to extract the exact rejection reason from the application data. Another possible explanation is the imbalanceness of the dataset, the number of rejected applications being lower than that of approved applications.

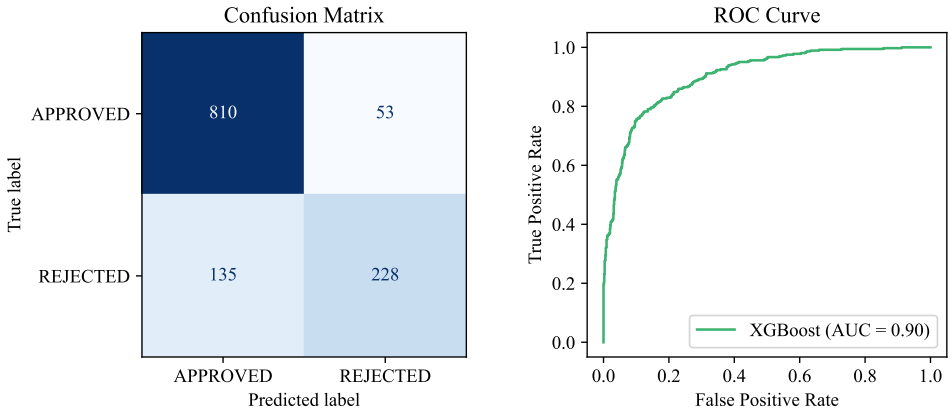


Figure 19. Performance of the hire-purchase model on test dataset of respective loan applications.

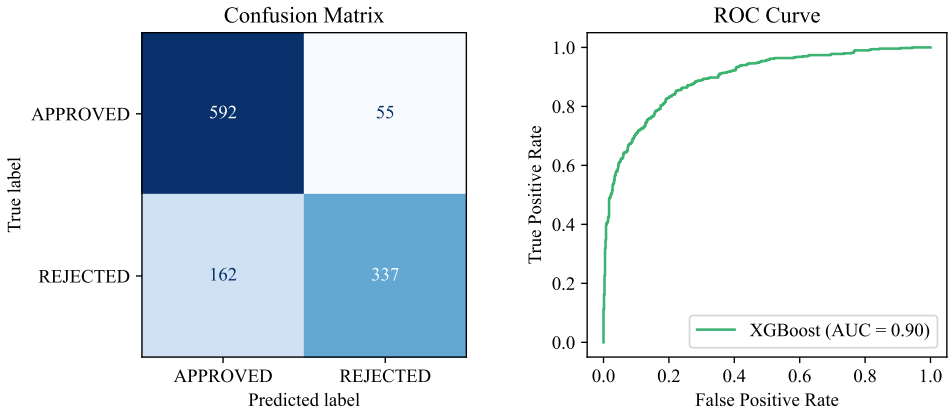


Figure 20. Performance of the consumer loan model on test dataset of respective loan applications.

The results also indicate that the efficiency of the models has remained stable, even despite the turbulent financial landscape in Estonia during the years 2022-23. Although the absence of a concept drift phenomenon is a rather unexpected finding, there exist hypotheses that explain the discovery, such as the possibility that the adapted business rules are not yet implemented, or the changed concepts have not so far been reflected in the available data, or the changes were not deemed significant enough in the broader context.

5.1.2 Explanation Quality

One of the crucial measures for evaluating XAI techniques is the *faithfulness* metric. The aim is to demonstrate whether the relevant features identified by an XAI method are indeed relevant for model prediction. One method of evaluating faithfulness is to analyse the impact of removing features on the model prediction and compute the correlation between probability drops and relevance scores assigned by the XAI method [40]. The results in this domain suggest that overall the LIME method's performance in terms of faithfulness is rather acceptable, with high variability and a median value of approximately 0.2 (see Figure 21). Furthermore, the impact of the number of features n, specified as a hyperparameter, on the faithfulness score is not significant, although $n = 10$ slightly improves the median value and reduces variability.

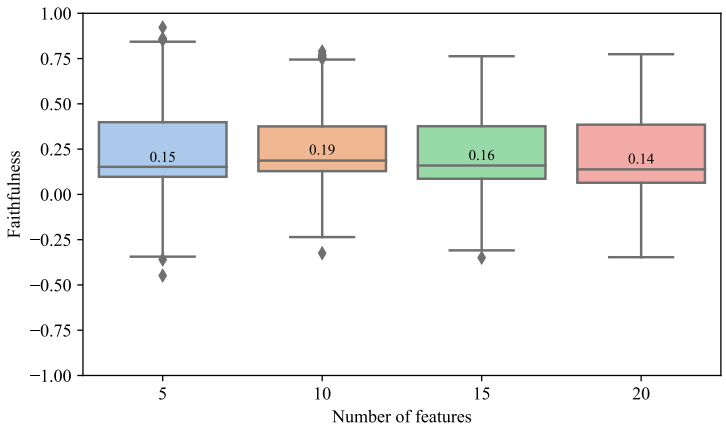


Figure 21. Distribution of LIME output faithfulness values.

The concept of *stability* refers to examining the similarity of explanations for similar or neighbouring data instances in the fixed model scope [37]. High stability implies that slight variations in the features of an instance do not substantially change the explanation, thereby increasing the robustness of the explanation. As one study points out, the conventional interpretability approaches, such as LIME and SHAP, are not robust enough as the minimal perturbations in data can have a drastic impact on the performance of these methods and significantly alter the explanation [40]. In this particular research domain, the LIME method has exhibited low variability only for a relatively small set of features, which limits its reliability (see Figure 22). Additionally, the number of features deemed universally important in all generated explanations is rather limited.

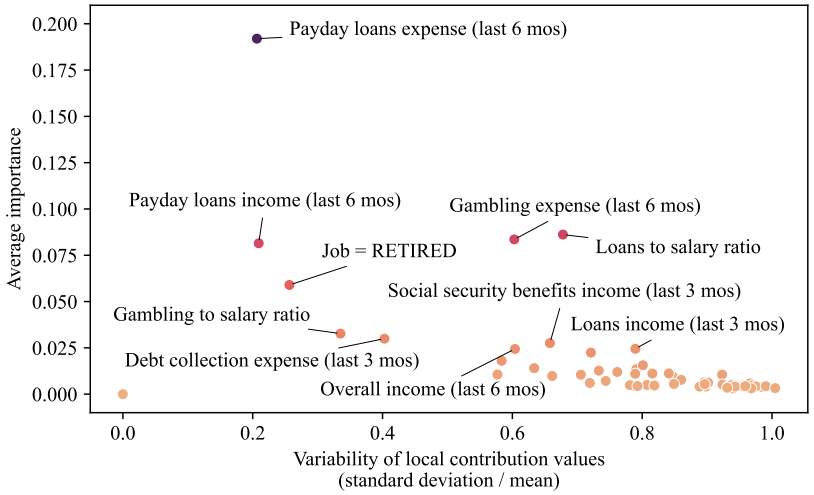


Figure 22. Importance and local stability of LIME output features.

Another way to assess a XAI method's output is *consistency*, which demonstrates the inter-XAI explanation consistency across multiple XAI methods on the same dataset [43]. The aim is to compare the explanation summaries produced by various XAI algorithms (or executions of the same algorithm), and one possibility is to calculate the L^2-norm distance between the explanations [44]:

$$L^2 = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}$$

The average L^2-norm distance between two LIME executions on the same dataset is $d_{L^2} = 0.21$, due to the nondeterministic behaviour of LIME. However, the average distance between the SHAP and LIME outputs is much higher ($d_{L^2} = 0.96$) – this implies that LIME and SHAP might offer fairly different justifications given the same input point (see Figure 23).

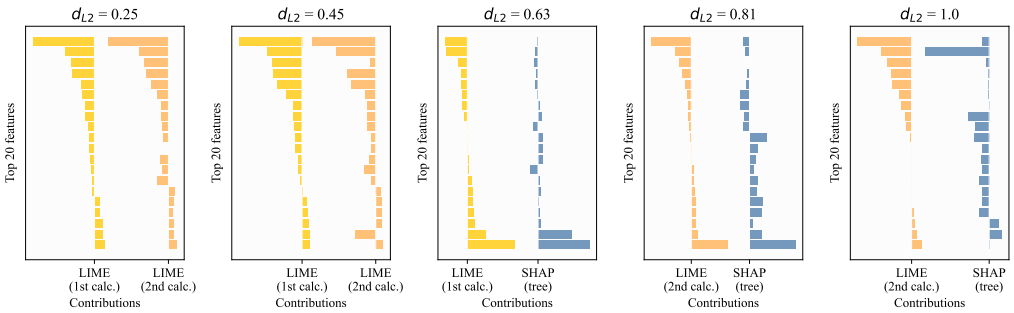


Figure 23. The L^2-norm distances between different XAI method outputs – example of 5 data instances.

Since XAI aims to improve human understanding of machine behaviour, one of the key criteria for explanation quality is sufficiency or conciseness: the explanation must generate an adequate explanation and enforce the concept of providing the minimum level of information necessary [45]. For this purpose, a framework suggests applying the *compacity* metric to assess the degree to which a set of n most important features describes the behaviour of a model [44]. The findings suggest that LIME is rather effective technique for achieving explanation sufficiency, even for models with a high number of distinct features (see Figure 24): for example, to explain at least 75% of model behaviour in more than 80% of cases, 10 features are sufficient. In order to obtain an explanation rate of at least 90% for 90% of the cases, approximately 20 features are required, albeit this may constitute excessive information in most cases.

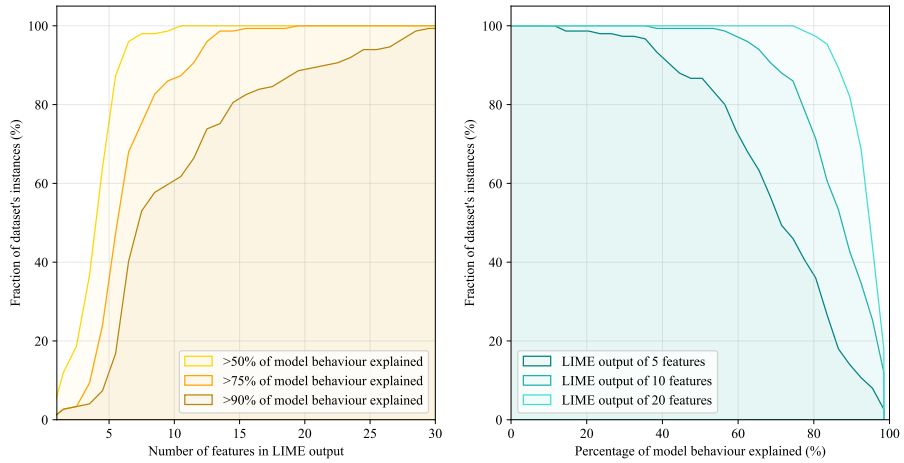


Figure 24. Relationship between the number of features in LIME output and percentage of model behaviour explained.

Based on the results of the XAI evaluation metrics, it is apparent that the LIME approach produces results that are moderately decent, albeit not excellent. This modest performance can largely be attributed to the nature of the data, which presents a difficulty in finding universally optimal features to explain the results. Consequently, the LIME technique yields output of higher variability, and more features are needed to accurately describe the behaviour of the model. Nevertheless, the aforementioned lower degree of explanatory accuracy is one of the drawbacks attributed to the utilisation of model-agnostic techniques and the consequent incapacity to capture the intrinsic model behavior accurately.

5.2 Social Evaluation

To fully comprehend the performance of an XAI model, it is essential to analyse not only quantitative metrics, but also invaluable feedback from end users and stakeholders. Being the primary beneficiaries of the model's advisory function and explainability, users can provide critical insights that contribute to a better understanding of its trustworthiness and efficacy. Multiple researches accentuate that the gathering of subjective human-centred evaluation of an XAI system (in the current scope, provided by the loan managers) allows researchers to validate that the model genuinely serves those who depend on its guidance [45, 46].

A particular study highlights the absence of a standardised evaluation framework and a shared understanding of the XAI evaluation taxonomy and proposes four primary components of human-centred evaluation categories, which serve as a foundation for evaluating XAI systems from the perspective of end users: trust, usefulness / satisfaction, understandability, and performance [46]. However, it is crucial to maintain a balance between technical specificity and conciseness when requesting feedback from users, ensuring that their input remains relevant and actionable.

To demonstrate the applicability of the proposed evaluation framework, the sample of 40 applications for a consecutive overview by a loan manager. To ensure that the sample was representative, 10 applications representing each outcome (TP, TN, FP, FN) were randomly chosen, mitigating potential biases and guaranteeing a diverse cross section of cases. Recognising that *trust* is a variable factor shaped by user interaction over time and usage, and considering that the LIME output exhibits a commendable level of *understandability*, it was decided to focus on two aspects in the questionnaire on each prediction.

- Performance – *"How well does the result reflect the actual situation?"* on a Likert scale of 1-5.
- Usefulness and satisfaction – *"Would the output be helpful for making a decision?"* on a binary scale.

Regarding performance, the results indicate a generally favorable perception of the XAI model (see Figure 25), with a substantial proportion of high scores on the Likert scale. In particular, a significant portion of the responses received the highest possible rating, demonstrating overall satisfaction with the system's performance. Expectedly, the results reveal that the average performance ratings were lower in cases where the prediction was incorrect. One of the primary concerns identified were the issues caused by the loss of data granularity during feature engineering, due to which some data characteristics were not reflected in the dataset. Moreover, the model's inability to account for certain aspects that it was not designed to analyse or detect from an account statement has also contributed to the lower subjective performance ratings.

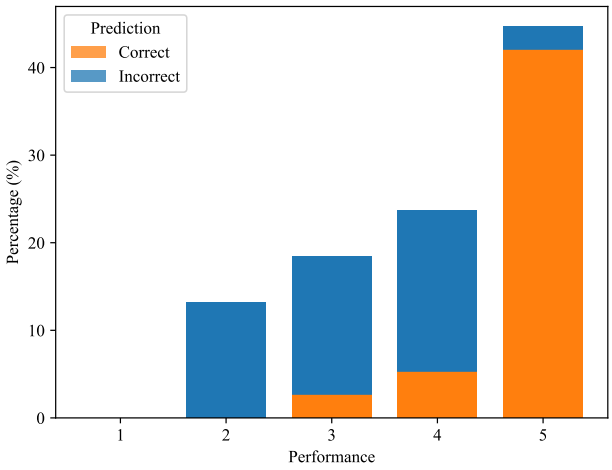


Figure 25. Perception of XAI model performance on a Likert scale.

The usefulness feedback results (see Figure 26) reveal that in approximately 2/3 of the cases the output was deemed useful, as the model effectively highlighted the correct aspects, thus fulfilling its advisory function. Intriguingly, even in instances where the model's predictions deviated from the actual outcomes, the results depict an acceptable satisfaction rate. It should be noted that the usefulness responses strongly correlate (≈ 0.82) with the performance ratings. However, this correlation is not universally consistent, as some applications exhibited lower performance evaluations due to users' expectations, but the output was still considered valuable.

Feedback on the model's decisions indicates that, in many cases, the decisions align with the actual outcomes and are considered useful, reflecting the real situation accurately. However, there are instances where the model fails to consider specific factors or misinterprets the

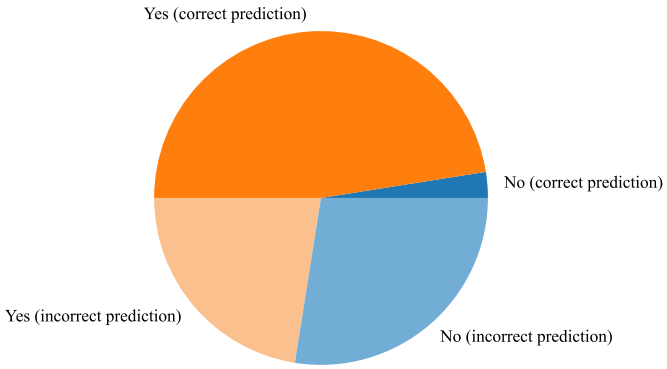


Figure 26. Perception of XAI model usefulness on a binary scale.

data, such as salary recognition, gambling transactions, and the end of financial obligations or country-specific income. Additionally, the model is not able to recognise account seizures or payment defaults. In some cases, the model's decisions do not match the actual decisions, but the provided explanations are still considered helpful or partially accurate. In general, although the model performs well in many cases, improvements in handling specific aspects could enhance its effectiveness in decision making.

5.3 System Integration

In light of the model's advisory nature, rather than offering definitive decisions, there exists an opportunity to interpret its probabilistic output in the Credit Decision System. As mentioned above, the primary outcomes of the current automated process are either automatic approval/rejection or redirecting applications for manual review. Therefore, a viable approach involves using the model to acquire the probability of loan rejection and, based on a predefined cutoff threshold, subsequently directing an application toward manual examination. This method allows for the incorporation of human expertise in cases where the model's prediction falls within the so-called "uncertainty" region, thereby ensuring a more informed and balanced decision-making process.

It is possible to analyse and evaluate the model output after employing various uncertainty thresholds. Analysing the outcomes of a consumer loan model, for instance, reveals that the model exhibits considerable certainty in the majority of its predictions, as probability values predominantly cluster near 0 or 1. For example, if the uncertainty or manual review

threshold is set to be within the range of 20% to 80%, the algorithm will still yield over 80% coverage of data automatically classified; the accuracy of the classified instances in this case will be approximately 90%, with the F_1-score over 82% (see Figure 27).

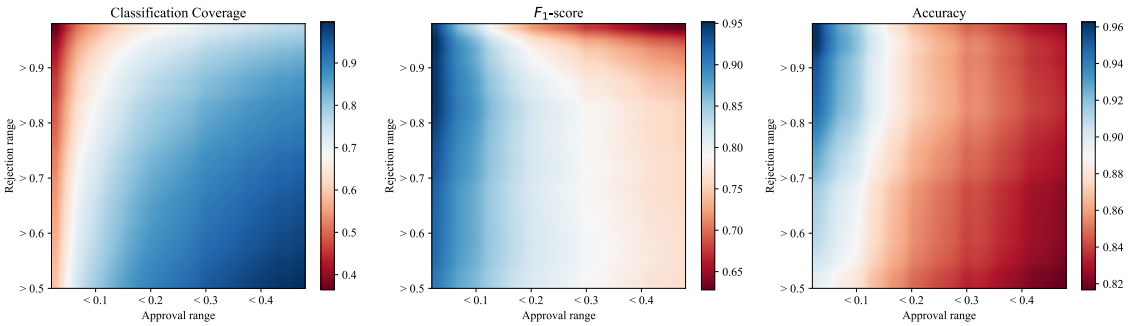


Figure 27. Performance of consumer loan model on empoying different probability cutoffs.

Consequently, the developed microservice has proven to be integrable with the Credit Decision System, where the XAI model analysis is requested and processed within the dedicated check, that is then displayed to users (see Figure 28). Initially, the initial limits for automatic decision making were established within the range of 20% to 90%: instances that exhibit rejection probabilities below 20% result in a passed account statement check, while those surpassing 90% yield a non-determining failed check, marking the need for a manual examination; the intermediate cases are labelled as the uncertain neutral outcome. To enhance the adaptability, the logic for adjusting aforementioned thresholds during the application's runtime was developed. This flexibility enables product managers to fine-tune these parameters in response to evolving requirements or emerging insights.

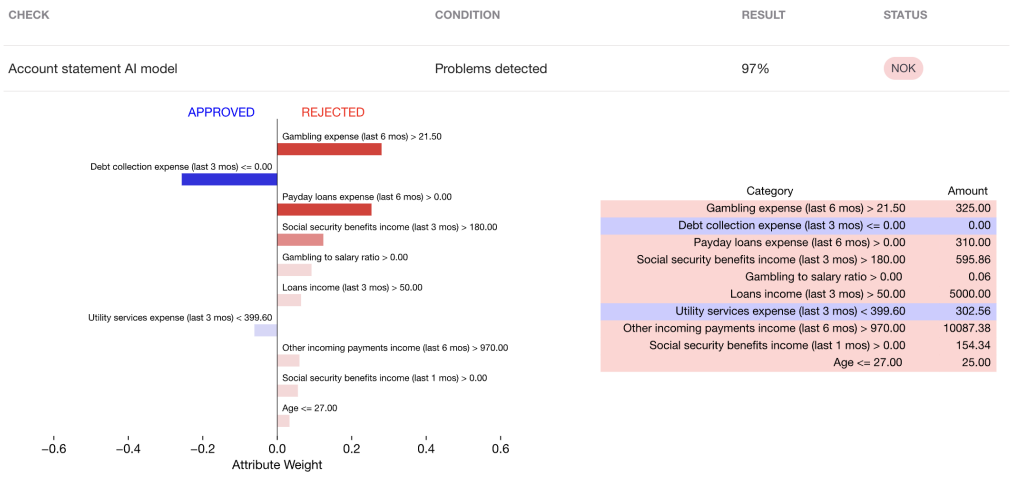


Figure 28. Check within the CDS displaying XAI model output.

5.4 Discussion

5.4.1 Limitations

Although remarkable results have been achieved in terms of model performance and the model has been successfully integrated into the existing business process, several limitations must be acknowledged. The main challenges encountered in the study were identified as follows.

- *Limited scope and feature space*. In this study, the primary focus was on the analysis of applicant-provided data: primarily, the account statement and with an allowed option of utilising personal information. Nonetheless, the loan application process involves the acquisition of supplementary data from a multitude of internal and external systems. By broadening the scope of features to include more variables, it is believed that the model's predictive performance would experience significant enhancement.
- *Data insufficiency*. The main limitation of the available data is the inability to easily acquire the exact reason for rejecting the loan application from the system, and thus detect whether an application was rejected specifically due to the account statement. Although it is theoretically possible to examine each rejection reason comment manually, this would entail a very time-consuming undertaking.
- *Data volume*. Despite being one of the largest financial institutions, the company still faces the issue of limited volumes of data available for analysis, especially in consumer financing. This scarcity of data may hinder the effectiveness of machine learning algorithms in developing high-quality generalisations, affecting the model's overall performance.
- *Human factor*. The current version of AI model encounteres difficulty identifying the data instances where the final outcome is not unambiguous. This phenomenon can be attributed to the inherent subjectivity associated with the decision making, as the data is susceptible to interpretation by the loan managers.
- *XAI technique performance*. Evaluation of the quality of the explanation has revealed that although the results are quite satisfactory, there remains considerable room

for improvement. LIME, a model-independent XAI technique, possesses certain advantages that are considered important in the scope of this study; however, it may not be able to capture the intrinsic behaviour of the model as accurately as model-specific interpretability methods. Thus, the application of such model-specific techniques could potentially yield superior outcomes in terms of explanation quality.

5.4.2 Further Improvements

A primary way for enhancement entails addressing the fundamental principle of input quality determining output quality. One possibility is the usage of more thorough feature engineering techniques. Incorporating domain knowledge and input from experts, such as risk analysts, is valuable in directing the feature engineering process and may help the models uncover more complex relationships and improve prediction performance. An implementation of so-called adaptive learning is also crucial: allowing the models to continuously improve based on new incoming data would help to account for any changes in trends and policies that may impact loan eligibility predictions.

Numerous data quality issues have emerged throughout the thesis, presenting challenges for both the analysis and interpretation of the results. One potential improvement option involves including the structured attribute that provides the reason for rejecting the loan application and is easily distinguishable by AI systems. This additional data would remove the need for manual examination of individual rejection commentaries. At the time of publication of this study, the proposed idea was referred to the business side for consideration and is currently being analysed.

A prevalent approach to assessing model performance is the cost-balancing technique, which systematically weighs the costs of different types of errors to determine the optimal trade-off between them. This approach is particularly popular in the financial sector, as it facilitates informed decision-making by addressing the economic implications of various outcomes [47]. However, modelling the precise costs of each error requires a deeper business analysis conducted at higher levels within the organisation.

Future research could explore and employ more sophisticated algorithms to improve the predictive performance and interpretability of the model. Modern advanced approaches (e.g., convolutional neural networks or deep learning) have demonstrated their efficacy in resolving a variety of challenging problems. These methodologies exhibit the capacity to manage data of intricate structure (such as account statements) ensuring that no granularity loss occurs during the analysis process. However, the application of these approaches was deemed too complex and resource-intensive in the scope of this thesis.

6 Summary

This thesis has demonstrated the development and integration of an XAI model for consumer financing applications within a large financial institution, where the use of machine learning is a novel approach. The primary focus of the research was on the analysis of data provided by the applicant, specifically account statements, to predict the outcome of the loan application. Incorporating the model into the current business process provides an additional layer of control, enabling the organisation to further mitigate the risks associated with loan eligibility assessment.

The developed model was based on a state-of-the-art XGBoost ensemble algorithm and used LIME for explainability. The performance of the XAI model was evaluated through both quantitative metrics and subjective human-centred evaluations from loan managers. The model was subsequently integrated into the Credit Decision System to seamlessly query the AI-powered predictions and provide a convenient solution for end-users. Despite the remarkable results achieved in the defined scope, there are several limitations and potential areas for improvement. Future work may be conducted within the company to address these limitations: improve accessible data quality, explore more advanced algorithms, and use more sophisticated XAI techniques.

Overall, this thesis has successfully exhibited the value of incorporating an XAI model into the loan eligibility assessment process. The research carried out emphasises the importance of explainability and transparency in AI models used within the financial sector. Through the provision of interpretable predictions, XAI systems can build trust and facilitate better communication between AI models and their end users. Ultimately, the achievements of this study have contributed to facilitating the attainment of the company's strategic goals to leverage machine learning as a means to enhance application throughput, further automate business processes, and simplify tasks for employees, highlighting the potential benefits and impact of integrating XAI models in the financial industry.

References

[1] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. "Consumer credit-risk models via machine-learning algorithms". In: *Journal of Banking & Finance* 34.11 (2010), pp. 2767–2787. ISSN: 0378-4266. DOI: 10.1016/j.jbankfin.2010.06.001.

[2] Ch. Naveen Kumar et al. "Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector". In: *2022 7th International Conference on Communication and Electronics Systems (ICCES)* (2022), pp. 1007–1012. DOI: 10.1109/ICCES54183.2022.9835725.

[3] Maisa Aniceto, Flavio Barboza, and Herbert Kimura. "Machine learning predictivity applied to consumer creditworthiness". In: *Future Business Journal* 6 (Dec. 2020). DOI: 10.1186/s43093-020-00041-w.

[4] Ram Machlev et al. "Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities". In: *Energy and AI* 9 (2022), p. 100169. ISSN: 2666-5468. DOI: 10.1016/j.egyai.2022.100169.

[5] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012.

[6] Xolani Dastile, Turgay Celik, and Moshe Potsane. "Statistical and machine learning models in credit scoring: A systematic literature survey". In: *Applied Soft Computing* 91 (2020), p. 106263. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2020.106263.

[7] Hussein Abdou and John Pointon. "Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature." In: *Int. Syst. in Accounting, Finance and Management* 18 (Apr. 2011), pp. 59–88. DOI: 10.1002/isaf.325.

[8] Francisco Louzada, Anderson Ara, and Guilherme B. Fernandes. "Classification methods applied to credit scoring: Systematic review and overall comparison". In: *Surveys in Operations Research and Management Science* 21.2 (2016), pp. 117–134. ISSN: 1876-7354. DOI: 10.1016/j.sorms.2016.10.001.

[9] Siddharth Bhatore, Lalit Mohan, and Y. Raghu Reddy. "Machine learning techniques for credit risk evaluation: a systematic literature review". In: *Journal of Banking and Financial Technology* 4.1 (Apr. 2020), pp. 111–138. ISSN: 2524-7964. DOI: 10.1007/s42786-020-00020-3.

[10] Maryan Rizinski et al. “Ethically Responsible Machine Learning in Fintech”. In: *IEEE Access* 10 (2022), pp. 97531–97554. DOI: 10.1109/ACCESS.2022.3202889.

[11] Basel Committee on Banking Supervision. *Basel III: A global regulatory framework for more resilient banks and banking systems*. Bank for International Settlements. Dec. 2010. URL: https://www.bis.org/publ/bcbs189.htm.

[12] Riigikogu. *Creditors and Credit Intermediaries Act*. June 2022. URL: https://www.riigiteataja.ee/en/eli/ee/513062022002.

[13] Riigikogu. *Money Laundering and Terrorist Financing Prevention Act*. Nov. 2017. URL: https://www.riigiteataja.ee/en/eli/ee/517112017003.

[14] LHV. *Principles of Processing Customer Data*. URL: https://www.lhv.ee/en/principles-of-processing-customer-data.

[15] Ethem Alpaydin. *Introduction to Machine Learning*. 3rd ed. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2014. ISBN: 978-0-262-02818-9.

[16] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, 2018. ISBN: 0262039249.

[17] Jason Brownlee. *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-to-end*. Jason Brownlee, 2016. ISBN: 9798540446273.

[18] Amazon SageMaker. *Machine Learning with Amazon SageMaker*. URL: https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-mlconcepts.html.

[19] Guo Yufeng. *The 7 Steps of Machine Learning*. Ed. by Towards Data Science. 2017. URL: https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e.

[20] Ronen Israel, Bryan T. Kelly, and Tobias J. Moskowitz. “Can Machines ’Learn’ Finance?” In: *Journal of Investment Management* (2020). DOI: 10.2139/ssrn.3624052.

[21] Jaydip Sen, Rajdeep Sen, and Abhishek Dutta. “Machine Learning in Finance – Emerging Trends and Challenges”. In: *Machine Learning - Algorithms, Models and Applications* (2021). DOI: 10.48550/ARXIV.2110.11999.

[22] Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. “Classification of imbalanced data: a review”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 23.4 (2009), pp. 687–719. DOI: 10.1142/S0218001409007326.

[23] Bartosz Krawczyk. “Learning from imbalanced data: open challenges and future directions”. In: *Progress in Artificial Intelligence* 5.4 (2016), pp. 221–232. ISSN: 2192-6360. DOI: 10.1007/s13748-016-0094-0.

[24] Haibo He and Edwardo A. Garcia. “Learning from imbalanced data”. In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

[25] Yiwen Shi et al. “Improving Imbalanced Learning by Pre-finetuning with Data Augmentation”. In: *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*. Ed. by Nuno Moniz et al. Vol. 183. Proceedings of Machine Learning Research. PMLR, Sept. 2022, pp. 68–82. URL: https://proceedings.mlr.press/v183/shi22a.html.

[26] Alberto Fernández et al. “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary”. In: *Journal of Artificial Intelligence Research* 61.1 (Jan. 2018), pp. 863–905. ISSN: 1076-9757. DOI: 10.1613/jair.1.11192.

[27] Pooja Pant and Prakash Srivastava. “Cost-Sensitive Model Evaluation Approach for Financial Fraud Detection System”. In: *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*. 2021, pp. 1606–1611. DOI: 10.1109/ICESC51422.2021.9532741.

[28] Yang Liu et al. “Alike and Unlike: Resolving Class Imbalance Problem in Financial Credit Risk Assessment”. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. CIKM ’20. Virtual Event, Ireland: Association for Computing Machinery, 2020, pp. 2125–2128. ISBN: 9781450368599. DOI: 10.1145/3340531.3412111.

[29] Charu C. Aggarwal. *Data Mining - The Textbook*. Springer, 2015. ISBN: 978-3-319-14141-1. DOI: 10.1007/978-3-319-14142-8.

[30] Iqbal Muhammad and Zhu Yan. “Supervised Machine Learning Approaches: A Survey”. In: *Soft Computing Models in Industrial and Environmental Applications*. 2015. DOI: 10.21917/ijsc.2015.0133.

[31] Salim Dridi. *Supervised Learning - A Systematic Literature Review*. Apr. 2022. DOI: 10.31219/osf.io/tysr4.

[32] Linwei Hu et al. “Supervised Machine Learning Techniques: An Overview with Applications to Banking”. In: *International Statistical Review* 89.3 (2021), pp. 573–604. ISSN: 0306-7734. DOI: 10.1111/insr.12448.

[33] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785.

[34] Susmita Ray. “A Quick Review of Machine Learning Algorithms”. In: *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. 2019, pp. 35–39. DOI: 10.1109/COMITCon.2019.8862451.

[35] Jarrod West and Maumita Bhattacharya. “Intelligent financial fraud detection: A comprehensive review”. In: *Computers & Security* 57 (2016), pp. 47–66. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2015.09.005.

[36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why should I trust you?": Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144. DOI: 10.48550/arXiv.1602.04938.

[37] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2022. URL: https://christophm.github.io/interpretable-ml-book/.

[38] Scott M. Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. DOI: 10.48550/arXiv.1705.07874.

[39] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. “Consistent Individualized Feature Attribution for Tree Ensembles”. In: *Computing Research Repository* abs/1802.03888 (2018). DOI: 10.48550/arXiv.1802.03888.

[40] David Alvarez-Melis and Tommi S. Jaakkola. *Towards Robust Interpretability with Self-Explaining Neural Networks*. 2018. DOI: 10.48550/arXiv.1806.07538.

[41] Jason Brownlee. *Nested Cross-Validation for Machine Learning with Python*. Nov. 2021. URL: https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/.

[42] Ajitesh Kumar. *Python - Nested Cross Validation for Algorithm Selection*. Aug. 2020. URL: https://vitalflux.com/python-nested-cross-validation-algorithm-selection/.

[43] Jun Huang et al. “The Analysis and Development of an XAI Process on Feature Contribution Explanation”. In: *2022 IEEE International Conference on Big Data (Big Data)*. 2022, pp. 5039–5048. DOI: 10.1109/BigData55660.2022.10020313.

[44] *Shapash – Make Machine Learning Models Transparent and Understandable by Everyone*. URL: https://maif.github.io/shapash/.

[45] Jonas Wanner et al. “A social evaluation of the perceived goodness of explainability in machine learning”. In: *Journal of Business Analytics* 5.1 (2022), pp. 29–50. DOI: 10.1080/2573234X.2021.1952913.

[46] Pedro Lopes et al. “XAI Systems Evaluation: A Review of Human and Computer-Centred Methods”. In: *Applied Sciences* 12.19 (Sept. 2022), p. 9423. ISSN: 2076-3417. DOI: 10.3390/app12199423.

[47] Charles Elkan. “The Foundations of Cost-Sensitive Learning”. In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI’01. Seattle, WA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 973–978. ISBN: 1558608125.

Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I, Artjom Pahhomov

1. grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Introducing an Explainable AI Model for Loan Eligibility Prediction: A Case Study of AS LHV Pank", supervised by Sven Nõmm and Indrek Arandi
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

08.05.2023

[1]The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.