Explaining black box decision-making

Master Thesis – Maarten Wijnands
EY Financial Services Technology Consulting
Digital & Emerging Technologies

UNIVERSITY OF TWENTE.

EY
Building a better working world

Explaining black box decision-making

Adopting explainable artificial intelligence in credit risk prediction for P2P lending

Author	M.P.J. (Maarten) Wijnands
Organization	University of Twente Faculty of Behavioural, Management and Social Sciences MSc. Industrial Engineering and Management
Specialization	Financial Engineering and Management
Examination Committee	Berend Roorda *Lead supervisor* Joerg Osterrieder *Supervisor*
Company	Ernst & Young
Department	Financial Services Technology Consulting
External Supervisors	Marco Campolo *Consultant - Digital & Emerging Technologies* Jiri Lammerts van Bueren *Manager - Digital & Emerging Technologies*
Date	September, 2021

MANAGEMENT SUMMARY

Within the financial services sector P2P lending platforms are rapidly expanding, since they offer the borrower an attractive alternative option for a loan compared to traditional lenders. In recent years it has caught the interest of researchers to develop models aiming to reduce the financial risks whilst increasing the related profits. Academics and practitioners are increasingly applying Artificial Intelligence (AI) and Machine Learning (ML) models in credit risk prediction to classify the creditworthiness of borrowers. These complex machine learning models have already shown remarkable performance but supporting real-world finance applications remains challenging. To a great degree because the increased performance comes at the cost of a clear interpretation and lack of explainability of these models.

In this paper, we explore three state-of-the-art post-hoc model agnostic explainable AI (XAI) techniques named Local Interpretable Model Agnostic Explanations (LIME), Anchors, and Shapley Additive exPlanations (SHAP) to assess their effect on interpretably and explainability. To do this, we come up with an exploratory framework called ARRGUS to assess the impact of the XAI techniques, as no assessment framework exists for it at present. We have based ARRGUS on the existing regulations of the Netherlands, the European Commission, and the GDPR law on the use of algorithm decision making. ARRGUS is composed of the indicators Accuracy, Readability, Robustness, Generalizability, Usability, and Stability to assess the effect of the XAI techniques.

To assess the effect of the XAI techniques we apply them to ML-based credit scoring models. The ML models consist of both transparent and black box models and are trained on a classification problem for determining whether a customer is creditworthy. We trained the ML models on an open-access dataset provided by the P2P lending platform LendingClub based on the period 2007-2018. We prepare the data by using some feature engineering techniques and apply the Synthetic Minority Oversampling TEchnique (SMOTE) to deal with the class imbalance problem in the dataset.

We use LIME and Anchors to explain the same instances locally since both techniques are only suitable for explaining individual predictions. We use SHAP to explain these same instances locally and explain the working of the ML models using global explanations. We present multiple comparisons by comparing the explanations for the same instances between the different XAI techniques. On top of that, we discuss the results in detail to see if explanations are in line with financial logic and compare them with the provided input data to get a better understanding of the functioning of the XAI techniques.

Our results show that all three XAI techniques provide fairly consistent explanations that are in line with financial logic and are supported by the input data. We observed some dominant features in all three techniques, which strengthens our confidence in stable outcomes provided by the techniques. Based on ARRGUS, the SHAP technique scores best on the indicators and is the most compliant.

We conclude that XAI techniques generate explanations that are understandable for all users involved or affected by the outcome of ML models and certainly add value to the outcome by indicating whether and to what extent each input parameter has contributed to the outcome of the prediction. XAI techniques show promising and useful results for improving the explainability of decision-making in AI models, however, these XAI techniques still need to overcome some practical challenges to support real-world finance applications.

While every dataset and use case is unique and has its own characteristics, the applied XAI techniques and exploratory framework ARRGUS in our research provide a helpful starting point for further research in the explanation and assessment of the working of different ML models.

Keywords: Explainable AI, Peer-to-Peer lending, Credit risk prediction, Machine Learning, LIME, Anchors, SHAP

PREFACE

This thesis marks the end of my master Industrial Engineering and Management at the University of Twente and finalizes my time as a student. During my time as a student, I have been able to continuously develop myself, had many accomplishments, and most importantly had a wonderful time. Both the bachelor's and master's programs of Industrial Engineering and Management at the University of Twente, combined with all the experience I gained, prepared me perfectly for the first steps of my professional career. By this means, I would like to take the opportunity to express my gratitude to all persons involved during the last period of me being a student at the University of Twente.

First of all, I would like to thank my lead supervisor Berend Roorda for his guidance during my thesis. His questions and feedback allowed me to improve the quality of my thesis and keeping me with the right focus. Not only does this apply to my thesis, but his supervision also helped me a lot during the whole specialization track of Financial Engineering and Management. I would also like to thank my supervisor Joerg Osterrieder for his in-debt knowledge and support that sharpened my view on the thesis.

Furthermore, I would like to thank EY for allowing me to conduct my master thesis within the Financial Services Technology Consulting department. I would like to thank my supervisor Marco for all his advice and directions throughout the period of my master thesis and Jiri for the opportunity to already work alongside him during the period of my master thesis. Through his support, I learned a lot from all the knowledge they shared on the financial world and giving me a good view of the job of a consultant in this department. I would like to thank all other colleagues that I have worked with for their support and the opportunities they gave me to work already alongside them.

Finally, I would like to thank my family and friends for supporting me not only during this last phase of writing my thesis but throughout my entire university period. To this end, I am looking forward to continuing to work within the Financial Services Technology Consulting department of EY.

I hope you find this thesis informative and a pleasure to read.

Maarten Wijnands
Utrecht, September 2021

CONTENT

Management summary i

Preface ii

Content iii

List of Figures v

List of Tables vi

List of Abbreviations vii

1. Introduction 1
1.1 Introduction to the Company 2
1.2 Problem Context 3
1.3 Research Design 4
1.4 Assumptions and scope 5
1.5 Methodology and Thesis Outline 6

2. Theoretical Review 8
2.1 Traditional credit risk modeling 8
2.2 Differences between traditional and P2P credit risk modeling 8
2.3 Explainable AI 9
2.3.1 The definition of Explainability and XAI 10
2.3.2 Taxonomy of explainability approaches 11
2.4 Post-hoc explainability techniques for machine learning models 13
2.5 Properties of Human-friendly Explanation 15
2.6 Conclusion on theoretical review 16

3. Evaluation Framework 17
3.1 Regulation on financial model explainability 17
3.1.1 Regulation by financial institutions in the Netherlands 17
3.1.2 Regulation by the European Commission 18
3.1.3 General Data Protection Regulation framework 18
3.1.4 Summarizing regulation on financial model explainability 18
3.2 Evaluating model explainability 19
3.2.1 Levels of evaluation of explainability 19
3.2.2 Properties of explanations 20
3.3 Model explainability evaluation framework: ARRGUS 22
3.4 Conclusion on the Evaluation Framework 22

4. Data & Models 25
4.1 Credit scorecard construction 25
4.2 Classification algorithms for credit scoring 25
4.2.1 Transparent classifiers 26
4.2.2 Black Box classifiers 26
4.3 Dataset description 28
4.4 Configuration of the data 28

4.4.1 Data leakage 28
4.4.2 Data cleaning 29
4.4.3 Data correlation 30
4.4.4 Data transformation 32
4.4.5 Normalizing the data 33
4.4.6 Resampling approach 33
4.5 Performance Criteria 34
4.5.1 Confusion Matrix 34
4.5.2 Accuracy 35
4.5.3 Receiver operating characteristics 35
4.5.4 Matthews Correlation Coefficient 36
4.5.5 G-Mean 36
4.6 Machine Learning Classification Results 37
4.7 Comparison with other LendingClub studies 38
4.8 Conclusion on Data & Modeling 39

5. Experimental results 40
5.1 Explainable model algorithms 40
5.1.1 LIME 40
5.1.2 Anchors 41
5.1.3 SHAP 42
5.1.4 The customer instances that we explain 43
5.2 Applying LIME to explain Local Instances 43
5.2.1 Explanation of LIME 44
5.2.2 Interpretation of LIME 46
5.2.3 Conclusion on LIME 52
5.3 Applying Anchors to explain Local Instances 53
5.3.1 Explanation of Anchors 53
5.3.2 Interpretation of Anchors 54
5.3.3 Conclusion on Anchors 60
5.4 Applying SHAP to explain Local Instances 61
5.4.1 Explanation of SHAP – Local 61
5.4.2 Interpretation of SHAP – Local 62
5.4.3 Conclusion on SHAP – Local 65
5.5 Applying SHAP to explain Global Instances 66
5.5.1 Explanation of SHAP – Local 66
5.5.2 Interpretation of SHAP – Global 69
5.5.3 Conclusion on SHAP - Global 71
5.6 Applying the ARRGUS framework to the XAI techniques 72
5.7 Conclusion on the Experimental results 73

6. Conclusion 74

7. Discussion and Further Research 76
7.1 Further Research based on our study 76
7.2 Discussion topics on model-agnostic post-hoc XAI 77
7.3 Contribution of this research 78

Appendix A. 79

References 80

LIST OF FIGURES

Figure 1: Methodology of this research. 6
Figure 2: The number of total publications whose title, abstract, and/or keywords refer to the field of XAI during recent years. *Data retrieved from Scopus (April 19th, 2021) by using the search terms indicated in the legend. 10
Figure 3: The scope of Explainable Artificial Intelligence, source: (Miller T. , 2019). 11
Figure 4: Conceptual diagram showing the different post-hoc explainability approaches available for an ML model Source: (Arrieta, et al., 2020). 14
Figure 5: Structure of ANN with one hidden layer 27
Figure 6: Simplified example of an RF 27
Figure 7: An example of an SVM 27
Figure 8: Heatmap of the correlation between numerical features from the LendingClub dataset, rounded off to one decimal place. 30
Figure 9: Correlation between the feature loan status and the other features 31
Figure 10: Example of a transformation of the categorical feature application type into a numerical feature. 32
Figure 11: Total amount of loans over the years. 32
Figure 12: An example of the working of SMOTE. Based on a k-nearest neighbors algorithm SMOTE connects with other minority class samples (black dots) and synthetically generates new instances (red dots) on these connections. Source: (Hu & Li, 2013) 34
Figure 13: ROC curve based on the result of our ML classifiers. The minimal AUROC score is 0.5 and represents the performance of a random classifier and the maximum value of 1.0 would correspond to a perfect classifier 36
Figure 14: Our Decision Tree visualized. With all the different branches is becomes unreadably on the size of this page, however it gives the reader an idea of the models working. 37
Figure 15: Example to present intuition for LIME, in which the dashed line is the learned linear regression explanation model that locally explains the bold red cross based on the instances that are captured by the learned linear regression explanation model. Source: (Ribeiro, Singh, & Guestrin, 2016). 40
Figure 16: Example of Anchors in which the presence of the words “not bad” almost guarantees a prediction of positive sentiment and the words “not good” a negative sentiment. Source: (Ribeiro, Singh, & Guestrin, 2018). 41
Figure 17: Comparison of two problems explained by Anchors and LIME. Perturbation space D is shown by a circle which both models use as their area/input to explain the model. LIME explains the prediction result by learning the line. Anchors uses a “local region” to learn how to explain the model. As we can see the “local region” refers to better construction of the generated data set for an explanation. Source: (Ribeiro, Singh, & Guestrin, 2018). 42
Figure 18: SHAP values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value E[f(x)] that would be predicted if we didn’t know any features, to the current output f(x). Source: (Lundberg & Lee, 2017). 43
Figure 19: Results from the LIME explainer for each ML classifier on Customer B. We observe incorrect predictions by the RF and ANN model. 46
Figure 20: XGBooster explanation for “Fully Paid” for Customer A by LIME. 47
Figure 21: Total counts of loans based on the earliest credit reported by the customer against the percentage of these loans that are Charged Off. In this figure, we also show the effect of SMOTE on the training set. Data is from our training set. 49
Figure 22: Total counts of loans based on the FICO score of the customer against the percentage of these loans that are Charged Off. Data is from our training set. 50
Figure 23: LR explanation for “Fully Paid” for Customer A by LIME. 50

Figure 24: ANN explanation for “Charged Off” for Customer A by LIME. 51
Figure 25: Results from the Anchor explainer for each ML classifier for Customer B. We observe an incorrect prediction by the RF model. 54
Figure 26: LR explanation for “Fully Paid” for Customer A by Anchors. 54
Figure 27: Total counts of loans based on the interest rate of the customer against the percentage of these loans that are Charged Off. Data is from our training set. 55
Figure 28: XGBooster explanation for “Fully Paid” for Customer A by Anchors. 55
Figure 29: Total counts of loans based on the loan amount of the customer against the percentage of these loans that are Charged Off. Data is from our training set. 56
Figure 30: Total counts of loans based on the DTI ratio of the customer against the percentage of these loans that are Charged Off. Data is from our training set. 56
Figure 31: Interactive dashboard of the Anchors framework. 58
Figure 32: SVM explanation for “Fully Paid” for Customer A by Anchors 59
Figure 33: Total counts of loans based on the number of mortgage accounts of the customer against the percentage of these loans that are Charged Off. Data is from our training set. 59
Figure 34: Results from the SHAP explainer. We observe an incorrect prediction for the RF and ANN model. 62
Figure 35: SHAP local explanation predicted “Fully Paid” for Customer A for the LR model 63
Figure 36: SHAP local explanation predicted “Fully Paid” for Customer A for the XGBooster model 63
Figure 37: SHAP local explanation predicted “Charged Off” for Customer A for the Neural Network model 64
Figure 38: Results from the SHAP explainer on the global impact 69

LIST OF TABLES

Table 1: Outline of the research 7
Table 2: Summary of the most important and applicable regulation on explainable models 19
Table 3: our proposed model explainability evaluation framework ARRGUS 24
Table 4: Sample features from LendingClub dataset 29
Table 5: Description of the final dataset 33
Table 6: Distribution of the target variable Loan Status 33
Table 7: Description of the training dataset before and after applying SMOTE 34
Table 8: Confusion Matrix with an explanation of all its classes 35
Table 9: Models, parameters settings, and performance-based on different metrics 38
Table 10: Results of other studies on the LendingClub dataset 39
Table 11: Two customer instances, Customer A and B, from different classes that we use in our research to see how the different XAI techniques are working. The features are presented with their transformed values, the values the ML/XAI models use, and their actual values for reader convenience. *We only listed the categorical features that are present in the user instances. 44
Table 12: All categorical features with their possible values presented. For each feature, the amount of loans with this feature in the training set is shown. Also, the percentage of these loans that are “Charged Off” are shown. 47
Table 13: Overview of the score per indicator for each XAI technique. 72

LIST OF ABBREVIATIONS

ACC	Accuracy
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic
DT	Decision Trees
EAD	Exposure at Default
FCRA	Fair Credit Reporting Act
FPR	False Positive Rate
GDPR	General Data Protection Regulations
LDA	Linear Discriminant Analysis
LGD	Loss Given Default
LIME	Local Interpretable Model-Agnostics Explanations
LR	Logistic Regression Analysis
MCC	Matthews Correlation Coefficient
ML	Machine Learning
P2P	Peer-to-Peer
PD	Probability of Default
RF	Random Forest
ROC	The Receiver Operating Characteristic
SHAP	Shapley Additive exPlanations
SMOTE	Synthetic Minority Over-Sampling Technique
SVM	Support Vector Machines
TPR	True Positive Rate
XAI	Explainable Artificial Intelligence
XGB	Extreme Gradient Boosting

1. INTRODUCTION

Presently, the financial services sector represents one of the major big data sources. The global digital lending platform market size was valued at 4.87 billion USD in 2020 and is expected to expand at a compound annual growth rate of 24.0% from 2021 to 2028 (Grand View Research, 2021). The outbreak of the COVID-19 pandemic has had a positive impact on the growth of the digital lending platform market since institutions have to better meet the needs of their customers.

Researchers try to develop models aiming to reduce the financial risks whilst increasing the related profits based on the available data. One of these risks is credit risk (Basel Committee on Banking Supervision, 2020). Credit risk is the risk that a loss will be experienced because of default by the counterparty in a transaction (Hull, 2018) and represents about 60% of the total risks that banks face (Buehler, Freeman, & Hulme, 2008). Credit risk is also the main risk that Social Lending Platforms, better known as *Peer-to-Peer* (P2P) lending, face.

Peer-to-peer (P2P) lending enables individuals to obtain loans directly from other individuals by cutting out the financial institution as the middleman. This makes it possible for any number and size of individual lenders to participate in the fundraising process of the borrower. In the Netherlands, the volume of P2P lending is limited with a total of 147.5 million EUR in 2018, though it is rapidly increasing (Ziegler, et al., 2020). For factors explaining the increasing role of P2P platforms in finance, we refer the reader to the work of (Giudici, 2018). Credit risk analysis for the lenders' risk concerning their investment is mainly based on the probability of the borrowers' failure to pay back the loans. The estimation of this probability is the credit risk assessment, the so-called credit scoring. The credit risk assessment problem of financial operations, including those supported by social lending platforms, is usually modeled as a binary classification problem based on debt repayment (Moscato, Picariello, & Sperlí, 2021). Fully paid loans are denoted as “0”, while default loans are represented as “1”.

There are two types of predictive models in credit risk assessment: i) Statistical approaches and ii) Artificial Intelligence methods (Namvar, Siami, Rabhi, & Naderpour, 2018). Statistical approaches have been developed for a long time, for example, by classifying solvent and insolvent companies using financial statement data (Altman, 1968). Since then, researchers have sought to improve bankruptcy forecasting models using various statistical approaches, applying logistic regression analysis to default estimation (Ohlson, 1980). Statistical approaches remain popular because of their high accuracy and ease of implementation; however, they have a major drawback since they do not properly cover non-linear effects among different variables (Moscato, Picariello, & Sperlí, 2021).

Therefore, current academics and practitioners are exploring and increasingly applying artificial intelligence (AI) and machine learning (ML) models in credit risk management (Giudici, Financial data science, 2018). Algorithms can be used to classify the creditworthiness of counterparties, since credit risk analysis is like pattern-recognition problems, by improving upon traditional models that are based on simpler multivariate statistical techniques (Kruppa, Schwarz, Arminger, & Ziegler, 2013). The application of these AI algorithms seems promising and complex machine learning models have already shown remarkable performance and made their way into a large number of systems (Miller T. , 2019) (Barboza, Kimura, & Altman, 2017).

The downside of this revolutionary performance is that it comes at the cost of a clear interpretation of the models' inner workings (Freitas, 2014). These models are better known as *Black Box* models and are described as systems that hide their internal logic to the user (Guidotti, et al., 2018). In real-world scenarios, there are numerous instances known which show that society

cannot rely on black box models because of their lack of transparency and the systematic bias they have shown (Larson, Mattu, Kirchner, & Angwin, 2016) (Liang, et al., 2018). The absence of proper explanations also has ethical implications, reported in the *General Data Protection Regulations* (GDPR), and approved by the European Parliament. Art. 22 of the GDPR provides restrictions on decisions based solely on automated processes, including profiling, which concerns or affects the data subject (General Data Protection Regulation, 2020). This means that GDPR introduces a right to meaningful explanations when one is subject to automated AI systems. The Fair Credit Reporting Act (FCRA) of 1970 requires lenders to explain the models they use to approve and deny credit applicants (Federal Trade Commission, 2012). Both GDPR and FRCA make black box models not suitable in regulated financial services without meaningful explanations. To meet the requirement of meaningful explanations, significant efforts are being made to make black box models more trustworthy and controllable by humans (Nassar, Salah, Rehman, & Svetinovic, 2019).

One of the most promising state-of-the-art approaches to meet this requirement is using the so-called *eXplainable Artificial Intelligence* (XAI) techniques, especially in the domain of being *model-agnostic* (Moscato, Picariello, & Sperlí, 2021). Model-agnostic means that the explanation is separated from the model and the explanation is extracted post-hoc by treating the original model as a black box (Ribeiro, Singh, & Guestrin, 2016). A model is called agnostic when it is technologically neutral and can be applied to the predictive output, regardless of which model generated it (Bussmann, Giudici, Marinelli, & Papenbrock, 2021). This involves learning an interpretable model on the predictions of the black box model, distressing inputs, and seeing how the black box model reacts, or both (Ribeiro, Singh, & Guestrin, Model-Agnostic Interpretability of Machine Learning, 2016). By separating the explanations from the original model, the constraint of restricting the model to be interpretable can be overcome. Otherwise, the result will be a limited original model which is less flexible, accurate, and usable.

In this research, we assess if the current state-of-the-art post-hoc model-agnostic XAI techniques are a viable solution in P2P lending to explain the decision for credit risk prediction. To achieve this, the XAI techniques will be tested on ML classifiers that are trained to predict if a loan will be repaid based on an open-access dataset from the P2P lending platform LendingClub. We evaluate the XAI techniques in what way and to what extent they can contribute to improving the explainability of decision-making by AI models.

1.1 Introduction to the Company

The research has been conducted with the professional guidance and support of the Amsterdam office of Ernst & Young (EY). EY is a worldwide firm with revenues of 31.4 billion USD and offers services in five areas: Assurance, Tax & Law, Consulting, Strategy & Transaction, and Core Business Services with its 247.000 employees globally. In the Netherlands, EY consists of 5000+ employees of which around 900 are active in Consulting. This research is carried out within Financial Services - Technology Consulting, and specifically within the team ‘Digital & Emerging Technologies’ (D&ET). D&ET provides services focused on the use of both new and existing technologies in the financial services sector. Results from this research provide the company with the latest insights on the application of AI models that can help to better serve its clients.

1.2 PROBLEM CONTEXT

Nowadays, there is a shift towards using black box models instead of simpler linear model architectures (Shim, 2019). As stated before, black models are less suitable in financial services, mainly because of their explanation problem.

One of the current solutions to this explanation problem is to use transparent models, also called "simple" statistical learning models, like linear models and simple decision trees. These types of models have the possibility of inspecting the components directly, such as a path in a decision tree or the weight of a specific feature in a linear model (Ribeiro, Singh, & Guestrin, Model-Agnostic Interpretability of Machine Learning, 2016) (Bussmann, Giudici, Marinelli, & Papenbrock, 2021). However, providing high interpretability could result in limited predictive accuracy. On the other hand, complex ML classifiers like neural networks provide high predictive accuracy at the expense of limited interpretability and currently require more research to be understood. Therefore, we focus on the more complex ML classifiers in this research.

Decision-making by AI models has become too complicated in human terms, which makes their adoption in many sensitive disciplines difficult, raising concerns about an ethical, privacy, fairness, and transparency perspective (Islam, Eberle, & Ghafoor, 2019). Instead of using transparent models, an alternative approach in machine learning is post-hoc evaluation in the form of model-agnostic models as we already introduced.

A lot of interest has been directed to the understanding of black box models via post-hoc explanations. However, the transition from explaining traditional models to explaining black box models is complicated. Problems within the P2P lending domain are mainly caused by:

- Non-uniform terminology (Arrieta, et al., 2020) (Ribeiro, Singh, & Guestrin, 2016) (Islam, Eberle, & Ghafoor, 2019);
- No assessment framework for model explainability (Bhatt, Weller, & Moura, 2020) (Singh, Sengupta, & Lakshminarayanan, 2020) (Islam, Eberle, & Ghafoor, 2019) (Nassar, Salah, Rehman, & Svetinovic, 2019) (Setzu, et al., 2021) (Molnar, Casalicchio, & Bischl, 2020) (Scholbeck, Molnar, Heumann, Bischl, & Casalicchio, 2020);
- Data quality problems in P2P lending (Zhang, Wang, Zhang, & Wang, 2020) (Namvar, Siami, Rabhi, & Naderpour, 2018) (Namvar & Naderpour, 2018) (Zhou, Zhang, & Luo, 2018) (Malekipirbazari & Aksakalli, 2015) (Ha, Lu, Choi, Nguyen, & Yoon, 2019).

Non-uniform terminology: what is explainability?

Throughout the literature, non-uniform terms are used for concepts such as model explainability, interpretability, and transparency. Some consider these as the same and stick to one, others differentiate among them, and still, others use them ambiguously. Until there is a consensus on the concrete definition of these terms, the quantification of explainability will remain another open and difficult challenge (Islam, Eberle, & Ghafoor, 2019). This research cannot solve this problem, but it provides the most used definitions of the terms often used in XAI communities in Chapter 2.

Evaluating XAI explanations: what model explains it better?

Post-hoc explanations are receiving great interest in understanding black box machine learning models since this type of explanation offers more freedom in modeling choices compared to transparent models (Bhatt, Weller, & Moura, 2020). Post-hoc explainability algorithms can be divided into Model-Agnostic and Model-Specific (Arrieta, et al., 2020). We focus on Model-Agnostic explainability algorithms in this research. Given the many candidate explanation functions, it is difficult to pick the explanation function that best captures the working of the model. The main problem is that there is no framework available to assess how well the model is explained.

Data quality of P2P lending platform: quantity versus quality?
Within P2P lending platforms, different challenges arise concerning traditional credit risk predictions methods due to the sparse and imbalanced data. Next to that, the default risk in P2P lending platforms increases compared to the traditional methods because of a lender, possibly, not being able to effectively evaluate the risk level of the borrowers. The main challenge concerns the evaluation of loan applicants' creditworthiness due to the lack of borrower's credit history whose results could not improve by adding more features. Therefore, P2P platforms produce a large amount of unlabeled data that require analysis for supporting lenders' real-time decisions (Moscato, Picariello, & Sperlí, 2021).

It has become clear that more research is needed on the working and decision-making process of AI models to become more transparent. Without a proper explanation, it is impossible to guarantee that the AI model makes the correct decisions and assumptions. Even if the AI model makes the correct decisions, poor data quality can cause an incorrect decision. XAI is a promising technique, however, the lack of a proper assessment framework makes it difficult to show the added value XAI techniques can add. In our opinion, the main problem can be summarized as:

"Users of ML classifiers and those affected by its outcomes lack the tools and an evaluation framework to assess the performance of models that are being used in P2P lending credit risk prediction, causing that the impact and fairness of the decisions from ML classifiers cannot be determined adequately."

1.3 Research Design

To solve the main problem, we formulate the main research question of this research as:

"In what way and to what extent can explainable AI algorithms improve the explainability of decision-making in AI models used in P2P credit risk prediction?"

We identify 'improve the explainability' as a solution in which the model's prediction becomes more explainable by a transferable, qualitative understanding of the relationship between model input and its prediction. In this research, we develop an evaluation framework that is suitable for practical implementation to measure to assess in what way and to what extent XAI leads to more explainable models.

We identified the following research objectives to answer the main question.

Identify: currently used credit risk prediction models, regulation on the use of AI models, most used ML classifiers to predict if a loan will be repaid in a P2P platform, and most promising XAI techniques.

Develop: an evaluation framework to assess the added value of the different XAI techniques on ML classifiers.

Investigate: the difference in how explainable an ML classifier becomes by applying different XAI techniques.

Conclude: on XAI techniques as a viable solution against the limited interpretability of black box models and the potential of more explainable models.

1.4 ASSUMPTIONS AND SCOPE

This research considers some decisions to ensure that the scope is manageable, valuable, and can be completed within time constraints. Therefore, the scope of this research is limited to:

- The current available XAI techniques that are classified as post-hoc. This research, therefore lets pre-model and in-model outside of the scope. Pre-model methods are independent of the actual model and In-models are explainable models that are integrated into the model itself. Both these types of models are too specific and are more concentrated on the actual development of an explainable model. The actual development of an explainable machine learning model is not feasible, given the time constraints on this research.

- The current available model-agnostic algorithms from the literature can be classified as a component of XAI. Based on (Arrieta, et al., 2020), this research lets model-specific algorithms outside of the scope. Model-Specific algorithms are outside of the scope since it would require this research to focus on a single ML classifier, while this research develops an evaluation metric from a broad perspective to assess the explainability of more ML classifiers. Therefore, it is decided to focus on model-agnostic algorithms.

- The currently available ML classifiers in the domain of credit risk prediction, specific to P2P platforms. The choice for P2P is because it is an emerging market and therefore research will likely help the market to become more mature. Next to that, the P2P market must base its models on different data compared to the traditional credit risk prediction models that financial institutions are using.

- Lastly, for this research, we are dependent on publicly available market data. Based on the availability of the dataset from LendingClub, period 2007-2018, the decision is made to focus on credit risk predictions in P2P platforms. This dataset is also used in other studies, for example: (Moscato, Picariello, & Sperlí, 2021).

For this research, a couple of assumptions have been made to simplify the real world, whilst keeping the research realistic. These assumptions are the following:

- As of 31-12-2020, LendingClub has retired its P2P platform (LendingClub , 2021). To the best of our knowledge, this retirement has nothing to do with the performance of the P2P platform. Since the dataset for this research contains data until 2018, we assume that LendingClub is a market example for other companies that are active as P2P lending platforms.

- As stated by (Islam, Eberle, & Ghafoor, 2019), the quantification of explainability is an open challenge until there is a consensus on the concrete definition of terms as explainability and transparency. We clarify those terms in Chapter 2 and assume that these are sufficient as a definition for this research.

1.5 Methodology and Thesis Outline

The use of ML classifiers in the financial services sector, and especially the use of XAI techniques, are still a relatively new topic, meaning that there is limited scientific literature available about the applications of XAI techniques on ML classifiers within credit risk prediction in a P2P lending domain. Therefore, we additionally consult research reports from other sectors on XAI. For the credit risk predictions that are currently used, we consult the related work in the literature. At the moment of writing the European Commission has released both a whitepaper on a European approach to AI (European Commission, 2020) and an Ethics Guideline for trustworthy AI (High-level expert group on artificial intelligence, 2019). In combination with the information, we can receive from the Dutch regulators, De Nederlandsche Bank (DNB) and Autoriteit Financiële Markten (AFM), we come up with the evaluation framework ARRGUS. We train the found ML classifiers on the publicly available dataset of the P2P platform LendingClub since a lot of information is available to support our research with this dataset. To assess the effect of the XAI techniques we apply them to ML-based credit scoring models. After which we will assess the effectiveness of XAI techniques based on the ARRGUS framework. All these steps are required to enable answering the main research question and are captured in Figure 1.

Table 1 presents the research questions with their research objective and the general thesis outline. In this chapter, we elaborated on the motivation for using post-hoc model-agnostic XAI techniques based on the problems that occur in the current situation. In Chapter 2 we discuss the current methods for credit risk prediction, an in-depth literature review on XAI techniques and we look at the properties of human-friendly explanations from a social science perspective. In Chapter 3 we develop the valuation framework ARRGUS to be able to assess the XAI techniques. In Chapter 4 we discuss all the choices made regarding the dataset and the chosen ML classifiers and present their results when trained on the dataset. In Chapter 5 we experiment with implementing the XAI techniques on the different ML classifiers and evaluate the results based on the ARRGUS framework. In Chapter 6 we present our conclusions, and the discussion and recommendations for further research are elaborated on in Chapter 7.

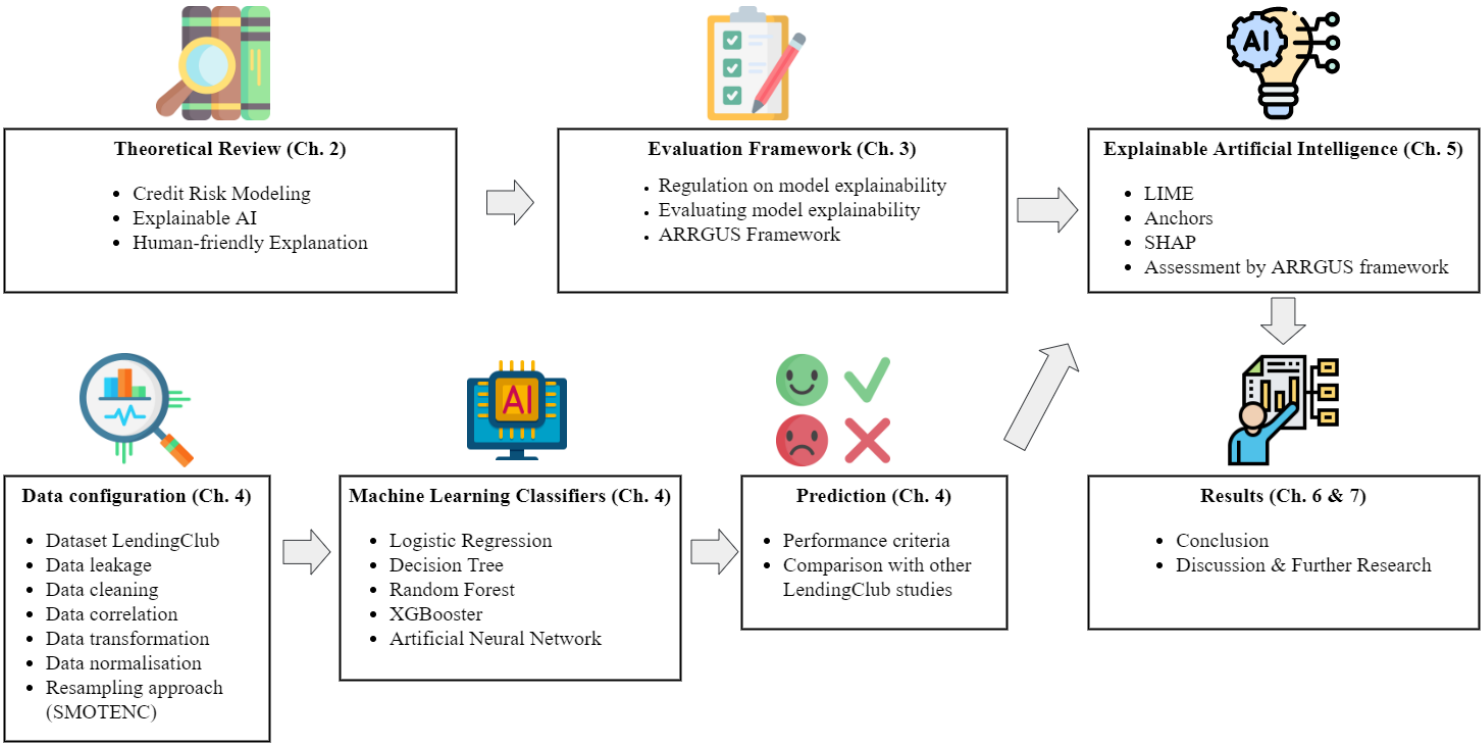


Figure 1: Methodology of this research.

Chapter	Research questions	Research objective
1. Introduction	What are the current challenges with using AI models in P2P lending credit risk prediction?	Summarize the current challenges with the use of AI models in P2P lending credit risk prediction.
2. Theoretical Review	What are the differences in credit risk prediction for traditional institutions versus P2P lending platforms?	To gain insight into credit risk prediction for P2P lending platforms.
	What is XAI and how can it play a role in credit risk prediction?	To gain insight into XAI and how it can effectively be used for credit risk prediction.
3. Evaluation Framework	How can we design a framework in which we can assess the explainability of a classifier?	Design a valuation framework to assess the explainability of an XAI technique outcome.
4. Data & Models	Which data and what ML classifiers do we need as a basis to experiment with the evaluation framework?	Select the most suitable dataset and ML classifiers to experiment with the different XAI techniques and be able to validate the outcomes of the ML classifiers.
5. Experimental results	Which XAI technique performs the best on explainability, based on our valuation framework ARRGUS?	Have an overview of how the different XAI techniques are performing on a P2P lending credit risk prediction dataset.
6. Conclusion	"In what way and to what extent can explainable AI algorithms improve the explainability of decision-making for ML classifiers used in P2P credit risk prediction?"	
7. Discussion and Further Research	Discussion and Further Research on the application of XAI in P2P credit risk prediction.	

Table 1: Outline of the research

.

2. THEORETICAL REVIEW

2.1 TRADITIONAL CREDIT RISK MODELING

Financial lending institutions, lenders, governments, and other players that participate in the financial lending market seek to develop and use models to efficiently assess the probability that the borrower, also called the counterparty, will show some undesirable behavior in the future (Barboza, Kimura, & Altman, 2017). These credit scoring systems aim to prevent bad debt loss by identifying, analyzing, and monitoring customer credit risk (Kruppa, Schwarz, Arminger, & Ziegler, 2013). The benefits obtained by developing a reliable credit scoring system are: i) reducing the cost of credit analysis, ii) enabling faster decision, and iii) ensuring credit collections and diminishing possible risk (Nanni & Lumini, 2009).

Lending institutions make use of credit scoring systems to provide them with the *Probability of Default* (PD) of the counterparty and to satisfy a minimum-loss principle for their sustainability. Therefore, credit scoring systems are important in the decision-making for credit applications to manage credit risks and are directly influencing the amount of non-performing loans that can lead to bankruptcy (Munkhdalai, Munkhdalai, Namsrai, Lee, & Ryu, 2019). Credit models from traditional financial institutions estimate creditworthiness based on a set of explanatory variables, also called features, from application forms, customer demographics, and transactional data from customer history (Lessmann, Basesens, Seow, & Lyn C. Thomas, 2015).

In regulated financial lending, The Basel III Capital Accord requires financial institutions to estimate, respectively, the *Probability of Default (PD)*, the *Exposure at Default (EAD)*, and the *Loss Given Default (LGD)* as inputs in the measurement of credit risk. PD models are well researched compared to EAD and LGD models and are used as the performance indicator for this research.

The main methods to develop PD models are classification and survival analysis. Survival analysis estimates not only whether but also when a counterparty might default. On the other hand, classification analysis represents the classic approach and benefits from an unmatched variety of modeling methods to estimate whether a counterparty defaults (Lessmann, Basesens, Seow, & Lyn C. Thomas, 2015).

Within classification analysis the credit risk models can be classified as a binary classification problem: borrowers that have a high probability of performing financial obligations are assigned to a “good credit” group and those that have a low probability of performing financial obligations to a “bad credit” (Akkoç, 2012).

2.2 DIFFERENCES BETWEEN TRADITIONAL AND P2P CREDIT RISK MODELING

Over the years, traditional financial institutions have divided their reference markets over specific business activities, increasing their expertise and the accuracy of their ratings. Differently, P2P platforms are based on a ‘universal’ banking model, which is a model that encompasses all banking activities, and makes developing an accurate rating model a more difficult task (Giudici, Financial data science, 2018).

However, P2P lending as a financial model has been studied extensively in recent years. Just like the classical financial institutions, credit risk evaluation in P2P lending commonly involves statistical approaches and machine learning methods that aim to predict the creditworthiness of

borrowers by considering loan evaluation as a binary classification problem. Compared to the plentiful literature on loan evaluation for traditional banking institutes, there are a limited number of studies on credit risk prediction in P2P lending (Namvar & Naderpour, 2018).

The main challenge for a lender in a P2P lending marketplace is making an appropriate risk assessment that can support its decision-making. The lenders aim to assess the expected return and credit risk of each loan accurately and this can be done by using traditional loan evaluation models, which are a subset of the earlier described classification methods. However, classification methods may be too unsophisticated to meet the needs of personal lenders in P2P lending (Guo, Jiang, Chen, Li, & Luo, 2019).

The traditional loan evaluation techniques used by financial institutions assume a balanced dataset and distribution of misclassifications. However, P2P lending usually occurs in situations with a high level of information asymmetry. Meaning that lenders do not have complete information about the borrowers' credit history. The result is an imbalanced dataset and can make it difficult for the model to effectively discriminate between good borrowers and potential defaulters (Namvar, Siami, Rabhi, & Naderpour, 2018). Next to that, P2P lending often contains irrelevant and redundant features which reduce the classification accuracy. Applying a feature engineering strategy helps to eliminate redundant features and select an optimal subset of relevant features (Ha, Lu, Choi, Nguyen, & Yoon, 2019).

Malekipirbazari & Aksakalli (2015) already found that using a machine learning approach is much more effective than relying on the existing financial metrics, like FICO grades, which LendingClub provides to help lenders making loan investment decisions. Nevertheless, it remains difficult to design new models for credit risk prediction in P2P lending due to the high number of missing values and class-imbalanced data (Moscato, Picariello, & Sperlí, 2021). Therefore, it tries to compensate for the high number of missing traditional used values by adding more features that can tell something about the borrower's creditworthiness like social media usage. In chapter 4 we discuss the different types of models that we used in modeling the credit risk prediction and discuss the feature engineering strategy we applied to the dataset of LendingClub. Adding more features is out of the scope of this research.

To conclude, although P2P lending is different from traditional loans from commercial banks in terms of lending style, they are both loan relations generated based on credit essentially, and the biggest risk is still the borrower's credit risks. By considering a feature engineering strategy, omitting available features, to increase the accuracy on an imbalanced dataset in P2P lending, the same methods can be applied as in credit risk modeling for traditional loans.

2.3 Explainable AI

In the previous sections, we discussed the concept behind credit risk modeling that can be used in both traditional financial institutions and P2P lending platforms. However, just showing the results if a borrower is 'good' or 'bad' is not satisfactory for most lenders and borrowers. Explaining how the models generate their answers is the next step in improving these models. The best explanation of a simple model is the model itself, since it perfectly represents itself, and is easy to understand. For complex models, the original model cannot be used as its own explanation because it is not easy to understand (Lundberg & Lee, 2017). Instead, a simpler explanation model must be used, which we define as an interpretable approximation of the original model (Lundberg & Lee, 2017).

Model-agnostic XAI is proposed as an interpretable approximation of the original model since these techniques do not limit the effectiveness of the current generation of machine learning models. XAI tries to create a collection of machine learning techniques that i) produce a more explainable model while maintaining a high level of prediction accuracy, and ii) enable humans to understand and effectively manage artificially intelligent models (Arrieta, et al., 2020). Figure 2 confirms the rising interest in XAI by showing the number of contributions in the literature on XAI in recent years.

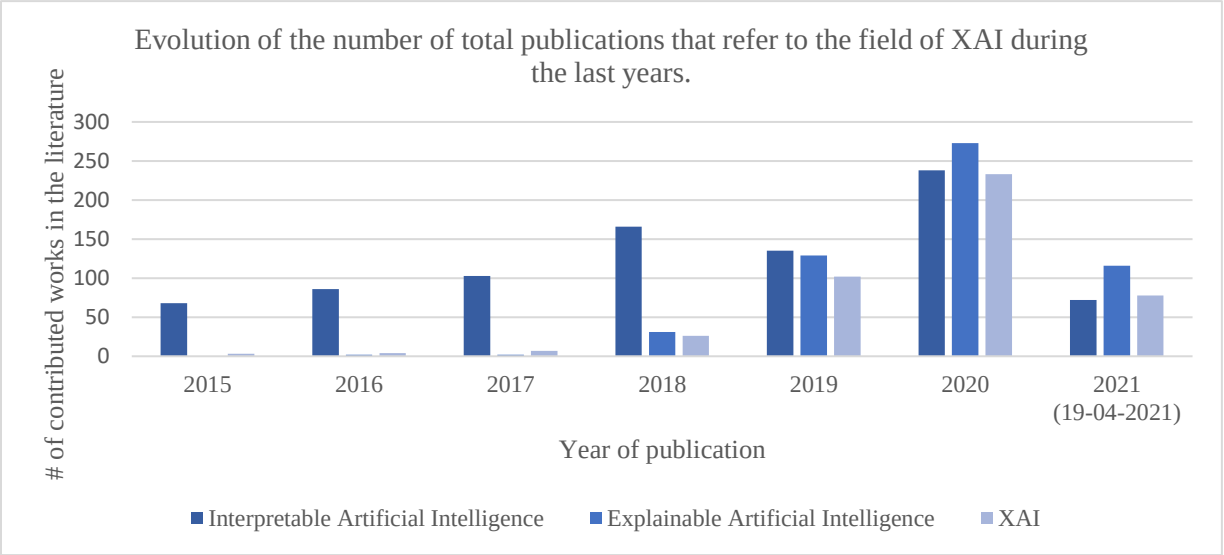


*Figure 2: The number of total publications whose title, abstract, and/or keywords refer to the field of XAI during recent years. *Data retrieved from Scopus (April 19th, 2021) by using the search terms indicated in the legend.*

2.3.1 The definition of Explainability and XAI

Before proceeding with our literature study, it is convenient to first establish a common point of understanding on what the term explainability stands for in the context of AI. We use the definition of explainability by (Islam, Eberle, & Ghafoor, 2019) because it captures clearly and extensively what an explainable model should contain. We extend this definition of explainability by the definition based on the extensive literature research by (Arrieta, et al., 2020) to get as clear a definition as possible against which we can measure the aim of this research. We define explainability for this research as follows:

Explainability: An AI model's prediction is explainable in the extent of transferable qualitative understanding of the relationship between model input and prediction that, at the same time, is both an accurate proxy of the decision-maker and comprehensible to humans.

Given the same lack of consensus on the definition of the term Explainable Artificial Intelligence, we make use of the definition of the term given by (Gunning & Aha, 2019) as a starting point:

XAI*:* Has the goal to create a collection of new or modified ML techniques that produce explainable models that, when combined with effective explanation techniques, enable end-users to understand, appropriately trust, and effectively manage the emerging generation of AI systems.

In our research, explainability is connected to post-hoc explainability because it covers the techniques used to convert a non-interpretable model into an explainable one by using an interpretable approximation of the original model. The design of a model is beyond the scope of this research, similarly to the focus on managing the emerging generation of AI systems.

Thus, the term 'Explainable Artificial Intelligence' refers to an explanatory agent revealing underlying causes to its or another agent's decision-making. It is important to realize that the solution to explainable AI is not to just have more AI. Ultimately, it can be seen as a human-agent interaction problem. Human-agent interaction can be defined as the intersection of artificial intelligence, social science, and human-computer interaction in which XAI is just one problem within human-agent interaction, as can be seen in Figure 3.

Therefore, we specify the definition of XAI for this research to:

XAI*:* Explainable Artificial Intelligence is an algorithm that produces explainable models that details or reason the working of ML models to make their functioning clear or easy to understand in an interaction between humans and an agent.

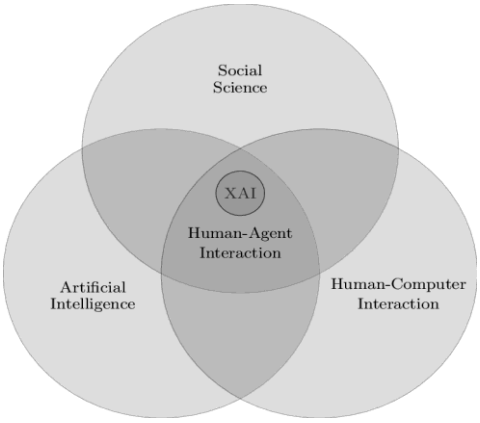


Figure 3: The scope of Explainable Artificial Intelligence, source: (Miller T. , 2019).

2.3.2 TAXONOMY OF EXPLAINABILITY APPROACHES

To classify different explainability methods the literature proposes several taxonomies. In section 2.2 we already concluded that we model credit risk as a classification problem. Therefore, we searched in the literature on XAI techniques that can be applied to a classification problem. Classification techniques are, generally, not absolute and can vary widely depending upon the characteristics of the methods; furthermore, they can be classified into several overlapping or non-overlapping classes simultaneously (Singh, Sengupta, & Lakshminarayanan, 2020). We briefly discuss different kinds of taxonomies based on the available literature, taking (Singh, Sengupta, & Lakshminarayanan, 2020) as a basis.

Model-Specific vs Model-Agnostic

Model-specific explanations methods are based on the parameters of individual models and therefore can only be used on this specific model type. As described earlier, model-agnostic means that the explanation is separated from the model, and the explanation is extracted post-hoc by treating the original model as a black box (Ribeiro, Singh, & Guestrin, Model-Agnostic Interpretability of Machine Learning, 2016). Model-agnostic models do not have direct access to the internal model weights or structural parameters and are therefore not limited to a specified model architecture.

In this research, we focus on Model-Agnostic explanation techniques.

Pre-model vs in-model vs post-model
Pre-model methods are independent of the actual model and do not depend on the specified model architecture to use it on. In-models are explainable models that are integrated into the model itself. Finally, post-models are implemented after building a model. Post-models can potentially develop meaningful insights about what exactly a model learned during the training.

In this research, we focus on Post-model, better known as Post-Hoc, explanation techniques.

Global Methods vs Local Methods
Global methods try to explain the behavior of the model in general by making use of the overall knowledge of the model, its training, and the associated data. An example of a global method is feature importance, which tries to determine the features that are in general responsible for better performance of the model among all the different features. Local methods are relevant for explaining a single outcome of the model. This can be obtained by designing methods that can explain the reason for a particular prediction or outcome. Local methods are, for example, interested in specific features and their characteristics.

In this research, we focus on Local Methods since we want to provide a borrower/lender with an explanation for its situation. However, we do not exclude global methods since the XAI technique SHAP, which will be later explained, also supports global explanations.

Surrogate Methods vs Visualization Methods
Surrogate methods include different models as a so-called 'ensemble' and are used to analyze other black-box models. The black box models can be understood better by interpreting the surrogate model's decisions by comparing the black-box model's decision and the surrogate model's decision. An example of a surrogate method is the decision tree. The visualization methods are not a different model; however, it helps to explain some components of the models by visual understanding like activation maps.

In this research, we focus on Surrogate explanation techniques.

To conclude, we will focus on post-hoc model-agnostic techniques, using local and surrogate methods. We do not exclude global and visualization methods since we believe that these can complement the other methods in gaining a more complete explanation. In the next section, we explore the different techniques covered by post-hoc model-agnostic techniques based on local and surrogate methods.

2.4 POST-HOC EXPLAINABILITY TECHNIQUES FOR MACHINE LEARNING MODELS

Post-hoc explainability targets models that are not readily interpretable by design. This can be done by applying different techniques to increase their interpretability, such as text explanations, visual explanations, local explanations, explanations by example, explanations by simplification, and feature relevance explanations techniques (Arrieta, et al., 2020). These techniques are inspired by the methods humans use to explain systems and processes by themselves. Below is a brief description of each of them.

Explanations by simplification

Explanations by simplification explain the model by building a new system based on the trained model to be explained. This new, simplified model usually follows an optimizing strategy to its antecedent functioning, while reducing its complexity, and keeping a similar performance score. The simplified model is, in general, easier to implement due to the reduction in complexity concerning the model it represents. Explanations by simplification are considered the broadest technique under the category of post-hoc model-agnostic methods (Arrieta, et al., 2020). Simplified models are, sometimes, only representations of certain sections of a model and therefore *Local explanations* are also present in the category of *Explanations by simplification.*

Most of the techniques used in *Explanations by simplification* are based on rule-extraction techniques. One of the most known contributions to this technique is that of Local Interpretable Model-Agnostics Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016), which will be further discussed in section 5.1.1. (Arrieta, et al., 2020) concludes that the popularity of model simplification is evident, given it temporally coincides with the most recent literature on XAI, including techniques such as LIME. This reveals that this post-hoc explainability approach is regarded to continue playing a central role in XAI.

Feature relevance explanations

Feature relevance explanation clarifies the inner functioning of the model by computing a relevance score for its managed variables. It aims to describe the functioning of an opaque model by ranking or measuring the influence, relevance, or importance each feature has in the prediction output by the model to be explained (Arrieta, et al., 2020). A variety of techniques is found within this category, each resorting to a different algorithmic with the same targeted goal. One of the most known contributions to this technique is that of *Shapley Additive exPlanations* (SHAP) (Lundberg & Lee, 2017), which will be further discussed in section 5.1.2.

Visual explanations

Visual explanations explain the model by applying techniques that aim at visualizing the model's behavior. According to the review of (Arrieta, et al., 2020), most literature makes use of dimensionality reduction techniques to make simple visualizations that can be easily interpreted by humans. Visual explanations are considered the most suitable way of introducing complex interactions within the variables involved in the model to users not acquainted with ML modeling (Arrieta, et al., 2020).

Local explanations

Local explanations explain the model by dividing the solution space into smaller subspaces and from there give explanations to less complex solution subspaces that are relevant for the complete model. This subspace can be a single or several outcomes.

Counterfactual examples

Counterfactual examples, also known as explanations by example, explains a single decision with a statement of how the value of the input features should change for a desirable outcome to occur. Its statement is a causal argument of the form: 'if A has not occurred then B would not have

occurred." (Nassar, Salah, Rehman, & Svetinovic, 2019). It is interesting to look up the smallest change to the feature values that would flip the outcome of the prediction. Counterfactual explanations are computed by minimizing a loss function which is composed of the norm of change in the feature vector and the norm that is composed of the difference between the instance prediction and the targeted prediction.

Text explanations
As the name suggests, text explanations explain the model by learning to generate text explanations, including generating symbols that represent the functioning of the model, that help to explain the outcomes of the model.

In this research, we use Explanations by simplification, Feature relevance explanations, Visual explanations, and Local explanations as post-hoc model-agnostic techniques. Counterfactual examples and Text explanations are out of scope for this research. A visualization of the different techniques is shown in Figure 4.

In this section, we described the post-hoc model-agnostic techniques to explain systems and processes that are based on methods humans use. We use these post-hoc techniques in our research in Chapter 5, since they are designed to be understandable by humans. However, both the research of (Molnar, Casalicchio, & Bischl, 2020) and (Islam, Eberle, & Ghafoor, 2019) state that these explanations are not necessarily human-friendly. We, therefore, discuss the properties of human-friendly explanations in the next section.

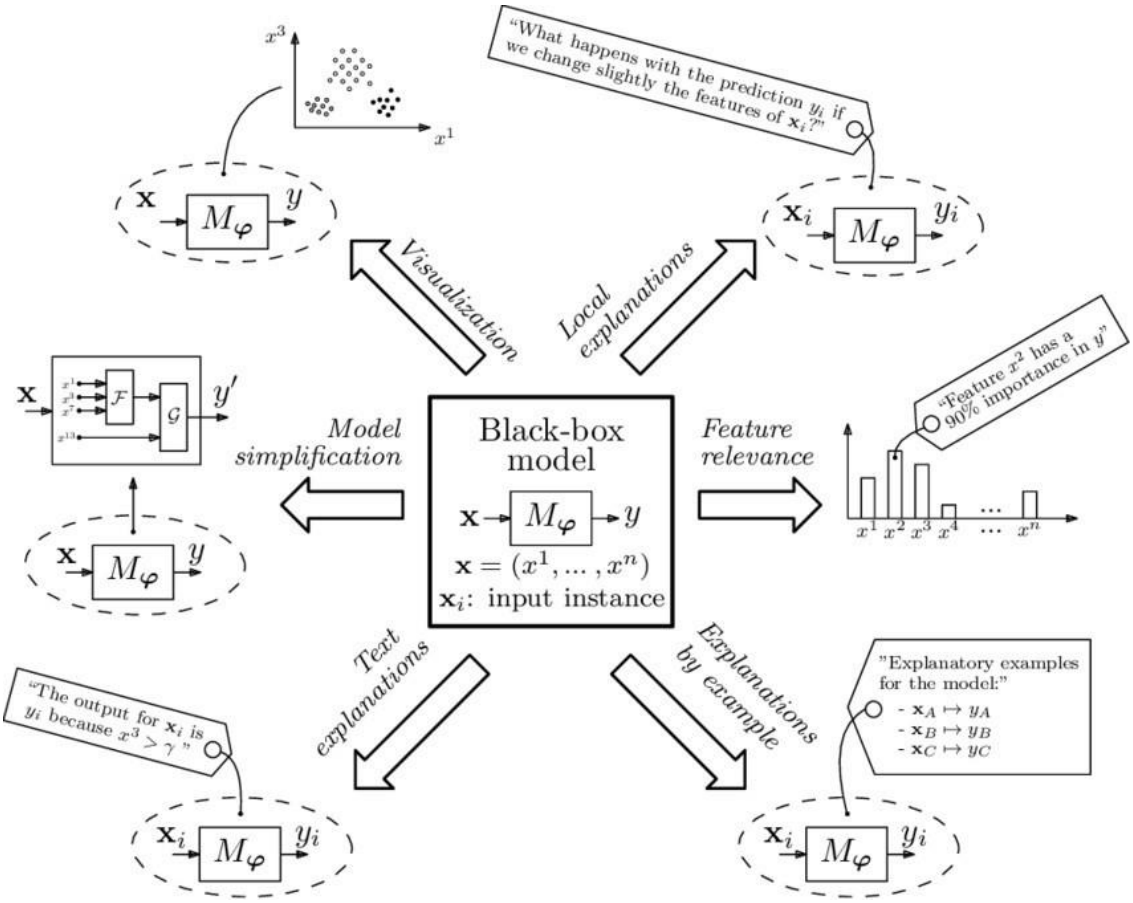


Figure 4: Conceptual diagram showing the different post-hoc explainability approaches available for an ML model Source: (Arrieta, et al., 2020).

2.5 PROPERTIES OF HUMAN-FRIENDLY EXPLANATION

As stated in section 2.3.1, XAI can be seen as a human-agent interaction problem and is based on social science interaction. Humans usually prefer short explanations that disagree with the current situation to a situation in which that event would not have occurred. Since explanations are a social interaction among the explainer and the recipient of the explanation, the social context can also be seen as an important factor for a good explanation. However, a human-friendly explanation is selective in nature and does not consider all factors for a particular prediction (Islam, Eberle, & Ghafoor, 2019).

(Miller T. , 2019) conducted a literature survey into this topic and based on their research and the research of (Islam, Eberle, & Ghafoor, 2019) we summarize the six properties that are important for explainable AI and the way explanations are constructed in a human-friendly way.

- Explanations should be **contrastive.**
  This means that people do not ask why event X happened, but they rather ask why even X happened instead of some event Y. This has important social and computational consequences for explainable AI.

- Explanations should be **selective.**
  People seldom, if ever, expect an explanation that consists of a complete and actual cause of an event. Humans are adept at selecting one or two causes from many causes to be the explanation. People tend to use inherent features rather than extrinsic features to explain the object. An inherent feature describes "how an object is established" and an extrinsic feature is for example a historical factor.

- Explanations should be **social.**
  Explanations are a transfer of knowledge, presented as part of social interaction between the explainer and the recipient of the explanation. Therefore, explanations are presented relative to the explainer's beliefs about the recipient of the explanation beliefs.

- **Probabilities** should be accompanied by a causal consequence.
  Truth and likelihood are important factors in explanations. Probabilities are also a factor that does matter, however referring to probabilities or statistical relationships in an explanation is not as effective as referring to causes. The most likely explanation is not always the best explanation for a person, and importantly, using statistical generalizations to explain why events occur is unsatisfying, unless accompanied by an underlying causal explanation for the generalization itself.

- Explanations should be **truthful & consistent.**
  An explanation should be as truthful as possible. However, selectiveness can sometimes come first which might exclude some of the true reasons.

- Explanations should be **general and feasible.**
  Acceptable explanations are general and feasible.

2.6 CONCLUSION ON THEORETICAL REVIEW

In this chapter, we focused on answering two research questions from the first of which read as follows: "*What are the differences in credit risk prediction for traditional institutions versus P2P lending platforms?".* Based on sections 2.1 and 2.2 we conclude, although P2P lending is different from traditional loans from commercial banks in terms of lending style, that they are both loan relations generated based on credit essentially, and the biggest risk is still the borrower's credit risks. By considering a feature engineering strategy to increase the accuracy on an imbalanced dataset in P2P lending, the same methods can be applied as in credit risk modeling for traditional loans.

We conclude the second research question: *"What is XAI and how can it play a role in credit risk prediction?"* as XAI being an algorithm that produces explainable models that details or reasons the working of ML models to make its functioning clear or easy to understand in an interaction between humans and an agent. This interaction is crucial for all the stakeholders in credit prediction to be able to understand why and on which basis decisions regarding credit and its associated risks are made. Based on sections 2.3, 2.4, and 2.5 this research use Explanations by simplification, Feature relevance explanations, Visual explanations, and Local explanations as post-hoc model-agnostic techniques to see if XAI can play a role for all stakeholders to better understand the working behind credit risk prediction.

In this chapter, we considered XAI with the question of devising viable criteria for evaluating the quality of explanations. Even though the usual consumer of explanations is the human end-user, frameworks are required when it is difficult to have a human in the loop to judge 'good' from 'bad' explanations. In the following chapter, we will construct a framework to be able to judge if an explanation, generated by an XAI algorithm, explains the model clearly and can thus be judged as good. We build this framework based on the information from this chapter and the guidelines given by the different regulations.

3. EVALUATION FRAMEWORK

3.1 REGULATION ON FINANCIAL MODEL EXPLAINABILITY

Regulation can be seen as the minimum standard a model must comply with. Therefore, we consult the current regulations on model explainability issued by regulators from the Netherlands and the European Commission. Next to that, we also consult the GDPR framework to understand what indicators these institutions and this regulation use to define a minimum standard for model explainability.

3.1.1 REGULATION BY FINANCIAL INSTITUTIONS IN THE NETHERLANDS

The DNB has formulated a few general principles, called 'SAFEST', for the responsible use of AI in the financial sector. These principles together constitute a framework in which companies can responsibly design the use of AI. However, compliance with these principles is not a hard requirement (DNB, 2019). The following principles are part of SAFEST:

- **Soundness**: AI applications must be reliable and accurate and operate predictably within the limits of applicable laws and regulations.
- **Accountability**: Organizations must be accountable if AI applications unexpectedly malfunction since this could harm different stakeholders.
- **Fairness**: AI applications must not unintentionally disadvantage certain groups of people.
- **Ethical**: Ensure that customers and other stakeholders are treated appropriately and not harmed using AI.
- **Skills**: Everyone in the organization must have the right level of expertise and must know the benefits and limitations of the AI systems they work with.
- **Transparency**: Organizations must be able to explain how and why they use AI in their business processes, and how exactly these applications work.

The SAFEST principles are general and leave room for their own interpretation and therefore do not provide a strong guideline. The DNB and AFM together have carried out an exploration of AI in the insurance sector and propose some points of interest on both model and social explainability (DNB & AFM, 2019). Summarizing the report, the main points to take into consideration are: i) To what extent is it possible to trace relationships between input parameters and model results and ii) what level of explainability is appropriate for that process?

The report states it is important for models that are used to accept or refuse a decision, to have the possibility to indicate for individual input parameters whether and to what extent they have contributed to the outcome of the model as well as which changes in input parameters are necessary to achieve a change in the model outcome. The reason behind this is that these kinds of models have a direct impact on the individual and are often used in automated decision-making. Social explainability goes beyond the ability to explain the technology or the outcomes of AI models. It touches on the question of whether the outcomes of models are seen as socially and ethically acceptable and fair. Having an explanation with an intuitive cause-and-effect relationship increases the social acceptance of the parameter.

To the best of our knowledge, the AFM on its own does not have any further documentation available on its view on model explainability. The AFM is optimistic that we can use the GDPR as a regulatory building block for financial regulation to provide more guidance on model explainability (Ethics in AI, a way to avoid regulation?, 2019).

3.1.2 Regulation by the European Commission

The European Commission recognizes that the transparency of algorithms is crucial since they affect more and more decisions in our lives (European Commission, 2018). Therefore, the European Commission proposed a regulatory proposal to provide AI developers, deployers, and users with clear requirements and obligations regarding specific uses of AI (European Commission, 2021). The proposed framework is a risk-based approach and consists of four layers of risk, namely: Unacceptable risk, High-risk, Limited risk, and Minimal risk.

Credit scoring is identified as a *High-risk AI system* in the category *of Essential private and public services* and therefore will be subject to strict obligations before such an AI system can be put on the market (European Commission, 2021). Below is a list of obligations the credit scoring AI system should adhere to:

- **Adequate risk assessment and mitigation systems.**
- **High quality of the datasets** feeding the system to minimize risks and discriminatory outcomes.
- **Logging of activity** to ensure traceability of results.
- **Detailed documentation** providing all information necessary on the system and its purpose for authorities to assess its compliance.
- **Clear and adequate information** to the user.
- **Appropriate human oversight** measures to minimize risk.
- **High level of robustness, security, and accuracy**.

Once the AI system will be on the market, authorities will oversee the market surveillance, its users will ensure human oversight and the monitoring of the AI system, and providers will have to have a post-market monitoring system in place. Providers and users also will have to have the opportunity to report serious incidents and malfunctioning. However, according to the European Commission, the earliest time this regulation could become applicable to operators is the second half of 2024 (European Commission, 2021).

At present, other national governments - such as the United States, the United Kingdom, Canada, China, Singapore, France, and New Zealand - are still making plans focusing on developing ethical standards, policies, regulations, or frameworks and are therefore not taken into consideration (Dutton, 2018).

3.1.3 General Data Protection Regulation framework

As stated earlier, the GDPR already includes some regulations to provide more guidance on model explainability and is often referred to as the minimum requirements that an interpretable model must meet. Considering GDPR Article 15.1(h) and Recital 71 the individual data subject, the individual whom it concerns, has the right to ask the organization how the system came to its decision. Next to that, the earlier mentioned GDPR Article 22 gives the individual the right not to be subject to a decision based solely on automated processing. Meaning that the individual may ask for a human to review the AI's decision to determine if the system made a mistake or not (General Data Protection Regulation, 2020). To be compliant with the GDPR, it requires all stakeholders to understand how the result or decision came about.

3.1.4 Summarizing regulation on financial model explainability

Based on section 3.1, we conclude that there is no clear overview or detailed framework on what makes a model explainable or how to measure this. We have consulted the most applicable institutions and regulations and summarized them in Table 2.

DNB	DNB & AFM	European Commission	GDPR
Safest Principle Soundness Accountability Fairness Ethical Skills Transparency	Model explainability For every individual input parameter whether and to what extent they have contributed to the outcome. For every input parameter which changes are necessary to achieve a change in the model outcome. Social explainability Having an explanation with an intuitive cause-and-effect relationship.	High-risk AI system: Essential private and public services Adequate risk assessment and mitigation systems. High quality of the datasets. Logging of activity. Detailed documentation on the system and its purpose for authorities. Clear and adequate information to the user. Appropriate human oversight. High level of robustness, security, and accuracy.	Article 15.1(h) & Recital 71 The individual data subject has the right to ask the organization how the system came to its decision. Article 22 The individual data subject has the right to ask for a human to review the AI's decision.

Table 2: Summary of the most important and applicable regulation on explainable models

Based on Table 2 we conclude that, from a regulatory standpoint, for a model to be explainable, it should at least present the following:

- The outcome shows for every individual input parameter whether and to what extent they have contributed to the outcome.
- The outcome shows for every input parameter which changes are necessary to achieve a change in the model outcome.
- The individual data subject can ask the organization how the system came to its decision and a human to review this decision.

The other requirements from Table 2 are achieved by fulfilling the requirements listed above or are outside the scope of this study. In the next section, we will come up with some indicators to measure model explainability to meet and extend these requirements, thereby enabling us to build a framework.

3.2 Evaluating model explainability

As we have seen in section 3.1, there is no real consensus on how to evaluate model explainability, nor it is clear how to measure this properly. Several exploratory studies have been carried out on this subject and will be used as a starting point in this section. Starting at a higher level, (Doshi-Velez & Kim, 2017) propose three main levels for the evaluation of explainability, as outlined in section 3.2.1.

3.2.1 Levels of evaluation of explainability

Application-level evaluation (real task): Put the explanation into the product and have it tested by the end-user. This level requires a good experimental setup and a human understanding of how to assess quality. How good a human would be at explaining the same decision can be used as a baseline.

Human-level evaluation (simple task): The difference from the application-level evaluation is that these experiments are not carried out with domain experts, but with non-professionals. An example would be to show a non-professional user a different explanation and let the user choose the best one.

Function level evaluation (proxy task): This level does not require humans and works best when the class of model that is used has already been evaluated by someone else in a human level evaluation. For example, if the user understands decision trees, an indicator for explanation quality may be the depth of the tree. A shorter tree would get a higher score in this case. Of course, a constraint can be added so that the predictive performance of the tree remains above a certain threshold and does not decrease too much compared to a larger tree.

We will focus on the explanations for individual predictions on the function level using the relevant properties of explanations that we consider for the model evaluation.

3.2.2 PROPERTIES OF EXPLANATIONS

We want to explain the predictions of a machine learning model. With the research of (Robnik-Šikonja & Bohanec, 2018) and (Arrieta, et al., 2020) we take a closer look at the properties of explanation methods and explanations. We will use these properties to judge how good an explanation method or explanation is. One of the challenges is to find a method to calculate and measure all these properties since that is currently not the case.

Properties of machine learning explanations
Expressive power: the structure of the explanations the method can generate. For example, an explanation method could generate IF-THEN rules, decision trees, or something else.

Translucency: describes the degree to which the explanation method relies on looking into the machine learning model. Explanation methods that rely on decomposing the internal representation of the model are highly translucent. Methods that treat the model as a black box, manipulating inputs and observing the predictions, are the lowest level of translucent. High translucency has the advantage that the method can make use of more information to generate explanations. Low translucency has the advantage that the explanation method is more portable.

Portability (or Transferability): is the range of machine learning models on which the explanation method can be used.

Algorithmic complexity: deals with the computational complexity of the method that generates the explanation.

Properties of explanation methods / Quality of explanations
Accuracy: the ability that an explanation of a given decision generalizes to other yet unseen data. For example, if explanations are in the form of rules, are these rules general, and do they cover unseen data. Low accuracy can be acceptable if the goal is to explain the working of the black-box model. In this case, only fidelity is important.

Fidelity: How well does the explanation approximate the prediction of the black box model? Having a high fidelity can be regarded as one of the most important properties of an explanation because an explanation with low fidelity is useless to explain the machine learning model. Fidelity and accuracy are closely related. A black-box model that has high accuracy and its explanation a high-fidelity score, results in its explanation also having high accuracy. Local fidelity means that

the explanation only approximates the model prediction for a subset of the data or an individual data instance. Local fidelity does not imply general fidelity.

Consistency: The degree to which similar explanations are generated on different models that have been trained on the same task. Similar models may produce similar predictions; however, explanations of similar instances may vary because of the variance of certain explanation methods. Highly consistent means that the explanations are very similar.

Stability: The degree to which similar explanations are generated for similar instances. While consistency compares explanations between models, stability compares explanations between similar instances for a fixed model. High stability means that a slight variation in one of the features of an instance does not considerably change the explanation.

Comprehensibility: The readability of explanations is difficult to define and measure since many people agree that it depends on the audience. Usually, a human can comprehend 7, plus or minus 2 pieces of information at a time (Miller G. A., 1956). Measuring comprehensibility can include the size of the explanations (e.g., number of decision rules) or by testing how well people can predict the behavior of the machine learning model from the explanation. Next to that, the comprehensibility of the features used in the explanation should also be considered.

Certainty: Does the explanation reflect the certainty of the machine learning model about its predictions? Several machine learning models give a prediction without a statement that tells how confident the model is about the correctness of the prediction. An explanation that includes the model's certainty on the 5% probability of default of one lender is as certain as the 5% probability of default on another lender, with different feature values.

Degree of Importance: To what degree does the explanation reflect the importance of features or parts of the explanation? For example, does the generated explanation reflect the importance of the explained features?

Novelty: To what degree does the explanation reflect whether an explained instance is from a new region, meaning that it is far removed from the distribution of training data. If that is the case, the model may be unreliable, and the explanation may be useless. Novelty and certainty are related in that a higher novelty is more likely to result in a low certainty of the model due to the lack of data.

Representativeness: How many instances, outcomes for different individuals, do an explanation cover? A model explanation may cover the entire model, just a part of it or only an individual prediction**.**

Fairness: From a social standpoint, explainability can be considered as the capacity to reach and guarantee fairness in ML models. In a certain literature strand, an explainable ML model suggests a clear visualization of the relations affecting a result, allowing for fairness or ethical analysis of the model at hand.

Accessibility: A minor subset of the reviewed contributions argues for explainability as the property that allows end-users to get more involved in the process of improving and developing a certain ML model.

3.3 Model explainability evaluation framework: ARRGUS

In this section, we propose an evaluation framework by which we evaluate the performance of the XAI models later in this research. Provided with the knowledge of sections 3.1 and 3.2, we conclude that there is currently no evaluation framework available that is based on measurable indicators on human-friendly explainability and considers the minimum requirements from a regulation perspective for model explainability evaluation. Using all the different properties of sections 3.1 and 3.2 on its own makes the evaluation, to our opinion, unclear and not user friendly, and is thus not sufficient for comparing different XAI models.

Therefore, we thought of what the end-user, an average borrower/lender on a P2P platform, would like to see in its explanation on the outcome of why a loan is granted or rejected. The most important factors for the end-user are that it would like to see how its score is build up, how the class of a good and bad borrower is defined, what needs to change to switch class, and potentially a visualization of the different aspects to facilitate the understanding. These conditions will at the same time help with the regulatory demand that a human must be able to justify how a decision is arrived at. We use these preferences as building blocks for our evaluation framework to determine if the XAI models help with obtaining more explainable model outcomes.

Based on the above, we propose our evaluation framework *ARRGUS* in which we combine the earlier mentioned properties into six different indicators namely: *Accuracy, Readability, Robustness, Generalizability, Usability, and Stability*. The objective of ARRGUS is to evaluate model explainability at a more general level. Future studies could look at the different indicators of ARRGUS and measure them quantitively at the level of the different properties. However, this is outside the scope of this research.

The goal behind ARRGUS is to measure to which extent the different XAI methods can fulfill the requirements mentioned earlier and validate the effectiveness of XAI models. The proposed model explainability evaluation framework ARRGUS can be found in *Table 3*.

3.4 Conclusion on the Evaluation Framework

In this chapter, we focused on answering the research question "*How can we design a framework in which we can assess explainability of a classifier?".* Based on Chapter 2, sections 3.1 and 3.2 we proposed the evaluation framework ARRGUS, by which we are able to assess the explainability of an XAI technique outcome on the indicators: Accuracy, Readability, Robustness, Generalizability, Usability, and, Stability. In the next chapter, we will discuss the different ML classifiers that are available for a classification problem and model them on the LendingClub dataset.

Indicator	Explanation	Based on properties	To what degree present? On a scale of 1 (Very Poor) to 5 (Very Good)
Accuracy	To what degree can the XAI technique explain how well the explanations reflect the behavior of the prediction model?	2.5: Explanations should be truthful & consistent. 3.1: The individual data subject can ask the organization how the system came to its decision and a human to review this decision. 3.2: Accuracy, Novelty, Translucency, Fairness, Fidelity.	Some questions to support the measurement: How well do the explanations reflect the behavior of the prediction model? How well does the XAI model generate similar explanations for similar instances? Are explanations reflecting the certainty of a model about its predictions?
Readability	To what degree can the XAI technique generate explanations that are understandable for the targeted audience group, the average borrower/lender?	2.5: Explanations should be contrastive, Explanations should be selective, Explanations should be social. 3.1: The individual data subject can ask the organization how the system came to its decision and a human to review this decision. 3.2: Accessibility, Expressive power, Comprehensibility.	Some questions to support the measurement: Can the explanation be customized to give a selective number of causes in the form of a top 5? Are the explanations understandable for the targeted audience group, the average borrower/lender?
Robustness	To what degree can the XAI technique explain whether and to what extent each individual input parameter has contributed to the outcome.	2.5: Probabilities should be accompanied by a causal consequence 3.1: The outcome shows for every individual input parameter whether and to what extent they have contributed to the outcome, the outcome shows for every input parameter which changes are necessary to achieve a change in the model outcome, the individual data subject can ask the organization how the system came to its decision	Some questions to support the measurement: Does the model give one or more causes why the customer is labeled as a good/bad credit group? Can the explanation be customized to show for each individual input parameter whether and to what extent they have contributed to the outcome? Can the explanation show for every input parameter which

		and a human to review this decision. 3.2: Degree of Importance.	changes are necessary to achieve a change in the model outcome? Can the probabilities used for the outcome be used as causes for the final explanation?
Generalizability	To what degree can the XAI technique generate similar explanations on Machine Learning models that are trained on the same task.	2.5: Explanations should be general and feasible. 3.1: The individual data subject can ask the organization how the system came to its decision and a human to review this decision. 3.2: Consistency, Representativeness.	Some questions to support the measurement: To what degree are similar explanations generated that are general and feasible?
Usability	To what degree can the XAI technique be used effectively on a range of machine learning models as an explanation method?	2.5: - 3.1: The individual data subject can ask the organization how the system came to its decision and a human to review this decision. 3.2: Algorithmic complexity, Portability.	Some questions to support the measurement: How many instances can the explanation of the XAI technique cover?
Stability	To what degree can the XAI technique generate similar explanations for similar instances and do these explanations reflect the same amount of certainty of a model about its predictions?	2.5: - 3.1: The individual data subject can ask the organization how the system came to its decision and a human to review this decision. 3.2: Stability, Certainty.	Some questions to support the measurement: To what degree are similar explanations generated on different models that have been trained on the same task?

Table 3: our proposed model explainability evaluation framework ARRGUS

4. DATA & MODELS

4.1 CREDIT SCORECARD CONSTRUCTION

We illustrate the development of a credit scorecard in the context of application scoring based on (Lessmann, Basesens, Seow, & Lyn C. Thomas, 2015) one of the most comprehensive classifier comparison studies to date. Let $x = (x_1, x_2, \dots, x_n) \in R^n$ be an n-dimensional vector with borrower application characteristics and let $y = \{0; 1\}$ be a binary variable that differentiates good ($y = 0$) and bad loans ($y = 1$). A credit scorecard estimates the probability $p(y = 1|x_i)$ that a default event will be observed for loan i. To decide on an application, a credit analyst compares the estimated default probability to a threshold τ. The gets approved if $p(y = 1|x_i) \leq \tau$ and rejected otherwise. The task of estimating $p(y = 1|x_i)$ belongs to the field of classification. A scorecard is a classification model that results from applying a classification algorithm to a dataset $D = (y_i, X_i)_{i=1}^n$ of past loans.

4.2 CLASSIFICATION ALGORITHMS FOR CREDIT SCORING

In this section, we focus on the different classification algorithms that are used as credit scoring models for predicting the default probability of new credits. In parallel with the growing credit volume of the financial sector, many different credit scoring models have been developed by banks and researchers to evaluate credit applications.

The industry standards are the traditional statistical methods: Linear Discriminant Analysis (LDA) and Logistic Regression Analysis (LR) (Akkoç, 2012) (Kruppa, Schwarz, Arminger, & Ziegler, 2013) (Lessmann, Basesens, Seow, & Lyn C. Thomas, 2015). An alternative to these models are the newer ML classifiers, like Decision Trees (DT), Random Forest (RF), Extreme Gradient Boosting (XGBooster), Support Vector Machines (SVM)) and Artificial Neural Networks (ANN) (Akkoç, 2012) (Kruppa, Schwarz, Arminger, & Ziegler, 2013) (Barboza, Kimura, & Altman, 2017) (Petropoulos, Siakoulis, Stavroulakis, & Klamargias, 2018) (Wang, Zhang, Lu, & Yu, 2020).

Both LDA and LR models are criticized because these make use of the assumption that there is a linear relationship among variables. This can result in lower predictive accuracy in credit risk prediction if there is no linear relationship present (Akkoç, 2012) (Namvar, Siami, Rabhi, & Naderpour, 2018). Looking at the last two decades, the mentioned ML classifiers are becoming more popular in research and show impressive results (Lessmann, Basesens, Seow, & Lyn C. Thomas, 2015). The analysis of (Doko, Kalajdziski, & Mishkovski, 2021) on ML classifiers for credit risk prediction shows that RF and Boosters, like XGB, are achieving the highest performance scores. The same analysis shows that the other mentioned ML classifiers are showing similar performance. In our research, we use the LR, DT, RF, XGB, SVM, and ANN as ML classifiers on which XAI algorithms are applied. The remainder of this section briefly explains the different ML classifiers that are applied in our research.

4.2.1 TRANSPARENT CLASSIFIERS

Logistic Regression Analysis

LR is a commonly used model in credit risk modeling and is one of the most popular statistical modeling techniques for classification problems, in which the probability of an outcome depends on a group of independent variables. The equation of an LR model is given by:

$$ln\frac{P(y=1 \mid x_i)}{1-P(y=1 \mid x_i)} = b_0 + \sum_m b_m x_{im} \quad (1)$$

where

$P(y_i = 1 \mid x_i)$	the probability of default for borrower i
x_{im}	a vector of the borrower-specific independent variables $x_{i1,}, \dots, x_{im}$
b_0	the intercept parameter
b_m	the regression coefficients $for\ m = 1, \dots, M$,

The objective of an LR model in credit scoring is to determine the conditional probability of a specific observation belonging to a class, given the values of the independent variables of that credit applicant (Lee & Chen, 2005). The LR performs well in many applications, however, its accuracy drops when the relationships in the data are non-linear.

Decision Tree

Based on the dataset, a decision tree can identify and derive a visualize with a series of if-then-else decision rules. The model is fitter if the tree is deeper since there are more complex if-then-else decision rules. A decision tree has a flowchart structure that mimics the activity of human thinking. This algorithm uses the tree representation to make classification choices. An example of a decision tree can be seen in Figure 6 since a Random Forest makes use of multiple decision trees.

4.2.2 BLACK BOX CLASSIFIERS

Random Forest

RF is based on decision tree models and can be used for both classification and regression problems. In the building process of decision trees, the RF models randomly choose *k* independent variables from all available variables and then uses these sub-variables to build a decision tree. This process is independently repeated *m* times to obtain *m* decision tree classifiers. The elements of each set have a label if they are correct or not. For each element, the *m* decision trees vote on a label. The label with the most votes is chosen as the preferred classification of the element (Zhang, Wang, Zhang, & Wang, 2020) (Barboza, Kimura, & Altman, 2017). Figure 6 shows an example of an RF that consists of three DTs.

Extreme Gradient Boosting

XGB is a decision-tree-based ensemble ML-classifier that uses a Gradient Boosting framework. Gradient Boosting is an approach where new models are trained to predict the errors of prior models. Something that a Random Forest does not consider. The XGBooster algorithm was developed as a research project at the University of Washington in 2016. Since then, it is being applied as the driving force for several cutting-edge industry applications. The extreme comes from a special case of boosting where errors are minimized by a gradient descent algorithm.

Support Vector Machines

The SVM method is currently one of the most popular ML algorithms for solving classification-related problems in almost every field (Zhang, Wang, Zhang, & Wang, 2020). The SVM separates the data into two regions, one for each class, by trying to locate a maximum hyperplane with the maximum margin width between instances of the two classes to avoid the misclassification of samples to the greatest extent possible (Malekipirbazari & Aksakalli, 2015). Figure 6 shows an example of an SVM. A larger margin width decreases the complexity of the model and the overall risk of errors. In practice, it is usually not possible to separate the data by a hyperplane and a soft margin is used instead. In this situation, a positive slack is added to the instances on the wrong side of the margin. This slack increase is proportional to how far the corresponding instance is from the margin. The goal is to minimize the sum of these slacks while maximizing the width of the margin (Malekipirbazari & Aksakalli, 2015). However, the SVM is less suited for large datasets because the training complexity of SVM is highly dependent on the size of the dataset.

Artificial Neural Networks

ANNs are mathematical representations inspired by the functioning of the human brain. ANN is composed of several processing elements and these elements come together within the frame of particular rules, the so-called nodes, or neurons. Generally, an ANN consists of three layers of interconnects neurons as shown in Figure 5. The number of layers can be increased by adding more 'Hidden Layers'. The first layer is called the 'Input Layer', where external information, corresponding to corresponding to independent variables in statistics, is received. From each neuron in the input layer, signals are sent to the second layer, called the "Hidden Layer". In the hidden layer, the information received from the input layer is processed. The third layer called the "Output Layer" transmits all the information outside the networks. This corresponds to a dependent variable in statistics. Since the 1990s, ANNs have been widely used in financial prediction studies and most of those studies report that prediction accuracies of ANNs are better than conventional statistical techniques, like LR and LDA.

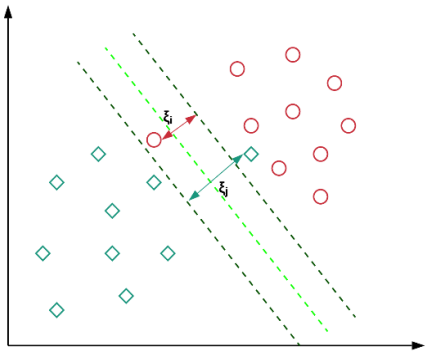


Figure 5: An example of an SVM

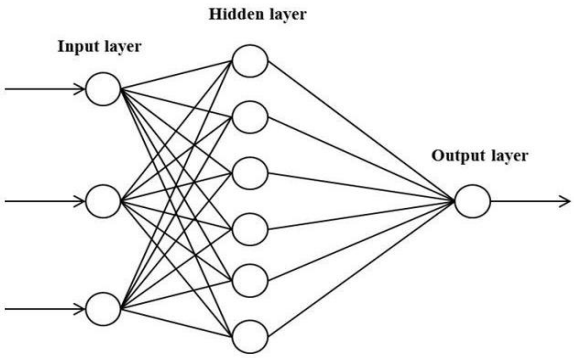


Figure 7: Structure of ANN with one hidden layer

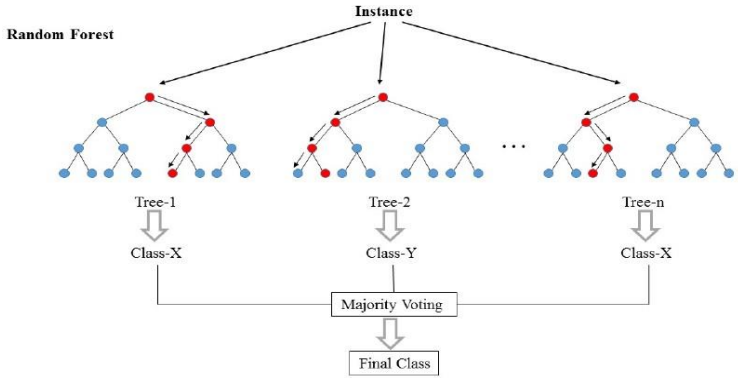


Figure 6: Simplified example of an RF

4.3 DATASET DESCRIPTION

The growth in P2P lending markets has generated a large amount of data on real-world P2P lending transactions. We used the dataset of LendingClub for our experimentation, which is obtained from the online data scientist community of Kaggle. The dataset contains over 2.25 million borrower records and 151 different features in total, all issued through the LendingClub platform. From each loan the outcome is known, either fully paid or charged off as a loss. With this information, we build the different ML classifiers that focus on achieving high accuracy in classifying customer instances as either the class "Fully Paid" or "Charged Off".

The dataset of LendingClub consists of 151 features that try to indicate the creditworthiness of the consumer to the best of its ability. The dataset from LendingClub is highly imbalanced, a common problem in credit risk prediction as described earlier, and a lot of features have missing values. This suggests that a limited number of available features can be used. Therefore, before we started configuring the data we plotted the distribution of the features to gain a better understanding of the values that are present in the dataset for each feature. Next to that, we plotted the boxplots of the continuous values to gain a better understanding of which values of the features are causing a default based on simple statistics. For the discrete features, we plotted the percentage of the total loans that are marked as "Charged Off" loans for the different values of the feature. By this, we could get a better understanding of how the distribution of the values of the features are impacting and/or causing the outcome of someone being a creditworthy borrower. To get useful input data, we performed several steps to configure the data as correctly as possible based on the information we obtained from our observations that we just described.

4.4 CONFIGURATION OF THE DATA

As discussed in section 2.2, we configure the data by applying a feature engineering strategy in our research. The goal of feature engineering is to improve data reliability by cleaning data and selecting the subset of features that have the most discriminatory power (Namvar, Siami, Rabhi, & Naderpour, 2018). In credit risk prediction, ignoring irrelevant features can increase classification accuracy and decrease computational complexity. Next to that, by applying feature selection the dimensionality of the data is reduced which helps to prevent the risk of overfitting. We applied six important steps to configure our data: removal of data leakage (4.4.1), data cleaning (4.4.2), correlation analysis (4.4.3), data transformation (4.4.4), normalizing the data (4.4.5), and oversample the minority class (4.4.6).

4.4.1 DATA LEAKAGE

We identified the features that may cause data leakage. Data leakage occurs when information is used in the model's training process which would not be expected to be available at the time when we would predict the outcome of the class in which the customer belongs, for example, by allowing the model to learn beyond the training set by also learning from the test set. This causes the predictive scores to overestimate the model's utility when run in a production environment. It sounds like "cheating" but most of the time users are not aware of it, so it is called "leakage". This 'leaking' of data defeats the purpose of having a test set since it is described as unseen data. Only the features in our dataset that are available for the lender at the moment of deciding to grant a loan to a borrower are kept in the dataset. The LendingClub has stopped as a P2P lending platform and no clear documentation is available on the available information the lender received at the time. Therefore, we have selected the features that we assumed available to the lender based on the explanation of the different features. A sample of the features we selected, including their description[1], can be found in Table 4. The total list of selected features can be found in Table A1.

[1] The LendingClub dictionary can be accessed via https://resources.lendingclub.com/LCDataDictionary.xlsx

The result is a new feature space of 31 features. We observed that only a few earlier studies that relied on data from LendingClub have considered data leakage. This could mean that their training models have made unrealistically good predictions. One of the studies that did take data leakage into account is that of (Namvar, Siami, Rabhi, & Naderpour, 2018), however, they also labeled the features grade and interest rate as leaky data. We have checked online videos and articles to conclude that both grade and interest rates were available for the lender before they decided to invest (Rose, 2018), (Hogue, 2019), (Rose, 2020). Therefore, grade and interest rate are not labeled as leaky data and kept in the dataset for this stage of our feature engineering strategy.

4.4.2 DATA CLEANING

The dataset is cleaned by removing missing and null values from the dataset. We plotted the missing values per feature based on the ratio of missing and "NaN" values. We observed that the features *employment length* and *mortgages accounts* had a too high percentage of missing values to drop these rows and required some additional work.

Feature	Description	Data Type
annual_inc	The self-reported annual income provided by the borrower during registration.	Numeric
earliest_cr_line	The month the borrower's earliest reported credit line was opened	Numeric
grade	LC assigned loan grade	Categorical
installment	The monthly payment owed by the borrower if the loan originates.	Numeric
loan_amnt	The listed amount of the loan is applied for by the borrower.	Numeric
loan_status	Current status of the loan	Categorical
mort_acc	The number of mortgage accounts.	Numeric
purpose	A category is provided by the borrower for the loan request.	Categorical
revol_bal	Total credit revolving balance	Numeric
term	The number of payments on the loan. Values are in months and can be either 36 or 60.	Categorical
title	The loan title provided by the borrower	Categorical

Table 4: Sample features from LendingClub dataset

For *employment length,* we assumed that a missing record means that the employment length is less than a year and filled the missing values with the minimum value of 0. For *mortgages accounts,* we assumed to fill the missing values with the mean value of the feature of 1.55. There are multiple ways of filling these missing values. For example, by filling the missing values with a mean value based on a strong correlation the feature has with another feature. However, we decided to keep it as simple as possible. All the remaining rows which contained missing values were dropped.

Next to removing the missing values from the dataset, we also checked the number of unique values for each feature to prevent overfitting. The features *id* and *employment title* had too many unique values and were removed from the dataset in this step.

4.4.3 DATA CORRELATION

The next step is removing the features that are highly correlated with each other. Correlated features will not necessarily worsen the model, but they will not always improve it either. We computed for both the numeric and categorical features the correlation with respect to each other, as well as the correlation to loan status to gain a better understanding of the data and its features.

The correlation between numerical features is obtained by calculating the Pearson's R correlation coefficient. It calculates the covariance between feature X and Y and divides this by the standard deviation of feature X times feature Y. The Pearson's R equation is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma X \sigma Y} \tag{2}$$

where,

$cov(X,Y)$	the covariance between feature X and Y
σX	the standard deviation of feature X
σY	the standard deviation of feature Y

We plotted the correlation between numerical features in Figure 8 in the form of a heatmap.

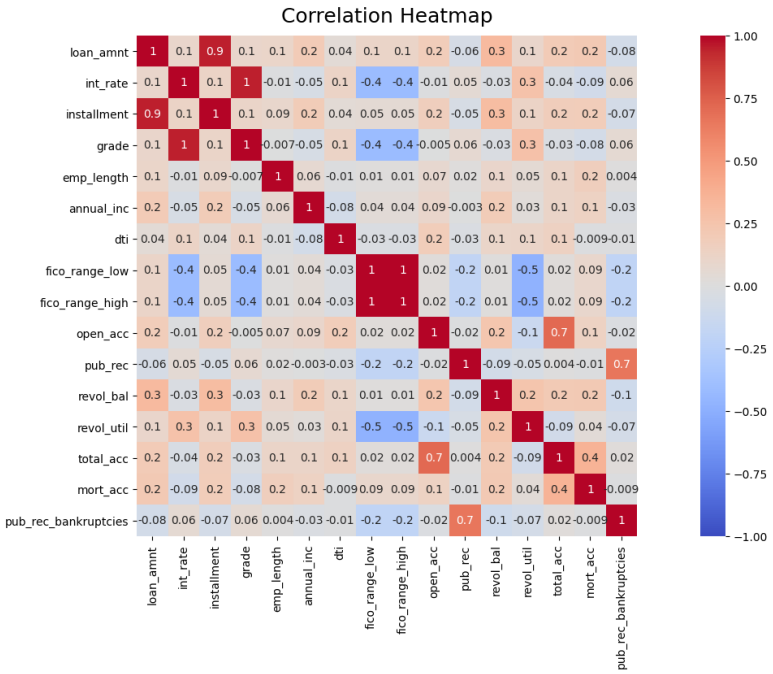


Figure 8: Heatmap of the correlation between numerical features from the LendingClub dataset, rounded off to one decimal place.

From Figure 8 we observe that the features *fico range low* and *fico range high* are perfectly correlated, and we, therefore, combined these into a new feature *fico score* with the equation:

$$fico\ score = 0.5 * \text{fico\_range}_{low} + 0.5 * \text{fico\_range}_{\text{high}} \tag{3}$$

The features *loan amount* and *installment* are also highly correlated. The description in Table A1 logically explains this correlation. The value for the installment, the monthly payment, depends on the total loan amount and explains the correlation between the two features. The features *interest rate* and *grade* are perfectly correlated, and we decided to drop the feature *grade*.

The correlation between categorical features is obtained by using Cramér's V correlation coefficient, which is based on a chi-square test. The equation for Cramér's V is:

$$V = \sqrt{\frac{\chi^2}{N \min(c-1, r-1)}} \tag{4}$$

where,

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{5}$$

N	is the sample size involved in the test
c	number of columns
r	number of rows
O_i	observed frequency counts in each category
E_i	expected frequency counts in each category
k	number of categories

The highly correlated categorical features, with a correlation value equal or greater than 0.9, are *grade* and *subgrade*, *purpose* and *title*, and *zip code* and *address state*. We decided to remove the feature *subgrade* to prevent the overfitting of the model. We decided to remove the feature *title* since this information can already be obtained from the fourteen fixed values of the feature *purpose*. In the United States, there is less regulation on the usage of personal information and therefore the features *zip code* and *address state* are present in the dataset. We decided to drop the feature *zip code* since it contains more unique values compared to *address state,* which could lead to overfitting the model. For Europe, it is questionable whether *address state* may be used since it contains personal information and may create conflicts to be compliant with the GDPR. Our experiments showed no significant predictive power from the feature *address state* and therefore we decided to also drop this feature.

The correlation to our target variable *loan status* is presented in Figure 9. From this figure, we observe that the features *interest rate*, *term,* and *dti* are having the highest correlation with *loan status*. Based on their low values we conclude that all the features have little if any linear correlation with our target feature *loan status* and therefore none of the features have to be removed based on this observation.

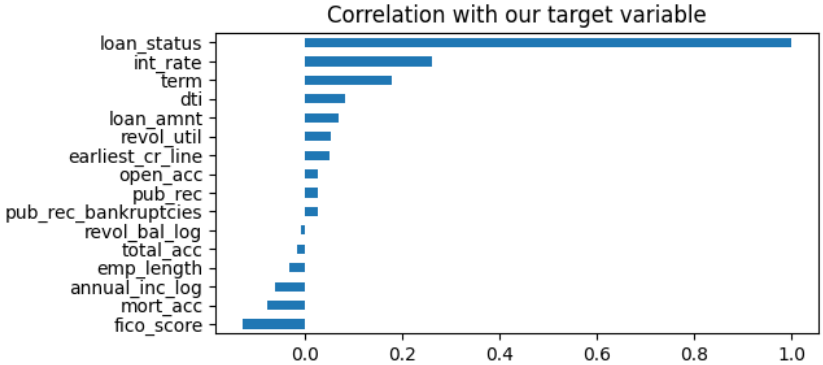


Figure 9: Correlation between the feature loan status and the other features

The features *installment, grade, fico range low, fico range high, subgrade, title, zip code,* and *address state* were removed from the dataset in this step.

4.4.4 DATA TRANSFORMATION

The next step is to transform the data from one format or structure into another format or structure that the model can interpret. We changed the features *term* from a categorical type into a numerical one. The feature *term* transformed 36 months and 60 months into the integers 36 and 60. For the feature *home ownership*, we added the categories 'None' and 'Any' to the category of 'Other' based on the variables that should be present in this feature according to the dictionary of Table A1.

We performed a log transformation on the features' *annual income* and *revolving balance* to reduces skewness in the data distributions. Through the log transformation, a normal distribution is created for these features.

Some classification algorithms, such as logistic regression, are unable to manage categorical features and these are therefore transformed into binarized data. This means that every feature is split into the different options and assigned the value '1' if it is present in this instance and '0' otherwise. We illustrated a transformation in Figure 10. The five categorical features of the dataset were transformed into binarized data: *verification_status* (3), *purpose* (14), *initial_list* (2), *application_type* (2), and *home_ownership* (4). The result is 3 + 14 + 2 + 2 + 4 = 25 numerical features that replaced the initial six categorical attributes.

Customer	application_type		application_type_INDIVIDUAL	application_type_JOINT
1	INDIVIDUAL	→	1	0
2	JOINT		0	1
3	INDIVIDUAL		1	0

Figure 10: Example of a transformation of the categorical feature application type into a numerical feature.

The features *issue date* and *earliest credit line* are transformed to a DateTime format to plot some time series analysis, an example is shown in Figure 11, to gain a better understanding of the data. From Figure 11 we see that the total amount of loans steadily increased over the years as a true emerging market. The feature *issue date* is removed after this since it can cause data leakage.

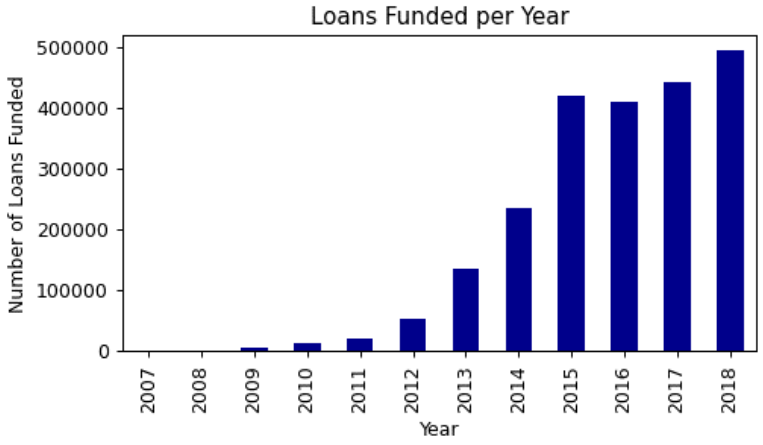


Figure 11: Total amount of loans over the years.

As a final step, we convert the target variable *loan status* into a binary numerical variable. The original nine categories with their count are displayed in Table 6. The new category "Fully Paid" consists of loans with the status "Fully Paid" and "In Grace Period" and is equal to "0". The new category "Charged Off" consists of all other statuses except that of "Current" and is equal to "1".

Loans with the status “Current” have not reached maturity and, therefore, do not contain information on the borrower’s creditworthiness.

The result and a brief description of our final dataset are presented in Table 5.

Number of instances	Features	% Fully Paid	% Charged Off	Imbalance Ratio
1.362.575	41	78.65	21.35	3.68

Table 5: Description of the final dataset

Status	Count
Fully Paid	1.063.380
Current	869.943
Charged Off	265.432
Late (31-120 days)	21.134
In Grace Period	8313
Late (16-30 days)	4276
Does not meet the credit policy. Status: Fully Paid	1528
Does not meet the credit policy. Status: Charged Off	543
Default	40
Total	2.234.589

Table 6: Distribution of the target variable Loan Status

4.4.5 Normalizing the data

After these feature engineering steps, we divided the dataset into a training and testing set based on a 75:25 ratio. After the splitting, we standardized the data by using the min-max normalization. This normalization ensures that all parameters use the same scale between a value of 0.0 and 1.0 using:

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{6}$$

where,

X_n the normalized value
X value of this feature for the data instance
$X_{min/max}$ the lowest/highest value of this feature in the dataset

4.4.6 Resampling approach

It has been shown that class imbalance obstructs classification (Lessmann, Basesens, Seow, & Lyn C. Thomas, 2015). The problem that occurs most is the classifier placing too much emphasis on the majority class while neglecting the minority class, resulting in a high accuracy score that shows an incorrect picture.

To deal with the imbalance problem in our dataset, we applied a resampling technique known as the synthetic minority over-sampling technique (SMOTE) since it has been proposed in the literature (Verbeke, Dejaeger, Martens, Hur, & Baessens, 2012) and already shown good results in (Namvar, Siami, Rabhi, & Naderpour, 2018). SMOTE uses a k-nearest neighbors’ algorithm to produce new instances based on the distance between the minority data and some randomly selected nearest neighbors, visualized in Figure 12.

We applied a special form of SMOTE, namely SMOTENC, which can take into account categorical variables. It puts a constraint on the categorical variables' *verification_status, purpose, initial_list, application_type,* and *home_ownership* to be either 0 or 1 and the sum of each feature must be equal to 1. This means that always exactly one of the values for each feature must be chosen. SMOTENC prevents the synthetic generated categorical variables to be a continuous value.

One of the drawbacks of SMOTE(NC) is that it does not take into consideration that neighboring examples can be from other classes. This can increase the overlapping of classes and can introduce additional noise. The use of SMOTENC results in a new training dataset and is described in Table 7.

Training dataset before SMOTE	Training dataset after SMOTE	Balance of positive and negative classes (%) before SMOTE	Balance of positive and negative classes (%) after SMOTE
Rows: 1.021.931 Features: 41	Rows: 1.607.540 Features: 41	0: 78.65% 1: 21.35%	0: 50.00% 1: 50.00%

Table 7: Description of the training dataset before and after applying SMOTE

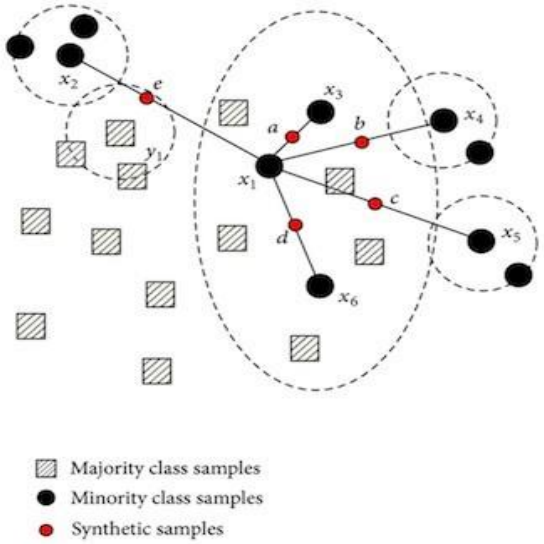


Figure 12: An example of the working of SMOTE. Based on a k-nearest neighbors algorithm SMOTE connects with other minority class samples (black dots) and synthetically generates new instances (red dots) on these connections. Source: (Hu & Li, 2013)

4.5 Performance Criteria

There are several performance metrics for binary classification. The confusion matrix is one of the most intuitive metrics used in statistical classification that allows visualization of the performance of an algorithm. The confusion matrix in itself is not a performance measure as such, but almost all the performance metrics are based on the confusion matrix and the numbers inside it. Therefore, we briefly explain the confusion matrix before we discuss the actual performance metrics.

4.5.1 Confusion Matrix

The confusion matrix is a table with two dimensions "True" and "Predicted" and sets of "classes" in both dimensions. Our True classifications are rows and Predicted ones are columns. In Table 8 the classes of the confusion matrix are explained. We may use a different convention for the axes

compared to other references, this may be because we kept the default axes from the Scikit-learn library used in Python.

	Predicted = Fully Paid (0)	Predicted = Charged Off (1)
True = Fully Paid (0)	**True Negatives (TN):** are the cases when the actual class of the data point was 0 and the predicted is also 0.	**False Positives (FP):** are the cases when the actual class of the data point was 0 and the predicted is 1.
True = Charged Off (1)	**False Negatives (FN):** are the cases when the actual class of the data point was 1 and the predicted is 0.	**True Positives (TP):** are the cases when the actual class of the data point was 1 and the predicted is also 1.

Table 8: Confusion Matrix with an explanation of all its classes

In our case of a binary classification problem with good (Fully Paid) and bad (Charged Off) class labels, the classification model is considered successful if the true negative and true positive values, of Table 8, are large and the values of false negatives and false positives are small (Gong & Kim, 2017). There are several performance metrics for binary classification that we will use both as a benchmark to compare our results with other studies as well as to make sure that our models are having a decent performance so we can test our XAI models for realistic models and outcomes. We measure if the classification model is successful using the following metrics: Accuracy (ACC), Area Under the Receiver Operating Characteristic (AUROC), the Matthews Correlation Coefficient (MCC), and the G-mean.

4.5.2 Accuracy

ACC is the most common performance metric to use in binary classification problems and is calculated by the equation:

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \quad (7)$$

The downside of accuracy is that, in imbalanced datasets, it tends to highlight the majority class and therefore neglect the minority class.

Another downside that accuracy does not consider, is that the false positives are more harmful than false negatives in the case of credit risk prediction (Caouette, Altman, Narayanan, & Nimmo, 2008). For the lender, it is more harmful to invest in a borrower who is incorrectly predicted as creditworthy compared to missing an investment opportunity through not investing in a borrower who incorrectly is predicted as not creditworthy since no money is lost in the latter case. Therefore, accuracy can be a misleading criterion that gives incorrect results. Since we applied SMOTENC to solve the imbalance problem, the accuracy metric will work better, and based on its popularity we will use it as one of the performance metrics.

4.5.3 Receiver operating characteristics

The Receiver Operating Characteristic (ROC) shows a curve by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The area under the curve (AUC) measures the area under the ROC curve and determines the AUROC score, which shows the ability of a classifier to distinguish between classes. The ROC curve based on the result of our ML classifiers can be seen in Figure 13.

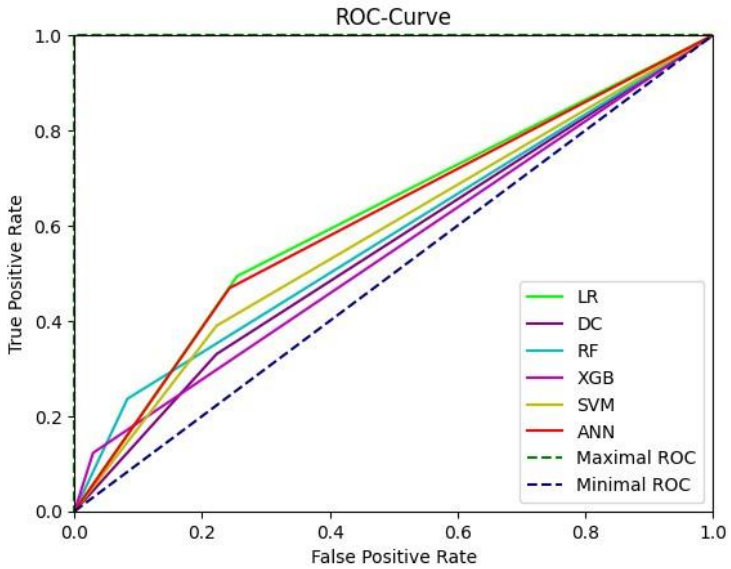


Figure 13: ROC curve based on the result of our ML classifiers. The minimal AUROC score is 0.5 and represents the performance of a random classifier and the maximum value of 1.0 would correspond to a perfect classifier

4.5.4 Matthews Correlation Coefficient

MCC is used instead of the more commonly used F-measure since it has been proven a more reliable statistical rate which produces a high score only if the prediction obtained good results in all the four confusion matrix categories (Chicco & Jurman, 2020). The equation of MCC is:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

where the denominator is a normalization constant so that result of the MCC is in the range -1 and 1, the only measure we used that is not between 0 and 1. A score of '-1' indicates the total dissimilarity between the prediction and the true value, '0' indicates a prediction that is no better than random and '1' indicates perfect similarity between the prediction and the true value. To the best of our knowledge, the MCC metric has not been previously used in any study considering LendingClub data.

4.5.5 G-Mean

The G-Mean measure makes use of both Sensitivity and Specificity. Sensitivity is calculated as the number of correct positive predictions divided by the total number of positives and Specificity is calculated as the number of correct negative predictions divided by the total number of negatives. Taking both measures into account by multiplying them against each other, the G-Mean is an effective performance metric for both balanced as imbalanced datasets (Namvar, Siami, Rabhi, & Naderpour, 2018). The equations for calculating the G-Mean:

$$G\text{-}Mean = \sqrt{Sensitivity * Specificity} \quad (9)$$

where,

$$Sensitivity = \frac{TP}{TP + FN} \quad (10)$$

$$Specificity = \frac{TN}{TN + FP} \quad (11)$$

4.6 Machine Learning Classification Results

The objective of our research is to determine the effectiveness of the post-hoc XAI techniques, by implementing XAI techniques on ML classifiers. With the use of feature engineering, we tried to create a realistic input dataset, so that we can test our XAI models on realistic data. Adjusting the hyperparameters for the best performance of the different models is outside the scope of our research and is left for future work. Therefore, we have not applied a Grid-search to find the optimal hyperparameters of a model which results in the most 'accurate' predictions and can be regarded as further research. We trained and tested an LR model, four ML classifiers, and one ANN binary classifier on the configured data described in section 4.2.

To compare the explainability of the different ML classifiers we trained a basic but transparent LR model. We trained a basic DT model that can hardly be classified as a transparent model since the decision tree consists of many nodes, each with its branches, that are not easily interpretable. Our decision tree is displayed in Figure 14. Next, are the ML classifiers that are not transparent by design but are widely used. These types of classifiers are the focus of this research and are represented by an RF, XGBooster, and SVM model. Next to that, we also wanted to expand our classification models even further into the direction of an ANN model, since these are becoming more popular in the financial sector (Barboza, Kimura, & Altman, 2017). For this, we build a Neural Network. After the training of the different models, we have evaluated them on the described criteria of section 4.5. The information about the hyper-parameter settings and the evaluation of the models on the model performance metrics are shown in Table 9.

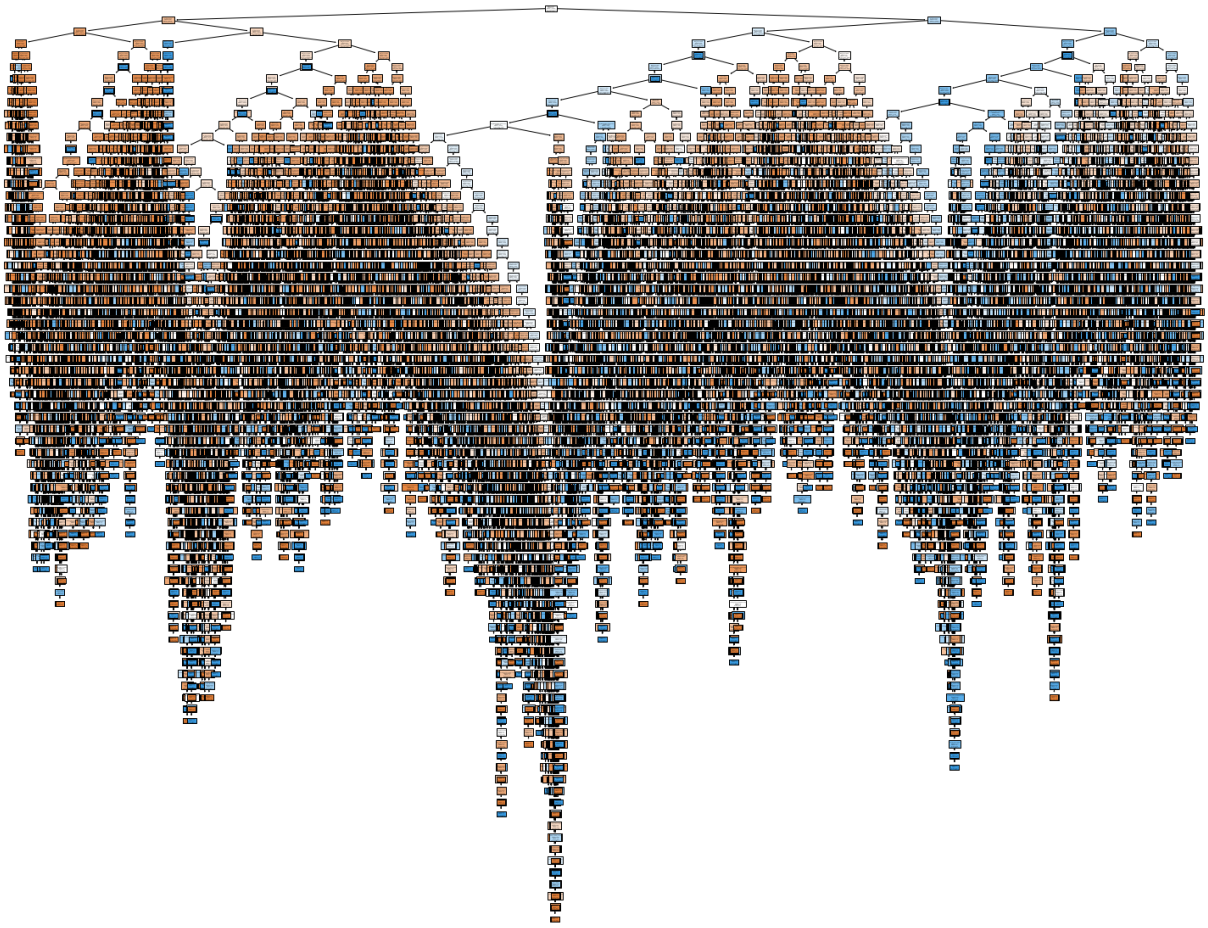


Figure 14: Our Decision Tree visualized. With all the different branches is becomes unreadably on the size of this page, however it gives the reader an idea of the models working.

Model	Hyper-Parameter	Performance					
		ACC	MCC	SE	SP	G-Mean	AUROC
LR	penalty = 'l2', random_state = 42, solver = 'lbfgs'	0.69	0.21	0.49	0.34	0.41	0.62
DT	criterion = 'entropy', max_depth = 'None', random_state = 42	0.68	0.10	0.33	0.29	0.31	0.55
RF	n_estimators = 100, random_state = 42	0.77	0.20	0.24	0.44	0.32	0.57
XGBooster	booster='gbtree', n_estimators = 100, max_depth = 3, random_state = 42	0.79	0.18	0.12	0.53	0.25	0.55
SVM	kernel = 'rbf', probability=True, class_weight='balanced', random_state = 42	0.69	0.16	0.39	0.32	0.35	0.58
ANN	n_hidden = 2, dropout = 0.2, neurons = [78, 39,19], activations = RelU, loss = binary_crossentropy, Optimizer = adam	0.70	0.20	0.47	0.34	0.40	0.62

Table 9: Models, parameters settings, and performance-based on different metrics

From Table 9 we can conclude that none of the classifiers performs significantly better. Their prediction performance is subpar to be applied effectively. The results of the research from (Wang, Zhang, Lu, & Yu, 2020) are showing the same ranking in the performance of some of the ML classifiers that we have used. However, the performance of our classifiers is poor compared to that of them.

4.7 Comparison with other LendingClub studies

As the final step, we compared the obtained results to the results in the work of (Namvar, Siami, Rabhi, & Naderpour, 2018) (Song, et al., 2020) (Moscato, Picariello, & Sperlí, 2021) and (Li, Cao, Li, Zhao, & Sun, 2020). We have chosen these studies as a comparison since they are currently the best ones in the field that also make use of either SMOTE or another oversampling technique.

It is difficult to compare the actual values of these studies with our study since they all have taken different data pre-processing steps and used other hyper-parameters settings. However, we can conclude that our values are in most cases in the same range as the values of these studies, except for (Li, Cao, Li, Zhao, & Sun, 2020) and therefore we can classify our results as representative. Next to that, their results give a good idea of the relative performance of the different ML classifiers. From Table 10, it is clear that RF has a high performance in most studies. The LR in these studies is performing similarly to the performance of the LR in our study. The study of (Li, Cao, Li, Zhao, & Sun, 2020) also confirms that the XGBooster classifier has a high performance on the LendingClub data.

The downside of these studies is that none of them used ANN classifiers to predict the creditworthiness of the borrowers in the LendingClub dataset. We found some studies that performed a classification study with a neural network, however, these used different and more complex variants of the neural network and are therefore not comparable with our ANN classifier.

Namvar (2018)	ACC	SE	SP	G-Mean	AUC
LR-SMOTE	0.65	0.64	0.64	0.64	0.70
RF-SMOTE	0.68	0.73	0.49	0.59	0.66
Song (2020)					
LR-Oversampling	0.56	0.56	0.56	0.56	0.56
DT-Oversampling	0.65	0.64	0.64	0.64	0.70
RF-Oversampling	0.68	0.73	0.49	0.59	0.66
Moscato (2021)					
LR-SMOTE	0.66	0.66	0.64	0.65	0.71
RF-SMOTE	0.77	0.98	0.1	0.31	0.71
Li (2020)					
LR-Oversampling	0.86	-	-	-	0.85
DT-Oversampling	0.87	-	-	-	0.80
RF-Oversampling	0.91	-	-	-	0.89
XGB-Oversampling	0.92	-	-	-	0.94

Table 10: Results of other studies on the LendingClub dataset

4.8 Conclusion on Data & Modeling

In this chapter, we focused on answering the research question "*Which data and what ML classifiers do we need as a basis to experiment with the evaluation framework?".* Based on section 4.2, we selected a few ML classifiers mentioned in the literature and selected the dataset of LendingClub to test these ML classifiers on. In line with Chapter 2, we first performed some feature engineering steps to configure the dataset and after this performed experiments with the ML classifiers. We compared those results with the results of other studies on the LendingClub dataset to verify the outcomes as trustworthy enough. In the next chapter, we will discuss the different XAI techniques we use to explain how our ML classifiers came to their predictions. After which we assess the XAI techniques based on our proposed ARRGUS framework.

5. EXPERIMENTAL RESULTS

5.1 EXPLAINABLE MODEL ALGORITHMS

In this section, we explain the most promising state-of-the-art post-hoc XAI techniques that are used in the literature or other practices and are available to us. In our view, it is important to test different post-hoc explainability techniques to see if the results are giving full transparency of decisions for our credit risk prediction problem. We do not get into the technicalities of them as it is beyond the scope of our research. Improvement of the technicalities of these algorithms can be regarded as a different research direction.

5.1.1 LIME

Locally Interpretable Model Agnostic Explanations (LIME) can be classified as a post-hoc model-agnostic explanation technique that aims to approximate any black-box machine learning model with a local, interpretable model to explain each prediction (Ribeiro, Singh, & Guestrin, 2016). LIME presents an explanation that is locally faithful, which means that it can give explanations for every specific observation it has. This can come in handy in the case the original model may be too complex to explain globally. LIME uses an approximation to fit a local model using sample data points that are like the instance being explained. This local model describes the local behavior of the model using a linearly weighted combination of the input features to provide explanations. The working of LIME is explained in Figure 15. Linear functions can capture the relative importance of features in an easy-to-understand manner. The drawback of linear explanations is that it is not clear whether they apply to an unseen instance since they are local, and it is unclear on which region its explanation applies. The human user may therefore think that the model is explaining an unforeseen instance when this may not be the case. Mathematically, the explanation produced by LIME can be expressed as follows:

$$explenation\ (x) = \ arg\min_{g \in G} L(f, g, \pi_x) + \ \Omega(g) \qquad (12)$$

Where the explanation model for instance x, or customer x in our case, is the model g (e.g., a linear regression model) that minimizes loss L (e.g., a mean squared error). Loss L measures how close the explanation of model g is to the prediction of the original model f, the ML classifier in our case, while the model complexity $\Omega(g)$ is kept low by preferring fewer features. G is the family of possible explanations, which can for example be all possible linear regression models. The proximity measure π_x defines how large the size of the neighborhood around instance x is that we consider for the explanation. For further details on LIME, we refer the reader to the research of (Ribeiro, Singh, & Guestrin, 2016).

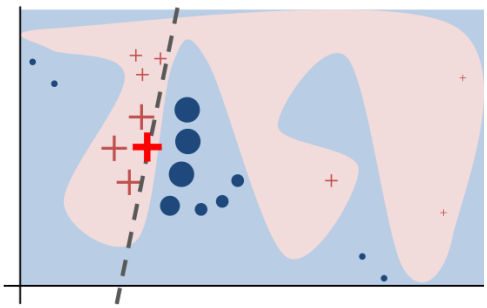


Figure 15: Example to present intuition for LIME, in which the dashed line is the learned linear regression explanation model that locally explains the bold red cross based on the instances that are captured by the learned linear regression explanation model. Source: (Ribeiro, Singh, & Guestrin, 2016).

5.1.2 ANCHORS

Anchors can be classified as a post-hoc model-agnostic explanation technique based on if-then rules. An anchor explanation is a rule that sufficiently "anchors" the prediction locally. In this way, a change to the rest of the feature values of the instance does not matter (Ribeiro, Singh, & Guestrin, 2018). In other words, for instances on which the anchor holds, the prediction is almost always the same. Anchors are intuitive, easy to understand, and have clear coverage. With coverage, the region on which the explanation applies is meant. Anchors only apply when all the conditions in the rule are met and can be based on a high precision depending on the preference of the user. The working of Anchors is explained in Figure 16. Formally defined, the explanation produced by an Anchor A can be expressed as follows:

$$E_{D_x(z|A)}\left[1_{f(x)=f(z)}\right] \geq \tau, A(x) = 1 \tag{13}$$

where,

x	represent the instance, in our case the customer, being explained.
A	is a set of rules, like $loan\ amount \geq 5000$, that returns '1' when all rules in the set correspond to x's feature values. The set of rules from A is also called the anchor.
f	denotes the classification model to be explained. In our case one of the ML classifiers.
D	denotes the perturbation space, a smaller representative area of the input.
$D_x(\cdot \mid A)$	indicates the area of neighbors' instances of x for which A holds.
So $D_x(z \mid A)$	indicates the area of neighbors' instances of x in which we predict sample z for which A holds
$1_{f(x)=f(z)}$	states the condition that the classification model is equal for data instance x and the random sample z
$0 \leq \tau \leq 1$	specifies a precision threshold. Only rules that achieve a local approximation of the prediction of the classification model of at least τ are considered as a valid result.

A is an anchor if the expected value of a random sample z, within the area D of neighbors' instances of x for which A holds, has the same prediction by the classification model as instance x $(1_{f(x)=f(z)})$ and has a precision greater than τ, given that all the rules are true $A(x) = 1$.

Since Anchors are developed by the same researchers as LIME, we also include the comparison between the working of LIME and Anchors in Figure 17 from the research of (Ribeiro, Singh, & Guestrin, 2018). From this figure, we see that LIME on the right gives a better local approximation of the black box model than LIME does on the left. Anchors gives a better local prediction in both cases and its boundaries are clearer compared to LIME. For further details on Anchors, we refer the reader to the research of (Ribeiro, Singh, & Guestrin, 2018).

Figure 16: Example of Anchors in which the presence of the words "not bad" almost guarantees a prediction of positive sentiment and the words "not good" a negative sentiment. Source: (Ribeiro, Singh, & Guestrin, 2018).

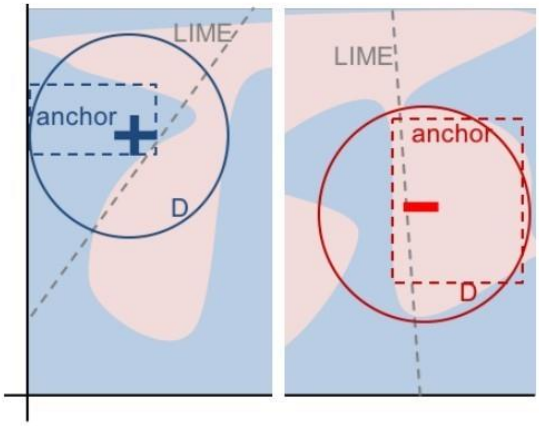


Figure 17: Comparison of two problems explained by Anchors and LIME. Perturbation space D is shown by a circle which both models use as their area/input to explain the model. LIME explains the prediction result by learning the line. Anchors uses a "local region" to learn how to explain the model. As we can see the "local region" refers to better construction of the generated data set for an explanation. Source: (Ribeiro, Singh, & Guestrin, 2018).

5.1.3 SHAP

Shapley Additive exPlanations (SHAP) is a method to explain the prediction of an instance x based on Shapley values from coalitional game theory, whose aim is to investigate how each feature affects the prediction (Lundberg & Lee, 2017). The research of (Lundberg & Lee, 2017) states that SHAP is a game-theoretic approach to explain the output $f(x)$ of any machine learning model. The feature values of a data instance x act as players in a coalition. Shapley values tell us how to fairly distribute the "payout", in our case the prediction among the features. In our case of tabular data, a player is the equivalent of an individual feature value. One of the innovations that SHAP offers is that it connects LIME and Shapley Values by representing the latter as an additive feature attribution method, a linear model. Mathematically, the explanation produced by SHAP can be expressed as follows:

$$g(\mathrm{z}') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j' \tag{14}$$

where g, again, is the explanation model, $z' \in \{0,1\}^M$ is the simplified feature, M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the feature attribution for a feature j, the Shapley Values. In the simplified features, an entry of 1 means that the matching feature value is "present" and 0 that it is "absent". To compute Shapley Values, SHAP simulates that only some of the feature values are "present" and some are "absent". This is equivalent to playing or not playing in the coalition from a game theory perspective. SHAP represents the coalitions a linear model to compute the ϕ's. Referring back to our instance x, the simplified features x' is a vector of all features that are "present". This simplifies the formula to:

$$g(\mathrm{x}') = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{15}$$

Several variants of SHAP are available and since this research focuses on post-hoc model-agnostic XAI techniques, we make use of KernelSHAP. KernelSHAP is a model-agnostic method to approximate SHAP values using ideas from LIME and Shapley values and can therefore be used on any ML classifier. KernelSHAP estimates for an instance x the contributions of each feature value to the prediction. Simplified, KernelSHAP creates sample coalitions z', get predictions for each z', compute the weight for each z' with the SHAP kernel and fit a weighted linear model on this. The coefficients from the linear model are the Shapley Values ϕ'. This is

translated into a base value $E[f(z)]$ that is the average of all predictions. SHAP values can be very complicated to compute since they are NP-hard in general. The working of SHAP is explained in Figure 18. An advantage of SHAP compared to LIME is that SHAP can obtain both local and global explanations compared to the local instances explained by LIME. For further details on SHAP, we refer the reader to the research of (Lundberg & Lee, 2017).

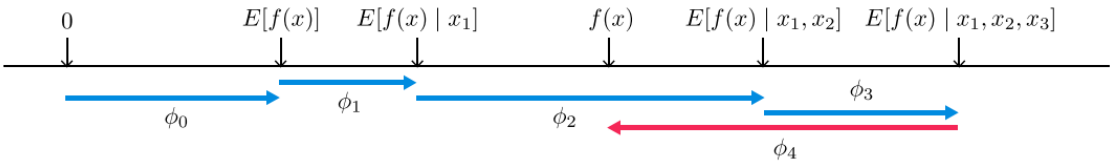


Figure 18: SHAP values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value E[f(x)] that would be predicted if we didn't know any features, to the current output f(x). Source: (Lundberg & Lee, 2017).

5.1.4 The customer instances that we explain

To be able to compare the performance of the different XAI techniques, we apply them to the same data instances. In this way, we can assess how the different XAI techniques are working and see where they differ from each other. We randomly choose two customer instances from the test data, one that belongs to the class "Fully Paid" and the other to the class "Charged Off". We presented these two customer instances with their values for each feature in Table 11. The transformed data represents the data that is used by the ML classifiers and the XAI technique and for reader convenience, we also provided the actual value of each feature. For the remainder of this chapter, we refer to the customer instance "Fully Paid" as Customer A and to the customer instance "Charged Off" as Customer B shown in Table 11.

We approach all three XAI models in the same way by first mentioning the additional steps and challenges we faced in the configuration of the explainer after which we show the results for all ML classifiers on Customer B. In the next section, we explain what the XAI technique is showing us after which we continue with an in-depth interpretation of the explanations for the LR (Transparent model), XGBooster (Tree model), and ANN (Neural Network model) to see if these explanations are in line with financial logic and/or can be justified from the input data. We end every section with an overall conclusion on the XAI technique and its performance.

5.2 Applying LIME to explain Local Instances

After performing the steps involved in getting the explanations from the LIME framework, we applied the procedure on the trained ML classifiers. The main challenge we faced in implementing the LIME framework was explaining our ANN classifier. The ANN classifier returns a 2D array for which LIME expects a 1D. We solved this problem by defining a new function that collapses the 2D array into 1D and updates it with both positive and negative probabilities so that the predict function understands the outcomes. Another challenge we faced was generating probabilities from the SVM output since an SVM training algorithm is a non-probabilistic binary linear classifier. We solved this problem by setting the hyperparameter "probability" to "True". This setting makes it possible to use the "predict_proba" function, which can obtain prediction probabilities.

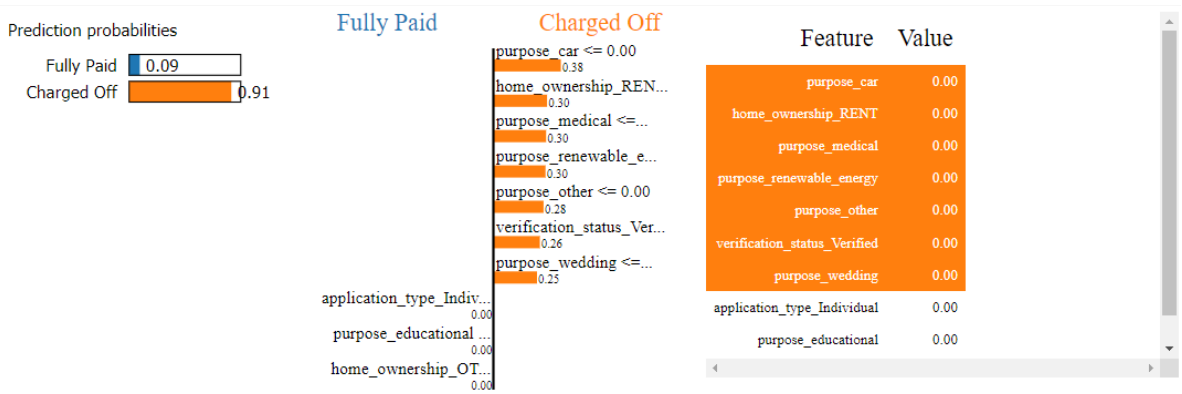
Customer A classified as "Fully Paid"			Customer B is classified as "Charged Off"		
Feature	**Transformed value**	**Actual value**	**Feature**	**Transformed value**	**Actual value**
loan_amnt	0,24	10000,00	loan_amnt	0,37	15000,00
term	0,00	36,00	term	1,00	60,00
int_rate	0,03	6,03	int_rate	0,53	19,03
emp_length	1,00	11,00	emp_length	0,00	0,00
dti	0,00	0,88	dti	0,11	111,52
earliest_cr_line	0,69	1990,00	earliest_cr_line	0,96	2012,00
open_acc	0,09	9,00	open_acc	0,13	13,00
pub_rec	0,00	0,00	pub_rec	0,00	0,00
revol_util	0,06	20,70	revol_util	0,11	39,50
total_acc	0,10	18,00	total_acc	0,07	14,00
mort_acc	0,02	1,00	mort_acc	0,00	0,00
pub_rec_bankruptcies	0,00	0,00	pub_rec_bankruptcies	0,00	0,00
fico_score	0,51	747,00	fico_score	0,22	687,00
annual_inc_log	0,63	4,88	annual_inc_log	0,48	4,00
revol_bal_log	0,58	3,72	revol_bal_log	0,58	3,73
verification_status_Not_Verified*	1,00	1,00	verification_status_Not_Verified*	1,00	1,00
purpose_debt_consolidation*	1,00	1,00	purpose_debt_consolidation*	1,00	1,00
initial_list_status_f*	1,00	1,00	initial_list_status_w*	1,00	1,00
application_type_Individual*	1,00	1,00	application_type_Joint*	1,00	1,00
home_ownership_MORTGAGE*	1,00	1,00	home_ownership_MORTGAGE*	1,00	1,00

*Table 11: Two customer instances, Customer A and B, from different classes that we use in our research to see how the different XAI techniques are working. The features are presented with their transformed values, the values the ML/XAI models use, and their actual values for reader convenience. *We only listed the categorical features that are present in the user instances.*

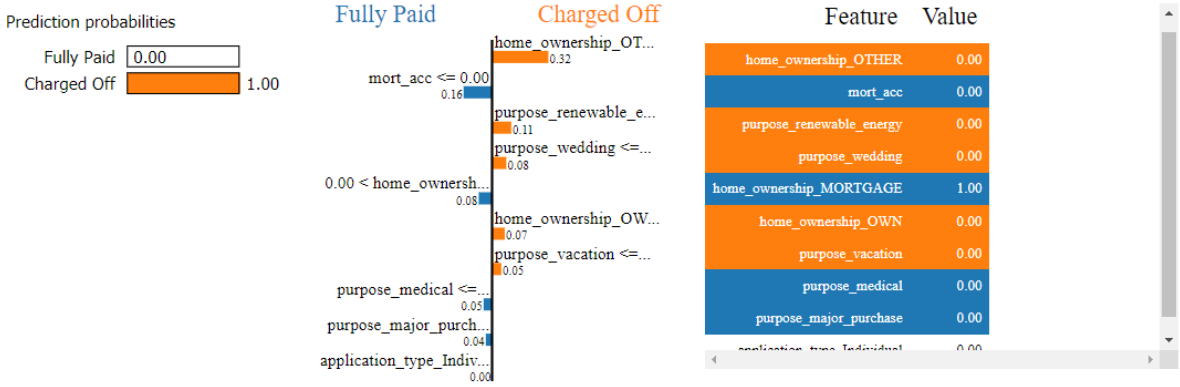
5.2.1 Explanation of LIME

To illustrate the working of LIME, we show the explanations given by the LIME technique in Figure 19 (a-f). The figure shows on the left side the model's confidence about its prediction. On the right side, the top ten features that contributed to the model's decision are presented. Lastly, in the middle, these same features are shown along with the contributions that they make in forcing the prediction made by the model towards that class.

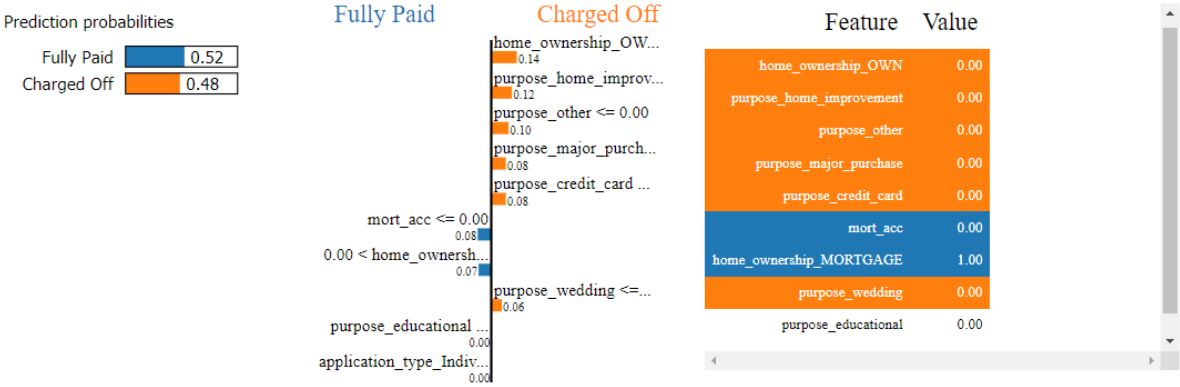
We interpret the LIME explanation of the LR, XGBooster and Neural Network models in section 5.2.2 and based on this explanation the other explanations by LIME can also be better understood.



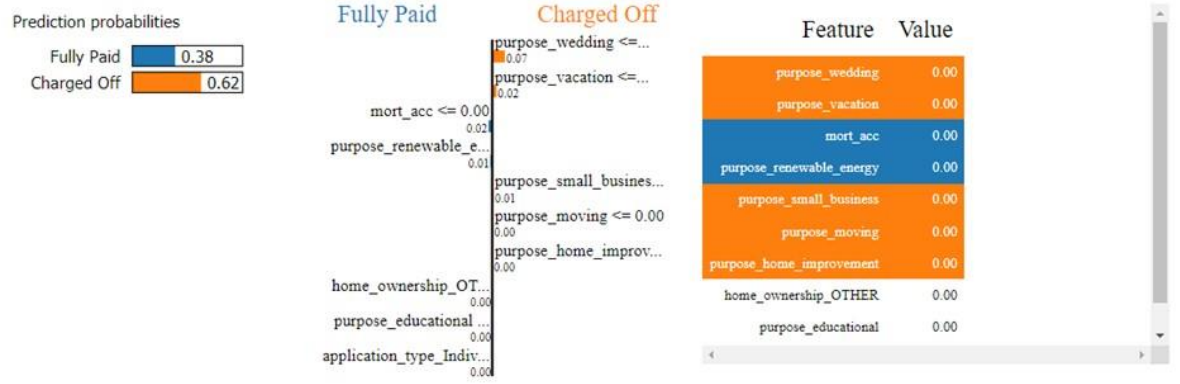
(a): LR explanation for "Charged Off" by LIME



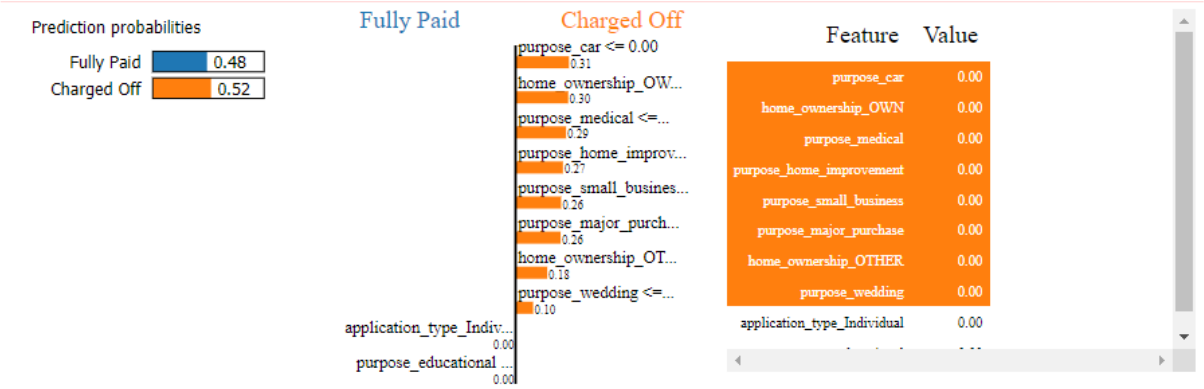
(b): DT explanation for "Charged Off" by LIME



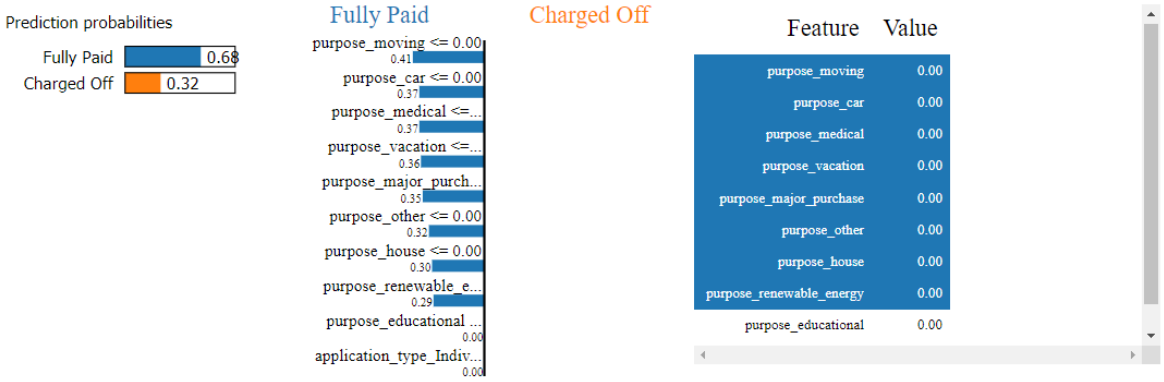
(c): RF explanation for "Fully Paid" by LIME



(d): XGBooster explanation for "Charged Off" by LIME



(e): SVM explanation for "Fully Paid" by LIME



(f): ANN explanation for "Fully Paid" by LIME

Figure 19: Results from the LIME explainer for each ML classifier on Customer B. We observe incorrect predictions by the RF and ANN model.

5.2.2 Interpretation of LIME

In this section, we discuss the local explanations given by the LIME explainer for the outcomes of the Logistic Regression, XGBooster, and Neural Network model on both customers A and B, described in detail in Table 11. We expand further on the interpretations of these explanations to check if they are in line with financial logic and compare them with the provided input data to get a better understanding of the functioning of LIME.

Table 12 present the total count of loans for each categorical feature and which percentage of these loans are signed off from the training data. We will refer to this table often since it provides us information on what data the models are trained for and help us interpret the explanations.

Feature: Purpose	**Count**	**% of loans Charged Off**
Debt Consolidation	1.033.427	55,53%
Credit Card	290.133	36,97%
Educational	273	34,80%
Small Business	11.944	32,34%
Other	66.860	31,23%
Moving	7.378	25,83%

Feature: Home Ownership	**Count**	**% of loans Charged Off**
Rent	689864	55,56%
Mortgage	755831	45,65%
Own	124749	30,96%
Other	408	27,45%
Feature: Initial List Status		
Fractional	639412	48,20%

Renewable Energy	683	24,60%
Medical	11.907	24,05%
Home Improvement	70.416	23,83%
House	5.648	23,78%
Major Purchase	22.827	21,92%
Vacation	6.904	20,22%
Car	11.131	16,44%
Wedding	1.742	13,49%
Feature: Application Type		
Individual	1583802	50,19%
Joint	23738	37,66%

Whole	968128	51,80%
Feature: Verification Status		
Verified	517500	54,06%
Source Verified	624572	50,68%
Not Verified	408973	36,91%
Feature: Term		
36 months	1111945	42.61%
60 months	495595	66,61%

Table 12: All categorical features with their possible values presented. For each feature, the amount of loans with this feature in the training set is shown. Also, the percentage of these loans that are "Charged Off" are shown.

XGBooster model

In Figure 20, we present an example of a loan contract that belongs to Customer A which the XGBooster model predicted correctly. We can interpret this explanation as follows:

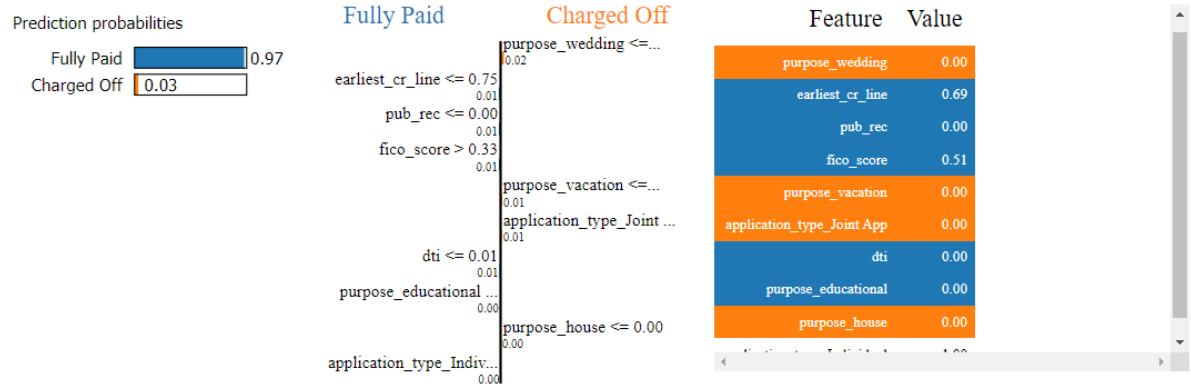


Figure 20: XGBooster explanation for "Fully Paid" for Customer A by LIME.

Customer A has been categorized as the class "Fully Paid", meaning that they are expected to fully repay the loan. The model is 97% sure about its prediction, as can be seen on the left. The fact that the purpose of the loan was not for "wedding", "vacation" or "house" forced the prediction toward the class "Charged Off". The same rule applies for the loan not being applied to as a type "Joint".

The earliest credit line is equal to 0,69 (1990). This is smaller than 0,75 (1995), the value used by the model for deciding. This difference forced the prediction towards the class "Fully Paid". The same applies for the features "public records" and "dti" having a value smaller than 0.00 and 0.01 respectively and therefore forcing the prediction towards the class "Fully Paid". The FICO score of this customer is equal to 0.51 (747), which is larger than 0.33 (712), and therefore forces the prediction towards the class "Fully Paid".

We observe that the weights for the top ten features are small, almost unimportant. This could mean that: i) The LIME technique does not present the weights correctly through modeling and/or input errors from our side. 2) The model treats the features as evenly important, which is unlikely.

In Figure 19 (d), we present an example of a loan contract that belongs to the class “Charged Off”, Customer B, which the XGBooster model predicted correctly. We can interpret this explanation as follows:

Customer B has been categorized as the class "Charged Off", meaning that they are expected to fail to fully repay the loan. The model is 62% sure about its prediction, as can be seen on the left. This low certainty of the model can be caused by a very poor predictive performance of the XGBooster.

The fact that the purpose of the loan was not for "wedding", "vacation", “small business”, “moving” or "home improvement" forced the prediction toward the class “Charged Off”. The fact that the purpose of the loan was not for "renewable energy" or "educational" forced the prediction toward the class “Fully Paid”. The same rule applies for the number of mortgage accounts for the customer of the loan not being larger than zero. We observe that, as opposed to the XGBooster model prediction for the “Fully Paid” instance, the last three features are not contributing towards a prediction of one of the two classes.

Based on Table 12 it makes sense that the fact that the purpose of the loan was not for "wedding", "vacation" or “House” forced the prediction toward the class “Charged Off” since these purposes have a relatively low value of loans that end up to be “Charged Off” compared to the other categories of purposes. Meaning that the purposes that have a higher percentage of being “Charged Off” remain and therefore this loan must be one of those riskier purposes.

The same logic applies to the application type. Application type “Joint” has a lower percentage of loans that are Charged Off compared to that of application type “Individual”. Application type “Joint” having the value zero, means that the loan application type is “Individual” and therefore forces the prediction towards the class “Charged Off”.

However, based on Table 12 all the described cases above have a lower amount of occurrence in the training set and therefore can have an unjustified impact. Significant fewer cases can mean that only a biased set of instances can be present in the dataset, while if all the features had the same number of occurrences the percentage of loans Charged Off could be higher for the described cases.

In Figure 21 the distribution of loans based on the earliest credit line of the customer is presented while on the second y-axis the percentage of “Charged Off” loans is visualized. In this figure, the impact of the synesthetic-generated instances by the SMOTE sampling technique can be seen. SMOTE generated the most “Charged Off” instances in the area 1995 – 2010 and this increases the percentage of loans that are classified as “Charged Off” for the same period in the original training set. The earliest credit line of the “Fully Paid” customer is equal to 1990 (0,69) which is smaller than 1995 (0,75), the value used by the model for deciding. The fact that the loans with the earliest credit line smaller than 1995 have a lower percentage of loans that are “Charged off” forced the prediction toward the class “Fully Paid”.

In case the SMOTE sampling technique was not used, the rule of the model to decide for which value it forces the prediction towards the class “Fully Paid” could have been the opposite since the original test set, presented by the red line, has a decrease in the percentage of loans that are Charged Off after 1995.

In Figure 22 the distribution of FICO scores of the customers is presented while on the second y-axis the percentage of “Charged Off” loans is visualized. We see fewer high FICO scores, the more creditworthy borrowers. These higher FICO scores also show a lower percentage of loans that are “Charged Off”, which is in line with financial logic. The FICO score of the “Fully Paid” customer is equal to 747 (0.51), which is larger than 712 (0.33), and therefore forces the prediction towards the class “Fully Paid”. From Figure 22 we see that the percentage of loans that are “Charged Off” become less than 30% from around a FICO score of 712 and it seems like an acceptable value used by the model for deciding. However, a more risk-averse person and/or model can of course choose a higher FICO score as a threshold and vice versa.

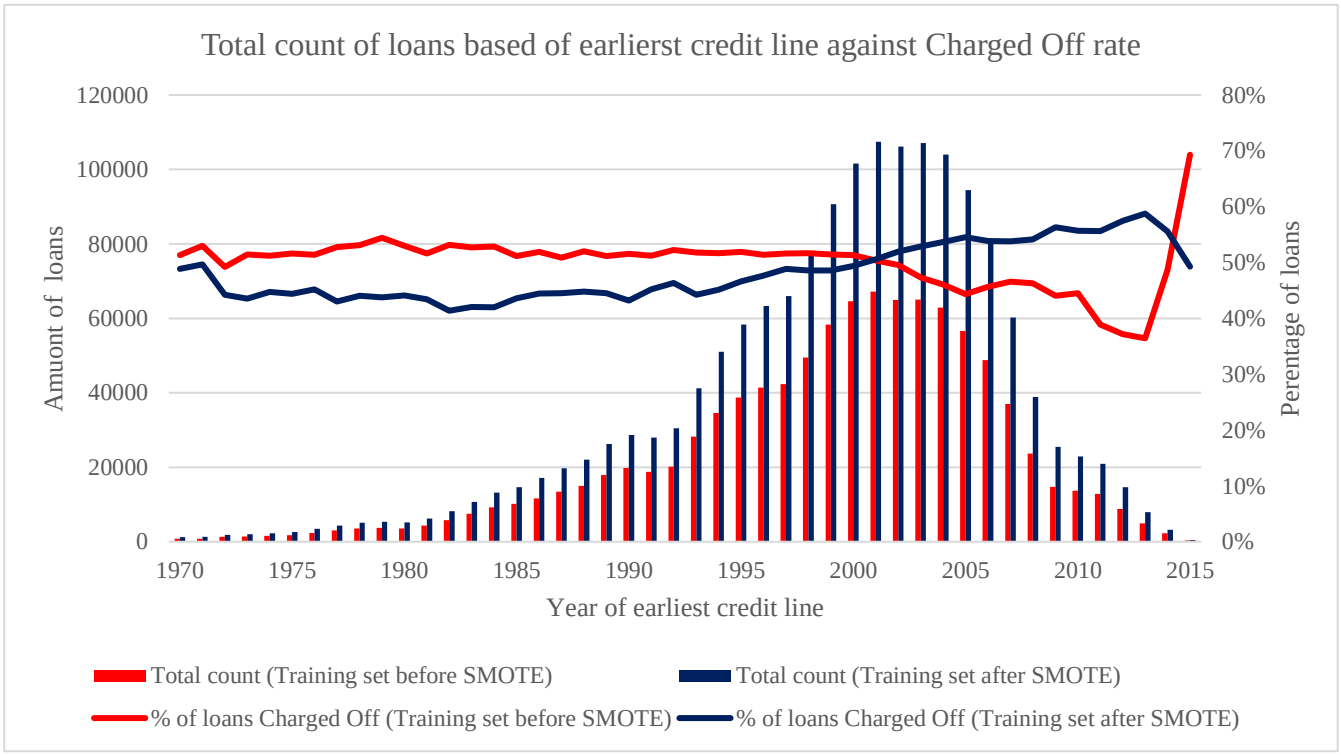


Figure 21: Total counts of loans based on the earliest credit reported by the customer against the percentage of these loans that are Charged Off. In this figure, we also show the effect of SMOTE on the training set. Data is from our training set.

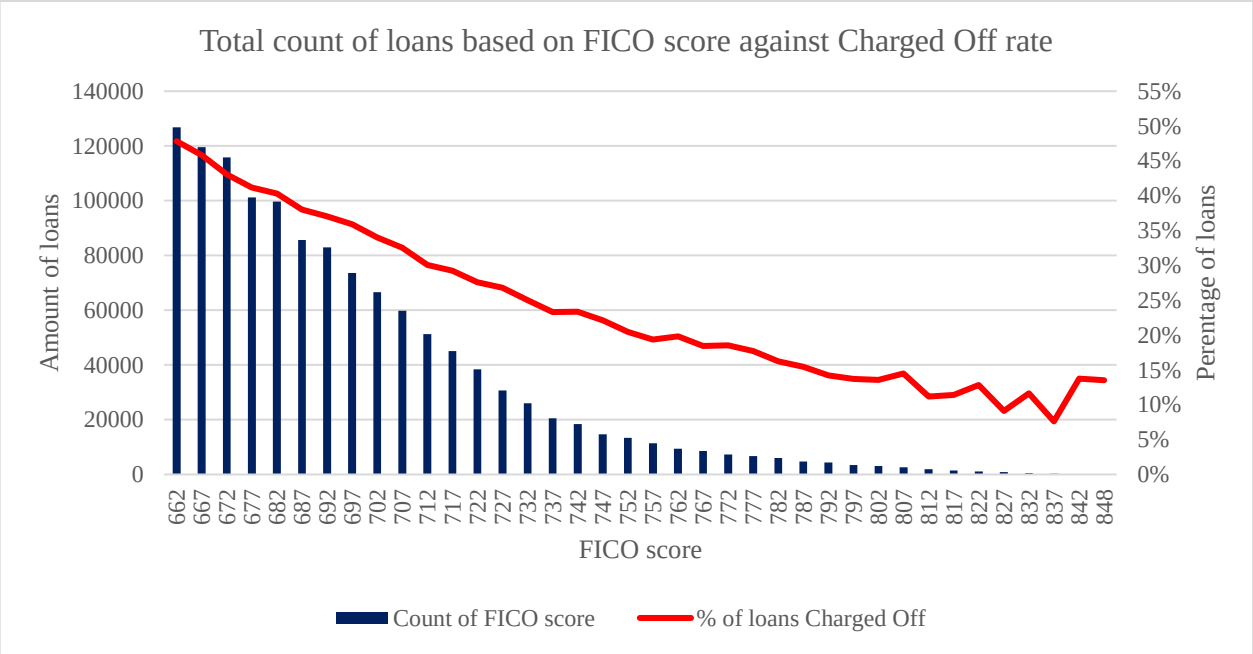


Figure 22: Total counts of loans based on the FICO score of the customer against the percentage of these loans that are Charged Off. Data is from our training set.

Logistic Regression model

In Figure 23 we present an example of a loan contract that belongs to Customer A which the LR model predicted correctly. Customer A has been categorized as the class "Fully Paid", meaning that they are expected to fully repay the loan. The model is 97% sure about its prediction, as can be seen on the left.

Again, based on Table 12, it makes sense that since the purpose of the loan was not for "vacation" or “car”, forcing the prediction toward the class “Charged Off” since these purposes have a relatively low value of loans that end up to be “Charged Off” compared to the other categories of purposes. This means that the purpose of this loan is more likely to have a higher likelihood of becoming “Charged Off”. However, this logic makes no sense for the three other purposes "other", " medical " and “moving” since these have a higher percentage of loans that are “Charged Off”.

We observe that even though the total count of the top ten features is in favor of pushing the prediction towards the class "Charged Off", the total sum of all the features combined is in favor of the class "Fully Paid". We also observe the similarity in features of the LR model used to predict the “Fully Paid” customer in Figure 23 and the “Charged Off” customer in Figure 19 (a).

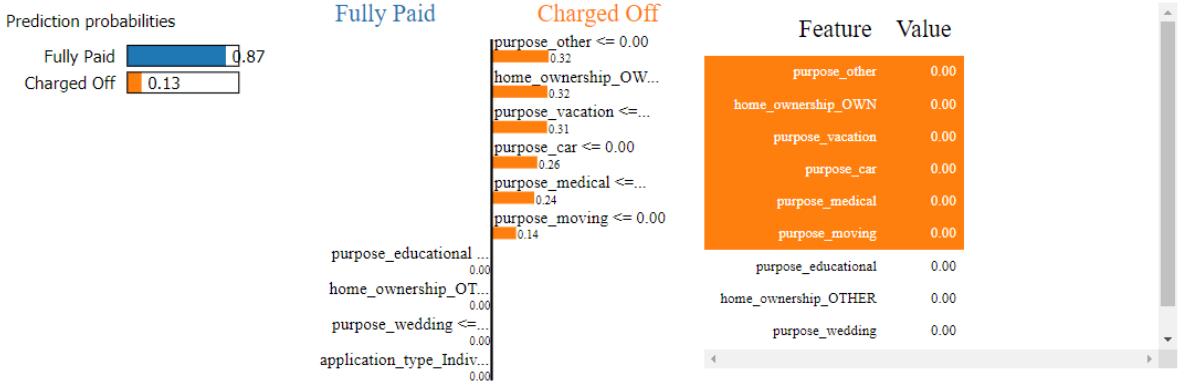


Figure 23: LR explanation for “Fully Paid” for Customer A by LIME.

Artificial Neural Network model

In Figure 24 we present an example of a loan contract that belongs to the “Fully Paid” customer whom the ANN model predicted incorrectly. Customer A has been categorized as the class "Charged Off", meaning that they are expected to fail to fully repay the loan. The model is 91% sure about its prediction, as can be seen on the left. However, the actual customer actual belongs to the class “Fully Paid”, and the Artificial Neural Network model is therefore wrong.

The fact that the purpose of the loan was not for "wedding", "medical", “car”, “moving” or "vacation" forced the prediction toward the class “Fully Paid” and this is opposite of what we observed from the other explainers that correctly predicted Customer A. From Table 12 we can see that the mentioned purposes are having a relatively low percentage of loans that are “Charged Off”. Since these purposes are not the reason for the loan of Customer A, it increases the likelihood that the loan has a purpose with a higher chance of being in the class “Charged Off” and therefore it should force the prediction of the model towards the class “Charged Off”. However, as we can see from the LIME framework, the ANN model uses these rules to force the prediction towards the class “Fully Paid”.

The feature initial list status not being “Whole” forced the prediction towards the class “Charged Off”. Based on Table 12, this seems to be another contradicting rule by the ANN model since the feature initial list status “Whole” has a higher percentage of loans that are “Charged Off” compared to the other option “Fractional”.

We observe that a small number of features is contributing a lot to the specific outcome of the model. This can explain the relatively poor prediction performance of the Artificial Neural Network. However, due to the described challenges in implementing LIME on the Artificial Neural Network model, it can also be that the LIME technique is not working correctly in our case.

Since an ANN model is a classic opaque Black-Box model we are not able to get a better insight into the inner working of the model than the LIME framework is currently providing us. The lacking amount of information and decision rules gives us only a tip of the iceberg of how the ANN model has made its predictions. However, this limited information is not enough to understand the model but suggests going back to the drawing board to improve the model before implementing it.

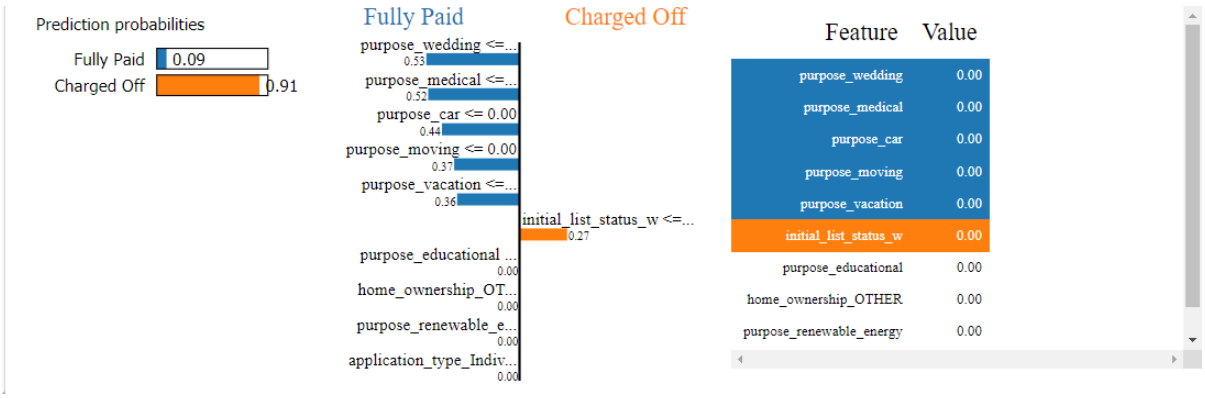


Figure 24: ANN explanation for “Charged Off” for Customer A by LIME.

5.2.3 Conclusion on LIME

Going in-depth to find the actual information behind the explanation and rules provided by the LIME framework resulted in a better understanding of how the different ML classifiers work. The explanations of the LIME framework were, except for the ANN model, in line with the logic, we would expect based on the provided input data. The reasoning of the models given by the model for making a prediction is also in line with financial logic. We would however expect that the continuous features make more of an impact than the categorical features. Customers with a purpose of the loan of "Vacation", "Car" or "Wedding" are marked as less risky compared to a "small business", a higher FICO score and people who own their own house are considered less risky and are more likely to have their loans approved. This was evident from the examples we discussed above. Similarly, high-risk customers can be identified by riskier purposes, lower FICO scores, and other ownerships of their houses. This information is helpful information for model developers, users that provide loans, and the end-users since it offers more transparency and increases trust in the predictions. The ANN model prediction shows some habits that are not in line with financial logic. Model developers therefore can observe these discrepancies to increase the model performance accuracy by having more knowledge on the rules and/or features that are causing these problems.

5.3 Applying Anchors to explain Local Instances

We followed the work of (Ribeiro, Singh, & Guestrin, 2018) and their guidelines on the GitHub page[2] for the additional configuration of the Anchors explainer. The main challenge we faced in implementing the Anchors technique was explaining our ANN classifier. Anchors expect the ML classifier to have the same dimensions for all the input arrays, however, our ANN classifier returns an array that has different dimensions for different indexes of the input array. We were not able to solve this challenge and therefore the Anchors explanation of the ANN is missing. The use of Anchors on an ANN classifier is left for further research.

5.3.1 Explanation of Anchors

Figure 25 (a-e) shows the Anchors' explanations on the prediction of the ML classifiers. The prediction is shown, followed by the conditions the Anchor framework identified that will result in the same prediction outcome, the precision of the Anchor framework, and the coverage on the region where the explanation applies to. As mentioned in section 5.1.2, Anchors works with easy-to-understand IF-THEN rules. We set a minimum threshold for the precision being at least equal to or larger than 95% to guarantee, with a high probability, that the predictions on instances where the anchor holds will be the same as the original prediction at least 95% of the time. A higher precision, most of the time, comes at the expense of a lower coverage since a higher precision explanation will be applicable on a smaller region of instances.

The generated explanations given by the Anchors framework are presented in Figure 25 (a-e). We explain the Anchors' explanation on the LR, XGBooster and Support Vector Machine models in section 5.3.2 and based on this explanation the other explanations by Anchors can also be better understood.

```
Prediction:  Charged Off
Anchor: application_type_Joint App > 0.00
Precision: 0.98
Coverage: 0.01
```

(a): LR explanation for "Charged Off" by Anchor

```
Prediction:  Charged Off
Anchor: int_rate > 0.46 AND 0.00 < term <= 1.00 AND earliest_cr_line > 0.86 AND purpose_home_improvement <= 0.00 AND home_owner
ship_OWN <= 0.00 AND purpose_credit_card <= 0.00 AND mort_acc <= 0.02 AND purpose_other <= 0.00 AND dti > 0.01 AND total_acc <=
0.17 AND loan_amnt > 0.32 AND 0.15 < fico_score <= 0.22 AND annual_inc_log <= 0.61 AND pub_rec_bankruptcies <= 0.00 AND initial
_list_status_f <= 0.00 AND purpose_small_business <= 0.00 AND purpose_moving <= 0.00 AND purpose_car <= 0.00 AND 0.08 < open_ac
c <= 0.15 AND purpose_major_purchase <= 0.00 AND verification_status_Verified <= 1.00 AND verification_status_Source Verified <
= 0.00 AND purpose_debt_consolidation <= 1.00 AND emp_length <= 1.00 AND purpose_wedding <= 0.00 AND purpose_educational <= 0.0
0 AND home_ownership_OTHER <= 0.00 AND home_ownership_MORTGAGE <= 1.00 AND application_type_Individual <= 1.00 AND 0.00 < initi
al_list_status_w <= 1.00 AND home_ownership_RENT <= 1.00 AND verification_status_Not Verified <= 1.00 AND purpose_house <= 0.00
AND purpose_vacation <= 0.00 AND purpose_renewable_energy <= 0.00 AND purpose_medical <= 0.00
Precision: 0.96
Coverage: 0.00
```

(b): DT explanation for "Charged Off" by Anchor

```
Prediction:  Fully Paid
Anchor: 0.00 < verification_status_Not Verified <= 1.00 AND application_type_Joint App > 0.00 AND mort_acc <= 0.00 AND 0.00 < h
ome_ownership_MORTGAGE <= 1.00 AND pub_rec <= 0.00 AND fico_score > 0.15 AND home_ownership_RENT <= 0.00 AND revol_util <= 0.15
AND loan_amnt <= 0.49 AND pub_rec_bankruptcies <= 0.00 AND 0.11 < open_acc <= 0.15 AND verification_status_Verified <= 0.00 AND
purpose_vacation <= 0.00 AND purpose_educational <= 0.00 AND emp_length <= 1.00 AND application_type_Individual <= 1.00 AND ver
ification_status_Source Verified <= 0.00 AND initial_list_status_w <= 1.00 AND initial_list_status_f <= 1.00 AND term <= 1.00 A
ND purpose_renewable_energy <= 0.00 AND home_ownership_OTHER <= 0.00 AND 0.00 < purpose_debt_consolidation <= 1.00 AND purpose_
wedding <= 0.00 AND purpose_small_business <= 0.00 AND purpose_house <= 0.00 AND purpose_other <= 0.00 AND purpose_moving <= 0.
00 AND purpose_car <= 0.00 AND purpose_medical <= 0.00 AND purpose_major_purchase <= 0.00 AND purpose_home_improvement <= 0.00
AND purpose_credit_card <= 0.00 AND revol_bal_log <= 0.59
Precision: 0.86
Coverage: 0.00
```

(c): RF explanation for "Fully Paid" by Anchor

[2] GitHub - marcotcr/anchor: Code for "High-Precision Model-Agnostic Explanations" paper

```
Prediction:  Charged Off
Anchor: verification_status_Verified <= 0.00 AND int_rate > 0.46 AND verification_status_Source Verified <= 0.00 AND term > 0.0
0 AND dti > 0.03 AND fico_score <= 0.22
Precision: 0.96
Coverage: 0.00
```

(d): XGBooster explanation for "Charged Off" by Anchor

```
Prediction:  Charged Off
Anchor: term > 0.00 AND int_rate > 0.21 AND mort_acc <= 0.00
Precision: 0.96
Coverage: 0.08
```

(e): SVM explanation for "Charged Off" by Anchor

Figure 25: Results from the Anchor explainer for each ML classifier for Customer B. We observe an incorrect prediction by the RF model.

5.3.2 Interpretation of Anchors

In this section, we discuss the explanations given by the Anchors explainer for the outcomes of the Logistic Regression, XGBooster, and Support Vector Machine model on the customer instances "Fully Paid" and "Charged Off" from Table 11. We replaced the ANN model with that of the SVM model since we were not able to get the explainer working for the ANN model. The SVM represents another class of models than the Logistic Regression and XGBooster models and is, therefore, a worthy replacement to test how "model-agnostic" these XAI techniques are. We expand further on the explanation to check if they are in line with financial logic and/or justify the explanation based on the input information.

Logistic Regression

In Figure 26, we see the correct prediction of the LR model for the customer instance that is in the class "Fully Paid". As we can see from the figure, this prediction is based on the rule that both the feature verification status must be "Not Verified", and the interest should be equal to or below 0.32 (= 13.6%). From Table 12 we see that a "Not Verified" loan that has this type of verification has a lower percentage of the loans that are "Charged Off" with only 37% compared to the 50%+ of the other values the feature verification status can take. From Figure 27 we see that the percentage of "Charged Off" loans for an interest rate with the value of 13.6 is around 45%. A lower interest rate means that the borrower is more creditworthy and therefore it makes sense to apply a rule that selects loans with low interest rates to use as an indicator for a customer that is in class "Fully Paid". With a coverage of 0.17, we can tell that this anchor is applicable on quite a region of instances.

However, the same does not hold for the prediction of the LR model on the customer instance that is in the class "Charged Off" in Figure 25 (a). As a result, we only have a single if-then rule that results in this prediction of "Charged Off" by having an application type with the value "Joint". From Table 12 we see that this application type has a lower percentage of loans that are "Charged Off" and therefore it is not a logical choice to use as an indicator for a customer to be classified as "Charged Off". Due to the low value of coverage, this explanation of the prediction can be unique for this case and does not tell us more about more general rules the LR classification model uses to predict if a customer is classified as "Charged Off".

```
Prediction:  Fully Paid
Anchor: verification_status_Not Verified > 0.00 AND int_rate <= 0.32
Precision: 0.96
Coverage: 0.17
```

Figure 26: LR explanation for "Fully Paid" for Customer A by Anchors.

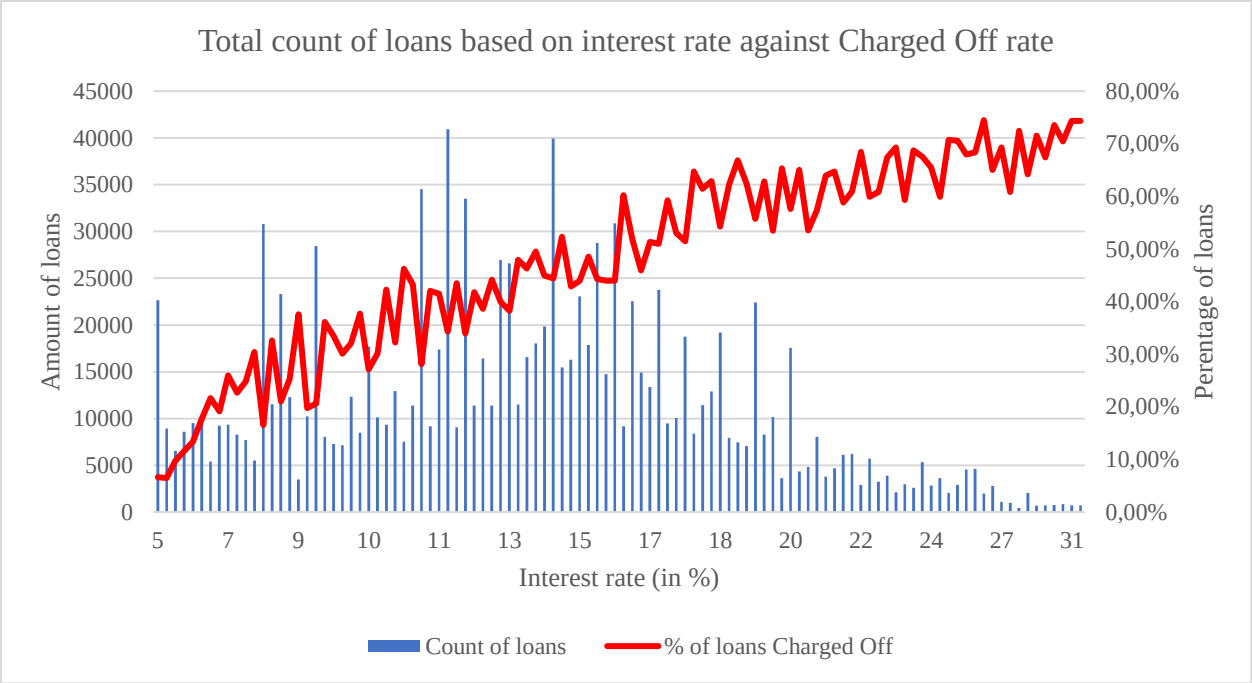


Figure 27: Total counts of loans based on the interest rate of the customer against the percentage of these loans that are Charged Off. Data is from our training set.

XGBooster model

In Figure 28 we see the correct prediction of the XGBooster model for the customer instance that is in the class "Fully Paid". As we can see from the figure, this prediction is based on more rules than we have seen in the earlier Anchor explanations. To classify a customer as "Fully Paid", it must have an interest rate equal to or smaller than 0.22 (= 11%) which is more risk-averse than we saw in the case of the LR model. Like the LR model, the verification status should be "Not Verified". Since it is a binary variable, the greater than zero values mean that it should be '1' and therefore the value "Not Verified". The home ownership not being "Other" seems to be contradicting. From Table 12 we see that home ownership "Other" has the lowest percentage of Charged Off predictions with 27,45%. The rule of employment length being smaller or equal to 1 is a redundant rule, since it applies to all cases. The same applies for the rules on initial list status, home ownership rent, and purpose debt consolidation.

The term being smaller than or equal to zero means that the term should be the shorter variant of 36 months. This makes sense since a loan with a shorter maturity is safer than one with longer maturity and can also be seen from the data in Table 12. The rule of the loan not having as purpose "Renewable Energy" is debatable. From Table 12 we see that this purpose is amongst the middle compared to the other purposes, so it would have made more sense to choose a rule that with a higher percentage of loans that are Charged Off. However, due to this loan already having the purpose with the highest percentage of loans that are Charged Off, namely debt consolidation with 55.53% and still being the class "Fully Paid", makes this explanation more complicated.

```
Prediction:  Fully Paid
Anchor: int_rate <= 0.22 AND 0.00 < verification_status_Not Verified <= 1.00 AND home_ownership_OTHER <= 0.00 AND emp_length <=
1.00 AND initial_list_status_f <= 1.00 AND home_ownership_RENT <= 1.00 AND purpose_debt_consolidation <= 1.00 AND term <= 0.00
AND purpose_renewable_energy <= 0.00 AND loan_amnt <= 0.49 AND dti <= 0.01
Precision: 0.86
Coverage: 0.03
```

Figure 28: XGBooster explanation for "Fully Paid" for Customer A by Anchors.

The loan amount being smaller or equal to 0.49 (= +/- 19.500) makes sense, since a smaller loan amount is less risky than a large one since it can be more easily repaid. This reasoning is confirmed from the data in Figure 29. The last line refers to a rule on the Debt-To-Income (DTI) ratio with a rule that this ratio must be smaller than or equal to 0.01 (= 10). The low value of 0.01 is caused by outliers with a maximum of 999 and these ratios are not representative for the DTI score. We, therefore, plotted the DTI ratios with at least 2000 records of loans, since SMOTE caused a lot of noise, and these are presented in Figure 30. In line with financial logic, we see a lower percentage of Charged Off loans for lower DTI ratios. A low DTI ratio indicates sufficient income relative to debt servicing and makes a borrower more attractive. This rule is therefore in line with what we expect for forcing the prediction towards the class of "Fully Paid". We observe that the Anchor explainer does not have enough precision compared to the minimum threshold we set, however, its reasoning is in line with financial logic.

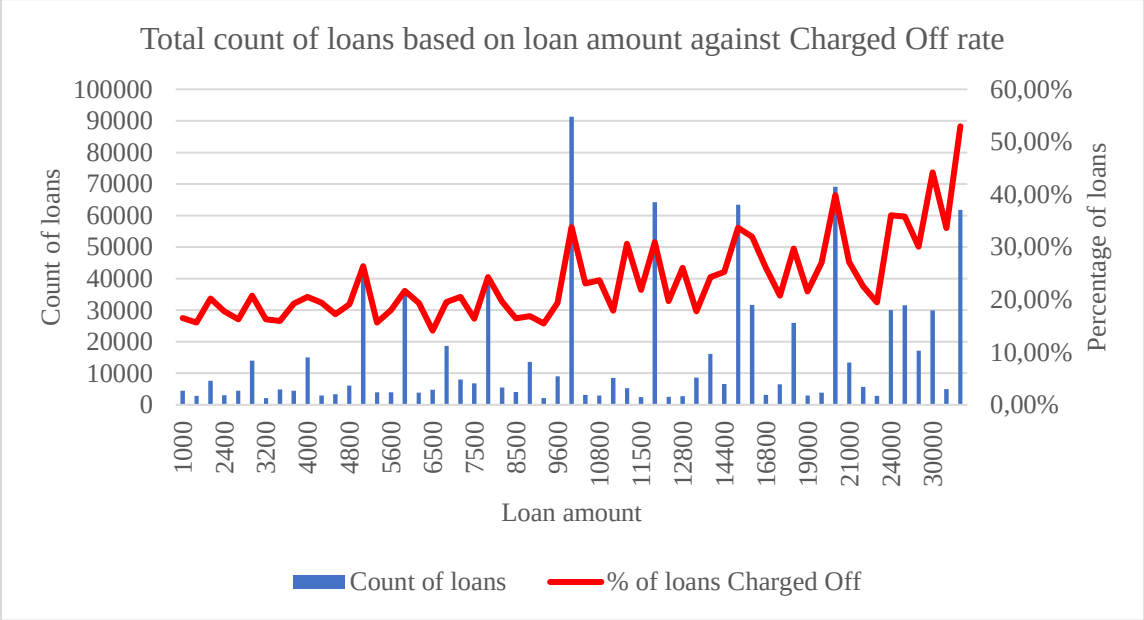


Figure 29: Total counts of loans based on the loan amount of the customer against the percentage of these loans that are Charged Off. Data is from our training set.

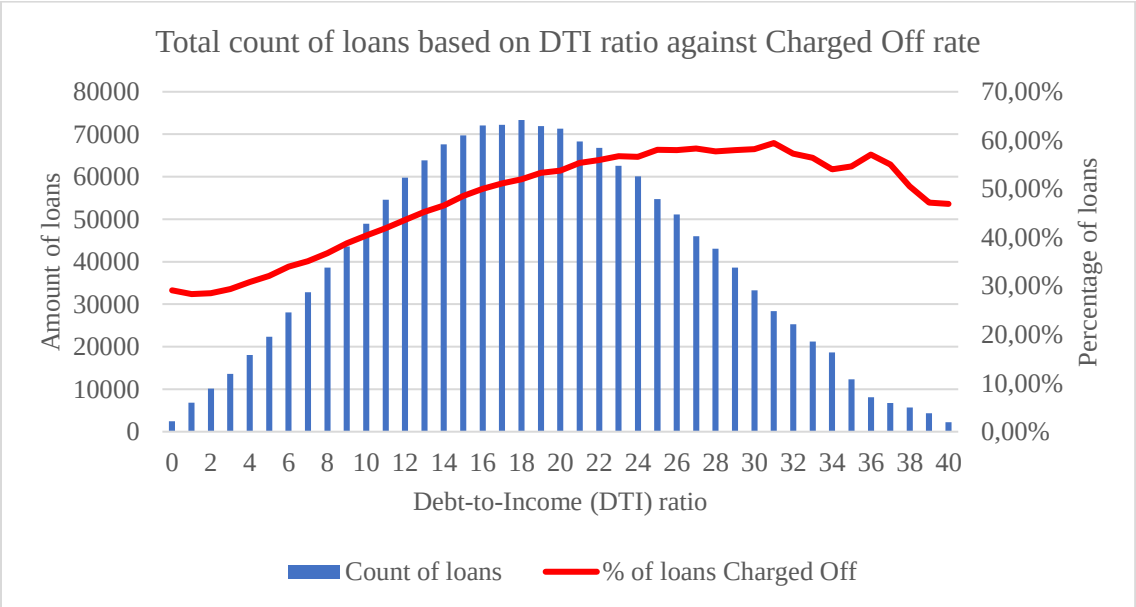
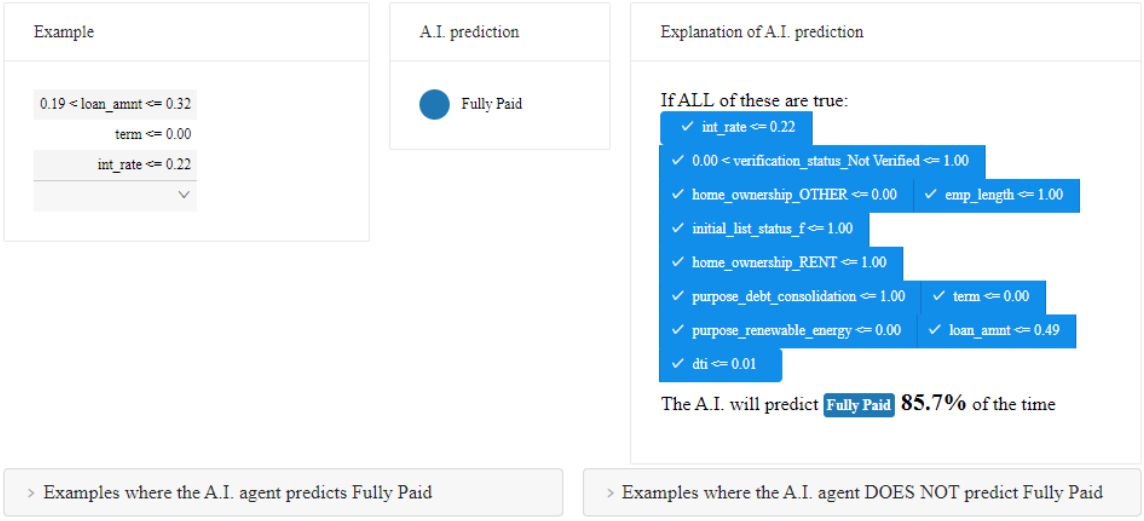


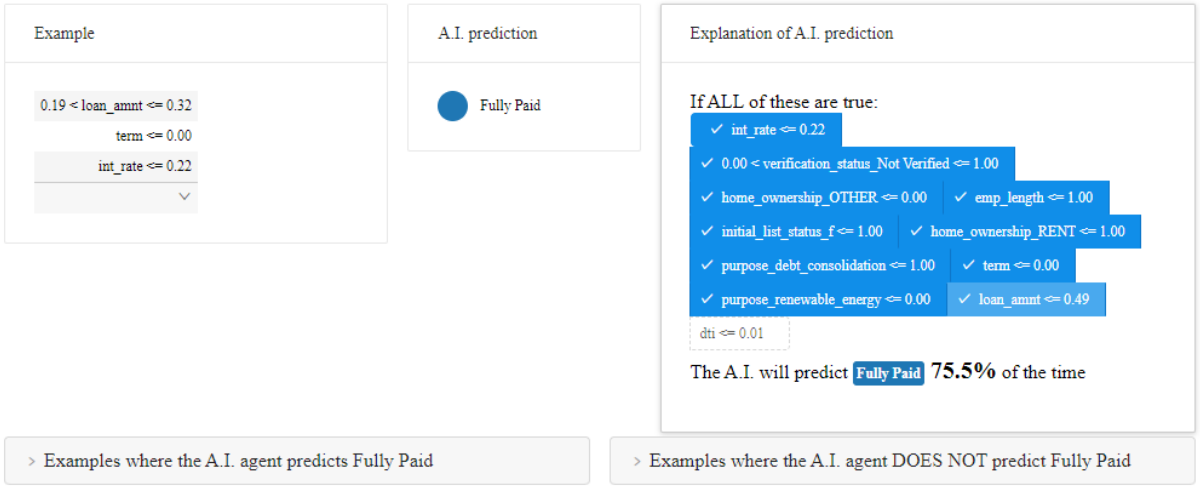
Figure 30: Total counts of loans based on the DTI ratio of the customer against the percentage of these loans that are Charged Off. Data is from our training set.

The Anchors framework also provides the user with a user-friendly, interactive dashboard that shows the precision based on the selected criteria of the user. In Figure 31 (a), we see on the left the instance that is taken as an example for this explanation, in the middle the prediction of the ML classifier, and on the right the precision of the ML classification based on certain rules. On the bottom side, the user can see more examples for both the class “Fully Paid” as for the class “Charged Off” when clicked on.

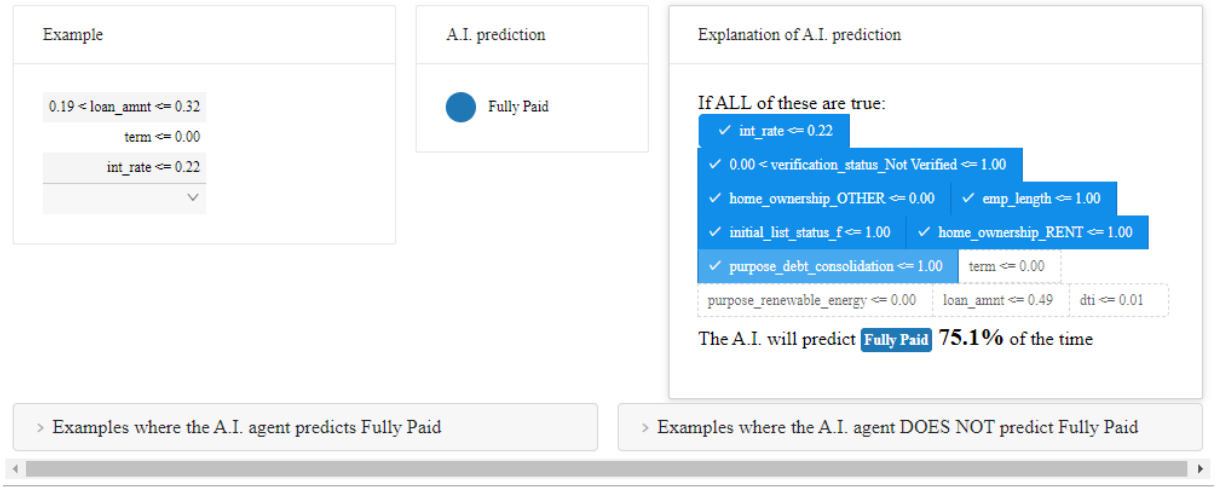
Now we deselect a rule, in this case, the rule of dti <= 0.01, the rule becomes a white box, and the precision of the ML classifier is changed accordingly as shown in Figure 31 (b). This dashboard also confirms our explanation about the rules of employment length, initial list status, home ownership rent, and purpose debt consolidation are redundant rules since the precision does not change as we can see from Figure 31 (c) and (d). The reason why these features are still in the explanation is probably caused by the fact that we could not map the categorical features in the correct way for the Anchors explainer. At the moment this would have been the case, we would expect that only one of the categorical features would have shown for each feature.



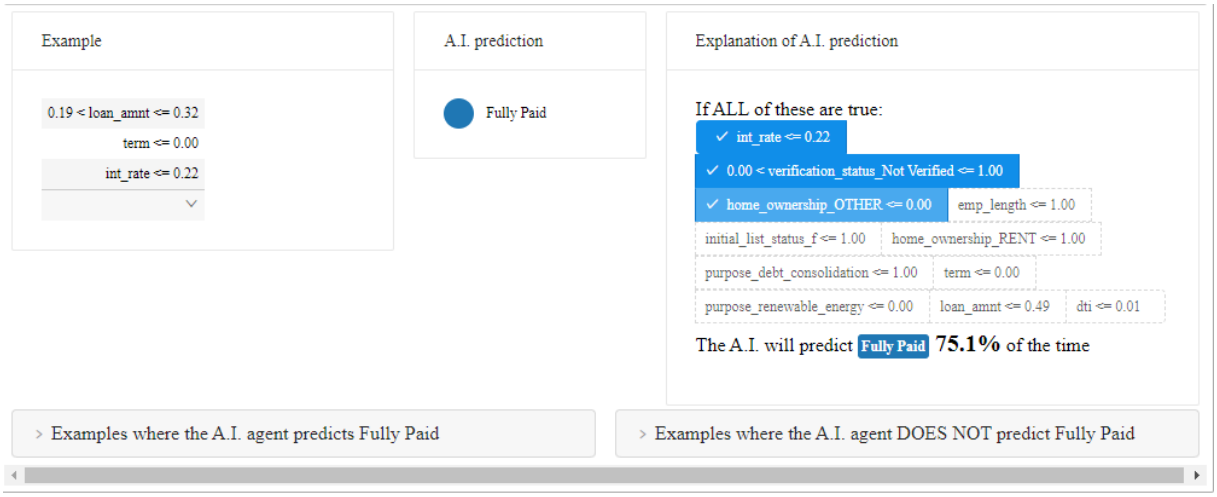
(a) The result of the Anchors explanation. The example can be folded out to show all feature rules.



(b) The result of the Anchors explanation when we do not fulfill the requirement of the dti.



(c) The result of the Anchors explanation when we do not fulfill several requirements.



(d) The result of the Anchors explanation when we do not fulfill the requirement of the dti.

Figure 31: Interactive dashboard of the Anchors framework.

For the prediction of the XGBooster model on the customer instance that is in the class “Charged Off” we see fewer rules that lead towards this prediction, however, the coverage of this explanation is minimal with a score of zero. Both verification statuses “Verified” and “Source Verified” must be zero, resulting in the verification status “Not Verified” must be one. Based on Table 12, we observe that this is not logical for predicting the customer as “Charged Off” since this status has the lowest percentage of loans that are Charged Off. This can be explained by the Anchor trying to explain this specific customer instance in which the verification status is “Not Verified” and therefore tries to generate an explanation that takes this into account. This also explains why the coverage is minimal. Next to that, the interest rate must be larger than 0.46 (= 17%) to predict this customer as “Charged Off”. This makes sense and is in line with financial logic since a higher interest rate in the case of the LendingClub means that the borrower is classified as less creditworthy, and a higher interest rate means that the loan is more difficult to pay off and therefore riskier. The term being equal to 60 months and a DTI score larger than 0,03 (= 30) is again in line with financial logic and can be explained in the same way as we did for the “Fully Paid” customer of the XGBooster model the opposite way around. The last rule of the FICO score having to be less than or equal to 0.22 (= 688) to predict this instance as “Charged Off” is also in line with the earlier described financial logic of a lower FICO score meaning that the borrower is less creditworthy and therefore granting a loan is riskier.

Support Vector Machine model

In Figure 32 we see the correct prediction of the SVM model for the customer instance that is in the class "Fully Paid". As we can see from the figure, this prediction is based on only two rules: the verification status must be "Not Verified", and the term must equal 36 months. We observed these rules already for the LR and XGBooster models and based on these explanations we can say that these rules are in line with the expected logic. We see a high value for the coverage of this explanation, this can be caused by the method of the SVM of applying different hyperplanes and meaning that these two features are in the same hyperplane for customers that are classified as "Fully Paid".

```
Prediction:  Fully Paid
Anchor: verification_status_Not Verified > 0.00 AND term <= 0.00
Precision: 0.96
Coverage: 0.21
```

Figure 32: SVM explanation for "Fully Paid" for Customer A by Anchors

For the prediction of the SVM model on the customer instance that is in the class "Charged Off" from Figure 22 (e), we observe rules that are in line with financial logic. We already discussed that a term of 60 months and a higher interest rate are riskier, especially the combination of the two is risky and therefore push the prediction towards the class of "Charged Off". The rule of not having more than 0 mortgages accounts seems to contradict financial logic. Having no mortgage accounts could mean that the customer already paid off their mortgage and are therefore more creditworthy. However, looking at the visualization of the data on the number of mortgage accounts against the percentage of loans that are Charged Off in Figure 33 we see the importance of data quality, almost half of the observations in our training set are customers that do not have any mortgage accounts, resulting in a highly imbalanced feature. This imbalance has a major impact on the prediction of the ML classifiers, resulting in a rule that is not in line with financial logic. If the number of cases per unit was the same, we would expect an upward trend instead of the current downward one. This will of course have a major impact on the way our ML classifiers are currently assessing the value of the feature mortgage accounts.

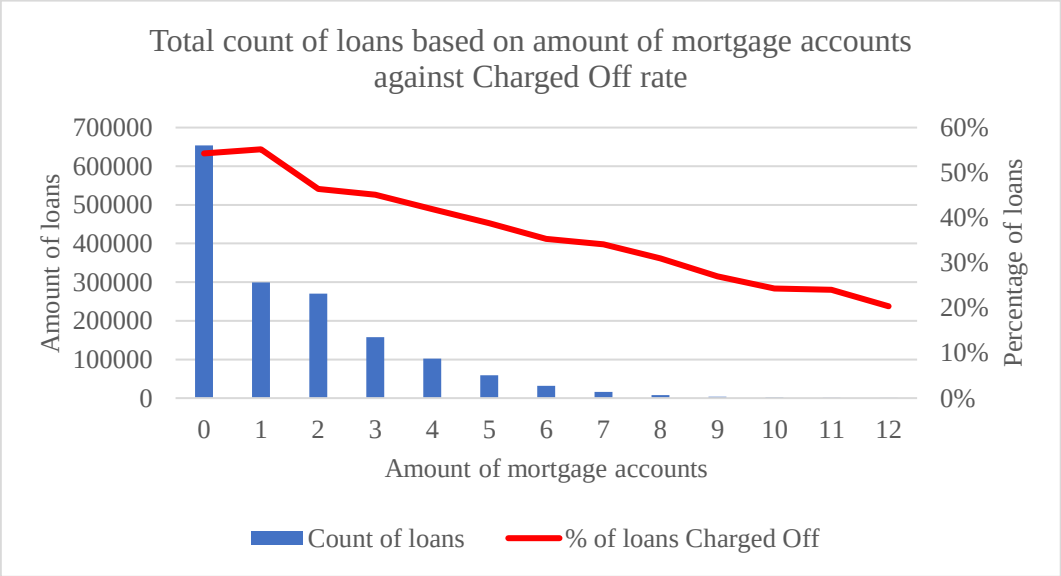


Figure 33: Total counts of loans based on the number of mortgage accounts of the customer against the percentage of these loans that are Charged Off. Data is from our training set.

5.3.3 Conclusion on Anchors

Going in-depth to find the actual information behind the explanation and rules provided by the Anchor framework resulted in a better understanding of how the different ML classifiers work. The explanations of the Anchors framework are in line with financial logic or what we would expect based on the provided input data. The if-then rules provided by the Anchors framework work the best with the tree models like the XGBooster and give these models the best understanding. It would have been nice if the Anchors framework explained the instances for other ML classifiers with some more rules since we were now limited with two rules. However, this is also caused by the high threshold of 95% precision we set for the Anchors explainer. One of the drawbacks of the Anchors explainer is the challenge to implement the ANN model, making the explainer less suitable as a practical model-agnostic explainer. We find the explanation generated by the Anchor completer and more useful compared to that generated by the LIME explainer. The coverage gives us a better understanding of the region the rules generated by the Anchor explainer apply to.

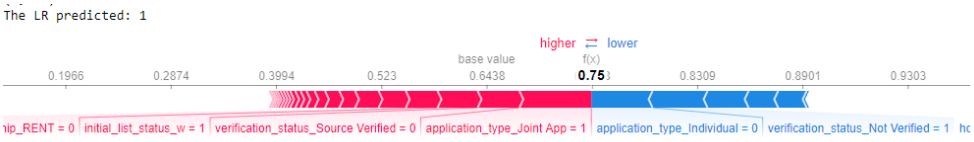
5.4 Applying SHAP to explain Local Instances

After we conducted the steps for getting the explanations from the SHAP framework, we applied the technique to the trained ML classifiers. Based on our research focusing on model-agnostic XAI techniques, we decided to use the SHAP kernel explainer from the SHAP library of explainers. The SHAP kernel explainer is model-agnostic and is, therefore, more suited for our research instead of the model-specific explainers that the SHAP library offers. A drawback of the SHAP kernel explainer is the slow exponential time complexity of this explainer and therefore we were not able to generate SHAP values for the entire dataset of 1.6 million data instances. We used a small sample of 1500 data points to generate the SHAP values. This results in an approximation rather than exact Shapley values. In this section, we focus on explaining the local instances by the SHAP framework, while in section 5.5 we focus on the global explanation of the different ML classifiers using the SHAP framework.

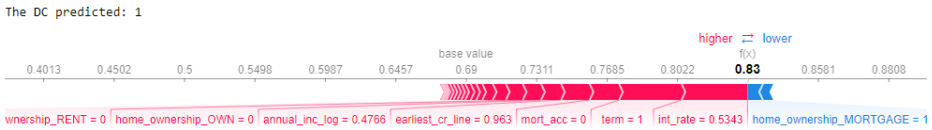
5.4.1 Explanation of SHAP – Local

To illustrate the working of the SHAP framework, we present in Figure 34 (a-f) the SHAP explanations on the different ML classifiers for Customer B. The base value is the value that would be predicted if we did not know any features for Customer B, meaning that the base value is equal to the mean prediction. The output value $f(x)$, shown in bold, is the actual prediction for the data instance that we selected, in our case Customer B. The output of the SHAP technique is presented in a log odds ratio instead of a probability space. In the figure, we see how each feature contributes to forcing the model output from the base value to the output value. The Shapley values in our figures are explanations with respect to the negative class "Fully Paid", meaning that we see if features are contributing positively or negatively to the prediction of our negative class "Fully Paid". The red bars present the top features that contribute positively to the prediction of the class "Fully Paid". The blue bars present the top features that contribute negatively to the prediction of the class "Fully Paid" or viewed differently, contribute positively to the prediction of the class "Charged Off". At the moment a red feature is dropped, the output value will move the length of the bar of that feature to the left. This means that prediction becomes less likely to be the class "Fully Paid". In case a blue feature is dropped, the prediction will move the length of that bar of that feature to the right and making it more likely the class "Fully Paid" is predicted.

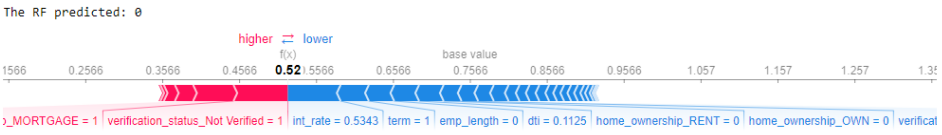
In section 5.4.2, we interpret the SHAP local explanation of the LR, XGBooster and Neural Network models in more detail and based on this interpretation the other explanations by SHAP can also be better understood.



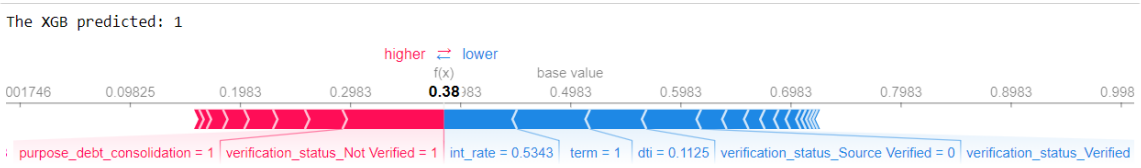
(a): Logistic Regression: SHAP explanation for "Charged Off"



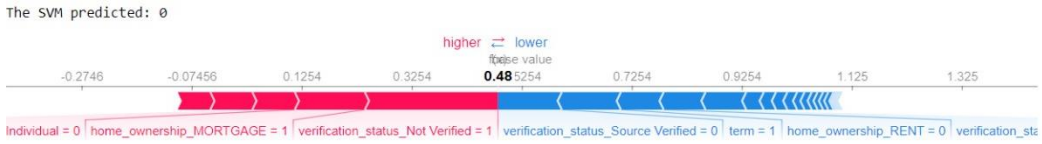
(b): Decision Tree: SHAP explanation "Charged Off"



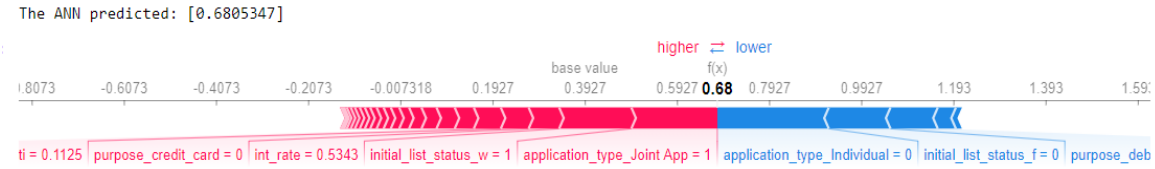
(c): Random Forest: SHAP explanation for "Fully Paid"



(d): XGBooster: SHAP explanation for "Charged Off"



(e): Support Vector Machine: SHAP explanation for "Charged Off"



(f): Neural Network: SHAP explanation for "Charged Off"

Figure 34: Results from the SHAP explainer. We observe an incorrect prediction for the RF and ANN model.

5.4.2 Interpretation of SHAP – Local

In this section, we discuss the local explanations given by the SHAP explainer for the outcomes of the Logistic Regression, XGBooster, and Neural Network model on both customers A and B, described in detail in Table 11. We expand further on the interpretations of these explanations to check if they are in line with financial logic and compare them with the provided input data to get a better understanding of the functioning of SHAP.

Logistic Regression

In Figure 35 we see the correct prediction of the LR model for Customer A. From the figure we see that the feature verification status being "Not Verified" has the highest positive contribution to the prediction being the class "Fully Paid". Next are the features of initial list status being not "Whole", home ownership being "Mortgage" and the purpose being "Debt Consolidation" that have a positive contribution of forcing the base value to the prediction being the class "Fully Paid". Based on Table 12 almost all features with their values are in line with what we would expect since these values have a lower percentage of loans that are Charged Off. However, the purpose being "Debt Consolidation" is contrary to what we would expect and cannot be logically explained from a financial point of view either. Borrowing to pay off debt makes one less creditworthy.

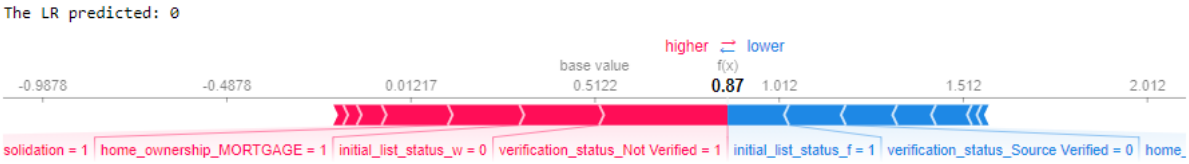


Figure 35: SHAP local explanation predicted "Fully Paid" for Customer A for the LR model

On the other hand, we see that the features initial list status being "Fractional" and verification status being not "Source Verified" having a negative contribution to the prediction being the class "Fully Paid". The initial list status being "Fractional" is the value with a lower percentage of loans that are Charged Off compared to the other value of "Whole". We therefore would expect that the initial list status being "Fractional" contributed positively to the prediction being the class "Fully Paid' and initial list status not being "Whole" having a negative contribution to the prediction. The opposite of the situation we see now. The verification status being not "Source Verified" increases the likelihood of verification status being "Verified" and could mean that it, therefore, has a negative contribution on the prediction being the class "Fully Paid".

In Figure 34 (a) we see the correct prediction of the LR model for Customer B. The low output value shows that most features are having a negative contribution to the prediction being the class "Fully Paid", which is correct for Customer B. We see that the feature application type not being "Individual" and verification status being "Not Verified" have a positive contribution to the outcome being in the class "Fully Paid". However, these features are dominated by the features that have a negative contribution to the class being "Fully Paid". In Figure 33 (a) only a few of the negative contributors are shown, the high interest rate and low annual income of Customer B are also having a negative contribution of the class being "Fully Paid" but are not shown in the figure. A high interest rate and low annual income are in line with financial logic to have Customer B classified as class "Charged Off" since a lower annual income makes it more difficult to pay off a loan, especially with a high interest rate.

XGBooster

In Figure 36 we see the correct prediction of the XGBooster model for Customer A. The high output value shows that the values of the features interest rate, FICO score, number of mortgages, and open accounts have the highest positive contribution to the prediction being the class "Fully Paid" followed by the verification status being "Not Verified" and the home ownership being "Mortgage". The values of these features resulting in a prediction of the class "Fully Paid" are in line with financial logic. Having a relatively low interest rate of 6% (see also Figure 27) with a relatively high FICO score of 747 (see also Figure 22) shows that Customer A can be considered creditworthy, supported by having only one mortgage account and a relatively low amount of open accounts with a value of 9. Having a smaller number of open credit lines means that the customer has less debt and is therefore financially healthier. Verification status being "Not Verified" and the home ownership being "Mortgage" have already been discussed as logical contributors to the score based on Table 12.

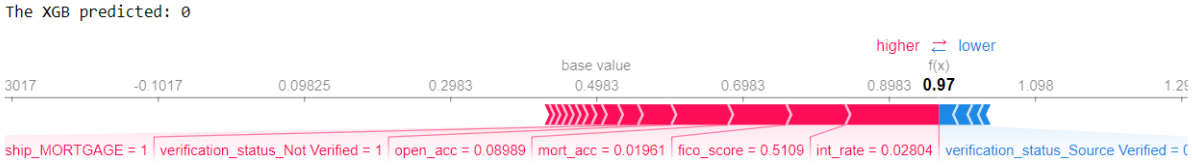


Figure 36: SHAP local explanation predicted "Fully Paid" for Customer A for the XGBooster model

In Figure 34 (d) we see the correct prediction of the XGBooster model for Customer B. The lower output value shows that more features are having a negative contribution to the prediction being the class “Fully Paid”, which is correct for Customer B. The feature verification status being “Not Verified” has a large positive contribution to the outcome being in the class “Fully Paid”, as can be seen from the size of the bar. However, the total of positive features is dominated by the features that have a negative contribution to the class being “Fully Paid”. From Figure 33 (a) we see that the high interest rate, the term being 60 months, and a high DTI ratio are having a negative contribution for the prediction being the class “Fully Paid”. The high interest rate is already discussed in the interpretation of the Logistic Regression. Having a loan for a longer period is riskier since there is a higher chance of something will go wrong and the borrower won't pay the loan back, this is also supported by Table 12. From a financial perspective, the riskier longer loan most of the time has a higher interest rate to compensate for this risk. This means that the features are likely to be related in some way. Another feature that has a negative contribution is the high DTI ratio of being 111. A DTI of 111 is highly unlikely to be an actual value since it means that Customer B spends more than his income on debt payments. Such a high DTI ratio is a good indicator for classifying Customer B as “Charged Off”. However, from this explanation, we can also conclude that we should have checked the data on outliers to increase the data quality and get a situation that best simulates the real world.

Neural Network

In Figure 37 we see that the prediction of the Neural Network model for Customer A is incorrect with a low output value for Customer A being in the class “Fully Paid”. The low output value shows that the feature initial list status being “Fractional” and the purpose not being “Credit Card” have the highest positive contribution to the prediction being the class “Fully Paid”. Based on Table 12, we see that the initial list status “Fractional” has a lower percentage of loans that are Charged Off compared to the other value of “Whole”. Purpose not being “Credit Card” has a positive contribution since this purpose has a high percentage of loans that are Charged Off. However, we would not expect that feature contribution in our explanation since the purpose is “Debt Consolidation” which is already shown as a feature that has a negative contribution to the prediction being the class “Fully Paid”. This negative contribution makes sense since the purpose “Debt Consolidation” has a high percentage of loans that are charged off. The initial list status not being “Whole” and verification status being “Not Verified” having a negative contribution to the prediction being the class “Fully Paid” is not supported by Table 12. Especially from the verification status being “Not Verified” we would expect that it contributed positively to the prediction being the class “Fully Paid”, as we also observed for the LR and XGBooster models. The same holds for the low interest rate and DTI ratio having a negative contribution to the prediction being the class “Fully Paid”. We would expect those feature values to have a positive contribution to the outcome of prediction being the class “Fully Paid”. This suggests that the Neural Network model in this case made the wrong prediction based on interpreting the feature values in a way that is not in line with financial logic.

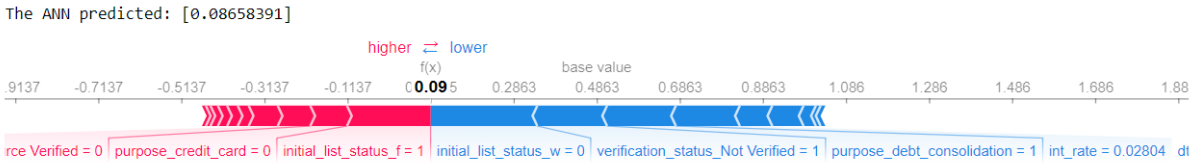


Figure 37: SHAP local explanation predicted “Charged Off” for Customer A for the Neural Network model

In Figure 34 (f) we see the incorrect prediction of the ANN model for Customer B. The high output value shows that the model predicts Customer B as the class “Fully Paid”. We see that most features are having a positive contribution to the prediction being the class “Fully Paid”, which is incorrect for Customer B. Based on Table 12, the initial list status “Fractional” can be expected to contribute positively to the prediction being the class “Fully Paid”. However, the high feature values of interest rate and DTI ratios are expected to contribute negatively to the prediction being the class “Fully Paid”. As we can see from the figure, the opposite is happening. Again, this suggests that the Neural Network model in this case made the wrong prediction based on interpreting the feature values in a way that is not in line with financial logic.

5.4.3 Conclusion on SHAP – Local

Going in-depth to find the actual information behind the local explanations and rules provided by the SHAP framework resulted in a better understanding of how the different ML classifiers work. Most of the explanations of the SHAP framework were in line with financial logic or what we would expect based on the provided input data. The figures provided by the SHAP framework are easy to read and clearly show how each feature contributes to a prediction. The user just needs to understand from which class we are viewing the Shapley values so that the direction of the contribution is correctly interpreted. We observed that the ANN model made the wrong prediction for both Customer A and B, by interpreting the feature values in a way that is not in line with financial logic. Therefore, it is interesting to see if the global explanation of the ANN model is explaining the instances in the same way or that our examples happen to be exceptions to the rules.

Customers with a low interest rate, high FICO score, low DTI ratio, and a low number of mortgages and open accounts are marked as less risky and are more likely to have their loans approved. This was evident from the examples we discussed above and is in line with financial logic. Similarly, high-risk customers can be identified by riskier purposes, lower FICO scores, higher interest rates, and a higher DTI ratio. This information is helpful information for model developers, users that provide loans, and the end-users since it offers more transparency and increases trust in the predictions.

5.5 Applying SHAP to explain Global Instances

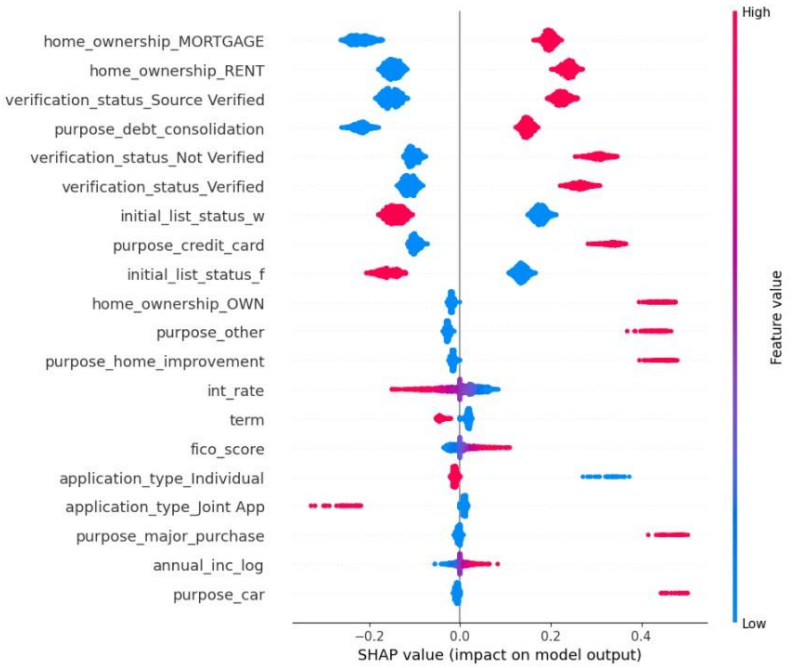
In this section, we focus on the global explanation of the different ML classifiers using the SHAP framework. The generated global explanations given by the SHAP framework are presented in Figure 38 (a-f). In the same way, as for the local explanations, we used the SHAP kernel explainer and a small sample of 1500 data points to generate the SHAP values. This results in an approximation rather than exact Shapley values.

5.5.1 Explanation of SHAP – Local

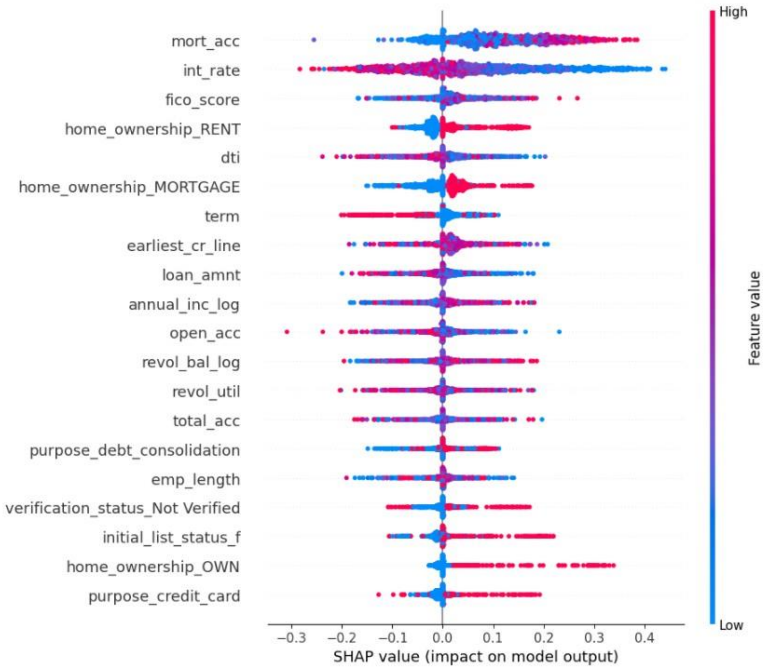
To illustrate the working of the SHAP framework, we present in Figure 38 (a-f) the global explanations of the different ML classifiers by SHAP. The figure shows the global impact of the top 20 features on the prediction of the different ML classifiers. The features that are presented on the top have more impact than the features that are presented below them. In this figure, the red and blue colors present the value of the feature. Red means a high value of the feature, in the case of loan amount it represents a high loan amount. Blue means a low value of the feature, in the case of loan amount it represents a low loan amount. On the X-axis the feature impact on the model predicting a data instance as the class "Fully Paid" is presented. Feature values that are plotted to the right have a positive impact on the prediction being the class "Fully Paid". On the other hand, feature values that are plotted to the left have a negative impact on the prediction being the class "Fully Paid", meaning it forces the prediction towards the class of "Charged Off".

Taking Figure 38 (b) as an example, we see that the interest rate is the most important feature after the amount of mortgages accounts and that a high feature value of interest rate contributes negatively to the prediction being the class "Fully Paid". The lower the feature value of interest rate becomes, the more it has a positive impact on the prediction being the class "Fully Paid" as we see by the color becoming more light blue.

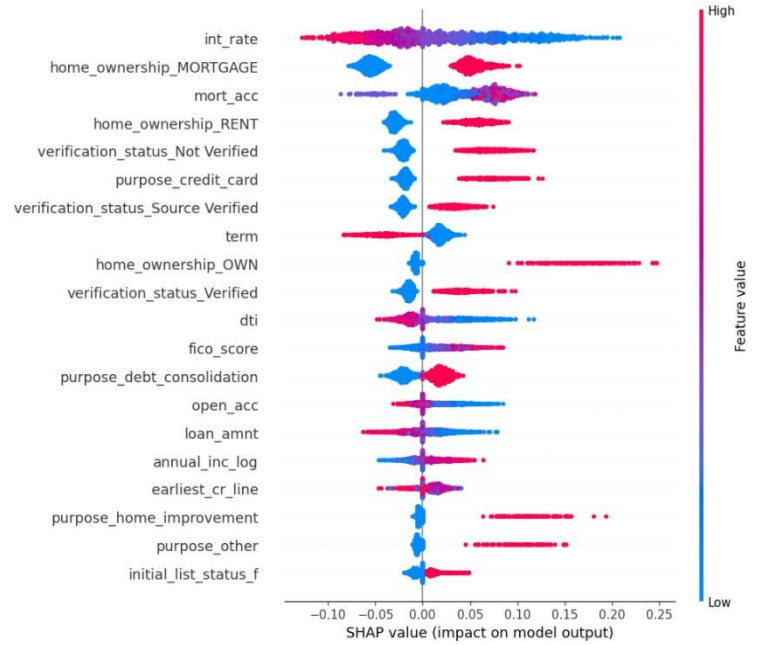
In section 5.5.2, we interpret the SHAP global explanation of the LR, XGBooster, and Neural Network models in more detail, and based on this interpretation the other global explanations by SHAP can also be better understood.



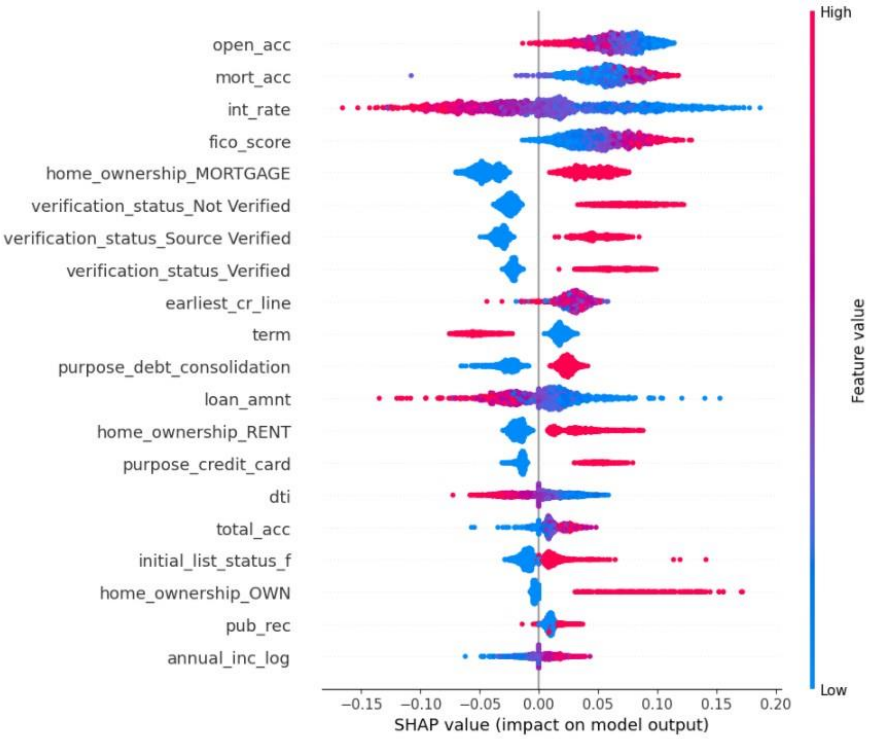
(a) Logistic Regression: Global SHAP explanation



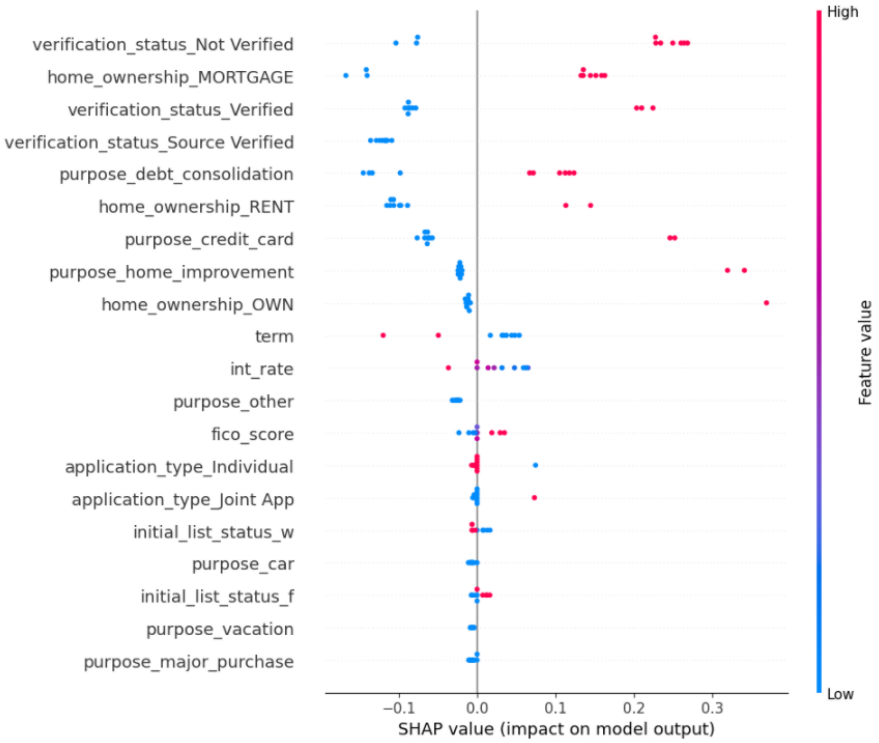
(b) Decision Tree: Global SHAP explanation



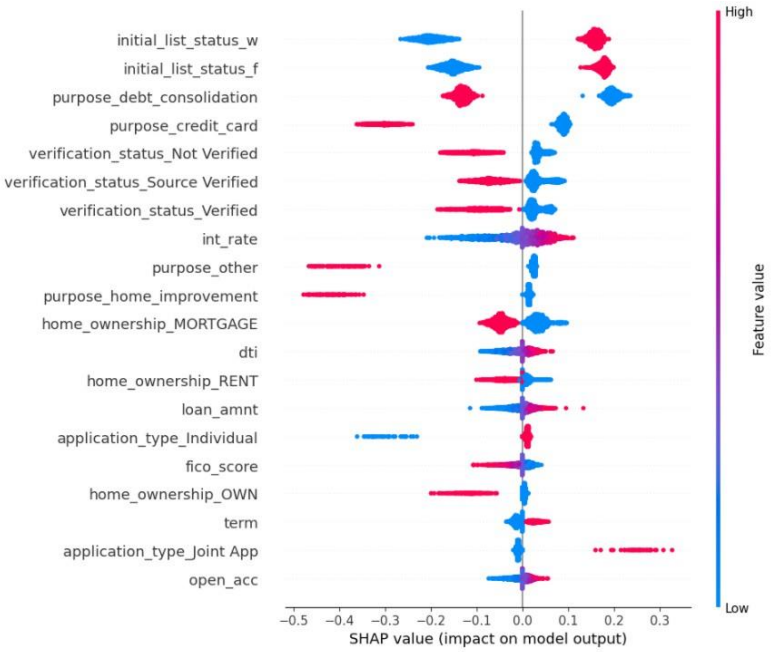
(c): Random Forest: Global SHAP explanation



(d): XGBooster: Global SHAP explanation



(e): Support Vector Machine: Global SHAP explanation. Due to the long computation time for the SVM model, only 10 data points are used instead of the 1500 instances.



(f): Neural Network: SHAP explanation for "Charged Off

Figure 38: Results from the SHAP explainer on the global impact

5.5.2 Interpretation of SHAP – Global

In this section, we discuss the global explanations given by the SHAP explainer for the outcomes of the Logistic Regression, XGBooster, and Neural Network model. We expand further on the explanation to check if they are in line with financial logic and/or justify the explanation based on the input information.

Logistic Regression

In Figure 38 (a) we see the global explanation of the LR model with the corresponding feature importance. From the figure, we see that the categorical features are having the most impact on the prediction of the model. In the case of these categorical features are a high value, so 1, they have a positive impact of the prediction being the class "Fully Paid" except for the feature initial list status. We would expect that a high value of the feature values purpose "Debt Consolidation" and "Credit Card" would have a negative impact on the prediction being the class "Fully Paid" if we take the high percentage of loans that are charged off from Table 12 into account. We see that a high feature value for the purposes "Major Purchase" and "Car" have a greater impact on the prediction of the model compared to that of the purposes "Debt Consolidation" and "Credit Card". Based on Table 12 this makes sense since these purposes have a lower amount of loans that are charged off and therefore are more likely to result in a customer paying back its loan. However, these features are considered of less importance by the LR model as we can see in the lower ranking.

In line with financial logic, we see that a high interest rate and a longer term have a negative impact on the prediction being the class "Fully Paid". The same holds for a high FICO score and a high annual income having a positive impact on the prediction being the class "Fully Paid". This is all in line with financial logic since it gives a good estimation of the financial health of a customer and therefore the chance that this customer will or will not pay back his loan.

XGBooster

In Figure 38 (d) we see the global explanation of the XGBooster model with the corresponding feature importance. From the figure, we see that the continuous features amount of open credit and mortgages accounts are considered the most important by the XGBooster model, followed by the interest rate and FICO score. We see that a lower amount of open credit accounts has a positive impact on the prediction being the class "Fully Paid", which is in line with financial logic. However, we see that only the highest values of this feature are having a negative contribution to the prediction being "Fully Paid". We would expect that this distribution was more shifted to the left so that any amount of open credit accounts greater than the average would have a negative contribution to the prediction being "Fully Paid". This can be caused by the training dataset not being represented with the situation we expect based on financial logic or it is a better distribution to present the creditworthiness of a customer that the XGBooster discovered.
For the feature mortgage accounts, we see the inverse distribution than we would expect based on financial logic. Having many mortgage accounts has a positive contribution to the prediction being "Fully Paid" and having no mortgage accounts result in less creditworthiness of the customer. We already discussed this observation in the interpretation of the Anchors model for the SVM model and the cause is probably the quality of the input data as we have shown in Figure 33. It shows us that this wrong input data if we are basing ourselves on financial logic, can have a great impact on the prediction results of our model. It also shows us where we can improve the models to improve their performance.

We see that a longer-term, higher loan amount, higher DTI ratio and lower annual income as feature values have a negative impact on the prediction being "Fully Paid". This is in line with financial logic since these are all factors that make a customer less creditworthy and the loan more difficult to pay back and therefore can be considered as riskier to pay back the loan. For the feature earliest credit line, we see a random distribution from which it is difficult to draw conclusions.

Neural Network

In Figure 38 (f) we see the global explanation of the Neural Network model with the corresponding feature importance. It confirms the explanations we observed for the local instances of Customer A and B that were incorrectly predicted. Comparing the global explanation of the Neural Network with all other ML classifiers, we observe that is presented exactly in the opposite way. Therefore, we think the configuration of the Neural Network for the XAI techniques is done in the opposite way. Meaning that it now predicts the importance of the feature to contribute to the class of "Charged Off". This is a modeling error by us and can explain the opposite explanation we observed for the LIME technique. We think this is the case, otherwise, the prediction accuracy of the ANN would have been considerably lower. However, this is a topic that can be researched in further research.

5.5.3 Conclusion on SHAP - Global

The SHAP framework providing us with the global explanations, and therefore inner workings, of the ML classifiers gives a better understanding of how the different ML classifiers work. The explanations the SHAP framework provided us on the global working of the models were mostly in line with financial logic. The impact of features that are not in line with financial logic we can explain based on the provided input data. It shows us the features that we still need to configure before we use them as training data for our ML classifiers to improve their performance. This is also needed before the ML classifiers may be considered for use since a current explanation of the working of the ML classifier to a customer affected by their outcome can have disastrous consequences. We take the feature mortgage accounts as an example.

Suppose a customer is rejected for a loan and goes to one of the model explainers and asks why the loan has been rejected. The model explainer explains why the model made this decision and the customer asks how it can improve its likelihood of being granted a loan. The model explainer then says, based on the explanation of the model, that the likelihood can be increased by having more mortgages accounts. If both parties follow this blindly, it will of course have major consequences for the customer who suddenly starts taking more risky mortgages.

The global explanation figures provided by the SHAP framework are easy to read and clearly show how each feature contributes to a prediction. Again, the user needs to understand from which class we are viewing the Shapley values so that the direction of the contribution is correctly interpreted.

From the global explanations, it becomes clear that customers with a low interest rate, high FICO score, low DTI ratio, short term of the loan, low loan amount, high amount of mortgages accounts are marked as less risky and are more likely to have their loans approved. This was evident from the examples we discussed above and is mostly in line with financial logic. Similarly, the opposite of these mentioned feature values results in customers being marked as riskier. This information is helpful information for model developers, users that provide loans, and the end-users since it offers more transparency and increases trust in the predictions.

5.6 APPLYING THE ARRGUS FRAMEWORK TO THE XAI TECHNIQUES

By applying the ARRGUS framework that we constructed in Chapter 3, we are be able to assess the XAI techniques on the different indicators an XAI technique must comply with. With the ARRGUS framework, we can give the XAI techniques scores on the different indicators based on our opinion. These scores together form the final total score for the XAI technique and based on that total score we can determine which XAI technique performs the best. We present the assessment of the XAI techniques in Table 13.

			LIME	Anchors	SHAP
Indicator		Explanation	Rating: 1 (Very Poor) to 5 (Very Good)		
1.	Accuracy	To what degree can the XAI technique explain how well the explanations reflect the behavior of the prediction model?	3	4	3
2.	Readability	To what degree can the XAI technique generate explanations that are understandable for the targeted audience group, the average borrower/lender?	3	4	2
3.	Robustness	To what degree can the XAI technique explain whether and to what extent each individual input parameter has contributed to the outcome.	2	3	5
4.	Generalizability	To what degree can the XAI technique generate similar explanations on Machine Learning models that are trained on the same task.	4	4	4
5.	Usability	To what degree can the XAI technique be used effectively on a range of machine learning models as an explanation method?	4	2	4
6.	Stability	To what degree can the XAI technique generate similar explanations for similar instances and do these explanations reflect the same amount of certainty of a model about its predictions?	3	3	4
Average score:			3,17	3,33	3,67

Table 13: Overview of the score per indicator for each XAI technique.

From Table 13 we conclude that SHAP has the best performance followed by Anchors and LIME in that order. We like to stress that this assessment is based on our opinion and therefore we support each score per indicator with a brief explanation on why we awarded the score. All these scores are based on our opinion based on how we experienced the whole process from implementation to the results of the XAI technique.

Accuracy
We awarded Anchors with the highest score for this indicator since it is the only XAI technique that shows a precision score on how well it reflects the behavior of the ML classifier. Therefore, we assigned a higher score of 4 to this technique and a general score of 3 to the LIME and SHAP technique since they still are able to reflect the behavior of the ML classifier but without an indicator on how accurate they are doing this.

Readability
SHAP is awarded the lowest score since the user must understand how Shapley values work and that they are based on the class that is actually predicted. This can lead to a lot of misinterpretation of what the explainer is telling us. Next to that, it can confuse the user with the base and output value. We awarded LIME with the middle score for this indicator since the rules

and the impact of the different features were not always clear from the provided explanation. Anchors is given a higher score since it presents the explanation in a more intuitive way and offers an interactive dashboard to experiment with the explanation and see other examples.

Robustness
SHAP is awarded the highest score for robustness since it shows perfectly how each feature contributes to the prediction and it can show for each feature what happens to the prediction if the input value is changed, both on a local and global level. Anchors shows for which value the feature contributes to the actual prediction and by the interactive dashboard the impact of each feature can be shown, resulting in the middle score for this indicator. On the other hand, LIME shows the top ten features that are contributing to the outcome but the impact that these features have on the prediction is most of the time not clear.

Generalizability
For this indicator, we awarded the XAI techniques with the same score, since we observed that they almost generate similar explanations for the different ML classifiers that we trained. We observe some differences in the importance of certain features, however in general they show the same explanations for the same ML classifiers.

Usability
Anchors is awarded the lowest score for this indicator since we could not configure the explainer to generate an outcome on the Neural Network model. This reduces the effectiveness of Anchors as a model-agnostic XAI technique enormously. Both LIME and SHAP did not have the exact problem, however, the configuration of the Neural Network model is a tricky one and results in questionable results.

Stability
SHAP is awarded a higher score since we can check this XAI technique on generating similar explanations for its local instances and its overall global explanation of the model. Based on this comparison we can observe how certain the model is about its prediction. This is an option that the local explanations of LIME and Anchors cannot provide.

5.7 Conclusion on the Experimental results

In this chapter, we focused on answering the research question "*Which XAI technique performs the best on explainability, based on our valuation framework ARRGUS?".* Based on our result we can answer this research question by stating that SHAP performs the best based on our ARRGUS framework, of which the scores are determined based on our opinion. If more research is done, the ARRGUS' indicators can be made quantifiable and the performance of the XAI explainers could be assessed on an objective basis to see which meets the requirements and expectations of both industry and stakeholders the best.

6. CONCLUSION

The main aim of this paper is to research the current knowledge on AI and ML applications in the credit risk prediction industry, to develop an exploratory framework to assess XAI techniques based on the current regulation, and finally to implement three advanced model-agnostic post-hoc explainability techniques (LIME, Anchors, and SHAP) on outputs obtained from current used ML-classifying credit scoring models. To achieve this goal, we came up with the following research question for our research:

"In what way and to what extent can explainable AI algorithms improve the explainability of decision-making in AI models used in P2P credit risk prediction?"

We designed our exploratory framework ARRGUS based on the current state-of-the-art research in the areas of Credit Risk, Machine Learning Classifiers, Explainable AI, and Social Sciences in combination with the regulation on the application of advanced AI models in financial services. With ARRGUS we assessed the different XAI algorithms on *Accuracy, Readability, Robustness, Generalizability, Usability, and Stability* to see how well they comply with our developed standards on the added value of applying XAI algorithms.

Our results show that all three XAI algorithms provide a fairly consistent explanation that we can justify based on the input data provided and are in line with financial logic. We observed some dominant features in all three explainers, which strengthens our confidence in stable outcomes provided by the explainers. From the explanations of our XAI algorithms, it became clear that customers with a low interest rate, short term for the loan, low loan amount, low DTI ratio, high FICO score, and a high amount of mortgages accounts are marked as less risky and are more likely to have their loans approved.

The XAI algorithms can provide all the stakeholders of AI usage with added value by providing some glimpse into what a black box model was previously. Our results show clearly that the XAI techniques can generate explanations that are understandable for the model builders, users of the model, and the end customers affect by its decision. Next to that, from our results, it is clear that the XAI techniques can explain whether and to what extent each input parameter has contributed to the outcome of the prediction. The Anchors and SHAP XAI techniques even provide the user with information on which features they can adjust to change the outcome of the prediction. Whether this is an option that we should want to pursue and/or make available to the public is another discussion, as it can lead to desired behavior or fraud/misrepresentation of the application form for example.

To conclude on in what way can XAI algorithms improve the explainability of decision-making in AI models, we conclude that the understandability of the explanations is different for each XAI technique and some require more explanation than others, but in general, we conclude that they are understandable for the user and certainly add value to a plain outcome of the model.

To conclude on what extent can explainable AI algorithms improve the explainability of decision-making in AI models we observe from our results that the XAI techniques can reflect the behavior of the prediction model, based on the outcome and rules we expect from the data and the Anchors explainer even give a precision score. From our results, we also see that the XAI techniques can generate similar explanations for the ML classifiers that are trained on the same task, which increases our confidence in the added value of using XAI techniques in practice. We also observe from our result that the XAI techniques can effectively be used on a range of ML classifiers models, but that there is still progress to be made to easily apply these correctly to all the different

ML models to be truly model-agnostic. Lastly, from our results, we see that the XAI techniques generate similar explanations for similar instances. However, as we can see from the Anchor explainer the explanations do not reflect the same amount of certainty of a model about these different predictions. Which leaves room for incorrect predictions based on the same set of rules. In this section we conclude that explainable AI algorithms show promising and useful results for improving the explainability of decision-making in AI models, however, these XAI techniques still need to take a maturity step in being more consistent in mimicking the inner workings of a model to be applied in practice.

We conclude that XAI techniques generate explanations that are understandable for all users involved or affected by the outcome of ML models and certainly add value to the outcome by indicating whether and to what extent each input parameter has contributed to the outcome of the prediction. XAI techniques show promising and useful results for improving the explainability of decision-making in black box AI models. Based on ARRGUS, the SHAP technique scores best on the indicators and is the most compliant and therefore we are of the opinion that this technique is the most promising to be applied in practice, especially given the variety of this technique, which was not discussed in this study.

However, these XAI techniques still need to overcome some practical challenges to support real-world finance applications. More research and practical examples are therefore needed before these techniques can be used responsibly and on a large scale. Given the growth in the use of AI applications, it would not surprise us if a new market emerged in the provision of services to make AI-solutions explainable and transparent. Therefore, the question will no longer be if we can explain black box decision-making but who can explain its black box decision-making the best with these XAI techniques.

7. DISCUSSION AND FURTHER RESEARCH

As concluded in Chapter 6, this research shows the possibilities and opportunities that XAI techniques have to offer in the financial ML industry for credit risk prediction and other fields where ML models are used. We conducted a research to investigate a possible solution to the current problems on the lack of algorithmic transparency in credit risk prediction. The lack of algorithmic transparency forms a barrier for adopting more automated, AI-based modeling solutions in credit risk prediction and many more. As humans, whether we are model builders, model users, or end customers, we must better understand the inner working of more advanced AI agents to trust them. The more we can prove the advantages of these advanced AI agents and overcoming the lack of transparency problem, the more the general public will trust AI. This will result in more credit providers deploying these advanced AI agents and encourage innovation in this area to keep developing more sophisticated systems

The results of our work are promising, however, there is still significant potential in developing more robust and reliable XAI techniques in the ML industry. During the research, we identified several topics that are left for further research, based on the shortcomings of our research and general topics that must be tackled before XAI can be successfully implemented. We divided our findings into further research topics based on our research and discussion topics based on the general field of XAI that we experienced and would like to shed some light on.

7.1 FURTHER RESEARCH BASED ON OUR STUDY

Hyper-parameter optimization
In this research, we have not focused on optimizing the hyper-parameter settings of our ML classifiers. We decided to leave the problem of choosing a set of optimal hyper-parameters outside the scope of this research since our main objective was to explore the potential of XAI techniques and not to find the optimal results for our ML classifiers. This can be regarded as a drawback of our research and we encourage any further research to use hyper-parameter optimization. To find the optimal hyper-parameter settings we suggest using a Grid-search with several iterations. When the ML classifiers can optimally solve the credit problem the results will differ and the XAI techniques could find different explanations for the working of the ML classifiers.

Prioritization on data preparation
With our in-depth analysis of the XAI techniques, we discovered several strange patterns in the training dataset that were either not representative of the real world or were wrongly provided by the customer. This showed, that even with a thorough preparation of the data, we still missed some inconsistencies within the data that may have caused some results not to be in line with financial logic and/or decrease the performance of the ML classifiers. These errors were amplified through the use of SMOTENC. The preparation of the data can be regarded as the most crucial step for the ML classifiers to work correctly and an enormously extensive investigation is required to pick out all the possible errors that could harm the correct working of the ML classifiers. Regardless of the research area, we recommend that further research pay more attention to ensuring the data quality rather than the performance of machine learning models. From the literature used, it also appears that there is more focus on the performance of machine learning models than on ensuring a correct data quality that these models say something about. Machine learning models that say something about incorrect data is of no practical use to us.

Explanations that are missing details / not making sense
In this research we only showed explanations for Customer A and B, two observations out of a test set of over 300.000 observations, to demonstrate the working of the different XAI techniques. By explaining two observations from different classes, we tried to prevent the classical error of only demonstrating the correct class since this can be misleading. Only showing the correct class can give unjustified confidence in the explanation technique and the black box. This is especially the case at the moment we miss a lot of details in the explanations or the explanations rules are not making sense. We observed this with the LIME explainer as we had to dive deeper into the training set to be able to understand why the ML classifiers had to make certain choices. Even after this deep dive, some of the explanations were not making sense to us making both the ML classifiers as the LIME explanations useless in practice.

Focus further research on the possibilities of the SHAP Explainer
We concluded that the SHAP XAI technique is the most promising technique based on our ARRGUS framework. In this research we only showed the model-agnostic Kernel explainer from the SHAP library, however, this library has many more explainers and functionalities to offer. For the tree models, a faster Tree explainer can be used for example, if one is interested in applying model-specific explainers. We showed only two types of plots for the SHAP explainer and encourage further research to also look into the other plots the SHAP explainer has to offer. These plots may help to get an even better idea of how the ML classifiers work or can be applied in a complete dashboard to show all possibilities attractively and clearly to all stakeholders. Therefore, we recommend further research to focus on the SHAP explainer to get the full potential out of this explainer for both local and global explanations.

7.2 Discussion topics on model-agnostic post-hoc XAI

Are XAI explanations faithful to what the original model computes?
Our used XAI techniques cannot have perfect fidelity concerning the original model, otherwise, the explanation was completely faithful to what the original model computes. This means that the XAI model would equal the original model and the original model would be already interpretable. The consequence is that any explanation method for a black box model can be an inaccurate representation of the original model and therefore providing incorrect information. An inaccurate, so low-fidelity, explanation model limits trust in the explanation and with that also a limited trust in the black box that it is trying to explain. To illustrate this principle, suppose we have an explainable model that has a 95% precision compared to the original model. This is the same as the minimum threshold we set for our Anchors explainer. The explanation model is correct 95% of the time and wrong 5% of the time, meaning that one-twentieth of the explanations are incorrect and these explanations cannot be trusted. As a result, the original black box model also cannot be trusted for these explanations. Without 100% precision, we cannot know which explanations are incorrect and therefore we cannot fully trust both the explanation and the original black box model.

Assessing Algorithmic Risk
One of the solutions that can be investigated for further research is by assigning a role to an algorithmic decision-making supervisor, as a data protection officer within the GDPR. Next to that, an algorithmic impact assessment, like the Third-Party Risk Assessment, can be made mandatory to estimate the risks of automated decision-making. If the risk is high, additional legal quality requirements can then be imposed.

The danger of typographical errors
Typographical errors are an important drawback of using overly complicated black box models for prediction purposes since they can cause incorrect calculations and therefore incorrect

outcomes in practice. Typographical errors can occur because people enter the data incorrectly in the application form. This can result in a type of procedural unfairness, whereby two identical individuals might be randomly granted different loan decisions. These kinds of mistakes would be hard to discover within the actual model and therefore have the potential to reduce the in-practice accuracy of complicated black box models. In our opinion, it is crucial to discuss how these typographical errors can be prevented or noticed in time as they may have far-reaching consequences.

The term Explanation
As already mentioned in Chapter 1 there is no consensus on the term "explanation", resulting in researchers using the term in different ways. We also want to make a small contribution to this discussion because explanatory models do not always try to mimic the calculations of the original model. An explanatory model seeks to explain the choices made by the original model. Even an explanatory model that performs almost identically to a black box model may use completely different features or relationships between features, and thus not be faithful to the black box calculation. Even without perfect fidelity, the XAI techniques still provide explanations that show useful trends in how predictions are related to the features. It might be an idea to calling these explanations "feature trends" rather than "explanations" to be less misleading.

Implementing XAI techniques in a P2P platform
A difficult challenge before XAI techniques can be implanted in a P2P platform and act a as real gamechanger is how the technique can be applied on a large scale without taking too long to generate an explanation. The time to generate explanations is based on the ML classifier, type of XAI technique and within the XAI technique, the type of Kernel used. These choices are also greatly affecting the type and accuracy of the explanations and with that the readability for the users of the P2P platforms. Next to that, the lender wants to see from the explanations what feature values are making the potential borrower worthy to invest while the borrower wants to see on which features values and rules he or she is assessed with a certain credit decision. Providing the borrower with all the information as to why he or she has been turned down for a loan may cause that person not to fill in the application form truthfully or otherwise engage in desirable behavior to cheat. However, under the GDPR, any individual subject to such a decision has this right. This could lead to a system that does not work as intended and such problems should be seriously considered before implementing XAI techniques in a P2P platform environment.

Human precision, the real bottleneck?
The core of interpretability and/or explainability is whether humans understand a model well enough to make accurate predictions about its behavior on unseen instances. At the moment humans can confidently predict the behavior of a model, let the 'human precision' be the part in which they are correct (so not the model precision). A high level of human precision is necessary for real interpretability. A human can hardly say they understand a model if they consistently think they know what it will do, but are often mistaken. So an important discussion is how do we get on with explaining black box models if we as humans are unable to understand them at all?

7.3 CONTRIBUTION OF THIS RESEARCH

The contributions of our research can be summarized by:
- The first study that used three different XAI algorithms to explain ML classification models.
- A set-up framework ARRGUS for assessing the explainability of machine learning models.
- Applying a feature engineering strategy, including an resampling technique, on the LendingClub dataset.

Our research can be seen as an exploratory study within the field of applying XAI techniques and from here more focused studies can be done to further prove the value of the XAI techniques.

APPENDIX A.

Table A1. Used features LendingClub dataset

Feature	Description	Data Type
addr_state	The state provided by the borrower in the loan application	Categorical
annual_inc	The self-reported annual income provided by the borrower during registration.	Numeric
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers	Categorical
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.	Numeric
earliest_cr_line	The month the borrower's earliest reported credit line was opened	Numeric
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.	Categorical
emp_title	The job title supplied by the Borrower when applying for the loan. *	Categorical
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.	Numeric
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.	Numeric
grade	LC assigned loan grade	Categorical
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER	Categorical
id	A unique LC assigned ID for the loan listing.	Numeric
initial_list_status	The initial listing status of the loan. Possible values are – W (Whole), F (Fractional)	Categorical
installment	The monthly payment owed by the borrower if the loan originates.	Numeric
int_rate	Interest Rate on the loan	Numeric
issue_d	The month which the loan was funded	Numeric
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.	Numeric
loan_status	Current status of the loan	Categorical
mort_acc	Number of mortgage accounts.	Numeric
open_acc	The number of open credit lines in the borrower's credit file.	Numeric
pub_rec	Number of derogatory public records	Numeric
pub_rec_bankruptcies	Number of public record bankruptcies	Numeric
purpose	A category provided by the borrower for the loan request.	Categorical
revol_bal	Total credit revolving balance	Numeric
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.	Numeric
sub_grade	LC assigned loan subgrade	Categorical
term	The number of payments on the loan. Values are in months and can be either 36 or 60.	Categorical
title	The loan title provided by the borrower	Categorical
total_acc	The total number of credit lines currently in the borrower's credit file	Numeric
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified	Categorical
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.	Categorical

References

Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*, 168-178. doi:http://dx.doi.org/10.1016/j.ejor.2012.04.009

Altman, E. I. (1968). FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY. *The Journal of Finance*, 589-609.

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

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 405-417. doi:https://doi-org.ezproxy2.utwente.nl/10.1016/j.eswa.2017.04.006

Basel Committee on Banking Supervision. (2020). *Basel III Monitoring Report.* Basel Committee on Banking Supervision. Retrieved from https://www.bis.org/bcbs/publ/d500.pdf

Bhatt, U., Weller, A., & Moura, J. M. (2020). *Evaluating and Aggregating Feature-based Model Explanations.* University of Cambridge. doi:https://arxiv.org/pdf/2005.00631.pdf

Buehler, K., Freeman, A., & Hulme, R. (2008, Septermber). The New Arsenal of Risk Management. *Harvard Business Review*. Retrieved from https://hbr.org/2008/09/the-new-arsenal-of-risk-management

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. *Computational Economics*, 203-216. doi:https://doi.org/10.1007/s10614-020-10042-0

Caouette, J. B., Altman, E. I., Narayanan, P., & Nimmo, R. (2008). *Managing Credit Risk: The Great Challenge for the Global Financial Markets.* Hoboken, New Jersey: John Wiley & Sons, Inc. Retrieved from http://www.untag-smd.ac.id/files/Perpustakaan_Digital_1/CREDIT%20RISK%20Managing%20credit%20risk%20%20the%20great%20challenge%20for%20global%20financial%20markets.pdf

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*(21, 6). doi:https://doi.org/10.1186/s12864-019-6413-7

DNB & AFM. (2019, July 25). *Artificiële Intelligentie in de verzekeringssector: een verkenning.* Retrieved from AFM: https://www.afm.nl/~/profmedia/files/rapporten/2019/afm-dnb-verkenning-ai-verzekeringssector.pdf

DNB. (2019, July 25). *DNB komt met richtlijnen voor gebruik kunstmatige intelligentie*. Retrieved from DNB: https://www.dnb.nl/actueel/algemeen-nieuws/dnbulletin-2019/dnb-komt-met-richtlijnen-voor-gebruik-kunstmatige-intelligentie/#:~:text=DNB%20heeft%20een%20aantal%20algemene%20principes%20geformuleerd%20voor,van%20AI%20op%20een%20verantwoorde%20manier%20kunne

Doko, F., Kalajdziski, S., & Mishkovski, I. (2021). Credit Risk Model Based on Central Bank Credit Registry Data. *Journal of Risk and Financial Management*, 138. doi:https://doi.org/10.3390/jrfm14030138

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning*. Retrieved from https://arxiv.org/pdf/1702.08608v2.pdf

Dutton, T. (2018, June 28). *An Overview of National AI Strategies*. Retrieved from Medium: https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

European Commission. (2018, April 25). *A European approach on Artificial Intelligence*. Retrieved from European Commission: https://ec.europa.eu/commission/presscorner/detail/lv/MEMO_18_3363

European Commission. (2020). *On Artificial Intelligence - A European approach to excellence and trust.* Brussels: European Commission. Retrieved from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission. (2021, April 26). *Regulatory framework proposal on Artificial Intelligence*. Retrieved from European Commission: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Federal Trade Commission. (2012). *Fair Credit Reporting Act.* Washington, D.C.: Federal Trade Commission. Retrieved from https://www.consumer.ftc.gov/articles/pdf-0111-fair-credit-reporting-act.pdf

Freitas, A. (2014). Comprehensible classification models: A position paper. *ACM SIGKDD Explorations Newsletter*, 1-10. doi:https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1145%2F2594473.2594475

General Data Protection Regulation. (2020). *Art. 22 GDPR: Automated individual decision-making, including profiling*. Retrieved from GDPR: https://gdpr.eu/article-22-automated-individual-decision-making/

Giudici, P. (2018). Financial data science. *Statistics & Probability Letters*, 160-164. doi:https://doi.org/10.1016/j.spl.2018.02.024

Giudici, P. (2018). Financial data science. *Statistics and Probability Letters*, 160-164. doi:https://doi.org/10.1016/j.spl.2018.02.024

Gong, J., & Kim, H. (2017). RHSBoost: Improving classification performance in imbalance data. *Computational Statistics & Data Analysis*, 1-13. doi:https://doi.org/10.1016/j.csda.2017.01.005

Grand View Research. (2021). *Digital Lending Platform Market Size Report, 2021-2028.* Grand View Research.

Guidotti, R., Monreale, A., Ruggiere, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*. doi:https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1145%2F3236009

Gunning, D., & Aha, D. W. (2019, 06 24). DARPA's Explainable Artificial Intelligence Program. *AI magazine*, pp. 44-58. doi:https://doi.org/10.1609/aimag.v40i2.2850

Guo, Y., Jiang, S., Chen, F., Li, Y., & Luo, C. (2019). Borrower-lender Information Fusion for P2P Lending: A Nonparametric Approach. *International Information and Engineering Technology Association*, 269-279. doi:https://doi.org/10.18280/isi.240307

Ha, V.-S., Lu, D.-N., Choi, G. S., Nguyen, H.-N., & Yoon, B. (2019). Improving Credit Risk Prediction in Online Peer-to-Peer (P2P) Lending Using Feature selection with Deep learning. *International Conference on Advanced Communication Technology (ICACT)* (pp. 511-515). PyeongChang: IEEE. doi:https://doi.org/10.23919/ICACT.2019.8701943

High-level expert group on artificial intelligence. (2019). *Ethics guidelines for trustworthy AI.* Brussels: European Commission.

Hogue, J. (2019, January 28). Lending Club Investment REVIEW | Investing for Beginners. Retrieved from https://www.youtube.com/watch?v=nXIE3TP_4-s

Hu, F., & Li, H. (2013). A Novel Boundary Oversampling Algorithm Based on Neighborhood Rough Set Model: NRSBoundary-SMOTE. *Mathematical Problems in Engineering*. doi:https://doi.org/10.1155/2013/694809

Hull, J. C. (2018). *Risk Management and Financial Institutions, 5th Edition.* Hoboken: Wiley.

Islam, S. R., Eberle, W., & Ghafoor, S. K. (2019). *Towards Quantification of Explainability in Explainable Artificial Intelligence Methods.* Tennessee: Association for the Advancement of Artificial Intelligence. Retrieved from https://arxiv.org/pdf/1911.10104v1.pdf

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 5125-5131. doi:https://doi.org/10.1016/j.eswa.2013.03.019

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How We Analyzed the COMPAS Recidivism Algorithm.* ProPublica. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lee, T.-S., & Chen, I.-F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, 743-752. doi:https://doi-org.ezproxy2.utwente.nl/10.1016/j.eswa.2004.12.031

LendingClub . (2021, February 11). *Important Updates to the LendingClub Notes Platform*. Retrieved from LendingClub : https://help.lendingclub.com/hc/en-us/articles/360050574891-Important-Updates-to-the-LendingClub-Notes-Platform

Lessmann, S., Basesens, B., Seow, H.-V., & Lyn C. Thomas. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal ofOperationalResearch*, 124-136. doi:http://dx.doi.org/10.1016/j.ejor.2015.05.030

Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020). XGBoost Model and Its Application to Personal Credit Evaluation. *IEEE Intelligent Systems*, 52-61. doi:https://doi.org/10.1109/MIS.2020.2972533

Liang, B., Li, H., Su, M., Bian, P., Li, X., & Shi, W. (2018). Deep Text Classification Can be Fooled. *International Joint Conference on Artificial Intelligence (IJCAI-18)*, (pp. 4208-4215). doi:https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.24963%2Fijcai.2018%2F585

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777). Long Beach: Curran Associates Inc.

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 4621–4631. doi:http://dx.doi.org/10.1016/j.eswa.2015.02.001

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63(2)*, 81–97. doi:https://doi.org/10.1037/h0043158

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 1-38. doi:https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Quantifying Model Complexity via Functional Decomposition for Better Post-hoc Interpretability. In C. P., & D. K., *Machine Learning and Knowledge Discovery in Databases* (pp. 193-204). Springer, Cham. doi:http://dx.doi.org/10.1007/978-3-030-43823-4_17

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, Volume 165. doi:https://doi.org/10.1016/j.eswa.2020.113986

Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J. Y., & Ryu, K. H. (2019). An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. *Sustainability*, 699. doi:https://doi.org/10.3390/su11030699

Namvar, A., & Naderpour, M. (2018). Handling Uncertainty in Social Lending Credit Risk Prediction with a Choquet Fuzzy Integral Model. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-8). Rio de Janeiro: IEEE. doi:https://doi.org/10.1109/FUZZ-IEEE.2018.8491600

Namvar, A., Siami, M., Rabhi, F., & Naderpour, M. (2018). Credit risk prediction in an imbalanced social lending environment. *International Journal of Computational Intelligence Systems*, 925-935. Retrieved from https://export.arxiv.org/ftp/arxiv/papers/1805/1805.00801.pdf

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 3028–3033. doi:10.1016/j.eswa.2008.01.018

Nassar, M., Salah, K., Rehman, M., & Svetinovic, D. (2019). Blockchain for explainable and trustworthy artificial intelligence. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*. doi:https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1002%2Fwidm.1340

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 109-131. doi:https://doi.org/10.2307/2490395

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Klamargias, A. (2018). *A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting.* Basel: Bank for International Settlements. Retrieved from https://www.bis.org/ifc/publ/ifcb49_49.pdf

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). San Francisco: Association for Computing Machinery. doi:https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Model-Agnostic Interpretability of Machine Learning.* New York: 6 ICML Workshop on Human Interpretability in Machine.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/11491

Robnik-Šikonja, M., & Bohanec, M. (2018). Perturbation-Based Explanations of Prediction Models. In J. Zhou, & F. Chen, *Human and Machine Learning* (pp. 159-175). Springer International Publishing AG. doi:https://doi.org/10.1007/978-3-319-90403-0_9

Rose, J. (2018, March 23). My Lending Club Investment Review (what I did wrong 😭). Retrieved from https://www.youtube.com/watch?v=zpAi9euMCJE

Rose, J. (2020, July 29). *Lending Club Reviews For Investors And Borrowers*. Retrieved from Good Financial Cents: https://www.goodfinancialcents.com/lending-club-review-for-investors-and-borrowers/

Scholbeck, C. A., Molnar, C., Heumann, C., Bischl, B., & Casalicchio, G. (2020). Sampling, Intervention, Prediction, Aggregation: A Generalized Framework for Model-Agnostic Interpretations. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, Cham. doi:10.1007/978-3-030-43823-4_18

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX - From Local to Global Explanations of Black Box AI Models. *Artificial Intelligence*. doi:https://doi.org/10.1016/j.artint.2021.103457

Shim, Y. (2019). *Five key trends illuminating AI's impact for financial services.* EY. Retrieved from https://www.ey.com/en_gl/innovation-financial-services/five-key-trends-illuminating-ai-s-impact-for-financial-services

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable Deep Learning Models in Medical Image Analysis. *Deep Learning in Medical Image Analysis*. doi:https://doi.org/10.3390/jimaging6060052

Song, Y., Wang, Y., Ye, X., Wang, D., Yin, Y., & Wang, Y. (2020). Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk

assessment in P2P lending. *Information Science*, 182-204. doi:https://doi.org/10.1016/j.ins.2020.03.027

van Beusekom, H. (2019, February 27). *Ethics in AI, a way to avoid regulation?* Retrieved from AFM: https://www.afm.nl/nl-nl/nieuws/2019/feb/hanzo-blog-fintech

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baessens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 211-229. doi:https://doi.org/10.1016/j.ejor.2011.09.031

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning —— a case study of bank loan data. *Procedia Computer Science*, 141-149. doi:https://doi.org/10.1016/j.procs.2020.06.069

Zhang, W., Wang, C., Zhang, Y., & Wang, J. (2020). Credit risk evaluation model with textual features from loan descriptions for P2P lending. *Electronic Commerce Research and Applications*. doi:https://doi.org/10.1016/j.elerap.2020.100989

Zhou, G., Zhang, Y., & Luo, S. (2018). P2P Network Lending, Loss Given Default and Credit Risks. *Sustainability*, 1010. doi:https://doi.org/10.3390/su10041010

Ziegler, T., Shneor, R., Wenzlaff, K., Wang, B. W., Kim, J., Odorovic, A., . . . Luo, D. (2020). *The Global Alternative Finance Market Benchmarking Report.* Cambridge: Cambridge Centre for Alternative Finance. Retrieved from https://www.jbs.cam.ac.uk/faculty-research/centres/alternative-finance/publications/the-global-alternative-finance-market-benchmarking-report/