DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2021

Comparison of Logistic Regression and an Explained Random Forest in the Domain of Creditworthiness Assessment

MARCUS ANKARÄNG

JAKOB KRISTIANSSON

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Comparison of Logistic Regression and an Explained Random Forest in the Domain of Creditworthiness Assessment

Marcus Ankaräng KTH, Jakob Kristiansson KTH

***Abstract*—As the use of AI in society is developing, the requirement of explainable algorithms has increased. A challenge with many modern machine learning algorithms is that they, due to their often complex structures, lack the ability to produce human-interpretable explanations. Research within explainable AI has resulted in methods that can be applied on top of non-interpretable models to motivate their decision bases. The aim of this thesis is to compare an unexplained machine learning model used in combination with an explanatory method, and a model that is explainable through its inherent structure. Random forest was the unexplained model in question and the explanatory method was SHAP. The explainable model was logistic regression, which is explanatory through its feature weights. The comparison was conducted within the area of creditworthiness and was based on predictive performance and explainability. Furthermore, the thesis intends to use these models to investigate what characterizes loan applicants who are likely to default. The comparison showed that no model performed significantly better than the other in terms of predictive performance. Characteristics of bad loan applicants differed between the two algorithms. Three important aspects were the applicant's age, where they lived and whether they had a residential phone. Regarding explainability, several advantages with SHAP were observed. With SHAP, explanations on both a local and a global level can be produced. Also, SHAP offers a way to take advantage of the high performance in many modern machine learning algorithms, and at the same time fulfil today's increased requirement of transparency.**

***Sammanfattning*—I takt med att AI används allt oftare för att fatta beslut i samhället, har kravet på förklarbarhet ökat. En utmaning med flera moderna maskininlärningsmodeller är att de, på grund av sina komplexa strukturer, sällan ger tillgång till mänskligt förståeliga motiveringar. Forskning inom förklarbar AI har lett fram till metoder som kan appliceras ovanpå icke-förklarbara modeller för att tolka deras beslutsgrunder. Det här arbetet syftar till att jämföra en icke-förklarbar maskininlärningsmodell i kombination med en förklaringsmetod, och en modell som är förklarbar genom sin struktur. Den icke-förklarbara modellen var random forest och förklaringsmetoden som användes var SHAP. Den förklarbara modellen var logistisk regression, som är förklarande genom sina vikter. Jämförelsen utfördes inom området kreditvärdighet och grundades i prediktiv prestanda och förklarbarhet. Vidare användes dessa modeller för att undersöka vilka egenskaper som var kännetecknande för låntagare som inte förväntades kunna betala tillbaka sitt lån. Jämförelsen visade att ingen av de båda metoderna presterande signifikant mycket bättre än den andra sett till prediktiv prestanda. Kännetecknande särdrag för dåliga låntagare skiljde sig åt mellan metoderna. Tre viktiga aspekter var låntagarens ålder, vart denna bodde och huruvida personen ägde en hemtelefon. Gällande förklarbarheten framträdde flera fördelar med SHAP, däribland möjligheten att kunna producera både lokala och globala förklaringar. Vidare konstaterades att SHAP gör det möjligt att dra fördel av den höga prestandan som många moderna maskininlärningsmetoder uppvisar och samtidigt uppfylla dagens ökade krav på transparens.**

***Index Terms*—Classification, Creditworthiness, Explainable Artificial Intelligence, Logistic Regression, Machine Learning, Random Forest, SHAP, XAI**

I. Introduction

THE development of modern machine learning has resulted in high performing models that have the potential to outperform more traditional methods. However, this increased performance often comes at the cost of lowered interpretability. The need for explanations of predictions are increasing. In some cases, a reliable model that produces accurate predictions is sufficient to solve a machine learning problem. In other cases, not only is it important to achieve high accuracy, but also to be able to explain how the model came to the conclusion. A field where explanations are of great importance, both because it is believed to increase user experience and also because it is required and regulated by law, is assessment of creditworthiness. Traditionally, statistical methods have been the most prominent ones in this area. However, there is an increasing interest being shown towards machine learning methods.

Without any complementary methods, the explainability of a machine learning model is dependent on the type of algorithm used. While some algorithms, like logistic regression through its feature weights, are interpretable by their inherent structure, other algorithms do not disclose any human-interpretable explanation for how the prediction was made. This lack of interpretation is largely due to the complex architectures and training methods of the algorithms. Uninterpretable models that do not uncover any human-interpretable relationship between the input variables and the predictions are referred to as black-box models [1].

Recent research has investigated and developed techniques that can be applied on top of black-box models in order to interpret underlying information about their predictions. These techniques have gained great attention since they offer a way to build high performing models that also can fulfill the need for interpretability. However, for managers the implementation of these new algorithms may not be straightforward and the implication of the algorithms on the business is relatively unknown. The aim with this paper is to investigate how explainable AI, more specifically SHAP, can be used to explain a non-explainable machine learning algorithm. Additionally, the result is compared in terms of both predictive performance and explainability with a model that is explainable through its inherent structure. After comparing the algorithms, they are also analyzed from a business perspective on their feasibility and practical implementation. The comparison is done in the field of creditworthiness.

II. Background and objectives

A. Objectives

The problem examined in this paper is threefold and formulated as follows:

1) *Which of the methods, logistic regression or random forest used in combination with SHAP, is more suitable in terms of predictive performance and explainability, to use when assessing loan applicants' creditworthiness?*
2) *Using the methods described in question one, what characterizes loan applicants that are likely to default?*
3) *How can management use the results of question one and two when implementing modern machine learning algorithms and explainable AI?*

In more detail, the first part in the study will be to establish the classifiers. Firstly, random forest is compared with logistic regression in terms of predictive performance. Secondly, an explainable AI model will be implemented on top of the random forest algorithm in order to explain its decisions. The models are then compared in terms of explainability. By comparing logistic regression and the combination of random forest and SHAP in terms of predictive performance and explainability, the findings are hoped to give further understanding on how to fulfil two of the most important requirements in assessment of creditworthiness, namely accuracy and trust.

In the second part of the study, the methods are used to identify characteristics of loan applicants that are likely to default.

For the last part of the study the results of the previous research questions and an extensive literature study are compiled and analyzed from a business perspective. Here, advantages and disadvantages of each algorithm in terms of feasibility and performance for managers are discussed.

B. Creditworthiness

A central part in today's economy is the ability for people and organizations to lend money by contractual agreements where the borrower assures that the borrowed sum and generally also an interest is repaid to the lender in the future. Providing credit to consumers is a worldwide industry and many firms even use the concept to increase their sales. However, if anyone applying for a loan would be granted one, the risk of customers not repaying the loan would increase and credit companies would most likely not be profitable. Assessment of creditworthiness is used by many companies, primarily financial ones, to mitigate risk in their business. The importance of creditworthiness assessment is noted by the EU, who has stipulated laws in the area. Following is an excerpt from the consumer credit directive (CCD) showing the importance of creditworthiness assessment: *"Member States shall ensure that, before the conclusion of the credit agreement, the creditor assesses the consumer's creditworthiness on the basis of sufficient information, where appropriate obtained from the consumer and, where necessary, on the basis of a consultation of the relevant database"* [2]. The background checks can be done in varying degrees of extensiveness and are usually performed by a third party.

C. Risk management

A central part in any company is to reduce their risk while improving profits. For financial institutions, the risk of over-indebtedness is especially important to reduce [2]. Irresponsible lending can be one of the reasons for over-indebtness. There are primarily two methods for this sector to reduce its risk, either through direct or indirect methods. Direct measures are either proactive, such is the case with creditworthiness assessment, or ex post, which could be done by creating payment plans. Indirect measures consist of risk management and mechanisms adopted by the financial institutions. With high requirements of risk stability in these companies, any change to risk management must be carefully planned. Logistic regression has been the staple of the creditworthiness industry for 70 years [3]. As modern machine learning algorithms have shown promising results, the need for further research in both algorithm performance and risk mitigation in implementation is evident.

D. Relevance

The results will be of interest to managers and machine learning modelers who work in environments where transparency is of great importance. As technologies such as machine learning are becoming more prevalent in decision making, especially in departments where regulations may restrict the usage of these algorithms, this area of research requires attention. One such field is the area of creditworthiness assessment. The dataset used in this study is further detailed in the Method section.

E. Ethical aspects

The data processed within creditworthiness is inherently sensitive as it details payment history and personal information. Although it is anonymized, it must still be processed in a responsible manner to ensure that it follows both legal and ethical rules. Legislation towards harmful AI is being worked on by the EU. The following right can be found in the European GDPR regulation: *"...the existence of automated decision-making should carry meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject"* [4]. Not only does the regulation specify when explanations should be given, it also informs about its purpose.

With a black-box model it is harder to understand and examine if the used machine learning model contains biases or other flaws that result in a discriminating behaviour. By extending the research area of explainable AI, the results will indirectly contribute to the global goals of reducing inequality. Implementation of a new AI algorithm will by itself introduce ethical issues as the explaining algorithms also must conform local laws and ethical rules by providing objective information.

III. Theory

A. Explainability and interpretability

Two common terms in the area of explainable AI are explainability and interpretability. Although they are similar, a clear distinction must be made. Explainability requires that the model can explain its reasoning. It should be able to gain the trust of the user on why a decision was made. Interpretability does not have as strict requirements as the algorithm only requires that it presents summarized information or visual queues on why the decision was made. While an explainable algorithm is always interpretable, the opposite does not necessarily have to be true [5].

B. Recipients of explanations

Explainability, which is seen as a mainly subjective measure, must take into consideration the recipient of the information when evaluating its effectiveness. In the context of creditworthiness, there are primarily three different types of recipients that each require different explanations [6]:

1) *Loan officers* — require explanations that are specific for a single user.
2) *Loan applicants (primarily the rejected ones)* — require an explanation for their application result.
3) *Regulators and data scientists* — require an explanation on the entire dataset to identify general trends, discrepancies and biases.

C. Levels of explanation

When providing explanations, a distinction can be made between global explanation methods and local explanation methods. The former intends to understand dependencies between the independent and dependent variables on the whole dataset. The latter refers to the understanding of specific decisions by looking at how the model would predict in a nearby subregion to the inputted data point [7].

D. Model agnostic explanations

Techniques used to make algorithms interpretable can either be model-specific or model-agnostic. While model-specific methods are heavily connected with the model used, agnostic methods are not bound to one single method or category of methods. An example of a model-agnostic method is SHAP. Model-agnostic methods utilize the inputs and predictions of the machine learning model to achieve interpretability [1].

E. Logistic Regression

Logistic regression provides a method for calculating the probability that a given input X belongs to a certain class Y [8]. In the binary case, the method is used with two classes, a positive and a negative, and in the multinomial case, the classes can be more than two. Binary logistic regression is used in this study. In classification, the probability $P(y = 1|x)$, meaning the probability that the given input x belongs to the positive class, is wanted. Many algorithms can solve these problems but unique for logistic regression is that it learns a set of feature weights along with a bias term. Each weight indicates the size and direction of the impact that a certain input feature has on the prediction. Logistic regression uses the logistic function, also called the sigmoid function, for calculating probabilities.

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}} \tag{1}$$

This function, $p(X)$, maps real-valued numbers in the form of features in X to a value in the range [0, 1]. The coefficients $\beta_0, \beta_1, ..., \beta_n$ determine the impact of feature values in terms of both size and direction. Output values are probabilities where $p(X) > \theta$ means that X belongs to the positive class and $p(X) <= \theta$ means that X belongs to the negative class. Here, θ is used as a notion for the threshold. The threshold can be set to 0.5 or any other value in the range (0, 1).

Minimizing the error of the predictions is done with the cross entropy loss function, which describes how much $p(X)$ differs from the true value y. The sigmoid function is denoted by σ in the cross entropy loss function below.

$$L(y, \sigma) = -[y \times log(\sigma) + (1-y) \times log(\sigma)] \tag{2}$$

F. Random forest

Random forest is an algorithm based on multiple decision trees that can be used for both regression and classification tasks [9]. By utilizing the low correlation across multiple trees, noise in data will have a low impact as it will statistically equalize itself among the trees. Decision rules for each tree are built top-down by considering a random sample of the original dataset as training data. As having individual training sets for each tree is rare, random sampling by bootstrapping is used for providing training data. Individual trees are built deep with high variance but low bias. From a subset of predictors in the training set, the binary split that leads to the lowest total variance is chosen as a decision rule. To find this split, the Gini index is commonly used:

$$I_g(p) = \sum_{i=1}^{K} p_i(1-p_i) \tag{3}$$

Binary splitting is repeated until the stop criterion is satisfied. The terminal nodes in the trees contain a predicted value. Decision trees form the basis of the random forest algorithm. Classification of an unseen data point is done by starting at the root node and following the branches down to a terminal node according to the decision rules. This is done through all trees and based on the most common prediction a classification can be made. Figure 1 displays an example of a basic decision tree.

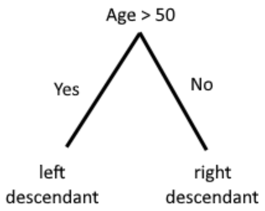


Fig. 1: Example of a decision tree

G. SHAP

SHAP (SHapley Additive exPlanations) was first proposed by Lundberg and Lee in 2016 [10]. The technique can be used to find the effect of a single feature on the prediction in terms of both magnitude and direction. SHAP is an extension of Shapley values, which can be summarized as the average contribution of a feature value over all combinations of feature set-ups [10]. Some distinctions between the original Shapley values and the more extensive SHAP method, are that SHAP provides several methods for producing explanations on a global level, and that it comes with two additional estimation approaches; KernelSHAP and TreeSHAP. The second one of these, TreeSHAP, is an approach for estimating tree-based models efficiently.

In more detail, the Shapley values determine, for each prediction and feature, how the prediction changes when that feature is dismissed from the model [10]. The effect of removing the feature is simulated by replacing it with the feature value from a randomly selected data point. A new prediction is made with this fictional data point. The difference between this probability and the expected value over the entire dataset is used as an estimation of the contribution of that feature. Estimations are improved by repeating the sampling step and averaging the result. The method follows coalition game theory, where each feature value is seen as a player who is evaluated based on its contribution to the total score, which in this case is the difference between the predicted probability and the overall expected value of the probability.

What is also characteristic about SHAP is that it represents the Shapley values by a linear model:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{4}$$

Here, g represents the explanation model, z the feature values and ϕ the Shapley values [10].

H. ROC curve and AUC

A ROC (Receiver Operating Characteristic) curve illustrates how well a classification model performs with different thresholds [11]. In a binary classification problem, a threshold indicates the probability score for which an entry is classified as positive. The ROC curve plots the rate of true positives and false positives, which are defined as follows:

$$True\,positive\,rate\,(TPR) = \frac{TP}{TP+FN} \tag{5}$$

$$False\,positive\,rate\,(FPR) = \frac{FP}{TP+FN} \tag{6}$$

AUC (Area Under the ROC Curve) gives an overall performance measure over all threshold levels. A benefit of using AUC as a measure in a classification problem, is that it focuses on the relative probabilities of instances rather than on their actual classification. Hereby, the AUC will increase if positive instances are assigned higher probabilities than negative instances, regardless of the threshold. AUC is therefore said to be classification-threshold-invariant [11].

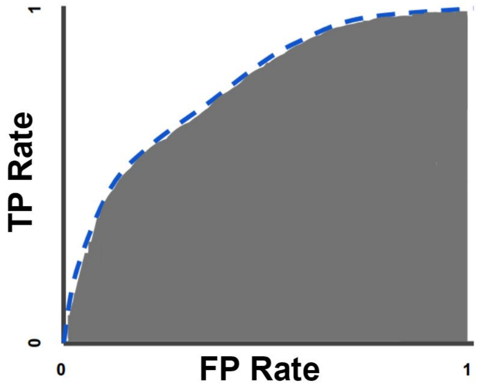


Fig. 2: Example of a ROC curve. AUC is marked in grey. The picture is taken from one of Google's machine learning courses [12].

I. Cross-validation

Cross-validation enables training and testing with all data and can be used to examine how well a model performs on unseen data. This can be done by randomly dividing the data into training and test sets. After having tested the model with one test set, a new set of training and test data is used. The process where this procedure is repeated k times and the measure of error is averaged is called k-fold cross validation. This technique is advantageous for avoiding overfitting and for producing a more reliable performance measure [13].

J. Product life cycle

The product life cycle in Figure 3 shows the different stages that a traditional product goes through during its time on the market. During the development phase, both the income and growth is low as focus is on improving the product [14]. The next phase, introduction, is the most critical part of the life cycle as the product needs to cross the chasm. By crossing the chasm the product goes from being a niche product to a mainstream product with increasing growth as a result. During the growth stage and early part of the maturity stage, the growth and the generated income is high. Towards the end of maturity the growth slows and and turns into a decline. In this phase, the product has served its purpose and been replaced by newer products.

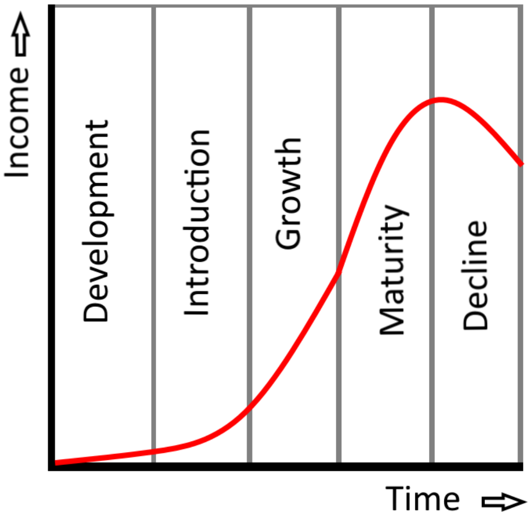


Fig. 3: A traditional view of the product life cycle.

IV. Previous Research

Even though logistic regression traditionally has been one of the most used methods in creditworthiness assessment, Lessman et al. highlighted that modern machine learning algorithms potentially can replace this method [15]. By conducting an extensive literature study involving a multitude of datasets and 41 different algorithms, the authors concluded that ensemble classifiers is preferred over single classifiers and that models based on deep learning generally perform better than traditional methods. However, they could not explain what made these algorithms function better. Furthermore, they researched if a higher accuracy equals higher profit. In their paper they found some evidence supporting this relationship. They also highlighted that data quality and feature selection are important aspects of acquiring high accuracy. Since they considered logistic regression to be outdated in performance, they caution against comparisons with only logistic regression as a benchmark.

Linhart et al. applied a few of the methods evaluated in the paper of Lessman et al. in a credit scoring contest hosted by PAKDD in 2009 [16]. In their paper, which is built around their contribution to the competition, they tested different algorithms including generalized boosting models, KNN and a combined model that used both logistic regression and KNN. In contrast to the findings of Lessman et al. this paper did not achieve good results with ensemble classifiers. The authors do however suggest that the poor performance of the generalized boosting models may be influenced by their lack of knowledge about these algorithms. The best performing method was the combined model of logistic regression and KNN, with which they achieved an AUC score of 0.677 and placed 3rd in the competition. The dataset used in the competition was similar to the one that will be used in this paper. Worth noting is that the AUC score reached by Linhart et al. is relatively low compared to what have been achieved in other papers that used different datasets. He et al. compared nine different machine learning models, including random forest and logistic regression, on six different datasets [17]. Overall, random forest performed just above logistic regression. However, they achieved AUC scores of over 0.90 with both methods on some of the datasets.

AI has within the last decade shaped decision making within many company sectors, but according to Dosilovic et al. many of these decisions are based on black-box models that lack transparency [18]. This happens despite there being evident ethical implications in using AI in real-life scenarios. By using explainable AI, an abstracted explanation for finding, checking and reasoning over useful properties is offered. The author furthermore claims that the technology can be used not only for verifying trust criteria, but also be helpful for further research. Similar to Dosilovic et al., Bussman et al. describe the issue of overcoming black-box models with explainable AI algorithms, as black-box models by themselves are unsuitable for regulated finance services [4]. The authors continue with implementing a SHAP explainer on top of an XGBoost classifier that they then compare to the result of a traditional logistic regression classifier. The conclusion of the paper shows that explainable AI algorithms give further explainability to decisions made by black-box models. Furthermore, it is highlighted that these advantages are especially noticeable in areas that are regulated. However, some issues are brought up in the form of understanding algorithm results and risks that arise when implementing new algorithms.

Within a similar field of research as the paper of Bussman et al., Olofsson explores the use of local explainable AI algorithms in the context of churn prediction for a finance app [19]. In her paper, she compares random forest with a number of different algorithms. The paper shows that the advantage of being able to explain a black-box model is that it gives insight into model bias and data imbalances. In her study, the method used to produce explanations was local interpretable model-agnostic explanations (LIME). Furthermore, she mentions that a lot of research has been made on global explanation algorithms but not as much on local. Zihni et al. also compares explainable machine learning algorithms to traditional logistic regression but in a medical domain [20]. The study compared a tree boosting algorithm and a neural network with traditional logistic regression models. Deep Taylor decomposition was used to explain the neural network and SHAP was used to explain the tree boosting algorithm. For the comparison, they compared both the performance and explainability of the models. They found that modern machine learning algorithms with explainable AI algorithms can provide explainability that is comparable with traditional method rankings. Furthermore, they suggest that further research should be conducted on more datasets within the area.

While several studies have compared classifiers based on their predictive performance, not as much research have been conducted that involves aspects of explainability in these models, especially not in assessment of creditworthiness. Also, many of the articles that have been produced does not take into account that there are different types of recipients to consider when explaining decisions. Hereby, this thesis intends to extend previous research on creditworthiness assessment by also including explainability as a perspective.

V. Method

A. Data

The dataset used in the study was collected by a Brazilian credit company between the years 2006–2009 and contained originally 50,000 entries and 54 features, including the target label [21]. The dataset was composed as part of the PAKDD 2010 Data Mining Competition. The objective in the competition was to classify loan applicants as either good or bad. A good applicant is someone who is likely to repay the loan and a bad one is someone who can not. In the data, which is historical, a bad loan applicant is defined as someone who failed to pay an invoice within 60 days over a one year period. Good applicants were represented by the number zero and bad applicants by the number one in the dataset. Inflation during the time of data collection was stable. Worth noting is that the data solely included the company's clients, meaning that rejected applicants were excluded.

B. Data preparation

The actions performed to clean the data varied depending on the type of variable and whether it was categorical or numerical. A difficulty with the dataset was that erroneous values had been encoded inconsistently as either empty cells, $NULL's$, $NaN's$ or sequences of $X's$ with varying length. Since the dataset description did not provide any further information about the different approaches, all variations were interpreted as missing values. A full list of all the original features can be found in Appendix A.

1) Missing values: In both numerical and categorical features, the number of missing values was evaluated. If more than 50% of the values in a feature was missing, that feature was removed. This was the case in for example the features for personal assets value and months in the job. In terms of numerical features, missing values were replaced with the mean over all values in that column. Missing values in categorical features were encoded as a specific category.

2) Correlating features: If two or more features showed strong correlation with each other, only one of the features was kept in order to reduce dimensionality.

3) Small sample sizes: Categories with few occurrences were grouped together and encoded as a specific category. As a rule, categories with fewer than 100 occurrences were merged together. Different values for the minimum occurrences in categories were tested, although no significant improvement in performance could be observed.

4) Encoding of categorical features: Categorical features with k distinct values were encoded into $k-1$ new categories using one-hot encoding. The same information in a categorical feature with k different categories can be represented by $k-1$ categories, why perfect collinearity would be achieved if not one category was dropped. Given a vector X, representing a feature with k distinct categories, $X = (x_1, ..., x_k)$, it can be divided into k binary features $X_1, ..., X_k$. Any arbitrary category X_n can then be represented by giving all other categories the value zero. This poses a problem for linear models, such as logistic regression, as it is not possible to separate the individual effects on a change in a variable when they change together.

5) Feature scaling: As the features of the dataset were of different scales, normalization was performed for the numerical features. The motivation behind normalizing the features was that different scales would affect the size of the feature weights in the logistic regression classifier and make the identification and comparison of important features more difficult. Scaling the features was done using MinMaxScaler from sklearn [22]. The method rescaled all values into the range [0, 1].

6) Other considerations: Several features in the dataset had identical values throughout the dataset, those features were removed. Also, a number of features in the dataset contained strings inputted by the loan applicants. All these features were geographical, either in terms of where the applicant lived or where the applicant worked. As these features contained many misspellings and several different values for the same area and since the rate of missing values in these was high, they were removed. Their exclusion was also motivated by the fact that the dataset contained other, more complete geographical features, such as zip codes. Instances containing outlier values and unreasonable values were also removed from the dataset. Example of such a value was an unreasonably young age.

After dropping columns that had too much missing data, overlapped or otherwise did not contribute to the prediction, 24 columns, excluding the target label, remained. These consisted of 5 numerical and 19 categorical features. The ratio of categorical variables was similar to that in the original dataset, before any data processing had been done. The features that remained after the data had been processed were the following:

1) User attributes
   a) **Categorical**: Sex, Marital status, Nationality, Flag residential phone, Residence type, Flag email, Flag professional phone, Profession code, Occupation type
   b) **Numeric**: Age, Quantity dependants, Months in residence, Quantity cars
2) Financial attributes
   a) **Categorical**: Flag visa, Flag mastercard, Flag other cards
   b) **Numeric**: Quantity bank accounts
3) Geographical attributes
   a) **Categorical**: Residential zip code, State of birth, Residential state
   b) **Numeric**: None
4) Application
   a) **Categorical**: Payment day, Application submission type, Supplied company name, Product type
   b) **Numeric**: None

C. Balancing data

Within the context of credit scoring, defaulting customers are usually less common than customers who can repay their loan, meaning that datasets are often imbalanced [23]. In this dataset, 26% of the applicants were classified as bad loan applicants, thus, the data was imbalanced. Training a model with an imbalanced dataset is difficult since it often results in overfitting on the majority class [24]. Without any technique for handling the uneven distribution, this issue was apparent when testing the logistic regression and random forest classifiers on unseen data. Two different approaches commonly used to handle imbalanced data are down-sampling and up-sampling [23]. Both of these methods were tested. For down-sampling, random samples from the majority class were removed until both classes were equally present. For up-sampling, synthetic minority oversampling technique (SMOTE) was tested. This technique extends datapoints in the minority class by considering their nearest neighbours [25]. However, up-sampling with SMOTE did not resolve the issue of overfitting in this study, why undersampling was used instead.

D. Logistic Regression

A logistic regression classifier was implemented using the library sklearn [26]. The solver used to optimize the model and thereby minimize the loss in the cost function was the default option set by sklearn, L-BFGS. Regularization, as implemented by default in the model, was used to penalize large class weights and thereby prevent overfitting. The maximum amount of iterations was increased from the default amount of 100 to 8000 in order for the optimization to converge.

After the model had been trained, its feature weights were extracted and sorted by their absolute values. Positive coefficients indicate a decrease in creditworthiness and negative indicate an increase, why they were sorted by their absolute values. This would explain what features had the most impact on the decisions made by the model.

E. Random Forest

The random forest model was also implemented using the library sklearn [27]. The model can be modified through a number of hyperparameters, two of those were tuned. The first parameter considered was $n\_estimators$, which decides the number of trees in the forest, and the second parameter was max_depth, which determines the maximum depth of the trees. These hyperparameters were optimized by following a simple strategy where the performance of the model was evaluated while one parameter at a time varied over a range of values and the rest of the model was kept constant. When tuning the hyperparameters, the high complexity in the SHAP explainer also had to be considered. Due to limited computing power, the random forest classifier was constrained to use a max depth of 24 and 600 estimators. The high complexity in SHAP is discussed in more detail in the following section.

F. SHAP

The explainable AI method implemented to interpret the random forest classifier was SHAP. In more detail, TreeSHAP was used [28]. Initialization of the explainer required the trained classifier as an argument. During the computations of the Shapley values one clear drawback, namely the complexity of the method, became obvious. As the depth of the trees increased, the time required to compute the Shapley values increased exponentially. More specifically, the complexity of the model is $\mathcal{O}(TLD^2)$, where T, L and D denotes the number of trees, the total number of leaves in a tree and the maximum depth of a tree respectively [10]. Hereby, too large hyperparameter values in random forest led to overly long computation times.

Using SHAP, two different explanations were implemented. The first one was a purely local explanation, where the Shapley values of the 10 most important features for a single instance was displayed in a waterfall plot. This instance was chosen randomly to avoid any selection bias. The second explanation implemented was a summary plot of Shapley values for all data points used to train the random forest model. The motivation for producing this plot was to explain the decisions of the classifier in a global aspect. This explanation would be comparable to the feature importances extracted from the logistic regression classifier.

G. Cross Validation

For evaluation of the predictive performance, 5-fold cross validation was implemented using the library sklearn [29]. In order to preserve the original ratio of negative and positive samples also in the test set, a stratified 5-fold cross validator was used. Since the ratio of bad loan applicants in the overall dataset was 26%, the test set also contained this percentage of bad loan applicants. This would make sure that each class was represented in each fold. Five folds were used for validation. Hereby, 80% of the samples in the dataset was used for training and the rest were set aside for testing the models in each round. The average AUC score from all folds was calculated and used as a measure of the performance of the models.

H. Faithfulness

Despite explainability being a partially subjective measure, it can still be evaluated in quantitative measures. Displaying correct information is a key part in creating user trust. Therefore the measure faithfulness, which examines if the features that are considered most important are the ones that affect the prediction the most, was implemented. For predictions made by random forest, the 10 most important features of a random instance were changed one by one in order to simulate them being dropped. Numerical feature values were changed to the mean over that column and binary categorical feature values were inverted. After modifying a feature, the change in prediction probability was noted. The feature value was changed back to its original value before the next feature was changed. According to the explainability algorithm these features should be the most contributing features for classification and as such the uncertainty should go up when modifying them. To ensure that the results were representative for the dataset, 100 applicants were tested and the results were averaged. In logistic regression, the features with the largest weights will by definition affect the prediction the most, why this test would not provide any new insights to this model. Therefore, this test was only conducted on the combination of random forest and SHAP.

I. *Literature study*

For the third objective in the thesis, a study of relevant literature was conducted in order to investigate how management can integrate explainable AI into a system for assessment of creditworthiness. Advantages and disadvantages in the perspective of companies and end users were considered. Also, findings from the literature study were combined with the results obtained in the experiments of the thesis.

VI. Results

A. *Predictive Performance*

One of the motives with this thesis was to compare a non-explainable machine learning model with a directly explainable classifier in prediction of creditworthiness. The non-explainable algorithm used was random forest and the explainable model used was logistic regression.

TABLE I: Predictive performance

Algorithm	AUC	Standard deviation
Logistic regression	0.6296	0.0032
Random forest	0.6325	0.0023

Table I shows the average AUC after 5-fold cross validation. Results indicated that the predictive performance of logistic regression and random forest was similar. Standard deviation was low with both methods, indicating that the AUC score did not fluctuate through the folds.

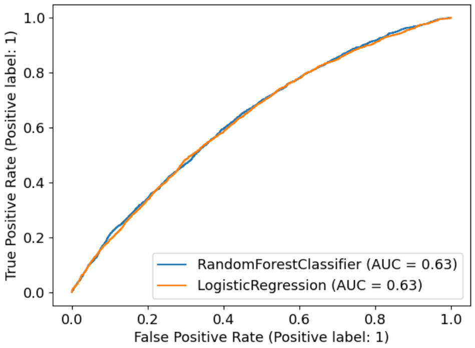


Fig. 4: ROC curves produced with logistic regression and random forest in one of the folds

One pair of ROC curves produced with logistic regression and random forest is displayed in Figure 4. For visual purposes, the plot only includes curves from one of the folds. However, it was noted that the shape of the curves was similar between the two models throughout all folds.

B. *Explainability*

1) Logistic Regression: The feature weights extracted from the trained logistic regression classifier are global explanations as they are computed with respect to the whole dataset and are describing a measure of how much an increase of a feature would change a prediction. A positive feature weight indicates that an increase of the feature value will increase the probability that the applicant is bad. In contrast, a negative feature weight indicates the opposite.

TABLE II: Logistic regression coefficients

Feature	Feature coefficient
AGE	-2.0587
RESIDENCIAL_ZIP_3_402	1.7907
RESIDENCIAL_ZIP_3_912	1.4588
RESIDENCIAL_ZIP_3_788	-1.3634
FLAG_RESIDENCIAL_PHONE	-1.2009
RESIDENCIAL_ZIP_3_917	1.1334
RESIDENCIAL_ZIP_3_689	-1.1018
RESIDENCIAL_ZIP_3_535	1.0431
RESIDENCIAL_ZIP_3_114	-1.0316
RESIDENCIAL_ZIP_3_454	-1.0206

Table II displays the 10 feature weights with the highest absolute values in the trained logistic regression model. A more extensive list of feature weights can be found in Appendix B. The most important feature according to the model was the age. The negative coefficient indicates that an increase in age lowers the probability that the applicant is bad. Noteworthy is that residential zip codes are eight of the most important features. While some zip code areas tend to increase the probability that the loan applicant is bad, other areas result in a decrease. Apart from zip codes, the feature determining whether the applicant has a residential phone was important. Having a residential phone decreased the probability of being a bad loan applicant.

2) SHAP: Unlike logistic regression, SHAP has the ability to give local explanations. As local explanations are unique to the inputted instance, the differences between local explanations can vary greatly between predictions. Presented is an example of an explained decision:

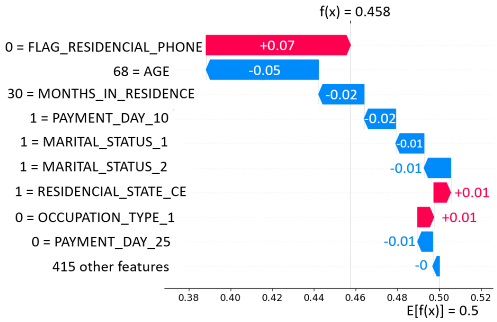


Fig. 5: Local explanation of an applicant with SHAP

Figure 5 highlights the 10 most important features of an application. The red boxes indicate a negative effect on the applicant's creditworthiness with not having a residential phone having the largest impact. Positive effects are indicated by the blue boxes with having an age of 68 affecting the classification the most among the positive features. $F(x)$ describes the probability of the applicant being bad. Since the value is less than 0.5, this instance is classified as a good applicant. $E[f(x)]$

is the expected value of the probability in regards to all training data. Since the training data was balanced, the expected value was 0.5.

In order to produce a global explanation of the random forest classifier, a summary plot of Shapley values for a larger sample of instances and their features was constructed, see Figure 6. In the graph, the 20 most important features are displayed. The colored regions are made up of many pixels, each one representing a Shapley value for an instance and its feature value. Pixels that represent the same Shapley value are clustered and aligned vertically, this gives an indication of how the Shapley values are distributed. As described in the color scale at the right hand side of the figure, red pixels represent high feature values and blue pixels low feature values. Hereby, binary features have only two distinct colors; red, which indicates when the feature has the value one and blue, which indicates when the feature has the value zero [10].

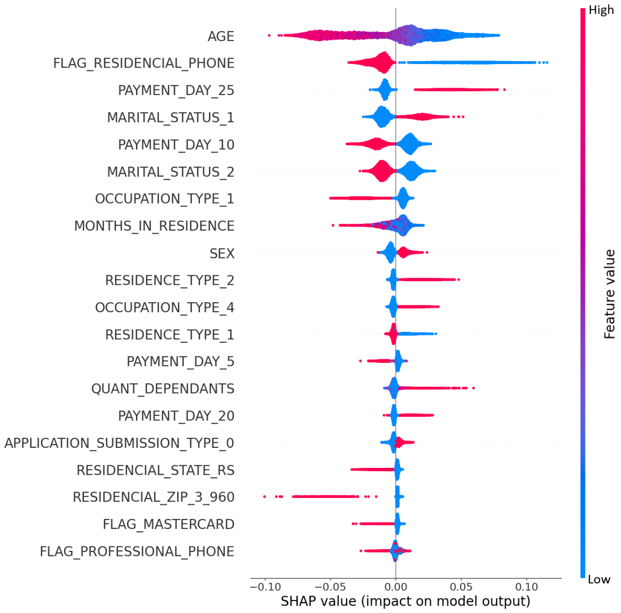


Fig. 6: Global explanation of the dataset with SHAP. A higher resolution version can be found in Appendix C.

According to the plot in Figure 6, the most important feature in the trained random forest classifier is the age of the applicant. Since lower feature values, meaning younger ages, are more often associated with positive Shapley values, this figure explains that younger people are more prone to be bad loan applicants and that the impact of the age is great. Whether the applicant has a residential phone or not was the second most important feature. In this case, where the feature is binary, the blue pixels represents not having a residential phone and is more associated with higher Shapley values and therefore also bad loan applicants. Following is a feature indicating that a payment of the applicant was delayed 25 days and a feature stating that the marital status of the loan applicant is of type one. The actual meanings of the marital status categories are not specified in the dataset description.

3) Faithfulness: As the 10 most important features were modified one by one for 100 instances, the average change in probability for the overall most important features in this sample was as described in Table III. A negative probability indicates the reduction in classification probability when considering each feature's importance. Faithful algorithms should have negative changes in probability for all features.

TABLE III: Faithfulness

Feature dropped	Average change in probability
FLAG_EMAIL	-0.0124
QUANT_CARS	0.0124
COMPANY	-0.0031
APP_TYPE_Carga	0.0209
MARITAL_STATUS_1	-0.0147
FLAG_PROF_PHONE	-0.0119
PAYMENT_DAY_25	-0.0436
MARITAL_STATUS_2	0.0395
PAYMENT_DAY_10	-0.0222
AGE	-0.0066

C. Literature study

From the literature study, four main points were identified as important in the adoption of machine learning algorithms and explainable AI in creditworthiness assessment. These aspects were the following:

1) *Its phase in the product lifecycle*
2) *Implications of being a first mover*
3) *Risk of gaming the system*
4) *Storing and processing user data*

These points are discussed in more detail under Business applicability in Discussion.

VII. Discussion

A. Interpreting the results

1) Model performances: Logistic regression and random forest provided very similar results in terms of AUC score. Due to this insignificant difference between the models, it is difficult to conclude that random forest is a better model than logistic regression in terms of predictive performance. As a previous study has proposed that modern machine learning algorithms, including random forest, can replace logistic regression for tasks like credit scoring, the result obtained in this study was somewhat unexpected [15]. Reasons for the different results may be insecurities in the data or that the tuning of hyperparameters in the random forest classifier was not optimal. Nevertheless does random forest not perform worse than logistic regression. Noteworthy is also that the overall results of both models are considered low. Linhart et al.'s paper, which is described above in the section Previous research, were able to reach an AUC score of 0.677 on a dataset similar to the one used in this study [16]. A difference in their method was that they combined a logistic regression model with a KNN function. Also, they used more extensive methods for feature engineering. For example, in terms of categorical values, the authors tested three different approaches for transformation of these. Since this thesis put more emphasis on explaining the model, such extensive feature engineering was considered out of scope. The added complexity to the model will however have to be compared to an increase of just about 0.045 in terms of AUC score. On the other hand, the AUC score achieved in this thesis is far lower than 0.90, which was achieved with completely different datasets in the study by He et al. [17]. The great differences in predictive performance across different datasets suggests that the data quality is highly important.

2) Explainability: Differences in presenting feature importances was apparent between logistic regression and SHAP. One of the more prominent differences is that SHAP explains how the decision is affected by high and low feature values, where logistic regression has constant feature coefficients. These more advanced explanations of SHAP comes however at a cost of increased complexity. With respect to the different recipients of explanations, a number of differences between the methods are noteworthy. In terms of loan officers and loan applicants, who both require explanations for specific instances, a disadvantage with logistic regression is that it does not offer any built-in way of extracting local explanations, this is however possible with SHAP. Regarding regulators and data scientists, who require explanations on the entire dataset, no model has a clear advantage for producing global explanations.

Regarding faithfulness, a truly faithful model would result in only negative changes of probability, which Table III shows is not the case for SHAP. However, these results does not necessarily mean that SHAP is not trustworthy at all. As the method used takes into consideration that the most important feature should have a larger impact than the next most important feature, the result only shows that the mutual order may not in all cases be correct. Trust can still exist in the algorithm. As features were not actually dropped, but instead simulated by using the mean, the method used could be improved. Therefore, the results of the faithfulness test is inconclusive.

3) Feature importances: When comparing which features are considered important in the models, differences can be found. While the 10 most important features in the trained logistic regression classifier are dominated by residential zip codes, a much greater variability of features are considered important in the trained random forest classifier according to SHAP. Here, important features included payment day, marital status and occupation type for example. However, both age and flag for residential phone was determined important in both methods. The differences between the feature importances in SHAP and logistic regression could be due to several reasons. Firstly, logistic regression and random forest are trained in two distinctive ways. While logistic regression uses the logistic function, random forest consists of multiple decision trees. Hereby, it is possible that the inherent structure of the two methods leads to different features being important for the predictions. Secondly, with random forest and SHAP, the values computed to explain the predictions not only relies on the trained random forest model, but also on the approximations used in SHAP.

Furthermore, notable findings from the SHAP explanations in Figure 6 are the impact of the applicants' age and how some binary features contribute to the predictions. In the figure, the age attribute displays a clear pattern where older ages are more often associated with good loan applicants and younger ages with bad loan applicants. Also, ages in between generally have a Shapley value close to zero, indicating that those are not as impactful. In terms of binary features, some of those, including flag for residential phone, has clustered Shapley values that are negative and close to zero when the feature value is one. However, in the case where the feature has the value zero, the Shapley values are more equally distributed and ranges between approximately 0.00–0.10. This indicates that having a residential phone has low impact on the classification but not having a phone can have a large impact. As this pattern, although sometimes in the opposite direction, is repeated on several categorical features, including the feature for a delayed payment of 25 days and the feature for being in occupation type one, it indicates that features can have both a large impact and a low impact.

B. *Business applicability*

1) Growth phase: From the literature study, it was noted that explainable AI is still in an early stage of the product life cycle, which must be taken into consideration when comparing it with traditional methods. In the Gartner hype curve for 2020, explainable AI just passed the peak of inflated expectations, indicating that explainability is in the growth stage and has a far way until maturity [30]. SHAP was released in 2016 with improvements and additions still being integrated. Logistic regression is therefore a much more refined method and may, with the presented results, make a compelling choice for managers at credit assessing companies. However, this assumes a binary relationship between explainable AI and logistic regression and that these two methods cannot coexist. Instead should the respective methods be used within areas they excel at. Given that there exists several types of recipients, the need for different types of explanations is clear. SHAP can be used for explaining specific applications for loan officers and end users, which is not possible at the same level of fidelity using logistic regression. For data scientists who require global explanations, comparisons between SHAP and logistic regression is easier and the use of both can give greater insights.

2) First mover implications: A player who is early on the market in a new field of technology is commonly referred to as a first mover [31]. As assessment of creditworthiness through machine learning models combined with explainable AI has not yet been fully exploited, credit assessing companies may be able to take advantage of being early to the market. One major benefit of being a first mover is that it can increase brand reputation as other companies will perceive the firm as a leader within the technology. This opportunity is especially prominent for third party credit assessing companies who sell their service to a wide range of financial institutions. Furthermore, being a first mover is advantageous for exploiting buyer switching costs. This can be done by integrating the creditworthiness assessment algorithms with the customer's system, making it difficult for them to switch. Lastly, being early in the field makes it possible to set standards and to collect user data at an early stage.

Being a first mover may also come with disadvantages as the exploration of new technology is associated with a risk. One of the large disadvantages can be high development costs in integrating the new algorithms in the current system. This is especially disadvantageous as limited research has been done on customer requirements within explainable AI, and the development costs may be larger than the increased income. Another disadvantage with being a first mover is that the structure and components in financial institutions might not support random forest and SHAP. However, as SHAP is model-agnostic and thereby not dependent on the underlying algorithm, this disadvantage is somewhat reduced.

3) Gaming the system: With local explanations provided to the end users there are also risks associated. One potential risk when using algorithms for assessment is for users exploiting or gaming the system to improve their score without actually improving their behaviour [32]. For example, the second most important feature in the global explanation made by SHAP in Figure 6, flag for residential phone, could easily be changed temporarily by an applicant. Despite there not being a behavioral change to the applicant, its score would increase. After having applied for a loan the residential phone could be sold, which would further indicate that a score increase was not justified. This issue is not relevant for all features as for example quantity cars is not as easily changeable. Presenting explanations to the users must therefore balance increased transparency versus giving away company secrets.

For companies in the creditworthiness sector it could also be used as a way of incentivising applicants to change their behavior. One example could be by giving higher scores to online applications which are cheaper to process compared to paper applications.

4) Storing and processing user data: Some ethical issues need to be addressed when storing and processing user data. These issues are present not only in the field of credit scoring, but also in fields such as user specific marketing and recommendations. It is especially critical when the data is easy to associate with a specific person and when it involves financial data, as this is extra sensitive. If a loan applicant is presented with explanations on how the model used the data, the applicant may experience more trust and therefore also be more willing to use the service. However, in a study where the authors investigated how explanations of predictions made by machine learning models affected the trust of a user, they found that accuracy of the model was more important than the explanations in terms of building trust [33]. It should however be noted that the study focused on determining offensive language on social media platforms and that the user group was teenagers. It is of interest to investigate this relationship between users' willingness to share information and explanations in other domains, including assessment of creditworthiness, and with other user groups as well.

C. Limitations

A major limitation in the study was the scarce descriptions of the attributes in the dataset. The limited information about certain features made handling of missing or invalid data difficult, especially when not familiar with Brazilian culture and geography. One of the methods employed in this paper was to average missing numeric values, which relies on the assumption that they statistically would be close to the mean. Without a clear understanding of how the data was gathered and what some feature values represented, the assumption may not have been justified. Similar issues arose regarding the categorical features, where grouping into a specific category was used for missing values. Other options could have been to change missing values to the most common category or to use predictive models like KNN to estimate the data. Due to the lack of background knowledge of the dataset it was difficult to verify that the most appropriate methods were used.

With the use of machine learning algorithms comes the need for tuning hyperparameters, which is a time consuming and sometimes complicated process. Limited resources in terms of time and computer capacity put constraints on the strategy used. This was further complicated by the use of a combined algorithm of random forest and SHAP. For example, when tuning hyperparameters in random forest, consideration also had to be taken to how these parameters would affect the computation of SHAP explanations, a task which turned out to be very resource intensive.

Another limitation lies in the lack of user related measurements when evaluating the explainability of the models. In this thesis, the focus when evaluating explainability was mainly functional. However, as proposed in a previous study that examined explainable AI literature in order to develop a set of criteria for evaluating explainable systems [34], usability should also be considered. Possible metrics could include user opinions through likert scales.

D. Future work

Future work within the area of creditworthiness and explainable AI is recommended to consider the points brought up in the Limitations section above. Out of these points, subjective metrics for evaluation of usability is particularly emphasized. Furthermore, this thesis focused only on two different classifiers, an ensemble classifier and logistic regression. As methods based on deep learning have shown promising results within this field, it would be interesting to see future research also include these in a comparison.

VIII. Conclusion

In conclusion, this study has shown that the modern machine learning algorithm random forest may not always outperform the more traditional logistic regression model when assessing creditworthiness of loan applicants. Furthermore, several advantages of using the explainable AI algorithm SHAP to interpret a black-box algorithm were identified. These advantages included the possibility to produce explanations on both a local and a global level. Also, with the use of frameworks like SHAP, areas where transparency is of great importance do not have to be restricted to methods that are explanatory by their inherent structure. The two methods evaluated in the study identified different features as important for determining bad loan applicants. Three of the important characteristics were the applicant's age, where they lived and whether they had a residential phone. The study also proposed several factors that are important in the adoption of explainable AI into an actual decision system. As explainable AI is still in an early stage of innovation, it could be advantageous to first implement the technology as a complement rather than a substitute to existing technology. There can also be both advantages and disadvantages with being a first mover in an emerging industry. Especially important in the field of creditworthiness assessment, is that while providing insufficient explanations to a customer might harm user experience, too much information could potentially enable gaming or other untruthful behaviour.

REFERENCES

[1] A. Rai, *Explainable AI: from black box to glass box.* J. of the Acad. Mark. Sci. 48, 137–141, 2020.

[2] C. Livada, *Assessment of consumers' creditworthiness*, ERA Forum, vol. 20, no. 2, pp. 225-236, 2019. Available: 10.1007/s12027-019-00574-w.

[3] M. Gouvêa, E. Gonçalves *Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models*, POMS 18th Annual Conference, 2007. Available: https://www.pomsmeetings.org/confpapers/007/007-0210.pdf

[4] N. Bussmann, P. Giudici, D. Marinelli and J. Papenbrock, *Explainable AI in Fintech Risk Management*, Frontiers in Artificial Intelligence, vol. 3, 2020. Available: 10.3389/frai.2020.00026.

[5] L. Gilpin, D. Bau, B. Yuan, A. Bajwa, M. Specter and L. Kagal, *Explaining Explanations: An Overview of Interpretability of Machine Learning*, 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018. Available: 10.1109/dsaa.2018.00018.

[6] L. Demajo, V. Vella and A. Dingli, *Explainable AI for Interpretable Credit Scoring*, Computer Science Information Technology (CS & IT), 2020. Available: 10.5121/csit.2020.101516.

[7] Provenzano, A. R., Trifirio, D., Datteo, A., Giada, L., Jean, N., Riciputi, A., Le Pera, G., Spadaccino, M., Massaron, L. and Nordio, C., *Machine Learning approach for Credit Scoring*, 2021. Available: https://arxiv.org/pdf/2008.01687.pdf.

[8] D. Jurafsky, J. Martin, "Logistic regression" in *Speech and Language Processing*, 2019. [Online]. Available: https://web.stanford.edu/ jurafsky/slp3/.

[9] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, Springer Texts in Statistics, 2013. Available: 10.1007/978-1-4614-7138-7.

[10] C. Molnar. *Interpretable Machine Learning*, 2021. [Online]. Available: https://christophm.github.io/interpretable-ml-book/index.html.

[11] Google Developers, *Classification: ROC Curve and AUC*, 2020, [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/thresholding.

[12] 'Classification: ROC and AUC' by Google Developers available at https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc under a Creative Commons Attribution 4.0. Full terms at http://creativecommons.org/licenses/by/4.0.

[13] D. Jurafsky, J. Martin, "Naive Bayes and Sentiment Classification" in *Speech and Language Processing*, 2019. [Online]. Available: https://web.stanford.edu/ jurafsky/slp3/4.pdf.

[14] M. Stone "The Product Life Cycle" in ´*Product Planning*, Palgrave Macmillan, London. 1976. Available: 10.1007/978-1-349-02250-2_4

[15] S. Lessmann, B. Baesens, H. Seow and L. Thomas, *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research*, European Journal of Operational Research, vol. 247, no. 1, pp. 124-136, 2015. Available: 10.1016/j.ejor.2015.05.030.

[16] C. Linhart, G. Harari, S. Abramovich and A. Buchris, *PAKDD Data Mining Competition 2009: New Ways of Using Known Methods*, New Frontiers in Applied Data Mining, pp. 99-105, 2010. Available: 10.1007/978-3-642-14640-4_7.

[17] H. He, W. Zhang and S. Zhang, *A novel ensemble method for credit scoring: Adaption of different imbalance ratios*, Expert Systems with Applications, vol. 98, pp. 105-117, 2018. Available: 10.1016/j.eswa.2018.01.012.

[18] F. Dosilovic, M. Brcic and N. Hlupic, *Explainable Artifical Intelligence: A Survey.*, Available: https://www.researchgate.net/profile/Mario-Brcic/publication/325398586_Explainable_Artificial_Intelligence_A_Survey/links/5b0bec90a6fdcc8c2534d673/Explainable-Artificial-Intelligence-A-Survey.pdf.

[19] N. Olofsson, *A Machine Learning Ensemble Approach to Churn Prediction*, 2017. Available: http://www.diva-portal.org/smash/get/diva2:1118767/FULLTEXT01.pdf.

[20] E. Zihni, V. Madai, M. Livne, I. Galinovic, A. Khalil, J. Fiebach and D. Frey, *Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome*, PLOS ONE, vol. 15, no. 4, p. e0231166, 2020. Available: 10.1371/journal.pone.0231166.

[21] PAKDD-2010 Data mining competition. *Re-Calibration of a Credit Risk Assessment System Based on Biased Data* , 2010. [Online]. Available: https://web.archive.org/web/20150925091413/http://sede.neurotech.com.br/PAKDD2010/arquivo.do?method=load

[22] Scikit learn, *MinMaxScaler*, 2020 [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[23] I. Brown and C. Mues, *An experimental comparison of classification algorithms for imbalanced credit scoring data sets*, Expert Systems with Applications, vol. 39, no. 3, pp. 3446-3453, 2012. Available: 10.1016/j.eswa.2011.09.033.

[24] B. Krawczyk, *Learning from imbalanced data: open challenges and future directions*, Progress in Artificial Intelligence, vol. 5, no. 4, pp. 221-232, 2016. Available: 10.1007/s13748-016-0094-0.

[25] Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, *SMOTE: Synthetic Minority Over-sampling Technique*, Cs.cmu, 2002. [Online]. Available: https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-html/chawla2002.html.

[26] Scikit learn, *Logistic Regression*, 2020 [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[27] Scikit learn, *Random Forest*, 2020 [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[28] Lundberg, S., 2018. *Python Version of Tree SHAP — SHAP latest documentation*. [Online] Shap.readthedocs.io. Available: https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Python%20Version%20of%20Tree%20SHAP.html

[29] Scikit learn, *Stratified K-Fold*, 2020 [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html.

[30] Gartner, *5 Trends Drive the Gartner Hype Cycle for Emerging Technologies, 2020*, 2020 [Online]. Available: https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020/.

[31] M. Schilling, "First-mover advantages" in *Strategic Management of Technological Innovation*, 2013. McGraw-Hill Irwin.

[32] C. Petre, B. Duffy and E. Hund, *"Gaming the System": Platform Paternalism and the Politics of Algorithmic Visibility*, Sage journals, 2019. Available: 10.1177/2056305119879995.

[33] A. Papenmeier, G. Englebienne and C. Seifert, *How model accuracy and explanation fidelity influence user trust in AI*, 2019. Available: https://arxiv.org/pdf/1907.12652.pdf.

[34] K. Sokol and P. Flach, *Explainability fact sheets*, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020. Available: 10.1145/3351095.3372870.

Marcus Ankaräng is currently studying Industrial Engineering and Management at KTH Royal Institute of Technology in Stockholm, Sweden. He contributed to all parts in the thesis and was responsible for implementing the algorithms.

Jakob Kristiansson is currently studying Industrial Engineering and Management at KTH Royal Institute of Technology in Stockholm, Sweden. He contributed to all parts in the thesis and was responsible for analysing previous research and developing the thesis structure.

Appendix

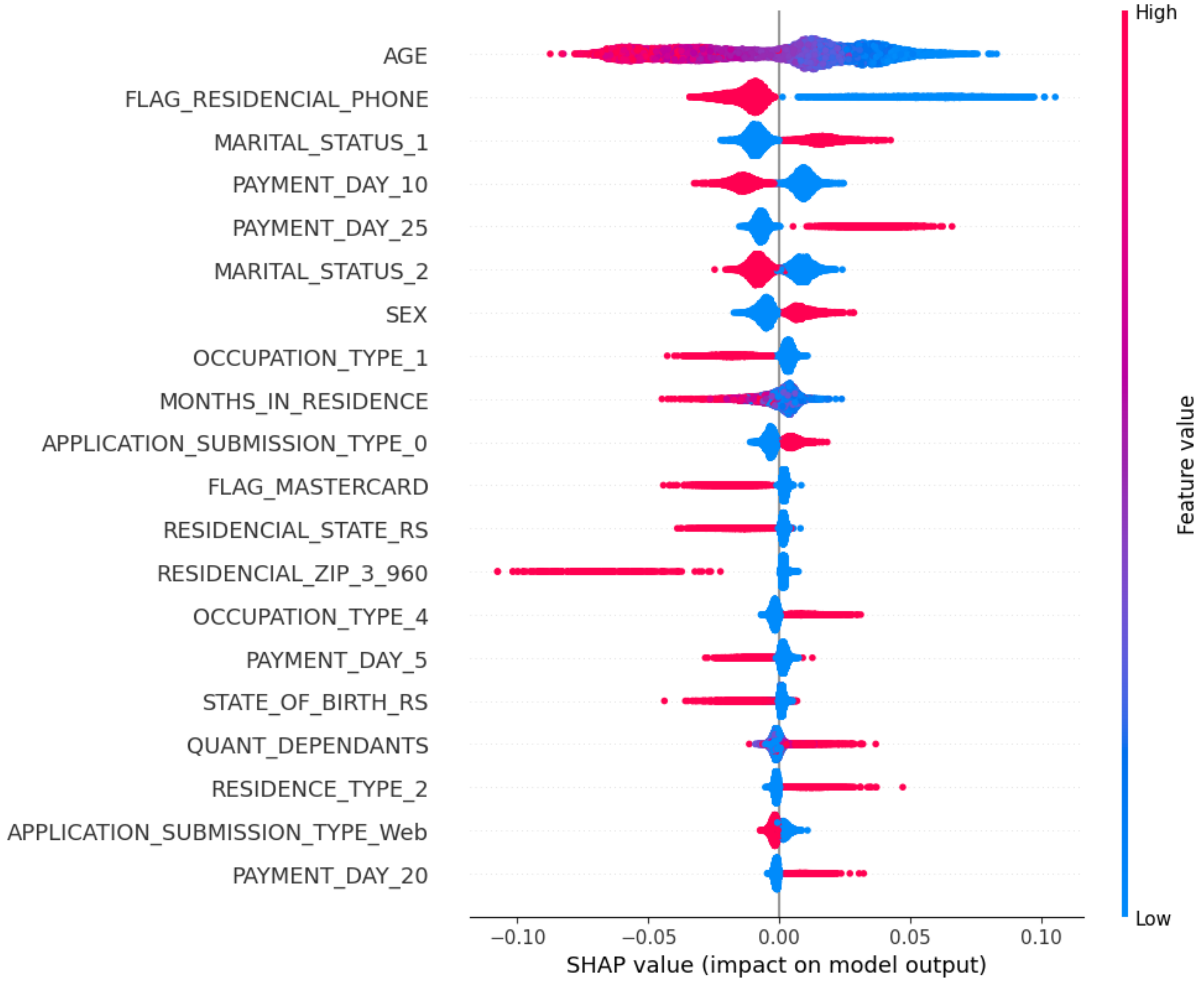
A. Full list of data features

ID_CLIENT
CLERK_TYPE
PAYMENT_DAY
APPLICATION_SUBMISSION_TYPE
QUANT_ADDITIONAL_CARDS
POSTAL_ADDRESS_TYPE
SEX
MARITAL_STATUS
QUANT_DEPENDANTS
EDUCATION_LEVEL
STATE_OF_BIRTH
CITY_OF_BIRTH
NACIONALITY
RESIDENCIAL_STATE
RESIDENCIAL_CITY
RESIDENCIAL_BOROUGH
FLAG_RESIDENCIAL_PHONE
RESIDENCIAL_PHONE_AREA_CODE
RESIDENCE_TYPE
MONTHS_IN_RESIDENCE
FLAG_MOBILE_PHONE
FLAG_EMAIL
PERSONAL_MONTHLY_INCOME
OTHER_INCOMES
FLAG_VISA
FLAG_MASTERCARD
FLAG_DINERS
FLAG_AMERICAN_EXPRESS
FLAG_OTHER_CARDS
QUANT_BANKING_ACCOUNTS
QUANT_SPECIAL_BANKING_ACCOUNTS
PERSONAL_ASSETS_VALUE
QUANT_CARS
COMPANY
PROFESSIONAL_STATE
PROFESSIONAL_CITY
PROFESSIONAL_BOROUGH
FLAG_PROFESSIONAL_PHONE
PROFESSIONAL_PHONE_AREA_CODE
MONTHS_IN_THE_JOB
PROFESSION_CODE
OCCUPATION_TYPE
MATE_PROFESSION_CODE
EDUCATION_LEVEL
FLAG_HOME_ADDRESS_DOCUMENT
FLAG_RG
FLAG_CPF
FLAG_INCOME_PROOF
PRODUCT
FLAG_ACSP_RECORD
AGE
RESIDENCIAL_ZIP_3
PROFESSIONAL_ZIP_3
TARGET_LABEL_BAD

B. Logistic regression coefficients

Feature	Feature coefficient
AGE	-2.0587
RESIDENCIAL_ZIP_3_402	1.7907
RESIDENCIAL_ZIP_3_912	1.4588
RESIDENCIAL_ZIP_3_788	-1.3634
FLAG_RESIDENCIAL_PHONE	-1.2009
RESIDENCIAL_ZIP_3_917	1.1334
RESIDENCIAL_ZIP_3_689	-1.1018
RESIDENCIAL_ZIP_3_535	1.0431
RESIDENCIAL_ZIP_3_114	-1.0316
RESIDENCIAL_ZIP_3_454	-1.0206
RESIDENCIAL_ZIP_3_681	-1.0052
RESIDENCIAL_ZIP_3_286	-0.9851
RESIDENCIAL_ZIP_3_607	0.9795
RESIDENCIAL_ZIP_3_790	0.9746
RESIDENCIAL_ZIP_3_750	-0.9467
RESIDENCIAL_ZIP_3_818	0.9098
RESIDENCIAL_ZIP_3_691	-0.8911
RESIDENCIAL_ZIP_3_651	-0.8886
RESIDENCIAL_ZIP_3_403	0.8848
RESIDENCIAL_ZIP_3_580	0.8655
RESIDENCIAL_ZIP_3_819	0.8649
RESIDENCIAL_ZIP_3_650	0.8458
RESIDENCIAL_ZIP_3_140	0.8407
RESIDENCIAL_ZIP_3_814	0.8341
RESIDENCIAL_ZIP_3_940	0.8155
RESIDENCIAL_ZIP_3_400	0.8072
RESIDENCIAL_ZIP_3_867	-0.7935
RESIDENCIAL_ZIP_3_138	-0.7703
RESIDENCIAL_ZIP_3_508	0.7696
RESIDENCIAL_ZIP_3_173	-0.7661
RESIDENCIAL_ZIP_3_486	-0.7445
RESIDENCIAL_ZIP_3_682	0.7381
RESIDENCIAL_ZIP_3_730	-0.7285
RESIDENCIAL_ZIP_3_931	-0.7108
RESIDENCIAL_ZIP_3_837	-0.6977
RESIDENCIAL_ZIP_3_510	0.6920
RESIDENCIAL_ZIP_3_603	0.6871
RESIDENCIAL_ZIP_3_375	-0.6742
RESIDENCIAL_ZIP_3_244	-0.6674
RESIDENCIAL_ZIP_3_980	0.6668

C. Enlarged SHAP summary plot



TRITA-EECS-EX-2021:368