Netspar
Network for Studies on Pensions, Aging and Retirement

NETSPAR ACADEMIC SERIES

Macroeconomic factors in loan default prediction

A machine learning based approach

Sjef Roijmans

MSc 12/2020-016

Macroeconomic factors in loan default prediction: A machine learning based approach

Sjef Roijmans
STUDENT NUMBER: 1274279

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:
Eric Postma
Peter Hendrix

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
December 2020

PREFACE

Dear reader,

Thank you for showing interest in my research on macroeconomic features in loan default prediction. I would like to thank my supervisor, Eric Postma, for his guidance and feedback.

Please enjoy reading my thesis,

Sjef Roijmans

ABSTRACT

This study focuses on the impact of macroeconomic factors within credit risk assessments in the peer to peer lending market. The credit risk industry faces a classification problem, i.e. predicting at initiation whether a loan will default or not. Existing literature obtained statistically significant logistic regression coefficients for macroeconomic variables. The predictive capability in terms of classification performance of macroeconomic variables in machine learning-based models has, however, not been tested. Besides, previous literature on explaining black-box algorithms in the credit risk industry mostly focused on the prediction fidelity of XAI tools and not on the economic interpretation of the SHAP values. This research attempts to improve state-of-the-art classification performance by including macroeconomic variables in the decision-making process. In addition, the importance and economic interpretation of the macroeconomic features are assessed by analyzing the SHAP values. The study uses the LendingClub loan data set from 2012 up to 2020. The main results include that machine learning algorithms are capable of leveraging macroeconomic features to improve overall classification performance. Furthermore, there appears to be a selection effect present in credit risk assessment, i.e. during economic downturns, default probabilities of originated loans decrease due to investors' demand for stricter lending conditions.

TABLE OF CONTENTS

1. Introduction

Since the recent financial crisis, a new form of lending has arrived, called peer-to-peer (P2P) lending. This practice, also known as crowdlending, consists of matching lenders with individuals or businesses through an online platform and attempts to eliminate the redundant financial intermediaries. Intuitively, the most important question for investors in the lending industry is whether a borrower will pay back its debt. The P2P lending market addresses this question more efficiently by leveraging the rich amount of alternative data in combination with the power of machine learning-based credit risk models. This helps them in constructing a more insightful and complete picture of the creditworthiness of borrowers, resulting in more favorable financing terms for both lenders as well as borrowers.

This thesis will focus on the additive value of macroeconomic variables in machine learning-based credit risk models. The analysis consists of a classification task, i.e. predicting at initiation whether an individual loan will default during the loan term. Two state-of-the-art models, the multi-layer perceptron (MLP) and the random forest (RF) algorithm (Moscato et al., 2021; Namvar et al., 2018), and one base model, the logistic regression (LR), are deployed to test the relevance of a set of common macroeconomic indicators in default prediction models.

Previous literature (Croux et al., 2020) established the importance of macroeconomic variables in traditional, logistic regression models and found machine learning-based models to outperform logistic models in the classification task. However, research on the inclusion of macroeconomic variables in machine learning-based models remained untouched. Therefore, this study will contribute to the existing literature by examining a more extensive and complete set of predictors, including macroeconomic variables in machine learning-based credit risk models. Moreover, this study incorporates new loan data as the LendingClub dataset is constantly updated. This allows for more robust analyses and more data points for data-hungry algorithms such as neural networks. Lastly, previous literature on explaining black-box algorithms in the credit risk industry mostly focused on the prediction fidelity of XAI tools and not on the economic interpretation of the SHAP values. This study adds to the existing literature by examining which macroeconomic variables affect the default prediction models most and which economic interpretation theory of the direction of the impact seems to prevail in these complex ‘black box’ algorithms.

Improving credit risk models is of practical relevance as it is beneficial to both borrowers and lenders. A more accurate default prediction model can prevent borrowers from taking on too much debt and avert any additional settlement and recovery fees. Secondly, research on the importance of a wide-ranging set of variables in default prediction models and general macroeconomic interpretation theories can aid investors’ understanding of the decision-making process of investors. Thirdly, investors face lower default risk if they have better screening capabilities, resulting in lower interest rates for borrowers. Lastly, enhanced credit risk models can help grow loan approval rates by evaluating borderline applicants more accurately. This is beneficial, especially to young borrowers who tend to have a short credit history and lower income and are therefore often not considered under traditional credit risk models.

The overall impact of macroeconomic variables on predicted default probabilities by machine learning models is assessed through the first research question. The second and third research questions are deployed to measure the impact of individual variables on default probabilities.

1. Are machine learning-based models capable of leveraging macroeconomic variables in default prediction to improve overall classification accuracy?
2. Are GDP growth, stock market returns, inflation, consumer confidence index, housing prices and average income negatively related to loan default probabilities in machine learning-based models?
3. Are the volatility index, unemployment rate, risk premium and economic policy uncertainty positively related to loan default probabilities in machine learning-based models?

This study established that macroeconomic conditions indeed play an important role in credit risk assessment. Adding macroeconomic variables to the classification models significantly improved the performance in terms of G-mean and area under the curve (AUC) compared to state-of-the-art classification results. The best performance on the test set was achieved under the neural network model using random oversampling. Two robustness checks, SHAP values and performance testing without the macroeconomic variables, confirmed that the classification improvement is indeed attributable to the adoption of macroeconomic factors. Also, the SHAP values showed that the consumer confidence index, inflation and average income per zip code are the most important macroeconomic factors. Lastly, it appears that a so-called ‘selection effect’ dominates the direction of the impact of macroeconomic factors. The selection effect argues that during economic downturns, default probabilities of originated loans decrease due to investors’ demand for stricter lending conditions.

The next section includes an overview of the current state of literature. Section three contains the methodology. Section four consists of the experimental setup and data used. Section five presents the results. Section six includes the discussion. Section seven contains the conclusion. At the end, a list of references and appendices is added.

2. Literature review

2.1. P2P Lending

In recent years, a new form of credit financing emerged, called peer-to-peer lending. This type of lending can be seen as a debt-based form of crowdfunding, connecting borrowers and financiers directly, without any intermediary parties. The explosive growth of the internet and social networks significantly lifted the scale of P2P lending by making borrowers less reliant on existing social relations for generating lending agreements. Besides the growth of the internet and interconnectivity, the main drivers of the success of P2P lending are its ability to leverage the rich amount of alternative data in combination with machine learning-based credit risk models. This helps them in constructing a more insightful and complete picture of the creditworthiness of a borrower. The following two subsections will elaborate further on the latter two value drivers and their roles in default prediction.

2.2. Alternative Data & Macroeconomic Determinants

The emergence of alternative loan data in credit risk models has greatly improved credit scoring models. Whereas traditional credit data is often intuitively related to a borrower's creditworthiness, e.g. fico score, current debt and the length of a person's credit history, alternative data includes a much broader and personal set of variables. This broader set can include information of a borrower's recurring transaction history, social media data, insurance claims, a person's profession, employment length or their educational background. The adoption of such alternative data sources by P2P platforms is also confirmed in terms of correlation by Jagtiani and Lemieux (2019). They observed that the correlation between LendingClub's loan grades and a borrower's FICO score, a widely used US credit score based on traditional metrics, decreased from 80% to 35% between 2007 and 2015, highlighting the embracement of alternative data sources by P2P platforms.

In addition to the abovementioned personal characteristics, macroeconomic conditions can also act as an alternative data source in credit risk assessment, capturing both county and country-level economic conditions. The most widely studied economic indicators in credit risk models are gross domestic product (GDP), inflation rate as measured by the change in the customer price index (CPI), stock market index return, the volatility index (VIX), unemployment rate, income, housing prices and policy-related uncertainty or disagreement measures (Croux et al., 2020; Jagtiani & Lemieux, 2018; Ramcharan & Crowe, 2013; Yoon et al., 2019).

GDP growth, stock market returns, inflation, average income and house price per county are intuitively negatively related to default risk as they all indicate a positive financial outlook for borrowers, and therefore decrease the default probability. Besides, if stock markets are performing well, investors might be more inclined to invest in stocks instead of P2P loans, therefore decreasing supply and the need to invest in more risky and obscure loans. On the other hand, unemployment, VIX, and policy uncertainty are assumed to be positively related to default probabilities, since high degrees of uncertainty or disagreement are often indicators of upcoming, adverse economic conditions. The above reasoning method, which assumes that general economic conditions are positively related to borrowers' financial condition, however, might be offset by a selection effect of lenders, i.e. during economic downturns, a negative economic outlook might result in demand for stricter lending conditions by investors, leading to lower expected default probabilities. This effect was observed by (Chen et al., 2017) during the economic crisis of 2008, during which the supply for small business credit decreased sharply. Unrelated to the above theories, high levels of inflation should decrease the real value of a loan and therefore lower the expected default probability. Lastly, the relative performance on zip code based macroeconomic variables can act as a signal of relative creditworthiness.

The importance of macroeconomic predictors in credit scoring and serving otherwise rejected loan applicants in economically less developed areas by P2P lending platforms is confirmed by Jagtiani and Lemieux (2018). They observed that the share of LendingClub's loans out of all loans in a local area is positively correlated with the unemployment rate and negatively correlated with average income and house price. Croux et al. (2020) examined the default determinants of LendingClub's P2P loans by incorporating a large set of loan, borrower and investor characteristics, and macroeconomic variables in their default prediction models. They concluded that alternative data had significant power in default prediction, even after controlling for obvious, traditional characteristics, e.g. fico score or income. Besides, they discovered that the risk premium, VIX and Russel 2000 return increase default probabilities and institutional investor characteristics and GDP growth decrease default probabilities. Another study by Yoon et al. (2019) examined default probabilities in the Chinese P2P lending market by including macroeconomic indicators in the model. They uncovered a negative relationship between the stock and real estate market and platform default risk, but the unemployment rate turned out insignificant. Ramcharan and Crowe (2013) analyzed the impact of housing prices on credit availability on Prosper, another major P2P lending platform. They pointed out that loan applicants from states with declining house prices face higher interest rates, credit rationing and more likely delinquencies, especially for subprime applicants whose balance sheet is most exposed to asset price fluctuations.

2.3. Machine Learning and Default Prediction

The field of credit risk assessment has known many different methods and developed rapidly over the past two decades. At the very start, loans were accepted or declined based on the personal judgements of loan originators, also known as expertise-based techniques. This method strongly depended on the officer's domain knowledge and proficiency, making this method a rather subjective and inconsistent one. Baklouti and Baccar (2013) concluded that even for an experienced loan officer, it remained tough to make an unbiased and rational decision in most cases. Along with computational advancements, more robust and consistent statistical models arrived in the default prediction field. Emekter et al. (2015) proved that a logistic regression model could improve accuracy in default prediction compared to expertise-based models on LendingClub's dataset. However, the main drawbacks of this linear model is its inability to capture more complex, non-linear relationships between input variables and the eventual prediction. Nonetheless, the logistic regression is often included in state-of-the-art papers and acts as a baseline model for comparison with other, more complex algorithms.

More recently, machine learning-based algorithms were introduced in credit scoring models. Moscato et al. (2021) and Namvar et al. (2018) compared the performance of different classifiers such as RF, Logistic Regression, MLP and Linear Discriminant Analysis (LDA) in default prediction models and both found the RF classifier to perform best in terms of overall performance, i.e. Area Under the Curve (AUC) and G-mean. Bastani et al. (2019) focused on loan profitability from an investors perspective and developed a two-stage model of which the first stage consisted of a default prediction model. They examined the performance of MLP, RF, SVM and found slightly superior performance for the MLP compared to the RF and SVM. Kim and Cho (2019) analyzed the ability of deep convolutional neural networks to automatically extract features from the dataset and make more accurate predictions. They show that the convolutional neural network effectively acts as a feature extraction mechanism and produces higher levels of accuracy. However, they achieved a less balanced performance in terms of precision and recall.

The overall outperformance of machine learning-based credit risk models compared to traditional, statistical models, demonstrates their capability of capturing more complex, non-linear relationships between a large set of input variables and the loan outcome. The current state of literature, however, restricted itself to loan and borrower characteristics as input features for machine learning-based models. This leaves an open gap for further research on the potential of including macroeconomic variables in these credit risk models. Especially since previous research demonstrated the importance of such variables in statistical credit risk models. Next to that, the impact of macroeconomic variables can differ largely per applicant, depending on personal and loan characteristics, as demonstrated by Ramcharan and Crowe (2013). This indicates the complexity of the ways macroeconomic conditions affect individual loan default and the potential of machine learning in modelling such relationships.

2.4. Sampling Approach

The LendingClub dataset suffers from a class imbalance problem. A class imbalance problem arises when the number of observations in one class, i.e. default, is much scarcer than another class, i.e. fully paid. Previous literature dealt with such problems by employing resampling techniques before training its models. Common techniques include Random UnderSampling (RUS), Random OverSampling (ROS), Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002), Instance Hardness Threshold (IHT) (Smith et al., 2014) and hybrid approaches between under- and oversampling techniques. Both Namvar et al. (2018) and Moscato et al. (2021) demonstrated strong performance using RUS, ROS and SMOTE, while Bastani et al. (2019) realized solid performance under IHT.

2.5. Research Questions

This research will contribute to the existing literature by examining the importance of macroeconomic variables in machine learning-based loan default models. Given the importance of such variables in traditional models (section 2.2) and the solid performance of machine learning algorithms in credit risk models (section 2.3), it is an appealing subject for further research and interesting to assess the capability of machine learning-based algorithms to also capture non-linear relationships between macroeconomic conditions and loan defaults. Besides, previous research on the impact of macroeconomic factors only established their relevance in terms of significance levels of the coefficients of the corresponding macroeconomic features. An assessment of the relevance of macroeconomic factors in terms of overall classification accuracy has not been conducted.

The performance of macroeconomic variables in loan default models is first assessed by comparing the classification models, including macroeconomic variables, against state-of-the-art classification results (Moscato et al., 2021). For robustness purposes, the SHAP values are checked to analyze the extent to which model decisions and potential improvements are attributable to macroeconomic conditions. This results in the following research question;

- Are machine learning-based models capable of leveraging macroeconomic variables in default prediction to improve overall classification accuracy?

Besides assessing the overall importance of macroeconomic variables, it is also interesting from an academic as well as practical perspective to analyze the impact of macroeconomic features separately. This can help both investors and borrowers in understanding which factors and underlying theories drive default predictions, stimulate acceptance of borderline applicants and prevent borrowers from taking on irrational loan amounts. Besides, the direction of the impact of macroeconomic factors on default probabilities remains inconclusive for certain factors, such as stock market returns. Croux et al. (2020) found a positive relationship, while Yoon et al. (2019) observed a negative relationship between stock returns and loan default probabilities. This study will build further on these previous studies by evaluating the direction of the impact of macroeconomic features in machine learning models, which can potentially discover more complex, non-linear relationships and evaluate the different economic interpretation theories, as described in section 2.2. Based on previous results from statistical models and the underlying theories on the impact of individual macroeconomic features on default probabilities, the following two research questions are used to test the importance and direction of these macroeconomic features. Again, SHAP values are used to assess the importance and direction of individual macroeconomic features.

- Are GDP growth, stock market returns, inflation, consumer confidence index, housing prices and average income negatively related to loan default probabilities in machine learning-based models?

- Are the volatility index, unemployment rate, risk premium and economic policy uncertainty positively related to loan default probabilities in machine learning-based models?

3. Methods

3.1. Resampling

Loan default datasets are characterized by their imbalanced nature, i.e. the number of loan defaults is considerably lower than the amount of fully paid loans. Previous research demonstrated that resampling techniques can vastly improve classification performance, especially in minority class prediction, i.e. loan defaults. Existing studies deployed many methods of which RUS, ROS, IHT and SMOTE proved to perform best. This study employs two oversampling techniques, ROS and SMOTE, and two undersampling techniques, RUS and IHT.

3.1.1. Random under and oversampling

The random undersampling and random oversampling techniques are the simplest strategies to pick loans for the transformed, balanced dataset. Random undersampling relies on randomly deleting observations from the majority class, whereas random oversampling relies on randomly duplicating observations from the minority class.

3.1.2. Instance hardness threshold

Instance hardness threshold is an undersampling technique which lessens the class imbalance problem by deleting hard to classify samples from the majority class. The idea behind removing 'hard' samples from the training set is that they often include outliers, noise or class overlap, which can slow down or hurt the training process. IHT resampling consists of an algorithm, e.g. Random Forest, which predicts the probability, p, that label y is assigned to the input vector, x. It then acts as a filter by removing instances with an instance hardness value higher than the threshold. The formula for calculating instance hardness, as defined by Smith et al. (2014), is presented below;

$$IH_h(\langle x_i, y_i \rangle) = 1 - p(y_i | x_i, h) \tag{1}$$

Where IH_h is the instance hardness value, $\langle x_i, y_i \rangle$ is the training data and $p(y_i | x_i, h)$ is the probability with which algorithm h assigns label y_i to the input x_i.

3.1.3. Synthetic minority oversampling technique

One of the drawbacks of ROS is that it involves duplicating minority samples, which does not create any new data for the model. SMOTE (Chawla et al., 2002) attempts to improve on this by synthesizing new samples from the existing minority class. Hence, SMOTE is considered a type of data augmentation for the minority group. The minority class is oversampled by creating new instances along the line segments between a minority sample and one of its k-nearest neighbors. The synthetic values are generated by taking the difference of the input vectors of the minority sample and its k-nearest neighbor. Hereafter, the difference between the input vectors is multiplied by a random number between one and zero and then added back to the original minority sample vector. This process is summarized below;

$$x_{new} = x_i + (x' - x_i) \times \delta \tag{2}$$

Where x_{new} is the newly generated sample, x_i is the existing minority sample, x' is the k-nearest neighbor of the minority sample and δ is a random number between one and zero. This method of resampling improves generalization as it forces a classifier to create wider and less specific decision regions.

3.2. Algorithms

The main goal of this study is to map a broad set of loan, borrower and macroeconomic characteristics to the output variable, i.e. default or fully paid. Previous literature (Bastani et al., 2019; Moscato et al., 2021; Namvar et al., 2018) demonstrated firmest performance under the random forest and multi-layer perceptron algorithm. Therefore, this study takes on the random forest and multi-layer perceptron as state-of-the-art models and the logistic regression will act as a baseline model.

3.2.1. Logistic regression

The logistic regression is a statistical classification method used to predict the probability of a certain class, e.g. the probability of default. The logistic regression model consists of taking the sigmoid function of a weighted sum of input variables, e.g. the loan characteristics, and will result in a probability between zero and one. This can be summarized as follows;

$$x = \theta \times weight + b \tag{3}$$

$$probability\ of\ default(x) = \frac{1}{1+e^{-x}} \tag{4}$$

Where x is a weighted sum of the input variables θ and a constant, b, and e is the natural logarithm base. The model is then optimized using an iterative optimization algorithm, e.g. gradient descent, to optimize the weights.

3.2.2. Random forest

The random forest algorithm is a tree-based classifier that incorporates many individual decision trees and operates as an ensemble. All the individual trees make independent predictions of the correct output class and the most common class is chosen as the output value. The individual trees split the training data based on a measure of informativeness of a split on a certain feature, such as Gini impurity. The split with the lowest level of Gini impurity, e.g. the split which performs best at distinguishing between defaulted loans and fully paid loans, is chosen as the next node. Gini impurity is calculated as follows;

$$G = \sum_{i=1}^{C} p(i) \times (1 - p(i)) \tag{5}$$

Where G is the level of impurity, C is the number of classes and p(i) is the probability of selecting a sample of class i. To be able to leverage the advantages of ensemble modelling of the random forest, the individual decision trees need to be sufficiently uncorrelated with each other. This is achieved by two concepts, bagging and feature randomness. Bagging consists of training each decision tree on a slightly different training set by randomly sampling observations from the training set with replacement. Feature randomness forces more variation among trees by restricting the number of features to consider for each split, whereas standard decision trees consider all possible features. This will result in a 'forest' of trees which are trained on 'random' data and 'random' feature sets.

3.2.3. Multi-layer perceptron

The multi-layer perceptron is a type of neural network, which has proven to be a powerful mapping algorithm in many areas, including credit risk management. The outstanding predictive capability of the MLP originates from the multi-layered structure of the model, through which it can learn complex, non-linear representations in the training data and combine lower level features into higher-order features and eventually a class prediction. The MLP finds the best mapping function that transforms the input variables into the correct class. The mapping function is summarized as follows;

$$y = f(x; \theta) \tag{6}$$

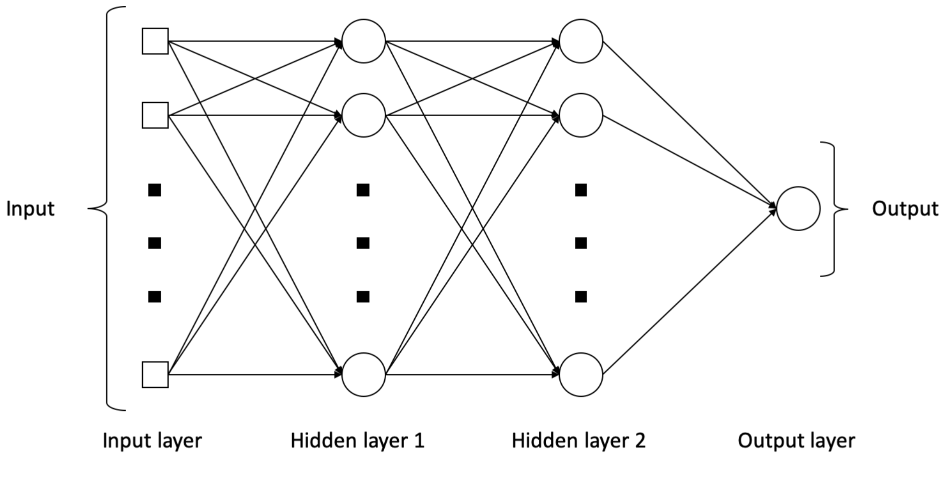
where y is the output label, x the input variables and θ the learnable parameters, i.e. weights and biases. The neurons in each hidden and output layer are summarized as follows;

$$y = g(b + \textstyle\sum_i x_i w_i) \tag{7}$$

where y is the neuron's output, g is an (often non-linear) activation function, b is the bias added to the neuron, x is the output of all the neurons of the previous layer and w represent the weights. A simplified overview of the MLP is presented in figure 1.

Figure 1

A simplified overview of the multi-layer perceptron. The white squares represent the input features, the circles represent the neurons in the different layers, the arrows represent the transformation of the output from the previous layer into the input of the next layer.



The training of the model consists of feeding batches of data into the model and comparing the predicted output with the actual output. The weights and biases of the neurons are optimized through backpropagation. The backpropagation algorithm calculates the gradient of the loss function with respect to the weights in the neurons using the chain rule. After training, the MLP is able to accurately map a set of input variables, e.g. loan characteristics, to the output variable, e.g. default or fully paid. In the remainder of this thesis, the multi-layer perceptron is referred to as the neural network.

3.3. SHAP values

The interpretability of a model can be just as important as the prediction accuracy, especially in heavily regulated businesses such as credit risk assessment. Besides, it is imperative to check whether any model improvement is attributable to one or more of the newly added macroeconomic variables in order to answer the three research questions of this study. Moscato et al. (2021) made a comparison of a set Explainable Artificial Intelligence (XAI) tools, including both rule-based and feature-based explainers, and found LORE and SHAP to perform best in terms of prediction fidelity. This study takes on the SHAP values as they proved to be a reliable explainer in previous research and are easily implemented using the SHAP package. The SHAP values are based on Shapley values from coalitional game theory and represent the contribution of the individual features to the model's output.

4. Experimental setup

4.1. Data

4.1.1. Data description and sources

The data used in this study originates from LendingClub's website and is made publicly available on Kaggle[1]. The loan data is updated on Kaggle with a small delay, i.e. the latest version available originates from 2020, Q1. Since the launch of LendingClub's platform in 2007, a total of 2883788 loans were successfully funded, each containing 141 loan observations. The Y variable is a dummy variable (1 = charged off, 0 = fully paid). The independent variables can be divided into 3 categories; loan and borrower characteristics, and macroeconomic variables. The loan and borrower characteristics are all included in the Kaggle dataset. The macroeconomic variables, 15 in total, are collected from various sources and can be divided into two types, country-level and zip code-level variables. The country-level variables include the S&P 500 and Russel 2000 returns, the volatility index, consumer confidence index, risk premium, policy uncertainty index and inflation (change in CPI), which are collected from Yahoo Finance, Quandl, policy uncertainty[2] and OECD[3]. The zip code-level variables include average income and income change, house price index and house price change, GDP index and GDP change, and the unemployment rate and unemployment rate change, which are collected from the Bureau of Labor Statistics (BLS), the Bureau of Economic Analysis (BEA) and the Federal Housing Agency (FHA). Moscato et al. (2021) achieved the state-of-the-art classification result. For comparability reasons, this study attempts to replicate their methodology. They use a subset of the data of this study, i.e. loans originated between 2016-2017.

4.1.2. Pre-processing

In this study, only loans with a definite outcome are considered, i.e. loans still outstanding are deleted, as the goal of this study is to predict whether a loan will default or not. This results in a dataset containing 1.859 million loans of which 1.496 million are fully paid and 0.363 million are charged off, i.e. a default rate of 19.5%. An overview of the monthly originated loans is displayed in figure 2. Forward-looking variables such as total payments, hardship flag or last fico score are deleted to avoid data leakage. Next, meaningless variables such as loan ID or URL are deleted and raw text-based variables such as loan description and employment title are removed as semantic analysis falls outside the scope of this study. LendingClub introduced new variables over the past decade such as joint variables and credit history indicators. This results in a large portion of missing values for older loans. Therefore, all variables added after 2012 are deleted and all loans before 2012 are dropped to incorporate some of the newly added variables. In the early years of LendingClubs business, i.e. 2007-2012, the amount of loans funded is relatively low, 67527. Deleting all loans funded before 2012,

[1] Source: https://www.kaggle.com/ethon0426/lending-club-20072020q1

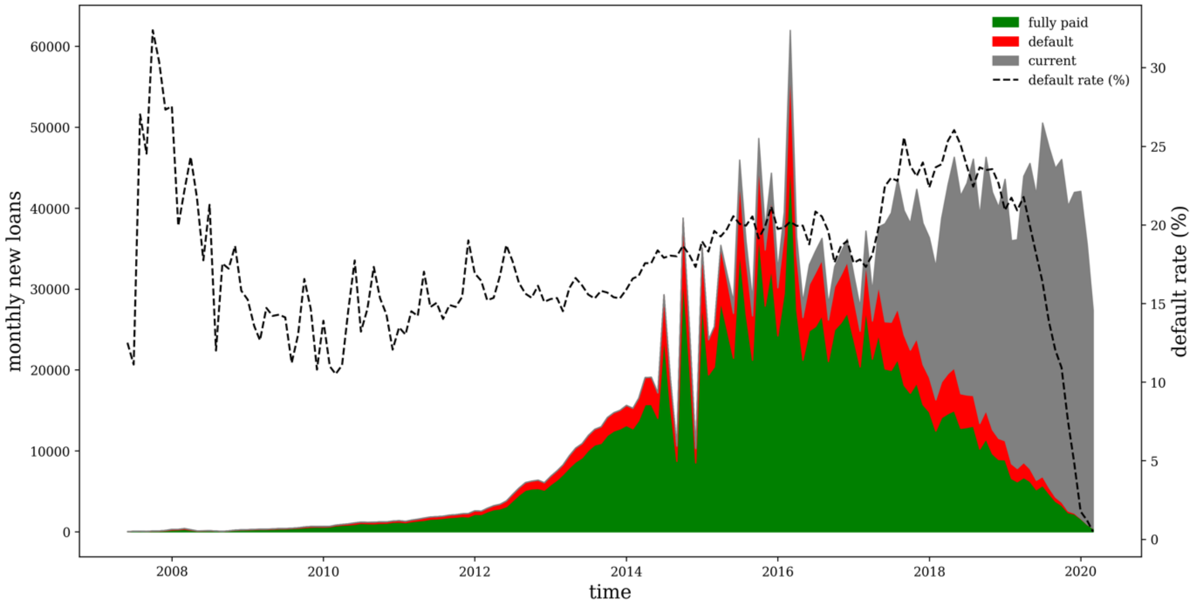
[2] Source: Baker, Bloom and Davis, Measuring Economic Policy Uncertainty, www.PolicyUncertainty.com.

[3] Source: Consumer Confidence index, https://data.oecd.org/leadind/consumer-confidence-index-cci.htm

therefore, has a relatively small impact on the total amount of loans. Besides, Moscato et al. (2021) also didn't incorporate loans before 2012. Next, loans with 3 or more missing values are dropped. The missing values within the employment length variable are imputed with the lowest possible value, 'less than 1 year' as it is likely that persons who don't have a job leave this section open. Also, the default rate for applicants with a missing value for employment length is 25% which is much higher than the overall rate of 19.5%. Similar to Moscato et al. (2021), the remaining missing values are median imputed, which is less affected by outliers than mean imputing. In total, 88,000 loans are deleted, leaving 1,771,316 loans to train and test the credit risk models. Moscato et al. (2021) also perform feature selection on their dataset based on missing values, and standard deviation and correlation of features. They used a cut-off point of 55% missing whereas this study deletes all variables with a missing-value percentage greater than 45% (these are all variables which were introduced in a later stadium by LendingClub, as mentioned above). An overview of the variables with missing values is presented in appendix 1. The implementation of feature selection based on correlation and standard deviation is not disclosed in their paper. They only disclose the deleted 11 variable names. Since this study includes a dataset over many more years compared to their study and since they did not disclose their implementation, this study did not perform feature selection based on correlation or standard deviation. Besides, 4 of the 11 variables are still deleted as they contained more than 45% missing values.

Figure 2

The figure includes a summary of the loans originated by LendingClub. The colored areas represent the number of newly issued loans in each month for each type. The green area represents the number of fully paid loans, the red area represents the number of defaulted loans and the gray area stand for the number of current loans. The dashed line depicts the monthly default rate, i.e. the number of defaulted loans divided by the sum of the number of fully paid and defaulted loans.



The LendingClub dataset includes many ‘month since’ variables, e.g. months since recent inquiry. The ‘month since’ variables are transformed by taking the inverse of them to convert the time interval into something that relates to frequency (Wurm, 2019). After taking the inverse, it is also possible to impute a sensible value, zero, for the missing values (when it never happened). The earliest credit line data is transformed by taking the difference between the earliest credit line date and the loan origination data.

Categorical features are either dummy encoded or transformed into an ordinal variable. Loan status, loan term, purpose and application type are dummy encoded as they cannot be logically ranked. Homeownership status, loan (sub)grade, employment length and verification status are ranked on an ordinal scale as they can be logically ordered from weak to strong. Dummy variables with very few observations are added to the ‘other’ group. Variables which took the form of dollar amounts included many extreme variables. Therefore the log of the dollar amount is taken to make the data less skewed. The fico high and low score are transformed into an average fico score.

Table 1

Descriptive statistics macroeconomic variables. The base year for the index values constitutes 2001.

Variable	Count	Mean	Std	Min	25%	Median	75%	Max
GDP growth (%)	1768387	2.14	2.92	-26.78	0.73	2.11	3.49	56.08
GDP index (%)	1768387	132.89	35.71	32.10	116.91	127.60	139.73	1054.28
Average income zip	1768387	48678	15552	24128	39488	45589	53867	251728
Average income change zip (%)	1768387	3.42	2.55	-22.35	2.15	3.62	4.97	38.99
Risk premium (%)	1768387	11.82	4.82	2.52	8.23	11.35	14.58	30.11
Unemployment rate (%)	1768387	6.00	1.91	1.40	4.61	5.67	7.00	18.33
Unemployment change (%)	1768387	-11.68	6.92	-38.89	-16.42	-12.38	-7.97	61.19
House price change (%)	1768387	4.95	4.45	-20.99	1.95	4.32	7.73	27.13
House price index (%)	1768387	149.79	34.80	61.96	125.61	143.78	168.35	393.32
Inflation (%)	1768387	1.37	0.84	-0.20	0.83	1.52	2.04	2.95
VIX	1768387	15.29	3.73	9.51	12.86	14.19	17.06	28.43
Policy uncertainty index (%)	1768387	81.18	26.80	46.40	63.20	72.45	92.19	193.70
Consumer confidence index (%)	1768387	100.58	0.77	98.40	100.39	100.71	101.18	101.62
S&P 500 return (%)	1768387	0.75	3.01	-9.18	-1.50	0.71	2.32	8.30
Russel 2000 return (%)	1768387	0.73	4.13	-10.91	-1.39	0.95	3.20	10.99

The country and zip level macroeconomic variables are merged based on the reporting year, month or zip code. To avoid any data leakage, the macroeconomic variables are set back one time period, either 1 month or 1 year. Stock returns and changes in variables are calculated by;

$$x_t = \frac{x_t}{x_{t-1}} - 1 \tag{8}$$

Where x is the value of a macroeconomic indicator and t the time period.

An overview of the macroeconomic variables and their descriptive statistics is presented in table 1. An overview of all the variables used for training and testing is included in appendix 2. Besides, a correlation matrix of the individual, scaled features is incorporated in appendix 3.

4.2. Algorithms

The classification task of this research consists of predicting, at initiation, whether a loan will default or not. First, the data is split in a training, validation and test set, using a ratio of 8:1:1. The split is performed by the train_test_split function from the sklearn model selection package. As previously stated in chapter three, this study takes on four resampling techniques to deal with the class imbalance problem. The implementation of these resampling techniques is provided by the imblearn package. For simplicity reasons, this study applied the default settings and algorithms of the resampling modules. After resampling, the independent variables are scaled using the standardscaler from the sklearn preprocessing package. The scaling is performed as it can stimulate convergence of the technique used for optimization, especially when using neural networks.

Three algorithms are tested, the logistic regression, random forest and the neural network. The logistic regression is the most simple model and will act as a baseline. The implementation of the logistic regression is based on sklearn linear models. The model is tested using no, L1 and L2 regularization. All three regularization forms produced similar results. Therefore, the default setting is chosen, i.e. L2. The solver is set to 'lbfgs', again the default setting. Other solvers produced similar results. Max iterations is set to 1000 to ensure model convergence. Moscato et al. (2021) also used the 'lbfgs' solver and 'L2' regularization. After training the logistic model on the training set and tuning the hyperparameters on the validation set, the model performance is assessed on the test set.

The second algorithm is the random forest classifier and is imported from the sklearn ensemble package. Moscato et al. (2021) only disclosed the number of estimators and the max depth. Since the full set of hyperparameters is not published, the classifier is optimized by performing a grid search cross-validation using the sklearn model selection package. The grid search is evaluated using the ROC AUC score as it is a balanced score of recall and accuracy. The hyperparameter settings are tuned on the RUS training and validation set and are used for all four resampling techniques. The grid search resulted in the following hyperparameter settings; max_depth: 22, max_features: the square root of the total features (sqrt), min_samples_leaf: 20, min_samples_split: 3, n_estimators: 150. The final out of sample performance is measured on the test set.

The third, neural network model is imported from Tensorflow based Keras. Moscato et al. (2021) only disclosed the solver, adam, the hidden layer sizes, 100 and alpha, 0.0001. Since the full set of hyperparameters is not published, the model's hyperparameters are optimized based on the performance on the validation set. The sequential model is utilized and tested using two, three and four hidden layers.

The amount of neurons in the input and hidden layers is based on best practices and a grid search. The following amount of neurons are used, from input to output layer: [128, 64, 32, 16, 1]. The broadly used 'relu' activation attained best performance, although the outperformance was only marginal compared to other activations such as 'elu', 'sigmoid' or 'tanh'. The batch size is set to 128, as it performed best under the grid search. The 'adam' optimizer performed best compared to 'Adamax', 'Adadelta' and 'RMSprop'. The model is regularized using three dropout layers, each using a dropout rate of 0.1. Also, the model is dynamically optimized during training through the Earlystopping and ReduceLROnPlateau callbacks. The maximum number of epochs without an accuracy improvement on the validation data using oversampled (undersampled) data is set to 4 (7) for the Earlystopping and 3 (5) for the ReduceLROnPLateau callback and the learning rate is reduced by a factor of 0.1. These callbacks reduce overfitting and dynamically tune the number of epochs needed.

4.3. Evaluation metrics

The LendingClub dataset exhibits an imbalanced classification problem. Besides, the costs of false positives are much greater than the costs of false negatives (García et al., 2019), as the downside risk of loan defaults is up to 100%. Therefore, a single evaluation measure is insufficient to make a thorough judgement. Past literature (Moscato et al., 2021; Namvar et al., 2018) included a set of performance, e.g. accuracy (ACC), recall (TPR), precision (TNR), false-positive rate (FPR), area under the curve (AUC) and G-mean, to establish a complete and informative image of each model's proficiency. This study uses the same metrics and they are calculated as follows;

$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN} \tag{9}$$

$$Recall = \frac{TP}{FN+TP} \tag{10}$$

$$Precision = \frac{TN}{FP+TN} \tag{11}$$

$$FP - rate = \frac{FN}{FN+TP} \tag{12}$$

$$G - Mean = (precsion \times recall)^{0.5} \tag{13}$$

Where TP is the amount of correctly classified loan defaults, TN the amount of correctly classified fully paid loans, FN the amount of defaulted loans which were predicted to be fully paid and FP the amount of fully paid loans which were predicted to default. The AUC is a tradeoff measure between TP and FP under different classification thresholds and counts the area under the curve. The confusion matrix used in this study is summarized in table 2.

The optimal model is picked based on the g-mean score of the model, similar to Moscato et al. (2021). After choosing the best model, its performance is compared to previous literature. This is done through 10-fold cross-validation to assess the robustness of the model's performance under different subsets of the data. After finishing the cross-validation, the g-mean scores are visualized on a boxplot and compared with scores from previous research.

Table 2

Depiction of the confusion matrix used in this study.

		True	
		Fully paid	Default
Predicted	Fully paid	TN	FP
	Default	FN	TP

4.4. SHAP values

The SHAP values of the most efficient model are studied to determine whether any improvement in the classification task can be attributed to macroeconomic variables. The SHAP values are estimated using the DeepExplainer module from the SHAP package. The individual feature importance is based on the average absolute SHAP value of a given feature over 3000 randomly selected loans. Besides the magnitude, i.e. the average impact of a feature on a model's output, the direction of the impact, i.e. positive or negative, is approximated by analyzing the correlation between the loan features and their corresponding SHAP values.

4.5. Software packages and versions

The programming language used in this study is python, version 3.7.4. An overview of all the ground packages, their versions and functions are displayed in table 3. The coding script is available on github[4].

Table 3

Packages, versions and functions used in this study

Package	Version	Functions
Pandas	0.25.1	Multiple basic function
Numpy	1.17.2	Multiple basic function
Matplotlib	3.2.1	Multiple basic function
Sklearn	0.23.2	Train_test_split, StandardScaler, StratifiedKFold, LogisticRegression, RandomForestClassifier, GridsearchCV, roc_auc_score, confusion_matrix
Tensorflow	1.15.0	Input, Dense, Flatten, Dropout, Sequential, EarlyStopping, ReduceLROnPlateau, set_random_seed
Shap	0.30.0	DeepExplainer, shap_values, summary_plot
Imblearn	0.7.0	RandomUnderSampler, RandumOverSampler, SMOTE, IHT
Seaborn	0.9.0	Boxplot

[4] Soruce: https://github.com/sjefkok/thesis

5. Results

5.1. Classification performance

Table 4 lists the classification results of the algorithms and resampling techniques examined. A confusion matrix of the best classifier is added to appendix 4. Overall, the neural network seems to be the best classifier as it achieves the highest performance in terms of G-mean, recall and AUC. The G-means of the random forest and the logistic regression are similar. The random forest scores significantly better at recall, whereas the logistic regression scores more evenly across precision and recall. As expected, the models with the highest levels of accuracy tend to predict the majority class, i.e. fully repaid. The different resampling techniques produce vastly different results. The highest G-mean and AUC scores are both achieved using random oversampling, although the performance is only marginally better when compared to the random undersampling method. All algorithms tend to overfit the majority class when SMOTE is applied to the training data. The accuracy and precision under SMOTE are both excellent but the recall score is very poor. Since the recall score is more important than precision in credit risk assessment, SMOTE appears to be unsuitable. Lastly, the balance between precision and recall under IHT is completely opposite to SMOTE, although less extreme. Algorithms using IHT produce the highest levels of recall, although the precision scores are considerably lower than under other resampling techniques. The AUC score is, however, lower than algorithms using RUS or ROS, meaning that algorithms using IHT are not better at separating defaults from non-defaults, but tend to favor a higher recall score over a higher precision score.

Table 4

Classification results. Recall measures the amount of correctly predicted loan defaults over all true loan defaults. Precision measures the amount of correctly predicted fully paid loans over all true fully paid loans. FP-rate is equal to 1-recall. AUC equals the ROC-AUC-Score. The best scores of each column are in bold.

Algorithm	Resampling	Accuracy	Precision	Recall	FP-rate	G-mean	AUC
Logistic regression	RUS	0.662	0.663	0.660	0.340	0.661	0.721
Logistic regression	ROS	0.663	0.663	0.661	0.339	0.662	0.721
Logistic regression	SMOTE	0.759	0.867	0.320	0.680	0.527	0.684
Logistic regression	IHT	0.491	0.400	0.863	0.137	0.588	0.714
Random forest	RUS	0.647	0.637	0.687	0.313	0.661	0.723
Random forest	ROS	0.705	0.735	0.582	0.418	0.654	0.725
Random forest	SMOTE	0.799	0.966	0.120	0.880	0.341	0.710
Random forest	IHT	0.513	0.432	0.840	0.160	0.602	0.712
Neural network	RUS	0.661	0.656	0.681	0.319	0.668	0.731
Neural network	ROS	0.658	0.651	0.687	0.313	**0.669**	**0.732**
Neural network	SMOTE	**0.797**	**0.951**	0.170	0.830	0.402	0.720
Neural network	IHT	0.478	0.380	**0.877**	**0.123**	0.577	0.719

5.2. Comparison with state-of-the-art performance

To assess whether the addition of macroeconomic variables as independent variables in the classification model improved overall performance, the obtained results are compared with previous studies. Table 5 includes the top three classifiers stated by Moscato et al. (2021) and their performance on the test set. Contrary to this study, Moscato et al. (2021) realized their best results under the random forest algorithm. However, their best G-mean and AUC score are respectively 1.3% and 1.5% lower (in absolute terms) than the neural network ROS classifier of this study. This performance improvement might be due to the inclusion of macroeconomic variables. Similar to this study, RUS and ROS bring about the highest performance in terms of G-mean and AUC. Bastani et al. (2019) also studied the performance of different classifiers on LendingClub's loan data, with a special focus on neural networks. They used slightly different performance measures, i.e. precision with respect to the positive class whereas this study calculates precision with respect to the negative class. They do, however, report the AUC which is directly comparable to the AUC of this study. They obtained the soundest performance using neural networks, compared to random forest and gradient boosting, with an AUC of 0.71. This is considerably lower than the best AUC of this study, 0.732.

Table 5

Classification results from Moscato et al. (2021). Recall measures the amount of correctly predicted loan defaults over all true loan defaults. Precision measures the amount of correctly predicted fully paid loans over all true fully paid loans. FP-rate is equal to 1-recall. AUC equals the ROC-AUC-Score.

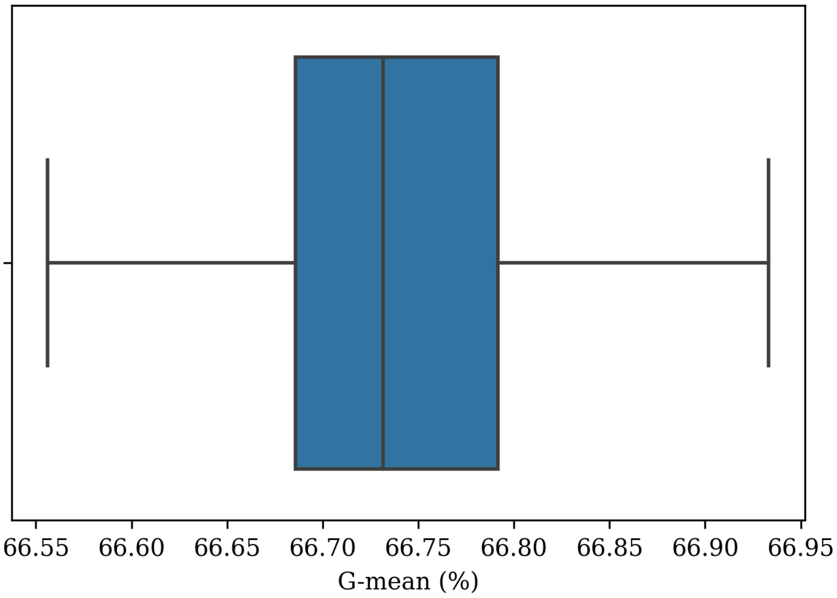
Algorithm	Resampling	Accuracy	Precision	Recall	FP-rate	G-mean	AUC
Random forest	RUS	0.640	0.630	**0.680**	**0.320**	**0.656**	**0.717**
Logistic regression	ROS	0.650	0.659	0.642	0.358	0.650	0.710
Logistic regression	SMOTE	**0.656**	**0.660**	0.639	0.360	0.650	0.709
Best own results							
Neural network	ROS	0.658	0.651	0.687	0.313	0.669	0.732

To assess whether the improvement in G-mean and AUC, respectively 1.3% and 1.5%, is significant, 10-fold cross-validation is run on the best performing model, i.e. neural network with ROS. The average G-mean score of the model equals 66.74% with a standard deviation of 0.10%. The distribution of the individual runs is depicted in figure 3. All the individual G-mean scores obtained during cross-validation are greater than the state-of-the-art results obtained by Moscato et al. (2021). Therefore it is highly likely that the newly proposed model significantly outperforms state-of-the-art models.

This study incorporates newly posted loans on LendingClub's website, up until Q1, 2020. Since the study by Moscato et al. (2021) incorporates only data from 2016 and 2017, any performance improvement might be due to more training data and not the addition of macroeconomic variables. In addition, there might be some differences in preprocessing or the quality of hyperparameter tuning. Therefore, for robustness purposes, the best performing model, neural network ROS, is run one more time, without including the macroeconomic variables. This resulted in a G-mean of 0.662, which is still below the lowest value of the 10-fold cross-validation and below the 99% confidence interval of the full model. In short, it seems that a significant portion of the model improvement is attributable to the inclusion of macroeconomic variables.

Figure 3

The figure depicts the 10-fold cross-validation G-mean score of the neural network using ROS. The minimum and maximum G-mean score equal 66.56% and 66.93%.



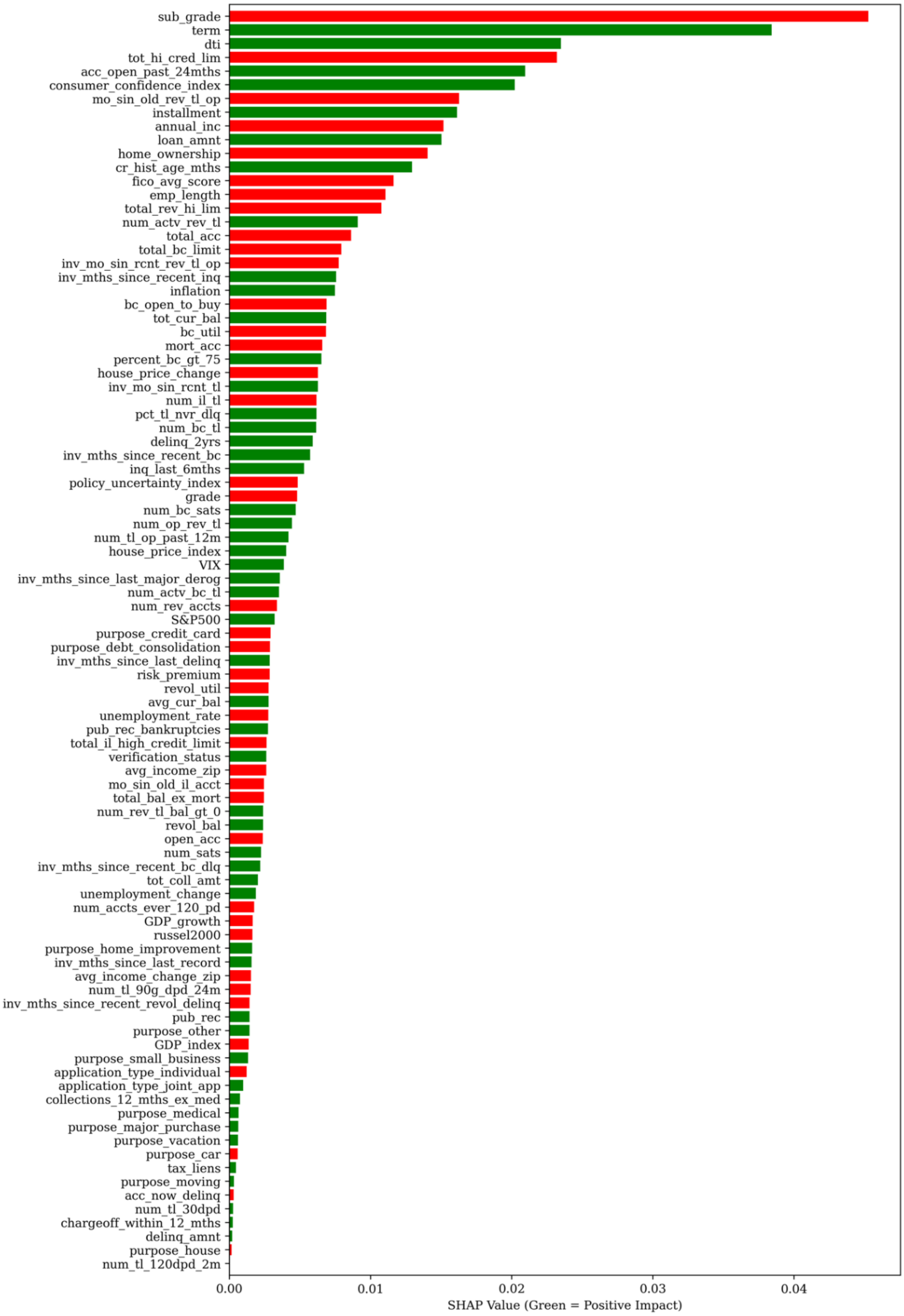
5.3. Economic interpretation - SHAP values

Another robustness check to confirm that the macroeconomic variables indeed influence the model's decisions is to analyze the SHAP values of the individual features. Besides, the SHAP values can provide insights into which macroeconomic features are most relevant in default prediction and which are not. Lastly, they can also shine a light on the direction of the impact of individual features on a model's output. Figure 4 contains the SHAP values of the individual features ordered by the magnitude and colored by the direction of the feature's impact on the model's output. The direction is determined by the sign of the correlation coefficient between feature values and their corresponding SHAP values (green if positive, red if negative). The macroeconomic variables are ranked as follows: consumer confidence index: 6, inflation: 21, house price change: 27, policy uncertainty index: 35, house price index: 40, VIX: 41, S&P 500: 45, risk premium: 49, unemployment rate: 52, average income: 56, unemployment rate change: 65, GDP growth: 67, Russel 2000: 68, average income change: 71, GDP index: 76.

Similar to the results obtained in section 5.2, it appears that a subset of macroeconomic variables has a considerable impact on the model's decisions. In terms of magnitude, the consumer confidence index (CCI), inflation and house price change are highly relevant in default prediction as they are all ranked

Figure 4

SHAP values individual features. The bars represent the absolute SHAP values. Red bars represent a positive correlation and blue bars a negative correlation on average between the feature values and their corresponding SHAP values. The best performing model, neural network with ROS, is used.



in the top 30 features and have SHAP values between 0.0063 – 0.02. Policy uncertainty index, house price index, VIX, S&P 500, risk premium, unemployment rate and average income are moderately important as they are ranked within the top 30 – 60 and have SHAP values between 0.0026 – 0.0049. Unemployment rate change, GDP growth, Russel 2000, average income change and GDP index seem to be irrelevant as their SHAP values are in the lowest tranche, i.e. 60 – 92, and their SHAP values are all below 0.0019.

Contrary to the hypothesis stated in chapter 2, the two most important macroeconomic features, CCI and inflation, are positively related to the probability of default, meaning that when consumer confidence or inflation is high, default probabilities are, on average, higher. The effect of house price change did turn out as expected, i.e. rising house prices decrease the probability of default.

The moderately important variables turned out partially as expected. The VIX and average income per zip code affected default probability as anticipated, i.e. low levels of stock market uncertainty and high levels of average income result in lower default probabilities. In contrast, the impact of policy uncertainty, house price index, S&P 500 returns, risk premium and unemployment rate on default likelihood ended up counterintuitive. High levels of policy uncertainty, risk premium and unemployment rate decrease default probabilities and high levels of house price index and S&P 500 returns increase default probabilities.

Lastly, the impact of all irrelevant variables turned out as projected. An increase in unemployment change increases default probabilities and an increase in average income change, GDP growth, GDP index or Russel 2000 return decreases default probabilities. The magnitude of the impact of these variables is, however, only trivial, meaning that the robustness of the direction is questionable.

6. Discussion

The main goal of this study is to assess the relevance of macroeconomic variables in credit default prediction in the P2P lending market. The credit risk business attempts to optimize a classification problem, i.e. predicting, at initiation, whether a loan will default or will be fully paid during the loan term. Previous statistical analysis demonstrated the significance of macroeconomic variables in logistic regressions (Croux et al., 2020). However, the performance of macroeconomic variables in default classification using machine learning algorithms has not been tested. Consequently, the main research question of this study is whether machine learning algorithms are capable of leveraging macroeconomic features to improve overall classification performance. Besides, this study will focus on the economic interpretation of the macroeconomic variables in machine learning algorithms through SHAP values.

6.1. Classification performance

This study incorporates four resampling methods to deal with the imbalanced data problem. Based on the results from table 4, it seems that random under and oversampling displays the strongest overall performance in terms of AUC and G-mean. The main purpose of SMOTE is to improve generalization by forcing classifiers to create wider and less specific decision regions for the minority class. However, the opposite effect is observed on the test set, i.e. compared to random oversampling, it significantly under classifies the minority class. This results in the lowest G-mean score across all resampling techniques. A possible explanation for this might be that the synthetic samples produced by SMOTE are unrealistic and vastly different from the default samples in the test set. In addition, the model is trained on fewer non-synthesized loan defaults, which appear to be most representative for the test set. This adverse effect might be mitigated by tuning the hyperparameters of the SMOTE algorithm, especially since Moscato et al. (2021) achieved a balanced performance between precision and recall. However, this study applied the default settings as resampling was not the focus of this study. The idea behind IHT resampling is to remove noise or reduce class overlap from the training set. This should improve generalization and training speed. The results indicate that IHT accomplishes much higher performance at recall. This is probably caused by reducing the class overlap and therefore classifying most borderline applicants as default. Precision bears the cost of this strategy as it is greatly reduced. Hence, the G-mean and AUC are worse compared to RUS or ROS.

In terms of classification algorithms, the neural network models demonstrate the most solid performance across both G-mean and AUC. This is surprising as Moscato et al. (2021) report the poorest performance under the MLP, although they recommend in their conclusion to further investigate the application of deep learning in credit risk models. On the other hand, Bastani et al. (2019) achieved the most solid performance using deep learning, in line with this study, compared to the random forest or support vector machines. Lastly, similar to Moscato et al. (2021), the random forest and logistic regression realized similar results in terms of AUC and G-mean.

In terms of absolute performance, the best classifier, neural network ROS, achieved an average G-mean score of 66,74% during 10-fold cross-validation. All the individual G-mean scores are well above the state-of-the-art performance by Moscato et al. (2021), meaning that the new model outperforms the existing state-of-the-art model with high certainty. This could be caused by the inclusion of macroeconomic features in the model, differences in feature selection, more data or better optimization of hyperparameters. The fact that the logistic regression in this study, which uses the exact same hyperparameter settings used in Moscato et al. (2021), still outperforms the logistic regression classifier from previous literature by over 0.5% in terms of G-mean, indicates that the outperformance cannot be fully attributed to model optimization. Besides, a robustness test demonstrated that, ceteris paribus, removing the macroeconomic variables decreased the G-mean to 66.2%, which is significantly lower than the performance with macroeconomic variables (66.74%). Although only part of the performance improvement is attributable to macroeconomic factors, machine learning algorithms are capable of leveraging macroeconomic features to significantly improve classification accuracy.

6.2. Economic interpretation - SHAP values

The last section of this research consists of analyzing the SHAP values of the independent variables in the model. The SHAP values confirm that, indeed, the macroeconomic conditions act as an important factor within default prediction. Whereas previous literature on explaining black-box algorithms in the credit risk industry mostly focused on the prediction fidelity of XAI tools such as SHAP or LIME, this study is centered on the economic interpretation of the SHAP values. Two main reasoning methods are proposed in chapter two. The first one argues that the macroeconomic variables act as an indicator of the overall economic outlook and therefore borrowers' financial outlook. Hence, if macroeconomic variables signal a solid economic outlook, it's expected that, on average, fewer borrowers will default on their loans. In addition, relative performance across zip codes can act as a signal of creditworthiness. Opposite to this, the second reasoning method argues that during economic downturns lenders might demand stricter lending conditions for borrowers, also called a 'selection effect', which results in lower default probabilities.

This study found the second theory of the selection effect to dominate over the first theory, although there are some exceptions. The two most important macroeconomic variables, CCI and inflation, both followed the theory of the selection effect. High inflation and consumer confidence are typically associated with economic expansion. However, they seem to increase the default probability by a great amount according to the SHAP values. The majority of the moderately relevant variables, 5 vs 2, also followed the selection effect theory. Exceptions in the most important and moderately important variable class include house price change, VIX and average income. The impact of the least important variable class, i.e. the irrelevant variables as mentioned in section 5.3, is ignored as the overall impact on model decisions is very low. Altogether, the opposite effect as stated in hypothesis two and three seems to be present, as the hypotheses were based on the first reasoning method, i.e. macroeconomic variables act as a proxy of the general economic outlook and therefore also borrowers' economic outlook.

6.3. Contributions to literature

This research contributed to the existing literature in three different ways. First of all, this study established that current, state-of-the-art, classification models can be improved by including macroeconomic variables. The significant, logistic regression coefficients of macroeconomic variables found in previous studies indeed bring about a significant performance improvement in terms of overall prediction accuracy. Secondly, this study demonstrated that neural network outperform logistic regression and random forest models, in contrast to Moscato et al. (2021). This signifies the capability of neural networks to capture more complex, non-linear relationships in credit data. Thirdly, the SHAP values of the best performing model improved the understanding of the driving variables in credit risk models and found a so-called 'selection effect' to dominate the overall direction of the impact of the macroeconomic features.

7. Conclusion

This study is centered on the impact of macroeconomic variables in default prediction in the P2P lending market. Previous statistical analysis displayed statistically significant logistic regression coefficients for macroeconomic variables. The predictive capability in terms of classification performance of macroeconomic variables in machine learning-based models has, however, not been tested. Besides, previous literature on explaining black-box algorithms in the credit risk industry mostly focused on the prediction fidelity of XAI and not on the economic interpretation of the SHAP values. This research aims to measure the magnitude and direction of the impact of macroeconomic variables in default classification, measured in terms of overall classification performance and individual SHAP values. The main results found in this study include that macroeconomic variables proved to be relevant in machine learning-based credit risk models. The neural network in conjunction with random oversampling performed best and the consumer confidence index, inflation and average income ended up as the most important macroeconomic predictors. Lastly, a selection effect is visible, i.e. investors seem to impose stricter lending conditions during economic downturns resulting in lower expected default rates. The opposite effect seems present during economic upturns.

This research identified three research directions for future research on credit risk assessment. Firstly, is it interesting to assess whether the produced results generalize to other P2P lending platforms in countries outside the United States. Secondly, this study discarded the use of text-based variables such as loan description, which can potentially further improve state-of-the-art results. Thirdly, research on the optimization of resampling techniques might further improve classification results.

8. References

Baklouti, I., & Baccar, A. (2013). Evaluating the predictive accuracy of microloan officers' subjective judgment. *International Journal of Research Studies in Management*. https://doi.org/10.5861/ijrsm.2013.343

Bastani, K., Asgari, E., & Namavari, H. (2019). Wide and deep learning for peer-to-peer lending. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2019.05.042

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.953

Chen, B. S., Hanson, S. G., & Stein, J. C. (2017). The Decline of Big-Bank Lending to Small Business: Dynamic Impacts on Local Credit and Labor Markets. In *Working Paper* (No. w23843). https://doi.org/10.3386/w23843

Croux, C., Jagtiani, J., Korivi, T., & Vulanovic, M. (2020). Important factors determining Fintech loan default: Evidence from a lendingclub consumer platform. *Journal of Economic Behavior and Organization*. https://doi.org/10.1016/j.jebo.2020.03.016

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*. https://doi.org/10.1080/00036846.2014.962222

García, V., Marqués, A. I., & Sánchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion*. https://doi.org/10.1016/j.inffus.2018.07.004

Jagtiani, J., & Lemieux, C. (2018). Do fintech lenders penetrate areas that are underserved by traditional banks? *Journal of Economics and Business*. https://doi.org/10.1016/j.jeconbus.2018.03.001

Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management*. https://doi.org/10.1111/fima.12295

Kim, J. Y., & Cho, S. B. (2019). Deep Dense Convolutional Networks for Repayment Prediction in Peer-to-Peer Lending. *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-3-319-94120-2_13

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

Namvar, A., Siami, M., Rabhi, F., & Naderpour, M. (2018). Credit risk prediction in an imbalanced social lending environment. *International Journal of Computational Intelligence Systems*. https://doi.org/10.2991/ijcis.11.1.70

Ramcharan, R., & Crowe, C. (2013). The impact of house prices on consumer credit: Evidence from an internet bank. *Journal of Money, Credit and Banking*. https://doi.org/10.1111/jmcb.12045

Smith, M. R., Martinez, T., & Giraud-Carrier, C. (2014). An instance level analysis of data complexity. *Machine Learning*. https://doi.org/10.1007/s10994-013-5422-z

Wurm, M. (2019, August). Intelligent Loan Selection for Peer-to-Peer Lending. *Towards Data Science*. https://towardsdatascience.com/intelligent-loan-selection-for-peer-to-peer-lending-575dfa2573cb

Yoon, Y., Li, Y., & Feng, Y. (2019). Factors affecting platform default risk in online peer-to-peer (P2P) lending business: an empirical study using Chinese online P2P platform data. *Electronic Commerce Research*. https://doi.org/10.1007/s10660-018-9291-1

9. Appendix

9.1. Appendix 1 – missing values

Missing Values percentages. All variables with more than 45% missing values are deleted

Variables	Missing Values	% of Total Values
sec_app_revol_util	1801501	96.9
revol_bal_joint	1800525	96.9
sec_app_collections_12_mths_ex_med	1800524	96.9
sec_app_open_acc	1800524	96.9
sec_app_earliest_cr_line	1800524	96.9
sec_app_fico_range_high	1800524	96.9
sec_app_fico_range_low	1800524	96.9
sec_app_inq_last_6mths	1800524	96.9
sec_app_mort_acc	1800524	96.9
sec_app_open_act_il	1800524	96.9
sec_app_num_rev_accts	1800524	96.9
sec_app_chargeoff_within_12_mths	1800524	96.9
verification_status_joint	1791845	96.4
dti_joint	1789666	96.3
annual_inc_joint	1789663	96.3
il_util	988328	53.2
all_util	846561	45.5
open_acc_6m	846424	45.5
inq_last_12m	846424	45.5
total_cu_tl	846424	45.5
inq_fi	846423	45.5
total_bal_il	846423	45.5
max_bal_bc	846423	45.5
open_rv_24m	846423	45.5
open_rv_12m	846423	45.5
open_il_24m	846423	45.5
open_il_12m	846423	45.5
open_act_il	846423	45.5
num_tl_120dpd_2m	135607	7.3
mo_sin_old_il_acct	121127	6.5
emp_length	117126	6.3
bc_util	68736	3.7
percent_bc_gt_75	68161	3.7
bc_open_to_buy	67721	3.6
pct_tl_nvr_dlq	67681	3.6
avg_cur_bal	67571	3.6
mo_sin_old_rev_tl_op	67528	3.6
num_rev_accts	67528	3.6
num_tl_90g_dpd_24m	67527	3.6
tot_coll_amt	67527	3.6

tot_cur_bal	67527	3.6
total_rev_hi_lim	67527	3.6
total_il_high_credit_limit	67527	3.6
tot_hi_cred_lim	67527	3.6
num_tl_op_past_12m	67527	3.6
num_accts_ever_120_pd	67527	3.6
num_tl_30dpd	67527	3.6
num_op_rev_tl	67527	3.6
num_actv_bc_tl	67527	3.6
num_actv_rev_tl	67527	3.6
num_il_tl	67527	3.6
num_bc_tl	67527	3.6
num_rev_tl_bal_gt_0	67527	3.6
num_bc_sats	55841	3
num_sats	55841	3
total_bal_ex_mort	47281	2.5
total_bc_limit	47281	2.5
acc_open_past_24mths	47281	2.5
mort_acc	47281	2.5
revol_util	1394	0.1
dti	1107	0.1
pub_rec_bankruptcies	697	0
chargeoff_within_12_mths	56	0
collections_12_mths_ex_med	56	0
tax_liens	39	0
zip_code	1	0
inq_last_6mths	1	0

9.2. Appendix 2 – variable descriptions

Descriptions of independent variables used for default prediction. This includes 19 loan characteristics, 58 borrower characteristics and 15 macroeconomic variables.

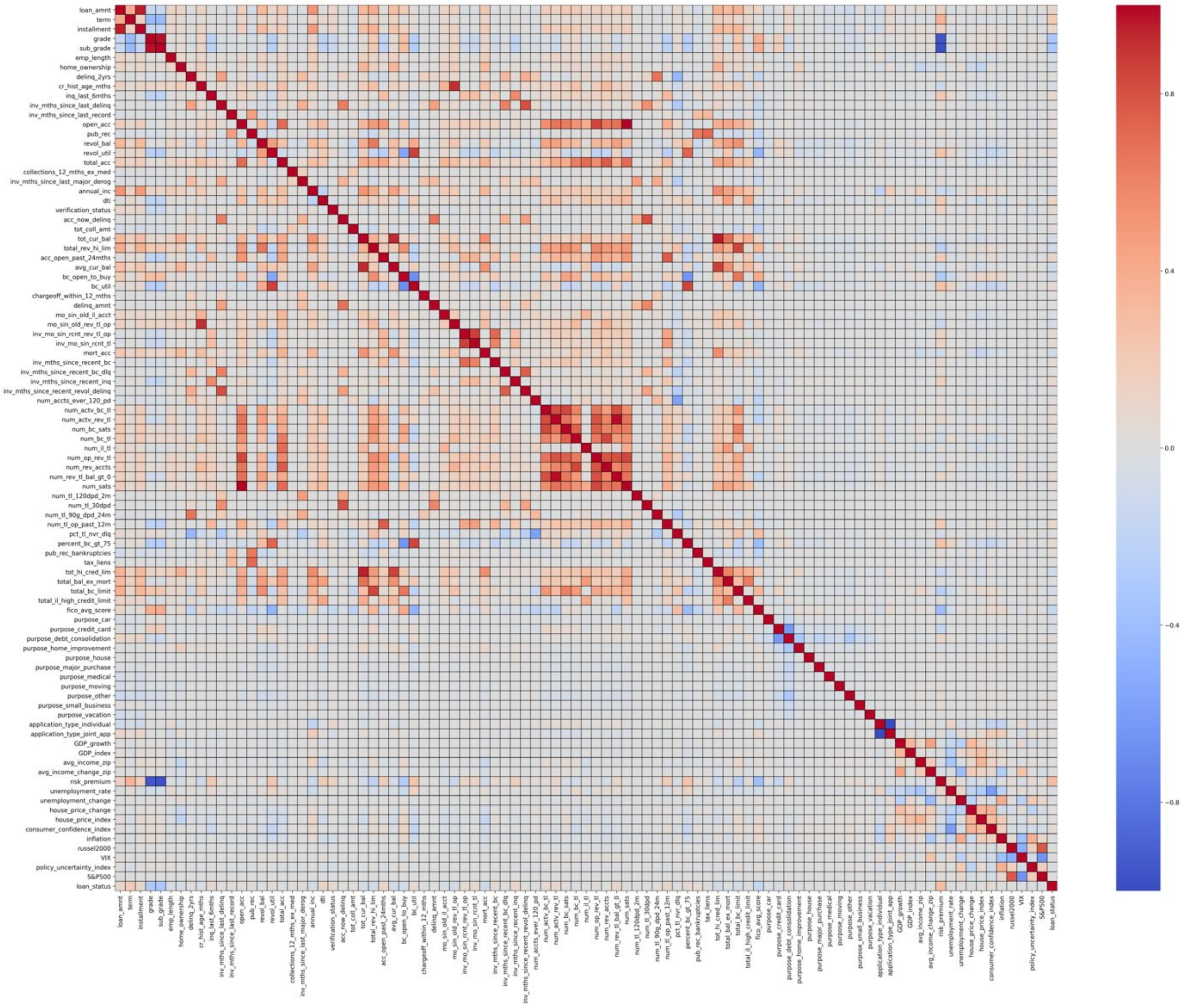
Abbreviation	Description	Type
loan_amnt	The loan amount	Loan
term	The term of the loan, 36 or 60 months	Loan
installment	The monthly payment owed by the borrower	Loan
grade	LC assigned loan grade	Loan
sub_grade	LC assigned loan subgrade	Loan
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections	Loan
purpose_car	Dummy purpose of loan is car	Loan
purpose_credit_card	Dummy purpose of loan is credit card	Loan
purpose_debt_consolidation	Dummy purpose of loan is debt consolidation	Loan
purpose_home_improvement	Dummy purpose of loan is home improvement	Loan
purpose_house	Dummy purpose of loan is house	Loan
purpose_major_purchase	Dummy purpose of loan is major purchase	Loan
purpose_medical	Dummy purpose of loan is medical	Loan

purpose_moving	Dummy purpose of loan is moving	Loan
purpose_other	Dummy purpose of loan is other	Loan
purpose_small_business	Dummy purpose of loan is small business	Loan
purpose_vacation	Dummy purpose of loan is vacation	Loan
application_type_individual	Dummy application type is individual	Loan
application_type_joint_app	Dummy application type is joint	Loan
emp_length	Employment length in years. Possible values are between 0.5 (<1 year) and 10 (10>years)	Borrower
home_ownership	The home ownership status. Other: 0, rent: 1, mortgage: 2, own: 3	Borrower
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years	Borrower
cr_hist_age_mths	Months since earliest credit history	Borrower
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)	Borrower
inv_mths_since_last_delinq	Inverse onths since last delinquency	Borrower
inv_mths_since_last_record	Inverse months since last record	Borrower
open_acc	The number of open credit lines in the borrower's credit file.	Borrower
pub_rec	Number of derogatory public records	Borrower
revol_bal	Total credit revolving balance	Borrower
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.	Borrower
total_acc	The total number of credit lines currently in the borrower's credit file	Borrower
inv_mths_since_last_major_derog	Inverse months since last major derogatory	Borrower
annual_inc	The self-reported annual income provided by the borrower during registration.	Borrower
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.	Borrower
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified	Borrower
acc_now_delinq	The number of accounts on which the borrower is now delinquent.	Borrower
tot_coll_amt	Total collection amounts ever owed	Borrower
tot_cur_bal	Total current balance of all accounts	Borrower
total_rev_hi_lim	Total revolving high credit/credit limit	Borrower
acc_open_past_24mths	Number of trades opened in past 24 months.	Borrower
avg_cur_bal	Average current balance of all accounts	Borrower
bc_open_to_buy	Total open to buy on revolving bankcards.	Borrower
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.	Borrower
chargeoff_within_12_mths	Number of charge-offs within 12 months	Borrower
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.	Borrower
mo_sin_old_il_acct	Months since oldest bank installment account opened	Borrower
mo_sin_old_rev_tl_op	Months since oldest revolving account opened	Borrower
inv_mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened	Borrower
inv_mo_sin_rcnt_tl	Months since most recent account opened	Borrower
mort_acc	Number of mortgage accounts.	Borrower
inv_mths_since_recent_bc	Months since most recent bankcard account opened.	Borrower
inv_mths_since_recent_bc_dlq	Months since most recent bankcard delinquency	Borrower
inv_mths_since_recent_inq	Months since most recent inquiry.	Borrower
inv_mths_since_recent_revol_delinq	Months since most recent revolving delinquency.	Borrower

num_accts_ever_120_pd	Number of accounts ever 120 or more days past due	Borrower
num_actv_bc_tl	Number of currently active bankcard accounts	Borrower
num_actv_rev_tl	Number of currently active revolving trades	Borrower
num_bc_sats	Number of satisfactory bankcard accounts	Borrower
num_bc_tl	Number of bankcard accounts	Borrower
num_il_tl	Number of installment accounts	Borrower
num_op_rev_tl	Number of open revolving accounts	Borrower
num_rev_accts	Number of revolving accounts	Borrower
num_rev_tl_bal_gt_0	Number of revolving trades with balance >0	Borrower
num_sats	Number of satisfactory accounts	Borrower
num_tl_120dpd_2m	Number of accounts currently 120 days past due	Borrower
num_tl_30dpd	Number of accounts currently 30 days past due	Borrower
num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in last 24 months	Borrower
num_tl_op_past_12m	Number of accounts opened in past 12 months	Borrower
pct_tl_nvr_dlq	Percent of trades never delinquent	Borrower
percent_bc_gt_75	Percentage of all bankcard accounts > 75% of limit.	Borrower
pub_rec_bankruptcies	Number of public record bankruptcies	Borrower
tax_liens	Number of tax liens	Borrower
tot_hi_cred_lim	Total high credit/credit limit	Borrower
total_bal_ex_mort	Total credit balance excluding mortgage	Borrower
total_bc_limit	Total bankcard high credit/credit limit	Borrower
total_il_high_credit_limit	Total installment high credit/credit limit	Borrower
fico_avg_score	Average fico score	Borrower
GDP_growth	GDP index zip code in %	Macro
GDP_index	GDP growth zip code in % (base level 2001)	Macro
avg_income_zip	Average Income per capita in zip code	Macro
avg_income_change_zip	Average Income change per capita in zip code	Macro
risk_premium	Loan interest rate minus the risk free rate	Macro
unemployment_rate	Unemployment rate in zip code	Macro
unemployment_change	Unemployment rate change in zip code	Macro
house_price_change	House price change in % in zip code	Macro
house_price_index	House price index in zip code (base level 2001)	Macro
consumer_confidence_index	Consumer confidence index USA	Macro
inflation	Inflation rate, i.e. Change in CPI	Macro
russel2000	Monthly Russel 2000 returns	Macro
VIX	Volatility index	Macro
policy_uncertainty_index	Policiy uncertainty index	Macro
S&P500	Montly S&P500 returns	Macro

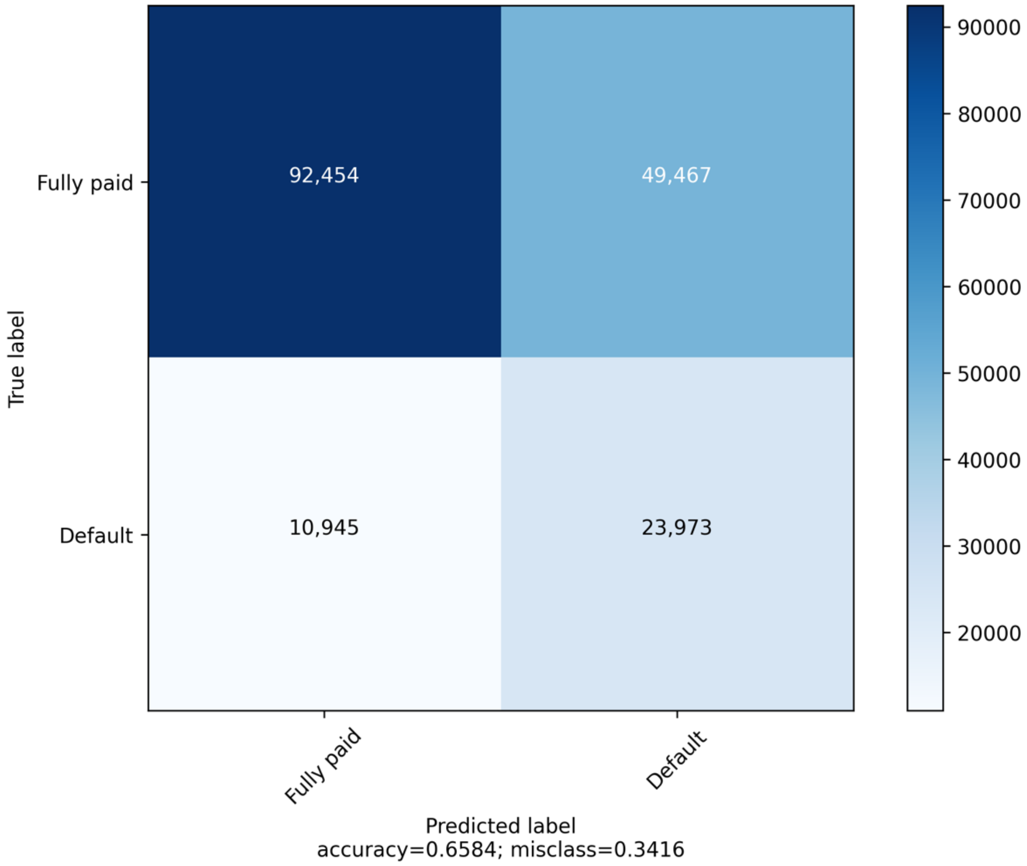
9.3. Appendix 3 – correlation matrix

The correlation matrix of the scaled, independent and dependent variables used in training and testing. Besides some clusters, e.g. loan amount and installment or 'number of' features, the correlation between the independent variables is relatively low.



9.4. Appendix 4 – confusion matrix

Confusion matrix of the best performing algorithm, i.e. neural network in combination with ROS. The results are based on the performance on the test set, similar to table 4 in chapter 5.1.


True label
Fully paid
Default
Fully paid
Default
Predicted label
accuracy=0.6584; misclass=0.3416