Applying Cost-Sensitive Machine Learning Models to Loan Default Prediction

Ben Hover
STUDENT NUMBER 2000240

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:

Supervisor: dr. R.J.C.M. Starmans
Second reader: dr. E. Fukuda

Word count: 8,594

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
January 2022

Acknowledgements

I wish to express my gratitude to my family, especially my father, for always supporting me, both mentally and financially. Their support was of great importance in completing not only this project, but more importantly the master's program as a whole. Furthermore, I want to give special appreciation to dr. Richard Starmans for providing excellent supervision throughout the entire process. Lastly, I would like to thank dr. Eriko Fukuda for the effort of being the second reader.

I hope you enjoy reading this thesis,

Ben Hover

Contents

List of Tables

List of Figures

List of Abbreviations

AUC Area Under the Curve

CS Cost-Sensitive

CSL Cost-Sensitive Learning

CS Cost-Sensitive

CV Cross-Validation

DT Decision Tree

FN False Negatives

FP False Positives

HPI House Price Index

LR Logistic Regression

ML Machine Learning

NPL Non-Performing Loan

RF Random Forest

SBA Small Business Administration

SHAP Shapley Additive exPlanations

TN True Negatives

TP True Positives

TPE Tree Parzen Estimator

U.S. United States

U.S.A. United States of America

VIX Volatility Index

XGBoost eXtreme Gradient Boosting

Abstract

This thesis investigated the predictive performance of logistic regression, decision tree, random forest and XGBoost for predicting loan default by using a dataset from the Small Business Administration. Random forest and XGBoost are state-of-the-art models and were compared against logistic regression and decision tree. The literature shows that correctly predicting non-performing loans continues to be difficult, as loan datasets typically suffer from the inherent class imbalance problem. Cost-sensitive learning is implemented to discover whether this method may prove to be an adequate solution to class imbalance. This thesis found that XGBoost exhibits the best results in loan default prediction. Furthermore, cost-sensitive learning considerably increases the number of correctly predicted non-performing loans in the logistic regression, decision tree and random forest. Moreover, this thesis explored the attribution of macroeconomic conditions and sectors (in addition to loan attributes) to loan default probabilities. This thesis established that both macroeconomic indicators and sectors improve predictive performance, with sectors contributing most.

Applying Cost-Sensitive Machine Learning Models to Loan Default Prediction

Ben Hover

1 Introduction

In this chapter, an overview of the background, purpose and scope of this thesis is provided.

1.1 Machine learning for loan default prediction

Firstly, machine learning (ML) based models have increasingly been adopted by numerous domains, including credit risk assessment. Potentially, the implementation of supervised ML techniques could improve traditional risk assessment because ML models offer a much broader view of a loan applicant than mere linear calculations of a small number of risk factors. Indeed, the traditional risk scoring model typically gives mixed and unreliable results (Ereiz 2019). Several ML algorithms have been investigated in the literature with regard to default prediction, including logistic regression, decision tree, random forest and XGBoost. Nevertheless, the highly imbalanced nature of loan data still remains a problem, as the cost of incorrectly predicting a fully paid loan is typically higher than the benefit of correctly doing so.

Secondly, ML models may prove a relevant remedy for predicting loan defaults for businesses that have little credit information available, such as small businesses. Due to this lack of information, information asymmetry arises and subsequently results in adverse selection and moral hazards (Cassar, Ittner, and Cavalluzzo 2015). Therefore, small businesses typically experience difficulties in accessing the credit market. The next section provides more information about the nature of small businesses and small business administration.

Thirdly, several empirical studies have confirmed that the number of non-performing loans (NPLs) is related to various key macroeconomic indicators. However, there is no consensus about the direction of the impact of these variables. In light of this lack of clarity, it was interesting to study the individual contributions of important macroeconomic indicators on the output of the ML models considered in this thesis. Rather than trying to choose among the large number of macroeconomic indicators a priori, the indicators were chosen from three classes: 1) general economic conditions (e.g. inflation, house price index and unemployment rate); 2) the direction in which the economy is moving (e.g. real GDP growth, policy uncertainty and business confidence); and 3) a set of indicators of the financial market conditions (e.g. risk premium rate, volatility index). Obviously, the sole use of macroeconomic variables is not optimal when predicting loan defaults; therefore, loan characteristics should also be included in ML models.

Lastly, knowledge on the persistence of shocks to specific sectors of the economy is relevant from a lender's point of view because such knowledge can help minimise the

impact of sector shocks to the number of NPLs. According to (Lee and Poon 2014), credit loss in the U.S. banking system is mainly caused by the real estate loan sectors. Gosh (2017) also found that the construction and agriculture sectors should be intensively monitored to reduce the number of NPLs. However, any business' credit risk is generally driven by shared risk factors affecting all sectors (Elizalde 2005). Consequently, it is relevant to investigate the attribution of the sectors as a whole to the model output and the individual contributions of the real estate, agriculture and construction sectors.

1.2 The Small Business Administration

A small business is formulated as privately owned corporation, partnership or sole proprietorship with fewer than 500 employees. The U.S. Small Business Administration (SBA) is a government agency devoted to stimulating economic growth and the development of small businesses and lenders across the U.S.—in part, through their loan programs. The agency does not lend directly to small business owners. Instead, it formulates guidelines for loans made by its partnering lending institutions. The SBA reduces the lending risk taken on by small business by guaranteeing a percentage of the loan. [1]

1.3 Purpose

This thesis focused on the performance of ML techniques for loan default prediction. An accurate model for default prediction is beneficial to both lenders and borrowers. On the one hand, an accurate prediction can prevent borrowers from taking on too high of a loan, thereby averting bankruptcy costs. Moreover, improving loan default prediction decreases the default risk, resulting in lower interest rates for borrowers. On the other hand, an accurate prediction ensures that lenders are better able to identify risky borrowers.

The first research question was developed to measure the performances of logistic regression (LR), decision tree (DT), random forest (RF) and XGBoost for loan default prediction. Specifically, it sought to confirm whether tree-based algorithms, which the literature has found to be effective for firms, is also effective for small business loans.

Research question 1:

> *"For a chosen set of machine learning algorithms, which algorithm exhibits the best performance in loan default prediction with respect to specific model evaluation metrics?"*

One novelty in this thesis is that the XGBoost model was compared to both cost-insensitive and cost-sensitive LR, DT and RF models to investigate whether the prediction of NPLs could be improved. This approach to loan data has not yet been used in the literature, and it is relevant since it could reduce the cost of false negatives, as the downside risk of default is up to 100%.

1 Source:https://www.sba.gov/

Sub-question 1.1:

"What is the effect of indirect cost-sensitive learning with cost-proportionate weights on the predictiveness of non-performing loans?"

Also, the existing literature has found statistically significant relationships between macroeconomic variables and the number of NPLs. However, the predictive capability of macroeconomic variables with regard to the binary classification performance of machine-learning based techniques has remained untouched. Consequently, the capability of macroeconomic variables in predicting loan default was investigated through research question 2.

Research question 2:

"To what extent is loan default prediction by machine-learning-based approaches attributable to macroeconomic variables?"

In addition, inflation, real GDP growth, business confidence and house price index tend to negatively relate to the number of non-performing loans as has been found in existing literature. However, the significance of these results differs across studies, therefore the direction of the impact of macroeconomic variables remains inconclusive. To examine whether these indicators are also negatively related to loan default probabilities, the relationship between these indicators and the probability of default is examined through sub-question 2.1.

Sub-question 2.1:

"How are real GDP growth, business confidence index, and house price index related to loan default probabilities?"

Furthermore, inflation, unemployment, economic policy uncertainty, risk premiums and the volatility index tend to positively relate to the number of NPLs in the empirical literature. However, the statistical significance of this direction differs across studies. Sub-question 2.2 was developed to explore the direction of the impact of these macroeconomic determinants on loan default probabilities.

Sub-question 2.2:

"How are inflation, unemployment, economic policy uncertainty, risk premiums and the volatility index related to loan default probabilities?"

As mentioned above, there is no consensus in the literature about the role of sector-specific risk on the number of NPLs. Rather, some studies insist on the systematic risk present across all sectors, whereas other studies emphasise the sector-specific credit risk. For that reason, the effect of sectors on loan default prediction was studied through research question 3.

Research question 3:

"To what extent is loan default prediction by machine-learning-based approaches attributable to sectors?"

Finally, to study the direction of the impact of the real estate, construction and agriculture sectors on loan default probabilities, the contribution of these sectors to loan default prediction was individually assessed.

Sub-question 3.1:

"How are the real estate sector, the construction sector and the agriculture sector related to loan default probabilities?"

1.4 Scope

The scope of this thesis was to investigate how various supervised ML techniques affect loan default prediction. The model evaluation metrics of special interest in this thesis were precision, recall, F_1-score, AUC score and Kappa measure. The classifiers used were as follows:

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost

The XGBoost and RF ensemble methods have proven to outperform the decision tree and logistic regression model in literature. However, due to the high interpretability and explainability of the results of LR and DT, it is relevant to include these models in this thesis.

1.5 Main findings

This thesis found that the XGBoost model exhibited the best performance with regard to loan default prediction. Next, this thesis has extended existing literature in two ways. First, applying a cost-sensitive approach to logistic regression, decision tree and random forest increases the number of correctly-identified NPLs. Second, loan default prediction is attributable to sectors and macroeconomic indicators. Furthermore, the results of the LR have shown that the agriculture and real estate sectors significantly positively relate to loan default probabilities. Meanwhile, the construction sector was not significantly related to loan default probabilities. This thesis also found that unemployment, the growth of real GDP, risk premiums, the house price index and the volatility index were significantly positively related to loan default probabilities. Similar to existing literature, the relationship between inflation and loan default probabilities was ambiguous, as the coefficient of inflation, even though significant, was approximately zero.

2 Related Work

In this chapter, small business lending is briefly discussed. Thereafter, the LR, DT, RF and XGBoost models and their performances in default prediction are discussed. Lastly, the importance of macroeconomic variables is reviewed, and the sector-specific and common risks concerned with loan default prediction are considered.

2.1 Credit risk assessment of small businesses

First of all, a distinguishing characteristic of the credit markets that supply small businesses loans is that they suffer from information and agency problems that are caused by a lack of information on the creditworthiness of small businesses. This limitation is mainly caused by the fact that small firms often do not have audited financial statements and are likely not monitored by credit rating agencies (Ortiz-Molina and Penas 2008). Credit scoring is a statistical approach to predicting the loan default of a loan applicant. Even though this method of assessment is firmly established in consumer credit markets, it has only been applied to small businesses' credit evaluation for a relatively short period of time.

2.2 Loan default prediction using machine learning models

In early approaches to statistical loan default prediction, the focus was mainly on linear classifiers, such as LR, where the predictors are used in the model in a linear combination. Indeed, LR is a benchmark model in the field of credit risk, especially since the lack of interpretability of ensemble methods conflicts with the needs for credit assessment (Dumitrescu et al. 2021). More specifically, for linear models such as LR, the coefficients and their significance show which predictors are important determinants for default prediction, whereas ensemble methods are often characterised as a 'black box' (Xia et al. 2021). More recently, nonlinear methods such as DT, RF and XGBoost have been proposed for default prediction. In contrast to linear models, tree-based models can learn nonlinearities, discontinuities and complex interactions.

Moreover, since DT, RF and XGBoost use trees as base learners, they are resistant to outliers in predictors and scale invariant to monotonic transformations for predictors (Sigrist and Hirnschall 2019). Next, tree-based learners is that their predictive performance is not decreased by the issue of multicollinearity, whereas linear learners cannot deal with high correlation between predictors (Kruppa et al. 2013). Besides, decision-tree-based classifiers can predict loan defaults via automated iterations without manual intervention, which has both practical value for loan prediction (Zhou et al. 2019). In many prediction tasks for imbalanced and high-dimensional data, these algorithms have achieved desirable prediction results. Accordingly, decision-tree-based algorithms have proven to be advanced algorithms developed for loan default prediction in recent years. Recently, though, extreme gradient boosting has become one of the state-of-the-art models. This relatively novel approach is known for its rapidness, efficiency and capability of parallelization.

2.3 Cost-sensitive learning

In loan default prediction, predicting positive cases as negative is more harmful than vice versa. Therefore, it is relevant to investigate how the minority class, that is, defaulted loans, can be better detected by ML models. (Zadrozny, Langford, and Abe 2003) propose cost-sensitive learning (CSL) to approach the common problem of class imbalance in loan default prediction. In CSL, there is a penalty associated when an observation is misclassified; this penalty is referred to as cost. The aim of CSL is to minimise the total cost of misclassification (i.e. the sum of the cost of incorrectly predicting the negative class and the cost of incorrectly predicting the positive class).

Indirect CSL aims to minimise misclassification costs by assigning weights to classes in the training set. There are two empirically set methods for assigning weights: weighting by inverse class frequency and weighting by a smoothed version of the inverse square root of class frequency (Cui et al. 2019). The class frequency is based on the relative number of examples of the negative and positive class. Thus, the cost of incorrectly predicting the negative class will be relatively high, as the inverse frequency of the positive class will be higher than the inverse frequency of the negative class. The smoothed version of weighting has resulted in better performances than weighting by inverse class frequency; therefore, this method is adopted (Cui et al. 2019).

2.4 Macroeconomic indicators for non-performing loans

The relationship between the number loan defaults and macroeconomic conditions has been discussed by connecting the economic boom and bust cycles with the financial distress of businesses. Intuitively, if the economy is growing fast, it is naturally in a better state than if it is declining. (Hussain, Khalil, and Nawaz 2013) and (Makri, Tsagkanos, and Bellas 2014) argue that an increase in GDP growth negatively relates to the number of defaults as the ability of firms to repay their loans is high. Still, the economy is able to grow most quickly when there is insufficient demand relative to what the economy is capable of producing, known as a 'period of economic slack' (Figlewski, Frydman, and Liang 2012). Therefore, from the perspective of (Figlewski, Frydman, and Liang 2012), it is less obvious that rapid growth in GDP should necessarily be associated with low risk of default. According to (Marcucci and Quagliariello 2008), there is a relationship between the change in business cycle and NPLs. More precisely, during economic booms, lenders tend to increase their lending activity in combination with relaxing their selection criteria, thus leading to fewer NPLs. In contrast, during downturns, the lending restrictions are tightened remarkably, resulting in an increase of NPLs. In addition, it is argued by (Figlewski, Frydman, and Liang 2012) that the unemployment rate is one of the most visible indicators of the health of an economy. High unemployment adversely affects income and thereby reduces an economy's output.

The impact of inflation on the number of NPLs is rather ambiguous. On the one hand, higher inflation can make repaying debt easier by diminishing the real value of loans; however, it can also diminish the borrowers' real income when wages are sticky-down (Klein 2013). That is, the nominal wages of employees of a borrowing firm are above the equilibrium because employees resist nominal wage cuts. Moreover, the impact of inflation depends on whether the interest rates are variable, as higher inflation can cause interest rates to increase due to the contractionary monetary policy strategy to limit the increase of inflation (Tanasković and Jandrić 2015). As a result, an increase in interest rates boost the costs of lending, thus increasing the risk of loan

default. Moreover, a rise in home prices boosts financial wealth and may help borrowers facing unexpected adverse shocks, or such a rise eases their access to credit by increasing the value of the houses used as collateral (Ghosh 2017). Consequently, positive changes in the housing price index (HPI) are expected to diminish the number of NPLs – in particular, for the real estate sector. However, an increase in the HPI accompanied by higher inflation may partially extinguish the decrease in the number of NPLs, as higher inflation diminishes the real value of houses.

Furthermore, (Vouldis and Louzis 2018) argue that the business confidence index, which is an indicator of business sentiment, is negatively related to the number of NPLs. On the other hand, economic policy uncertainty, which reflects the frequency of articles containing negative sentiments in U.S. newspapers, positively relates to the number of NPLs (Karadima and Louri 2021). According to (Karman et al. 2016), there is also a significant positive relationship between the risk premium rate and the number of NPLs. Another primary determinant of NPLs is the unemployment rate, which is significantly positively related to the number of NPLs (Vouldis and Louzis 2018). The inflation rate, which measures the change in the prices of a basket of goods and services, can be either positively or negatively related to the number of NPLs (Nkusu 2011). The volatility index (VIX), which measures the volatility of the U.S. stock market, also positively relates to the number of NPL (Makri, Tsagkanos, and Bellas 2014).

2.5 Sector-specific and common credit risk factors in non-performing loans

In various empirical studies, the role of economy-wide risks in relation to default risks have been frequently discussed. However, a scarce number of studies have investigated the influence of sector-specific risks. (Lee and Poon 2014) found that in addition to economy-wide credit risk, the real estate sector is one of the main contributors to the credit risk in the U.S. banking system. This significant contribution is mainly due to the subprime crisis of 2008, where the real value of houses dropped dramatically.

Next, an increase in real GDP has been proven to significantly reduce NPLs for the real estate, construction and agriculture sectors (Ghosh 2017). Moreover, increases in the HPI diminishes NPLs for all sectors in general – again with real estate, construction and agriculture showing the highest sensitivity.

Conversely, there is also evidence that firms' default probabilities are driven most by a number of common macroeconomic risk factors. Unfortunately, most of these risk factors are unobservable and therefore difficult to incorporate into ML models. (Elizalde 2005) shows that a sole common risk factor accounts for more than 50% of the business' credit risk levels, with an average of 68% across firms. Another common and observable factor is the evolution of the structure of interest rates, which is also associated with the credit risk of the firms' sector of activity. Nevertheless, this association seems to be less influential than the unobservable risk factors.

3 Methods

In this chapter, the characteristics of the ML models are provided, the evaluation metrics that are of interest are discussed and the robustness test is explained.

3.1 Logistic regression

Logistic regression is a linear supervised learning algorithm that is used to calculate the probability of a loan to default. The probability is calculated by taking the sigmoid function of a linear combination of predictors and results in a value ranging from 0 to 1. The predicted probability of X belonging to the positive class:

$$P(y = 1|X) = \frac{1}{1 + e^{-(a+b^T X)}} \tag{3.1}$$

where e is Euler's number, a is an unknown parameter, b is an unknown vector with parameters and X is a predictor vector.

The goal of the LR model is to optimise the unknown parameters a and b so that Eq. 3.1 will give an output as close to 1 as possible for the predicted values that are correctly-identified as members of the positive class. The LR is of interest in this thesis because the results of this algorithm show the coefficients of the predictors with their respective significance. Thus, the contribution of the predictors on the probability of default can be individually assessed, which is an advantage over ensemble methods. Both the DTs and LR use decision boundaries to separate classes. However, a DT is able to divide the decision space into increasingly smaller regions with nonlinear decision boundaries, whereas logistic LR only fits a single linear boundary to divide the decision space (Kim 2016).

3.2 Decision tree

Unlike LR, DTs are nonlinear classifiers that identify ways to classify observations based on how previous questions are answered. A DT is a tree-like graph: The base of the tree is the root node, and from the root node flows a sequence of decision nodes showing possible decisions and finally resulting in a leaf node with the predicted class label (e.g. a defaulted loan). For classification, the Gini impurity is a widely used measure to choose the best predictor to split on at each step in building the tree. The Gini impurity measures the probability of misclassifying an observation. A DT splits the nodes on all predictors and then chooses the variables with the largest decrease in Gini impurity.

$$I_G = \sum_{i=0}^{1} p(i) * (1 - p(i)) \tag{3.2}$$

where I_G is the probability of misclassifying an observation, i denotes the class and $p(i)$ is the probability of observing an observation of the i-th class.

A DT has some advantages when classifying loans. First, the interpretation of the results in a DT is simple. Once the DT has been fitted, new observations can be rapidly

predicted by only using a number of if-then statements. This attribute is useful for addressing loan default prediction since the DT can provide insight into the relationship between the predictors and the target. Second, since a DT is nonparametric, it is able to generate if-then statements which do not require any implicit assumption about the underlying distributions of the variables. Third, a DT is capable of learning the nonlinear behavior of predictors, which is relevant since loan data predictors typically behave in a nonlinear way (Zhao and Zou 2021).

Even though introducing nonlinearity usually increases complexity, a DT is able to implement nonlinearity in a interpretable manner. Furthermore, overfitting of the DT can be avoided by decreasing the size of the DT, this technique was used in this thesis. DTs are useful for loan default prediction since they are intuitive and interpretable, compared to the ensemble models. However, a DT is more sensitive to data patterns than LR, as any minor change in the training set can drastically change the tree and significant change in the predictions.

3.3 Random forest along with hyperparameters

Similar to the DT, RFs are nonlinear classifiers. RF is based on an ensemble technique of many individual DTs, also known as bagging. In bagging, a random subset of the training data is created with replacement. For classifications, all individual trees independently make predictions of the output class and the majority class is chosen as output value. The majority vote can be calculated as follows:

$$\bar{f}(X) = sign(sign(\sum_{i=0}^{T} f_i(x))) \tag{3.3}$$

Where f_i represents the *i*-th decision tree and sign denotes the signum function, that is, sign: $[0, \infty) \rightarrow \{0, 1\}$, $sign(x) = 0 \Leftrightarrow x = 0$

As mentioned in section 3.2, each internal node of a DT determines the decrease in the Gini impurity of each predictor. The predictor with the largest decrease in impurity is selected for the internal node. In RFs, each individual tree is grown to the largest extent possible. If each individual tree is unstable, the aggregated classifier has a smaller variance compared to the individual trees. Therefore, RFs are less prone to overfitting than individual DTs. Moreover, due to bootstrapping and the random predictor selection, RFs achieve uncorrelated trees. Tuning the number of variables randomly sampled at each split provides the biggest improvement of the AUC (Probst, Wright, and Boulesteix 2019). Therefore, the number of variables at each split has been tuned on the validation set, with the evaluation measure AUC, the evaluation strategy 10-fold cross-validation and the search domain ranging from 1 to 15 variables. The remaining hyperparameters are based on the default settings in R. The main advantage of this ensemble technique is that it produces better predictive performance than an individual DT, as the combination of weak learners form a stronger learner. Therefore, an RF typically outperforms a single DT. However, the RF model is less interpretable than the LR and DT models.

3.4 XGBoost along with hyperparameters

Gradient boosting (GB) is a method for creating an ensemble by combining weak learners into a stronger learner. More clearly, it starts by fitting a DT, and then it builds a second model that focuses on the poorly predicted cases of the previous model, continuing to do so for many models. The rationale behind this technique is that each successive model corrects for the shortcomings of all previous models, thus creating a more powerful model. The XGBoost algorithm is the more advanced version of gradient boosting, as it also incorporates regularization, which improves the model's generalisation capabilities (Wang and Ni 2019).

According to (Wang and Ni 2019), the optimisation of XGBoost can be explained as follows. Let the dataset be denoted as $D = \{x, y\}$ with n observations, where x and y are the predictors and the response variable, respectively. Let $\hat{y}_i$ be the prediction and y_i the actual value of the *i*-th instance at the *b*-th boost, having $l(y_i, \hat{y}_i)$, which measures the difference between the predicted and the real value. f_b represents a DT q with leaf j having a weight measure of w_j. The regularization term $\Omega(f_b)$ penalises the model's complexity. The hyperparameter γ indicates the minimum loss reduction required to make a split. On condition that the loss reduction is smaller than γ, XGBoost stops adding trees, thus reducing the complexity. λ is a fixed coefficient, T indicates the number of leaves that are contained by the tree and $||w||^2$ constitutes the L2 regularization norm of the weight of the leaf. Comparable with γ, the hyperparameter w_{mc} reduces the model's complexity by controlling the depth of the tree, where a large w_{mc} causes the model to be more conservative in splitting. Another manner in which XGBoost avoids overfitting is the column subsampling. It is considered that column subsampling is more effective in avoiding overfitting than the traditional row subsampling used in GB (Bergstra and Bengio 2012). The goal of XGBoost is to minimise the loss function L_b given in Eq. 3.4 below.

$$L_b = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{b=1}^{B} \Omega(f_b) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \gamma T + 0.5\lambda ||w||^2 \tag{3.4}$$

Since various studies have shown that the hyperparameters can be efficiently optimised by adopting the Bayesian Tree Parzen Estimators (TPE), I also used this method for hyperparameter optimisation (Hutter, Hoos, and Leyton-Brown 2011)(Thornton et al. 2013). TPE relies on Bayes' Theorem given in Eq. 3.5, to direct the search for the minimum of the objective function:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.5}$$

Since TPE is used to optimise the objective function instead of calculating the conditional probability, the normalizing value P(B) can be removed (Wang and Ni 2019). Instead, the conditional probability can be formulated as a proportional quantity as in Eq 3.6:

$$P(A|B) = P(B|A)\, P(A) \tag{3.6}$$

The search domains of the hyperparameters in Table 1 were based on the proposition of (Wang and Ni 2019). The remaining hyperparameters were based on the default settings in R. The hyperparameters were tuned on the validation set, with the evaluation metric AUC and the evaluation strategy 10-fold cross-validation.

Table 1: Search domains of hyperparameters in XGBoost

Hyperparameter	Description	Domain
Eta (η, learning rate)	Step size shrinkage of predictor weights	(0.005, 0.2)
Subsample	Ratio of observations to be randomly sampled for each tree	(0.8, 1)
Max_depth	Maximum depth of a tree	(5, 30)
Gamma (γ)	Minimum loss reduction required for further partition	(0, 0.02)
Colsample_bytree	Ratio of features used for fitting an individual tree	(0.8, 1)
Min_child_weight(w_{mc})	Minimum sum of weights of all observations required in a child	(0, 10))

While both RF and XGBoost use an ensemble method of DTs, the XGBoost model provides better predictive performance. First, XGBoost considers the similarity between the parent node and its child nodes and stops increasing the depth of the tree when the gain from a node is found to be minimal, thus reducing the complexity of the model. In contrast, the RF potentially overfits the data if trees are provided with similar samples. Second, when XGBoost fails to predict the minority class for the first time, it gives more weight to the minority class in the upcoming iterations. Meanwhile, each tree in an RF is based on a random sample from the data. Consequently, each DT will be biased in the same direction by class imbalance.

3.5 Evaluation metrics

In this thesis, the designations true positive (TP) and false positive (FP) denote correctly and wrongly classified NPLs, respectively. Similarly, true negative (TN) and false negative (FN) indicate correctly and wrongly classified loans paid in full, respectively. Accuracy is a widely used metric in binary classifications. However, because the dataset used in this thesis is imbalanced, accuracy can be a misleading metric (Vuttipittayamongkol, Elyan, and Petrovski 2021). As considered in (Begueria 2006), recall and precision are weighted more heavily than accuracy in the risk-modelling domain. At the same time, recall is weighted more heavily than precision because a false negative error may signify a loss in loan default prediction (Wang and Ni 2019). Moreover, the F_1-score is another relevant measure in this thesis since it is the harmonic mean of precision and recall.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.7}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.10}$$

$$F_1 - score = 2 * \frac{precision * recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3.11}$$

Finally, the Kappa coefficient is a measure of agreement between the predictions and the actual values. Practically, Kappa removes the possibility of agreement between the classifier and a random guess and calculates the number of predictions that cannot be explained by randomness. The Kappa measure is relevant for ranking models when using imbalanced data (Fatourechi et al. 2008).

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{3.12}$$

where p_0 is the number of agreed cases divided by the total number of cases and p_e is the probability of random agreement for both the prediction and the real value.

The confusion matrix, which is formatted in Table 2, provides the accuracy, precision, recall, specificity, F_1-score and Kappa coefficient. The confusion matrix of the best performing model is given in Figure 1 in chapter 5.

Table 2: Confusion matrix

		Actual	
		Fully paid	Defaulted
Predicted	Fully paid	TN	FN
	Defaulted	FP	TP

3.6 Receiver operating characteristic curve and area under the curve

The receiver operating characteristic (ROC) curve shows the tradeoff between the false positive rate (FPR) and the true positive rate (TPR). [2] The ROC curve does not emphasise one class over the other, thus is not biased against the minority class. The area under the curve (AUC) is the quantitative evaluation metric of the ROC curve, which ranges from 0 to 1. A larger AUC value indicates a better classification result and therefore a better capability of distinguishing between fully paid and defaulted loans. When AUC is approximately equal to 0, the model is predicting a negative class as a positive class and vice versa. Since one of the goals of this thesis is to minimise the number of incorrectly predicted negative classes, it is relevant to include the AUC metric. Also, according to (Wardhani et al. 2019), the AUC is a robust metric for measuring imbalanced data. When the decision threshold varies, the AUC can be calculated by using Eq. 3.13 as was proven by (Bradley 1997).

$$AUC = \sum_i (1 - \beta_i \Delta\alpha) + \frac{1}{2}[\Delta(1-\beta)\Delta\alpha] \tag{3.13}$$

where α_i and $(1 - \beta_i)$ denote the FPR and TPR for threshold i, respectively.

$$\Delta(1-\beta) = (1-\beta_i) - (1-\beta_{i-1}),$$
$$\Delta\alpha = \alpha_i - \alpha_{i-1}$$

3.7 Indirect cost-sensitive learning by cost-proportionate case weighting

To address class imbalance, cost-sensitive learning is included in this thesis. The advantage of this technique is that the original data remains untouched, whereas other techniques such as oversampling the minority class introduce bias (Kim, Kwon, and Paik 2019). To implement indirect cost-sensitive learning, weights should be assigned to the positive and negative class in the training data. The weights of the negative class and positive class are given in Eq. 3.14 and 3.15, respectively:

$$w_0 = \sqrt{\frac{\sum_{i=0}^{1} n_i}{n_0}} \tag{3.14}$$

2

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + TN}$$

$$w_1 = \sqrt{\frac{\sum_{i=0}^{1} n_i}{n_1}} \tag{3.15}$$

where the number of observations of the i-th class in the training data is denoted by n_i, $i = 0$ denotes the negative class and $i = 1$ represents the positive class.

The re-weighted classes were used in the LR, DT and RF models. In XGBoost, class weighting was already accounted for in the model itself. Including these class weights in the binary cross-entropy function of LR gives the weighted cost function stated in Eq. 3.16:

$$Cost_{LR} = \frac{1}{n} \sum_{i=1}^{n} -w_1 y_i \log(\hat{y}_i) - w_0 \left(1 - y_i\right) \log(1 - \hat{y}_i) \tag{3.16}$$

where y_i and $\hat{y}_i$ are the actual class and the predicted class of the i-th example, respectively.

The number of examples is denoted as n. Regarding the DT and RF, the weights are incorporated in the process of splitting the nodes. The weight of the observations in a potential child node c is:

$$t_c = \sum_{i=0}^{1} w_i * \ n_i \tag{3.17}$$

where n_i is the number of examples of the i-th class in child node c and w_i is the weight assigned to the i-th class.

In addition, the impurity of child node c is formulated in Eq. 3.18:

$$I_c = 1 - \sum_{i=0}^{1} (\frac{w_i \ * \ n_i}{t_c})^2 \tag{3.18}$$

Finally, the impurity of the entire split is:

$$I_t = \sum_{c} \frac{t_c}{t_p} * \ I_c \tag{3.19}$$

where t_p is the total weight of all examples in the parent node that is being split.

3.8 SHAP values

The SHAP values method is used to quantify how much individual predictors contribute to XGBoost predictions (Meng et al. 2021). Where positive SHAP values indicate a positive impact on prediction, the resulting model predicts the positive class. In the same way, negative SHAP values represent a negative impact, leading the model to predict the negative class. This method is of interest in thesis since it increases the interpretability of XGBoost.

3.9 Robustness test

First, the model with the highest performance on the dataset including all variables had to be determined. Thereafter, three robustness checks were done to confirm whether the inclusion of macroeconomic indicators and sectors considerably improved the predictive performance. The first check only used loan attributes to train and test the best model. Second, another check was done with data that included loan attributes and macroeconomic indicators. Finally, the last check was done by training and testing data containing loan attributes and sectors. Consequently, the robustness of the individual contributions of macroeconomic variables and sectors was checked.

4 Experimental Setup

This chapter contains information about the origin of the data, the processing of the data, the exploratory data analysis, the software packages used and the manner in which the data was split.

4.1 Data description

The SBA loan dataset used in this thesis originated from the U.S. Small Business Administration and is publicly available on Kaggle.[3] The SBA dataset, which includes historical data from 1987 through 2014, contains 897,994 unique loans with 27 variables. The target variable, which shows whether a loan was paid in full or defaulted, was treated as a dummy variable (0 = paid in full, 1 = defaulted). 157,558 loans were charged off, and 739,609 loans were paid in full (i.e. a default rate of approximately 18%).

The unemployment rate, unemployment rate change, GDP growth rate, the economic policy uncertainty index, volatility index and the house price index were collected from the Federal Reserve Economic Data website.[4] The risk premium rate on lending (lending rate minus treasury bill rate) and inflation rate were collected from the World Bank website.[5] The business confidence index was retrieved from the website of the Organization for Economic Co-operation and Development.[6] The descriptive statistics of the loan variables and macroeconomic indicators are given in Table 3.

The predictors can be divided into three groups: loan, sector and macroeconomic indicators. The loan variables were determined by the loan characteristics, such as the loan amount. The sector variables show which sector the loan was distributed to. The macroeconomic variables contain information about general economic conditions (inflation, house price index, unemployment); information relating to the direction in which the economy is moving (real GDP growth, policy uncertainty and business confidence); and a set of indicators of the financial market conditions (risk premium, volatility index). The full description of all variables is given in Appendix A.

3 Source: https://www.kaggle.com/kerneler/starter-should-this-loan-be-approved-b812e39d-a/data
4 Source: https://fred.stlouisfed.org/
5 Source: https://data.worldbank.org/
6 Source: https://data.oecd.org/

Table 3: Descriptive statistics of loan attributes and macroeconomic indicators

Variable	Mean	SD	Min	25%	Median	75%	Max
Term	83.75	58.91	1	51	85	85	411
NoEmp	8.33	31.57	0	2	8.33	8	8,000
CreateJob	2.05	14.33	0	0	0	2	5,085
RetainedJob	5.68	20.08	0	1	2	6	7,250
DisbursementGross	161,787	285,373	4,000	27,800	63,000	160,000	1,144,635
GrAppv	138,842	268,902	1,000	25,000	50,000	120,000	5,000,000
SBA_Appv	101,853	220,986	500	12,500	25,000	75,000	4,500,000
GDP growth (%)	4.68	2.08	-2.00	3.70	4.80	6.40	6.70
Risk premium (%)	3.52	0.23	2.90	3.10	3.20	3.50	3.60
Unemployment rate (%)	5.67	1.55	4.00	4.60	5.10	5.80	9.60
Unemployment rate change (%)	0.14	0.95	-1.20	-0.50	0.00	0.30	3.50
Inflation (%)	2.73	0.92	-0.40	2.30	2.90	3.40	3.80
VIX	19.93	6.63	12.39	12.81	17.54	24.20	32.70
House price index	331.40	44.78	182.60	312.60	349.10	374.10	378.30
Policy uncertainty index	92.38	36.20	58.00	62.40	73.00	130.60	155.60
Business confidence index	99.89	1.42	95.70	99.30	99.90	100.80	102.20

4.2 Data processing

Initially, all sector codes were translated to the corresponding sector names. Thereafter, since XGBoost cannot handle categorical variables, the sectors were one-hot encoded, that is, each sector was converted into a new column and was assigned a binary value of 0 or 1. Three variables containing the bank name, the state in which the business is situated and the state in which the bank is situated are also categorical; however, one-hot encoding these would have resulted in a considerable increase in dimensions and therefore an increase in overfitting. In addition, the relative differences between states and between banks are not of interest in this thesis, therefore these three categorical variables were removed. Furthermore, variables that were not of interest were removed – namely, loan ID, city, zip code, disbursement data, approval date and charged off date. Furthermore, the missing values were removed, and the (relative) number of missing values is given in Appendix B. Finally, the macroeconomic indicators that are of interest in this thesis were added column wise for similar years to the SBA dataset, thereafter the final dataset consisted of 362,970 observations and 40 variables.

4.3 Exploratory Data Analysis

One of the aims of this thesis was to investigate the individual contribution of macroeconomic indicators and sectors. However, when there are correlations between predictors, a problem may arise in determining the impact of predictors on the response variable. On the one hand, the precision of the estimated coefficients is reduced. On the other hand, the confidence intervals will be wide (Paul 2006). Therefore, it was relevant to determine whether there were strong correlations between predictors. The correlation matrix, given in Appendix C, provided insight into the correlations between predictors. There were strong correlations between the loan variables concerning the loan amount

(i.e. SBA_Appv, GrAppv, DisbursementGross), to avoid (multi-)collinearity, only the disbursement gross was kept for classification. Furthermore, there was a strong correlation between the change in unemployment rate and the growth rate of real GDP. For that reason, the change in unemployment rate was also not considered for classification.

4.4 Software packages

In this thesis, I used the programming language R and the program RStudio version 4.1.1 to perform the data manipulation and ML. The packages used are listed in Table 4 below.

Table 4: R packages

Package	Usage
Caret	Machine learning techniques
Dplyr	Data manipulation
Fastshap	SHAP Values
Ggplot2	Visualization purposes
Metan	Correlation matrix
ParBayesianOptimization	Bayesian Tree Parzen Estimator
pROC	ROC-curve and AUC
Randomforest	Random forest algorithm
Readr	Importation of CSV file
Rpart	Decision tree
Rpart.plot	Decision tree plot
Tidyr	Data manipulation
XGBoost	XGBoost algorithm

4.5 Training and test set

For training and evaluation of the ML models, it is necessary to split the data. In this thesis, the data was split: 80% of the data was used for training, and 20%, for the test set. The training set was used to fit the model parameters, and the test set was used for model evaluation. With regard to the RF and XGBoost models, 10% of the training set was used to tune the hyperparameters. The training set contained 290,376 observations, and the test set contained 72,594 observations. For comparability purposes, all models were trained on the same training set and tested on the same test set. This was achieved by fixing the seed (i.e. a pseudo-random number) before the data was split. Cross-validation (CV) was used to assess the generalisability of the ML models and to prevent overfitting (Berrar 2019). In this thesis, due to the class imbalance, stratified CV was used to preserve the class distribution in each fold (Purushotham and Tripathy 2011). Considering the large size of the dataset, 10-fold CV was adopted.

5 Results

In this chapter, the tuned values of the hyperparameters are given, the results from the ML models on the test are presented, the significance of the results are discussed, and the robustness of the findings is tested.

5.1 Hyperparameter values of random forest and XGBoost

First, the tuned number of variables to randomly sample at each split in RF was 6. Table 5 contains the tuned values of hyperparameters in XGBoost. The tuned values for the hyperparameters of RF and XGBoost were used in the training of the models.

Table 5: XGBoost's tuned settings for hyperparameters

Hyperparameter	Tuned value
Eta (η, learning rate)	0.170
Subsample	0.971
Max_depth	9
Gamma (γ)	0.001
Colsample_bytree	0.838
Min_child_weight(w_{mc})	5

5.2 Classification with all predictors

The results of the classifications on the test set are displayed in Table 6 below, where LR, DT and RF are also used with weighted classes. Considering the metrics of most interest in this thesis, XGBoost exhibited the best predictive performance with regard to recall, F_1-score and Kappa. However, although accuracy and specificity are both excellent for cost-insensitive models, precision is relatively low, indicating that the positive class is poorly predicted. CSL improves precision for all models, which shows that CSL makes ML models more reliable in classifying samples as positive. Since precision and recall are inversely proportional to each other, recall diminishes when adopting CSL, which results in more false negatives. Still, the F_1-score is higher for cost-sensitive LR and DT compared to cost-insensitive LR and DT, which means that the total number of incorrect classes is reduced by CSL for these models. CSL also improves Kappa and AUC for LR and DT. This shows better agreement between predictions and actual values and better capability of distinguishing between classes, respectively. All the same, CSL proves to be a suitable method to increase the number of true positives, albeit at the cost of having more false negatives.

Table 6: Classification results of ML models with the best scores of each metric in bold

Algorithm	Classes	Accuracy	Precision	Recall	F_1-score	Specificity	Kappa	AUC
LR	Standard	0.860	0.411	0.623	0.495	0.889	0.418	0.681
LR	Weighted	0.847	0.664	0.533	0.592	0.929	0.499	0.774
DT	Standard	0.868	0473	0.646	0.546	0.899	0.472	0.863
DT	Weighted	0.829	**0.858**	0.494	0.627	0.967	0.527	0.886
RF	Standard	0.901	0.700	0.707	0.703	0.940	0.644	**0.937**
RF	Weighted	0.884	0.793	0.619	0.696	0.956	0.625	0.934
XGBoost	Weighted	**0.940**	0.842	**0.806**	**0.824**	**0.968**	**0.788**	0.901

5.2.1 Confusion matrix of XGBoost

The confusion matrix of XGBoost in Figure 1 shows all predictions in combination with the actual classes. Even though XGBoost takes class imbalance into account, a considerable share of the NPLs was still predicted incorrectly: Approximately 20% of all defaulted loans were predicted as fully paid.

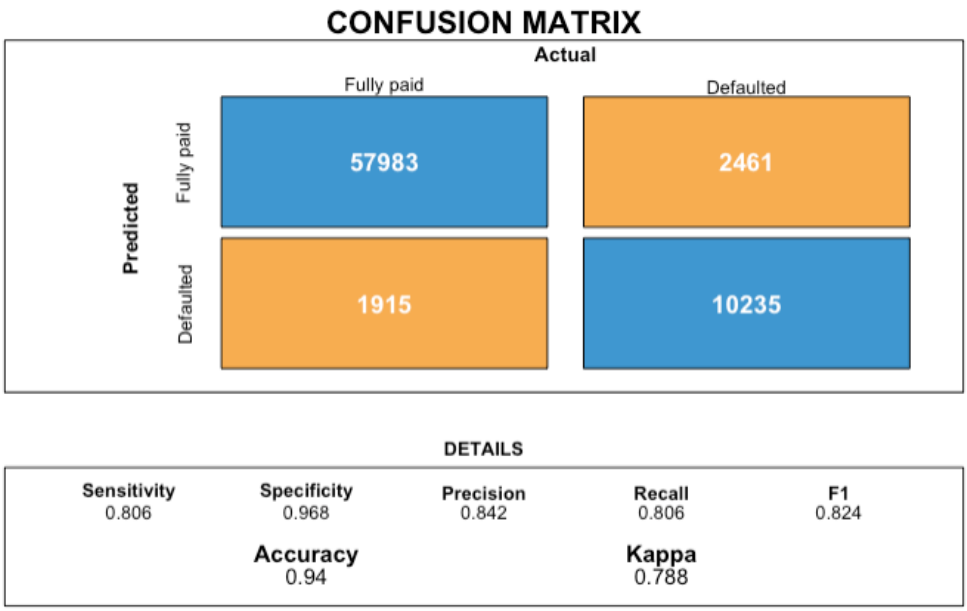


Figure 1: Confusion matrix of XGBoost

5.2.2 Prediction with logistic regression

The results of the CS-LR were considered to test the magnitude and significance of the individual contributions displayed in Table 7. The magnitude of the contribution was investigated through the estimated coefficient and the significance through the p-value. For continuous predictors, the estimated coefficient (β_i) is the expected change in the odds of default by a factor of e^{β_i} per unit change in that predictor.[7] For binary predictors, changing predictor j from the reference group 0 to 1 changes the estimated odds of default by a factor of e^{β_j}. To judge whether a coefficient was statistically significant, the p-value was compared to the significance criterion (α) that was set at 5%. When the p-value was smaller than α, the coefficient estimate was statistically significant. The intercept is interpreted assuming a value of 0 for all the predictors, which was not realistic in this thesis, therefore the intercept was omitted from Table 7. The loan attributes were all statistically significant. With regard to the macroeconomic indicators, the unemployment rate, real GDP growth, risk premium rate, house price index and volatility index were all positively related to loan default probabilities. The most important contribution, with respect to magnitude, was given by the risk premium rate. The relationship between default probabilities and inflation, even though significant, was ambiguous because the coefficient estimate was approximately zero. With regard to the sectors of interest, agriculture and real estate were significantly positively related to loan default probabilities, whereas construction was not significantly related.

7

$$\text{Odds of default} = \frac{\text{probability of default}}{1 - \text{probability of default}}$$

Table 7: Summary of logistic regression output

Variable	Estimate	SE	95% confidence	P-value
ApprovalFY	-0.250	0.011	(0.762, 0.796)	0.000*
Term	-0.037	0.000	(0.963, 0.964)	0.000*
NoEmp	-0.005	0.000	(0.995, 0.996)	0.000*
NewExist	-0.183	0.007	(0.821, 0.845)	0.000*
CreateJob	0.003	0.000	(1.002, 1.003)	0.000*
RetainedJob	-0.013	0.000	(0.986, 0.989)	0.000*
UrbanRural	-0.408	0.010	(0.654, 0.677	0.000*
RevLineCr	-0.720	0.007	(0.479, 0.493)	0.000*
DisbursementGross	0.000	0.000	(1.000, 1.000)	0.000*
Agriculture	0.332	0.115	(1.114, 1.745)	0.004*
Mining	-0.848	0.136	(0.328, 0.560)	0.000*
Construction	0.170	0.112	(0.951, 1.477)	0.129
Manufacturing	-0.027	0.113	(0.780, 1.213)	0.808
Wholesale	0.131	0.113	(0.914, 1.421)	0.246
Retail	0.367	0.112	(1.159, 1.799)	0.001*
Transportation	0.029	0.113	(0.825, 1.284)	0.800
Information	0.234	0.114	(1.010, 1.581)	0.041 *
Finance	0.424	0.114	(1.221, 1.912)	0.000*
Estate	0.552	0.114	(1.389, 2.169)	0.000*
Professional	0.007	0.112	(0.808, 1.255)	0.948
Management	0.288	0.238	(0.836, 2.126)	0.227
Administrative	0.112	0.113	(0.897, 1.394)	0.322
Education	0.212	0.116	(0.985, 1.551)	0.068
Health	-0.445	0.113	(0.514, 0.800)	0.000*
Recreation	0.288	0.114	(1.066, 1.670)	0.012*
Accommodation	0.471	0.112	(1.284, 1.996)	0.000*
Other	0.199	0.112	(0.979, 1.521)	0.076
Public	0.233	0.209	(0.838, 1.902)	0.264
UnemploymentRate	0.179	0.019	(1.152, 1.241)	0.000*
GDPgrowth	0.122	0.019	(1.114, 1.197)	0.000*
Inflation	-0.072	0.014	(0.920, 0.972)	0.000*
UncertaintyIndex	0.000	0.000	(0.999, 1.001)	0.837
Riskpremium	1.206	0.044	(3.066, 3.641)	0.000*
BusinessConfidence	0.023	0.000	(0.997, 1.050)	0.080
HousePriceIndex	0.029	0.001	(1.028, 1.030)	0.000*
VIX	0.070	0.004	(1.062, 1.078)	0.000*

*Indicates significant at $\alpha = 0.05$

5.2.3 Prediction with decision tree

As mentioned above, two advantages of a DT are the interpretability and rapid manner in which new observations can be classified. The optimal number of splits was 3, as can been observed in Appendix D. As can be seen in Figure 2, the term and house price index were the most important variables for splitting the nodes. The nodes contained the positive or negative class, the probability of default and the percentage of observations. At the root node, the overall probability of loan default was given, which was 29%. The

root node asked whether the term was greater than 79 months. If so, then the next node was the child node down on the left. This leaf node shows that 53% of the loans had a term greater than 79 months, with a probability of default of 6%. This process was similar for the other nodes.

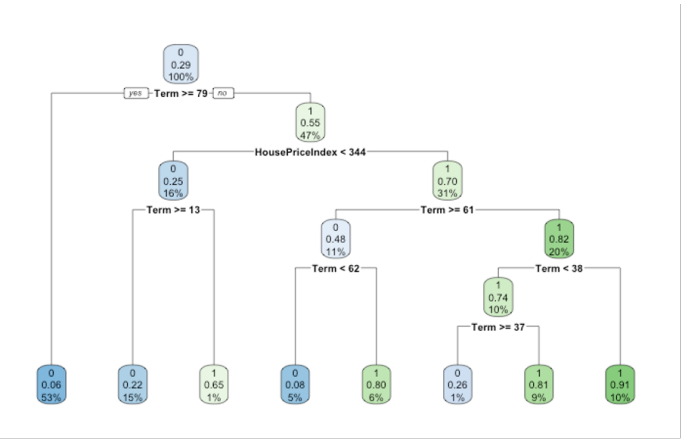


Figure 2: Graphical depiction of pruned DT

5.2.4 SHAP values for XGBoost

The SHAP values show the average of the marginal contribution of predictors to prediction considering all possible order of the contributors arrival. As can be observed in Figure 3 below, the HPI had a contribution of approximately 0.48. More specifically, the inclusion of HPI on average increased the log-odds ratio by approximately 0.48. With regard to the other macroeconomic indicators, only the uncertainty index seemed to have some (negative) impact on default prediction; other macroeconomic contributions to prediction were negligible. With respect to the sectors, no individual sector had a considerable contribution to prediction.

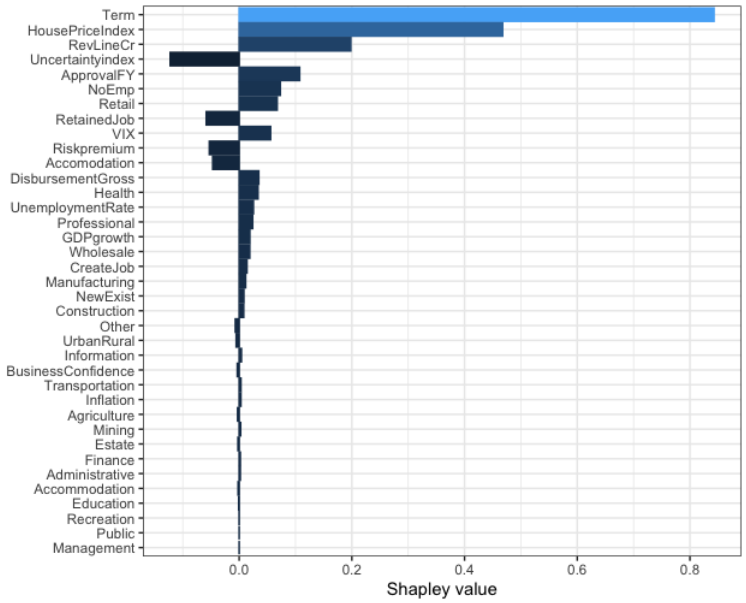
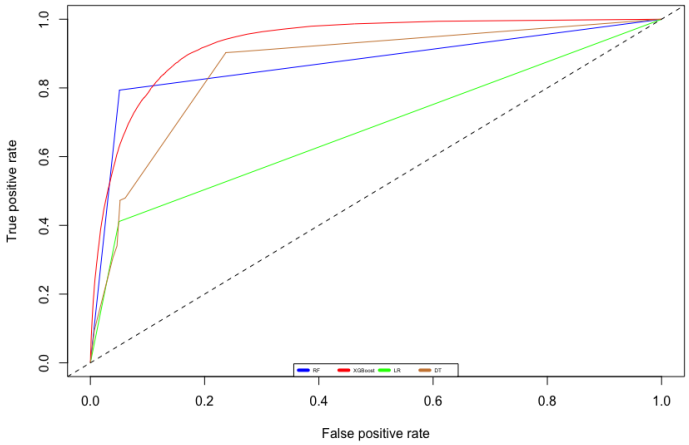


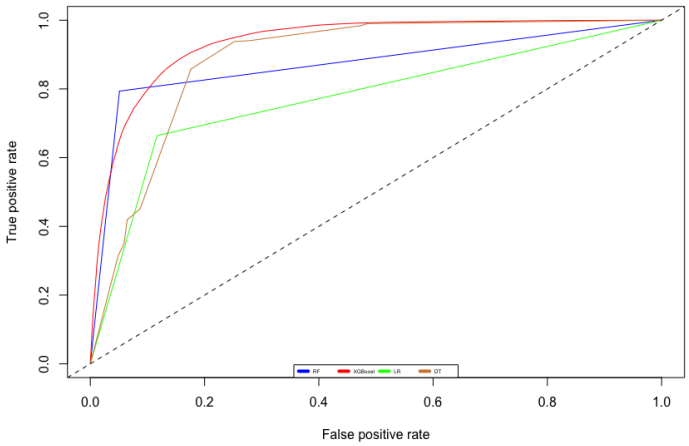
Figure 3: SHAP values of individual predictors.

5.2.5 Evaluation in the ROC-plane

As mentioned earlier, the amount of True Positives is relevant, as the costs of not detecting a loan default are high. From the ROC-curves (Figure 4), it can be observed that the LR had the worst performance for both cost-insensitive and cost-sensitive models, as LR had a lower True Positive Rate (TPR) than the other models. However, the TPR considerably increased due to the CSL approach. With respect to the DT and RF, the effect of CSL also increased the TPR, thereby increasing the AUC.



(a) Cost-insensitive



(b) Cost-sensitive

Figure 4: ROC-curves.

5.3 Results of robustness test

In Table 8, the results of the robustness test are given. The predictive performance of XGBoost increased when macroeconomic indicators or sectors were added to loan attributes. The strongest improvement occurred when sectors were added. So, the findings in literature that suggest that macroeconomic variables and sectors have a relationship with the number NPLs are similar to the findings of this thesis, as the predictive performance of XGBoost was improved by including these variables.

Table 8: Robustness test for XGBoost model, with different sets of predictors and best scores in bold

Predictors included	Accuracy	Precision	Recall	F_1-score	Specificity	Kappa	AUC
Loan attributes	0.928	0.842	0.756	0.797	0.967	0.753	0.894
Loan attr. + macro	0.938	0.839	0.799	0.818	0.967	0.781	0.896
Loan attr. + sectors	**0.940**	**0.855**	**0.801**	**0.827**	**0.970**	**0.791**	**0.901**

6 Discussion

In this chapter, the results of the classifications in the context of existing literature are discussed. The contributions of this thesis to existing literature is given as well as the limitations and recommendations for further research.

6.1 Classification performance

The differences in performance measures of classifications by LR, DT, RF and XGBoost are in line with the findings in the existing literature. (Kruppa et al. 2013)and (Zhou et al. 2019) found that tree-based algorithms outperform the benchmark LR model in loan default prediction, which is consistent with the results of the classification results in this thesis. Moreover, (Malekipirbazari and Aksakalli 2015) proposed cost-sensitive approach for class imbalance was found to positively influence the prediction of correctly identified NPLs. However, the number of incorrectly predicted fully paid loans increased in this method. Furthermore, it is suggested that there is a significant positive relationship between real GDP growth and the number of NPLs (Hussain, Khalil, and Nawaz 2013);(Makri, Tsagkanos, and Bellas 2014). Comparable with these findings, this thesis found that real GDP was significantly positively related to loan default probabilities. Lastly, similar to the findings of (Lee and Poon 2014) and (Ghosh 2017), the agriculture and the real estate sectors were both significantly positively related to loan default probabilities.

6.2 Contributions to literature

This thesis contributes to existing literature through its robust results relating to the cost-sensitive learning method in cost-insensitive ML models for loan default prediction, which have been untouched by past studies. Also, this thesis extended literature by investigating the attribution of important macroeconomic variables and risky sectors. The relationships between macroeconomic indicators and the number of NPLs and between several sectors and the number of NPLs had already been discussed in literature. This thesis extended literature by investigating the relationships between these variables and loan default probabilities. The results of the robustness test showed that these variables are indeed important with regard to loan default prediction. Moreover, the sector-specific risk was assessed, considering the attribution of sectors to the prediction of loan default probabilities, which showed that sectors substantially contribute to the predictive performance of XGBoost.

6.3 Limitations and recommendations for further research

The foremost limitation of this thesis is that the SBA loan data did not contain any information about the incentives for loan approval or loan rejection. This deficit limited the ability of this thesis to consider the loan application process. Also, the SBA dataset only runs to 2014, thereby limiting the generalisation of the results to the present. For further research on loan default prediction, this thesis suggests three directions. First, a future study should investigate whether the results of this thesis can be generalised to small businesses in countries outside the United States. Second, the optimisation of cost-sensitive learning for loan data might be useful for better detecting NPLs. Lastly,

research on the inherent risks of sectors could be expanded, possibly leading to a better understanding of the riskiness of sectors.

7 Conclusion

In this chapter, the research questions of this thesis are restated and answered.

Research question 1:

"For a chosen set of machine learning algorithms, which algorithm exhibits the best performance in loan default prediction with respect to specific model evaluation metrics?"

In conclusion, the XGBoost model exhibited the best predictive performance for loan default prediction. The other state-of-the-art model, the RF model, outperformed LR and the DT, as has also been found in the existing literature. At the same time, the DT outperformed LR, which is also similar to the findings in literature.

Sub-question 1.1:

"What is the effect of indirect cost-sensitive learning with cost-proportionate weights on the predictiveness of non-performing loans?"

Cost-sensitive learning considerably increased precision compared to cost-insensitive models for the LR, DT and RF models, thereby increasing the ability of ML models to detect true positives. With regard to CS-LR and CS-DT, the F_1-score, Kappa measure and AUC also improved.

Research question 2:

"To what extent is loan default prediction by machine-learning-based approaches attributable to macroeconomic variables?"

Sub-question 2.1:

"How are real GDP growth, business confidence index, and house price index related to loan default probabilities?"

Sub-question 2.2:

"How are inflation, unemployment, economic policy uncertainty, risk premiums and the volatility index related to loan default probabilities?"

The inclusion of macroeconomic indicators improved the predictive performance of XGBoost. With regard to the individual contributions of macroeconomic indicators, the unemployment rate, real GDP growth, risk premium rate, house price index and volatility index were all significantly positively related to loan default probabilities.

In contrast, the impact of inflation was negligible; nonetheless, it had a statistically significant impact. Meanwhile, business confidence and economic policy uncertainty were not significantly related to loan default probabilities.

Research question 3:

"To what extent is loan default prediction by machine-learning-based approaches attributable to sectors?"

Sub-question 3.1:

"How are the real estate sector, the construction sector and the agriculture sector related to loan default probabilities?"

The contribution of sectors improved the predictive performance of XGBoost. Also, considering the individual contributions of agriculture, construction and real estate sectors, which have been suggested by the literature, to bear more risk of default, only agriculture and real estate are significantly positively related to loan default probabilities.

References

Begueria, Santiago. 2006. Validation and evaluation of predictive models in hazard assessment and risk management. *Natural Hazards*, 37(3):315–329.

Bergstra, James and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

Berrar, Daniel. 2019. Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, page 542–545.

Bradley, Andrew P. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Cassar, Gavin, Christopher D Ittner, and Ken S Cavalluzzo. 2015. Alternative information sources and information asymmetry reduction: Evidence from small business debt. *Journal of Accounting and Economics*, 59(2-3):242–263.

Cui, Yin, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Dumitrescu, Elena-Ivona, Sullivan Hué, Christophe Hurlin, et al. 2021. Machine learning or econometrics for credit scoring: Let's get the best of both worlds.

Elizalde, Abel. 2005. Do we need to worry about credit risk correlation? *The Journal of Fixed Income*, 15(3):42–59.

Ereiz, Zoran. 2019. Predicting default loans using machine learning (optiml). In *2019 27th Telecommunications Forum (TELFOR)*, pages 1–4, IEEE.

Fatourechi, Mehrdad, Rabab K Ward, Steven G Mason, Jane Huggins, Alois Schloegl, and Gary E Birch. 2008. Comparison of evaluation metrics in classification applications with imbalanced datasets. In *2008 seventh international conference on machine learning and applications*, pages 777–782, IEEE.

Figlewski, Stephen, Halina Frydman, and Weijian Liang. 2012. Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Economics & Finance*, 21(1):87–105.

Ghosh, Amit. 2017. Sector-specific analysis of non-performing loans in the us banking system and their macroeconomic impact. *Journal of Economics and Business*, 93:29–45.

Hussain, Altaf, Ambar Khalil, and Maryam Nawaz. 2013. Macroeconomic determinants of non-performing loans (npl): Evidence from pakistan. *Pakistan Journal of Humanities and Social Sciences*, 1(2):59–72.

Hutter, Frank, Holger H Hoos, and Kevin Leyton-Brown. 2011. Sequential model-based optimization for general algorithm configuration. In *International conference on learning and intelligent optimization*, pages 507–523, Springer.

Karadima, Maria and Helen Louri. 2021. Economic policy uncertainty and non-performing loans: The moderating role of bank concentration. *Finance Research Letters*, 38:101458.

Karman, Hafiz Waqas, Shameer Malik, Hamas Butt, M Hamza, Umair Afzal, and Shahzaib Maqbool. 2016. Risk premium and its effect on bank's non-performing loans. *International Journal of Innovation and Economic Development*, 1(6):79–89.

Kim, Kyoungok. 2016. A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. *Pattern Recognition*, 60:157–163.

Kim, Young-geun, Yongchan Kwon, and Myunghee Cho Paik. 2019. Valid oversampling schemes to handle imbalance. *Pattern Recognition Letters*, 125:661–667.

Klein, Nir. 2013. *Non-performing loans in CESEE: Determinants and impact on macroeconomic performance*. International Monetary Fund.

Kruppa, Jochen, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. 2013. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13):5125–5131.

Lee, Yongwoong and Ser-Huang Poon. 2014. Forecasting and decomposition of portfolio credit risk using macroeconomic and frailty factors. *Journal of Economic Dynamics and Control*, 41:69–92.

Makri, Vasiliki, Athanasios Tsagkanos, and Athanasios Bellas. 2014. Determinants of non-performing loans: The case of eurozone. *Panoeconomicus*, 61(2):193–206.

Malekipirbazari, Milad and Vural Aksakalli. 2015. Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10):4621–4631.

Marcucci, Juri and Mario Quagliariello. 2008. Is bank portfolio riskiness procyclical?: Evidence from italy using a vector autoregression. *Journal of International Financial Markets, Institutions*

and Money, 18(1):46–63.
Meng, Yuan, Nianhua Yang, Zhilin Qian, and Gaoyu Zhang. 2021. What makes an online review more helpful: an interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(3):466–490.
Nkusu, Ms Mwanza. 2011. *Nonperforming loans and macrofinancial vulnerabilities in advanced economies*. International Monetary Fund.
Ortiz-Molina, Hernan and Maria Fabiana Penas. 2008. Lending to small businesses: The role of loan maturity in addressing information problems. *Small Business Economics*, 30(4):361–383.
Paul, Ranjit Kumar. 2006. Multicollinearity: Causes, effects and remedies. *IASRI, New Delhi*, 1(1):58–65.
Probst, Philipp, Marvin N Wright, and Anne-Laure Boulesteix. 2019. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301.
Purushotham, Swarnalatha and BK Tripathy. 2011. Evaluation of classifier models using stratified tenfold cross validation techniques. In *International Conference on Computing and Communication Systems*, pages 680–690, Springer.
Sigrist, Fabio and Christoph Hirnschall. 2019. Grabit: Gradient tree-boosted tobit models for default prediction. *Journal of Banking & Finance*, 102:177–192.
Tanasković, Svetozar and Maja Jandrić. 2015. Macroeconomic and institutional determinants of non-performing loans. *Journal of Central Banking Theory and Practice*, 4(1):47–62.
Thornton, Chris, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2013. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855.
Vouldis, Angelos T and Dimitrios P Louzis. 2018. Leading indicators of non-performing loans in greece: the information content of macro-, micro-and bank-specific variables. *Empirical Economics*, 54(3):1187–1214.
Vuttipittayamongkol, Pattaramon, Eyad Elyan, and Andrei Petrovski. 2021. On the class overlap problem in imbalanced data classification. *Knowledge-based systems*, 212:106631.
Wang, Yan and Xuelei Sherry Ni. 2019. A xgboost risk model via feature selection and bayesian hyper-parameter optimization. *arXiv preprint arXiv:1901.08433*.
Wardhani, Ni Wayan Surya, Masithoh Yessi Rochayani, Atiek Iriany, Agus Dwi Sulistyono, and Prayudi Lestantyo. 2019. Cross-validation metrics for evaluating classification performance on imbalanced data. In *2019 international conference on computer, control, informatics and its applications (ic3ina)*, pages 14–18, IEEE.
Xia, Yufei, Yinguo Li, Lingyun He, Yixin Xu, and Yiqun Meng. 2021. Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending. *Electronic Commerce Research and Applications*, 49:101095.
Zadrozny, Bianca, John Langford, and Naoki Abe. 2003. Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE international conference on data mining*, pages 435–442, IEEE.
Zhao, Selena and Jiying Zou. 2021. Predicting loan defaults using logistic regression. *Journal of Student Research*, 10(1).
Zhou, Jing, Wei Li, Jiaxin Wang, Shuai Ding, and Chengyi Xia. 2019. Default prediction in p2p lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and its Applications*, 534:122370.

Appendix A: Description of variables

Description of variables

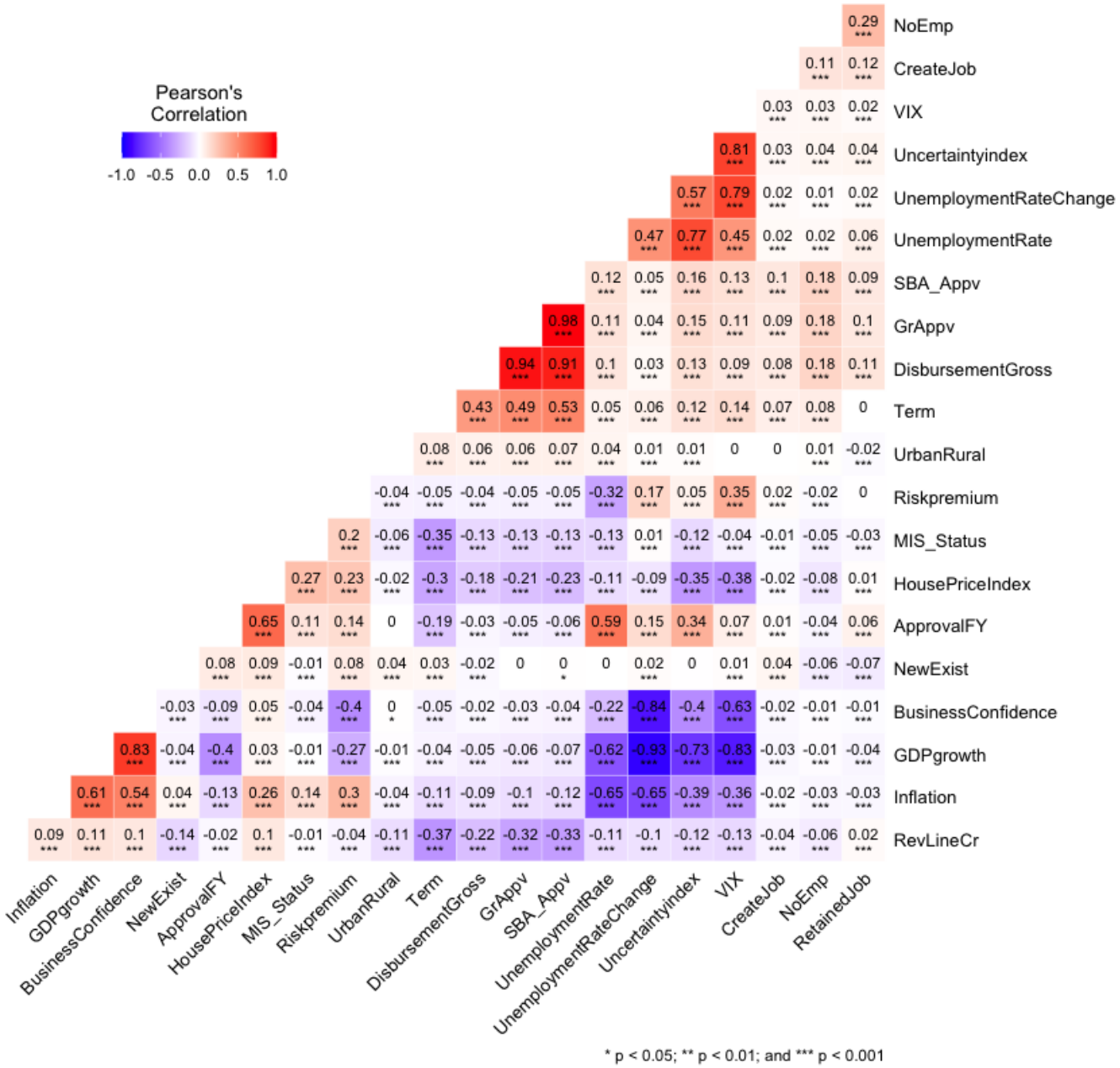
Abbreviation	Description	Type
ApprovalFY	Year of loan approval	Loan
Term	Loan term in months	Loan
NoEmp	Number of employees before loan	Loan
NewExist	0 = existing, 1 = new	Loan
CreateJob	Number of jobs created	Loan
RetainedJob	Number of jobs retained	Loan
UrbanRural	0 = urban, 1 = rural	Loan
RevLineCr	0 = repeated credit, 1 = new credit	Loan
DisbursementGross	Loan amount in dollars	Loan
GrAppv	Loan amount in dollars approved by bank	Loan
SBA_Appv	Loan amount in dollars guaranteed by SBA	Loan
MIS_Status	0 = paid in full, 1 = defaulted	Loan
Agriculture	0 = non-agricultural, 1 = agriculture	Sector
Mining	0 = non-mining, 1 = mining	Sector
Construction	0 = non-construction, 1 = construction	Sector
Manufacturing	0 = non-manufacturing, 1 = manufacturing	Sector
Wholesale	0 = non-wholesale 1 = wholesale	Sector
Retail	0 = non-retail, 1 = retail	Sector
Transportation	0 = non-transportation, 1 = transportation	Sector
Information	0 = non-informational, 1 = information	Sector
Finance	0 = non-financial, 1 = finance	Sector
Estate	0 = non-estate, 1 = estate	Sector
Professional	0 = non-professional, 1 = professional	Sector
Management	0 = non-managerial, 1 = management	Sector
Administrative	0 = non-administrative, 1 = administrative	Sector
Education	0 = non-educational, 1 = education	Sector
Health	0 = non-health, 1 = health	Sector
Recreation	0 = non-recreational, 1 = recreation	Sector
Accommodation	0 = non-accommodation, 1 = accommodation	Sector
Other	0 = defined sector, 1 = other	Sector
Public	0 = non-public, 1 = public	Sector
UnemploymentRate	Annual unemployment rate (%)	Macro
UnemploymentRateChange	Change in annual unemployment rate (%)	Macro
GDPgrowth	Annual growth of real GDP (%)	Macro
Inflation	Annual inflation rate (%)	Macro
UncertaintyIndex	Risk of future government policies	Macro
Riskpremium	Loan interest rate minus risk-free rate	Macro
BusinessConfidence	Business confidence towards future economic situation	Macro
HousePriceIndex	Price development of single-family property prices	Macro
VIX	Volatility of the stock market	Macro

Appendix B: Information about missing values in loan data

Information about missing values in loan data

Variable name	N	%
ApprovalFY	18	0.00
Term	0	0.00
NoEmp	0	0.00
NewExist	136	0.02
CreateJob	0	0.00
RetainedJob	0	0.00
UrbanRural	0	0.00
RevLineCr	4528	0.50
DisbursementGross	0	0.00
MIS_Status	1997	2.20
GrAppv	0	0.00
SBA_Appv	0	0.00

Appendix C: Correlation matrix of loan variables and macroeconomic variables. Except for some clusters (e.g. disbursement gross and amount approved), the correlation between predictors is relatively low



Appendix D: Complexity plot for decision tree, where the bottom horizontal axis shows the complexity level, the vertical axis shows the error rate, and the top horizontal axis shows the number of splits