NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Peer-to-peer lending: evaluation of credit risk using machine learning

Francisca Viçoso Vila Verde

Dissertation presented as partial requirement for obtaining the Master's degree in Statistics and Information Management specialised in Risk Management and Analysis

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

PEER-TO-PEER LENDING: EVALUATION OF CREDIT RISK USING MACHINE LEARNING

by

Francisca Viçoso Vila Verde

Dissertation presented as partial requirement for obtaining the Master's degree in Statistics and Information Management specialised in Risk Management and Analysis

Advisor: Professor Dr. Jorge Miguel Ventura Bravo

May 2021

ABSTRACT

Peer-to-peer lenders have transformed the credit market by being an alternative to traditional financial services and taking advantage of the most advanced analytics techniques. Credit scoring and accurate assessment of borrower's creditworthiness is crucial to managing credit risk and having the capacity of adapting to current market conditions. The Logistic Regression has long been recognised as the benchmark model for credit scoring, so this dissertation aims to evaluate and compare its capabilities to predict loan defaults with other parametric and non-parametric methods, to assess the improvement in predictive power between the most modern techniques and the traditional models in a peer-to-peer lending context. We compare the performance of four different algorithms, the single classifiers Decision Trees and K-Nearest Neighbours, and the ensemble classifiers Random Forest and XGBoost against a benchmark model, the Logistic Regression, using six performance evaluation measures. This dissertation also includes a review of related work, an explanation of the pre-processing involved, and a description of the models. The research reveals that both XGBoost and Random Forest outperform the benchmark's predictive capacity and that the KNN and the Decision Tree models have weaker performance compared to the benchmark. Hence, it can be concluded that it still makes sense to use this benchmark model, however, the more modern techniques should also be taken into consideration.

KEYWORDS

Machine Learning; Credit Scoring; Peer-to-peer lending; Ensemble methods; Single classifiers.

INDEX

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AUC Area Under the ROC Curve
B-Net Bayesian network
BS Brier Score
DT Decision Tree
FN False Negative
FP False Positive
GBM Gradient boosting machines
GD Grid Search
H H-measure
IPO Initial Public Offering
KNN K-Nearest Neighbour
KS Kolmogorov-Smirnov
LC Lending Club
LDA Linear Discriminant Analysis
LOG Logistic Regression
LS-SVM Least-squares Support Vector Machine
MS Manual Search
NB Naïve Bayes
NN Neural Network
P2P Peer-to-peer
PCC Percent Correct Classification
PGI Product Gini Index
QDA Quadratic Discriminant Analysis
RMSE Root Mean Square Error
ROC Receiver Operating Characteristic Curve
RS Random Search
SVM Support Vector Machine
TAN Tree augmented naive Bayes
TN True Negative
TP True Positive
TPE Tree-structured Parden estimator
UK United Kingdom
US United States
XGBoost Extreme gradient boosting

1. INTRODUCTION

Machine learning is defined as "the computer's ability to learn without being explicitly programmed" (Samuel, 1959). Advancement in this area has had a massive impact on the financial world as it promises to unbundle core functions such as settling payments, performing maturity transformation and sharing risk, along with an enhanced allocation of capital due to new participants' entrance such as payment service providers, aggregators, and Robo-advisors, peer-to-peer lenders, and innovative trading platforms. Together with the support of FinTech, the big data revolution, and Machine learning developments, the financial services industry has experienced a great deal of disruption (Bachmann et al., 2011).

The term "FinTech" encompasses technology-enabled services and solutions with the use of integrated IT. FinTech payment innovations offer a new landscape in the digital era of the financial industry. FinTech also provides online platforms for banks and non-banks to facilitate cross-network transfers and payment services (Shim & Shin, 2015; Thompson, 2017).

Online peer-to-peer lending describes the loan origination process between private individuals on online platforms where the financial services institutions operate only as intermediates required by law. Initialised by groups in online social networks, the first commercial online P2P lending platforms started in 2005 (Munusamy et al., 2013). Commercial and non-commercial are the two types of P2P lending platforms. The main difference between them is the lender's general intention and expectations concerning returns.

While lenders seek opportunities to invest money in the most profitable way possible at a certain level of risk, borrowers with different default risks look for liquidity sources. P2P platforms act as intermediaries and bring these groups together. They try to match the expectations of both parties. Lenders or borrowers sometimes engage in groups and form small communities to concentrate their interests (Wang & Greiner, 2011; Herrero-Lopez, 2009).

Peer-to-peer lending gained more relevance in 2010 due to the trend of growing trust in online transactions, increasing consumer expectations of immediacy and proliferation of public data and the use of analytics for credit scoring purposes (Tomlinson et al., 2016). Peer-to-peer lending impacts the way consumers and small to medium enterprises can access credit.

Its development is quite visible. However, the volume is distant from the numbers of the traditional sector. It is still unclear how the industry will reshape the financial landscape in the long-term impact (Jagtiani & Lemieux, 2018). Governments and regulators have become informed of the implications for these organisations' economy, and regulation has more than doubled its magnitude in the US, UK, and China between 2010 and 2015 (Aveni et al., 2015). The traditional retail banks have been taking one of two different approaches: (i) backing these organisations or (ii) creating similar solutions.

In some cases, these organisations perceive them as competitors; however, they also recognise the potential of these innovations. For instance, this dissertation's analyses one of the biggest peer-to-peer lenders in the US, Lending Club, where the retail banking sector represents more than 25% of total clients. Lending Club is a peer-to-peer lending company that matches borrowers with investors through an online platform—serving people that need personal loans between $1,000 and $40,000. Borrowers receive the entire amount of the issued loan minus an origination fee paid to the corporation. Investors purchase notes backed by personal loans and pay Lending Club a service fee.

The most appealing aspect is the democratisation of financial services (Herzenstein & Andrews, 2008). Since consumers will make more informed decisions and have access to better-targeted service, the P2P lenders will empower them, increasing the transparency and giving control to lenders. The level of diversity will grow as small and medium-sized businesses will get access to new credit. Therefore, P2P enhances the financial stability of the sector.

On the other hand, system risks will evolve due to new underwriting models, consequently changing credit quality and even macroeconomic dynamics. These risks are associated with credit intermediation, including maturity transformation, leverage, and liquidity mismatch. Despite not facing material systemic risk due to their small scale and business models, it is essential to understand the extent of the peer-to-peer lending growth and how these types of risks are considered (Carney, 2017).

Anderson (2007) defines credit scoring as using statistical models to transform relevant data into numerical measures that guide credit decisions, which is an ever-evolving area. It is crucial to adapt to the most recent developments and take advantage of the available data.

A well-performing credit scoring model is pivotal for P2P lending (Polena, 2017). Therefore, the use of new forms of data from behavioural data to social metric data, biometric data, psychometric data, and even social media data can be the key to improve the credit scoring models and ensure the survival of this marker.

This industry faces several challenges that restrain the growth of the market. Including the existence of fraudulent activities, difficulty in evaluating the creditworthiness of borrowers due to both sides meeting anonymously through the internet (Nigmonov et al., 2020), and most recently, the economic downturn caused by the COVID19 pandemic, creating the need to understand the dynamics of financial distress in the P2P market since this industry did not exist in the last worldwide financial crisis.

As mentioned before, the COVID-19 pandemic has brought several additional challenges to risk managers. They have to understand the pandemic's impact on credit and market portfolios to mitigate their operations' negative effects. Redesigning underwriting can be a solution to improve efficiency and effectiveness and, according to Arroyo et al. (2020), can represent a ten to twenty-five percent improvement in the accuracy of underwriting predictions through advanced analytics.

Anderson (2007) defined default as the failure to honour financial commitments, in this case, a loan. Calculating the probability of default is one of those critical subjects, especially after a wave of defaults affected the Chinese market in 2018. Causing the withdrawal of funds by investors, which resulted in the reduction of Chinese peer-to-peer lenders to only twenty-nine from about six thousand and, according to Zhu et al. (2020)—translated into a loss of more than US$ 115bn.

We aspire to infer if changing the credit scoring models' paradigm for more modern techniques creates value for the stakeholders through this research.

This dissertation aims to model the default probability by training five different models and comparing it with the traditional credit scoring model, the Logistic Regression, by calculating six evaluation metrics. Most of the related studies did not have such a broad spectrum of historical data and complete dataset and the pre-processing approach that aims to achieve state-of-the-art results. The models being compared are the single classifiers – Logistic Regression (LG) and Decision Trees (DT) and the ensemble classifiers – Random Forest (RF), K-Nearest Neighbours (KNN) and Extreme Gradient Boosting (EGB). Several studies have adopted them in traditional loan default prediction. We contribute to the literature by analysing their performance in P2P credit risk classification.

The importance, explanation of the concepts and impact on the accuracy of hyperparameter tuning were also analysed in this study, together with the influence of imbalanced datasets, which is a common problem in credit scoring problems (see, e.g., Ashoften and Bravo (2021a,2021b) and references therein).

The R statistical software was used to perform the exploratory analysis, whereas Python statistical software was used in the remaining steps. This includes using the Scikit-learn library to train (Random Forest, the KNN and the Decision Trees), and the XGBoost package.

Regarding the thesis outline, Section 1 introduces peer-to-peer lending and the credit scoring models and explains the subject's relevance. Section 2 reviews the relevant literature, where the most important papers are discussed, and similar research results are presented.

The third section starts with a detailed explanation of the algorithms used in this study and how their predictive performance will be evaluated. Proceeded by an explanation on experimental design and a detailed description of the dataset plus all the steps conducted during the exploratory analysis, data pre-processing, and data training.

In the fourth section, the performance evaluation metrics will be reviewed. Section followed by the presentation of the empirical results, and the conclusions are drawn from the results in section five. Lastly, the reporting of the limitations of this research and recommendations for future works will be examined. Bibliographic references are included at the end of the manuscript.

2. LITERATURE REVIEW

2.1.1. A Glimpse from Parallel Studies

This dissertation revolves around two different themes: peer-to-peer lending and credit scoring models. Therefore, this research will analyse both subjects.

Single-period classification techniques to classify borrowers into diverse risk categories and to estimate the probability of default continue to be the most used data mining techniques in this field (Chamboko & Bravo, 2020). Altman (1968) first approached this theme by developing the Z-score discriminant analysis model, based on five financial ratios, to predict corporate bankruptcies. His work was followed by the development of diverse techniques based on a combination of traditional statistical methods and more advanced modelling approaches that make use of individual classifiers, homogenous and heterogenous ensembles in addition to alternative predictive features. Chamboko and Bravo (2016, 2019a, 2019b) explore various ways of modelling and forecasting recurrent delinquency and recovery events on consumer loans using survival analysis, including the underestimation of credit losses due to the classical assumption of independence of default event times and multiple defaults.

The majority of the proposals focus on introducing new classification algorithms, performance measures, statistical hypotheses and attempts to minimise the decision-relevant costs. Comparing the performance of parametric and non-parametric techniques is also a critical subject. Over the years, since this subject gained relevance, several approaches have been implemented. In the following paragraphs, the most relevant studies will be analysed and summed up in the end.

In addition to quantitative research that studies the dynamics of successful P2P lending, empirical studies such as (Bachmann et al., 2011; see also E. Lee & Lee, 2012; Serrano-Cinca et al., 2015; Weiss et al., 2012) also contributed positively to the literature to a limited extent since these do not incorporate quantitative research.

Emekter et al. (2015) detail important aspects regarding P2P loan characteristics, evaluates their credit risk assessment and measures loan performances. By exploring Lending club data, Emekter et al. (2015), observe a selection bias since high-income borrowers (highest FICO scores) do not borrow from Lending Club. In comparison, higher interest rates charged on the higher risk borrowers are not worth the risk. Besides answering the following questions: What are some of the borrowers characteristics that help determine the default risk? Is the higher return generated from, the riskier borrower large enough to compensate for the incremental risk? Emekter et al. (2015) predicted the default probability by modelling a Logistic Regression. According to this research, the ratio of default loan declines as the credit grade increases. Simultaneously, the total loan default ratio stood at 6.3% for the period taken into consideration, from May 2007 to June 2012. This study does not include the model accuracy in classifying “good” and “bad” borrowers.

Iyer et al. (2009) evaluate whether lenders in P2P markets can use borrower information to infer creditworthiness by exploiting the fact that researchers can analyse the exact credit score of a borrower instead of the aggregated credit category that **Proper.com** lenders have access to. Detecting the interest spread between the least and the most creditworthy credit rating can be explained by other features other than the credit rating itself.

According to the authors, the lenders assess the borrower's creditworthiness within credit-rating categories mainly from other standard-banking variables like the debt-to-income ratio, the number of current delinquencies or the number of credit inquiries. Non-standard variables have a lower impact on the interest spread.

Malekipirbazari and Aksakalli (2015) aspire to confirm if Random Forest, Support Vector Machines, Logistic Regression and K-Nearest Neighbour classification methods outperform the FICO based scores and the Lending Club grades to identify good borrower status. They concluded that lenders would be better off to lend only to the safest borrowers with the highest Lending Club scores.

This research employed five fold Cross-Validation combined with seven performance metrics that include: (i) accuracy rate on the test slice in the fold; (ii) the ROC area under the curve (AUC); (iii) the Root Mean Square Error (RMSE), and lastly (iv) the confusion matrix.

Lessmann et al. (2015) compare different traditional techniques such as Logistic Regression, linear to quadratic discriminant analysis with linear programming support vector machines (SVMs), neural networks, (tree augmented) naive Bayes and nearest-neighbour classifiers, helping to understand the power and usefulness of the SVM, least squares support vector machines (LS-SVM), and tree augmented naive Bayes (TAN), which was still in development at the time. The authors concluded that RBF LS-SVM and NN classifiers yield superior performance (in PCC and AUC). However, it highlighted that simple, linear classifiers such as linear discriminant analysis and Logistic Regression also delivered solid performances, which indicates that most credit scoring data sets are only weakly non-linear. The experiment also suggested that many classification techniques are quite competitive with each other. Only a small percentage of classifiers presented weakened results (namely QDA, NB and C4.5rules).

A recent master's thesis is ElMasry (2019) modelling the probability of default of mortgage lending using machine learning techniques, evaluating its results by analysing the ROC curve and computing the AUC rate.

Thanawala (2019) intends to boost credit risk models' performance by using German and Australian credit data to train machine learning algorithms. Another example of a study that pursues the same goals is Mezei et al. (2018) that uses a combination of traditional machine learning algorithms such as Neural networks, Decision Trees together with KNN and linguistic fuzzy set theory based data transformation.

Some proposals adopted reject inference methods, combining the data of the rejected and accepted loan applications. This has been done either by proposing novel reject inference methods such as Semi-supervised Support Vector Machines (Li et al., 2017) or by using OD-LightGBM that combines an outlier detection algorithm and state-of-the-art Boosting Decision Tree (Xia, 2019).

The credit risk assessment's main challenges are the evaluation speed, the concept drift problem, and the class imbalance problem (Zhang & Liu, 2019), which pointed out that only a few machine learning algorithms focus and solve the concept drift and class imbalance problems.

The majority of these proposals face imbalanced class distribution in the client credit assessment (Zhang & Liu, 2019) since the bulk of the number of "good" clients is larger than the number of risk clients. Brown and Mues (2012) dealt with this issue by applying simple random undersampling

and oversampling methods. Alternatively, Härdle et al. (2018) used the synthetic minority oversampling technique (SMOTE) to generate risk clients and achieve better performance than random sampling methods (Zhou et al., 2008).

Every year new machine learning algorithms are introduced. A portion combines existing algorithms and newly created ones. There is ample information available regarding ensemble, supervised, semi-supervised, reinforcement and unsupervised machine learning algorithms.

2.1.2. Results from related work

Results from previous related work, such as Malekipirbazari and Aksakalli (2015), Akindaini (2017), Polena (2017), Xia et al. (2017), Mezei et al. (2018), ElMasry (2019), and Thanawala (2019), are included in Table 1.

The author column represents the reference of the article, the model column refers to the name of the models studied, followed by the accuracy column and the AUC column that represents the AUC statistic obtained in each study. It is important to mention that not all of the studies have the same type of evaluation measures, but since the most common ones are the accuracy and the AUC score, these were chosen to be a part of this table.

Lessmann et al. (2015), Malekipirbazari and Aksakalli (2015) and Thanawala (2019) concluded that Random Forest performed nicely with an accuracy of 78.00%, 85.20% and 88.00%, respectively. Akindaini (2017) and Polena (2017) observed that Logistic Regression was the algorithm that performed better, while Xia et al. (2017) classified XGBoost-RS as the top performer. Mezei et al., (2018) categorised KNN as the algorithm with weaker performance. ElMasry (2019) classified Random Forest and SVM as the best classifiers.

Table 1 - Summary of the empirical results obtained in previous related studies

Author	Model	Accuracy	AUC	Dataset
(Malekipirbazari & Aksakalli, 2015)	1. Random forest 2. K-Nearest Neighbor 3. Support vector machine 4. Logistic regression	78.00% 70.10% 63.30% 54.50%	71.00% 55.00% 68.00% 68.00%	Lending Club (January 2012 - September 2014)
(Lessmann et al., 2015)	1. CART 2. ELM 3. LDA 4. Logistic regression 5. ADT 6. Bag 7. Boost 8. Random Forest	66.40% 69.80% 78.90% 80.70% 79.80% 76.80% 81.00% 85.20%	-	Australian credit dataset
(Akindaini, 2017)	1. Logistic regression 2. Multinomial Logistic Regression 3. Naïve Bayes 4. K-Nearest Neighbour (K=5)	95.15% 74.08% 70.74% 83.14%	-	Mortgage loan data from the Fannie Mae, Unemployment rate data from the US Department
(Polena, 2017)	1. Logistic regression 2. ANN 3. LDA 4. L-SVM 5. Random Forest 6. SVM with basis kernel function 7. Naïve Bayes 8. CART 9. K-Nearest Neighbor	- - - - - - - - -	69.79% 69.75% 69.55% 69.67% 69.28% 67.87% 65.19% 66.89% 63.73% 63.60%	Lending Club (2009-2013)
(Xia et al., 2017)	1. AdaBoost 2. AdaBoost-NN 3. Bagging-DT 4. Bagging-NN 5. DT 6. LR 7. NN RF 8. SVM GBDT 9. XGBoost-MS 10. XGBoost-GS 11. XGBoost-RS 12. XGBoost-TPE	61.25% 64.09% 62.43% 65.34% 60.11% 64.74% 63.65% 63.20% 60.67% 66.25% 66.70% 66.31% 67.08% 66 .97%	-	Lending club
(Mezei et al., 2018)	1. Neural network 2. Classification tree 3. K-Nearest Neighbor	80.02% 77.20% 75.80%	85.50% 80.10% 77.00%	P2P platform Bondora (March 2009 – February 2015)
(ElMasry, 2019)	1. Decision tree 2. Random forest 3. K-Nearest Neighbor 4. Support Vector Machine	88.40% 89.04% 88.84% 89.04%	-	Dataset provided by Freddie Mac (January 1999 – March 2017)
(Thanawala, 2019)	1. K-Nearest Neighbor 2. Logistic Regression 3. Naïve Bayes 4. Support Vector Machine 5. Decision tree 6. Random forest 7. Artificial Neural Networks	Australian data (85.00%) German data (75.00%); 87.00%; 77.00% 87.00%; 75.00% 87.00%; 77.00% 85.00%; 69.00% 88.00%; 75.00% 77.00%; 75.00%	-	Australian and German credit dataset

Source: Author's preparation.

3. MATERIALS AND METHODS

This section is divided into eight subsections. We start by providing an explanation of the individual classifiers and ensemble learning approaches applied in this study, offering additional clarification about the hyperparameters tuning procedures adopted in each one of them. We follow by describing the dataset used alongside with the experiment design and the exploratory analysis. Whenever possible, we use visualisation tools to analyse several types of data. After the exploratory analysis, it is time for the explanation of the pre-processing stage, which includes all the steps taken and the reasoning behind each one of them, such as the missing values treatment, feature selection and feature engineering. In the modelling techniques subsection, the different stages from splitting the data into different subsets to the definition of important concepts related such as K-fold Cross-Validation, class reweight, and the technique used for hyperparameters optimisation.

3.1. SINGLE CLASSIFIERS

3.1.1. Logistic Regression

Logistic Regression (Cox, 1958) is a parametric method that is generally used in credit scoring among financial institutions because it is easily interpretable and can directly predict probabilities, which is one of the main differentiation factors concerning Linear Regression. The Linear Regression pursues finding estimates of the parameters so that the sum of the squared errors of a function of the sum of the squared error can be divided into components of model variance and bias (Kuhn & Johnson, 2013). The Linear Regression model's downsides include restrictive expressiveness (Molnar, 2019) since the solution obtained is a flat hyperplane, and the interactions must be added manually. Simultaneously, the interpretation can be more complex than other models, such as Decision Trees, because the weights are multiplicative and not additive. It is an iterative and calculation-intensive methodology that may need about six training epochs to reach convergence, using one of several convergence criteria (Garson, 2012). Logistic Regression is the go-to option for developing credit-scoring models (Lessmann et al., 2015), in particular, because: (i) the final probability cannot fall outside of the range 0 to 1; and (ii) it provides a relatively robust estimate of the actual likelihood, given available information (Anderson, 2007).

Logistic Regression models the relationship between one or more independent variables (predictors) and categorical variables (output) by employing a logistic function to evaluate the probabilities. Logistic Regression can be binary (the categorical response has just two possible outcomes), multinomial (three or more categories without ordering) or ordinal (with three or more classes with ordering) (Akindaini, 2017).

According to Kuhn and Johnson (2013), the probability of a default can be written in the following mathematical form

$$p = P\,(loan\ default\ status\ =\ 1)\ =\ \frac{1}{1\ +\ e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_K X_k),}} \qquad (1)$$

where β_0 is the Linear Regression intercept $X_j (j = 1, \ldots, k)$ are explanatory (independent) variables with $\beta_j\ (j = 1, \ldots, k)$ the corresponding coefficients to be estimated from regressing the

model on data. In this case, the response (output) is a binary variable equal to 1 when the loan status is default and 0 when the loan is fully paid. Hence, the Logistic Regression formula is:

$$\frac{f(x)}{1 - f(x)} = e^{-(\beta_0 + \beta_1 x + \ldots + \beta_K x_K)} \quad (2)$$

Note that the probability of loan default status equal to 0 is equivalent to $1 - p$.

Logistic Regression models are fitted via maximum likelihood estimation, which is a technique that can be used when one is willing to make assumptions about the probability distribution of the data. The likelihood function is a probability statement that can be made towards a particular set of parameter values when identifying two sets of parameters, the set with the bigger likelihood is likely to be considered more consistent with the observation data (Kuhn & Johnson, 2013). This process is composed of three different steps: (i) transform the dependent variable into a log function; (ii) guess what the coefficients should be; and (iii) determine changes to the coefficients to maximise the log-likelihood (Anderson, 2007).

Logistic Regression's simplicity can be maintained whilst improving the predictive performance using univariate and bivariate threshold effects.

The Logistic Regression model is trained with stochastic gradient descent, which has several advantages. First, it only requires the parameter and a single training example to be stored in memory. It is also computationally fast and works well with large datasets. The gradient of the loss function is estimated one sample at a time, and the model is updated along the path with a decreasing learning rate. The training examples are sorted in random order, and the parameters are updated for each example sequentially (Elkan, 2012). On the other hand, it can also lose the advantage of vectorised operations, as it deals with only one single example at a time. The minima's frequent updates are very noisy and can lead the gradient descent into other directions (Kapil, 2019).

When tuning this classifier, two parameters were taken into consideration the penalty and the alpha. Penalty corresponds to the norm used in penalisation, while the alpha corresponds to the constant that multiplies the penalty. In terms of penalty, two different types were evaluated: l1 corresponding to Lasso regression and l2 that corresponds to Ridge regression. These are shrinkage methods that regularise the coefficient estimates to improve the fit while managing at the same time to reduce the coefficients' estimates variance.

The Ridge regression adds a "squared magnitude" of coefficient as a penalty to the loss function. According to James et al. (2013), the ridge regression coefficient estimates $\widehat{\beta^R}$ are the values that minimize:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2 \quad (3)$$

where $\lambda \geq 0$ is a tuning parameter, in our case the alpha, determined separately, has the purpose of controlling the relative impact of the two terms of the equation on the regression coefficient estimates. When $\lambda = 0$, the penalty term does not affect, and the ridge regression will produce the least squares estimates. However, when $\lambda \to \infty$, the impact of the shrinkage penalty, the second term of the equation, grows and the ridge regression coefficient estimates will approach zero (James et al., 2013).

According to James et al. (2013), the Lasso regression is a newer alternative to Ridge regression, which adds an "absolute value of magnitude" of coefficient as a penalty term to the loss function. Where the lasso coefficients, $\hat{\beta}_\lambda^L$ minimize the quantity:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (4)$$

Similarly to the Ridge regression, the Lasso shrinks the coefficient estimates towards zero however the penalty has the influence of forcing some of the coefficient estimates to be exactly equal to zero, when tuning parameter λ is sufficiently large, thus it performs variable selection. Therefore, the results are easier to interpret.

3.1.2. Decision Trees

An example of a non-parametric approach. Appropriate for categorical analysis since it can identify patterns, including finding and exploiting interactions. The results are transparent and easily interpretable. Additionally, it is computationally simple, using only one measure to choose variables and to determine the splits, but it is relatively inflexible. The Decision Tree theory is appropriate for credit scoring modelling and has been used extensively (Lee & Chen, 2005). Another essential aspect is that it does not make any assumptions about the underlying distribution.

To explain this subject more easily, figure 16 is a graphical representation of a Decision Tree. First, the root node contains a sample of good and bad credit applications. Then, the algorithm attempts all possible binary splits to find the attribute x and corresponding cut-off value giving the best separation in terms of discriminating individuals according to the class they belong to. This process is repeated for the new nodes until a stopping criterion is satisfied. To understand the splitting criteria, it is important to know the concepts of entropy and information gain. When building a classification tree, either the Gini index or the entropy are generally used to evaluate the quality of the split (James et al., 2013). The Gini index is defined as a measure of node purity, which means that a small value indicates that a node contains mainly observations from one class. According to James et al. (2013), the formulation is as follows:

$$G = \sum_{k=1}^{K}\hat{p}_{mk}\,(1-\hat{p}_{mk}) \qquad (5)$$

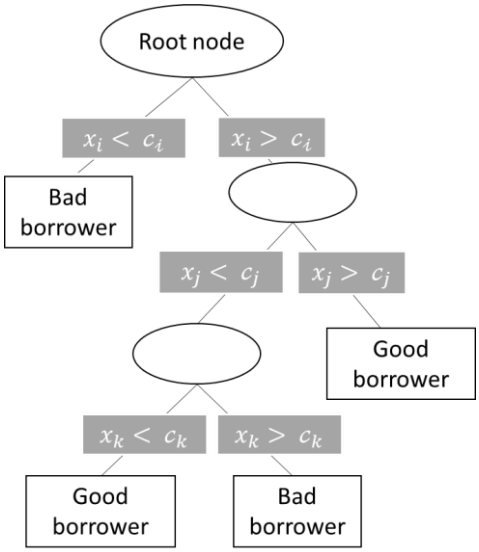
Defined as a measure of total variance across the k class. Whereas entropy is defined by James et al. (2013) as:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \, log\hat{p}_{mk} \quad (6)$$

Where $\hat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class. Since $0 \leq \hat{p}_{mk} \leq 1$, it follows that $0 \leq \hat{p}_{mk} log\hat{p}_{mk}$. Meaning that the entropy will take on a value near zero if the $\hat{p}_{mk}$'s are all near zero or near one.

The information gain from each split within the several features options is assessed, and the feature that maximises the information gain is the chosen one to do the split.

Figure 1 - Decision Tree diagram



Source: Author's Preparation based on Bastos (2008).

Decision Tree algorithms have been gaining popularity because of the features above mentioned. This model will find the splitting rule that efficiently distinguishes the bad from the good borrowers in terms of their probability of default (Amaro, 2020). However, there has been some evidence of a tendency to overfit the training data in a single classifier format since the trees are constructed through recursive partitioning, which often results in an intricate tree with many internal nodes (Lessmann et al., 2015). The set of partition values is determined by calculating the midpoint of each set of consecutive unique responses along with each feature. For each p unique responses, $p - 1$ possible partition values are calculated. Then, a scoring criteria is used to evaluate and compare each of the possible partition values.

In terms of hyperparameters used in the optimisation, both the Gini index and entropy were evaluated in addition to different "*min_samples_split"* which is the minimum number of samples required to split an internal node.

3.1.3. K-Nearest Neighbours

KNN is considered a relatively simple method since it merely identifies the K-nearest observations from the training sample to classify a new observation from a testing set (Polena, 2017). A distance measure is used to define which of the K instances in the training dataset are more similar to the

new input. Distance metrics include Euclidean, Manhattan, Chebyshev, Hamming and Minkowski distance. The Euclidean distance is calculated as the square root of the sum of the squared differences concerning a new point and an existing point across all input variables (Härdle et al., 2018). While the Hamming distance calculates the distance between binary vectors, the Manhattan calculates the distance between real vectors using the sum of their absolute distance. The Minkowski is a generalisation of Euclidean and Manhattan.

The accuracy of the KNN classification relies heavily on the value of K, i.e., the neighbourhood's size, which means that classification with a large value of k is more robust and less prone to outliers, implying lower variance but increased bias. In comparison, a small K will restrain the prediction region and lead to high variance with low bias.

It does not learn a discriminative function from the training data but memorises the training dataset instead (ElMasry, 2019), which entails a possible memory problem. Kuhn and Johnson (2013) state that the KNN can have a poor performance when the "local" predictor structure is not relevant to the response. This can be referred to as a curse of dimensionality, which entails the various difficulties a larger number of predictors can cause to model fitting or computation (Härdle et al., 2018). Hence, removing irrelevant or noisy predictors is a critical pre-processing step for KNN. Therefore, in this research, it was decided to include in this model, Principal component analysis to reduce the feature number, as is further explained in the next sector.

3.2. ENSEMBLE LEARNING APPROACHES

The difference between single classifier machine learning approaches and ensemble methods is that the first try to learn one hypothesis from training data. In contrast, the second one tries to construct a set of hypotheses and combine them (Zhou, 2013). An ensemble containing several base learners can be a neural network or other learning algorithms. There are homogeneous and heterogeneous ensembles depending if there the learners are of the same type or different kinds. The most common types of ensembles include Bayes optimal classifier, Bootstrap aggregating (Bagging), Boosting, Bayesian model averaging or ensemble (BME), Bayesian model combination, Bucket of models and Stacking ensembles (see Bravo (2019, 2020, 2021), Bravo et al. (2021a,b,c), Ayuso et al. (2021a,b), and Bravo and Ayuso (2020, 2021) for concrete examples on the use of BME for time series forecasting).

One of the benefits of this methodology is the ability to boost weak learners into vital learners. Two stages are necessary to produce an ensemble. Firstly, several base learners are generated either in a parallel or sequential style.

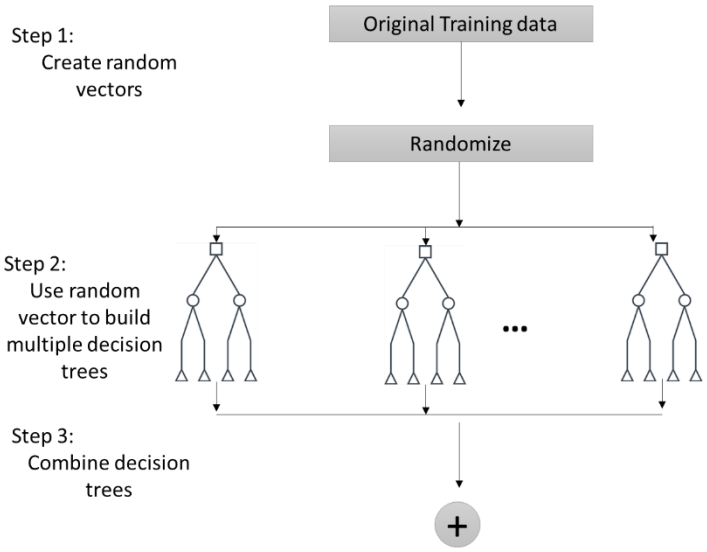
Ensemble approaches are considered better than single because the training data may not have sufficient information to select one single best learner and minimise the search process's deficiencies. However, it is essential to mention that it can bring a lack of comprehensibility of the knowledge acquired. Boosting and Bagging are two well-known ensemble approaches. Boosting has proven its capacity to reduce the generalisation error and not suffer from overfitting. Simultaneously, Bagging can significantly diminish bias and reduce variance (Zhou, 2013).

3.2.1. Random Forest

Random Forest is a homogeneous ensemble classifier introduced by Bregman in the early 2000s that uses a Decision Tree as base learners. The model is trained with samples drawn randomly with replacement (bootstrap samples) from the training dataset and using random feature selection in each single tree generation (Brown & Mues, 2012).

Each tree generates a class prediction, and the class with the majority of votes becomes our model's prediction, which can be characterised as the wisdom of crowds. To achieve this, we have to make sure that the models diversify each other. The trees of the forest and, more importantly, their predictions need to be uncorrelated (Yiu, 2019), and features with some predictive power need to be included.

Figure 2 - Random Forest methodology



Source: Author's Preparation based on Malekipirbazari and Aksakalli (2015).

Random Forest improves on Bagging because it decorrelates the trees with the introduction of splitting on a random subset of features.

According to Brown and Mues (2012), the Random Forest's ability to concentrate on "local" features in the imbalanced data is valuable. The advantages of this approach include: being insensitive to skewed distribution, outliers and missing values, the fact that the predictive variables can be of any type (Carvajal et al., 2018). Also, it can judge variable importance by ranking each variable's performance (ElMasry, 2019).

3.2.2. Extreme Gradient Boosting

Chen and Guestrin (2016) proposed a decision-tree-based ensemble algorithm that uses a gradient boosting algorithm. Gaining popularity and considered as an efficient open-source implementation of this algorithm.

Extreme Gradient Boosting has been responsible for winning numerous Kaggle competitions, achieving state-of-the-art results in various real-world applications (Chen & Guestrin, 2016). One

of its more favourable attributes is its scalability in all scenarios since it can handle various data types, relationships, distributions and fine-tune a variety of hyperparameters. XGBoost has multiple application regression, classification and ranking problems.

Unlike Random Forest, Extreme Gradient Boosting uses boosting that combines weak learners, usually Decision Trees with only one split – decision stumps, sequentially so that each new tree corrects the errors of the previous tree (Dhingra, 2020).

Firstly, it starts with one Decision Tree. Then, using a loss function, the performance of the tree is evaluated. There are different loss functions, such as Cross entropy or Logarithmic loss, which penalises false classifications by considering the probability of classifications. Cross entropy is a similar metric and the loss associated with it increases as the predicted probability diverges from the actual label (Saha, 2018). Upon completing the first tree and the loss function, the next tree added to lower the loss than the first tree alone. The core problem of XGBoost is to determine the optimal tree structure, employing a greedy search algorithm to achieve this (Xia et al., 2017).

Choosing a learning rate is a technique used to slow down the model's adaptation to the training data by applying a weighting factor for the corrections by new trees when added to the model.

In our case, the Grid Search capability was used to evaluate the effect of a different number of estimators and the learning rate on the logarithmic loss of training a gradient boosting model with diverse learning rate values.

Execution speed and model performance are two of the most appealing benefits. By allowing parallelisation, possible due to the interchangeable nature of loops used for building base learners, the outer loop enumerates a tree's leaf nodes and the second inner loop that calculates the features. Regarding tree pruning, the stopping criterion within the GBM framework is greedy and depends on the negative loss criterion. Uses the maximum depth of a tree parameter, allowing for a computational performance enhancement, as specified instead of criterion first and performs the pruning trees backwards.

Hardware Optimisation achieved by allocating internal buffers in each thread to store gradient statistics. In addition to the availability of 'out-of-core computing to optimise available disk space.

3.3. PERFORMANCE EVALUATION

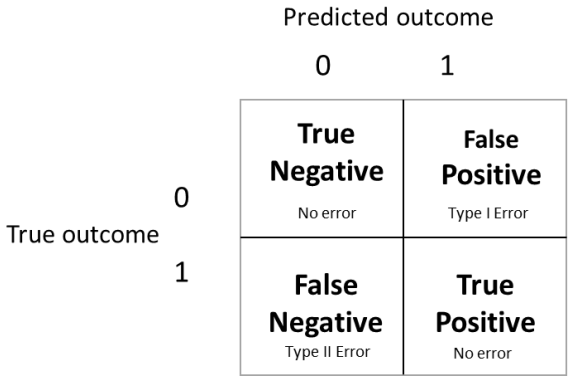
There is a variety of performance evaluation criteria used in credit scoring applications. According to ElMasry (2019) there are three types of performance measures: Discriminatory ability, the accuracy of probability predictions, and predictions' correctness. Several techniques include the area under the curve that is the most commonly used in prediction models with binary outcomes.

The correctness of predictions can be measured using Kolmogorov-Smirnov Statistic (KS) and Percent Correctly Classified (PCC).

The confusion matrix's introduction is necessary to explain the percentage correctly classified (PCC) and the area under the curve (AUC). Percent Correctly classified is the portion of the observations that are classified correctly. The correctly classified cases are called true positives (TN) and

negatives (TN). On the other hand, if it does not correspond, they are labelled false positives (FP) equivalent to type I error and negative resulting in a Type II error, illustrated in the following table. One similarity between these two mechanisms is the fact that they measure accuracy to a single reference point.

Figure 3 - Confusion Matrix diagram



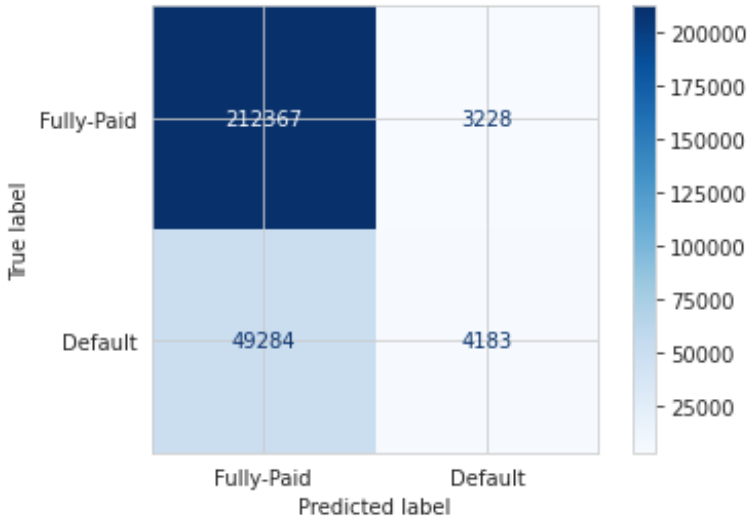
Source: Author's Preparation based on Polena (2017).

- True Positive (TP): The truthful outcome is one, and the prediction is also one. The borrower defaulted, and the same outcome was predicted in our classification.
- True Negative (TN): The real outcome is 0, and the predicted outcome is also 0. Meaning that the borrower fully paid his, hers or their loan, and our classification model predicted the same outcome.
- False Positive (FP): The truthful outcome is zero, and the predicted one is one. The borrower fully paid his, hers or their loan, although the classification model predicted otherwise.
- False Negative (FN): The real outcome is one, and the predicted fallout is zero. The loan was charged off, yet the opposite outcome was predicted.

The formulation of percentage correctly classified corresponds to the following:

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

Figure 4 - Confusion Matrix of XGboost



Source: Author's Preparation.

By recurring to the confusion matrix, it is possible to calculate a few more metrics such as sensitivity, specificity, precision and the F1 score.

Sensitivity or recall is the measure of the correctly identified positive cases from all the actual positive cases:

$$Sensitivity = \frac{TP}{TP + FN} \quad (8)$$

While specificity identifies the portions of non-default cases that are classified as so:

$$Specificity = \frac{TN}{TN + FP} \quad (9)$$

Precision quantifies how many of the loans labelled as defaulted in our prediction defaulted in reality. This metric is beneficial in cases where the classification of True Positives is a priority. In our case, since a TP represents a risk of loss, it is highly recommended precision.

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

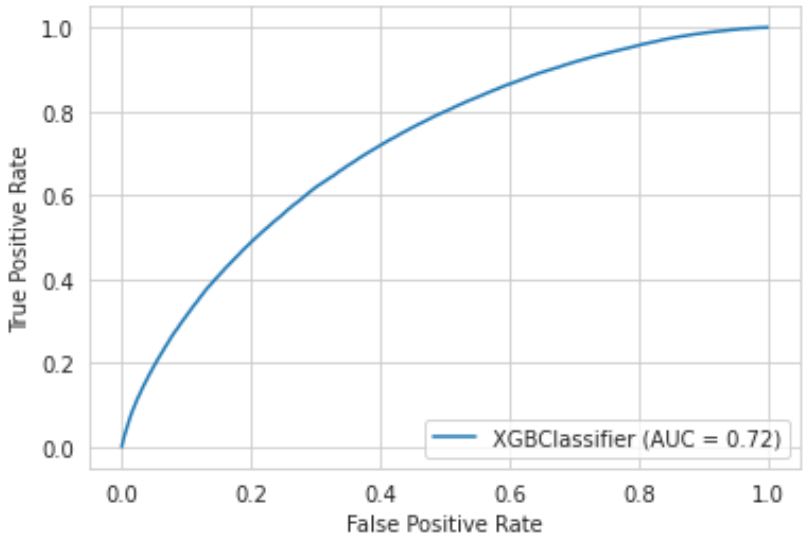
The F1 score, also known as F-score or F-Measure, is a combination of Precision and Recall. Useful when there is a class imbalance.

$$F1 = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (11)$$

AUC Curve equals the probability that a randomly chose positive case receives a score higher than a randomly chosen negative case. Lessmann et al. (2015) stated that the AUC performs a global assessment in that it considers the whole score distribution. It uses relative (to other observations) score ranks. This measure works by calculating the relative numbers of correctly and incorrectly identified predictions across all possible classification thresholds (Redding et al., 2017).

Lessmann et al. (2015) and Polena (2017) used six performance measures from three different performance measurement groups, which are: Area under the curve (AUC); Percentage correctly classified (PCC); Brier score (BS); H-measure (H); Partial Gini index (PGI) and Kolmogorov-Smirnov statistic (KS) showing adequate results. Nevertheless, using many instead of one statistic allows for a robust and complete evaluation of the competing models' relative performances. Therefore, this methodology is going to be reproduced in this research.

Figure 5 - ROC curve of XGBoost



Source: Author's Preparation.

3.4. DATASET

The dataset belongs to the P2P lending organisation "Lending Club". It contains data for all loans issued on Lending Club from 2007 until the fourth quarter of 2018.
The dataset can be found on the website below:
https://www.kaggle.com/wordsforthewise/lending-club#accepted_2007_to_2018Q4.csv.gz
and it was accessed on 9/10/2019.
Together with the accepted loan data, there is also the data related to the rejected loans. Nevertheless, for our classification problem, only the accepted loan data was applied.

In the following section, there is a description of the origination data.

3.4.1. Origination data

A total of 151 variables are present in the original Lending club dataset. 112 variables are numerical and 39 are categorical.

The data focuses on three different aspects:

1. Personal details (for example address, employment, homeownership).
2. Credit history (for example: whether the borrower has filed for bankruptcy, the balance of all accounts, inquiries, revolving and currently past due accounts).
3. Loan characteristics (for example issue date, the FICO score, LC grade, subgrade, status, type of candidature, policy, payment dates, purpose, term, late fees and principal amount).

Table 2 - Data dictionary of the final variables

Variables	Description	Allowable values
addr_state	The state provided by the borrower in the loan application	Categorical: Shortaname of the US States
annual_inc	The self-reported annual income provided by the borrower during registration.	Numeric (0 - 1.1e+08)
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers	Categorical: Joint, Individual and Direct_pay
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.	Numeric (-1 until 9999)
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.	Numeric (610 - 845)
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.	Numeric (614 - 850)
grade	LC assigned loan grade	Categorical: A; B; C; D ; E; F; G
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER	Categorical: Mortage; Rent; Own; Other; None; Any
installment	The monthly payment owed by the borrower if the loan originates.	Numeric (4.93 - 1719.83)
int_rate	Interest Rate on the loan	Numeric (5.31 - 30.099)
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.	Numeric (500-40000)
open_acc	The number of open credit lines in the borrower's credit file.	Numeric (0- 1.01e+02)
pub_rec	Number of derogatory public records	Numeric (0 - 8.61e+01
pub_rec_bankruptcies	Number of public record bankruptcies	Numeric (0 -1.2e+01)
purpose	A category provided by the borrower for the loan request.	Categorical: Car; Credit Card; Debt Consolidation; Educational; Home improvement; House; Major Purchase; Medical; Moving; Other; Renewable energy; Small business; vacation; Wedding;
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.	Numeric (0 - 8.023e+02)
total_acc	The total number of credit lines currently in the borrower's credit file	Numeric (1 - 1.76e+02)
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified	Categorical: Verified; Not verified; Source Verified

Source: Author's Preparation retrieved from Lending Club resources.

There is a complete dictionary available online that contains information regarding all the variables but for this research, it was decided not to include all the variables in our data dictionary and focus on the variables that were a part of the model.

In table 2 it is possible to examine a table that contains information about the variables that were chosen to be a part of the final model, which includes descriptions of variables along with allowable values and indication if they are categorical or numerical.

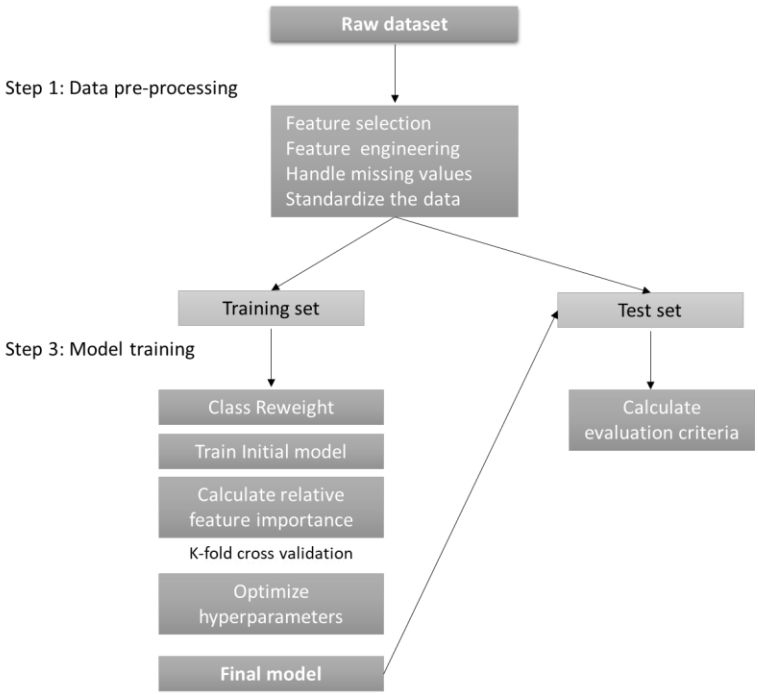
There are two types of disbursement methods: cash and direct pay. When the borrower does not fulfil their obligations, two things can happen - a settlement plan or a hardship plan where there is an additional interest payment with an associated status. In this case, the organisation can opt by pulling off the loan. There is information regarding the hardship plan and about the settlement. The loan can be applied either as an individual or joint loan. All the loans are unsecured, so the investor assumes the liability if the borrower defaults on their loan.

3.5. EXPERIMENT DESIGN

The experiment design provides an introduction for the following sectors so that it is easier to visualise and understand the methodology.

This flowchart is a representation of the experiment design. It helps in introducing the topics that are presented in detail next in this chapter. The experiment starts with the raw dataset that was introduced in the previous sector, followed by the pre-processing, then split the data into two samples and then again, several steps are conducted to prepare for the training and evaluation of the models, which are further explained in the subsequent sections.

Figure 6 - Flowchart of the methodology



Source: Author's Preparation.

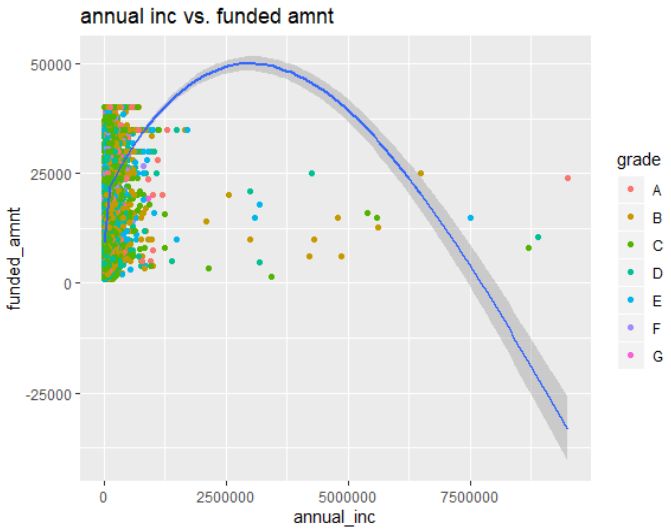
3.6. EXPLORATORY ANALYSIS

This section revolves around getting familiar with the data and present graphical visualisations so that it is easier to understand the key takeaways.

3.6.1. Visualisation of the data

To help visualise and understand better the dataset, several graphs were plotted from figure 5 to figure 12.

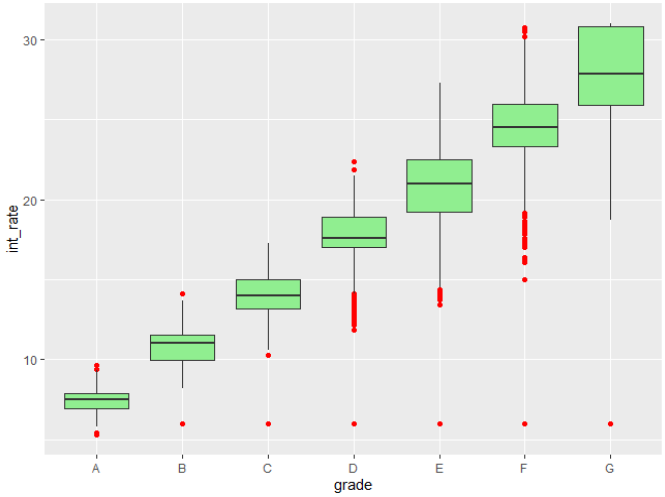
Figure 7 - Annual income vs Funded amount



Source: Author's Preparation.

Figure 7 can be interpreted by looking at the curve that represents the relation between the funded amount and the annual income, where it is visible that until a certain threshold, the funded amount increases with higher annual income. By looking at the dots, the majority seems to be represented by the green colour, which can be interpreted as a sign that most of the loans are characterised as grade C. It is possible to observe the existence of some potential outliers.

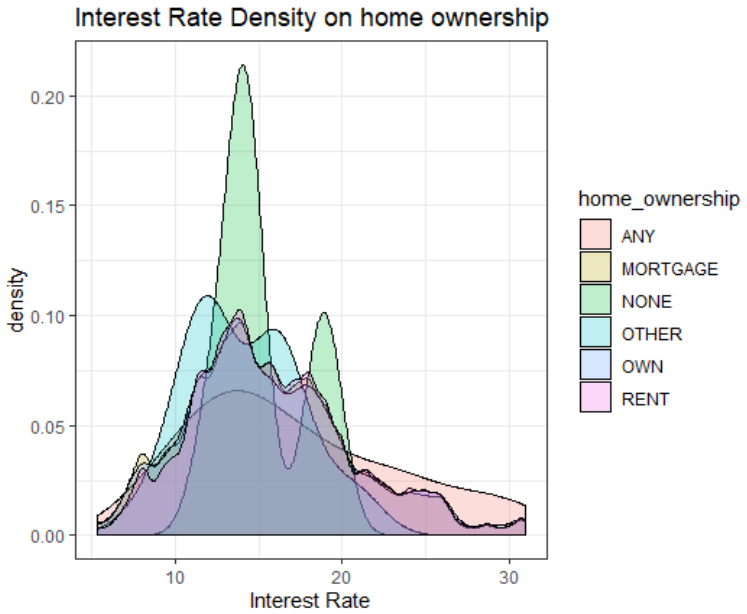
Figure 8 - Density plot of the grade vs the interest rate



Source: Author's Preparation.

In figure 8, it is possible to observe the density of the different loan interest rates and the respective grades. A conclusion that can be drawn from this is that the higher-grade loan correlates to a lower interest rate.

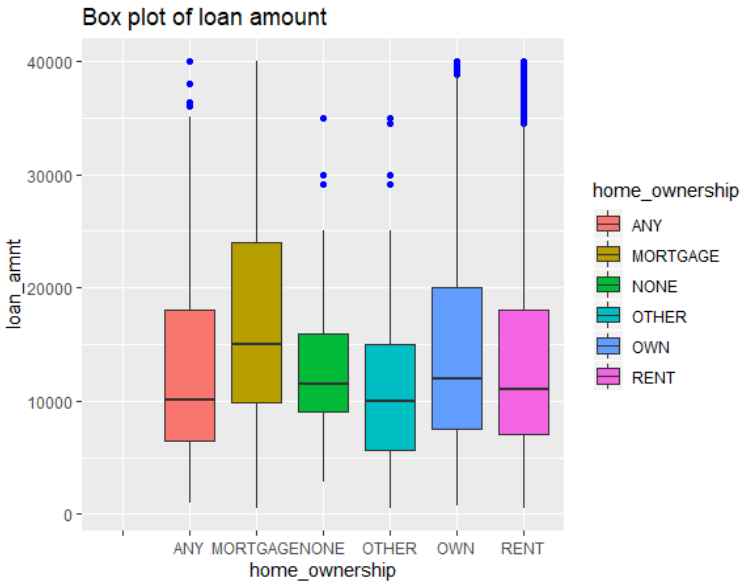
Figure 9 - Visualisation of Interest Rate Density on Homeownership



Source: Author's Preparation.

In figure 9, the different colours correspond to different types of home ownership of the applicants. The density is the proportion of loans in a specific interest rate level. What can be observed by looking at this graph is that the highest density value corresponded to having home_ownership status as None and interest rate between 10 and 20.

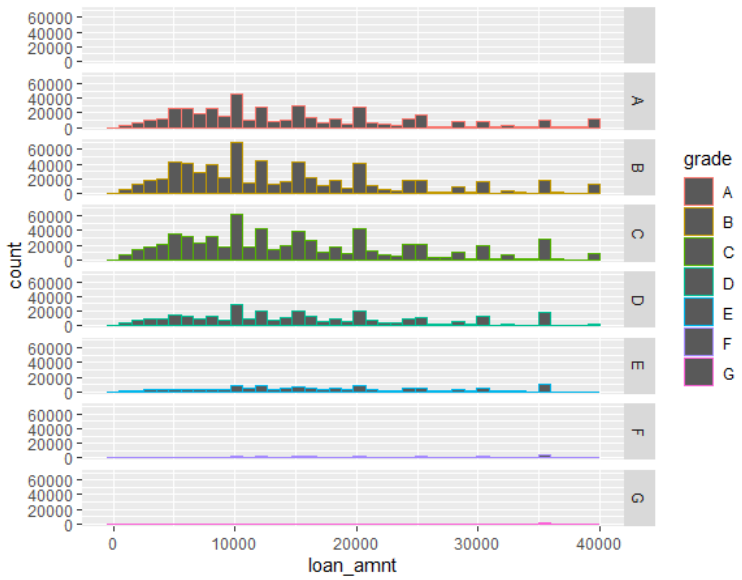
Figure 10 - Box plot of Loan amount and Homeownership



Source: Author's Preparation.

In figure 10, the different colours represent different types of homeownership and the y-axis represents the loan amount. What can be inferred from this graph is that for most cases the loan amount does not fluctuate much.

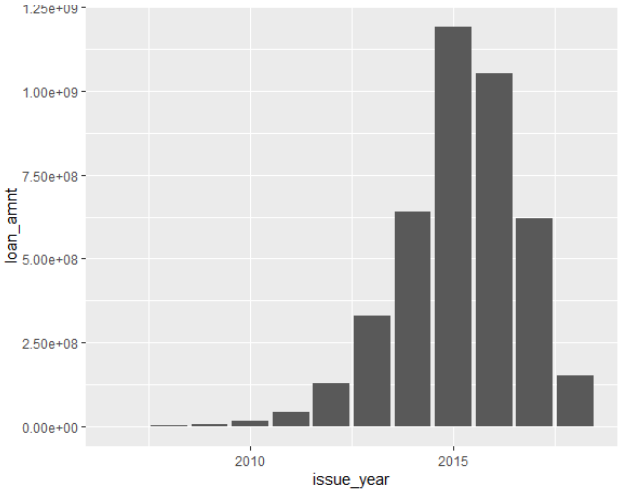
Figure 11 - Histogram of loan amount vs grade vs number of loans



Source: Author's Preparation.

In figure 11, one of the most notable aspects is the decreasing number of loans as one reaches the bottom of the figure, which means that there are fewer loans with grade G to E than with grade A, B, C and D.

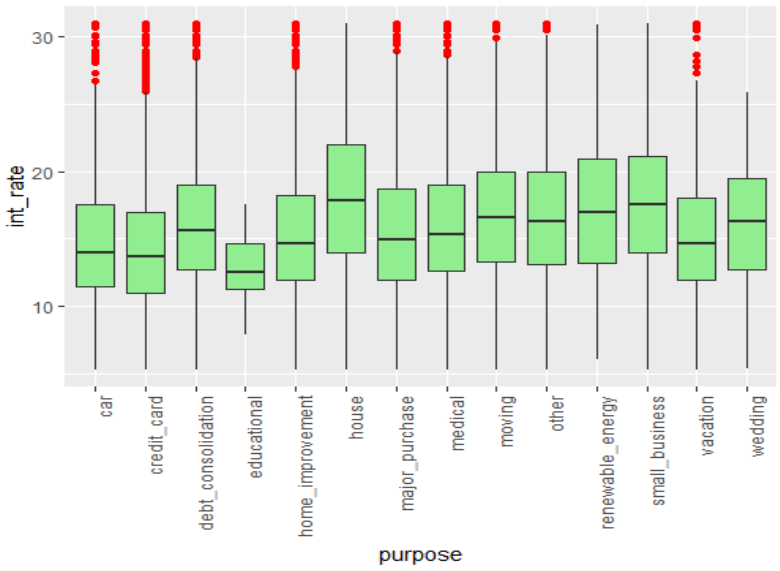
Figure 12 - Plot of Issue year vs Loan Amount



Source: Author's Preparation.

From analysing figure 12, it is possible to infer that the loan amount reached its peak in 2015 and started to decrease since then.

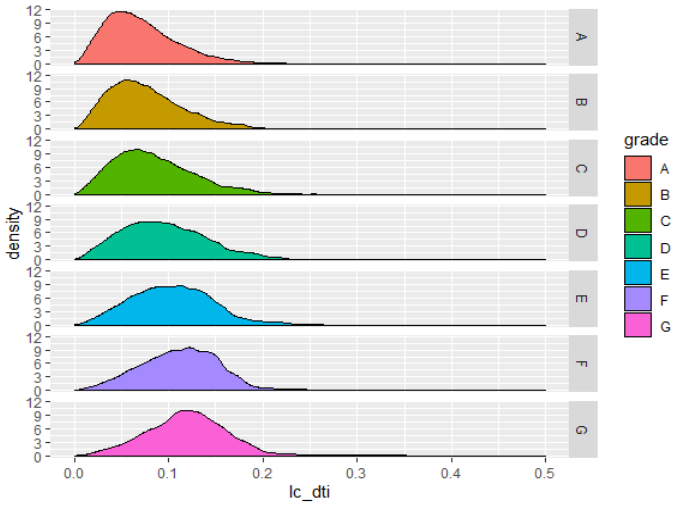
Figure 13 - Box plot of purpose vs interest rate



Source: Author's Preparation.

In figure 13, there are indicators of the presence of outliers and there is not a direct correlation between the interest rate and the purpose of the loan.

Figure 14 - Density plot of Debt-to-Income ratio vs Grade



Source: Author's Preparation.

In figure 14, the x-axis represents the debt-to-income ratio and the y-axis represents the density while the different colours represent different loan grade. It is possible to suggest that the debt-to-income ratio is widely held between 0 and 0.2.

3.7. DATA PRE-PROCESSING

The models considered in this research require diverse types of preparation. For instance, it is important to remove correlated inputs in the Logistic Regression because the estimation process is more susceptible to failure if this is not considered.

Detailed pre-processing is presented in the next segment.

3.7.1. Missing values

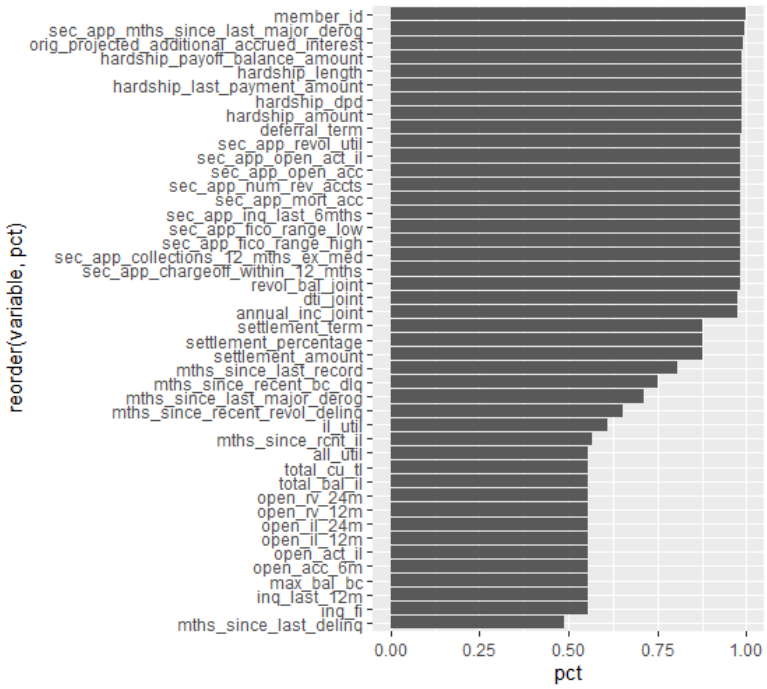
Variables with more than seventy percent of missing values were deleted.

Regarding the variables with a percentage below seventy percent, the course taken to remove the missing values was the following:

- For numerical features: replace the missing values with the median of the non-missing values;
- For categorical features: replace the missing values with the most frequent non-missing value;

To visualise the missing variables, figure 15 was plotted, where axis x represents the percentage of missing values, and axis y presents the variables' name. Merely the variables with more than thirty percent of missing values are presented because it would be impossible to make a perceptible plot with the original 151 variables.

Figure 15 - Missing values plot



Source: Author's Preparation.

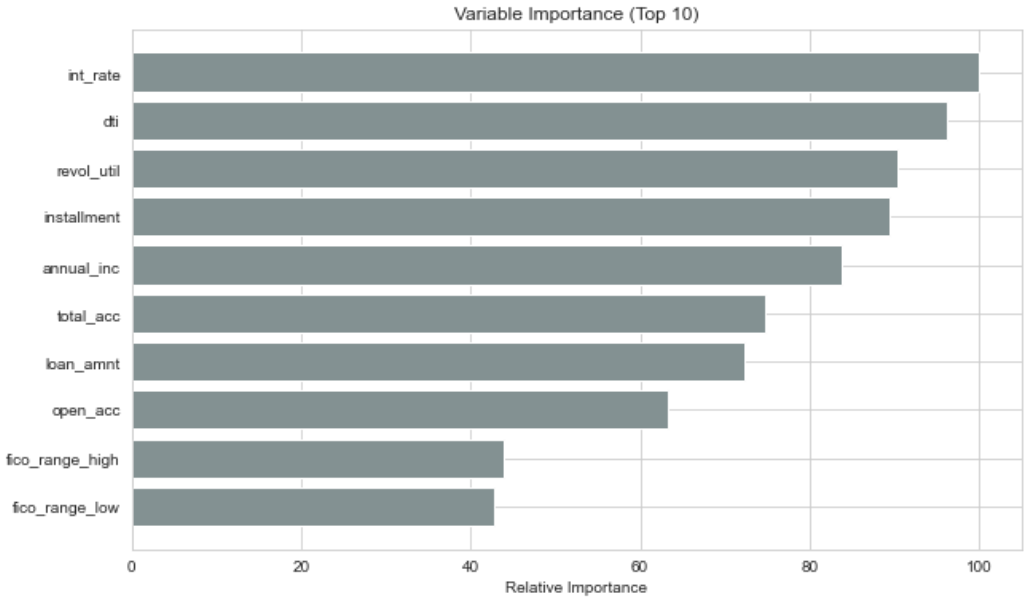
3.7.2. Feature selection

Feature selection aims to avoid overfitting, reduce noise and redundancy, reduce computing effort, and interpret the output more easily. The different stages of the process were: evaluating the variable importance, correlation analysis, and analysis of the optimal number of features assessing the performance on the test set.
The criteria applied to include:

- Features that would not have been available at the time of the loan;
- Convert strings to numerical values;
- Drop superfluous attributes;
- Zero (or nearly zero) variance predictors;
- The high number of missing values. Namely with more than 70% of missing values.

To evaluate each variable's importance in building the model, the Scikit-learn Package was used to remove redundant variables by creating a correlation matrix. This package can build a matrix and rank the importance of each variable when making a model.

Figure 16 - Variable importance



Source: Author's Preparation.

Zero or nearly zero variance predictors were assessed and consequently removed.

After the pre-processing and the creation of dummy variables, the dataset was left with eighty-nine variables, so in order to optimise the results and the learning time on the algorithms that work better with a lower number of variables – KNN, a Principal Component analysis was performed. A PCA analysis is a popular approach for deriving a low dimensional set of features from a large set of variables (James et al., 2013). It combines input variables in a specific way that allows for the "least important" variables to be dropped while retaining the most valuable parts of all the variables. Nevertheless, each of the "new" variables after PCA are independent of one another. The PCA package used to perform this was from the Scikit-learn library, which identifies the combinations of attributes or principal components that account for the most variance in the data by decomposing a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of the variance (Halko et al., 2009).

3.7.3. Feature engineering

Feature engineering includes cleaning and formatting the data, checking each feature for its relation to the target variable and transforming features.

The pre-processing differed according to the type and cardinality of the variable.

- Numerical variables:

The variables "policy_code" and "pymnt_plan" have mostly one level of data in all rows, so it is better to remove it as well.

- Categorical variables with a low number of categories:

The date columns were in a categorical format, so they were converted into a date format.

- Categorical variable with a high number of categories:

The “id”, “member_id” and “URL” were removed to avoid overfitting as they have unique values for each loan for the sole purpose of identification. Removing the “emp_title” variable was necessary because it had almost one unique value for each loan, which would slow down the learning time.
The title and purpose have a similar description. Therefore to reduce redundancy, the best option was to remove one of these features.

- Default Status Variable:

Initially, the target variable was Loan status, which could assume the following values: Charged off, current, default, does not meet the credit policy. Status: Charged off, does not meet the credit policy, Fully Paid, Fully paid, In grace period, Late (16-30 days) and Late (31-120 days). Since this research’s primary goal is to study the accuracy in distinguishing good from bad loans, we will only include the loans being fully paid or charged off, assuming that they both meet the credit policy. A fully paid loan corresponds to entirely repaid loans. Either by a prepayment or at the maturity of a three or five-year term.
In comparison, the charged-off category corresponds to a loan for which there is no longer a reasonable expectation of further payments. This variable was later filtered to contain the fully paid, representing 80% of the total observations, and charged-off loans, followed by its transformation into a binary variable for modelling.

The dataset is unbalanced because of the disparity between the number of default and non-default loans. Previous studies have shown the advantages of undersampling and oversampling to deal with this matter. We note that a further explanation about these techniques is available in the section Class Reweight.

- Dummy Variables

The remaining factor variables (“grade”, “home_ownership”, “verification_status”, “purpose”, “addr_state” and “application_type”) were transformed into dummy variables because the Logistic Regression requires that all data are numbers (Nisbet et al., 2018).

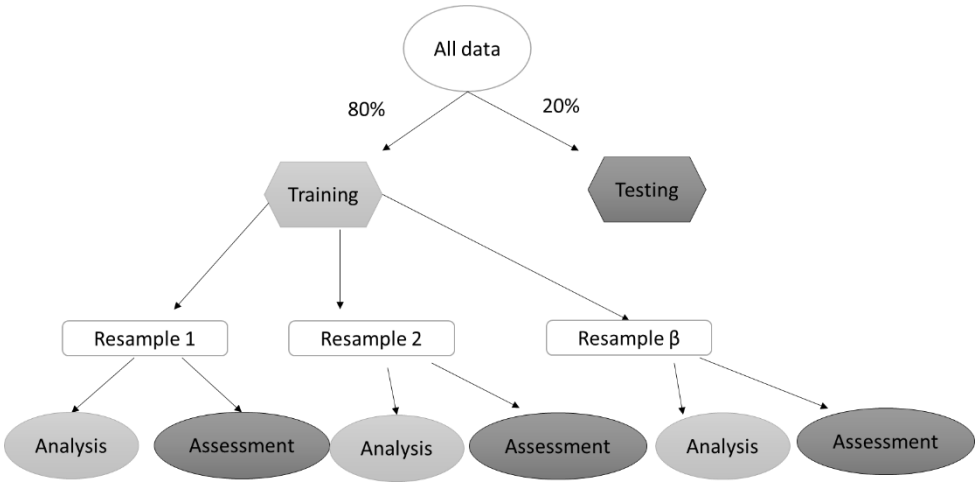
3.8. MODELLING TECHNIQUES

3.8.1. Sample split

The chosen method splits the original dataset into two parts – training and testing dataset. A proportion of 80% attributed to the training set and the remaining to the testing. This separation helps mitigating problems such as overfitting the model.

The test section ensures enough data to be statistically meaningful results.

Figure 17 - Diagram of the predictive modelling workflow



Source: Author's Preparation.

3.8.2. K-fold Cross-Validation

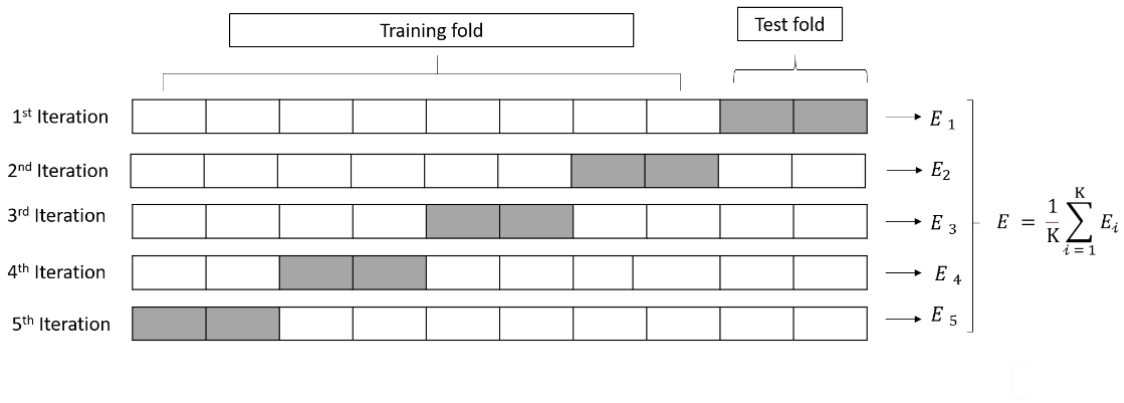
Using resampling is essential because there are times where assessing the effectiveness of the model without using the test set is necessary. Resampling methods are imperative because they can generate different versions of our training dataset to train our model in different subsets and simulate how the model would perform in the new data. There are diverse techniques of resampling, such as Bootstrapping and cross-validation.

The K-fold Cross-Validation method divides the original dataset into k subsets. Each of the k subsets is used as testing data in one of the k iterations. The remaining k-1 subsets are used for model training and fine-tuning, according to Polena (2017). The positive aspects include reducing the impact of data dependency since the risk of a classifier's performance depending on the testing set's choice is mitigated and guarantees the results legitimacy.

Kuhn and Johnson (2013) state that the choice of k is usually 5 or 10, but there is no formal rule. As k gets larger, the difference between the training set and the resampling set becomes smaller, meaning that a higher value of k leads to a less biased model; however, large variance might lead to overfitting. In this dissertation, given the size of the dataset, the chosen k was 5.

The following figure presents a representation of five-fold cross-validation. The dataset was divided into ten parts, and eight of them represent the training data and two represent the test data. The average value E of the five-group test results is calculated as an estimate of the model accuracy and is used as a performance indicator for the current K-fold cross-validation model (Niu et al., 2018).

Figure 18 - K-fold Cross-Validation diagram



Source: Author's prepared based on Niu et al. (2018).

3.8.3. Class reweigh

Dealing with an imbalanced dataset to obtain an equilibrium between the default and non-default observations in our training dataset can be handled through several approaches. One of the options is to collect more representatives. However, this is not possible in this case. Thus, the most reasonable alternative is to generate synthetic data or perform either undersampling or oversampling.

Undersampling consists of sampling from the majority class to keep only part of these points while oversampling consists of replicating some points from the minority class to increase its cardinality. Oversampling consists of reproducing some points from the minority class to increase its cardinality (Rocca, 2019).

The methodology chosen for our modelling was a combination of both random sampling techniques. Meaning that a modest amount of oversampling is applied to the minority class, improving the bias to the minority category. The majority class also uses a fair amount of undersampling to reduce the bias on the majority class. Implementing this technique can partially avoid the problems of increased learning time and lost information from deleting observations. This step was implemented using the imbalanced-learn framework—both oversampling and undersampling techniques used as sampling strategy attributes.

3.8.4. Hyperparameters optimisation

The tuning of the hyperparameters was performed using a cross-validated grid search. It selects the parameters on a specified parameter grid, maximising the underlying estimator's score (Pedregosa et al., 2011). The performance obtained in each combination of hyperparameters measured using the accuracy metric.

All the hyperparameters were explained in section 3 at the end of the explanation of each classifier.

The hyperparameters value given as input for the grid search is stated in the tables below (Table 3). The tuned hyperparameters column corresponds to the parameters that were used to train the

final model. The set of values column corresponds to the values used to perform the optimisation, while the Hyperparameter column refers to the name of parameters used.

Table 3 - Hyperparameters optimisation

KNN		Tuned Parameters
Hyperparameter	Set of Values	
n_neighbors	3,5,7,9,11,13	13
metric	minkowski	minkowski
p	1,2,3	1
Logistic Regression		**Tuned Parameters**
Hyperparameter	Set of Values	
alpha	10**-5, 10**-1, 10**2	1e-05
penalty	l1, l2	l1
Decision Tree		**Tuned Parameters**
Hyperparameter	Set of Values	
Criterion	Gini, entropy	Gini
Min_samples_split	2 ,4 ,6 ,8 ,10 ,15	15
Random Forest		**Tuned Parameters**
Hyperparameter	Set of Values	
N_estimators	50	50
Class_weight	0:1, 1:1	0: 1, 1: 1
XGBoost		**Tuned Parameters**
Hyperparameter	Set of Values	
N_estimators	50, 100, 150, 200, 250, 500	200
Learning_rate	0.05, 0.01, 0.5, 0.1, 1	0.5

Source: Author's Preparation.

4. EMPIRICAL RESULTS

4.1.1. Detailed results of the Hyperparameters optimisation

In Table 4 we report the accuracy results for each method before and after tuning procedures. The results show that the largest accuracy increase in the accuracy occurred in the KNN algorithm. In comparison, the smallest increases belong to the Logistic Regression. The algorithm with the highest accuracy is XGBoost. Note that this accuracy was computed against the training set. The hyperparameters used to calculate the before tuning accuracy were the default values.

Table 4 - Accuracy before and after tuning

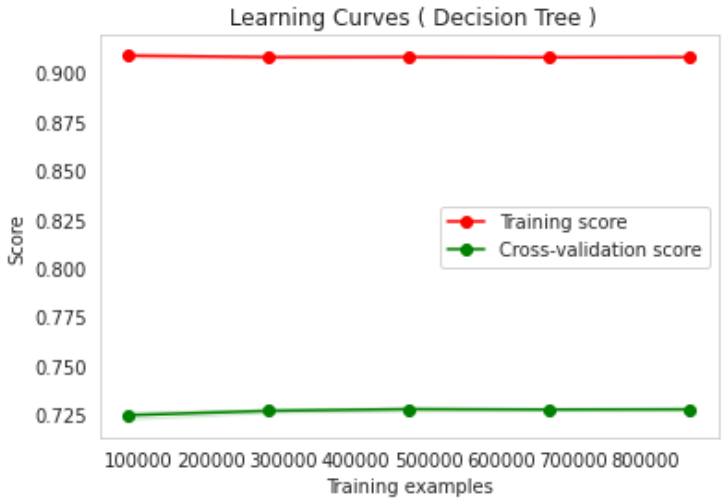
	Accuracy before tunning	Accuracy after tunning	%
XGBoost	80.22%	80.39%	0.19%
KNN	77.24%	79.20%	2.54%
Decision tree	72.07%	72.80%	1.02%
Logistic Regression	80.09%	80.11%	0.02%
Random Forest	79.42%	80.07%	0.82%

Source: Author's Preparation.

Learning curves are deemed valuable tools for monitoring workers' performance exposed to a new task (Anzanello & Fogliatto, 2011).

In the figure below, it is possible to observe the Learning curve of the Decision Tree that shows the score's evolution throughout different subsets of the training data.

Figure 19 - Learning curve (Decision Tree)



Source: Author's Preparation.

As we increase the training samples, the model is learning at a slow rate, and in the end, there is a large gap between the training curve and cross-validation. What can be concluded from this is that the model is overfitting the data. Zhou (2013) defines overfitting as the phenomenon that the learning result performs well on training data but poorly on the test data, caused by the learning

approach fitting the training data too much. The learning result has also captured some malign particularities that prevent a good generalisation. A few techniques can be used to limit overfitting, and one of them was used in our case, using a resampling technique, k-fold cross-validation, to estimate model accuracy.

4.1.2. Detailed results of the evaluation measures

In table 5, the six performance measures are present for the five algorithms. All the values are presented as a percentage.

Table 5 - Results

%	Decision Tree	Logistic Regression	Random Forest	XGBoost	KNN
AUC Train set	72.80469	69.87798	80.12744	80.37246	79.17803
AUC	55.20886	69.84504	69.35870	53.16313	53.01979
Sensitivity	81.99834	80.17612	80.83173	81.16422	81.09768
Specificity	29.51912	57.61099	51.68785	56.44313	40.68839
Precision	29.51912	57.61099	51.68785	56.44312	40.68839
PCC	72.77505	80.09678	80.03917	80.48331	79.21706
F1	72.15699	71.48866	73.57574	74.04255	73.66623

Source: Author's Preparation.

To understand the results of the performance measures, Table 6 presents the average of all the performance metrics across all classifiers.

Table 6 - Evaluation measures' average

Evaluation Measure	Average (%)
AUC Train set	76.47212
AUC	60.11910
Sensitivity	81.05362
Specificity	47.18990
Precision	47.18989
PCC	78.52227
F1	72.98603

Source: Author's Preparation.

Concerning the AUC calculated in the Train set, the Decision Tree and Logistic Regression presented a performance below average. Regarding the AUC, there are three classifiers below average – KNN, XGBoost and Decision Tree. Concerning the sensitivity, three classifiers outperform the average – Decision Tree, XGBoost and KNN. In terms of specificity and precision, KNN and Decision Tree are below average. While in terms of PCC, the Decision Tree was the only classifier below average. In terms of the F1 score, the average was superior to the Logistic Regression and Decision Tree

performance. The sensitivity measurement presented the highest average (81.05%), while the AUC curve presented the lowest (60.12%).

4.1.3. Global Performance Analysis

To have the ability to compare and rank our classifiers throughout every performance metrics, we have built Table 7. This table displays the ranking position for every performance measurement and the global performance for each classifier. The best performance classifier gets ranking one while the second-best gets ranking number two and so on, which means that the classifier with the least total amount of ranking is the best.

Table 7 - Model's global performance analysis

	Decision Tree	Logistic Regression	Random Forest	XGBoost	KNN
AUC Train set	4	5	2	1	3
AUC	3	1	2	4	5
Sensitivity	1	5	2	3	4
Specificity	5	1	3	2	4
Precision	5	1	3	2	4
PCC	5	2	3	1	4
F1	4	5	3	1	2
Total	**27**	**20**	**18**	**14**	**26**
Average	**3.86**	**2.86**	**2.57**	**2.00**	**3.71**
Ranking	#5	#3	#2	#1	#4

Source: Author's Preparation.

To sum up, XGBoost is the best classifier, followed by Random Forest, Logistic Regression. The worst performance models are KNN and Decision Tree. Compared to the benchmark model, the Logistic Regression, XGBoost and Random Forest have superior performance, while KNN and Decision Tree have weaker performance.

In order to make a comparative analysis with similar studies and cross matching the results of Table 1 and Table 7, it is possible to infer that similar to what happens in Lessmann et al. (2015), Malekipirbazari and Aksakalli (2015), Thanawala (2019) where Random Forest delivers superior results, in this case, Random Forest is placed in second place. Besides, ElMasry (2019), similarly classified the KNN and Decision Tree as the less accurate classifiers.

Although the use of XGBoost has not been present in most related studies, Xia et al. (2017), the XGBoost-RS classifier with the best predictive power followed by XGBoost-TPE and XGBoost-GS.

When comparing in terms of percentage, Malekipirbazari and Aksakalli (2015) presents a Random Forest AUC score (around 2% higher) than ours, while the KNN presents a value of 4% higher, and the Logistic Regression presents a value 3% lower. In Polena (2017), the AUC score of Random Forest and Logistic Regression equals ours and the KNN's value is 20% higher.

Where most studies lack detail in the presentation of the results instead of presenting just one measure as "accuracy", the studies should include other types of performance measure similar to

what is presented in this research or another example such as Amaro (2020). Trying to approach newer and less institutionalised classifiers and learning approaches would also be interesting and could bring value or also using different ramifications of the same classifiers and study the effect on the performance measures.

Table 8 presents the pros and cons of each classifier as well as a sum-up of its performance. The Pros and cons column represent with a "+" sign the upside of each classifier and with a "-" sign the downside of each classifier.

Table 8 - Key Findings

Total Ranking	Name	Performance	Pros & Cons
1	XGBoost	High	+ Fast training time + Outliers have minimal impact - Difficult interpretation
2	Random Forest	High	+ Good performance on imbalanced data + Good at handling high dimensional datasets + No overfitting + Useful to extract feature importance - Black Box
3	Logistic Regression	Medium	+ Model's interpretability + Simple to implement + Tuning of the hyperparameter not needed
4	K-Nearest Neighbour	Low	- Long execution time - Easy model training + No assumptions about data + Need to only tune one hyperparameter
5	Decision Tree	Low	- Prone to overfitting + Normalization not needed + Easy to explain

Source: Author's Preparation.

What can be concluded from this table is that each method has its up and downsides and that is worth it to try combining different methods because it is always useful for the researcher to have a broader perspective of the data and the results.

5. CONCLUSION

Credit scoring models have been around since the beginning of times as a subject of great importance, but for a long time, there were no ground-breaking developments. However, the big data revolution is changing this. The world is facing uncertain times, and it is still unknown the future impact of the COVID19 pandemic on the credit market. The industry must use all the resources available to ensure the credit market's stability and liquidity. One of these resources is alternative lending and the use of analytics and artificial intelligence. The peer-to-peer lenders offer unsecured loans, so, therefore, to ensure the survival of these organisations, the underwriting policies and credit risk monitoring should be very thorough.

Peer-to-peer lenders have explored uncharted territory in terms of credit risk, while introducing several positive novelties that allow for better customer experience due to the faster processes, less bureaucracy and reduction of costs, given lower operational expenses because the platforms are not required, in some cases, to respect and follow the same solvency and regulatory guidance as traditional banks, for instance, they do not need to respect bank capital requirements or pay fees associated with state deposit insurance practices. On the other hand, they are much more susceptible to other types of hazards.

In order to achieve the purpose of this dissertation and be able to infer if it's worth changing the credit scoring paradigm for alternative options and model the probability of default in a peer-to-peer lending context. In other words, studying the possible replacement of the benchmark model, the Logistic Regression, for different types of classifiers in terms of predicting the loans' probability of default. This research underwent rigorous pre-processing, k-fold cross-validation, and hyperparameter tuning to ensure the robustness of the results. The data used for this approach belongs to Lending Club, a US peer-to-peer lending platform. Given the complexity of the data, it was extremely important to understand how the credit scoring process is conducted, how the industry works, studying the variable importance and understand how can the pre-processing stage could be adapted to each model while making sure that state-of-the-art results can be obtained. As with most credit datasets, the Lending Club dataset is also unbalanced therefore a class reweight technique was implemented. Besides, it was necessary to conduct feature engineering so that the data is consistent, deal with missing values, and define criteria for feature selection. After the modelling, it was time to evaluate the different classifiers with several performance measures and compare the results with related studies.

Through this research, we concluded that the model XGBoost outperforms the industry's benchmark and that peer-to-peer lenders should be receptive to change. Also, another appealing factor is that this algorithm can be easily implemented using a library such as *Scikit-learn* in Python or either caret or *tidyverse* in R. Not only the results are more accurate in some performance measures, given a PCC of 80.4% versus a corresponding value of 80.1% regarding the Logistic Regression, but the computing time is less demanding. Nevertheless, it is essential to mention that the Logistic Regression outperformed two of the five classifiers. However, it is also perceptive to understand that these improvements may not be enough to change the paradigm and is crucial to continue the study and develop techniques that may offer greater advantages that can be more easily explainable and understandable.

What can be inferred in terms of limitations is the use of one dataset and considering two particular types of outcomes, either default or no-default, not performing financial impact analysis nor studying different alternatives in terms of pre-processing which one possible example could have been combining different class reweight or hyperparameter tuning techniques. Restricting the loan data to the accepted loans and not taking advantage of the rejected loans as well. Some researches have also shown promising results with outlier treatment techniques such as isolation forest, local outlier factor or stochastic outlier selection.

Additionally, the value could have been added if the analysis could have also transposed the current and historic market conditions and consequently, it could have been inferred if and how that affected the distinction between good and bad borrowers, moreover the credit scoring. If we could access more recent data, it would have been interesting to study the effect of the COVID-19 pandemic and how the market reacted to the increase in unemployment and overall uncertainty.

Beyond this, if the data contained more information regarding the borrower's behaviour and psychological traits, either for example, social media data or spending habits, that could have provided some insights and predictive capacity in terms of modelling the probability of default.

The main difficulty was related to computational power given the size of the dataset and the demands for each classifier. If this had not been an issue, it would have been possible to deepen the methodology in terms of evaluating more thoroughly what different types of pre-processing or modelling techniques could have impacted the final result.

In terms of recommendations for future works, it would be attractive to combine the same research datasets from different peer-to-peer lenders so that the study could envision a broader picture of the sector. Adding other variations of the ensemble learning approach would also add more value or retrieving information from the rejected loans, to learn from both sides of the spectrum, and study if it increases the models' predictive capacity. Another example could to not only consider two possible outcomes for the classification models but also consider that sometimes there might be scenarios that were not represented in this case, for instance including transform the default variable into a continuous variable and consider not only 0 and 1 but also taken into consideration values between these numbers that correspond to scenarios such as a debt settlement plan or an alternative payment plans.

Evaluating the impact of the increasing regularisation in the sector would also be interesting and studying its impact regarding underwriting policies, credit scoring models, sustainability of the market or even model the probability of a loan being accepted under these circumstances.

To conclude, according to experts, the next big thing in Artificial Intelligence is Deep learning hence using this in a credit scoring and classification context and comparing its performance with ensemble classifiers to assess if this is suitable for the future of credit scoring would also be something of great value in terms of providing new insights and the by having the capacity to use more data.

6. REFERENCES

Ashofteh, A., & Bravo, J. M. (2021a). A Conservative Approach for Online Credit Scoring. Expert Systems with Applications, Volume 176, p. 1-16, 114835. https://doi.org/10.1016/j.eswa.2021.114835

Ashofteh, A. & Bravo, J. M. (2021b). Spark Code: A Novel Conservative Approach for Online Credit Scoring [Source Code]. https://doi.org/10.24433/CO.1963899.v1. Associated Publication: "A Conservative Approach for Online Credit Scoring", Expert Systems with Applications, https://doi.org/10.1016/j.eswa.2021.114835.

Akindaini, B. (2017). *Machine Learning Applications in Mortgage Default Prediction* (Issue November) [University of Tampere]. https://trepo.tuni.fi/bitstream/handle/10024/102533/1513083673.pdf

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(1), 589–609. https://doi.org/10.1111/j.1540-6261.1974.tb00057.x

Amaro, M. M. (2020). *Credit Scoring: Comparison of Non-Parametric Techniques against Logistic Regression*. [Nova Information Management School] http://hdl.handle.net/10362/99692

Anderson, R. (2007). *The Credit Scoring Toolkit - Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press. https://ideas.repec.org/b/oxp/obooks/9780199226405.html

Anzanello, M. J., & Fogliatto, F. S. (2011). Learning curve models and applications: Literature review and research directions. *International Journal of Industrial Ergonomics*, *41*(5), 573–583. https://doi.org/10.1016/j.ergon.2011.05.001

Arroyo, J. M., Chiapolino, M., Freiman, M., Gabruashvili, I., & Pancaldi, L. (2020). *A fast-track risk-management transformation to counter the COVID-19 crisis*. *October*, 5–6. Retrieved from https://www.mckinsey.com/business-functions/risk/our-insights/a-fast-track-risk-management-transformation-to-counter-the-covid-19-crisis

Aveni, T., Qu, C., Hsu, K., Zhang, A., & Lei, X. (2015). New Insights Into An Evolving P2P Lending Industry: how shifts in roles and risk are shaping the industry. In *Positive Planet Group* (Issue August). https://www.findevgateway.org/paper/2015/08/new-insights-evolving-p2p-lending-industry

Ayuso, M., Bravo, J. M. & Holzmann, R. (2021a). Getting Life Expectancy Estimates Right for Pension Policy: Period versus Cohort Approach. Journal of Pension Economics and Finance, 20(2), 212–231. https://doi.org/10.1017/S1474747220000050

Ayuso, M., Bravo, J. M., Holzmann, R. & Palmer, E. (2021b). Automatic indexation of the pension age to life expectancy: When policy design matters. Risks, in press.

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehman, M., & Tiburtius, P. (2011). Online Peer-to-Peer Lending - A Literature Review. *Journal of Internet Banking and Commerce*, *16*(2), 1–14. https://doi.org/10.1007/978-3-531-92534-9_12

Bastos, J. (2008). Credit scoring with boosted decision trees. *Munich Personal RePEc Archive*, *8156*. https://mpra.ub.uni-muenchen.de/8156/

Bravo, J. M. (2019). Funding for longer lives: Retirement wallet and risk-sharing annuities. Ekonomiaz, 96(2), 268-291.

Bravo, J. M. (2020). Longevity-Linked Life Annuities: A Bayesian Model Ensemble Pricing Approach. Atas da Conferência da Associação Portuguesa de Sistemas de Informação, CAPSI 2020 Proceedings, 29. https://aisel.aisnet.org/capsi2020/29

Bravo, J. M. (2021). Pricing Participating Longevity-Linked Life Annuities: A Bayesian Model Ensemble approach. European Actuarial Journal. https://doi.org/10.1007/s13385-021-00279-w

Bravo, J. M., Ayuso, M. (2020). Previsões de mortalidade e de esperança de vida mediante combinação Bayesiana de modelos: Uma aplicação à população portuguesa. RISTI - Revista Iberica de Sistemas e Tecnologias de Informação E40, 128-144. DOI: 10.17013/risti.40.128–145.

Bravo, J. M., Ayuso, M. (2021). Forecasting the retirement age: A Bayesian Model Ensemble Approach. In: Rocha Á., Adeli H., Dzemyda G., Moreira F., Ramalho Correia A.M. (eds) Trends and Applications in Information Systems and Technologies, pp 123-135. WorldCIST 2021. Advances in Intelligent Systems and Computing, vol 1365. Springer, Cham. https://doi.org/10.1007/978-3-030-72657-7_12

Bravo, J. M., & Nunes, J. P. V. (2021a). Pricing Longevity Derivatives via Fourier Transforms. Insurance: Mathematics and Economics, 96, 81-97.

Bravo, J. M., Ayuso, M., Holzmann, R., & Palmer, E. (2021b). Addressing the Life Expectancy Gap in Pension Policy. Insurance: Mathematics and Economics, 99, 200-221. https://doi.org/10.1016/j.insmatheco.2021.03.025

Bravo, J. M., Ayuso, M., Holzmann, R., & Palmer, E. (2021c). Intergenerational actuarial fairness when longevity increases: Amending the retirement age. Scandinavian Actuarial Journal, Preprint to submit.

Brown, I., & Mues, C. (2012). Expert Systems with Applications An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Carney, M. (2017). The Promise of FinTech – Something New Under the Sun? Speech given by Governor of the Bank of England Chair of the Financial Stability Board Deutsche Bundesbank G20 conference on "Digitizing finance, financial inclusion and. *Deutsche Bundesbank G20 Conference on "Digitising Finance, Financial Inclusion and Financial Literacy", Wiesbaden*, *January*, 1–14. Retrieved from https://www.bankofengland.co.uk/speech/2017/the-promise-of-fintech-something-new-under-the-sun%0Ahttp://www.nber.org/papers/w22476.pdf

Carvajal, G., Maučec, M., & Cullick, A. (2018). *Intelligent Digital Oil & Gas Field Concepts, Collaboration and Right-time decisions*. Gulf Professional Publishing. https://doi.org/https://doi.org/10.1016/B978-0-12-804642-5.00004-9

Chamboko, R., & Bravo, J. M. (2016). On the modelling of prognosis from delinquency to normal performance on retail consumer loans. Risk Management, 18(4), 264–287. https://doi.org/10.1057/s41283-016-0006-4.

Chamboko, R. & Bravo, J. M. (2019a). Modelling and forecasting recurrent recovery events on consumer loans. International Journal of Applied Decision Sciences, Vol. 12, No. 3, 271-287.

https://doi.org/10.1504/IJADS.2019.10019811.

Chamboko, R. & Bravo, J. M. (2019b). Frailty correlated default on retail consumer loans in developing markets. International Journal of Applied Decision Sciences, Vol. 12, No. 3, 257–270. https://doi.org/10.1504/IJADS.2019.10019807.

Chamboko, R., & Bravo, J. M. (2020). A Multi-State Approach to Modelling Intermediate Events and Multiple Mortgage Loan Outcomes. *Risks*. https://doi.org/https://doi.org/10.3390/risks8020064

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *ArXivLabs*, *13-17-August*, 785–794. https://doi.org/10.1145/2939672.2939785

Cox, D. (1958). Journal of the Royal Statistical Society. *Journal of Royal Statistical Society*, *10*(2), 215–242. https://doi.org/10.1002/0471667196.ess7018

Dhingra, C. (2020). *A Visual Guide to Gradient Boosted Trees (XGBoost)*. Towards Data Science. [Blog post]. Retrieved February 21, 2021, from https://towardsdatascience.com/a-visual-guide-to-gradient-boosted-trees-8d9ed578b33

Elkan, C. (2012). Maximum Likelihood, Logistic Regression, and Stochastic Gradient Training. *Tutorial Notes at CIKM*, *7–9*, 11. http://www.ats.ucla.edu/stat/stata/dae/mlogit.htm

ElMasry, M. (2019). *Machine learning approach for credit score analysis: a case study of predicting mortgage loan defaults*. [Nova Information Management School] http://hdl.handle.net/10362/62427

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, *47*(1), 54–70. https://doi.org/10.1080/00036846.2014.962222

Garson, G. D. (2012). Discriminant Function Analysis. In *Statistical Associates Publishing*. http://www.statisticalassociates.com/discriminantfunctionanalysis.htm

Halko, N., Martinsson, P. G., & J.A, T. (2009). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev., Survey and Review Section*, *53*(2), 217–288. https://arxiv.org/abs/0909.4061v2

Härdle, W. K., Lu, H. H.-S., & Shen, X. (2018). *Handbook of Big Data Analytics*. Springer International Publishing. https://www.springer.com/gp/book/9783319182834

Herrero-Lopez, S. (2009). Social Interactions in P2P Lending. *Association for Computing Machinery*, *09*. https://doi.org/10.1145/1731011.1731014

Herzenstein, M., & Andrews, R. L. (2008). The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities. *Boston University School of Management Research Paper*, *14*(6), 1–45. Retrieved from http://www.prosper.com/downloads/research/democratization-consumer-loans.pdf

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2009). Screening in New Credit Markets: Can individual lenders infer borrower creditworthiness in Peer-to-peer lending. *SSRN Electronic Journal*, *15242*(AFA 2011 Denver Meetings Paper). https://doi.org/10.2139/ssrn.1570115

Jagtiani, J., & Lemieux, C. (2018). Do fintech lenders penetrate areas that are underserved by traditional banks? *Journal of Economics and Business*, *100*(March), 43–54.

https://doi.org/10.1016/j.jeconbus.2018.03.001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R* (8th ed.). Springer New York Heidelberg Dordrecht London. https://doi.org/10.1007/978-1-4614-7138-7

Kapil, D. (2019). *Stochastic vs Batch Gradient Descent*. Medium. [Blog post] Retrieved March 10, 2021, from https://medium.com/@divakar_239/stochastic-vs-batch-gradient-descent-8820568eada1

Kuhn, M., & Johnson, K. (2013). Applied predictive modelling. In *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

Lee, E., & Lee, B. (2012). Herding behavior in online P2P lending: An empirical investigation. *Electronic Commerce Research and Applications*, *11*(5), 495–503. https://doi.org/10.1016/j.elerap.2012.02.001

Lee, T. S., & Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, *28*(4), 743–752. https://doi.org/10.1016/j.eswa.2004.12.031

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using Semi-supervised Support Vector Machines. *Expert Systems with Applications*, *74*, 105–114. https://doi.org/10.1016/j.eswa.2017.01.011

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, *42*(10), 4621–4631. https://doi.org/10.1016/j.eswa.2015.02.001

Mezei, J., Byanjankar, A., & Heikkilä, M. (2018). Credit Risk Evaluation in Peer-to-peer Lending with Linguistic Data Transformation and Supervised Learning. *Proceedings of the 51st Hawaii International Conference on System Sciences*, *9*, 1366–1375. https://doi.org/10.24251/hicss.2018.169

Molnar, C. (2019). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/

Munusamy, J., Run, E. C., Chelliah, S., & Annamalah, S. (2013). Adoption of Retail Internet Banking: A Study of Demographic Factors. *Journal of Internet Banking and Commerce*, *17*(3). https://www.researchgate.net/publication/256051102_Adoption_of_Retail_Internet_Banking_A_Study_of_Demographic_Factors

Nigmonov, A., Shams, S., & Alam, K. (2020). *Born in Crisis: Early Impact of COVID-19 Pandemic on P2P Lending Market*. 1–38. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3721406

Nisbet, R., Miner, G., & Yale, K. (2018). *Handbook of Statistical Analysis and Data Mining Applications* (K. Y. Robert Nisbet, Gary Miner (ed.); Second Edi). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-416632-5.00009-8

Niu, M., Li, Y., Wang, C., & Han, K. (2018). RFAmyloid: A web server for predicting amyloid proteins. *International Journal of Molecular Sciences*, *19*(7).

https://doi.org/10.3390/ijms19072071

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *127*(12), 3–4. https://doi.org/10.1289/EHP4713

Polena, M. (2017). *Performance Analysis of Credit Scoring Models on Lending Club Data* (Vol. 39, Issue 3) [Charles University]. http://dx.doi.org/10.1016/j.eswa.2011.08.093

Redding, D. W., Lucas, T. C. D., Blackburn, T., & Jones, K. E. (2017). Evaluating Bayesian spatial methods for modelling species distributions with clumped and restricted occurrence data. In *bioRxiv*. https://doi.org/10.1101/105742

Rocca, B. (2019). *Handling imbalanced datasets in machine learning*. Towards Data Science. [Blog Post] Retrieved January 24, 2021, from https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28

Saha, S. (2018). *Understanding the log loss function of XGBoost*. Medium. [Blog Post] Retrieved February 21, 2021, from https://medium.datadriveninvestor.com/understanding-the-log-loss-function-of-xgboost-8842e99d975d

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *44*(1–2), 207–219. https://doi.org/10.1147/rd.441.0206

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PLoS ONE*, *10*(10), 1–22. https://doi.org/10.1371/journal.pone.0139427

Shim, Y., & Shin, D. (2015). Analyzing China's Fintech Industry from the Perspective of Actor-Network Theory. *Telecommunications Policy*, 1–14. https://doi.org/10.1016/j.telpol.2015.11.005

Thanawala, D. D. (2019). *Credit Risk Analysis using Machine Learning and neural networks*. http://arxiv.org/abs/1907.03044

Thompson, B. S. (2017). Can Financial Technology Innovate Benefit Distribution in Payments for Ecosystem Services and REDD +? *Ecological Economics*, 1–8. https://doi.org/10.1016/j.ecolecon.2017.04.008

Tomlinson, N., Foottit, I., & Doyle, M. (2016). Marketplace lending: A temporary phenomenon? In *Deloitte LLP*. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/de/Documents/financial-services/deloitte-uk-fs-marketplace-lending.pdf

Wang, H., & Greiner, M. E. (2011). Prosper — The eBay for Money in Lending 2.0. *Communications of the Association for Information Systems*, *29*. https://doi.org/10.17705/1CAIS.02913

Weiss, G. N. F., Pelger, K., & Horsch, A. (2012). Mitigating Adverse Selection in P2P Lending – Empirical Evidence from Prosper.com. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1650774

Xia, Y. (2019). A Novel Reject Inference Model Using Outlier Detection and Gradient Boosting Technique in Peer-to-Peer Lending. *IEEE Access*, *7*, 1–2. https://doi.org/10.1109/ACCESS.2019.2927602

Xia, Y., Liu, C., Li, Y. Y., & Liu, N. (2017). A boosted Decision Tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Yiu, T. (2019). *Understanding Random Forest*. Towards Data Science. [Blog Post] Retrieved March 10, 2021, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Zhang, H., & Liu, Q. (2019). Online learning method for drift and imbalance problem in client credit assessment. *Symmetry*, *11*(7), 523–538. https://doi.org/10.3390/sym11070890

Zhou, X., Zhang, D., & Jiang, Y. (2008). A new credit scoring method based on rough sets and Decision Tree. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5012 LNAI*, 1081–1089. https://doi.org/10.1007/978-3-540-68125-0_117

Zhou, Z.-H. (2013). *Ensemble Learning*. 25–32. https://doi.org/10.1007/978-3-642-38652-7_3

Zhu, C., Luo, J., & Li, Z. (2020, August 14). China's Peer-to-Peer Lending Purge Leaves $115 Billion in Losses. Bloomberg News. Retrieved from https://www.bloomberg.com/news/articles/2020-08-14/china-s-peer-to-peer-lending-purge-leaves-115-billion-in-losses