Predicting loan status for Risk management in Peer-to-peer lending market: The Lending Club case

HAOUFADI, Aouatef
MAI, Thu Ha
TEHAR, Maria Yamina

22th January 2022

Contents

1 Introduction

1.1 Overview of the study

Peer-to-peer lending started out as a relatively simple online system for facilitating loans between individuals but since has grown into a complex ecosystem of technologies, institutions, and auxiliary start-ups. While by definition, the term "peer-to-peer" designates exchange between individuals, the term has increasingly become a misnomer for this industry, which is increasingly referred to as "marketplace" lending. Initially, borrowers could crowdfund loans by appealing to multiple small investors. But today, the majority of peer-to-peer loans are purchased by large investors like banks, hedge funds, and wealth management firms. The entry of these investors has motivated the growth of start-ups and other actors dedicated to advising investors, performing loan data analysis, and automating the investment process. The promise of disintermediation, or removing the banks from the equation, has given way to a wide range of intermediaries, including but not limited to banks.

Peer-to-peer lending is used to describe online marketplaces where lenders (also referred to interchangeably as investors) can lend to individuals or small businesses. In 2005, the first peer-to-peer lending platform, Zopa, was established in the UK, followed shortly in the U.S. by Prosper, Lending Club, and others. Today, there are over a dozen peer-to-peer lending companies in the U.S., but Lending Club, headquartered in San Francisco, California, is the largest. It is an online lending platform connecting borrowers and investors, where borrowers are able to obtain loans and investors can purchase notes backed by payments based on loans. It allows consumers and small business owners to lower the cost of their credit and enjoy a better experience than traditional bank lending, and investors earn attractive risk-adjusted returns.

Peer-to-peer lending is also pushing the sector of finance toward greater credit accessibility in many ways. Making credit available for small businesses means economic growth, support for local entrepreneurs, and potentially more job opportunities. It has also helped free up funds for consumers that would otherwise be exploited by credit card debt. These funds can be put to good use such as investing, which is beneficial to the economy.

However, peer-to-peer lending is quite risky for investors since there is credit risk, liquidity risk and even platform-collapse risk. The most important one is the credit risk of borrowers. Since many loans come without collateral, it is difficult for the platform to freeze the other assets of borrowers when it defaults, which leads to low loan recovery rates and high recovery costs.

Loan status prediction, especially for loan defaulters, has been a significant problem in the financial domain because overdue loans may incur significant losses. Machine learning methods have been introduced to solve this problem, but there are still many challenges including feature multicollinearity and imbalanced labels, etc.

In this study, we used a dataset found on Kaggle which was retrieved from the official website of Lending Club company. This dataset contains 396,030 observations and 27 features in the period from 2007 to 2015. Our objective is (1) to predict the loan status based on the applicants' profiles using Logistic Regression as a statistical model and Random Forest as a machine learning model, (2) after dealing with the multicollinearity problem using two Information Reduction methods, Elastic Net and Autometrics, (3) and then to identify the driving factors behind loan default, or in other words, the variables which are strong indicators of charged-off loans.

The rest of this report is composed as follows. In the next part, "Materials and Methods", we will explain briefly the dataset and its features, then give some remarks about the data preprocessing and introduce the theoretical backgrounds of all methods, empirical strategy and metrics we used in the following part, "Results and Discussion". In this part, we will highlight some key findings in three sections, Exploratory Data Analysis, Information Reduction and Prediction models. The report will then follow with a conclusion where we summarize our key findings, discuss our work's contribution and suggest further research. The References and Annex can be found afterwards.

1.2 Literature review

One of the most well-liked research subjects in recent years has been the development of decision-making models for peer-to-peer lending. For example, in order to explore the likelihood of defaults, Emekter et al. (2015) used logistic regression. They discovered some correlations between loan defaults and credit scores, debt-to-income ratios (DTI), FICO scores, and revolving credit lines. Another study by Malekipirbazari and Aksakalli (2015) classified good and poor loans using various machine learning techniques, including support vector machines, logistic regression, random forest, and k-nearest neighbour. They discovered that utilizing a machine learning strategy is significantly more efficient than relying on the current financial criteria, such as FICO and LC grades, that the Lending Club offers to assist lenders in making loan investment decisions.

To assist banks in deciding whether or not to issue a credit to loan applicants, a number of decision support systems based on credit scoring models have been created. Credit scoring uses two different types of predictive models: statistical techniques and artificial intelligence techniques. Logistic regression and linear discriminant analysis are examples of statistical techniques. Some of the artificial intelligence techniques are decision trees, random forests, support vector machines, neural networks, and naive Bayes. Despite the development of more sophisticated AI methods for assessing the credit risk of borrowers, straightforward statistical procedures like logistic regression and linear discriminate analysis continue to be popular due to their high accuracy and simplicity, according to several studies by Ala'raj et al. (2016), Byanjankar et al. (2015), Siami et al. (2011) (2014).

Studies that combine machine learning with an explainability or interpretability study are quite uncommon, nevertheless. In this regard, Jin and Zhu (2015) suggested using artificial neural networks, decision trees, and SVM to predict the probability of default. Still, they only included a relative importance analysis of the variables for the artificial neural network and decision tree methodologies as an element of interpretability. Similarly, Li et al. (2018) compared various ensembles and individual algorithms using a collection of techniques including XGBoost, deep neural networks, and logistic regression. In their comparison, they classified their ideas according to interpretability. They showed that the ensemble of the three approaches, which is the best prediction methodology, performs poorly when it comes to interpretation. In addition, while using XGBoost technique, the researchers also included a feature importance analysis.

On the other hand, works that incorporate new variables into a logistic regression model using artificial intelligence (or other cutting-edge techniques) can be determined to have interpretability. The interpretability of the new variables' coefficients is examined, and statistical inference is used to confirm the significance. The study by Yao et al. (2019) is an example of this. For the purpose of a loan, a logistic regression model is provided with variables extracted through text mining. Similar to this, Ahelegbey et al. (2019) created segments using latent factors models and connectivity networks and then estimated various logistic regression models for a collection of small and medium-sized businesses using those segments. However, the model is a statistical one in both instances, and its meaning is the same.

The understandability of analytical models for credit risk analysis becomes increasingly crucial, according to Andriosopoulos et al. (2019), especially from a supervisory perspective. However, as evidenced by the Explainable Machine Learning Challenge sponsored by FICO, a well-known provider of credit scores, private businesses are focusing on interpretability, according to Chen et al. (2018) and Holter et al. (2020). The SHAP values were then used by Ariza-Garzón et al. (2020) to describe machine learning models in the context of credit risk in P2P lending.

2 Material and Methods

2.1 Dataset and Feature Understanding

The data we used for this study contains information about past loan applicants and whether they defaulted or not. The period for the loan covers from 2007 to 2015 with the geographical area of

the borrowers covering different states of the U.S.A. The dataset is retrieved from the official website of Lending Club which was found on Kaggle [1]. It has 396,030 observations and 27 features. In this study, the primary target is to predict if a loan will result in default or non-default based on borrowers' profiles. For this, we select the current loan status as the target variable with two classes as follows.

- Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
- Charged-off: Applicant has not paid the instalments in due time for a long period of time (defaulted on the loan)

The dataset includes detailed records of the borrowers who applied for the loan. The raw dataset includes a large dimension that captures various sources of information. However, in general, the variables can be divided into two segments:

- Features related to the borrowers (e.g., employment title, employment length, home ownership, annual income, verification status, address)
- Features related to the loan characteristics (e.g., loan amount, interest rate, term, instalment, grade, subgrade, purpose, title, DTI).

The data dictionary is provided in Table 1 below.

No.	Features	Description
0	loan_amnt	The listed amount of the loan applied for by the borrower.
1	term	The number of payments on the loan: 36 or 60 months
2	int_rate	Interest Rate on the loan
3	installment	The monthly payment owed by the borrower if the loan originates
4	grade	LC assigned loan grade
5	sub_grade	LC assigned loan subgrade
6	emp_title	The job title supplied by the Borrower when applying for the loan
7	emp_length	Employment length in years. Possible values are between 0 and 10
8	home_ownership	The home ownership status provided by the borrower during registration
9	annual_inc	The self-reported annual income provided by the borrower during registration
10	verification_status	Indicates if income was verified by LC, not verified, or verified income source
11	issue_d	The month in which the loan was funded
12	loan_status	Current status of the loan
13	purpose	A category provided by the borrower for the loan request
14	title	The loan title provided by the borrower
15	zip_code	The first 3 numbers of the zip code extracted from address
16	address	The state provided by the borrower in the loan application
17	dti	A ratio between total monthly debt payments divided by monthly income
18	earliest_cr_line	The month the borrower's earliest reported credit line was opened
19	open_acc	The number of open credit lines in the borrower's credit file
20	pub_rec	Number of derogatory public records
21	revol_bal	Total credit revolving balance
22	revol_util	Revolving line utilization rate
23	total_acc	The total number of credit lines currently in the borrower's credit file
24	initial_list_status	The initial listing status of the loan. Possible values are W and F
25	application_type	Indicates whether the loan is an individual or joint application
26	mort_acc	Number of mortgage accounts
27	pub_rec_bankruptcies	Number of public record bankruptcies

Table 1: Data dictionary of the Lending Club set

[1] https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction/data

2.2 Data Preprocessing

Detection and Treatment of duplicates and missing values

As a first step, we checked for duplicate rows and missing values. The result shows that our data does not have any duplicate rows, while it has missing data in the following features: emp_title with 5.78 %, emp_length with 4.62 %, title with 0.44 %, revol_util with 0.07 %, mort_acc with 9.54 %, pub_rec_bankruptcies with 0.13 %.

After detecting these missing values, we did some exploratory data analysis to help us handle this problem. The mean for the mort_acc per total_acc value can be imputed for missing values in mort_acc since mort_acc correlates the most with total_acc. Rows with missing values in pub_rec_bankruptcies and revol_utils are removed because they account for less than 0.5 % of the total data. Additionally, we have dropped the title columns since this is simply a string subcategory/ description of the purpose column. In addition to that, we decided to delete the emp_title column because it has too many unique jobs to try to convert this to a dummy variable feature.

Feature Engineering

After examining our data, we decided to drop some unnecessary columns. For example, the grade column is dropped because it is just a sub-feature of the sub-grade column. We also dropped the issue_d column because it is considered as data leakage as we should not know beforehand whether a loan would be issued or not. In addition, we chose to create a zip_code column that is extracted from the address column, which turns out to be a critical feature for our prediction models in the following part.

After some basic cleaning, we converted our categorical features to numerical ones either by performing one hot-encoding or label encoding depending on the kind of data they represent. For example, label encoding was performed on sub_grade and as it is ordinal in nature, whereas one hot-encoding was performed on the other categorical features to turn them into dummy variables.

Detection and treatment of outliers

After splitting our data into training, validation and test sets, we move to one of the most important steps in data preprocessing, detecting and treating outliers. This process is vital as it can affect the statistical analysis and the training process of the machine learning algorithm. We chose to detect and treat the outliers only on the training set to avoid leaking information from the test set. We then followed the Inter Quartile Range (IQR) approach to filter and detect the outliers. IQR is the range between the first and the third quartiles of Q1 and Q3, and any data points lying beyond the upper limit $(Q3 + 3 * IQR)$ and below the lower limit $(Q1 + 3 * IQR)$. are considered outliers. We used the trimming method to treat the outliers since the outliers account for only 0.5 to 1 % of the training set.

2.3 Methodology

2.3.1 Information Reduction Methods

In our Correlational Analysis, we detected that our dataset has multicollinearity problem which is a common phenomenon in high dimensional settings, in which two or more predictor variables are highly correlated. To deal with it and handle high-dimensional setup, we decided to use Elastic Net and Autometrics methods.

2.3.1.1 Elastic Net

The Elastic Net method is a popular technique for parameter estimation and variable selection. Moreover, the elastic net method uses the adaptive weights on the penalty function based on the elastic net estimates. The adaptive weight is related to the power order of the estimator. Normally, this method focuses to estimate parameters in terms of linear regression models that are based on the dependent variable and independent variable as a continuous scale.

The Elastic Net regularisation is used for feature selection to minimize the total number of features. The main purpose of using elastic net regularisation is that it solves the limitations of Lasso

regression in losing the relevant and independent features. It adds a quadratic part to the penalty which is obtained after modifying the Ridge Regression.

Elastic Net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models. And this technique is most appropriate where the dimensional data is greater than the number of samples used.

2.3.1.2 Autometrics

Autometrics is a new algorithm for automatic model selection within the general-to-specific framework (also known as the 'Hendry' or 'LSE' methodology). It improves on previous implementations through an enhanced search method which essentially makes a presearch simplification unnecessary. In addition, the algorithm is presented within a likelihood framework, allowing for applications beyond regression models. The tuning parameters of the Autometrics algorithm are determined through Monte Carlo experiments, and evidence regarding its performance is presented.

Autometrics comprises the following five basic stages:

(i) In the first stage, the linear model known as the so-called general unrestricted model (GUM) is formed

(ii) In the second stage, the parameters are estimated along with testing the statistical significance of the GUM

(iii) In the third stage, the presearch process performed

(iv) The fourth stage produces the tree-path search

(v) In the last stage, the final model is selected

Doornik elaborated the entire algorithm of Auto-metrics whereas the steps to run Autometrics are as follows. Start off by considering all the candidate variables in a linear model (GUM), estimate it by the least-squares method, and then verify through diagnostic tests. In the case of insignificant coefficients, then simpler models are estimated utilizing a tree-path reduction search and validated by diagnostic tests. If some terminal models are detected, Autometrics undertakes its union testing. Rejected models are deleted, and the union of those terminal models that survived induces new GUM for another tree-path search iteration. This inspection procedure continues, and the terminal models are statistically assessed against their union. If two or more terminal models clear the encompassing tests, then the prechosen information criterion is a gateway to a final decision.

2.3.2 Prediction Methods

For prediction methods, we chose to use Logistic regression as a statistical model to predict the likelihood of defaulting on a loan. It is the most appropriate because the outcome we are trying to predict is a binary response, either they will default or they will not. And we used Random Forest over other machine learning algorithms because it gives a high classification accuracy and it is more efficient on large databases.

2.3.2.1 Logistic Regression

Logistic regression (LR) modelling has been a popular statistical tool in healthcare analysis and medical research for the last three decades. its origin was established in the 19th century, It is the most common statistical method to predict the dichotomous dependent variable using one or more than one independent variable. The French mathematician Pierre François Verhulst invented the logistic function in the 19th century for the description of the growth of the human population. Between 1838 and 1847 Verhulst published his suggestions which were edited by Quetelet. Pearl and Reed discovered a new logistic function in 1920 in the USA for a study of population growth.

Regression is a technique very commonly used to describe the existing relationship between a qualitative variable to be explained and one or more explanatory variables. And the logistic regression model makes it possible to estimate the strength of the association between a qualitative variable

with two variants (dichotomous) called the dependent variable and variables that can be qualitative or quantitative called the explanatory or independent variables. If the variable to be explained has only two modalities (variants), binary logistic regression is used. If it has more than two modalities and if these are not ordered, nominal polychotomous logistic regression must be used. Finally, if the variable to be explained has more than two modalities and these are ordered, the method to be used is ordinal polychotomous regression. Using logistic regression, we seek to estimate the probability of success of this variable by linearizing explanatory variables

Logistic regression models were created based on the number of categories in the dependent variable. On the one hand, a binary model was created for the dichotomous dependent variables and on the other hand, a multinomial model for the dependent variables with more than two modalities. All the explanatory variables must be introduced in each model. This is the top-down step-by-step method. Non-significant variables can either be kept in the model or removed by the Akaike Information Criterion (AIC) minimization procedure. This step-by-step method makes it possible to identify the independent variables of the model that better explain the dependent variable. The lower the AIC, the better the model. Since the analysis of the logistic regression results is done as for a simple linear regression, we will calculate each binary model's coefficient of determination (R^2). The coefficient of determination designates the part explained by the variance of the explanatory variables on the dependent variable. This coefficient of determination also called pseudo-R2 varies between 0 and 1, the value 1 reflecting the perfect fit of the model and 0 the poor fit of the model. The Logistic regression equation is the same as the multiple regression but adding the logarithmic transformation, it is like follows:

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 + ... + b_p)}}$$

Where:

- P(Y): the probability that Y occurs.
- b_0, ... , b_p: the coefficients of the model.

2.3.2.2 Random Forest

Random Forest is a famous machine learning algorithm that uses supervised learning methods and can be applied to both classification and regression problems. It is based on ensemble learning, which integrates multiple classifiers to solve a complex issue and increases the model's performance. It has several hyperparameters that have to be set by the user, such as the number of observations drawn randomly for each tree and whether they are drawn with or without replacement, the number of variables drawn randomly for each split, the splitting rule, the minimum number of samples that a node must contain, and the number of trees.

Random Forest can build prediction models using random forest regression trees, which are usually unpruned to give strong predictions. The bootstrap sampling method is used on the regression trees, which should not be pruned. Optimal nodes are sampled from the total nodes in the tree to form the optimal splitting feature. The random sampling technique used in selecting the optimal splitting feature lowers the correlation and hence, the variance of the regression trees. It improves the predictive capability of distinct trees in the forest. The sampling using bootstrap also increases independence among individual trees.

Random forest theory is based on the use of many decision trees that have been grown using the bagging method. Each tree plays as a weak learner in the algorithm and the combination of these weak inputs builds up a robust ensemble learning model. Results obtained from the RF model are based on the majority votes in the ensemble models.

Random Forest is one of the best high-performance strategies widely applied in numerous industries due to its effectiveness. It can handle data very effectively, whether it is binary, continuous, or

categorical. And it can accommodate missing values, making it an excellent solution for anyone who wants to create a model quickly and efficiently.

Random Forest uses the Gini index taken from the CART learning system to construct decision trees. The Gini index of node impurity is the measure most commonly chosen for classification-type problems. If a dataset T contains examples from n classes, the Gini(T) is defined as:

$$Gini(T) = 1 - \sum_{j=1}^{n} (p_j)^2$$

Where:

- p_j : the relative frequency of class j in T.

2.3.3 Empirical Strategy

2.3.3.1 Model Selection

Model selection is the process of selecting the machine learning model most appropriate for a given issue. It may be used to compare models of the same type that have been set up with various model hyperparameters and models of other types.

There are two main classes of techniques to approximate the ideal case of model selection; they are the probabilistic measures that involve analytically scoring a candidate model using both its performance on the training dataset and the complexity of the model, and the resampling methods that seek to estimate the performance of a model on out-of-sample data.

The probabilistic measures are as follows:

Akaike Information Criterion (AIC): it is a single numerical score that may be used to distinguish across many models the one that is most likely to be the best fit for a given dataset.

Minimum Description Length (MDL): according to the (MDL) the explanation that allows for the most data compression is the best given a small collection of observed data.

Bayesian Information Criterion (BIC): was derived using the Bayesian probability idea and is appropriate for models that use maximum likelihood estimation during training.

And the Resampling methods are as follows:

Cross-validation: it is a resampling procedure to evaluate models by splitting the data.

Bootstrap: it involves replacing the data with random samples, in other words, used to sample a dataset using replacement to estimate statistics on a population.

2.3.3.2 Cross Validation (CV)

Cross-validation is among the most popular procedures for estimating the out-of-sample predictive performance of statistical models fitted on data sets randomly sampled from a population. Generally speaking, cross-validation estimates the out-of-sample prediction accuracy by fitting and assessing a fitted model on separate subsets of data. One of the most common forms of cross-validation is V-fold cross-validation, where data are partitioned into V folds (sets) of identical size; then, each fold is used to assess the error of the model fitted using the other V1 folds. Finally, the average of all V estimates is used to create the cross-validation risk estimate.

Cross-validation is a general method of statistical analysis with various purposes, but mainly for model validation. It traditionally involves multiple rounds with three steps: the bi-partitioning of the data set, analysis of one partition, and validation/testing on the other one. Results from these rounds may be analysed after some aggregation, or individually.

Cross-validation methods can be broadly classified into two categories: exhaustive and non-exhaustive methods. The exhaustive methods strive to test all possible ways to divide the original data sample into a training and a testing set. But the non-exhaustive ones don't compute all ways of partitioning the original data into training and evaluation sets.

The five common types of cross-validation are the holdout method, the k-fold cross-validation method, stratified k-fold cross-validation, leave-p-out cross-validation (LpOCV) and the leave-one-out cross-validation (LOOCV) approach.

2.3.3.3 Grid Search Cross Validation

Grid SearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. As mentioned above, the performance of a model significantly depends on the value of hyperparameters. There is no way to know in advance the best values for hyperparameters so ideally, we need to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus we use GridSearchCV to automate the tuning of hyperparameters.

Grid search cross-validation (GridSearchCV) is used to choose the best model for each ML approach. With the parameters that produced the better cross-validation performance, a new model is automatically fitted using this method to the whole training dataset. This method aids in obtaining a more accurate generalization performance estimate.

GridSearchCV is a function that comes in Scikit-learn's (or SK-learn) model selection package. This function helps to loop through predefined hyperparameters and fit the estimator (model) on the training set. Thus, in the end, we can select the best parameters from the listed hyperparameters.

2.3.3.4 SMOTE-NC

Synthetic Minority Over-sampling Technique (SMOTE) is one of the derivatives of oversampling which was first introduced by Chawla et al in 2002. SMOTE can handle imbalanced data by replicating minority data, the result is known as synthetic data. In numerical data, SMOTE will work by finding k-nearest neighbours for each data in the minority class using Euclidean distance. For categorical and numerical data, the Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC) will be used. According to Li Sun, if the proportion of minority class is less than 35 % of the total data, then it is categorized as an imbalanced dataset.

2.3.4 Performance Evaluation Metrics

Evaluating the performance of a Machine learning model is one of the important steps while building an effective ML model. To evaluate the performance or quality of the model, we use different types of metrics, and these metrics are known as performance metrics or evaluation metrics. These metrics help us to understand how well our model has performed for the given data. So, we can improve the model's performance by tuning the hyperparameters. The choice of performance metrics is very crucial, this choice depends on the type of model and its application.

Accuracy: This is the most intuitive model evaluation metric, it is the measure of correct predictions made by the model. It is equal to the number of correct predictions made upon the total number of predictions made by the model. The higher the accuracy, the more accurate the model. It is good to use the Accuracy metric when the target variable classes in data are approximately balanced, but it is not recommended to use it when the target variable majorly belongs to one class.

Precision: This metric is also known as Positive Predictive Value (PPV). this metric is used to overcome the limitation of Accuracy. The precision determines the proportion of positive prediction that was actually correct, it is the ratio of predictions that are actually true to the total positive predictions.

Recall: This metric shares some commons with the Precision metric; however, it aims to calculate the proportion of actual positive that was identified incorrectly. It can be calculated as a True Positive or a prediction that is actually true to the total number of positives, either correctly predicted as positive or incorrectly predicted as negative (True Positive and False Negative).

F1 Score: This is a metric to evaluate a binary classification model on the basis of predictions that are made for the positive class. It is a type of single score that represents both Precision and

Recall. It should be used if both of them are important for evaluation, but one (precision or recall) is slightly more important to consider than the other. Thus, the F1 Score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them.

AUC: This metric calculates the performance across all the thresholds and provides an aggregate measure. The value of AUC ranges from 0 to 1, which means that a model with 100 % wrong prediction will have an AUC of 0, whereas models with 100 % correct predictions will have an AUC of 1. It is used to measure how well the predictions are ranked rather than their absolute values.

Specificity: it is defined as the proportion of actual negatives, which got predicted as the negative (or true negative). This implies that there will be another proportion of actual negatives, which got predicted as positive and could be termed as false positives. The sum of specificity and false positive rate would always be one. The higher value of specificity means a higher value of true negative and lower false positive rate.

3 Results and Discussion

3.1 Exploratory Data Analysis

In this section, we will give some key findings in three parts, (1), Univariate Analysis where we focused on exploring the target variable, (2) Multivariate Analysis where we analyzed the relationships between the target variable and other features and (3) Correlational Analysis where we explored the correlations among numerical and categorical variables separately. All the subsections will include both visual and statistical analysis.

3.1.1 Univariate Analysis

Loan status is our dependent variable which is a categorical variable that takes only two values: Fully paid or Charged Off. The figure below shows the existence of an imbalanced dataset where borrowers with charged-off status only represent 20 % of the total observations, this problem can affect the capability of machine learning models to predict the true negative value.

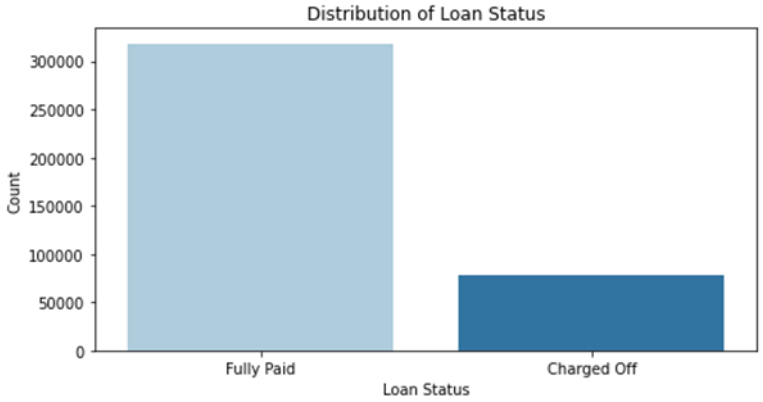

Figure 1: Distribution of loan status

3.1.2 Multivariate Analysis

Loan characteristics feature analysis

Loan Amount

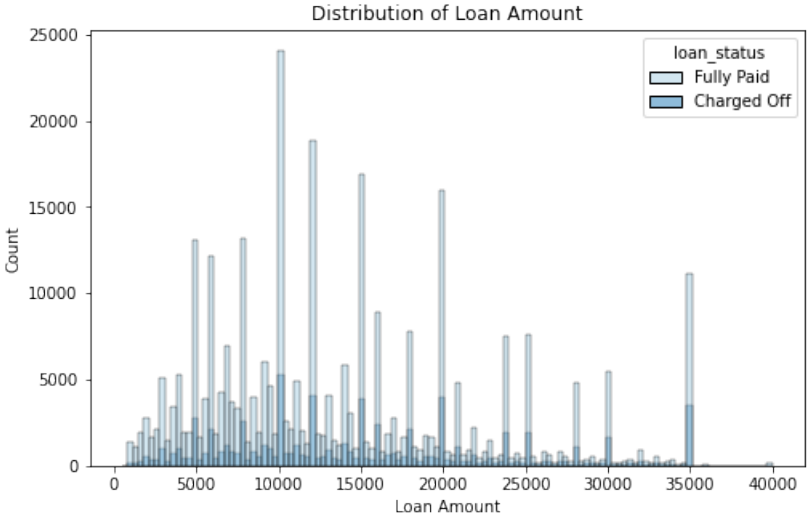

Figure 2: Distribution of Loan Status by Loan Amount

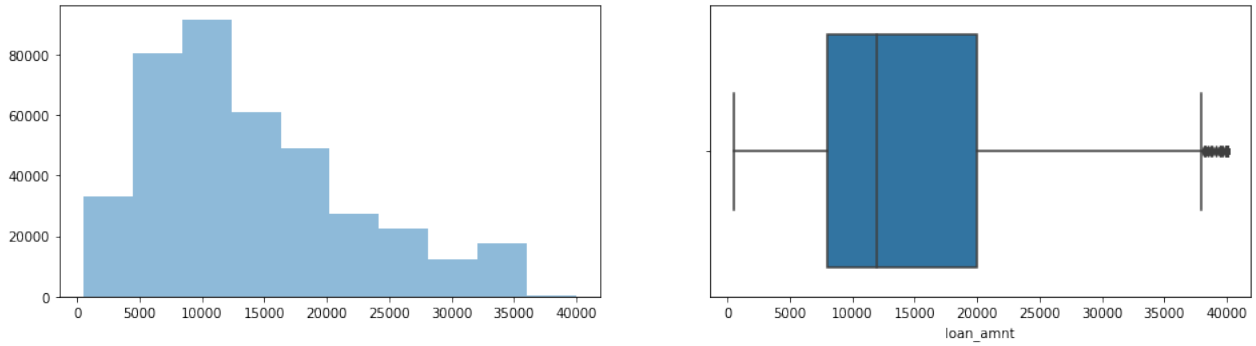

Figure 3: Boxplot for Loan Amount

The loan amount is one of the most important continuous variables. It is denoted as loan_amnt in the data set. Lending Club enables borrowers to create personal loans between 500.00 and 40,000 US dollars, with a mean of 14.114 US dollars. And the standard deviation is 85,357.44 US dollars. From Figure 3, we conclude that loan amount is normally distributed, slightly right skewed where most of the borrowers are seeking loan amounts ranging from 5,000 US dollars to 20,000 US dollars. Our loan amount shows no outliers point.

Interest Rate

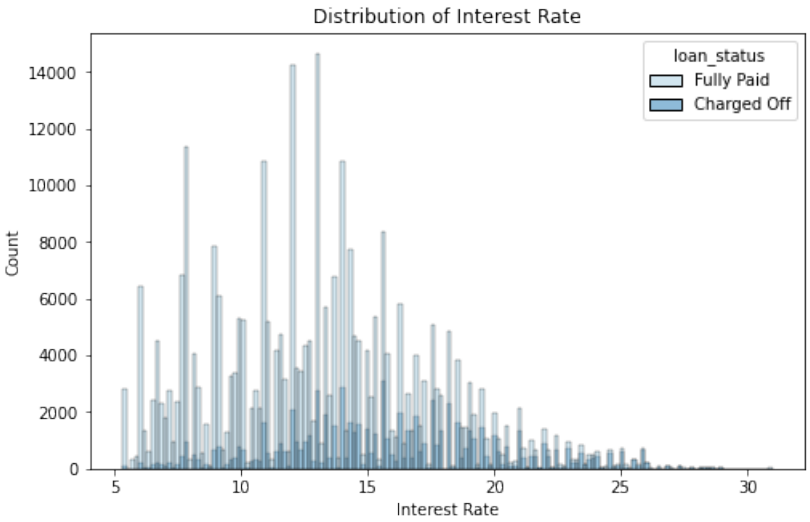

Figure 4: Distribution of interest rate

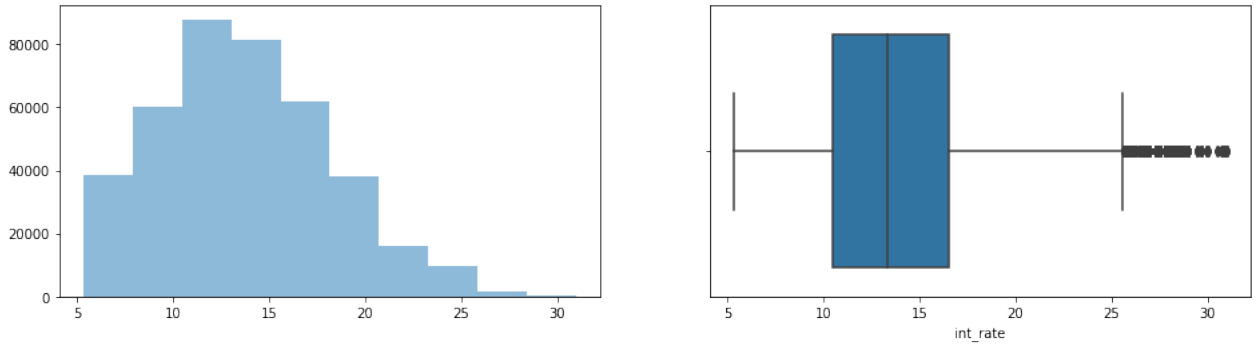

Figure 5: Boxplot of Interest Rate

This variable is denoted as int_rate. Normally, loans become more lucrative when the interest rate is lower. From Figure 5 above, we noticed that the average interest rate is 13.64 % and that the interest rate seems to be normally distributed, and slightly skewed. For most borrowers, the interest rate is ranging from 10 % to 15 %. However, as can be seen in Figure 5, interest rates show a few outliers in the upper bounds, which is probably due to poor credit ratings.

Grade and Subgrades

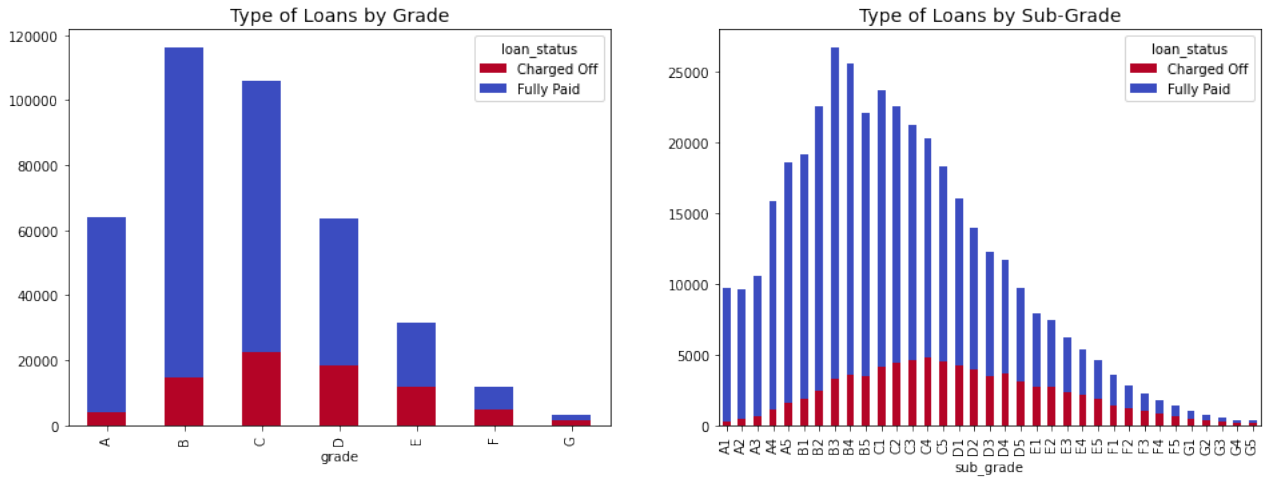

Figure 6: Distribution of Loan Status by Grades and Sub-grades

After determining whether the borrower is creditworthy, the lending Club assigns a credit grade that determines the payable interest rate and fees to their approved loans. These grades are assigned within an alphabetical range from A to G. Each of these letter grades has five finer-grain sub-grades, numbered 1 to 5, with 1 being the highest category within the grade. Loan interest rates are inversely proportionate to the credit grade. 'A' being the highest grade, therefore low-interest rate and vice-versa.

In Figure 6, we can see that the distribution of sub-grade seems to follow the normal distribution but has quite a heavy right tail. There is a trend of a higher fraction of charged-off loans as the grade goes from A to G, with some small variations among sub-grades. For most loans, the grades are ranging from B to C and are more likely to be fully paid.

Loans by purpose

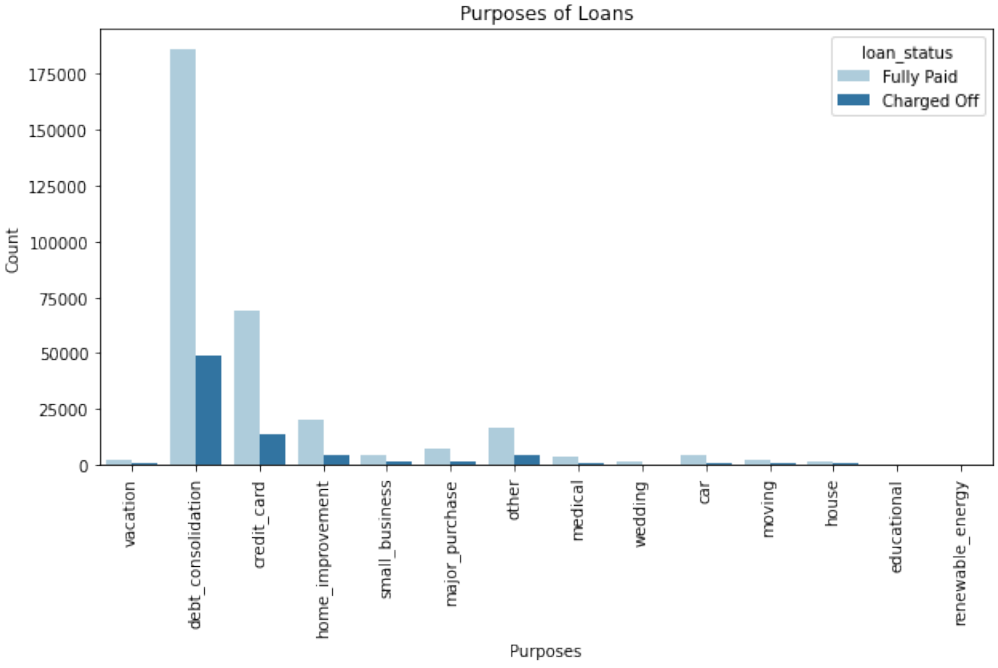

Figure 7: Distribution of Loan Status by Purposes

In Figure 7, we can see the purposes why the borrowers apply for loans from the Lending club. We noticed that most borrowers decided to take loans for debt consolidation, credit card and home

improvement purposes.

Debt to income ratios (DTI)

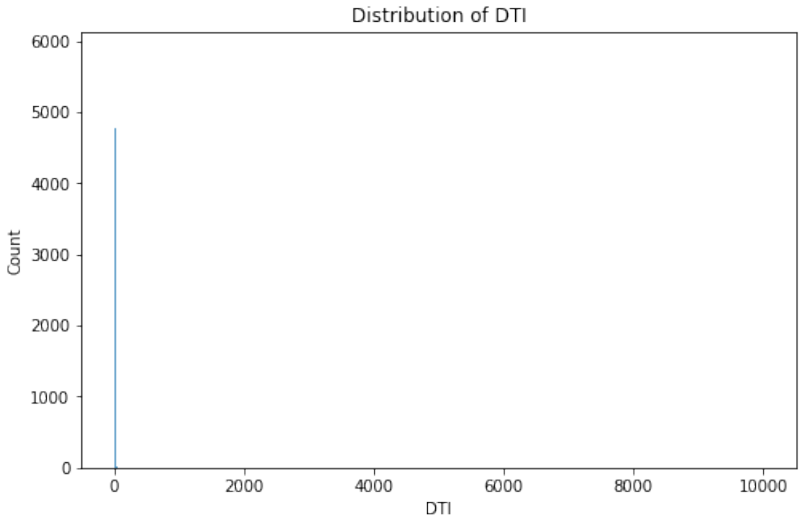

Figure 8: Distribution of Loan Status by DTI

Figure 8 shows that the debt-to-income ratios feature donated as dti in our dataset has an outlier issue. After looking at the box plot and using the Inter Quartile Range to detect the outliers, we noticed that the outliers lie in the upper range. Hence, we will plot the histogram of dti for the upper limit to get more reasonable bounds for the data.

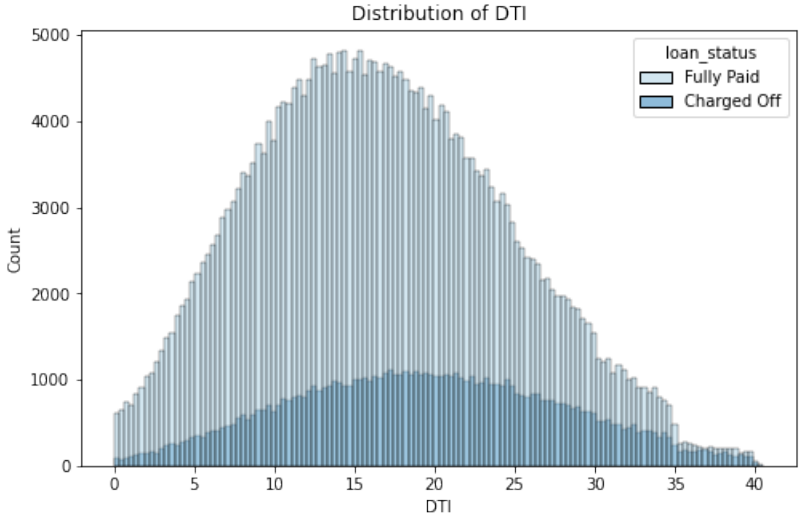

Figure 9: Distribution of Loan Status by DTI constraint by the upper limit

Figure 9 above shows that the majority of the density is captured between 0 and 40, which seems to be reasonable since the lower bound is zero and it should be zero since you cannot have less than no debt.

Term

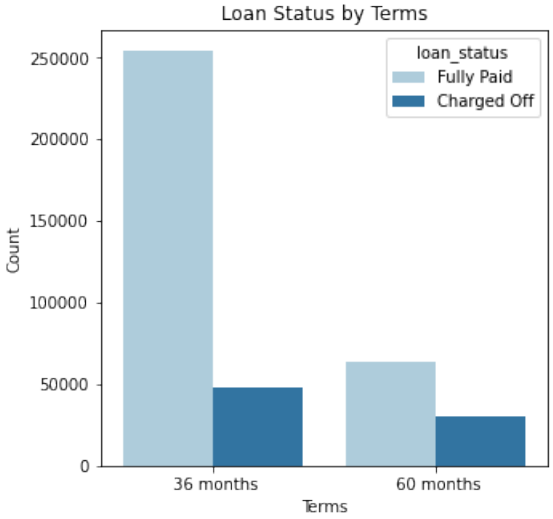

Figure 10: Distribution of Loans Status by Term

The term variable represents the duration of the loan. This categorical value takes only two values: 36 months and 60 months. A loan with 60 months tends to have a lower fraction of being fully paid.

Applicants' demographic features analysis

Employment title and Employment length

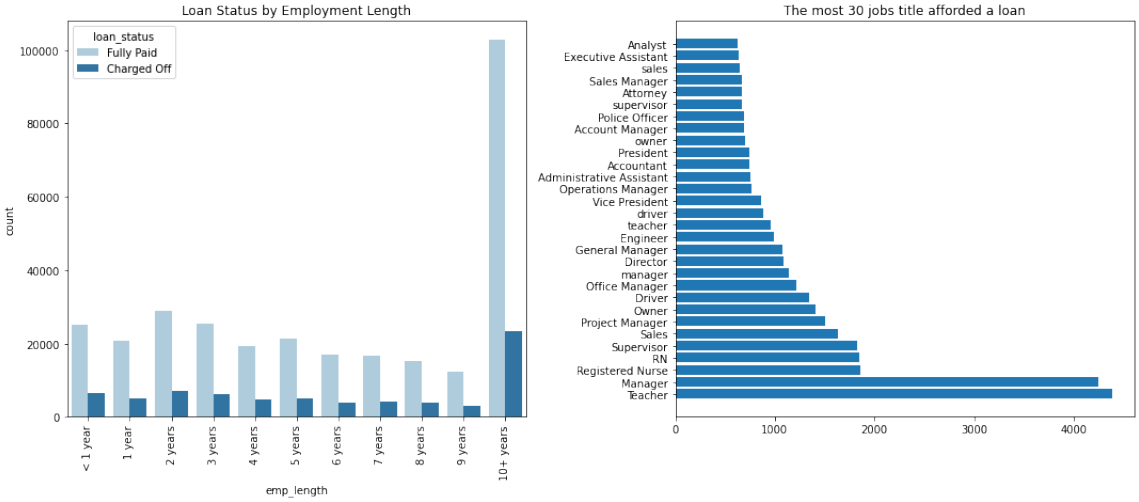

Figure 11: Distribution of Loans Status by Employment Length and Top 30 jobs afford a loan

Figure 11 above shows that most individuals have more than 10 years of job experience. We can see that the job title distribution has a strong right tail and that Teachers, Managers, and Registered Nurses take loans more often compared to other jobs.

Home ownership

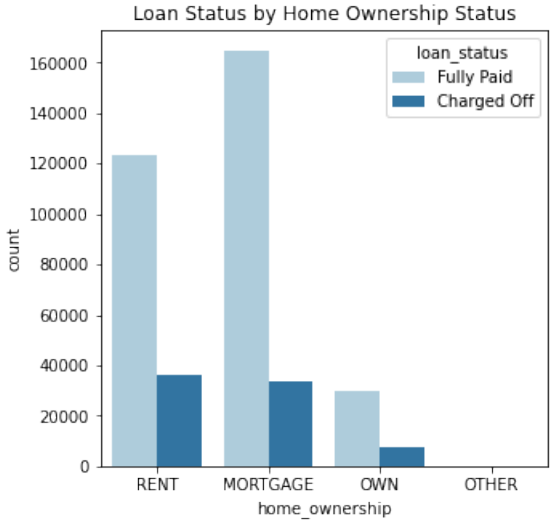

Figure 12: Distribution of Loan Status by Home Ownership

Home ownership represents the residential situation of the borrowers. This categorical variable takes four values: Rent, Own, Mortgage, and Other. From Figure 12, it appears that there is a small difference in charge-off rates by home ownership status, and most borrowers have mortgage homes. The mortgage has less probability of being charged off.

Annual income

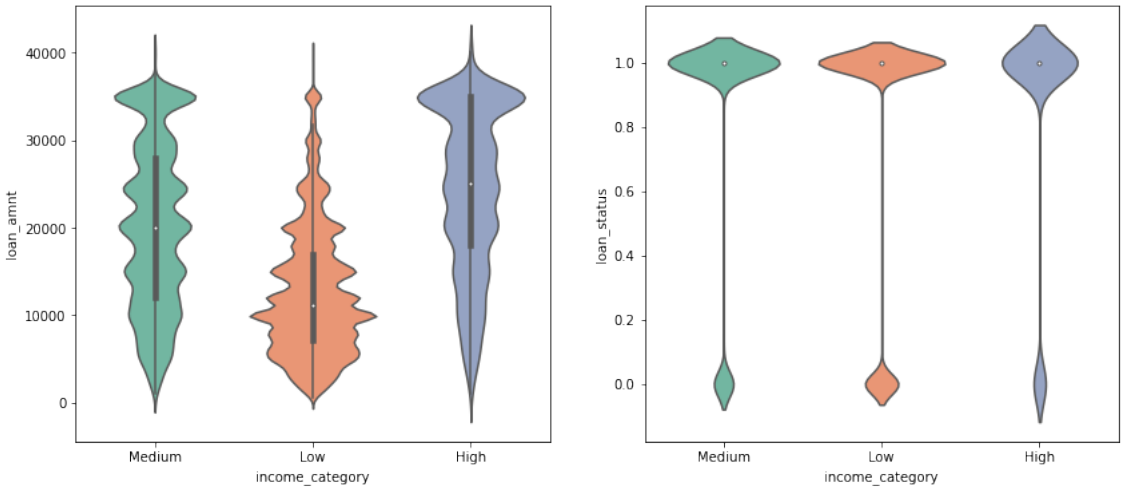

Figure 13: Loan Amount by Annual Income Category

This continuous variable is denoted as annual_inc. It is the annual income provided by the borrower during registration. The annual income is on average 74,203 US dollars with a maximum amount of 870,6582 US dollars. To show the distribution of annual income, we used the violin plot. A violin plot is used to visualize the distribution of the data and its probability density. We noticed from the figure above that borrowers who made part of the high-income category took higher loan amounts than those from low- and medium-income categories. Loans borrowed by the Low-income category had a slightly higher chance of becoming a charge-off loan.

Verification status

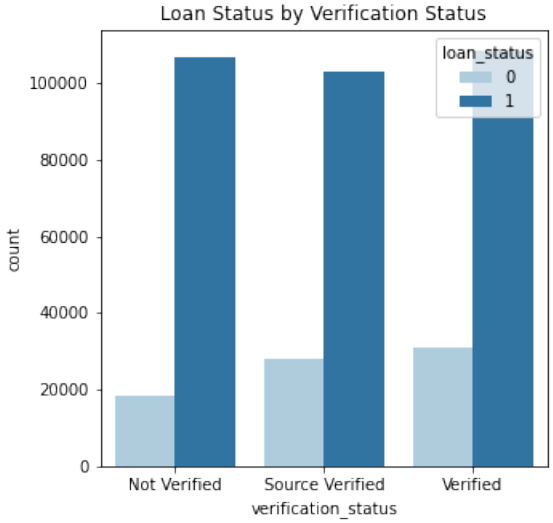

Figure 14: Distribution of Loan Status by Verification Status

This variable indicates if income was either verified or not by Lending Club, or if the income source was verified. From Figure 14, we see that, surprisingly, verified loans have a higher chance of being charged-off.

3.1.3 Correlational Analysis

In order to check the existence or multicollinearity problem within the features, we assessed the correlation between variables and made use of the Variance Inflation Factor.

.

Correlation between numerical features

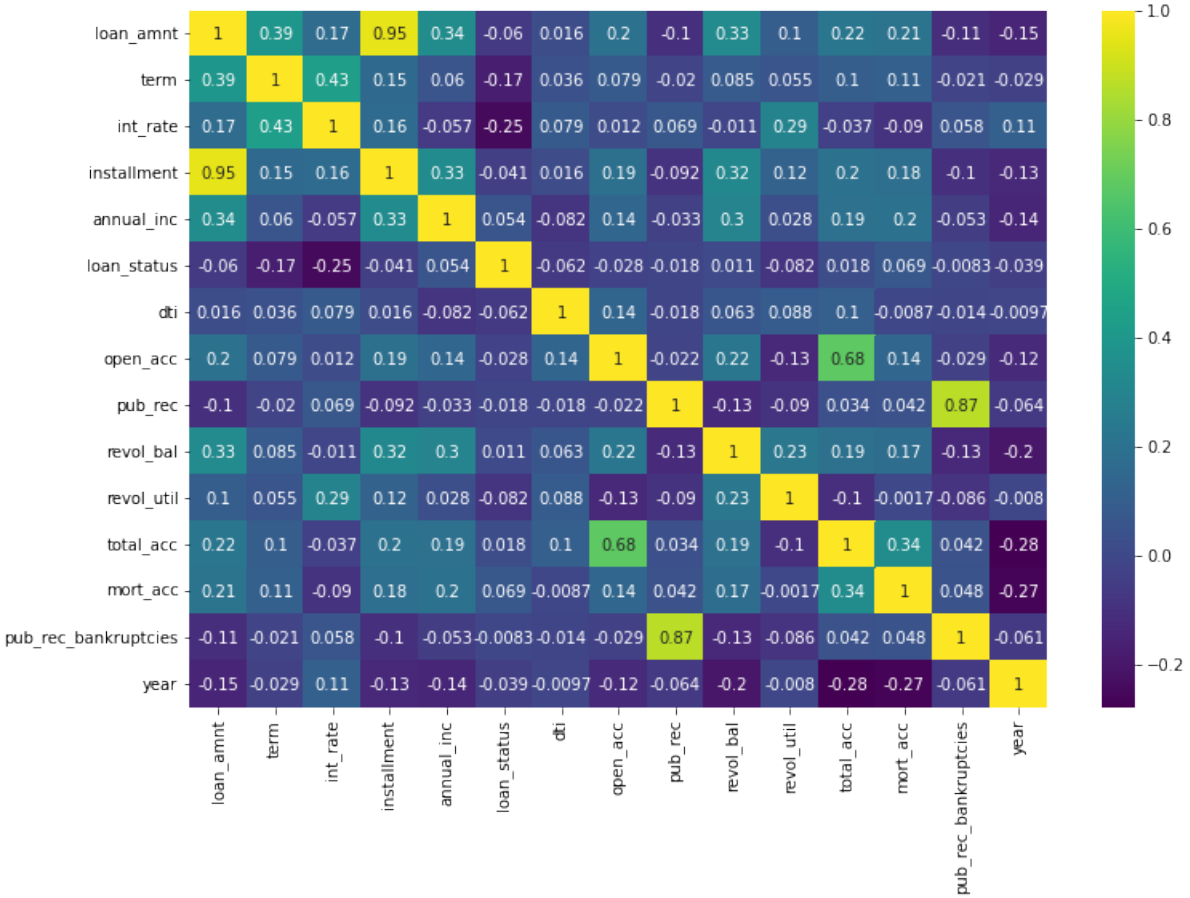

Figure 15: Pearson correlation for numerical features

Figure 15 above represents a heat map of the Pearson correlation coefficients among our numerical features. The heat map shows that there is a low correlation between most of our features. However, we notice that there is a strong correlation between instalment (the monthly payment owed by the borrower) and loan amount at 95 %, pub_rec and pub_rec_bankruptcies at 87 %, and between total_acc and open_acc at 68 %.

Next, we compute the Variance Inflation Factor to better detect the multicollinearity problem as follows.

The Variance Inflation Factor analyses the correlation between one variable Xi with the rest of the variables. A VIF value above 4 indicates the possible existence of multicollinearity and a value greater the 10 indicates the significant presence of multicollinearity that needs to be corrected.

Feature name	Variance Inflation Factor
loan_amnt	226.958690
term	116.444380
int_rat	22.867651
installment	203.377241
annual_inc	3.033681
loan_status	5.514942
dti	2.025930
open_acc	11.760811
pub_rec	4.778457
revol_bal	2.095968
revol_util	7.291668
total_acc	11.901645
mort_acc	3.159499
pub_rec_bankruptcies	4.613555
year	76.540778

Table 2: Variance Inflation Factor by Feature

From Table 2, we notice that the VIF is very high for loan_amnt, term, instalment and year; fairly high for int_rate, open_acc and total_acc. Hence, considering these features together leads to a model with high multicollinearity, so to explosive variance when performing parametric models. This obstacle was addressed by applying dimensionality reduction and regularization techniques.

.

Correlation between categorical features

The Chi-Square test is a statistical test that is used to find out the difference between the observed and the expected data we can also use this test to find the correlation between categorical variables in our data.

While conducting the chi-square test, we have to initially consider two hypotheses, the Null Hypothesis, and the Alternative Hypothesis.

- H0 (Null Hypothesis): The two variables being compared are independent.
- H1 (Alternative Hypothesis): The two variables are dependent.

Now, if the p-value obtained after conducting the test is less than 0.05, we reject the Null hypothesis and accept the Alternative hypothesis and if the p-value is greater than 0.05 we accept the Null hypothesis and reject the Alternative hypothesis.

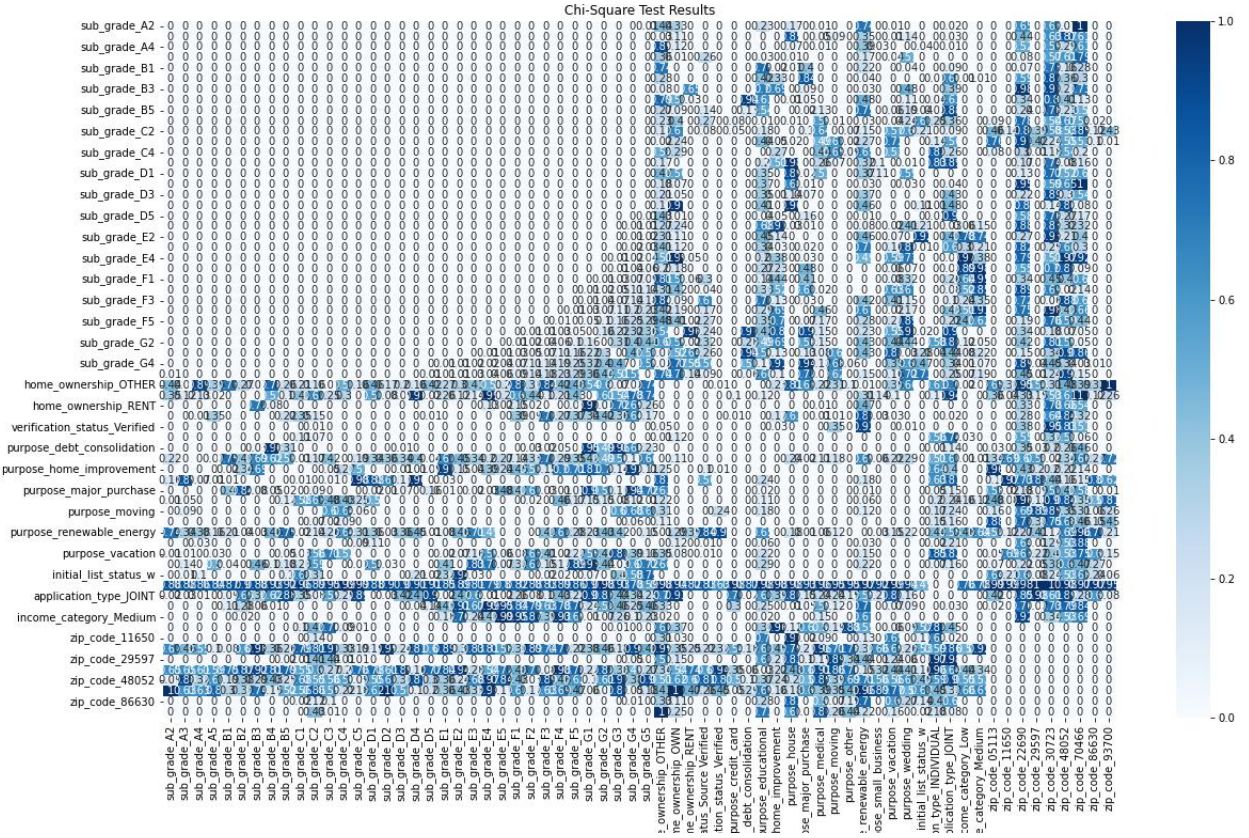

Figure 16: Correlation for categorical features

A heat map of the p-values for the Chi-Square test is represented in the figure above, which shows p-values lower than 0.05. Therefore, we conclude a high correlation between most of our categorical variables. We noticed a high correlation between verification status verified and home ownership "Own" and purpose "Major purchase" and zip code 93700.

3.2 Information Reduction models

In order to build a model that predicts whether a loan will be fully paid or charged off, we first divided the data into 03 subsets, namely training, test and validation sets. The purpose of the validation set is to tune the hyperparameters without the concern of leaking the information from the training set. As mentioned in subsection. . . , the SMOTE-NC technique was applied to the training set only with the aim to solve the problem of a mildly imbalanced dataset. Hence, we can see the effectiveness of this technique in our case when comparing the results before and after SMOTE-NC. In addition, we also tested the impact of adding polynomial features degree 02 on the performance of parametric models (i.e., ElasticNet and Logistic Classification; we could not test this with Autometrics since the database with polynomial features is too large for OxMetrics to render), to check whether there is nonlinearity existing in the models.

This section will discuss separately two aspects of our modelling process, first the Information Reduction techniques with ElasticNet and Autometrics and second, Prediction models with Logistic Classification and Random Forest. Each part will include two sections, before and after SMOTE-NC to highlight the impact of this oversampling technique. The After SMOTE-NC subsection will discuss further each model's performance metrics, normally the default model and the version with tuned hyperparameters. And in the case of parametric models, there is a part where we compared the result

of models with and without polynomial features.

3.2.1 Before SMOTE-NC

Table 3 shows the evaluation metrics for 02 information reduction methods using the original training dataset. We can notice that these models perform pretty well in classifying the positive classes when the Recall, Precision and F1 scores are relatively high (approximately 0.98, 0.88 and 0.93 respectively). The AUC score remains decent at 0.73, which means that there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. This is actually what we would like to witness in our classifier because if the applicant is likely to pay off the loan, then not approving it will result in a profit loss for the Lending Club.

However, as can be seen from the table, the Specificity score of these models is just slightly higher than 0.5, which implies that the ability to detect actual negatives is quite poor. If the Lending Club issues loans to an applicant who is not likely to repay the loan, this also leads to a financial loss for the company. As mentioned in the Introduction, in the scope of our study, we would like to also pay the attention to detecting the default loan. That's the main reason why we adopted the SMOTE-NC technique with the aim to solve the mildly imbalanced dataset with only 20% of default loans.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
Default ElasticNet Classification before SMOTE-NC	0.887771	0.989005	0.934104	0.884978	0.572985	0.730378
Tuned ElasticNet Classification before SMOTE-NC	0.887759	0.989036	0.934099	0.884943	0.572836	0.730297
Autometrics before SMOTE-NC	0.887506	0.988706	0.93394	0.884922	0.572823	0.730164

Table 3: Evaluation metrics for Information Reduction models before SMOTE-NC

3.2.2 After SMOTE-NC

3.2.2.1 ElasticNet

Default and Tuned Models

In this section, we tested the default Logistic Regression classifier and the classifier with the tuned hyperparameter l1_ratio=0.1 using ElasticNetCV. The result indicates that these two models perform quite similarly, but noticeably the Specificity scores for these models are better with a rate of 0.7, compared to 0.57 with the before SMOTE models. The ROC score also slightly improves to 0.74 versus the previous 0.73 rates. However, as a trade-off, the Recall and F1 scores decrease quite significantly by 8.6% and 18.4% compared to the models using the original training set.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
Tuned ElasticNet Classification before SMOTE-NC	0.887759	0.989036	0.934099	0.884943	0.572836	0.730297
Default ElasticNet Classification after SMOTE-NC	0.783171	0.808428	0.85709	0.911987	0.704636	0.743904
Tuned ElasticNet Classification after SMOTE-NC	0.783133	0.808381	0.857062	0.911982	0.704627	0.74388

Table 4: Evaluation metrics for ElasticNet models before and after SMOTE-NC

The following graph shows the feature importance of the tuned ElasticNet model. We can clearly notice that the zip_code feature plays a vital role in our classifier's predictive power. There are 03 neighbourhoods on the upper left side of the graph means that the applicants living there are less likely to pay off their loans, while the other 02 neighbourhoods on the down right side imply the opposite. Secondly, if the borrowers apply for individual loans, there is a high chance that they will not be able to pay off the loans while the joint type says the opposite. Thirdly, the higher the sub-grade, the lower the likelihood that the applicants will fully pay the loans. This aligns with the way we feature-engineered this variable, with types A graded as 1 and then increased with types B to G. We can see a similar thing happening to the loan amount variable. This makes sense in reality when the higher the loan amount, the more difficult to pay it off. Additionally, the DTI feature performs as expected since this ratio is calculated using the borrower's total monthly debt payments on the total

debt obligations, excluding mortgage and the requested Lending Club loan, divided by the borrower's self-reported monthly income. In other words, if the applicants' monthly debts are higher than their monthly income, it makes sense when they are less likely to repay the loan issued by the Lending Club.

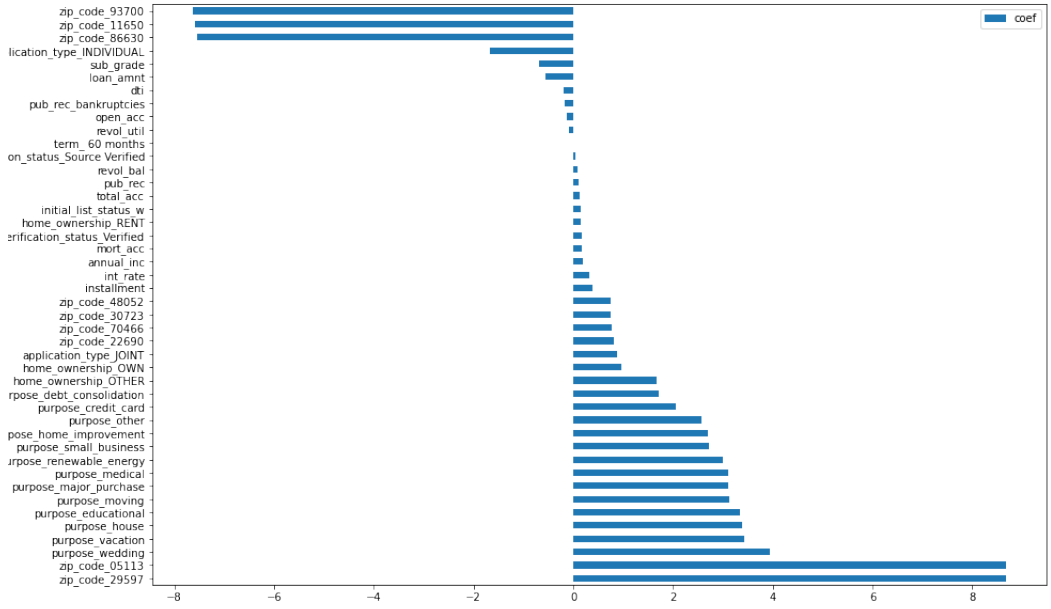

Figure 17: Feature Importance of the tuned ElasticNet after SMOTE

Polynomial Features Model

In order to assess the impact of adding nonlinear components on the models' performances, we first estimated the ElasticNet model with polynomial features and interactions degree 02. This increases the number of variables to 1035. As a result, the AUC score increases very slightly by 0.003 while the Specificity score remains quite the same. This may imply that there is nonlinearity existing in our model, but since the difference between the two models is very small, we need to look at the result of the Logistic Regression with the polynomial features to make a more confident conclusion.

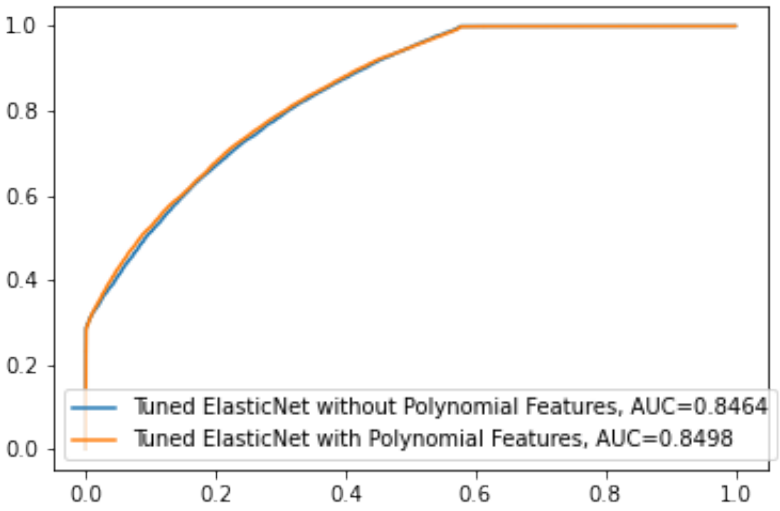

Figure 18: ROC curves of ElasticNet models with and without Polynomial Features

3.2.2.2 Autometrics

Since there is no equivalent package available in Python to run Autometrics, we decided to first run Autometrics on OxMetrics, then used the selected variables to estimate the Logistic Classification models' performances for both before and after SMOTE datasets. The regressor significance level is chosen as α=0.1.

The terminal model using the original training set contains 28 variables (not including the constant) while that of SMOTE set contains almost all the variables, 42 out of a total of 44 features. Details about which variables are selected can be found in Table 9 in the Annex section. As can be seen from Graph 19 and Table 5, the evaluation metrics for the Logistic Regression model using the variables selected by Autometrics are quite similar to those of ElasticNet. This may imply that regarding the power of information reduction, there is no superior technique in the case of our study.

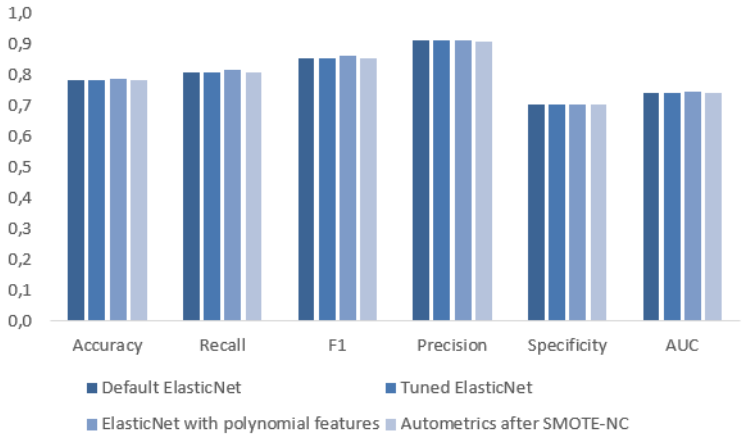

Figure 19: Evaluation Metrics for all Information Reduction models

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
Default ElasticNet Classification after SMOTE-NC	0.783171	0.808428	0.85709	0.911987	0.704636	0.743904
Tuned ElasticNet Classification after SMOTE-NC	0.783133	0.808381	0.857062	0.911982	0.704627	0.74388
ElasticNet Classification after SMOTE-NC with polynomial features	0.789674	0.81741	0.862098	0.911956	0.70343	0.746552
Autometrics after SMOTE-NC	0.782944	0.808664	0.856998	0.911476	0.702967	0.742955

Table 5: Evaluation metrics for Information Reduction models after SMOTE-NC

3.3 Prediction models

3.3.1 Before SMOTE-NC

Table 6 below shows the overall performance of 02 Prediction models, Logistic Regression and Random Forest. We can see that the two models achieve pretty high Recall, Precision and F1 scores (approximately 0.99, 0.88 and 0.93), which is quite similar to the performance of the 02 Information Reduction techniques' models. The AUC and Specificity scores witness the same thing also, with a slightly better performance of Logistic Regression compared to the Random Forest classifier regarding these evaluation metrics.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
Logistic Classification before SMOTE-NC	0.887695	0.989099	0.934068	0.884838	0.57238	0.730038
Random Forest before SMOTE-NC	0.888164	0.993488	0.934596	0.882295	0.560657	0.72441

Table 6: Evaluation metrics for Prediction models before SMOTE-NC

3.3.2 After SMOTE-NC

3.3.2.1 Logistic Regression

Default and Tuned Models

In this section, we tested the default Logistic Regression classifier and the classifier with the tuned hyperparameter c_values=0.1 using GridSearchCV on the validation set. The result indicates that the tuned model performs mildly better than the default one with all the evaluation metrics higher than those of the default model. We also noticed that the Logistic Regression model after SMOTE can detect far more true negatives (i.e., actual charged-off loans) compared to the before SMOTE model. This is what we expected when we synthetically increased the number of charged-off loans by using SMOTE-NC.

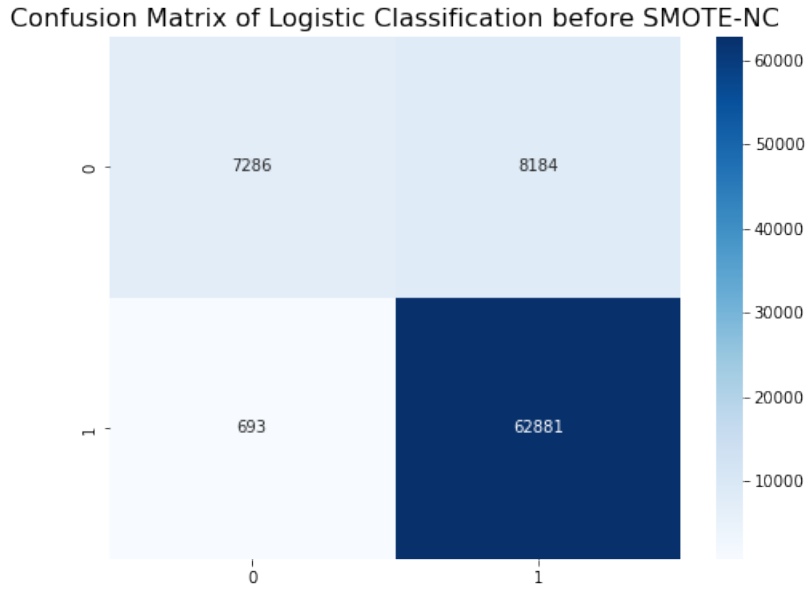

Figure 20: Confusion matrix of Logistic Regression before SMOTE

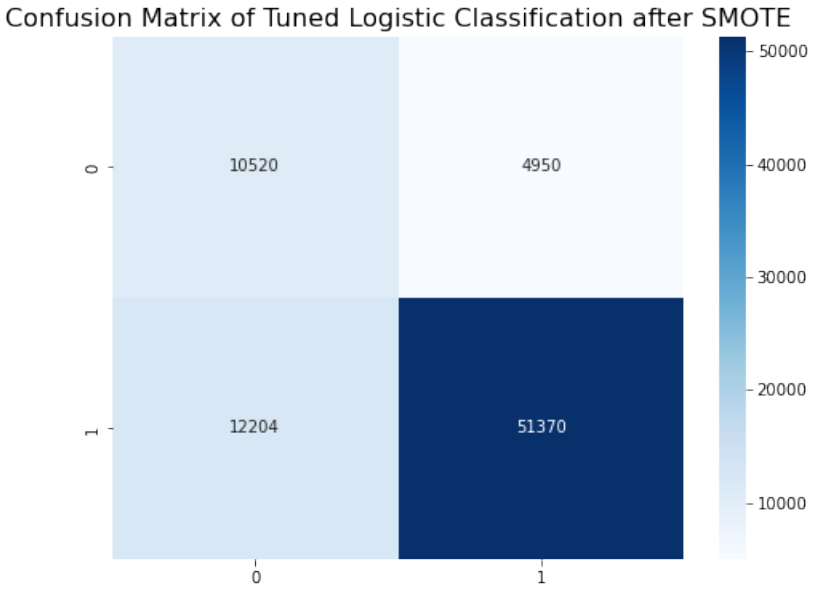

Figure 21: Confusion matrix of tuned Logistic Regression after SMOTE

The down below graph indicates the feature importance of the Logistic Classification after SMOTE. We can see that this graph provides quite similar information to that of ElasticNet we interpreted above. One significant difference is that the coefficient magnitudes of Logistic Regression are higher than those of ElasticNet. This is due to the multicollinearity issue we mentioned in the Correlational Analysis section. This also implies that ElasticNet performs well in cases with highly-correlated data like ours.

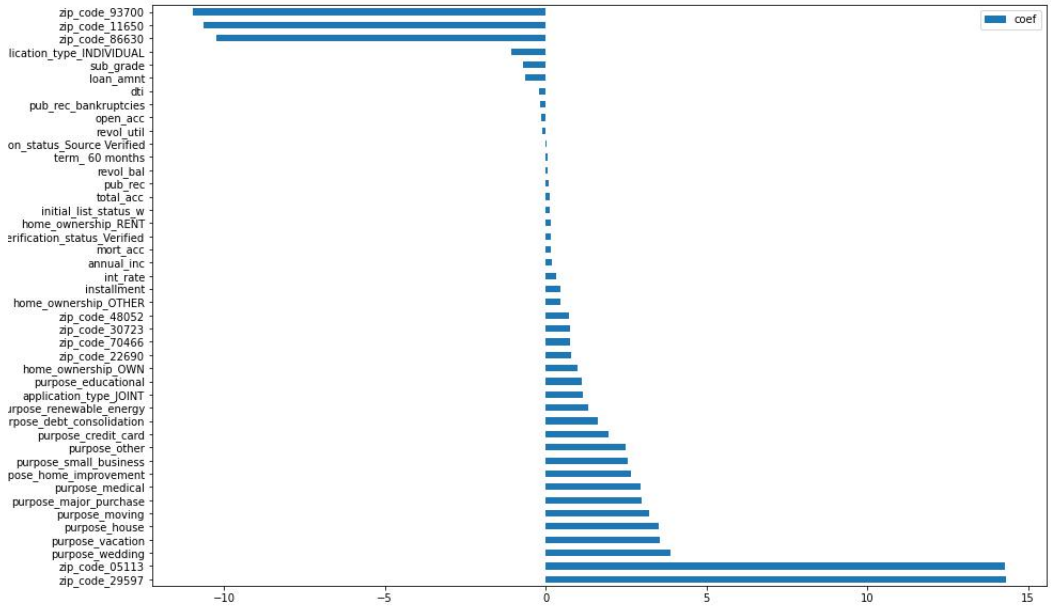

Figure 22: Feature Importance of Logistic Regression model after SMOTE

Polynomial Features Model

After adding polynomial features and interactions degree 02, we can see that the Recall and F1 scores are improved while the Specificity and AUC scores decrease. Along with the mild results from ElasticNet with polynomial features analyzed in the section above, this result suggests that nonlinearity may not exist in our true Data Generation Process.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
Tuned Logistic Classification after SMOTE-NC	0.782982	0.808035	0.856924	0.912109	0.705079	0.74403
Logistic Classification after SMOTE-NC with polynomial features	0.785309	0.812675	0.858936	0.910782	0.700216	0.742763

Table 7: Evaluation metrics for Logistic Regression after SMOTE with and without polynomial features

3.3.2.2 Random Forest

This section estimated the default Random Forest on the SMOTE dataset. The evaluation metrics Recall and F1 show improvement compared to the other models using the new artificial observations. The Recall and F1 scores for Random Forest are 0.92 and 0.91, respectively while they are approximately 0.81 and 0.85 for the rest of the after-SMOTE models.

However, despite the creation of synthetic observations in the training set, the Specificity score of Random Forest is lower than those of other models, only 0.64 compared to 0.70.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
Logistic Classification after SMOTE-NC	0.782918	0.807988	0.856881	0.912072	0.704966	0.743942
Tuned Logistic Classification after SMOTE-NC	0.782982	0.808035	0.856924	0.912109	0.705079	0.74403
Logistic Classification after SMOTE-NC with po...	0.785309	0.812675	0.858936	0.910782	0.700216	0.742763
Random Forest after SMOTE-NC with polynomial f...	0.856181	0.924985	0.911861	0.899104	0.642236	0.749209

Table 8: Evaluation metrics for all Prediction models after SMOTE

With regards to the interpretability of this model, we can notice quite a big difference in the choice

of the most important features, compared to the other models estimated on the SMOTE dataset. Sub-grade and interest rate are the two most vital variables, which totally align with what can be expected in reality. Other variables which are not considered as important in other models appear in this plot of Random Forest, for example, annual income, open account (i.e., the number of open credit lines in the borrower's credit file) and other loan characteristics features.

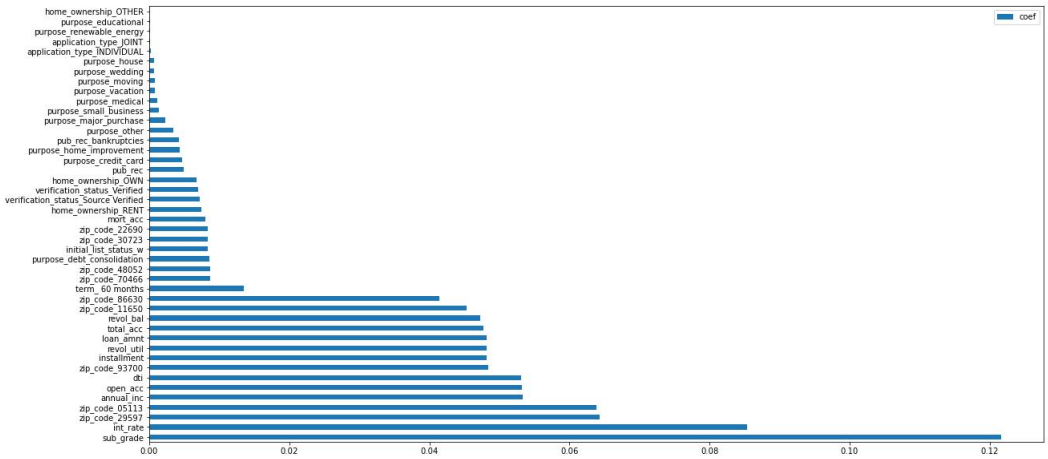

Figure 23: Feature Importance of Random Forest model after SMOTE

4 Conclusion

In this study, we aimed at predicting loan status by exploring and analysing different features related to loan characteristics and borrowers' profiles. We detected Multicollinearity problems in our dataset which we tried to resolve by using two Information Reduction techniques, Elastic Net and Autometrics. The prediction results of these two methods are quite similar, which suggests that in our case, there is no superior technique in terms of information reduction power.

Moreover, for the sake of Risk Management, we would like to pay more attention to the default loans although they do not account for a high proportion of observations in our dataset. This is also one of our main challenges which we tried to tackle by using SMOTE-NC technique. The Specificity scores for both Logistic Regression and Random Forest are quite low for the imbalanced training set, approximately at 0.57 despite pretty decent positive predictive power. The same thing can be witnessed when we compared evaluation metrics for the predictive capability of the two aforementioned Information Reduction models. Hence, by switching to the SMOTE-NC dataset, the Specificity scores for all prediction models increased to 0.7, except for the Random Forest model. However, as a trade-off, the positive prediction metrics for models using SMOTE-NC dataset have dropped quite significantly.

In addition, we added polynomial features degree 02 to test its impact and the results show that these non-linear components do not help improve the models' performances. This aligns with the fact that our parametric models outperform the non-parametric model since the relationships between the target variable and features do not contain non-linearity.

Regarding the interpretability of our models, we found several insights that are worth highlighting about the important features. The zip code feature extracted from the original address variables seems to be the strongest indicator for both default and fully-paid loans, in all our parametric models. Other variables, despite having smaller signalling power, behave according to what we expected, for example, the sub-grade and DTI features. However, the Random Forest's feature importance analysis suggests

a different story. By using GiniIndex, Random Forest considers sub-grade and interest rate are the two most important features, then the zip code features as we see in other parametric models.

The credit risk prediction modelling covered in this study can be used by P2P lenders to make better decisions when assessing loan applications. Preventing financial loss, in fact, requires early detection of defaulting debtors. Therefore, more precise calculations of the likelihood of default may be helpful for formulating risk-reduction plans like raising interest rates. Additionally, the accuracy of prediction models may be impacted by future changes given that the underlying distribution of incoming data released regularly by the Lending Club may vary unexpectedly over time. Hence, it would be interesting to test our existing models' performances and adapt them if necessary. Furthermore, we suggest future research on this topic focus on the profit/ cost analysis of models' performance. Our study results show that while some models are better at recognizing positive situations, others are more capable of correctly classifying negative ones. When misclassification errors occur, there are associated costs. These costs may be direct, as in the case of a loan given to a borrower who would default, or indirect, as in the case of a P2P platform that would lose out on tax and interest revenue.

5 References

Adnan, Md Nasim. (2022). On Reducing the Bias of Random Forest. 10.1007/978-3-031-22137-8-14.

Ala'raj, M., and Abbod, M.F. (2016). Classifiers consensus system approach for credit scoring. Knowledge-Based Systems, 2016, 104, pp. 89-105.

Alhakeem, Z.M.; Jebur, Y.M.; Henedy, S.N.; Imran, H.; Bernardo, L.F.A.; Hussein, H.M. (2022). Prediction of Ecofriendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques. Materials, 15, 7432.

Amaratunga, Dhammika Cabrera, Javier Lee, Yung-Seop. (2008). Enriched random forests. Bioinformatics (Oxford, England). 24. 2010-4. 10.1093/bioinformatics/btn356.

Araveeporn, Autcha. (2021). The Higher-Order of Adaptive Lasso and Elastic Net Methods for Classification on High Dimensional Data. Mathematics. 9. 1091. 10.3390/math9101091.

Brownlee, J. (2019) A gentle introduction to model selection for Machine Learning, Machine-LearningMastery.com. Available at: https://machinelearningmastery.com/a-gentle-introduction-to-model-selection-for-machine-learning/ (Accessed: January 22, 2023).

Buhayar, Noah. (2015). Where Peer-to-Peer Loans Are Born. Bloomberg, April 16, 2015. (Accessed: January 22, 2023).

Byanjankar, A., Heikkilä, M., and Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach. IEEE, 2015, edn., pp. 719-725.

C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang and T. Wang. (2018). An interpretable model with globally consistent explanations for credit risk. arXiv:1811.12615 [online]. Available at: http:// arxiv.org/abs/1811.12615.

Cao, Yiyang Chen, Haoyu Lin, Bochun. (2022). Wine Type Classification Using Random Forest Model. Highlights in Science, Engineering and Technology. 4. 400-408. 10.54097/hset.v4i.1032.

Chang, Shunpo, Simon D. Kim, and Genki Kondo. (2015). Predicting default risk of lending club loans. Machine Learning (2015): 1-5.

Chaudhary, M. (2022) Random Forest algorithm - how it works and why it is so effective, Random Forest Algorithm - How It Works and Why It Is So Effective. Turing Enterprises Inc. Available at: https://www.turing.com/kb/random-forest-algorithm (Accessed: January 22, 2023).

D. Andriosopoulos, M. Doumpos, P. M. Pardalos and C. Zopounidis. (2019). Computational approaches and data analytics in financial services: A literature review. J. Oper. Res. Soc., vol. 70, no. 10, pp. 1581-1599, Oct. 2019.

D. F. Ahelegbey, P. Giudici and B. Hadji-Misheva. (2019). Latent factor models for credit scoring in P2P systems. Phys. A Stat. Mech. Appl., vol. 522, pp. 112-121, May 2019.

Doornik, Jurgen A., 'Autometrics', in Jennifer Castle, and Neil Shephard (eds). (2009). The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry. Oxford Academic. https://doi.org/10.1093/acprof:oso/9780199237197.003.0004.

Elastic Net (2022) Corporate Finance Institute. Available at: https://corporatefinanceinstitute .com/resources/data-science/elastic-net/ (Accessed: January 22, 2023).

Emekter, R., Tu, Y., Jirasakuldech, B., and Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Applied Economics, 2015, 47, (1), pp. 54-70.

Gad, A.F. (2021) Accuracy, precision, and recall in deep learning, Paperspace Blog. Paperspace Blog. Available at: https://blog.paperspace.com/deep-learning-metrics-precision-recall-accuracy/ (Accessed: January 22, 2023).

Hafezi, M. H., Liu, L., Millward, H. (2018). Learning Daily Activity Sequences of Population Groups using Random Forest Theory. Transportation Research Record, 2672(47), 194–207. https://doi.org/10.1177/0361198118773197.

Hyperparameters and tuning strategies for Random Forest - Probst - 2019 ... (no date). Available at: https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1301 (Accessed: January 22, 2023).

J. A. Doornik, Econometric Model Selection with More Variables than Observations. (2009). Economics Department, University of Oxford.

J. Yao, J. Chen, J. Wei, Y. Chen and S. Yang. (2019). The relationship between soft information in loan titles and online peer-to-peer lending: Evidence from RenRenDai platform. Electron. Commerce Res., vol. 19, no. 1, pp. 111-129, Mar. 2019.

Joby, A. (no date) What is cross-validation? Comparing machine learning models, Learn Hub. Available at: https://learn.g2.com/cross-validation (Accessed: January 22, 2023).

Justin, L. (2020) 8 popular evaluation metrics for Machine Learning Models, Just into Data. Available at: https://www.justintodata.com/machine-learning-model-evaluation-metrics/ (Accessed: January 22, 2023).

Khan, Faridoon Urooj, Amena Ullah, Kalim Alnssyan, Badr Almaspoor, Zahra. (2021). A Comparison of Autometrics and Penalization Techniques under Various Error Distributions: Evidence from Monte Carlo Simulation. Complexity. 2021. 1-8. 10.1155/2021/9223763. Kissel, Nicholas Lei, Jing. (2022). On High-Dimensional Gaussian Comparisons For Cross-Validation.

Lala, A. (2021) Performance evaluation metrics, Medium. Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/performance-evaluation-metrics-5c9104b2a407 (Accessed: January 22, 2023).

Malekipirbazari, M., and Aksakalli, V. (2015). Risk assessment in social lending via random forests. Expert Systems with Applications. 42, (10), pp. 4621-4631.

Mateescu, Alexandra. (2015). Peer-to-Peer lending. Data Society Research Institute 2 (2015): 19-25.

Model selection techniques: How to select a suitable machine learning model? (no date) Censius. Available at: https://censius.ai/blogs/machine-learning-model-selection-techniques (Accessed: January 22, 2023).

Moïse, Musubao Kapiri, Musubao Eloge, Kambale Florence, Kahambu Gilbert, Paluku Jean, Saambili Musivirwa, Paluku Paul, Jean Kahambu, Mbafumoja Florence, Nzenda Gilbert, Saambili Jean, Paluku Musivirwa, Jean Paul,. (2022). Perceptions des agriculteurs sur la culture de chia (Salvia hispanica L.) en ville de Butembo: Essai d'application du modèle de régression logistique [Perceptions of farmers on the cultivation of chia (Salvia hispanica L.) in the city of Butembo: Trial application of the logistic regression model].

International Journal of Innovation and Applied Studies. 36. 468-492. Montesinos-López, Osval Montesinos, Abelardo Crossa, Jose. (2022). Random Forest for Genomic Prediction. 10.1007/978-3-030-89010-0-15.

(no date) Decoding The Chi-Square Test. Available at: https://www.analyticsvidhya.com/blog/2021/06/decoding-the-chi-square-test (Accessed: January 22, 2023).

Panda, Nihar. (2022). A Review on Logistic Regression in Medical Research. National Journal

of Community Medicine. 13. 265-270. 10.55489/njcm.134202222.

Performance metrics in machine learning (no date) www.javatpoint.com. Available at: https://www.javatpoint.com/performance-metrics-in-machine-learning (Accessed: January 22, 2023).

Random Forest (2022) Corporate Finance Institute. Available at: https://corporatefinanceinstitute.com/resources/data-science/random-forest/ (Accessed: January 22, 2023).

S. Holter, O. Gomez and E. Bertini. (2020). FICO explainable machine learning challenge creating visual explanations to black-box machine learning models. [online] Available at: http://www.ml-explainer.com/static/images/FICO-paper.pdf.

S;, S. (no date) Understanding logistic regression analysis, Biochemia medica. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/24627710/ (Accessed: January 22, 2023).

Shrestha, Kshitiz Alsadoon, Omar Alsadoon, Abeer Rashid, Tarik Ali, Rasha P.W.C, Prasad Jerew, Oday. (2021). A novel solution of an elastic net regularisation for dementia knowledge discovery using deep learning. Journal of Experimental Theoretical Artificial Intelligence. 1-23. 10.1080/0952813X.2021.1970237.

Siami, M., Gholamian, M.R., and Basiri, J. (2014). An application of locally linear model tree algorithm with combination of feature selection in credit scoring. International Journal of Systems Science, 2014, 45, (10), pp. 2213-2222.

Siami, M., Gholamian, M.R., Basiri, J., and Fathian, M. (2011). An Application of Locally Linear Model Tree Algorithm for Predictive Accuracy of Credit Scoring. Springer, 2011, edn., pp. 133-142.

Stratified K fold cross validation (2023) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/stratified-k-fold-cross-validation/ (Accessed: January 22, 2023).

Sziklai, Balázs Baranyi, Máté Héberger, Károly. (2022). Testing Rankings with Cross-Validation.

Team, G.L. (2022) Hyperparameter tuning with GRIDSEARCHCV. Available at: https://www.mygreatlearning.com/blog/gridsearchcv/ (Accessed: January 22, 2023).

Thorat, Arvind. (2021). Simply Logistic Regression.

W. Li, S. Ding, Y. Chen and S. Yang. (2018). Heterogeneous ensemble for default prediction of Peer-to-Peer lending in China. IEEE Access, vol. 6, pp. 54396-54406, 2018.

Y. Jin and Y. Zhu. (2015). A data-driven approach to predict default risk loan for online Peer-to-Peer (P2P) lending. Proc. Int. Conf. Commun. Syst. Netw. Technol., pp. 609-613, 2015.

6 Annex

No.	Variables	Before SMOTE-NC	After SMOTE-NC
0	loan_amnt	NS	NS
1	int_rate	S	S
2	installment	S	S
3	sub_grade	S	S
4	annual_inc	S	S
5	dti	S	S
6	open_acc	S	S
7	pub_rec	S	S
8	revol_bal	S	S
9	revol_util	S	S
10	total_acc	S	S
11	mort_acc	S	S
12	pub_rec_bankruptcies	S	S
13	term_ 60 months	S	S
14	home_ownership_OTHER	NS	S
15	home_ownership_OWN	S	S
16	home_ownership_RENT	S	S
17	verification_status_Source Verified	S	S
18	verification_status_Verified	S	S
19	purpose_credit_card	S	S
20	purpose_debt_consolidation	NS	S
21	purpose_educational	NS	S
22	purpose_home_improvement	NS	S
23	purpose_house	NS	S
24	purpose_major_purchase	NS	S
25	purpose_medical	NS	S
26	purpose_moving	NS	S
27	purpose_other	NS	S
28	purpose_renewable_energy	NS	S
29	purpose_small_business	S	S
30	purpose_vacation	NS	S
31	purpose_wedding	S	S
32	initial_list_status_w	NS	S
33	application_type_INDIVIDUAL	NS	S
34	application_type_JOINT	S	NS
35	zip_code_05113	NS	S
36	zip_code_11650	S	S
37	zip_code_22690	S	S
38	zip_code_29597	NS	S
39	zip_code_30723	S	S
40	zip_code_48052	S	S
41	zip_code_70466	S	S
42	zip_code_86630	S	S
43	zip_code_93700	S	S

Table 9: Features selection by Autometrics for dataset before and after SMOTE-NC
[2]