2-2022

Loan Default Prediction System

Ali Abdullatif Ali Albastaki

Loan Default Prediction System

by

Ali Abdullatif Ali Albastaki

A Capstone Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology
RIT Dubai
2022 – February

RIT

Master of Science in Professional Studies:
Data Analytics

Graduate Capstone Approval

Student Name: Ali Abdullatif Ali Albastaki

Graduate Capstone Title: Loan Default Prediction System

Graduate Capstone Committee:

Name: Dr. Sanjay Modak Date:
Chair of committee

Name: Dr.Khalil Al hussaeni Date:
Member of committee

Acknowledgments

I would firstly like to sincerely thank my mentor and professor, Khalil Al Hussain, with whose constant support and mentoring I was able to progress in this journey of the course. With his all-round guidance and support I am able to complete the project with ease and in a timely manner. I would also like to thank all the members and chair of the committee for making this amazing coursework possible for the students, including myself. This coursework and the related projects have helped me understand the field in a better manner, with the help of which I can work in the real-world problems with a new perspective of learning.

I would also like to thank my family for their trust and support in me throughout my journey in life and would like to continue my appreciation towards them for everything. Having said that, I also have my friends and colleagues to thank for who have supported me with my endeavors and journey and contributed in some way or another to my progress and learning.

Abstract

Financial institutions have battled with handling and determining the creditworthiness of their clients in the recent past. The ever-increasing customer base makes it hard for financial institutions, especially banks to follow due process of determining if a customer qualifies for a loan or not depending on one's credit history manually. As a result, there have been delays in processing customer loans, making banks and other financial institutions inefficient. Automation of these tasks has come as a factor of necessity, to improve the speed, cost, and efficiency of processing loans. Thus, an AI web-based application that predicts the probability of borrowers' failure to repay the loans is a handy solution for this time.

The system will auto-collect historical borrowing and repaying data for that particular borrower within the shortest time with high precision whenever an individual uploads the personal data. The prototype will apply cloud cutting-edge AI and machine learning services to analyse the borrowers' creditworthiness and apply the recommendation to achieve the following: Identifying personal information of the proposed borrower, evaluating the prerequisite information for loan approval or decline, determining the credibility, notifying the lender if any loan default history, and recommending approval or disapproval based on the history of the borrower. This application will save financial institutions stress and time, avoid losses in the lending business, reduce the loan process time, decrease the likely risks associated with loans, and save the costs of the admission department.

Keywords: ***creditworthiness, financial institutions, Artificial Intelligence, default, loan.***

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

1.1Background Information

In the modern world of business, lending and borrowing money from financial institutions bring new opportunities to financial organisations, yet there is a challenge of incurring losses following the risk of loan defaulters. This business of lending money to people becomes a frequent business activity for financial institutions. Every day people are seeking to borrow money from different financial institutions for different reasons. Conversely, not every person in need of a loan is reliable, and not all people can be given loans. Moreover, a reasonable number of people on yearly basis do not repay the amount advanced from lending institutions, hence these institutions incur huge losses.

In addition to many people seeking loans from different lending financial institutions, bad loans are significantly affecting the financial sector across the globe. Developing a load predictive model can greatly help these financial institutions to deal with the challenge of giving loans to historically loan defaulters and minimising the risk of incurring huge losses from the number of money defaulters end up not paying back. Using historical data of borrowers can be a great tool in developing a better way of predicting the likely behaviour of a loan applicant and being able to classify the person as a defaulter or non-defaulter.

The process of approving or declining an application for a loan is a significant process for any lending institution. In response, the technological advancement of e-commerce and big data technology can be applied to create a predictive model to categorise each borrower as a defaulter or not using these technologies when financial institutions want to give loans. Therefore, this project is based on the concept of artificial intelligence and machine language techniques to use client's data accessed from reliable financial analytics data websites to ascertain relevant information and predict whether a loan application would be able to repay the loan or not, that is, predict whether the loan applicant in question would be a loan defaulter or not.

1.2 Statement of the problem

Loan processing and approval in financial institutions is a major issue and bottleneck problem, yet integrating machine learning and artificial intelligence (AI) technologies can make a significant change in this important business activity. There is an urgency to develop an efficient loan default predicting system that will become a game-changer in loan processing and approval that can be used by all lending institutions to have fair and successful loan approval systems with the least minimal ratio value of loan defaulters by applying emerging technology to perform time-intensive tasks and making problem-solving more efficient (Tariq et al., 2019). In classifying a loan applicant, the borrower's history about finance would be used. It implies that some predictive data variables to predict the targeted variable as defaulter (delayed or failed to repay the loan on time, 1) or non-defaulter (repaid the loans in time, 0) would be used.

1.3 Project Definition and Goals

- This project aims to build a predictive model to categorise a loan application as a loan defaulter or not using the relevant personal data collected from the historical loan default website when lending is processed and given.
- Minimise the default risk of borrowers ending up not repaying the loans using this created model.

1.4 Methodology

The project proposes the utilisation of emerging technologies such as artificial intelligence and machine learning integration to develop a predictive loan default system in loan processing and approval within lending institutions. The predicting model uses data from the loan default database information shared over financial platforms that aim to minimise the credit risk and bad loans in the financial industry. Additionally, other leading individuals in peer-to-peer platforms will find this proposed predictive loan default application relevant in handling the likelihood of losing money through lending historically known loan defaulters.

In data collection to evaluate the loan applicant's historical borrowing records, data will be gathered from reliable and credible databases of renowned institutions to achieve the initial objective of this proposal of minimising the value of money lost through lending to defaulters. Using the data provided by some research websites like the S&P Global Market Intelligence will help in developing this project. This data will be meaningful in testing the functionality of this application in evaluating the creditworthiness of a loan applicant and aiding decision-making as to whether is a defaulter or a non-defaulter. The study employs data analysis algorithms such as Neural networks, decision trees, random forest, and k-nearest neighbours to analyse the data. The techniques help in developing the trends, patterns, and insights about one’s financial status based on history. After analysis, the final results of each algorithm are compared with the others. The algorithm with the highest accuracy, reliability was chosen as the most suitable for the study.

1.4.1 Data understanding

This entails exploratory data analysis to give an overview of the data sample collected through the presentation in graphs relative to the problem background.

1.4.2 Data preparation

This entails data cleaning to remove the non-useful sets. Data splitting is also done to give the two data sets, one for training the models and another set for testing them. The training data comprise 80% of the total while the testing set comprises 20% of the total data.

1.4.3 Modeling

The 80% training data set is useful in developing the models using algorithms such as Neural networks, Logistic regression, Random Forest, Decision tree, and k-nearest neighbour. The model parameters are chosen by parameter tuning using the cross-validation method.

1.4.4 Evaluation

The evaluation process entails prediction, testing data, and calculation of model accuracy. A comparison of the result of the models is done to determine the best and most efficient model of choice.

1.5 Limitations of the Study

- The data used included both personal loans and joint loans resulting in inaccurate results.
- The dataset contained loan records with various attributes with missing cases.
- The study utilized secondary data sources with second hand information.
- The study focused on the data source from a single source.

Chapter 2 - Literature Review

2.1 Literature Review

The credit score is an influential metric in loan origination and is used in loan processing and approval in many financial lending institutions. According to Sengupta and Bhardwaj (2015) credit scoring metric is useful in ascertaining the borrower's creditworthiness in the current loan application. This is the continued use of credit scoring information across lending institutions aimed at minimising the default loss that these financial institutions incur. Further, the credit scoring metric can be meaningful in observing the loan performance because of its ability to determine the likely credit risk the lending institution can presume to incur in the event the borrower defaults the loan approved.

Lending in finance is increasing and as one way of getting monetary support to meet personal needs without the old credit or bank union (Zhu et al., 2020). So, developing a loan default predictive system is becoming a necessity in evaluating the type of borrowers to give and not give a loan because of the bad loan resulting in huge financial losses to lending institutions. Having a good credit score for a loan applicant is vital for one to get a loan approval or else get declined. Different criteria are used to minimise the risk of a financial institution losing its resources when a loan applicant fails to repay the loan given. Majorly, the lending firms use the historical data of a loan applicant to ascertain whether an individual qualifies to get a loan or not.

Further, individuals and investors seek loans from financial institutions unlike those obtained from a bank. Another type of platform that readily lends money to borrowers includes individual to individual (P2P) in need of cash. Currently, there are online platforms for different financial institutions that offer this lending service to new applicants because of reducing lenders' risk of losing the monies to loan defaulters. Tariq et al. (2020) assert that using technological advancement and information sharing across lending institutions and individuals is taking a new perspective in the lending decision-making process. This proves how this proposed loan default predictive system application will be beneficial to lenders.

Several mobile-based systems have been also developed to help microfinance institutions predict the creditworthiness of their clients. By utilising spatial data, such as travel and expenditure behaviour, these organisations can analyse, determine and classify any customer to a credit level. The system then automatically recommends the amount of credit a customer qualifies for.

However, this project aims to develop a web-based solution for large financial institutions like banks, dealing with a large volume of customers and data.

Alomari & Zakaria. (2017) used machine learning classifiers to predict loan default based on 188,124 loan records from lending club. Random forest classifiers yielded the best performance (71.75%) followed by Naïve bayes classifier (61.44%).The worst was 1R with 59.9%. In a similar study (Xu et al, 2021), they used random forest (RF), extreme gradient boosting tree (XGBT), gradient boosting model (GBM), and neural network (NN) to predict loan default. Data from Renrendai.com was used. Random forest was found to be more superior than the rest of the models. All models achieved over 90% in accuracy.

(Zhu,2019) Used Random Forest, Decision Tree, SVM and Logistic Regression to predict loan default in more than 115,000 records lending club records. Random forest (98%) scored the best followed by Decision tree (95%) and SVM (75%). Logistic regression scored (73%). (Nowshath et al, 2019) used Decision tree, Logistic regression and Neural networks to predict Loan default on another sample drawn from Lending club, Neural networks proved to be the best with with 83.07% followed by logistic regression (80.9%) while decision tree had 79.8% accuracy. (Turiel & Aste, 2020) conducted a similar analysis on lending club data and found Neural networks (DNN) to be the best with 75% recall rates.

From the above reviewed literature, we found evidence that machine learning models can be used to predict loan default, with high accuracy results in most of the scenarios. Most of the reviewed results used lending club datasets in their analysis. The results varied significantly which is to some extent attributable to change in time among other factors. More recent research is therefore needed to provide a current picture of the situation.

(Zhu et al., 2019) used machine learning to develop a new loan default prediction based on a random forest algorithm. The literature also used the SMOTE method to deal with class imbalance problems in the data set.

To predict defaulters, (Aditya Sai Srinivas et al., 2022) employed machine learning algorithms such as KNN, decision tree, SVM, and logistic regression. Metrics such as log loss, Jaccard similarity coefficient, and F1 Score were used to assess the accuracy of various approaches. The metrics were compared to see how accurate the prediction was.

(Aditya Sai Srinivas et al., 2022) employed Random Forest and Decision Tree machine learning models to by examining specific qualities, banking authorities can anticipate if an individual should be granted a loan, enabling them in selecting eligible individuals from a pool of loan applicants.

To forecast factors impacting repayment, the researchers utilised extreme gradient boosting tree, random forest, neural network and gradient boosting model (Xu et al., 2021). The accuracy and kappa value of all four approaches surpass 90%, and RF outperforms the others.

(Aniceto et al., 2020) This study compares the prediction accuracy of Bagging, Support Vector Machine, AdaBoost, Decision Trees and Random Forest models to a Logistic Regression model benchmark. The results of the comparisons are compared using standard categorization performance indicators. When compared to other models, the results reveal that Random Forest and Adaboost are superior. However, utilising both linear and nonlinear kernels, Support Vector Machine models perform poorly.

(Turiel & Aste, 2020) The study applies logistic regression and support vector machine methods to lending data, as well as linear and nonlinear deep neural networks, in order to mimic lender acceptance of loans and estimate the likelihood of default of provided loans.

(Zhao & Zou, 2021) employed logistic regression to forecast the likelihood of loan default using multiple loan characteristics as predictor variables. AIC, AUC, and projected accuracy were used to test and cross-validate the models. Because the loan dataset was stratified, we also examined weighted accuracy.

The research employs logistic regressions, naive bayes, and decision trees (Kisutsa, 2021). The best machine learning algorithm for predicting loan default is then chosen after their performance is compared using performance criteria.

Bagherpour (2017) uses machine learning methods to forecast mortgage default on a huge dataset. To predict loan default, methods used included Support-Vector Machines, K-Nearest Neighbors, Factorization Machines and Random Forest. The study claims that non-linear, nonparametric techniques outperform the classic logistic regression model.

Based on real-life peer-to-peer transactions from Lending Club, Xiaojun, M., et al. (2018) employ unique machine learning methods dubbed LightGBM and XGBoost to forecast consumer default. The methods were used since they have a strong theoretical foundation and practical applicability.

Kvamme, H.et al. (2018) suggest a method for predicting mortgage default based on time series data. Convolutional Neural Networks were used to create the analytical algorithm, which is a sort of Deep Learning model (CNN).

Koutanaei, F.N., et al. (2015) used several selection algorithms. For feature selection, PCA was the best option (Principal Component Analysis). ANN-AdaBoost, an artificial neural network adaptive boosting technique, was shown to be the best model for classification.

Khandani, A.E. et al. (2010) provide a set of variables that may be utilised as input for the model, ranging from the basic credit score debt-to-income ratio to more comprehensive characteristics, and suggest that the latter considerably boosts its predictive potential.

Khashman, A. (2011) presents an approach to predicting credit risk for application by scoring a neural network that considers anxiety and confidence during the learning process.

Beque, A., Lessmann, S. (2017) the study introduces Extreme Learning Machine which compares its performance to that of decision trees, artificial neural networks, support vector machines, and RLR. They suggest that this strategy is a step forward since it combines a high level of prediction performance with a noticeable increase in processing efficiency.

Harris, T. (2013) studied credit risk prediction using a support vector machine by considering a broader rule for up to 90 days and narrow rule for only customers who were 90 days late. He believes that the model employed for the larger definition is more accurate than the other and is more dependable and accurate.

Zhang, T. et al. (2018) present a methodology which uses Multiple Instance Learning for developing a credit score model history. This approach allows for the extraction of features from transactional data.

Papouskova, M., Hajek, P. (2019) presents a two-stage credit risk model: uses ensemble classifiers to differentiate between good and bad payers to predict PD. The second one uses a regression ensemble to determine EAD. The two models are then integrated to forecast the anticipated loss.

2.2 Summary of Literature Review

From the above research on different papers and works of researchers in the same topic, we were able to identify pain points in the finance industry and loan default area. The researchers have also summarised various solution approaches to these problem statements which has helped us understand possible solutions and outcomes for our study as well. We have studied the use of different modelling approaches using Machine Learning models like SVM, Logistic Regression, Random Forest etc. to determine if a customer would default on a loan or not, based on certain factors and the predictive capabilities of these models. The authors have also described different approaches to factor analysis and feature engineering to obtain improved scoring mechanisms at the end. The common problem areas can be summarised in the above research works in the loan market along with the best fit model that can be used for solving such problems.

Chapter 3 - Project Description

In this project, we will source the best available dataset for the related problem statement of predicting loan defaults within customers. This would involve using multiple steps to proceed with the problem statement which is termed as the CRISP-DM method of solving a data analytics or data science problem. It involves multiple steps of solution like business understanding, data collection, data understanding, data cleaning and preparation, modelling and then providing the insights and recommendations to the stakeholders. The below figure summarises the steps that would be involved within the course of this project which we plan to perform one by one.

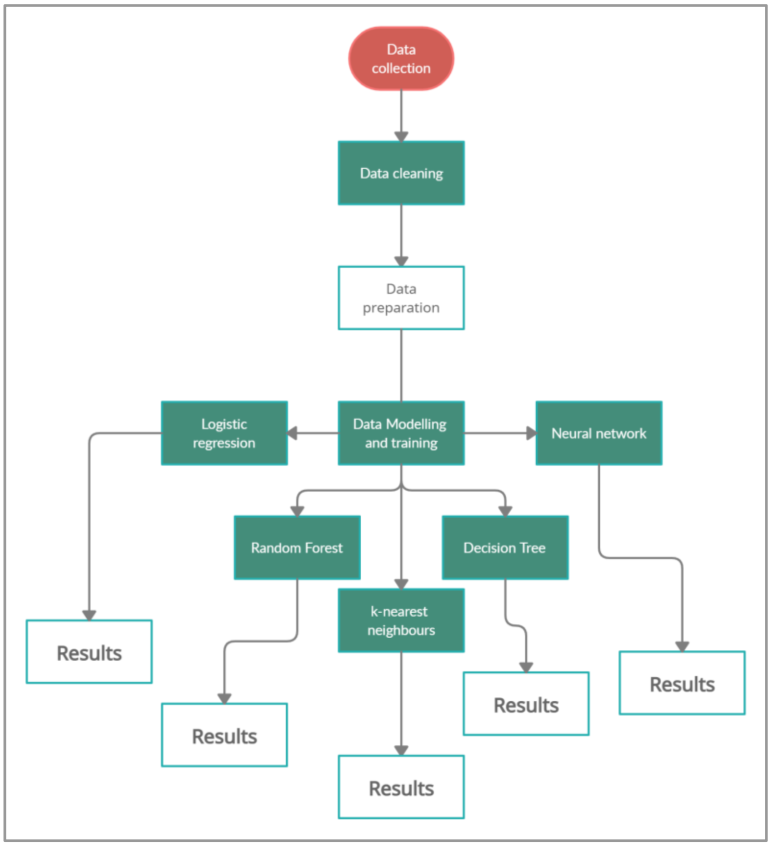


Figure 1: system flow chart

To summarise the steps that would be involved in the course of the project, we outline the major steps below. This would help our readers understand the core steps that involve solving an analytics project in a framework manner.

Business Understanding

It is important that the domain of the report is understood before moving forward with the solution approach. The financial industry along with the loan business has to be properly understood

Data Understanding

Here, the dataset which is obtained from an online repository has been explored well by using data analysis and various statistics to understand the data in depth. Different data description needs to be identified like averages, standard deviations as well as other skewness of the variables in the data

Data Preparation

The obtained data is then prepared by using data cleaning methods to treat missing value and other inconsistencies. This would help us obtain a standardized data for the modeling approach as well as any data analysis

Data Modelling

The cleaned data is then used for the modeling purpose wherein the data is fed into the machine learning models with a split of train and test to the ratio 4:1 for train to test. This helps us to validate the prediction results at the end

Evaluation

In the final step of the entire pipeline, we want to validate and compare the different models that have been obtained. Various evaluation methods are used to identify the best fit model for the solution approach

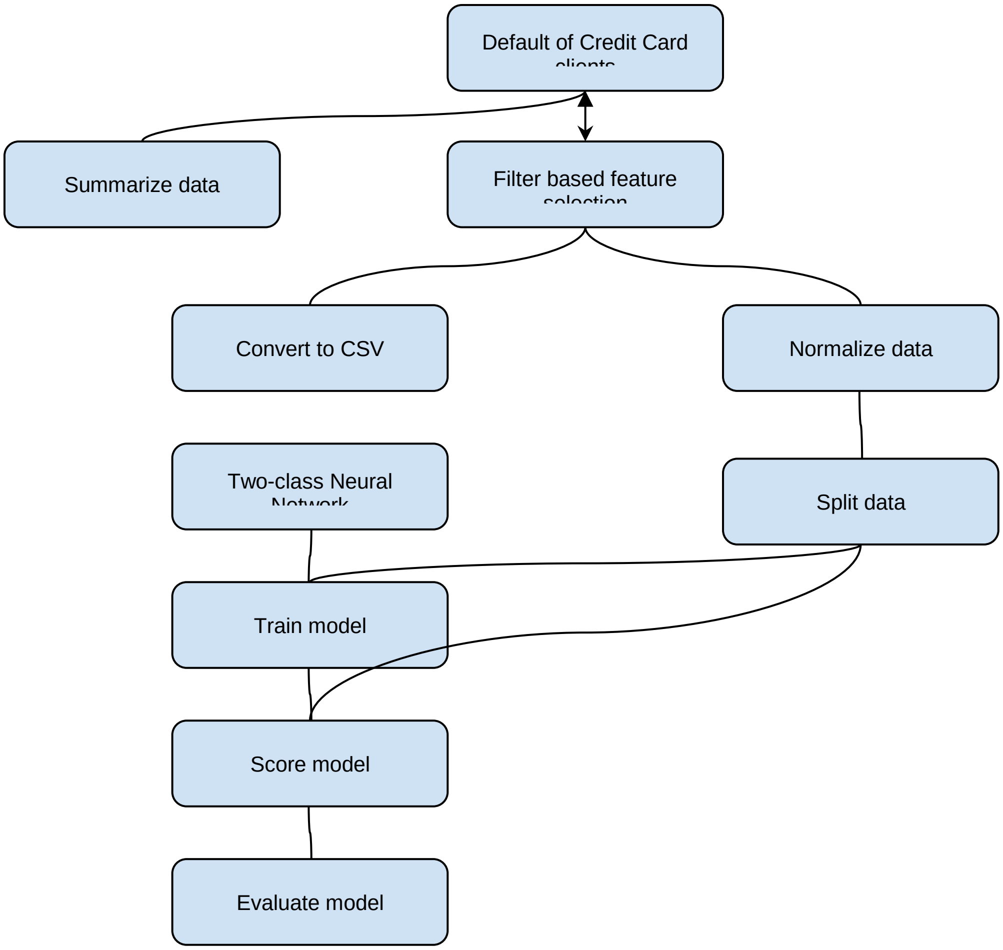


Figure 2: A prediction model

After data collection, to gain useful information different data analytics techniques such as Excel, R programming, and Tableau to process data. Then, in filtering and recognizing meaningful patterns, artificial intelligence, data mining, machine learning, and modeling help in determining prediction in this project. In this project, I will use sample data of loan application forms, and the identity of applicants approved to the lending institution in the past few years ago. I will use data from the financial institution data repository database to predict if a loan applicant will fail to repay or not based on the objective data and whether a lending institution should lend to a loan application or not.

Project variables

An AI system that readily works on historical borrowing data to precise information for decision making.

- R scripts having fitted models from the data.

- The results of the study and the research publication.
- Collected data in a CSV file format.
- Dashboard for the updated creditworthiness of a loan applicant
- Recommendation based on the classification of a loan applicant on the predetermined variables.
- An efficient and fast predictive model that improves loan processing speed.

Chapter 4 - Project Analysis

4.1 Dataset

Data used in this study comes from LendingClub.com, which is a peer-to-peer lending organisation based in San Francisco, California. It consists of details of 2,925,493 loan records for the period between 2007 and the third quarter of 2020. For each loan record, the data consist of 141 attributes, measuring individual and group borrowing and repayment behaviour. The data is available for download on the Kaggle machine learning repository.

4.1.1 Data cleaning and pre-processing

Most machine learning methods use listwise deletion to deal with missing cases. This means cases (loan records) with at least one missing attribute would be excluded from the analysis. Some attributes on the data had so many missing cases and would lead to exclusion of many cases. To avoid such a scenario attributes with more than 40% missing cases were excluded. Conversely, the data includes both individual and joint loans. Some attributes like details of the co-borrower are only possible for group loans. Such details are not available for individual loans and would lead to exclusion of all individual loans if listwise deletion happens; we dropped such variables too to avoid losing much data.

4.1.2 Data partitioning

The data was partitioned randomly into 80% training and 20% testing sets. The training set will be used to fit the models while the remaining 20% will be used to evaluate and compare the performance of the models.

4.2 Exploratory data analysis

Summary statistics and graphing methods were used to understand better borrowers on this sample and their borrowing behaviours.

4.2.1 Frequency distribution of loan outcomes

Figure 1 below shows that 90% of the loan records used to train our models were properly serviced. The remaining 10% of the loans were either on default or had been charged off. The difference between the two loan statuses is; the organisation treats loans with more than 120 without payments as defaulted and charges off defaulted loans if there are no hopes of receiving further payments.

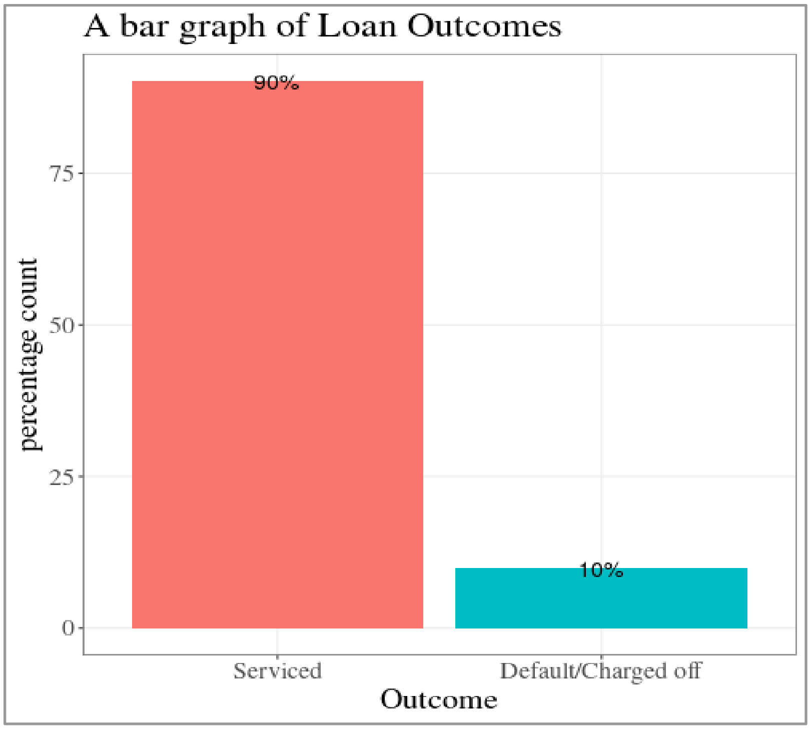


Figure 1: Distribution of loan outcomes

4.2.2 Distribution of loan outcome across independent variables

By comparing loan outcomes across application types we see that defaults/charge offs were slightly higher on individual loans (9.96%) as opposed to group loans (7.80%). On the other hand, borrowers who had a charge off and were working with debt settlement companies had a significantly higher chance of defaulting (99.04%) compared to those who were not working with a settlement company (8.38%). Borrowers on hardship plans had a lower chance of repaying their loan (0.04%) compared to those not on hardship plans (10.40%). For homeownership, people with rented homes had the highest risk of not repaying loans (11.54%) followed by people who own houses (10.06%) and then people on mortgaged (8.30%). Long-term loans had a higher chance of being defaulted (12.18%) compared to loan term loans (8.31%). Verified clients on the other hand had higher default rates (15.02%) compared to source verified (10.44%) and unverified ones (6.48%). See *figure 2* below.

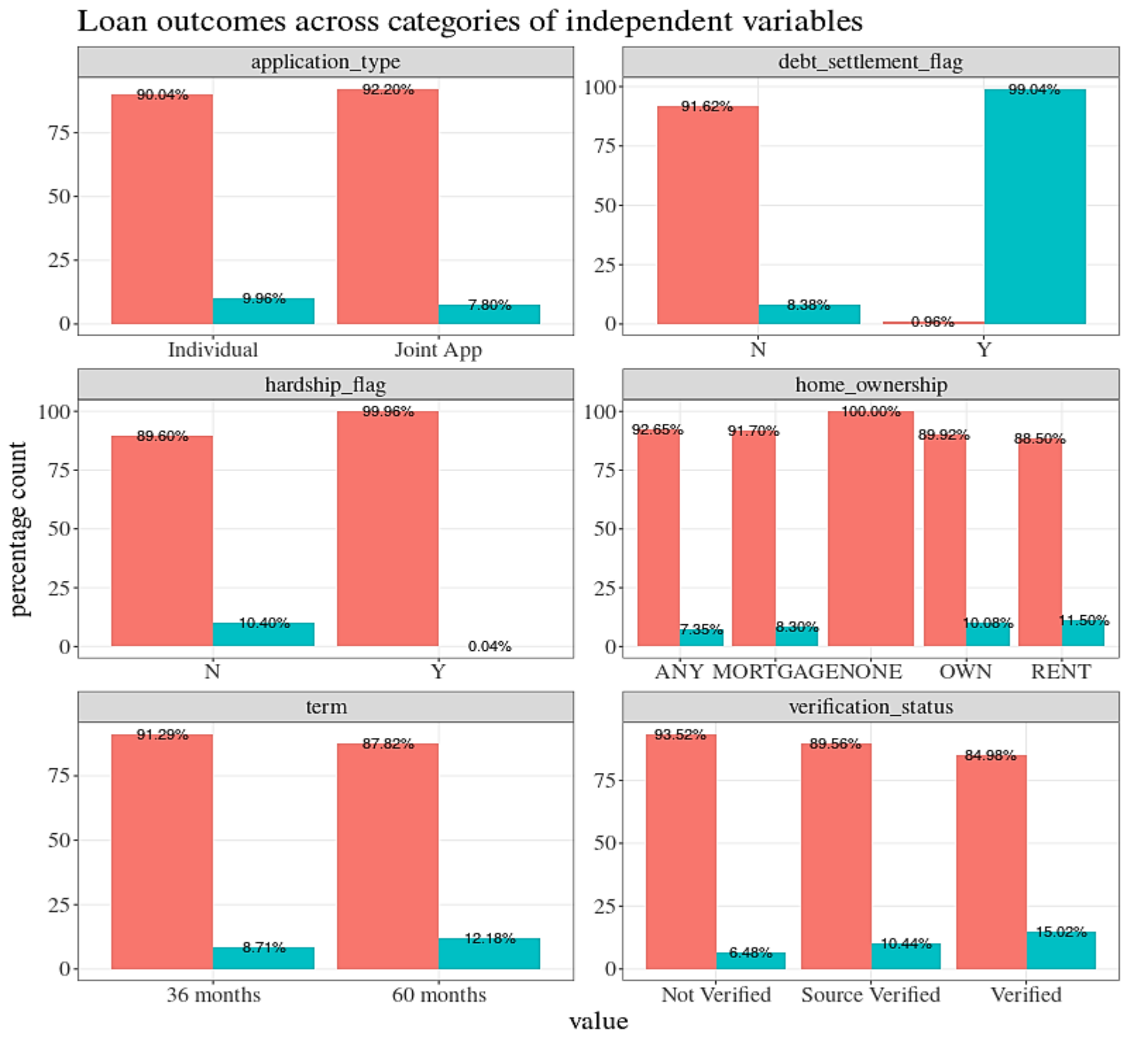


Figure 2: distribution of loan defaults/charge offs across categorical independent variables.

4.2.3 Distribution of independent variables across outcome categories

From *Figure 3* below we see that most of the continuous variables are right-skewed, which means the median is a good measure of central tendency compared to the mean. Loan limit utilisation was seen to be approximately normal while the percentage of trades never delinquent is left-skewed. See *appendix 1* for summary statistics corresponding to the graph panels in *figure 3*.

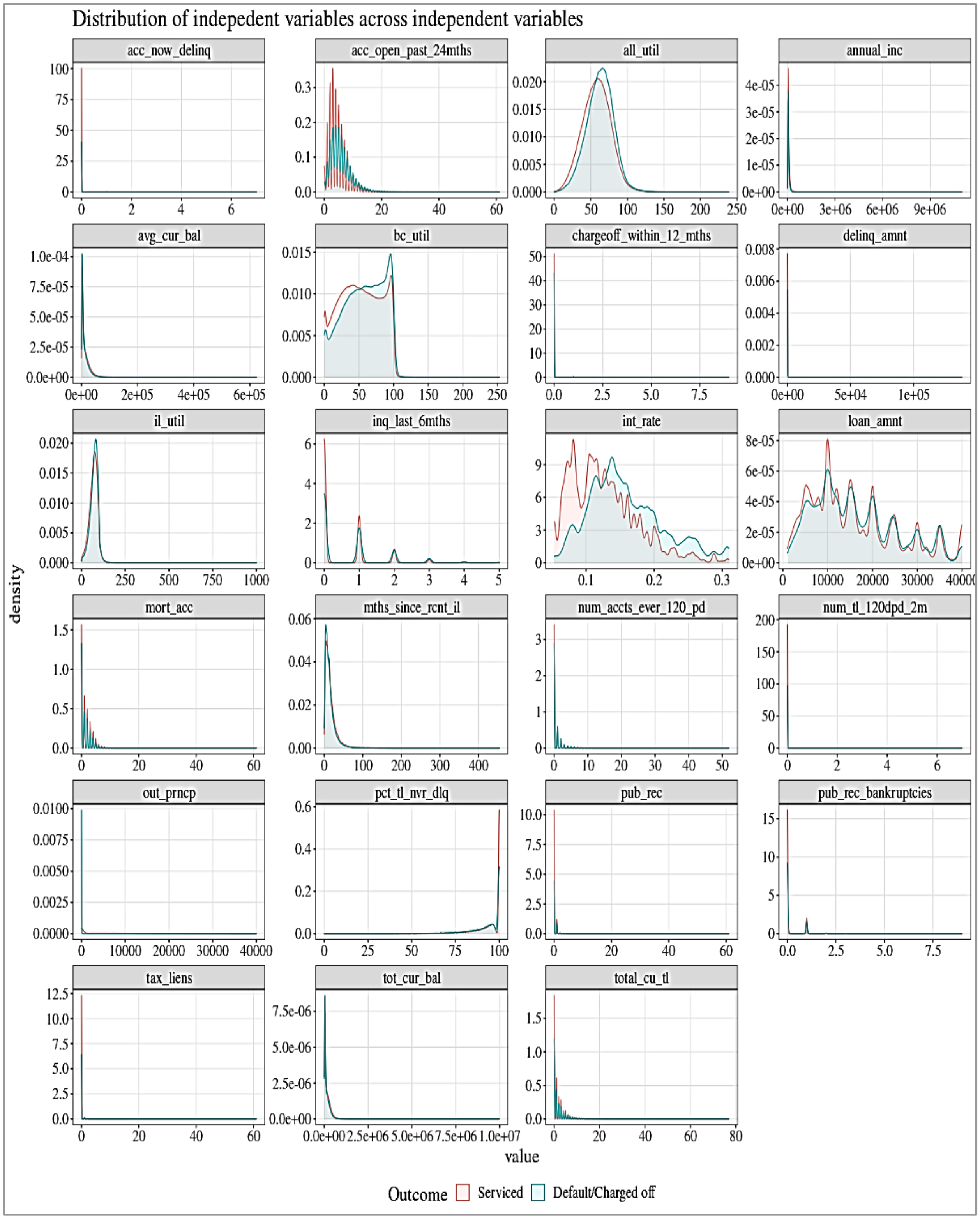


Figure 3: Distribution of independent variables across outcome by outcome categories

4.3 Analysis

4.3.1 Logistic regression

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta(\text{Age})$$

A logistic regression model was built to predict whether an application was likely to result in default or proper repayments. The model assumes a linear relationship between the dependent variables and logs the odds of default/charge off. The prediction equation is:

Logit =

Where $x_1 \dots x_n$ are the independent variables (borrowing and repayment attributes) and $\beta_1 \dots \beta_n$ estimates of effect from each variable have.

Table 1 below reports the estimates of log odds, odds of default/charge off, and their significance. Positive estimates of log odds indicate that the feature is a risk factor to default or charge offs, they can be interpreted by subtracting 1 from the odds. Conversely, negative estimates of log odds indicate protective factors and can be interpreted by subtracting 1 from the odds, for instance; we can see that while the rest of the features are constant, every extra account opened within the last 2 years (acc_open_past_24mths) increase the odds of default by 100*(1.050-1) = 5% times. Similarly, high utilisation of loan limit is a red flag to default, for every extra unit of utilisation, as the other features stay the same, odds of default increase by 100*(1.006-1) = 0.6% of the time. The risk of default also increases with; Reported annual income, average current balance, number of charges off within one year, interest rates, etc *see Table 1 for significant positive estimates*.

Mortgage accounts were among the protective factors, as the number increased by one account while the other factors remained unchanged, odds of default decreased by about 6.6%times. Conversely, the odds for people on hardship plans to default are 93.2% times less compared to borrowers not on hardship plans. *see Table 1 for significant negative estimates*.

Table 1: Logistic regression Estimates

Variable		**Estimate**	**Std. Error**	**z value**	**Exp(B)**	**Pr(>|z|)**	

(Intercept)		-3.001	0.117	-25.607	0.05	< 2e-16	***
acc_now_delinq		0.006	0.06	0.092	1.006	0.93	
acc_open_past_24mths		0.048	0.001	38.417	1.05	< 2e-16	***
all_util		0.006	0	15.502	1.006	< 2e-16	***
annual_inc		0	0	-14.597	1	< 2e-16	***
`application_typeJoint App`	Joint App	0.056	0.014	3.874	1.058	<0.001	***
avg_cur_bal		0	0	-3.259	1	<0.001	**
bc_util		0.001	0	5.91	1.001	<0.001	***
chargeoff_within_12_mths		0.081	0.034	2.399	1.084	0.02	*
delinq_amnt		0	0	4.086	1	<0.001	***
`emp_length	1 year	-0.068	0.017	-3.945	0.934	<0.001	***
	10+ years	-0.149	0.013	-11.146	0.862	< 2e-16	***
	2 years	-0.098	0.016	-6.103	0.907	<0.001	***
	3 years	-0.06	0.017	-3.655	0.941	<0.001	***
	4 years	-0.071	0.018	-3.949	0.932	<0.001	***
	5 years	-0.075	0.018	-4.199	0.927	<0.001	***
	6 years	-0.133	0.02	-6.544	0.876	<0.001	***
	7 years	-0.086	0.022	-3.912	0.918	<0.001	***
	8 years	-0.092	0.022	-4.232	0.913	<0.001	***
	9 years	-0.148	0.023	-6.558	0.862	<0.001	***
home_ownership	MORTGAGE	-0.112	0.104	-1.071	0.894	0.28	
	NONE	0.531	882.7	0.001	1.701	1	

	OTHER	NA	NA	NA	#VALUE!	NA	
	OWN	0	0.1	-0.019	0.998	0.98	
	RENT	0.11	0.1	1.078	1.119	<0.001	
il_util		0	0	-8.716	0.998	< 2e-16	***
inq_last_6mths		0.08	0	18.762	1.088	< 2e-16	***
int_rate		8.26	0.08	104.781	3854.513	< 2e-16	***
loan_amnt		0	0	52.153	1	< 2e-16	***
mort_acc		-0.07	0	-23.348	0.934	< 2e-16	***
mths_since_rcnt_il		0	0	-1.981	1	0.05	*
num_accts_ever_120_pd		-0.01	0	-4.292	0.987	<0.001	***
num_tl_120dpd_2m		-0.11	0.15	-0.706	0.897	0.48	
out_prncp		0	0	-70.13	0.998	< 2e-16	***
pct_tl_nvr_dlq		-0.01	0	-10.404	0.995	< 2e-16	***
pub_rec		0.03	0.02	1.627	1.03	0.1	
pub_rec_bankruptcies		0.04	0.02	1.838	1.038	0.07	.
tax_liens		0.01	0.02	0.729	1.015	0.47	
term	60 months	0.57	0.01	62.569	1.769	< 2e-16	***
tot_cur_bal		0	0	-12.086	1	< 2e-16	***
total_cu_tl		-0.01	0	-10.626	0.986	< 2e-16	***
verification_status	Source Verified`	0.15	0.01	16.419	1.158	< 2e-16	***
	Verified	0.15	0.01	15.04	1.167	< 2e-16	***
hardship_flag	Y	-2.68	0.17	-15.437	0.068	< 2e-16	***

debt_settlement_flag	Y	7.75	0.14	55.191	2326.22	< 2e-16	***
significance codes 0.05 ‘’ 0.01 ‘**’ 0.001 ‘***’*							

4.3.2 Decision tree

A decision tree model works by recursive partitioning to classify whether the outcome of a loan application will be a default/ charge off or it will be fully repaid. The model was trained with a 10 fold cross-validation, Complexity parameter controls how deep the tree grows, a small value allows for the splitting of even smaller nodes which doesn’t improve prediction fit by a significant amount. This might lead to a deeply rooted tree that would likely overfit. Conversely, a large value would mean a split must improve model fit with a huge margin, for it to be considered. To choose an appropriate value the cp parameter was tuned. The best model had cp = 0.0002133307, which corresponds to a cross-validation accuracy of 91.77%. *see figure 4 below*.

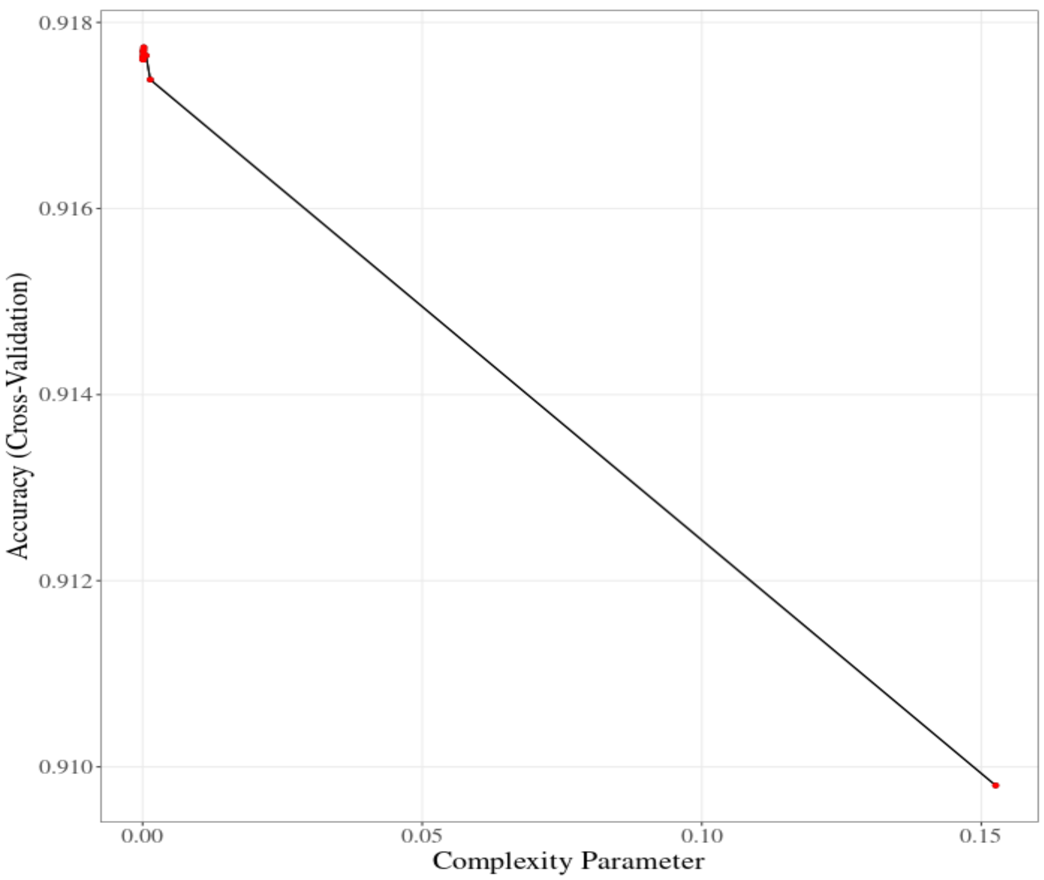


Figure 4: visualisation of cp tuning

The model shows that overall there is a 10% probability that a loan be defaulted/charged off. For borrowers on a debt settlement plan, there is an 8% probability of default/charge off. the probability is 99% for the other group. For borrowers with a debt settlement plan and remaining outstanding principal for total amount funded equal to or less than 0.005, the probability of default is approximately 0%, with greater than remaining outstanding principal for total amount funded equal to or less than 0.005, the probability is 17%. Borrowers under debt settlement plan and with greater than remaining outstanding principal for total amount funded equal or less than 0.005 has a 10% chance of defaulting if the interest rate is less than 14%. If the interest rate is greater than 14%, the probability is 27%.

It is also seen that if a borrower is on a settlement plan, has remaining outstanding principal for total amount funded equal to or less than 0.005 greater than 0.05, the interest rate is greater than 14% and the loan term is 60 months, the predicted probability of default/charge off is 22%. If the loan term is 36 months the probability of default/charge off is 35%. Going deeper we see that if the borrower has more than one mortgage the probability is 31%.see figure 5.

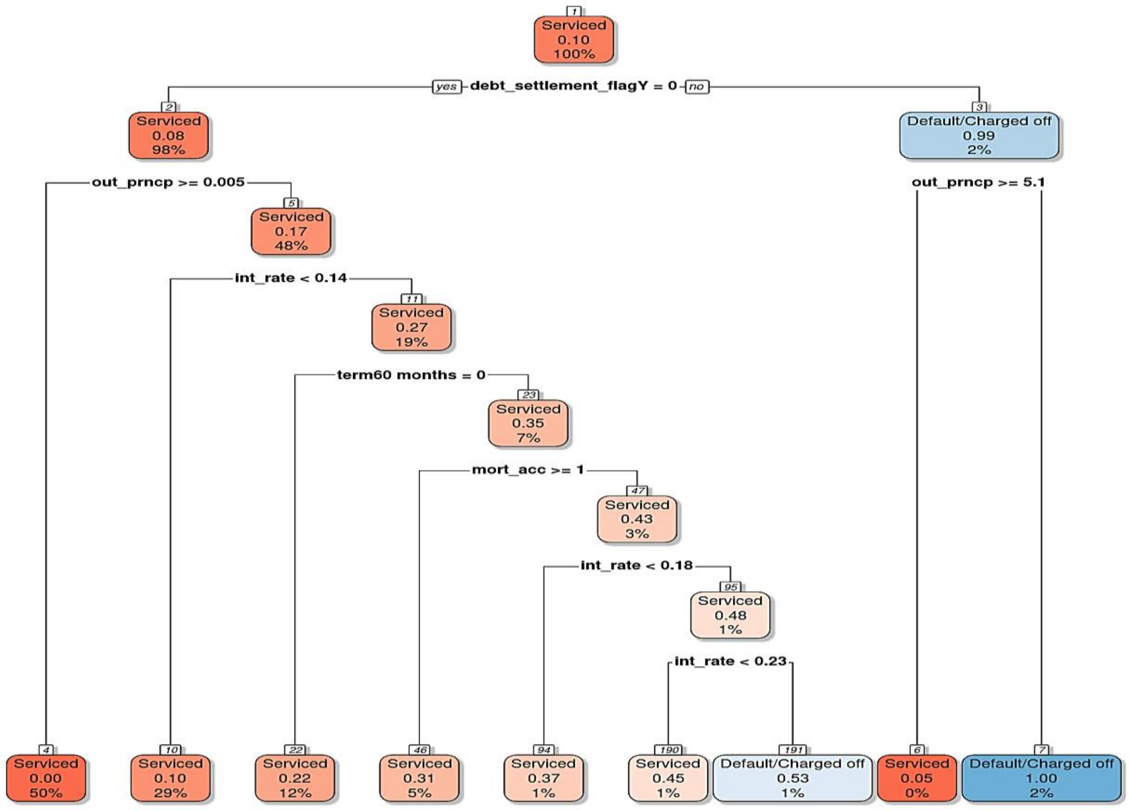


Figure 5: Visualisation of decision tree nodes

A variable importance plot shows that the remaining outstanding principal for the total amount funded is the most important followed by the knowledge of whether a borrower is on a settlement plan and interest rate. Verification status is the least important. This importance ranking is based on how the inclusion of the variable improves mean accuracy. See figure 6 below.

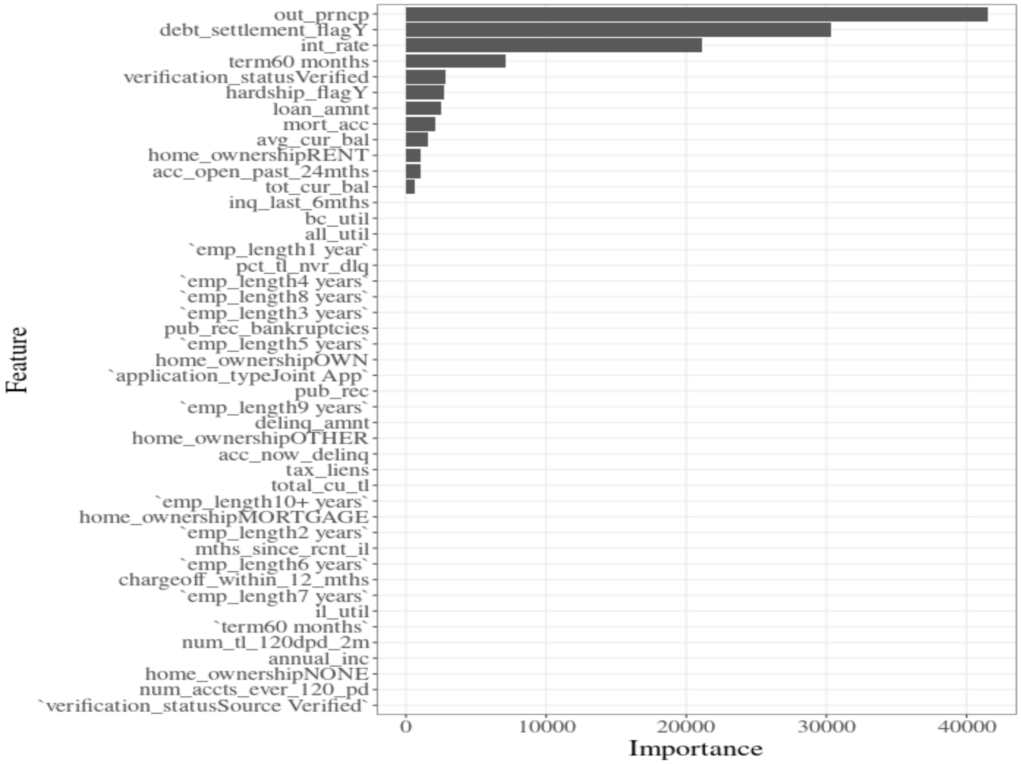


Figure 6: Variable importance.

4.3.3 Random forest

Random forest is a machine learning model which is similar to decision trees but just that it is a collection of decision trees. In this case, the training data is used and fed into multiple decision trees. The number of samples (number of trees) can be adjusted although the tricky part is that the range is too wide to come up with a reasonable search grid. The minimum number of features to include in each sample (mtry) was tuned with 10 fold cross-validation, which suggested that 5 features were optimal. It yields cross-validation accuracy equal to 91.86%. *See figure 7 below*.

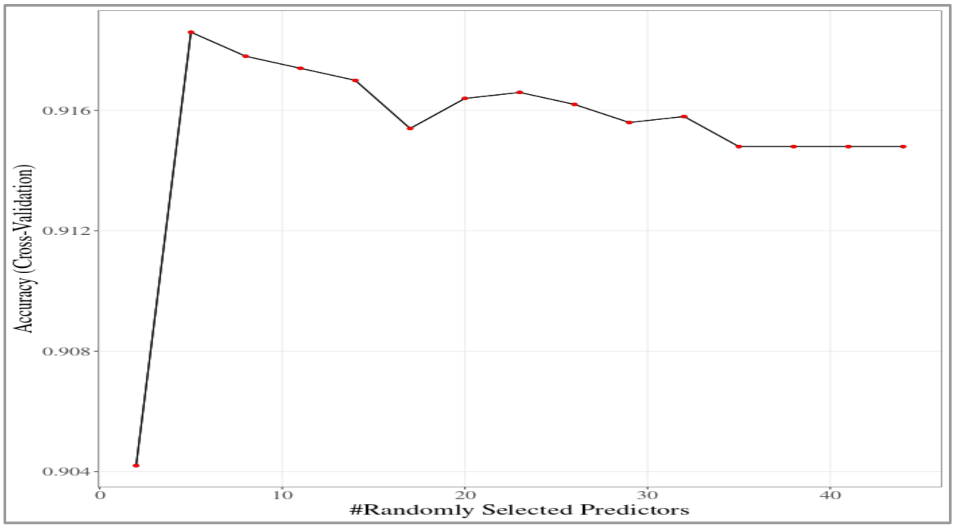


Figure 7: Random forest tuning results

The variable importance for the model ranks the debt settlement flag as the most important variable in prediction default. The remaining outstanding principal for the total amount funded is the second most important while, knowledge of whether homeownership is none is the least important. See

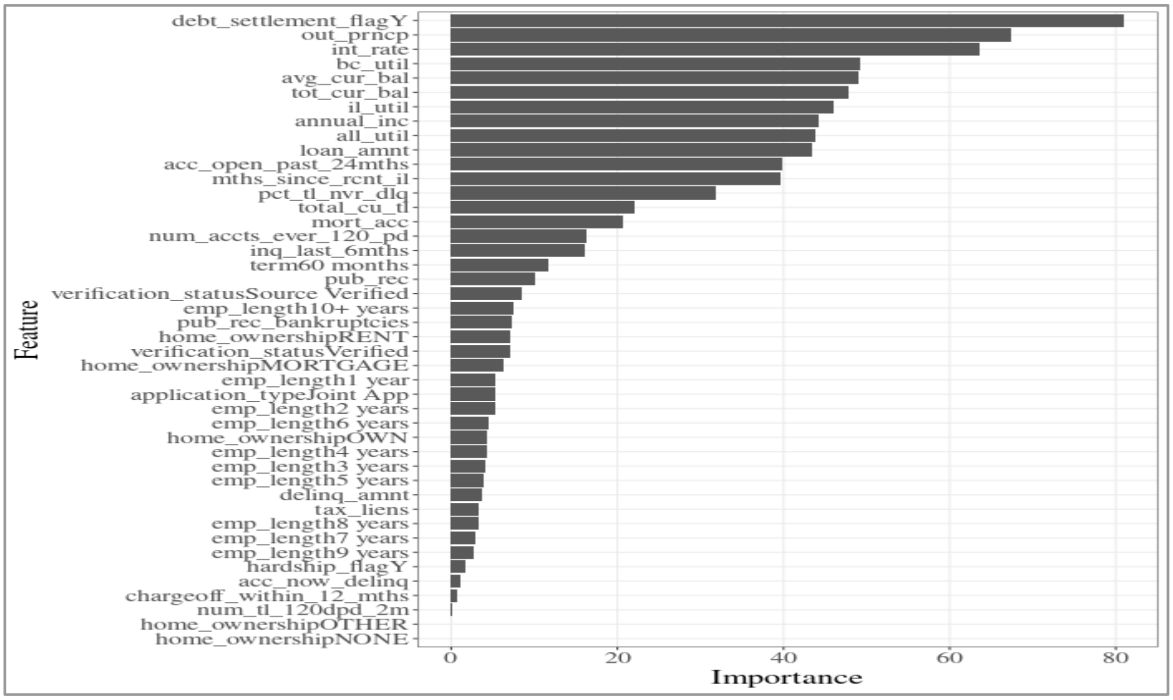
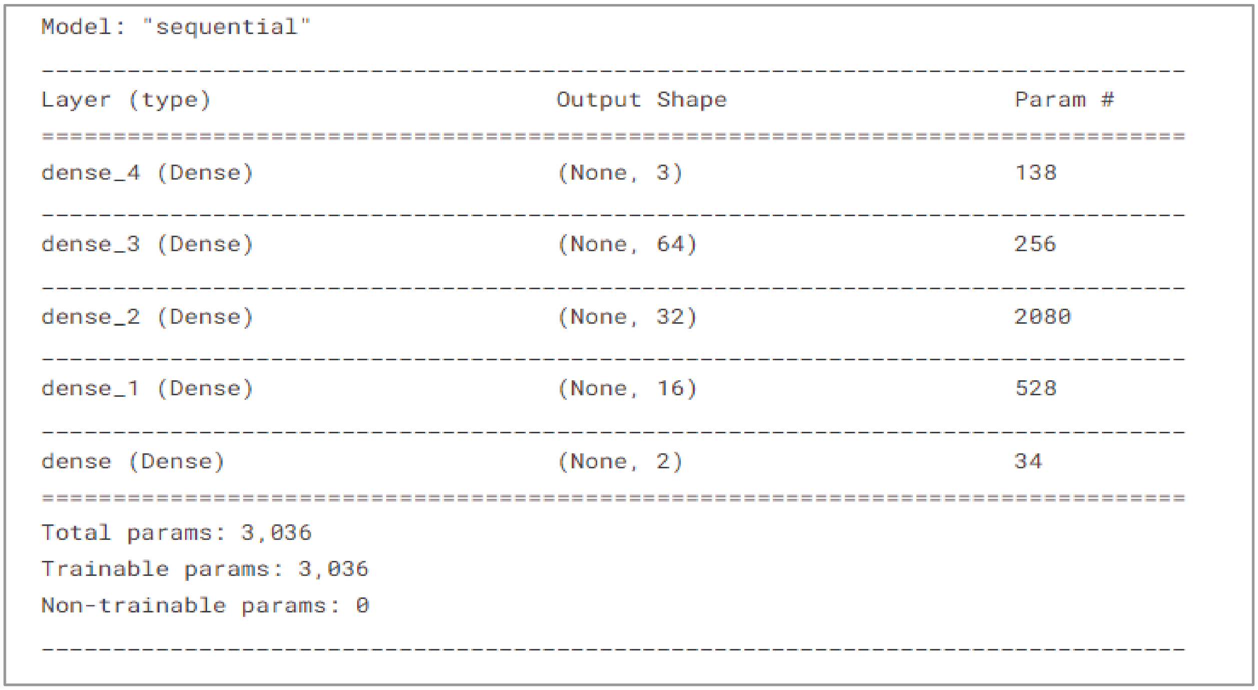


Figure 8: Random forest feature importance

4.3.4 Neural networks

An artificial neural network model with 4 hidden dense layers was fit to classify loan outcomes. The number of units (neurons) was 3, 64, 32 and 16. The activation method used is real for the hidden layers while the hidden layer uses a sigmoid activation function. See *figure 9* below. The model yields around 90.15% on the training set.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 3)                 138
_________________________________________________________________
dense_3 (Dense)              (None, 64)                256
_________________________________________________________________
dense_2 (Dense)              (None, 32)                2080
_________________________________________________________________
dense_1 (Dense)              (None, 16)                528
_________________________________________________________________
dense (Dense)                (None, 2)                 34
=================================================================
Total params: 3,036
Trainable params: 3,036
Non-trainable params: 0
_________________________________________________________________
```

Figure 9: Specified neural network layers

4.3.5 K nearest Neighbour

A k nearest neighbour model was implemented to classify loan application outcomes as default or properly serviced. The model classifies new loan applications with the outcome of the k nearest cases on the training set. During training, model tuning a search fork was done by trying different values of k with a 10 fold cross-validation. A k =31 was found to be optimal, it corresponds to 90.09% accuracy on training data. This means that to classify new loan applications, the model picks the most similar 31 historical loan applications from the training set. If most of them resulted in default then the new case is predicted to result in default. The default similarity index is Euclidean distance.

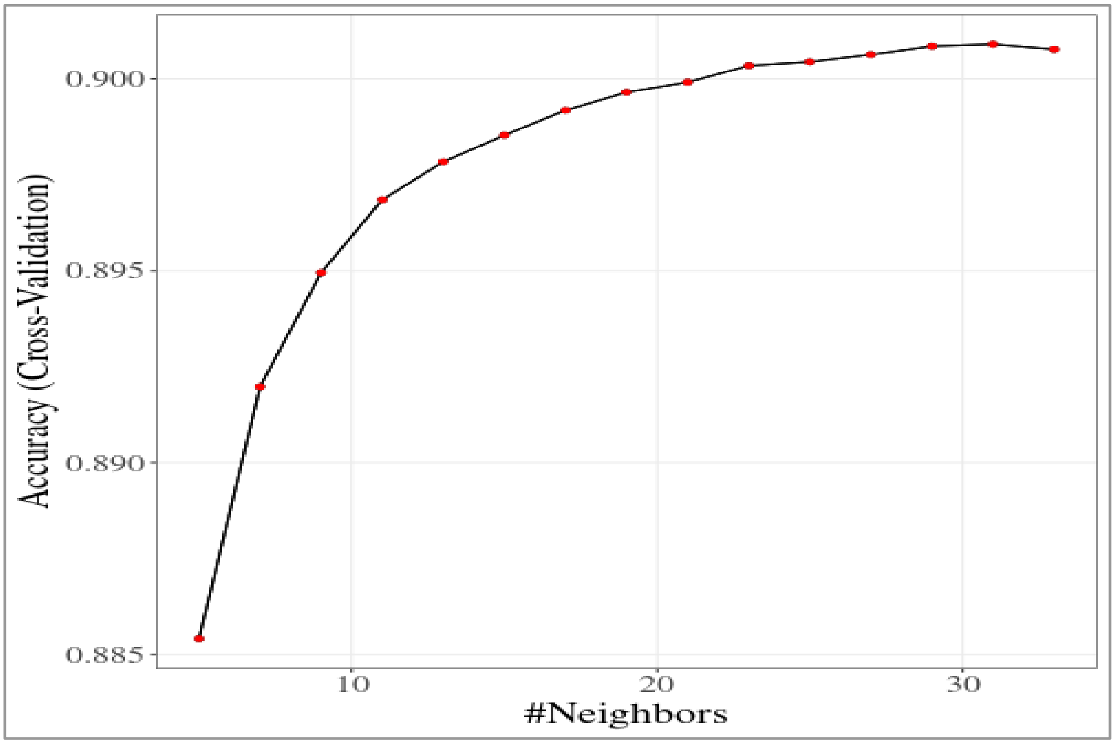


Figure 10: Knn tuning results

4.4 Comparing model performance

The random forest model and decision tree were the best in predicting the outcome of new loan applications; the two models were able to correctly predict 91.86% of the testing data. The logistic regression model came third with 91.84% accuracy; KNN had 90.29% accuracy while neural networks scored 90.14%. In terms of sensitivity. Logistic regression was the best in terms of detecting loan applications that would lead to default/charge offs. Of all loans on the testing set that resulted in default/charge off the model was able to detect correctly 22.08%. The second best is the decision tree with 19.32% I sensitivity. Figure 11 below reports performance measures for all the classification models.

Metric	Logistic regression	Decision tree	Neural networks	KNN	Random forest

Accuracy	0.9184	0.9186	0.9014	0.9029	0.9186
Sensitivity	0.2208	0.1932	0.0000	0.0046	0.1612
Specificity	0.9931	0.9963	1.0000	0.9991	0.9998
Pos Pred Value	0.7750	0.8481	NaN	0.3644	0.9855
Neg Pred Value	0.9224	0.9201	0.9014	0.9035	0.9175
Prevalence	0.0968	0.0968	0.0987	0.0968	0.0968
Detection Rate	0.0214	0.0187	0.0000	0.0004	0.0156
Detection Prevalence	0.0276	0.0221	0.0000	0.0012	0.0158
Balanced Accuracy	0.6070	0.5948	0.5000	0.5019	0.5805

Figure 11: Comparing models

Chapter 5 - Conclusion

5.1 Conclusion

There is ever-increasing lending in finance as one of the ways of receiving financial support to cater to personal needs in the absence of bank unions or old banks. Currently, various financial institutions have established online platforms that offer money lending services to new loan applicants as a way of minimising the potential risks of money loss to loan defaulters. Besides, the microfinance institutions have also introduced various mobile-based systems that utilise spatial data, such as travel and expenditure behaviour to help in predicting an individual's creditworthiness as well as determine and classify any customer to a credit level. Even though most of the financial institutions are currently leveraging the benefits of credit score as an influential metric in loan processing and approval, the absence of fair and successful loan approval systems with the least minimal ratio value of loan defaulters is still a major stumbling block relating to loan processing and approval in the financial institutions.

The escalating instances of loan defaults cause massive losses in money lending companies thus creating an urgency to introduce effective strategies for addressing the identified issue. Developing a model for predicting loan default is critical in minimising the risks related to loan defaults after giving loans to individuals who end up not paying back the money. Emerging technologies such as Machine learning techniques are at the heart of addressing the issue. The techniques are helping in developing a practical predictive model which utilises an individual's historical data to predict their behaviours and classify them either as a loan defaulter or non-defaulters before giving them loans. Such approaches are significant in making useful decisions in financial institutions as far as minimising losses from loan defaults is concerned.

The current research study is associated with various limitations. Firstly, the study utilised a secondary data source that may contain inaccurate information thus producing unreliable results. Secondly, the dataset contained loan records with various attributes with missing cases thus affecting the final results of the study. Thirdly, the study focused on the data source from one money lending company which affected the outcome of the study. Lastly, the data used included both personal loans and joint loans which affected the study outcomes due to the unavailability of attributes such as the details of co-borrowers on individual loans.

5.2 Recommendations

- Based on the limitations of the study, I would recommend the use of primary data sources instead of secondary data sources since it gives first-hand information which could produce more accurate results.
- Also, I would recommend the use of data sources with more records as this would produce more reliable outcomes.
- Finally, I recommend the use of various data sources for different money lending companies to enable a comparison of the results.

5.3 Future Work

The current study was associated with various limitations which creates a need for future studies to address the identified shortcomings. I suggest the following future studies;

- Firstly, a study should be conducted using primary data sources to obtain first-hand information as this is likely to give more accurate results compared to secondary data sources, usually associated with various limitations.
- Secondly, it is critical to carry out another study that utilises a dataset with more records as this would give more reliable study results.
- Finally, there is a need to conduct another study that utilises datasets from various money lending companies to attain reliable results after comparing the outcomes.

Bibliography

Aditya Sai Srinivas, T., Ramasubbareddy, S., & Govinda, K. (2022). Loan Default Prediction Using Machine Learning Techniques. *Innovations in Computer Science and Engineering*, 529–535. https://doi.org/10.1007/978-981-16-8987-1_56

AIS Electronic Library (AISeL) - ICIS 2016 Proceedings: Credit-worthiness Prediction in Microfinance using Mobile Data: A Spatio-network Approach. (2016). AIS Electronic Library. https://aisel.aisnet.org/icis2016/EBusiness/Presentations/28/

Aniceto, M. C., Barboza, F., & Kimura, H. (2020). Machine learning predictivity applied to consumer creditworthiness. *Future Business Journal*, *6*(1). https://doi.org/10.1186/s43093-020-00041-w

Aslam, U., Tariq Aziz, H. I., Sohail, A., & Batcha, N. K. (2019). An Empirical Study on Loan Default Prediction Models. *Journal of Computational and Theoretical Nanoscience*, *16*(8), 3483–3488. https://doi.org/10.1166/jctn.2019.8312

Bagherpour, A. (2017),Predicting Mortgage Loan Default with MachineLearning Methods; University of California, Riverside;

Beque, A., Lessmann, S. (2017), Extreme Learning Machines for CreditScoring: An Empirical Evaluation;Expert Systems with Applications, 86,pp.42-53;

Harris, T. (2013), Quantitative Credit Risk Assessment Using SupportVector Machines: Broad versus Narrow Default Definitions; ExpertSystems with Applications, 40, pp.4404-4413;

Kakouris, R. (2020, June 3). US Loan Default Rate Tops Historical Average -Finally -Led by Retail, Telecom. S&P Global Market Intelligence. https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/us-loan-default-rate-tops-historical-average-8212-finally-8212-led-by-retail-telecom-58895219

Khandani, A.E. et al. (2010), Consumer Credit-Risk Models via Machine-Learning Algorithms;Journal of Banking & Finance, 34, pp.2767-2787;

Khashman, A. (2011), Credit Risk Evaluation Using Neural Networks:Emotional versus Conventional Models;Applied Soft Computing, 11,pp.5477-5484;

Koutanaei, F.N. etal. (2015), A Hybrid Data Mining Model of FeatureSelection Algorithm and Ensemble Learning Classifiers for CreditScoring;Journal of Retailing and Consumer Services, 27, pp.11-23;

Kvamme,H. et al. (2018), Predicting Mortgage Default UsingConvolutional Neural Networks; Expert Systems With Applications, 102,pp.207-217;

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012042. https://doi.org/10.1088/1757-899x/1022/1/012042

Motwani, Anand & Chaurasiya, Prabhat & Bajaj, G.. (2018). Predicting Credit Worthiness of Bank Customer with Machine Learning Over Cloud. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING. 6. 1471-1477. 10.26438/ijcse/v6i7.14711477.

Papouskova, M., Hajek, P. (2019), Two-stage Consumer Credit RiskModelling Using Heterogeneous Ensemble Learning;Decision SupportSystems, 118, pp.3

Sengupta, R., & Bhardwaj, G. (2015). Credit Scoring and Loan Default. *International Review of Finance*, *15*(2), 139–167.

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of Default in P2P Lending. *PloS One*, *10*(10), e0139427–e0139427.

Tariq, Hafiz Ilyas; Sohail, Asim; Aslam, Uzair; Batcha, Nowshath Kadhar (2019). Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (SEMMA). *Journal of Computational and Theoretical Nanoscience*, 16(8), 3489-3503.

Top 12 Data Analyst Tools - Best Software For Data Analysts. (2020). Datapine. https://www.datapine.com/articles/data-analyst-tools-software

Turiel, J. D., & Aste, T. (2020). Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society Open Science*, *7*(6), 191649. https://doi.org/10.1098/rsos.191649

Xiaojun, M.et al. (2018), Study on a Prediction Of P2P Network LoanDefault Based on the Machine Learning Lightgbm and XgboostAlgorithms According to Different High Dimensional Data Cleaning;Electronic Commerce Research and Applications, 31, pp.24-39;

Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: a machine learning methodology. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-98361-6

Zhang, T. et al. (2018), Multiple Instance Learning for Credit RiskAssessment with Transaction Data;Knowledge-Based Systems, 161, pp.65-77;

Zhao, S., & Zou, J. (2021). Predicting Loan Defaults Using Logistic Regression. *Journal of Student Research*, *10*(1). https://doi.org/10.47611/jsrhs.v10i1.1326

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science, 162*, 503–513. https://doi.org/10.1016/j.procs.2019.12.017

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2020). A Study on Predicting Loan Default Based on the Random Forest Algorithm. *International Conference on Information Technology and Quantitative Management.* 162, 503–513.

Appendices

Appendix 1: ***summary statistics for continuous variables across categories of the outcome variable***

name	Outcome	n	mean	sd	median	IQR	min	max
acc_now_delinq	Serviced	1113037	0.002	0.051	0	0	0	7
	Default/Charged off	120314	0.004	0.067	0	0	0	3
acc_open_past_24mths	Serviced	1113037	4.636	3.187	4	4	0	61
	Default/Charged off	120314	5.613	3.573	5	4	0	56
all_util	Serviced	1113037	57.839	19.089	58	26	0	239
	Default/Charged off	120314	62.613	18.106	63	24	1	204
annual_inc	Serviced	1113037	86346.28	93236.623	72000	50000	0	10999200
	Default/Charged off	120314	76978.62	72289.691	65000	42000	0	9573072
avg_cur_bal	Serviced	1113037	13967.91	15610.719	8087	16043	0	623229
	Default/Charged off	120314	11323.6	13104.182	6055	12209	39	288165

bc_util	Serviced	1113037	52.167	28.692	51.8	48	0	252.3
	Default/Charged off	120314	57.347	28.384	59.3	46.8	0	201.9
chargeoff_within_12_mths	Serviced	1113037	0.007	0.095	0	0	0	9
	Default/Charged off	120314	0.009	0.103	0	0	0	4
delinq_amnt	Serviced	1113037	6.467	485.245	0	0	0	138474
	Default/Charged off	120314	15.848	803.726	0	0	0	65000
il_util	Serviced	1113037	68.918	23.251	71	30	0	1000
	Default/Charged off	120314	72.898	22.104	75	28	0	384
inq_last_6mths	Serviced	1113037	0.473	0.757	0	1	0	5
	Default/Charged off	120314	0.637	0.881	0	1	0	5
int_rate	Serviced	1113037	0.126	0.049	0.118	0.064	0.053	0.31
	Default/Charged off	120314	0.159	0.055	0.15	0.07	0.053	0.31
loan_amnt	Serviced	1113037	15726.04	9895.002	13600	13000	1000	40000

	Default/Charged off	120314	16425.37	9531.8	15000	12975	1000	40000
mort_acc	Serviced	1113037	1.443	1.763	1	2	0	61
	Default/Charged off	120314	1.193	1.628	1	2	0	27
mths_since_rcnt_il	Serviced	1113037	16.138	16.265	12	14	0	454
	Default/Charged off	120314	14.664	15.848	11	13	0	397
num_accts_ever_120_pd	Serviced	1113037	0.507	1.446	0	0	0	52
	Default/Charged off	120314	0.602	1.496	0	1	0	34
num_tl_120dpd_2m	Serviced	1113037	0	0.021	0	0	0	7
	Default/Charged off	120314	0.001	0.026	0	0	0	2
out_prncp	Serviced	1113037	6502.885	8708.409	2254.94	10611.69	0	40000
	Default/Charged off	120314	11.645	451.13	0	0	0	37003.92
pct_tl_nvr_dlq	Serviced	1113037	94.434	8.995	100	8	0	100
	Default/Charged off	120314	93.793	9.231	97.4	9.1	12.5	100

pub_rec	Serviced	1113037	0.158	0.478	0	0	0	61
	Default/Charged off	120314	0.248	0.639	0	0	0	61
pub_rec_bankruptcies	Serviced	1113037	0.118	0.342	0	0	0	9
	Default/Charged off	120314	0.163	0.41	0	0	0	8
tax_liens	Serviced	1113037	0.028	0.288	0	0	0	61
	Default/Charged off	120314	0.058	0.433	0	0	0	61
tot_cur_bal	Serviced	1113037	157976.6	169032.243	94286	200813	1	9971659
	Default/Charged off	120314	129104.5	142604.298	67079	159266.75	119	2881652
total_cu_tl	Serviced	1113037	1.629	2.79	0	2	0	77
	Default/Charged off	120314	1.635	2.822	0	2	0	54