Prediction of Charged-off Loans for P2P Online Banking using Classification Models and Deep Neural Network

MSc Research Project
Master's in Data Analytics

Bharat Bhardwaj
Student ID: X18186424

School of Computing
National College of Ireland

Supervisor: Dr Rashmi Gupta

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Bharat Bhardwaj		
Student ID:	X18186424		
Programme:	Master's in Data Analytics	**Year:**	2019-2020
Module:	Research Project		
Supervisor:	Dr Rashmi Gupta		
Submission Due Date:	25 September 2020		
Project Title:	Prediction of Charged-off Loans Using Classification Models and Artificial Neural Network for P2P Online Banking		
Word Count:	8849	**Page Count**	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Bharat Bhardwaj

Date: 25 September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Charged-off Loans for P2P Online Banking using Classification Models and Deep Neural Network

Bharat Bhardwaj
X18186424

Abstract

Peer-to-Peer (P2P) lending operation is an innovative way of lending activity to invest and borrow money than traditional banking operations. P2P online banking gives an internet platform where investors and borrowers meet directly without any middleman which could be a shared benefit of high return and low interest between investors and borrowers. Social lending operations are based on peers, where investors face the direct risk of damages in case borrowers does not pay the loan amount. Hence, accurate prediction of charged-off loan is necessary in P2P online banking. This research paper presents a study to predict the charged-off loans accurately on using unlabelled data of P2P LendingClub online platform. Our study utilises multi-dimensional loan data for mining activities and presents statistical observations for the data. In further research, the study applies modern Logistic Regression (LR), Random Forest Classifier (RFC), K-Nearest Neighbour (KNN) along with hyperparameter tuning and Artificial Neural Network (ANN), which uses real transactional data of LendingClub. The study compares the outcomes of the classification model and artificial neural network results to identify the suitable model in prediction of charged-off loans for LendingClub investors.

Contents

1 Introduction

Finance operations over the internet have rapidly changed due to the advancement of internet technologies and the arrival of big data. Finances over the internet have gained popularity whether it is a government or an organisation or a financial institution. Online P2P lending also known as social lending, is establishing as an alternate option to a bank where an individual member does borrow and lend the money without the help of intermediate banks. Social lending comes up with a great benefit of mutual profitability in lending operations. It focused on the direct connection of borrowers and investors, and assure that both could benefit the profits. P2P lending comes with higher return benefits and low-interest rates among many convenient channels than traditional banking operations. P2P lending operations gives a platform where a borrower may get the money at lower interest rates and investors may earn better loan interest rates in comparison to investing in traditional banking operations. P2P lending platform allows lenders, where they can find potential borrowers and choose one of a borrower to whom they wish to lend the money. As P2P lending operations operate over the internet, it does not ask much efforts to reach a smaller number of communities such as towns, villages, religious or ethnic groups. Besides, P2P lending operations are more convenient, transparent, and faster in terms of lending operations when compared with traditional banking operations.

LendingClub Corp.[1] and Prosper[2] popular P2P online banking platforms in the US. Besides, there are many more P2P platforms available around the world such as UK based Zopa Ltd.[3], Germany based Smava GmbH[4]. Each of these online banking platforms relies on the credit reporting agency's credit score such as TransUnion LLC, Experian, Equifax Inc., and Schufa Holding AG, respectively.

Information economics believes that asymmetric information could result in adverse selection (Akerlof, 1970) and moral hazard (Stiglitz,1981) that could provide a theory to cause a credit risk due to limited information of borrower's creditworthiness. This situation could be relating to P2P online banking operations where investors do not have much information about the borrower whether he could repay the loan or having enough creditworthiness. Further, In P2P microfinance operations, most borrowers are having either small business or individuals who may under poor economic conditions. This makes investors focus on a high rate of interest regardless of borrowers creditworthiness. This cause a problem in front of a lending platform, how they could help members in their adverse selections? According to Iyer

[1] http://www.lendingclub.com.

[2] http://www.prosper.com.

[3] http://www.zopa.com.

[4] http://www.smava.de.

et. al. (2009), lenders with soft information such as communication with the borrower is much helpful in successive borrowing operation than hard information (information comes from documents) of traditional banking operations. According to Lin et. al (2013), social networking could help lenders in identifying the borrowers for lending operations from largest online P2P Prosper platform. There are multiple works of literature which are focused on behaviour analysis of borrowers and investors during the lending process. While some of the literature presented machine learning approaches to accommodate the information and performed statistical analysis on P2P lending operations to predict the default risk. As per, Olson et. al. (2012), Decision tree model performed well in prediction of default risk. While according to Moro et. al. (2014), the neural network performed well when compared with decision tree models. These differences in the prediction of default credit risk highlight the problem scenario and justify applying different machine learning models in prediction of charged-off loans for online banking operations.

Our study is focused to use classification models with hyperparameter and neural network to predict charged-off loans for online lending operations. This study infers the Logistic Regression, Random Forest Classifier and K-Nearest Neighbour. To tune the hyperparameter study applies the GridSearchCV that will help to pick the optimal parameter using the grid search techniques. The study also applies the ANN model to predict the charged-off loans for online lending operations to understand how deep dense layers could help to improve the model performance in prediction of charged-off loans. The goal of this study is to build a machine learning model to predict the probability that a loan is charged-off. The study will use only the data that are available to investors of LendingClub Corporations such as FICO score, income, debt-to-income ratio, employment length, loan amount, loan grade, loan purpose, interest rates, instalments etc.

The main objective of this study is to build a machine learning model which could help investors in making decisions for their investments and save them from the credit risk or asset loss. The study is also aimed to identify the important features of borrowers which could be helpful for investors in their loan approval decisions. The study is aimed to improve machine learning model performance using hyperparameter tuning via GridSearchCV techniques. This study also observes computational cost in the process of data mining activities from complex data. The structure of this paper considers the loans that are accepted by LendingClub Corporation under its creditworthiness.

1.1 Research Question

How well can classification models such as logistic regression, random forest classifier, K-nearest neighbour and artificial neural network identify charged-off loans for online P2P banking?

The study is aimed to improve the performance of classification models and observe the computational cost while processing the data mining activities on a large amount of data. The study develops a classification model along with hyperparameter tuning and evaluates the models using 5-folds cross-validation techniques. We believe that by changing the class weight and the number of layers for an artificial neural network the performance could be improved in the prediction of charged-off loans.

1.2 Document Structure

This proposed study is organised as follows: Section 2 will review the literature and explores the research study on borrowers and investors. Section 2 also explores researches on different machine learning models application that are inspired by a different objective to explore more information from online P2P lending operations. Section 3 will present the methodology that is applied to develop this study. Section 4 will explain the architecture design and implementation techniques for the study. Section 5 will give brief details of machine learning modes evaluation techniques and share the results. Section 6 will discuss the state of the art for the study. Finally, section 7 will share the inference of research work.

2 Literature Survey

A good amount of economic and information technology research has been carried out on P2P lending operation since their transactional data are opened for research work. In recent years, academic research is gaining popularity in P2P lending operations.

2.1 Research on borrowers and investors

According to Klafft (2008), borrowers who are having less credit score is having less rate of success, high-interest rates and high default rates. According to Iyer et. al (2010), analysed the default rate as per the credit score of borrowers and finds that investors in P2P lending operations had a good credit screening ability. Borrower default rate prediction was 45% accurate in comparison to take a decision, based on credit score and their 87% of prediction accuracy are based on the economic condition of borrowers. According to Freedman et. al. (2011), analysed the internal rate of return (IRR) and finds that IRR was initially declined as Prosper Corporation lenders underestimated the market credit risk and concluded that lenders can identify the high-risk borrowers by calculating internal rate of return. While much of lenders do prefer to lend them money in locals only. Yuelei Li et al. (2013), analysed heard behaviour where investors know creditworthiness of borrowers by some local middleman reference, which caused an important role in successful lending operations for china P2P platform. As per Everett (2015), default risk could be reduced in case investors and borrowers know each other. The investors may able to judge the default risk based on borrower's public information while in other cases borrower is having more information in their hand.

As per the above research review, it has been evident that investors are much conscious to lend them money and have their own criteria to offer a loan other than borrowers credit score. Above research review also gives a hint that geographical area also affects the investors' decisions and tend to high chance in loan succession rates. It also indicates that investors tend to have as much information about the borrowers as they can and hence high priority to give loan in locals. It will be helpful for this study to identify the investors feature selection in terms of charged-off loan decision prediction.

Next, Study will explore the prediction models research review to understand what machine learning models could play an important role in prediction, in terms of financial data analysis.

2.2 Research on prediction models

Credit risk forecasting can be classified as per the type of forecasting methods such as statistical methods, mathematical programming, and Artificial Intelligence (AI). Discriminant analysis was the original way to forecast credit risk on P2P lending operations. But this requires strict assumptions such as predictive variables should follow multivariate correlation and should not be correlated significantly to get a good accuracy in prediction, which makes it difficult in a real-time scenario. Regression analysis is one of the popular among the researchers. In terms of mathematical programming, Mangasarian (1965), was the first to propose linear programming in pattern separation. Tam and Kiang (1992), applied a neural network to predict the default loans and compared the results with logistic regression, linear classifier, and decision tree. Results showed that the neural network outperformed in compare to other methods. Shi et. al (2002), extended the research work and applied multiple criteria linear programming on credit card problem. This comes with the limitations that mathematical programming is optimal for the small dataset samples and showed the slow convergence.

Recently many researchers have done statistical analysis on social lending operations. Most of the research has been carried in identifying the relationship with default state by identifying the major attributes using statistical analysis. Emekter et. al (2015), analysed attributes of borrowers using logistic regression discovered the important features of borrowers and loan products which could be helpful for the P2P investors.

In recent years studies have been carried out to predict the default prediction and to improve the performances of online banking operations. Bitvai and Cohn (2015), developed a model to improve the performance of default prediction including the number of words in borrower’s loan application form. Guo et. al (2016), developed a framework which was based on a data-driven methodology for the P2P market. They developed a model which was instance-based to assess the credit risk and used the logistic regression that is helpful to assess the risk and return for every individual loan. However, this framework has limitations where it focused on improving the performance in default prediction rather than defining a relationship to predict the default and borrower's additional data. Ge et. al (2017), developed a method to predict the obligation attributes of borrowers by using their social media information. Lin et. al. (2017), developed a model for credit risk assessment for Yooli dataset which is a P2P lending platform in China. They used a non-parametric statistical method to assess the demographic data of borrowers and extracted the data which could affect the default loans

Machine learning is used to solve complex problems such as credit scoring, debt scoring and default prediction as well. Serrano-Cinca and Gutiérrez-Nieto, (2016), developed a credit score system using a decision tree. According to Xia et.al (2017), developed an XGBoost model to predict the default by analysing the internal return rate (IRR). Huo et. al (2017), analysed the behavioural scoring issues for P2P lending operations and used the Neural Network along with the logistic regression model. However, some of the research has been carried out to improve the performance by using ensemble techniques.

Recently, few researchers used deep learning techniques to improve the performance of default prediction. Fu (2017), combined neural network and random forest model to predict the defaults. Jiang et. al. (2017), suggested the combination of P2P loan and text description could be helpful when support vector machine and random forest model are used combinedly to predict the loan defaults. There was one more research has been made on semi-supervised learning for P2P lending operations. Ma et al. (2018), applied XGBoost and LightGBM machine learning model using multidimensional and multi observation data cleaning algorithms. However, the limitation was that the experimental data was lesser in the amount in consideration of the complexity of the model application.

As per the above research review, it has been found that logistic regression and neural network application on financial data was much popular among the researchers. While another machine learning model such as random forest, XGBoost, LightGBM and decision tree was also applied to improve the accuracy in default prediction. Researchers have applied different techniques along with machine learning model applications such as ensemble techniques, statistical analysis, and mathematical programming to identify the patterns of relationship among the loan data attributes.

Next, this study will summarize the learning from the literature review of investors, borrowers and prediction models and analyse the suitable models to apply for the P2P online banking domain.

2.3 Literature review output summary

Research on the prediction model says that there are many new machine learning models has been applied in terms to improve the accuracy of default prediction. Most of the applied techniques have used the traditional way of the machine learning model. Table 1 summarizes the literature on default prediction for P2P lending operations. We see that logistic regression and the random forest is widely used to predict default on financial data. This study finds that a large amount of research has been carried out on social lending operations. Much of the research has used logistic regression but very few are executed on a large amount of data. The closest research to our study is Emekter et. al (2015) and Byanjankar et. al (2015). Emekter et. al (2015) observed logistic regression but not on a large amount of data. Also, the authors aimed to identify attributes that could play a key role in the loan approval decision of investors. Only Fu (2017) et. al has done research on a large amount of dataset and applied random forest. We believe that very few researches such as Huo et al. (2017) been taken place with a neural network but that is also done on a very small amount of dataset. Nonetheless has researched on a prediction of charged-off loans but carried research on loan default prediction. Researchers were having different objective while observing the social lending operation data such as default prediction, credit risk assessment, behavioural scoring and reject inference in credit scoring.

Table 1: Default prediction study in P2P lending

Author	Task	Methods	Data characteristics	Data	Results
Emekter et. al (2015)	Credit Risk	Logistic regression	P2P transaction	58,864	debt-to-income ratio, revolving line, FICO score, Credit grade play key role in defaults prediction.
Byanjankar et. al (2015)	Credit Risk	Logistic regression and Neural Network	P2P transaction	16,037	Results say that the neural network-based credit scoring model performs well in screening default applications.
Bitvai and Cohn (2015)	Default prediction	Bayesian non-linear regression	Sentences	43,881	Bayesian non-linear methods outperformed for predicting market rates and generates substantial profit in a trading simulation.
Guo et al. (2016)	Default prediction	Logistic regression	P2P transaction	20,16,128	Results revealed that the logistic regression model overperformed and effectively improved investment performances in P2P lending.
Serrano-Cinca and Gutiérrez-Nieto (2016)	Internal rate of return	Decision tree	P2P transaction	40,901	profit scoring system using a decision tree model outperforms the results of logistic regression.
Ge et al. (2017)	Default prediction	Logistic regression	Social media information	35,457	Results suggested that borrowers' social information can be used not only for credit screening but also for default reduction and debt collection.

Lin et al. (2017b)	Default prediction	Logistic regression	P2P transaction	48,784	Results say that age, gender, educational level, monthly payment, loan amount, marital status play a key role in defaults loan.
Xia et al. (2017)	Default prediction	XGBoost	P2P transaction	49,795	XGBoost enhanced the capability of discriminating potential default borrowers.
Huo et al. (2017)	Default prediction	Logistic regression, Neural network	P2P transaction	4,518	the combination neural network and logistic regression slightly overperformed than the individual model application of neural network or logistic regression model
Fu (2017)	Default prediction	Combination of random forest and neural network	P2P transaction	13,20,000	combination of neural network and random forest model overperformed than single neural network but less performed when compared to only random forest model.
Jiang et al. (2017b)	Default prediction	Support vector machine, random forest	P2P transaction	39,538	Random forest model overperformed than Support vector machine

It is to be noted that unlikely, default prediction, this study will predict ‘Charged-Off’ loans i.e. loans that are classified to be in default state for some time and financial institution has not found any way to recover the amount from borrowers (ex.- in case of insolvency or bankruptcy). A loan firstly becomes ‘Default’ and after some time if banks unable to recover the amount, the loan becomes ‘Charged-Off’. However, banks do write-off that type of loans, but it does not mean that borrower need not to pay the amount in case of loan is charged-off. Charged-off loans are truly asset loss for the investors and hence a better prediction model is necessary to identify a true borrower.

The closest research to ours is Emekter et al. (2015) work, where authors researched on LendingClub data from 2007 to 2012 and used logistic regression to predict the default for borrowers. They also identified that revolving line utilization, FICO score, debt-to-income ratio and credit grade does play an important role in predicting the default for borrowers. It is to be noted that our study uses data from 2007 to 2018 and instead of applying simple logistic regression, this study applies classifiers (Logistic regression, Random forest classifier and k-nearest neighbour) along with hyperparameter search techniques using GridSearchCV. This study extends the model application by applying artificial neural network to predict the charged-off loans and explore new borrowers features which could play an important role for investors in loan approval decisions.

In the next section 3, Study will give brief details of applied methodology in the prediction of ‘Charged-Off’ loans for the online P2P lending operations.

3 Research Methodology

Data mining is a tedious process that involves many steps to gain information from data. There are many processes evolved to simplify the data mining process like CRISP-DM, SEMMA and Knowledge Discovery in Databases (KDD). This project is following KDD in terms to gain the information from LendingClub dataset. KDD is a way to get meaningful information from data. To gain knowledge from data, KDD gives a technique to develop methods, so data could be translated into knowledge. Our study uses the KDD process as shown in Figure 1. This starts with data collection to gain data knowledge. Down the line study will explore all steps in details.

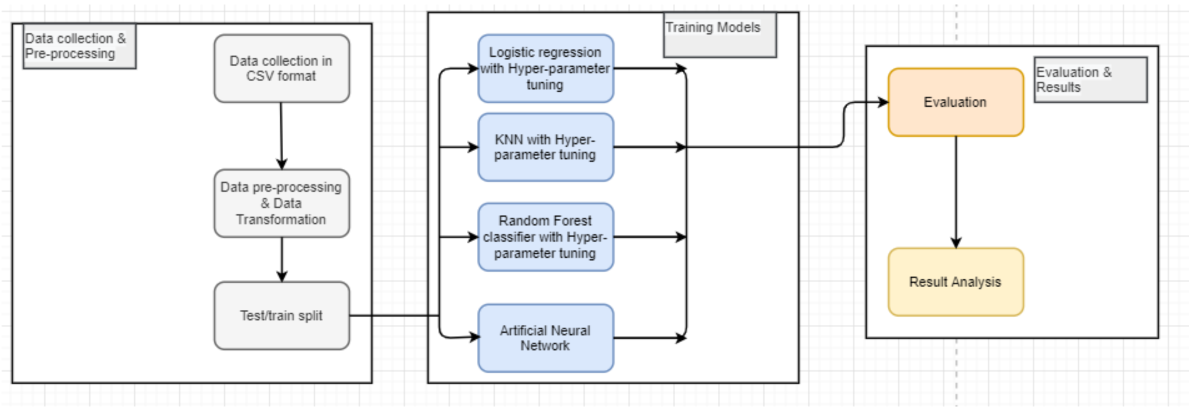

Data collection &
Pre-processing
Data collection in CSV format
Data pre-processing & Data Transformation
Test/train split
Training Models
Logistic regression with Hyper-parameter tuning
KNN with Hyper-parameter tuning
Random Forest classifier with Hyper-parameter tuning
Artificial Neural Network
Evaluation &
Results
Evaluation
Result Analysis

Figure 1: KDD process for charged-off loan prediction

3.1 Dataset

Our study uses 2260701 loan application of LendingClub from the year 2007 to 2018. The data is sourced from LendingClub [5]. The target variable could be identified as 'loan_status' which is classified as 'Fully Paid' or 'Charged-Off'. As per the statistics of the dataset, it has been observed that almost 80% of loans are fully paid, and 19% of loans are charged-off, which indicates a potential risk for the investors. The dataset is having 151 features of each loan. The dataset contains the information about the borrowers such as income, employment length, debt-to-income ratio, FICO score etc and loan product details like loan purpose, loan amount, interest rate, loan grade, instalment etc. This study limited only those loans which are either fully paid or charged-off and does not considers loans that are current, default, does not meet the credit policy. In the second step, to limit the feature space, the study identifies potential features which could be useful for the investors to make their decision in a loan approval. The definition of features has been given in LendingClub data dictionary and could be identified at LendingClub online portal. The data dictionary is having the definition of each feature which was recorded by LendingClub.

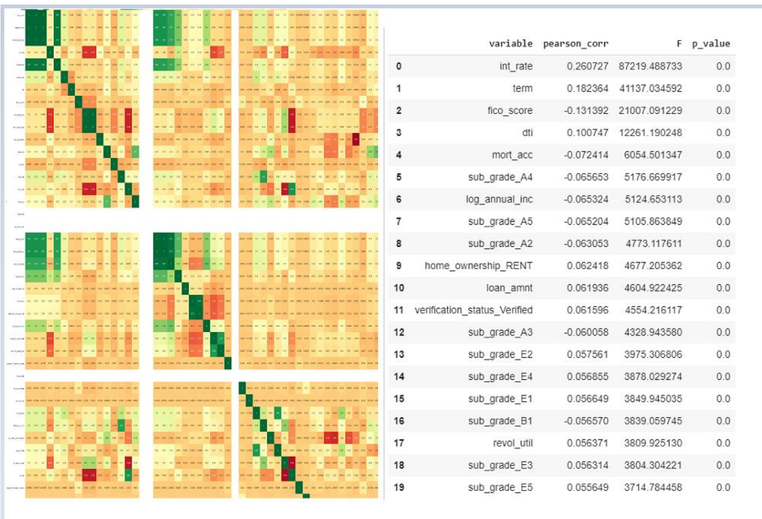
In the next section, the study will discuss the potential feature of borrowers and loan products which could be helpful for the investors in identifying the true borrowers.

3.2 Feature Selection

This study uses the correlation matrix and the Pearson correlation coefficient to identify the potential features for charged-off loans predictions. Correlation matrix and Pearson correlation statistics help to identify the correlation between the features and hence only those features have been taken under consideration which is closely correlated and could affect the predictor variable i.e. loan status. As per the Figure 2, There are many features which are not important for investors such as id, policy_code etc. as these all features doesn't play an important role in loan approval decisions and so we can drop these type of features. Besides, features such as instalment, loan amount, interest rate etc. are identified as closely correlated and could

[5] https://www.lendingclub.com/statistics/additional-statistics?

participate in loan approval decisions for investors, hence these features taken under consideration to predict the charged-off loans.



	variable	pearson_corr	F	p_value
0	int_rate	0.260727	87219.488733	0.0
1	term	0.182364	41137.034592	0.0
2	fico_score	-0.131392	21007.091229	0.0
3	dti	0.100747	12261.190248	0.0
4	mort_acc	-0.072414	6054.501347	0.0
5	sub_grade_A4	-0.065653	5176.669917	0.0
6	log_annual_inc	-0.065324	5124.653113	0.0
7	sub_grade_A5	-0.065204	5105.863849	0.0
8	sub_grade_A2	-0.063053	4773.117611	0.0
9	home_ownership_RENT	0.062418	4677.205362	0.0
10	loan_amnt	0.061936	4604.922425	0.0
11	verification_status_Verified	0.061596	4554.216117	0.0
12	sub_grade_A3	-0.060058	4328.943580	0.0
13	sub_grade_E2	0.057561	3975.306806	0.0
14	sub_grade_E4	0.056855	3878.029274	0.0
15	sub_grade_E1	0.056649	3849.945035	0.0
16	sub_grade_B1	-0.056570	3839.059745	0.0
17	revol_util	0.056371	3809.925130	0.0
18	sub_grade_E3	0.056314	3804.304221	0.0
19	sub_grade_E5	0.055649	3714.784458	0.0

Figure 2: Correlation matrix and Pearson correlation table with F-statistics for feature selection

As per Everett (2015) literature review, we find that investors tend to give loans in locals where they know borrowers, Hence this study keeps borrowers geographical details (addr_state and zip_code) under consideration in terms to predict charged-off loans. Finally, the study takes subsets of 31 features for exploratory data analysis.

Next, the Study will explore the data insights and share the bivariate visualization for outcome variable and independent variables.

3.3 Data Pre-processing & Data transformation

In this section, the study will explore important feature individually and drop the feature in case it is not useful. Here, the study will explore the statistics summary, data visualization and plots against the dependent variable. We will also do modify the data if necessary.

On visualizing loan status against the loan amount, the study finds that higher loan amount tends to have high chances to become charged-off loans (see Figure 3).

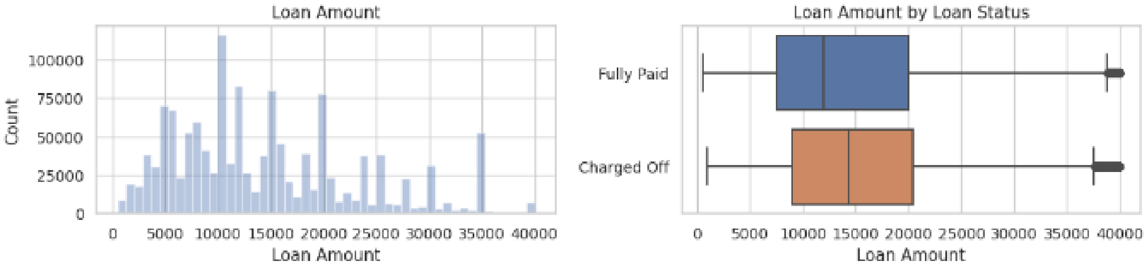


Figure 3: loan status and loan amount

While on visualizing the loan status with interest rates, it has been found that higher interest rate loans are having a high chance to become charged-off loans (see Figure 4).

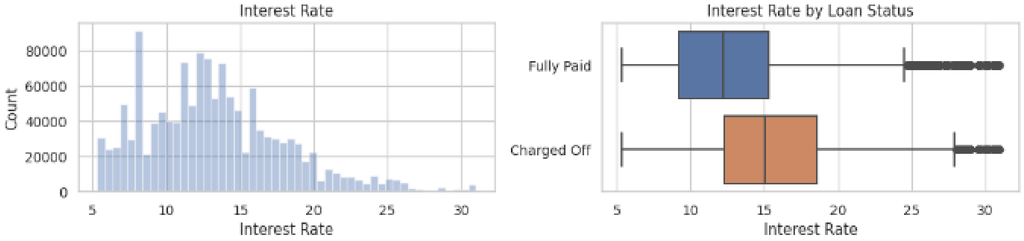


Figure 4: loan status and interest rates

On exploring the insights for loan status and instalments it has been found that higher loan instalments tend to become the loan as charged-off (see Figure 5).

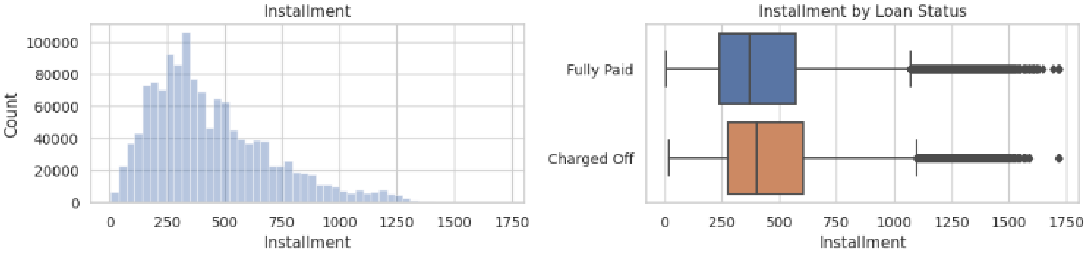


Figure 5: loan status and instalments

Study implies the log transformation due to the high gap in between the min and max income with a median of 65000 dollars (see Figure 6).

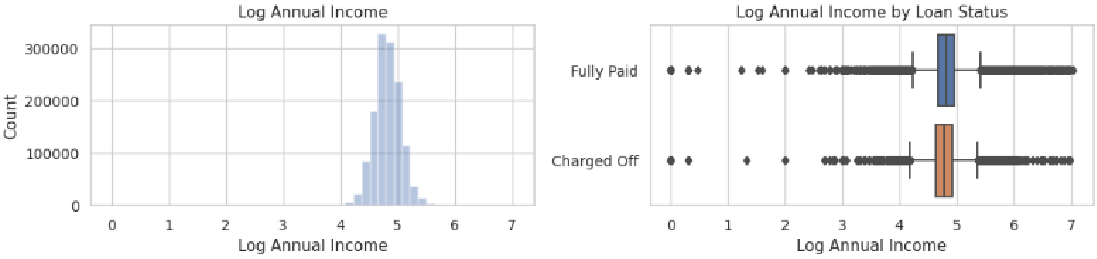


Figure 6: loan status and instalments

As the study is much focused on features that are available to investors before the loan was funded. So, issue_d will be excluded from the model building. But the study will keep this feature for test/train split. Next, on analysing the 'earliest_cr_line' feature, it has been found that shorter credit line borrowers loan is having a high chance to become charged-off (see Figure 7).

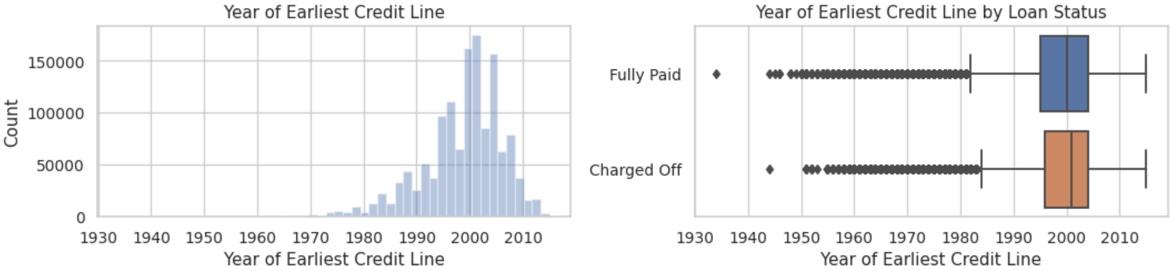


Figure 7: loan status and earliest credit line

Next, instead of two features fico_range_low and fico_range_high, the study takes the average of both and name it fico_score. On analysing the fico_score features, it has been found that charged-off loans tend to have 10 points lower than the average fico score (see Figure 8).

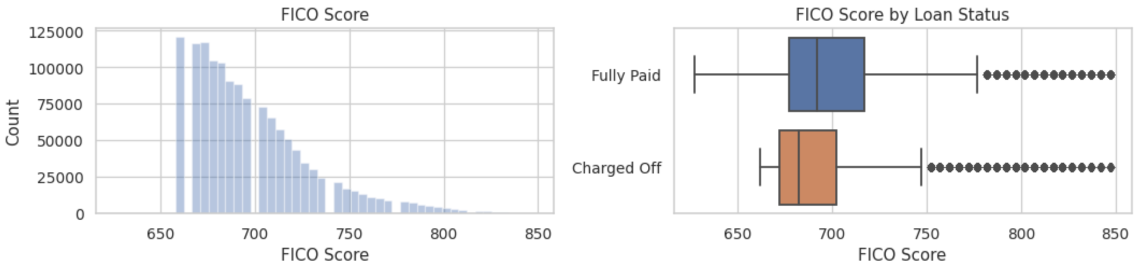


Figure 8: loan status and fico score

In summary, this section explained data insights via univariate and bivariate data visualization, removed non-significant features, handled outliers via log transformation of high dimensional data and does the feature engineering for few of borrowers features wherever necessary. Once the pre-processing parts are completed, the study does split the data into test and train set.

Next, the study will explain the data modelling part to build a model to predict the charged-off loans.

3.4 Data Modelling

As per the notation in James et. al. (2013), this study applies the mathematical model on loan status prediction problem as below:

Let the binary dependent variable is denoted as Y, where –

$$Y = \begin{cases} 0 & If\ Fully\ Paid \\ 1 & If\ Charged-off \end{cases}$$

Our mathematical model is presented such as –

$$Y\ = f(X) + \in$$

Here, X= $(X_{1,}\, X_2, \ldots X_p)$ is a feature vector and $\in$ is denoted as irreproducible error i.e. measurement errors and noise in data. Additionally, even if a function is known there is still a chance of errors in prediction as at each X=x there may be a distribution of Y values. Hence, the function could be defined as-

$$f(x) := E(Y|X = x) + \in$$

Here, right-hand-side (RHS) represents expected Y values for a realization of x for feature vector X. f(x) denotes the probability of charged-off borrowers and that can be described as below-

$$f(x) := E(Y|X = x) = \Pr\,(Y = 1|X = x)$$

Thus, the P2P lending problem is to identify a good prediction of f(x) that could minimize the reducible error. This study considers logistic regression (LR), random forest classifier (RFC), k-nearest neighbour (KNN) and artificial neural network (ANN) to minimize the error and to predict the charged-off loans based on the selected features.

3.4.1 Logistic Regression

A linear regression estimator function can be defined as a linear combination of individual attributes such as below:

$$f_{LR}(x) := \sum_{i=0}^{p} \beta_i\, x_i$$

The useful feature of logistic regression is that the output values in this model lie between 0 and 1 and helpful in class-conditional probability for classification problem.

This study uses SGD classifier estimator in the scikit-learn library to implements the linear classifier such as logistic regression with stochastic gradient descent (SGD) training. The study chose a linear classifier through the loss hyperparameter. As this study develops logistic regression model so loss parameter has been set as 'log'. SGD is an optimization technique and an efficient approach to fit the linear classifier under convex loss function such as logistic regression. In a further development of the logistic regression model, this study implements machine learning pipeline to automate the machine learning flow. Machine learning pipeline enables a sequence of data to be transformed and correlated with each other in a model which helps to test and evaluation of model on the positive or negative outcome. This study trains the logistic regression model via machine learning pipeline. This pipeline runs iteratively as every step try to improve the accuracy in terms to achieve a good model. Applying pipeline techniques makes the model scalable. Then finally, the study applies the GridSearchCV to train the final model on the whole dataset. GridSearchCV hyperparameter techniques help to identify the best hyperparameter and check the mean cross-validated AUROC score for the logistic regression model.

3.4.2 Random Forest Classifier

Random forest model uses a collection of decision tree to split the data node for training and test data. The splitting is done using the GINI index. For an individual attribute split x_i, which denotes levels as $L_1, L_2, \dots \dots . L_J$, Gini index for this attribute can be calculated as below:

$$G(x_i) \coloneqq \sum_{j=1}^{J} \Pr(x_i = L_j)\left(1 - \Pr(x_i = L_j)\right) = 1 - \sum_{j=1}^{J} \Pr\,(x_i = L_j)^2$$

Once the Gini index is calculated for every candidate's split attribute, the split is made based on the highest Gini indices. Random forest model gives powerful techniques to build a forest of decision tree randomly. It helps in reducing the variance when accounted on tree averages (Breiman, 2001).

Like, logistic regression model, this study develops random forest classifier using machine learning pipeline and GridSearchCV. The study checks how well the GridSearchCV could help to identify the best hyperparameter for the random forest model and check the mean cross-validated AUROC score for random forest classifier model.

3.4.3 K-Nearest Neighbour (KNN)

KNN is widely used to solve a classification problem. This algorithm takes input as training instances and test instance with respect to distance function, typically define k value between the range of 1 to 10. The classification is based on the nearest neighbour majority vote that is helpful to define the weight in terms of neighbour contribution. The weight is defined in a manner where closer neighbour does participates more in prediction decision. The function of KNN can be understood as below:

$$f_{knn}(x) \coloneqq Majority(Y|X \in \mu_k(X))$$

Here, Majority refers the majority vote function and μ_k denotes the closest neighbour X with respect of Euclidean distance for p dimensional space.

This study applying the KNN model using linear discriminant analysis (LDA) to reduce the number of variables 10 or lesser so the model could perform well. Like logistic regression and random forest classifier model, this study develops machine learning pipeline with

GridSearchCV hyperparameter to tune the model and check the mean cross-validated AUROC score for KNN model.

3.4.4 Artificial Neural Network (ANN)

ANN is based on Artificial Intelligence techniques which enable the human brain to intimate computer and perform tasks like pattern recognition, evaluation, prediction, and classification. ANN consists of a large number of neurons which are structured in a layer like Input, output, and hidden layers. ANN model can perform a massive computational unit parallelly those are interconnected via weighted connections. Every computation unit is known as a neuron that is having a set of connection. Each neuron receives an input signal from other neurons that are the set of the transfer function and weighed inputs.

Along with the classification model such as logistic regression, random forest and k-nearest neighbour, this study applies an Artificial Neural Network (ANN) too. This is used to see, does the neural network performs well over the supervised machine learning models. To reduce the computational time under-sampling has been performed on training data. Later, we do scale the data and create a validation test using sklearn libraries. This study applies 7 layers to train the neural network using keras libraries and uses sigmoid activation function. As the dataset is highly imbalanced, hence weights have been adjusted for classes. ANN model has been run over the 50 epochs and uses activation function as 'relu'. This approach has never been used by any other researchers on LendingClub dataset.

4 Design and Implementation Specification

This section shares details of project architecture along with different phases of development and implementation specifications.

4.1 Architecture Design

This section describes the architecture of this research study. It is necessary to understand the flow of the project. Maximum steps of architecture design have been well explained in section 3. Section 4 will explain the evaluation and results of the study. Figure 9 explains the graphical representation of project architecture.

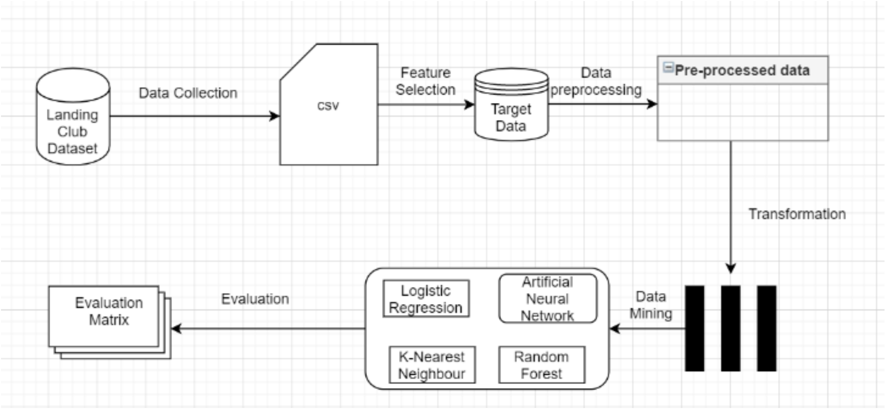


Figure 9: Architecture for P2P charged-off loan prediction

4.2 Implementation Specification

This study identifies 'loan status' as the dependent variable and other remaining variables as independent variables. Study keeps only those transaction whose loan status is either 'charged-off' or 'fully paid'. Rest all loan status such as 'current', 'late' etc. are out of scope for this study. In data pre-processing and data transformation (Section 3.3), the study presents the insights of the dependent variable and independent variables and identifies the outliers. Along with plots, the study also compares the statistical summary between dependent variables and independent variables. As per learning from data analysis, study drops the features which are accounted via other features. To handle the outliers study applies the log transformation for the features like 'annual_inc', 'revol_bal' etc. Furthermore, the study creates a dummy variable for loan status such as 0 for 'fully paid' and 1 for 'charged-off'. Once data transformation and data pre-processing completed, the study does the test/train split and train the Logistic Regression (LR), Random Forest Classifier (RFC), K-Nearest Neighbour (KNN) and Artificial Neural Network (ANN).

This study applies the classification models along with hyperparameter tuning. The study applies the gridSearchCV techniques with 5-cross validation and finds the best hyperparameter so model performance could be improved. Once all models are applied with hyperparameter then based on the AUROC score, we identify LR model overperformed among the RFC and KNN. Then we do tune the LR model with the best hyperparameter and checks the model accuracy but in this case, the model accuracy was similar as we found earlier without tune the model.

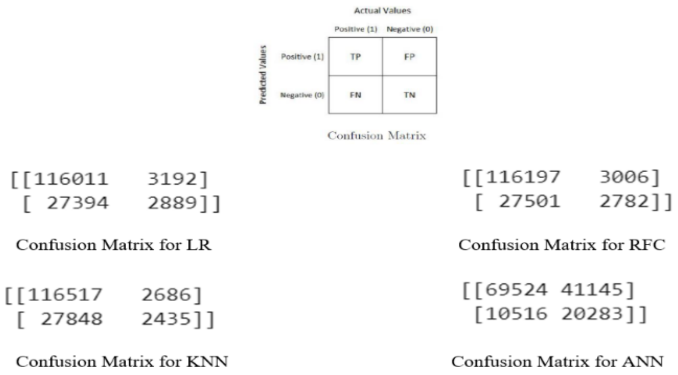
The last model applies using the Keras library. The deep learning models x and y both are used for testing and training data. the x variables consist of all variables but not the dependent variable. All variables are scaled too before the data mining process. Initially, a sequential model is initialized with 'relu' activation function. Dense layers are also used to avoid the overfitting for the model. In the last layers, we used 'sigmoid' activation function. Next, the study applies RMSprop gradient-based optimizer and binary cross-entropy as a loss to compile the model. Then the model is fitted using the epochs and batch size. Then prediction for testing data is performed once an evaluation is done for the model. As the dataset is highly imbalanced, hence we adjust the class weight of the outcome variable. In the next section, we will evaluate the model based on different matrices.

5 Evaluation

This section will describe brief details about the model evaluation process for the study. This study majorly calculates the accuracy, classification report and Area Under the Curve (AUC). Classification report will help to provide the statistics such as F1 Score, recall and precision for the applied models.

5.1 Confusion Matrix

Confusion matrix gives a summary of prediction results for a classification problem. Misclassification of false negative could be a major loss for online P2P lending operations or financial organisation.



Confusion Matrix

```
[[116011   3192]
 [ 27394   2889]]
```

Confusion Matrix for LR

```
[[116197   3006]
 [ 27501   2782]]
```

Confusion Matrix for RFC

```
[[116517   2686]
 [ 27848   2435]]
```

Confusion Matrix for KNN

```
[[69524 41145]
 [10516 20283]]
```

Confusion Matrix for ANN

Figure 10: Confusion Matrix

Here,
TP (True Positive): These are the numbers of good borrowers who were classified as a good borrower by model.
TN (True Negative): these are the numbers of bad borrowers who were classified as a bad borrower by model.
FP (False Positive): These are the numbers of bad borrowers who were classified as a good borrower by model.
FN (False Negative): These are the numbers of good borrowers who were classified as a bad borrower by model.

5.1.1 Accuracy

Accuracy can be measure by calculating the sample ratio which is rightly classified and total no of samples as per the test dataset. Accuracy can be defined as per below formula:

$$\text{Accuracy} = \frac{(TP+TN)}{TP+TN+FP+FN}$$

5.1.2 Precision and Recall

Precision can calculate by division of true positive with true positive and false positive. Mathematically, precision calculation may understand as per below formula:

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

A recall is the fraction of all positive 'charged-off' instances that classifier identifies correctly as true. This is also known as true positive rate.

5.1.3 F1 Score

F1 score gives a balance average of precision and recall. This is also named as balanced F1-score. The computation for F1 score is as below-

$$\text{F1} = 2 * \frac{precision * recall}{precison + recall}$$

5.1.4 Mathew Correlation Coefficient (MCC)

MCC plays an important role to identify the correlation between true class and predicted class. MCC values lie between -1 and +1. For perfect classifier say 1, means FP=FN=0. MCC can be computed as below:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The results of the study are shown in Table 2. The models were evaluated via 5-fold CV for the historical data of worlds' largest online P2P banking system namely LendingClub. Overall, the study observes that LR, RFC and KNN give test accuracy nearly similar i.e. 79% approximate. And besides the precision, recall and F1 measures are also having a very small gap in terms of comparison between respective results for classifier models (LR, RFC, KNN). While it is way hard to observe the best performed model based on the test accuracy, precision, recall or F1 measures as there is very minimal difference in resultant data. Hence study observes the cross-validation accuracy in terms to observe the model performance. As shown in Table 2, surely LR has outperformed when compared to RFC, KNN models. Here, the LR model has recorded the mean validation accuracy as 71% while RFC and KNN have recorded 69% and 70% respectively. It has been observed that ANN model performance was significantly low when compared in terms of accuracy but by seeing overall measures of classification report, the study finds that ANN model performance is well among applied classification models.

The low score of precision for classifier models is due to the large amount of TP and TN score as per the confusion matrix. The gap between the FP and FN values are not too large hence the precision score for classifier models are also nearly similar. Besides on analysing the ANN model, the FP values is a large enough as 41145 which causes a significantly low score of precision for the ANN model. While recall values are very poor for the classifier models when compared to the ANN model. The poor recall score caused due to a large amount of TP values for the classifier models. As TP score for ANN model is quite high hence the recall score for ANN model is way better when compared to classifier models. The F1 score presents the balance between the precision and recall for the models. F1 score for the classifier models is way poor when compared to ANN models as precision and recall are not well balanced for the classifiers. Which also explains that classifier models are not much competitive when compared to the ANN model. Also, the MCC score for the classifier model is poor than the ANN model score. It means that predicted class is poorly correlated with a true class for classifier models but comparatively better correlated when talks about the ANN model.

Table 2: Performance comparison of classifiers

Classifier	Test Accuracy	Precision	Recall	F1 Score	AUC	MCC	Validation Accuracy
Logistic Regression with SGD Training	0.7953	0.475	0.095	0.159	0.534	0.139	0.7114
Random Forest Classifier	0.7959	0.481	0.092	0.154	0.533	0.138	0.6943
KNN with LDA	0.7957	0.475	0.08	0.138	0.529	0.138	0.7029
Artificial Neural Network	0.6466	0.333	0.619	0.433	0.692	0.238	-

5.1.5 Area Under the Curve (AUC) and ROC

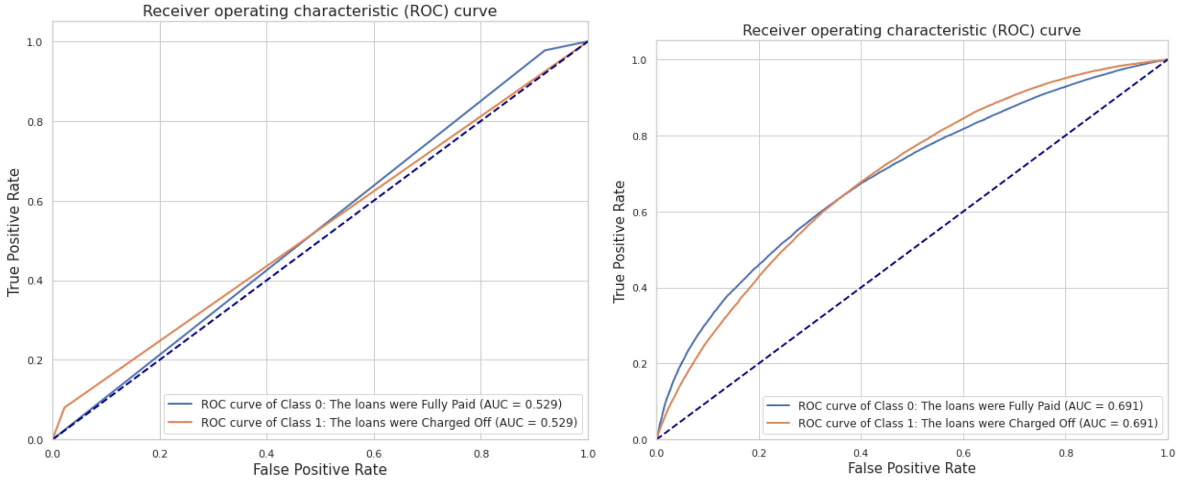
AUC is the percentage of ROC (Receiver Operating Characteristic) plot that is underneath the curve. Assume the ROC is formed using sequential points such as (X_1, Y_1), (X_2, Y_2), (X_3, Y_3), $\ldots(X_m, Y_m)\}$, then the AUC computation would be presented as:

$$\text{AUC} = ½ \sum_{i=1}^{m-1} (X_{i+1} - X_i) * (Y_{i+1} - Y_i)$$

AUC is having the range of between 0 to 1. The higher AUC values denote the better performance of the model. The performance of binary classifier can be evaluated using ROC technique. It develops a 2-D plot and uses false positive rate (FPR) and true positive rate (TPR). ROC computation can be defined as below:

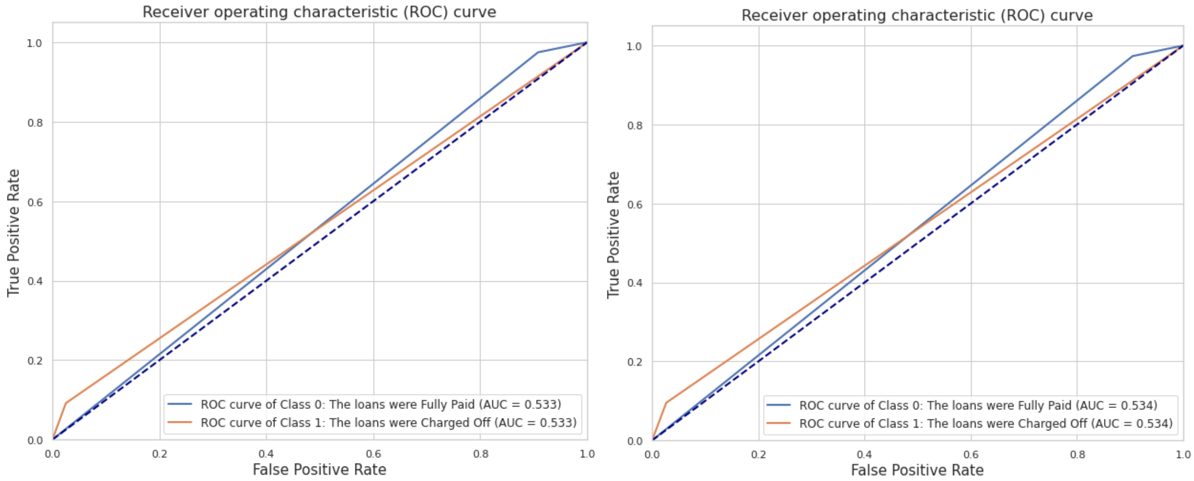
$$\text{FPR} = \frac{FP}{FP+TN} \quad ; \text{TPR} = \frac{TP}{TP+FN}$$

Figure 11 and Figure 12 represents the ROC curves and AUC score for all applied classifiers along with the ANN model. While, if we compare the ROC-AUC curve, it has been found that ANN model gives a good AUC score of 69% when compared to classifier models LR (0.534), KNN (0.529) and RFC (0.533)



(1) (2)

Figure 11: ROC-AUC curve for KNN (1) and ANN (2) model



(1) (2)

Figure 12: ROC-AUC curve for LR (1) and RFC (2)

As per the above ROC curve, we find that the ANN model performs well to discriminate the classes. Means, ANN predictions are way better by calculating the TPR and FPR. While the classifier models give a poor ROC curve as classifier models predict the majority of classes under all thresholds points of the curve. Which means classifier models such as LR, RFC and KNN have not a good skill in prediction of classes. Due to poor skill of prediction of classes the classifier models LR, KNN and RFC getting poor AUC score when compared to the ANN model.

6 Discussion

This section discusses the performance of each model as per the different evaluation measures. This will give the insights of best and worst performed model. In the event of classifier model application such as logistic regression (LR), random forest classifier (RFC) and k-nearest neighbour (KNN), we find that LR model performed well when compared with mean validation accuracy. LR model gets the highest validation accuracy as 71% while the lowest validation accuracy was obtained by RFC i.e. 69%. However, there are poor recall and F1 score for all classifier models (LR, RFC, KNN) that is due to a high amount of true positive values and poor balance between recall and precision scores for the models. The MCC score was also poor for all classifier models i.e. nearly 13% that is because predicted class are poorly correlated with true class. While in terms of accuracy all classifiers models give an almost similar score that is 79%. So, it is hard to say that what classifier model performed well, based on the accuracy score. Besides the poor MCC and F1 score also does not infer the best performed model. So, the study calculates the AUC and ROC score for all classifier models and finds that the AUC score is also not as significant good to decide which model performed well in the prediction of charged-off loans. The AUC score for all three classifiers is nearly 53% that is not a good score as the majority of predictions are made near threshold points of curve.

While on the other side, ANN performance was quite good in terms of ROC and AUC score. The AUC score for the ANN model was 69% as most of the predicted classes are above the threshold point of the curve. While MCC, recall and F1 score was also good when compared to LR, RFC and KNN models. However, precision score i.e. 33% was not good when compared with LR, RFC and KNN models. The less precision score was due to a high volume of false positive values for the classification matrices.

Most relevant study to our work is Byanjankar et. al (2015) research, they researched on European based online peer to peer lending operations namely Bondora[6]. They applied LR and ANN on 16037 loan applications. They found logistic regression model accuracy as 64.5% and ANN model accuracy as 63.6%. Our study finds LR model accuracy as 79.53% and ANN model accuracy is nearly like their study but better than their results i.e. 64.66%. While the related research has been applied to a smaller amount of data with 16037 loan observations only, while our study observed 2260701 loan transactions. The study way varied than the Byanjankar et. al (2015) research work, as our study explored a large amount of data and explored multiple classification models (LR, RFC and KNN) with hyperparameter tuning. Also, our study explored the ANN model with a smaller number of layers than Byanjankar et. al (2015) applied a neural network.

Although logistic regression model gets a good accuracy score than all other classifier model and neural network but also gets the poor score in terms of precision, F1 score, recall and AUC. While ANN model gives a fair accuracy, recall, F1 score, MCC and AUC score when compared with LR, KNN and RFC models. And a model cannot be good based on just accuracy score while other classification scores are not significant good. So, our study says that ANN model performs well than other classifier models as this model is more skilled in the prediction of 0 or 1 due to a good score of AUC. Further, KNN took the maximum time followed by random forest and logistic regression model. Logistic regression model took less time and performed better than the RFC and KNN models. Misclassification of False Negative (FN) could cause a big loss for a financial institution. Our study finds that LR models give less amount of FN than other models but due to low ROC-AUC score seems the model has less skill in prediction of class.

7 Conclusion and Future work

To function an online P2P banking operation healthily it is necessary to identify true worthiness of a potential loan borrower. This study is aimed to predict a charged-off loan using the borrowers and loan product attributes. To predict charged-off loans, this study develops logistic

[6] https://www.bondora.fi/

regression (LR), random forest classifier (RFC), k-nearest neighbour (KNN), and artificial neural network (ANN). Instead of implementing simple LR, RFC and KNN, this study develops all these models using hyperparameter tuning with the help of GridSearchCV algorithm.

This study infers that the artificial neural network model performs well when compared with classifier models such as logistic regression, random forest classifier and k-nearest neighbour. Our study does not choose the best model based on the accuracy score only but best performed model was evaluated based on the overall performance of models in terms of AUC, ROC, F1-score, recall and Mathew co-relation co-efficient etc. This study uses Pearson correlation between the predictors and response variables and finds that loan term, debt-to-income ratio, interest rate and fico score play an important role in charged-off loans prediction for the P2P online banking system.

In future work scope, we can extend this study by doing more data pre-processing to improve the model performance. There are tons of borrowers and loan features which yet to analyse and could play an important role in terms to predict charged-off loans. The limitation of time consumption during model train could be reduced by applying to randomize hyperparameter tuning. Researchers may implement the other model with hyperparameter tuning to identify the improved model accuracy for the online financial P2P lending operations.

References

G. Akerlof, “The market for lemons: quality uncertainty and the market mechanism,” Quarterly Journal of Economics, vol. 84, 1970, pp. 488- 500, doi: 10.2307/1879431.

J. Stiglitz and A. Weiss, “Credit rationing in markets with imperfect information,” American Economic Review, vol. 7, 1981, pp. 393-419.

M.Lin, N.Prabhala and S.Viswanathan, “Judging borrowers by the company they keep: friendship networks and information asymmetry in online P2P lending,” Management Science, vol. 59, 2013, pp. 17-35, doi:10.1287/mnsc.1120.1560.

R. Iyer, A. Khwaja, E. Luttmer et al. “Screening in new credit markets: can individual lenders infer borrow creditworthiness in P2P lending?” NBER Working Paper No. 15242, NBER, Cambridge, MA, 2009.

S. Moro, P. Cortez, P. Rita, “A data-driven approach to predict the success of bank telemarketing,” Decision Support Systems, vol. 62, 2014, pp.22-31.

D L. Olson, D. Delen, Y. Meng, “Comparative analysis of data mining methods for bankruptcy prediction,” Decision Support Systems, vol. 52, 2012, pp.464-473.

Klafft, M., 2008. Peer to peer lending: auctioning microcredits over the internet [C]. Technology and Management, A. Agarwal, R. Khurana, eds, IMT, Dubai.

Iyer, Rajkamal, Khwaja, Asim Ijaz, Luttmer, Erzo F.P., Shue, Kelly, 2010. Screening in New Credit Markets: Can Individual Lenders Infer Borrower Creditworthiness in Peerto-Peer Lending? AFA 2011 Denver Meetings Paper.<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1570115 >.

Freedman, S., Jin, G.Z., 2011. Learning by Doing with Asymmetric Information: Evidence from Prosper. com NBER Working, Paper No. 16855.

Everett, C.R., 2015. Group membership, relationship banking and loan default risk: the case of online social lending. Appl. Econ. 47 (1), 54–70.

Li, Yuelei, Guo, Yang, Zhang, Wei, 2013. Analysis on factors influencing the success rate of P2P small loan market in China [J]. J. Financial Res. (07), 126–138.

Ge, R., Feng, J., Gu, B., Zhang, P., 2017. Predicting and deterring default with social media information in P2P lending. J. Manage. Inf. Syst. 34, 401–424.

Bitvai, Z., Cohn, T., 2015. Predicting P2P loan rates using Bayesian non-linear regression. In: AAAI Conference on Artificial Intelligence. pp. 2203–2209.

Lin, X., Li, X., Zheng, Z., 2017b. Evaluating borrower's default risk in peer-topeer lending: Evidence from a lending platform in China. Appl. Econ. 49 (35),3538–3545.

Guo, Y., Zhou, W., Luo, C., Liu, C., Xiong, H., 2016. Instance-based credit risk assessment for investment decisions in P2P lending. European J. Oper. Res. 249 (2), 417–426.

Serrano-Cinca, C., Gutiérrez-Nieto, B., 2016. The use of profit scoring as an alternative to credit scoring systems in Peer-to-Peer (P2P) lending. Decis. Support Syst. 89, 113–122.

Xia, Y., Liu, C., Liu, N., 2017. Cost-sensitive boosted tree for loan evaluation in P2P lending. Electron. Commer. Res. Appl. 24, 30–49.

Huo, Y., Chen, H., Chen, J., 2017. Research on personal credit assessment based on neural network-logistic regression combination model. Open J. Bus. Manag. 5, 244.

Fu, Y., 2017. Combination of random forests and neural networks in social lending. J. Financ. Risk Manag. 4 (6), 418–426.

Jiang, C., Wang, Z., Wang, R., Ding, Y., 2017b. Loan default prediction by combining soft information extracted from descriptive text in online P2P lending. Ann. Oper. Res.1–19.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., Niu, X., 2018. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electron. Commer. Res. Appl. 31, 24–39.

O L. Mangasarian, "Linear and nonlinear separation of patterns by linear programming," Operations research, vol. 13, 1965, pp. 444-452.

Y. Shi, Y. Peng, W. Xu, et al. "Data minging via multiple criteria linear programming: applications in credit card portfolio management," International Journal of Information Technology & Decision Making, vol. 1, 2002, pp. 131-151.

K. Tam, M. Kiang, "Managerial Applications of the Neural Networks: The Case of Bank Failure Predictions," Management Science, vol. 38, 1992, pp. 416-430.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. NY: Springer.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

A. Byanjankar, M. Heikkilä and J. Mezei, "Predicting Credit Risk in P2P Lending: A Neural Network Approach," 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, 2015, pp. 719-725, doi: 10.1109/SSCI.2015.109.

Emekter, R.; Tu, Y.; Jirasakuldech, B.; Lu, M. Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. Appl. Econ. 2015, 47, 54–70.