Machine Learning on Credit Risk for Lending Club Business

Xianfeng Cui [1, †], Yuting Huang[2, †], Sizhe Lai[3,*,†], Chi Iek Keong[4, †], Di Wu[5, †]

[1]School of International Trade, Shanxi University of Finance and Economics, Shanxi, China

[2]Huntsman School of Business, Utah State University, Beijing, China

[3]School of Physical Sciences, University of California, Irvine, Irvine, the U.S

[4]Faculty of Business Administration, University of Macau, Macau, China

[5] College of Letters and Science, University of California, Santa Barbara, the U.S.

*Corresponding author: sizhel1@uci.edu

[†]These authors contributed equally

Abstract: Under the big data era, human beings attach much importance on machine learning. The purpose of this paper is to identify the way to use different methods of machine learning for data analysis and data mining when the companies in lending industry is faced with credit risk. To be specific, the methods are judged from the indicators (e.g., accuracy score and AUC score), and utilized to recognize which loans are bad loans, so as to facilitate the company to make better decisions. The methods including Random Forest, Gaussian Naive Bayes, and Artificial Neural Network are discussed. According to the analysis, all models have an accuracy around 60-70%, while each show different tendency in classifying results. Further optimization that can be applied in the future studies is suggested in the paper. The overall value and reputation of a company will be improved with good credit risk management. Therefore, a good method of credit risk management (e.g., accurately identifying good loans and bad loans) is classifying and analyzing the existing data of the company via different algorithms, and eventually compare them. These results shed light on guiding further exploration of evaluating the creditworthiness in the lending field.

Keywords: Credit Risk, Feature Engineering, Machine Learning, Binary Classification.

1. Introduction

Credit risk refers to the lender, securities investor or both buyer and seller uncapable or unwilling to fulfill the contract conditions for different reasons, which leads to a breach of contract, and the losses if financial institutions, investors and counterparties. It is generally considered as the risk of loss caused by the borrower's breach of contract or the decline in his or her credit. On this basis, it makes the borrower unable to fully recover the interest due. For most banks, loans are the largest and most obvious source of credit risk [1]. Credit risk is also one of the main risks that banks often face, which is embodied in the risk that the other party of the transaction cannot fulfill the contract as agreed. This risk not only exists in loan business, but also in the acceptance, guarantee and securities investment businesses.

With a rating mechanism, lenders (e.g., banks, or the lending club, which is a P2P lending company in this paper) can keep the credit risk exposure within the acceptable parameter range while maximizing the rate of risk adjusted return and achieve the goal of managing credit risk [2]. Generally, there are 5 types of risks, which are default risk, concentration risk, country risk. institutional risk, and downgrade risk. To be specific, default risk refers to the risk that the bond issuer cannot pay the interest and principal on time; concentration risk means that the value of the financial institution may be considered to be about to lose when the exposures move in a negative direction; country risk refers to the possibility that foreign governments or foreign countries' default on account of the political turmoil or economic recession; institutional risk denotes for the risk of regulators not being capable of achieving their organizational goals; downgrade risk refers to the decline of the issuer's reputation.

To figure out the way to reduce or avoid credit risk, the relationship between credit risk and bad loan should be recognized. The causes of bad loans and loan defaults can be enumerated by collecting

and analyzing empirical data [3]. However, banks sell the non-performing loans at significant discounts [4]. It is one of the ways lenders deal with non-performing loans (NPLs). Another method is that the lender can hire a collection institution to forcibly recover the loans that are defaulted, in exchange for a certain amount of recovered amount.

ML is most often used in anti-money laundering and fraud-detection applications [5]. The ability of ML model is to reveal the subtle relationship between different data, deal with unstructured data, and catch different nonlinearities. For instance, a predefined structure will not be needed for a fraud detection analysis. Random forest is a classification algorithm in machine learning, and decision tree is a classifier. A decision tree is constructed through the training set, so that the classification of new data can be predicted. The random forest algorithm can be seen as a classifier consists of many decision trees. Multiple sub data sets are constituted from the original data set through put back sampling, and then each sub data set is used to construct a decision tree. The effect of classification will be determined by the mode predicted by multiple decision trees [6].

Gaussian Naive Bayes classifies the data by using the method of statistics and probability. its misjudgment rate is very low since its mathematical foundation. Naive Bayes is a learning algorithm with greater bias, but lower variance, than Logistic Regression [7]. This algorithm is relatively simple, and is commonly used for text classification, and is more friendly to missing data. It deals with multi classification problems and is also good at incremental training. At the same time, it also has relatively stable classification efficiency.

Smote was proposed by Chawla in 2002 to deal with the issue of unbalanced data. KNN technology was used for simulation process of SMOTE. SMOTE analyzes and simulates small scales of category samples, adding the manually simulated new samples to data set, which fixes the problem of categories in original data set being unbalanced. However, SMOTE algorithm also has its disadvantage such as over generalization of the minority class space [8]. This means that if majority samples surround the selected minority samples, this may be "noise", i.e., classification difficulty will occur because the newly synthesized samples overlap with surrounding majority samples.

With the increasing popularity of credit card use, there are more and more credit card related fraud activities all over the world, which undoubtedly caused huge losses to financial institutions. However, according to Tomi Himberg's paper, machine learning has become an important tool in following regulations and enabling an agile business environment [9]. Today's form of credit risk management and supervision reduces the risks of bank regulators and financial institutions, ensuring the normal operation of the financial market and avoiding losses. This is because machine learning technology can provide accurate and efficient credit risk analysis and decision-making. As mentioned above, credit risk has a great impact on the lending businesses. Not only banks, but many companies in the lending industry are also trying to find out how to carry out risk control, so as to improve productivity and performance, and improve the value and reputation of the company. Machine learning, as an idealized tool of data analysis and data mining, should be discussed and studied. Therefore, this paper will take the data of lending club as the basis, and discuss it combined with the algorithm of machine learning. The rest part of the paper is organized as follows. The second part will be introducing the basic information of the notebook data, and the implementation steps of the algorithms as well, the results analysis and limitations will be discussed in the third part, and conclusion of the paper will be elaborated in the forth part.

2. Methodology

2.2 Data

The project uses the lending club dataset from Kaggle that contains 396030 loans from 2007 to 2016. The dataset has 27 variables which are either numerical or categorical (including binary variable “loan status” represents the bad and good loans). Figure 1 and Table 1 show the distribution of good loans versus bad loans and the stats of some important numerical features in the dataset.

Table. 1. Statistics of Important Numerical Features

	loan amount	interest rate	annual income	debt to income
Mean	14113.89	13.64	74203.18	17.38
Standard Deviation	8357.44	4.47	61637.62	18.02
Minimum	500.00	5.32	0.00	0.00
25%	8000.00	10.49	45000.00	11.28
50%	12000.00	13.33	64000.00	16.91
75%	20000.00	16.49	90000.00	22.98
Maximum	40000.00	30.99	8706582.00	9999.00

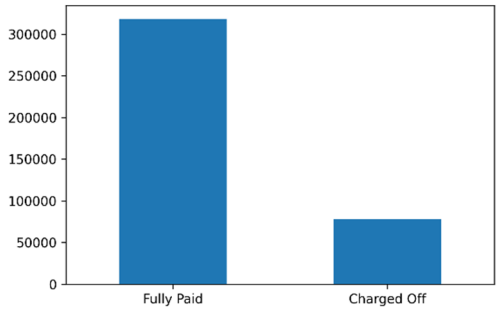

Figure. 1. Good Loan V.S. Bad Loan Distribution

Figure. 2. Heat Map of Original Numerical Dataset

Among the 26 variables (exclude the "loan status"), some variables are highly correlated or functionless for the study. We perform feature engineering on the dataset before building the classification models. To decide the features to keep, we first draw a heat map in Figure 2 to analyze the correlation coefficient among the features. One noticed that among all numerical features, "loan amount" and "installment", "total accounts" and "open accounts", "public bankruptcies record" and "public records" are highly correlated to each other (with scores larger than 0.65), so we look through the definitions of these variables. As defined above, these variables are corresponding to each other which will affect weighting process in some model training process, we only keep one in each pair for feature engineering. The installment is dropped since it is based on the loan amount, while loan

amount can reflect the changes synchronously. As presented in Fig. 2, “loan amount” and “installment” scores 0.95 which is enormously high, it is promising to drop out one of the variables. The “total accounts” is kept since it shows the status of the borrower’s credit file, where the status is preferred in the study. In the public record pair, we deleted the bankruptcies variable since it is included in the derogatory public record and the models need as much information as possible.

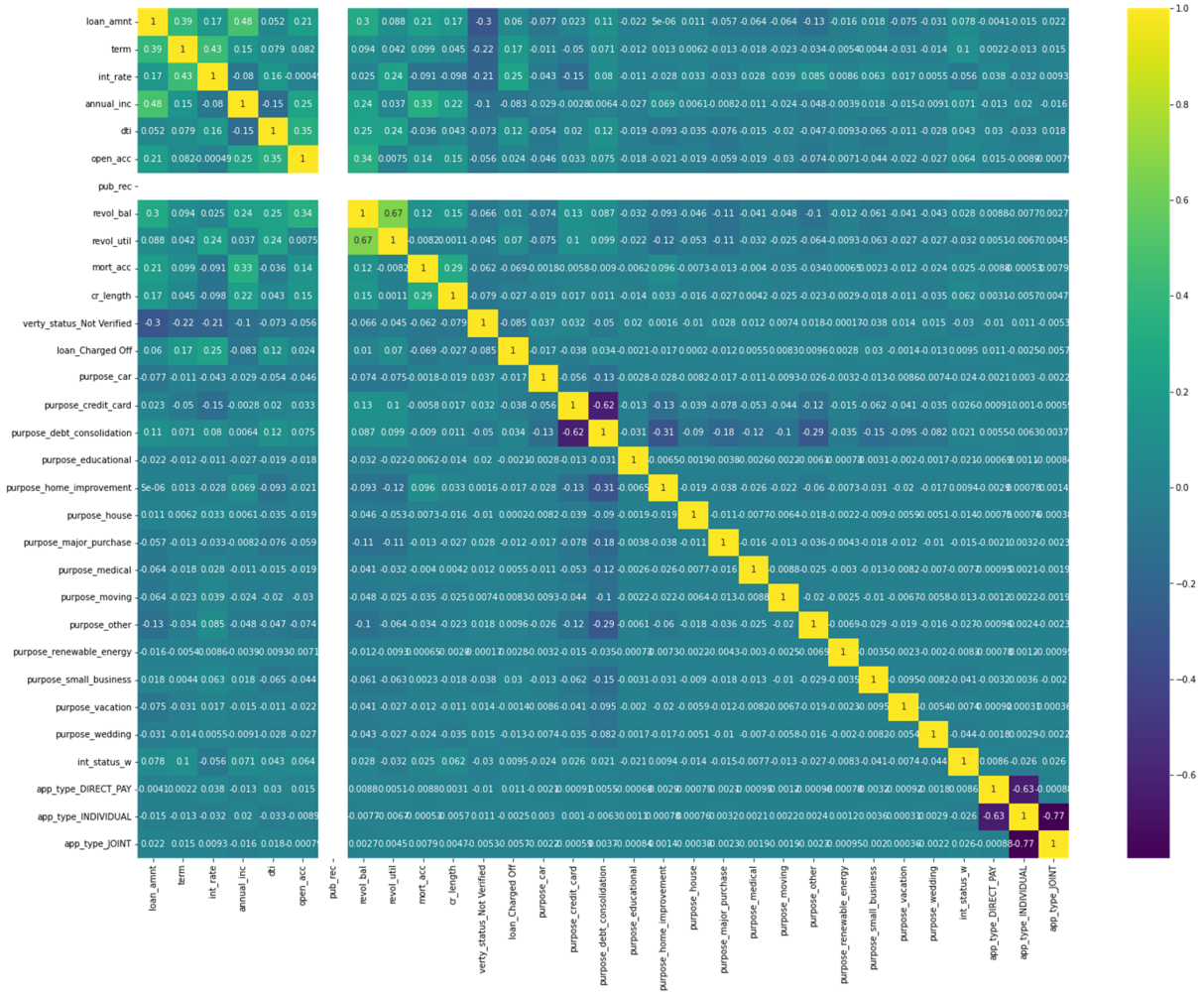

Figure. 3.Heatmap of the Training Set

The notebook tries to distinguish the geographic variables based on the address variables. However, some addresses from the dataset are not formally written and some addresses like USCGC (which stands for the United States Coast Guard Cutter) all make the geographical labeling hard to perform, so the notebook drops the address variable in feature engineering. Apart from the loan status, the dataset also has loan grade and loan subgrade calculated by Lending Club. The grade and subgrade have a linear relationship with interest rate. Importantly, the grading measurement is unknown but possibly imply Lending Club’s risk management algorithm, the notebook drops the two grading categories. Furthermore, the notebook notices the mortgage account has null values, and it is based on the home ownership feature, where a borrower only has mortgage account if one owns a real estate. Therefore, the notebook handles the mortgage account by setting all missing values as zero for the “rent” and “other” home ownership borrowers, setting the rest missing mortgage account by rounded the average mortgage account to integers. With the new mortgage account, the notebook deletes home ownership feature since it is shown in the numerical data mortgage account.

The notebook also creates a new variable call credit length, which measures the length of the borrowers’ credit history by subtracting the loan issued year from the earliest loan application year. With all the numerical variables setting up, the notebook plots the distribution of the variables. Most variables have a wide range but uniform distribution: either left skew or right skew. The notebook combines the distribution and the quartile statistics and decided to clip some numerical data in 75

percent to remove the outliers. Later, based on the distribution shape, the notebook either applies logarithm or z-scale to the numerical data as normalization in feature engineering.

The last process is to one hot encoding the remaining categorical features, so that all the remaining variables are numerical. By feature engineering, the final variables for training and the correlation coefficient are shown in the heat map (seen from Figure 3). To split the imbalance training and testing dataset, the notebook uses two methods: stratified shuffle split and synthetic minority oversampling technique (SMOTE). Stratified shuffle split splits dataset to different folds and applies cross validation so that the train and test set follows the same ratio as the origin dataset. Two commonly used methods are change the size of the origin dataset: either undersamples the majority set or oversamples the minority set. Mohammed et al. used the resampling technique to do a comparison between oversampling and undersampling, the results show that most of the oversampling did better than undersampling and oversampling also gained higher scores in different evaluation metrics [11]. Therefore, the SMOTE is applied to analyze the bad loan samples and manually synthesize them to enlarge the data set by the newly synthesized bad loan samples.

2.2 Models & Algorithms

Machine learning, which can be categorize into two methods, i.e., supervised and unsupervised learning methods. Surprisingly, the supervised methods can turn to unsupervised methods in some ways. As Shi and Horvath written in their paper, one can create a manual label which can distinguish “observed " data (original and unlabeled) and "synthetic" data (extracted from a reference distribution) [10]. A random forest is a classifier with multiple decision trees, and the output class is determined by the plurality of the output classes of the individual trees. It is composed of many decision trees, and there is no association between different decision trees. Breiman and Cutler are the first to propose this model, and they proposed to distinguish the observed data and the synthetic data by using the Random Forest (RF).

Gaussian Naive Bayesian algorithm is originated from classical mathematical theory with a solid foundation in mathematics. It makes an assumption that all features have a Gaussian distribution (normal/bell-shaped curve).

Artificial neural network mimics human brain. The network is formulated by nodes and layers, each with different processors and weights. Dongare et al. draws a comparison between biological terminology and ANN terminology [12]. The layer processors are activation functions like rectified linear units (ReLU), sigmoid or hyperbolic tangent functions (tanh), which outputs numerical data. Among the mentioned activation functions, sigmoid outputs number in the range of 0 to 1, tanh outputs from -1 to 1, and ReLU outputs from 0 to infinity. Note the activation functions can be combined with different layers, each layer applies by identical weights calculated through training process. Since the final output layer is numerical, it is required to set thresholds in classifier problem. ANN can apply dropouts as optimization to avoid overfitting, which randomly drops out layers or data when training.

3. Results & Discussion

3.1 Analysis procedure

This section will introduce and analyze the results of models from Section 2, then discuss the limitation and potential improvement based on the results. The notebook uses accuracy score, f beta score and roc curve to measure the model performance. Frist of all, four commonly used quantities in binary classification models are introduced: true positive, true negative, false positive and false negative. As the names shown, these quantities are the counts on the four only possible outcomes. For example, true positive is the number of cases where the model correctly classifies a positive output, while the false positive counts the cases when model claims positive incorrectly. Noticing the four quantities represent all possible solution, hence the sum of four quantities is equivalent to the total test set size.

$$\text{Accuracy Score} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

where TP, TN, FP, FN are true positive, true negative, false positive, false negative, respectively. It shows the measurement of accuracy score. Since the denominator is the test set size and the numerator is the total count of correct predictions, accuracy score directly calculates the overall correct rate of the model. Note that accuracy score cannot show the distribution of positive and negative results of the prediction, which can be meaningless in some imbalance data set. For example, the good loan/ bad loan rate of the lending club data set is approximately 8:2. Using stratified shuffle introduced in Section 2, the test set follows the distribution of 8:2. If we have a model that always predict good loans regardless of the input variable, the accuracy score will always be 80, the percentage of good loans in the test set. Although 80 is a reasonable accuracy score, this model is functionless. Therefore, this notebook calculates the accuracy score to show the overall model performance but focuses on the other scoring method when analyze the results.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

In the previous example, the model is not capable to classify bad loans, which is useless for lending club because the default loans bring loss to the business. Therefore, instead of measuring accuracy score, the model should pay more attention on the bad loan prediction. Based on the four quantities introduced previously, two further calculations are developed: precision (function 2) and recall (function 3, also known as sensitivity). Precision is the correct rate among the positive prediction and recall is the correct rate among the real positive values. Function 4 shows how f beta score measures the imbalance data model based on the actual target. By setting the weight beta in calculating f beta, it represents that recall is beta times important as precision when determine the model performance. By setting the weight beta in calculating f beta, it represents that recall is beta times important as precision when determine the model performance. Since the bad loans harm the business more, the notebook set beta equal to two to avoid misclassifying good loans.

$$f_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \quad (4)$$

The notebook also conducts another measuring method—Receiver Operating Characteristic (ROC curve). ROC curve works on binary classification problems. Eq. (5) shows false positive rate which calculus the incorrect positive label among the total negative cases while function 3.2 precision measures the true positive rate correspondingly. In a ROC curve, the x-coordinate is the false positive rate and the y-coordinate is the true positive rate. Fawcett explains graphing a ROC curve by setting up different threshold [13]. Furthermore, he also introduces calculating the area under curve (auc score) as a quantity of measuring the model performance. In general, a curve that close to the upper domain stands for a good classification, with a large auc score that is closed to 1. Note that an auc score of 0.5 means a random choice classifier and model should not expect an auc score under 0.5.

$$\text{False Positive Rate} = \frac{\text{FP}}{\text{FP+TN}} \quad (5)$$

3.2 Empirical analysis

With the three performance methods introduced above, we are presenting the output of all three methods conducted in the notebook, including both with oversampling and without oversampling. In addition, the result also contains an accuracy score conducts on the training set to detect if the model is overfitting.

As summarized in Table. 2, random forest has an around 70 percent accuracy score and the auc score is approximately 60 percent, which is slightly better than a random guess. Oversampling increases the F2 score so it has a better performance on classifying bad loans. However, the accuracy score on the training set is both relatively high, which means random forest is overfitting. Recall the introduction of random forest from section 2, oversampling increases the distribution of the training

set hence change the nodes when selecting good loans and bad loans. Since the test set is imbalance, random forest is more likely to choose good loans instead of bad loans, which explains the lower accuracy score but higher F2 score after overfitting.

Table. 2. Model Performance Scores

Models	Random Forest	Gaussian Naïve Bayes	Artificial Neural Network
Before Oversampling			
Accuracy Score	0.7494	0.7484	0.6160
F2 Score	0.2333	0.3514	0.5484
AUC Score	0.5875	0.6658	0.7034
Accuracy Score on Training Set	0.9494	0.7488	0.6183
After Oversampling			
Accuracy Score	0.7144	0.2986	0.6433
F2 Score	0.3191	0.5488	0.5303
AUC Score	0.6145	0.6641	0.6582
Accuracy Score on Training Set	0.9817	0.5410	0.7001

Gaussian Naïve Bayes has an accuracy score around 70 percent and an AUC score around 60 percent which means it has a slightly better performance than the random guess. The accuracy score on training set is reasonable hence not overfitting. Additionally, oversampling increases the F2 score, shows that the model develops better bad loans classification. Nevertheless, the accuracy score drops to about 30 percent and the AUC score has little difference after oversampling, which means it has lower performance labeling the results. Since the bad loans only take a small percentage in the data set, the tendency of labeling more bad loans increases the true negative values but also the false negative values. Although the aim is to predict possible bad loans rather than good loans, a model with overall accuracy scores less than 30 percent is not tolerated.

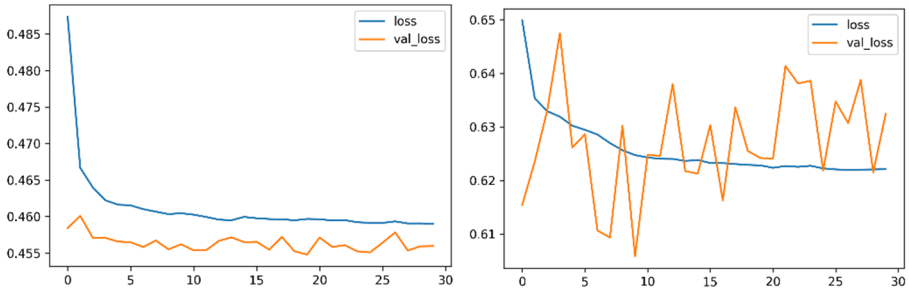

Figure. 4. Loss and Validation Loss (Left: without Oversampling; Right: Oversampling).

Since the ANN applies sigmoid function in the output layer, the result ranges in 0 and 1. A threshold value is required to perform the classification. Notice the notebook utilizes stratified shuffling to split the train and test set we choose 0.2, the bad loan rate in the train set as the threshold. The overall accuracy score is 63 percent, which is only slightly better than a random guess, but ANN has the highest f2 score of .54. Comparing the accuracy score and f2 score applying on the training set, the overall performance is consistent, so the model is not overfitting. Figure 4 plots the loss and validation loss. The model converges quickly in the first few iterations hence the training is done in a first few iterations, and the validation loss is always lower than the loss. According to Ref. [14], the possible reasons are the regulation (dropouts) applied during training, the delay of validation loss measurement and a simple chosen validation set. After oversampling, the notebook changes the threshold to 0.5 since the new training ratio after oversampling is 0.5. Note that the final bad loan and good loan prediction maintains the same ratio before oversampling because the model is applied in the same test set, which means the threshold should be set on the ratio of the training set but not the

test set. The accuracy score and f2 score has little variation after oversampling. In fact, the f2 score decreases. However, figure 4 shows that after oversampling, both loss and validation loss increase to over 0.6 and the validation loss diverges. Therefore, oversampling is not necessary for ANN, and ANN has over 60 percent accuracy on both good loans and bad loans. Given the imbalance distribution of good loans and bad loans, the ANN model has the best performance in labeling good loans among all the models in the notebook.

3.3 Limitations

In general, since the data set is imbalanced, all three models tend to label good loans over bad loans. For instance, random forest and Gaussian Naïve Bayes both have a high accuracy score around 74 percent, but both have a less than 0.5 F2 score, representing the models are over labeling the good loans. Oversampling increases F2 score of random forest, which means the imbalance affects the random forest performance. The future study should take closer look at the imbalance and the over/under sampling process when applying random forest algorithm. The feature engineering process is also an impediment for getting a practical result. The notebook handles the outliers by replacing the top 25 percent numerical with the 75 percent quantile. Furthermore, this paper only analyzes standard addresses and does not use different methods to study in one area, nor does it use different methods to study the same area and different areas in a period, so the research on the two variables of time and area is not particularly thorough. The accuracy rate of classifying bad loans is less than 50 percent. There are three possible reasons. First, the oversampling method might not simulate the characteristic of bad loans. Second, the bad loans are not classified well when splitting the training set, and the overall characteristic of bad loans are not displayed. Third, compared with good loans, bad loans themselves do not have obvious characteristics. Bad loans may appear in the good loans with similar characteristics. For example, an A1 loan taker runs into accident and eventually defaulted the loans, while normally this kind of loan takers are not likely to have bad loans. In such case, the models are unlikely to classify the difference between bad loans and good loans.

Besides, the model chosen in the paper should be revised in future study. As introduced in Section 2.2, random forest only makes predictions that are the average of previously observed labels. If any time period of the training data is missing, then depending on the trend, the random forest model will under- or over-predict examples outside the time range in the training data. This will be very apparent if the model's predictions are plotted against their true values. In the notebook, random forest on the training set shows that the model is overfitting.

Further conclusion can draw from comparing the Kaggle notebook [15] from SAYAH, where he drops the training data with outliers and performs ANN, XG Boost and random forest on the same data set, his accuracy score and f2 score are both better than the notebook, meaning that the capping 75 percent quantile does not exclude the outliers of the dataset. Notice that SAYAH's random forest has a 100% accuracy score on the training set, meaning that the random forest on this notebook is also overfitting. The comparison suggests that in the future study, a closer look at the data outliers should be handled carefully to improve the study. Additionally, SAYAH splits the dataset randomly in the notebook without considering the data imbalance. The model performance in [15] suggests the original dataset has outliers or noise that affect the prediction. After removing the outliers, the data shows good tendency that help classifying the final results.

In summary, all three models show poor performance on loan prediction. Although the feature engineering does not remove the outliers that influence the prediction, all three models shows potential on classifying. Random forest trains fast and oversampling improves its classifying ability, but hyper parameters should be carefully selected to avoid overfitting. Gaussian Naïve Bayes has great training speed and better recall rate, it also requires no oversampling methods. ANN trains slowly, but converges rapidly in a few iterations. It has the best classification performance and does not require oversampling. With drop outs implement, it avoids overfitting. ANN also has the best AUC score so it is preferred in the future study.

4. Conclusions

In summary, this paper investigates machine learning based on credit risk evaluation. Specifically, we did a background investigation to determine the credit risk assessment primarily. Subsequently, the 2007-2016 data is used to conduct empirical research based on the models of Random Forest, where Gaussian Naive Bayes, Z score, Logarithm, Straitified Shufflesplit and One Hot Encoding methods are studied. According to the analysis, the best results appeal on random forests. In the future, in the one hot encoding with more dimensions, one can try to use PVC to reduce the dimensions, and in terms of dealing with data deviations, the overall data is available to be determined as much as possible to reduce the two kinds of deviations. These results offer a guideline for theoretical guidance for evaluating whether to borrow or not, and this paper provides an analytical tool for evaluating users' creditworthiness for the lending field.

References

[1] Basle Committee on Banking Supervision, Bank for International Settlements. Principles for the management of credit risk[M]. Bank for International Settlements, 2000.

[2] Information on https://www.gdslink.com/what-are-the-different-types-of-credit-risk/

[3] Kwabena A B M. Credit risk management in financial institutions: A case study of Ghana Commercial Bank Limited[J]. Research journal of finance and accounting, 2014, 5(23): 67-85.

[4] Information on: https://corporatefinanceinstitute.com/resources/knowledge/finance/non-performing-loan-npl/

[5] Information on: https://www.spglobal.com/marketintelligence/en/documents/machine_learning_and_credit_risk_modelling_november_2020.pdf

[6] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[7] Permission S. Generative and discriminative classifiers: naive bayes and logistic regression[J]. 2005.

[8] Huang P J. Classification of imbalanced data using synthetic over-sampling techniques[M]. University of California, Los Angeles, 2015.

[9] Himberg T. Loan Default Prediction with Machine Learning[J]. 2021.

[10] Shi T, Horvath S. Unsupervised learning with random forest predictors[J]. Journal of Computational and Graphical Statistics, 2006, 15(1): 118-138.

[11] Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results[C]. 2020 11th international conference on information and communication systems (ICICS). IEEE, 2020: 243-248.

[12] Dongare A D, Kharde R R, Kachare A D. Introduction to artificial neural network[J]. International Journal of Engineering and Innovative Technology (IJEIT), 2012, 2(1): 189-194.

[13] Fawcett T. An introduction to ROC analysis[J]. Pattern recognition letters, 2006, 27(8): 861-874.

[14] Rosebrock, A., 2022. Why is my validation loss lower than my training loss? - PyImageSearch. Information on: https://pyimagesearch.com/2019/10/14/why-is-my-validation-loss-lower-than-my-training-loss/> [Accessed 30 May 2022].

[15] Kaggle.com. 2022. Lending Club Loan Defaulters Prediction. Information on: https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction.