Classification Model Comparison and Improvement Regarding Credit Risk

Simin Yu, Yuan Tan, Yuying Zhang, Kexin Sheng

Abstract

Credit risk measurement is always an important topic in the finance market since the precise prediction of borrowers' defaults can help banks or companies to maximize profits. Our main goal is to compare models' accuracy to better classify their borrowers with different grading systems. This paper access credit risk through fives models: Multiple Logistic Regression, K-Nearest-Neighborhood (KNN), Support Vector Machine (SVM), Decision Tree and Random Forest by using relevant data from *Kaggle*, which overall contain 12 columns in the data with the information about borrowers' background and some traits of their loans. Firstly, we compared the models regarding three main features -- sensitivity, specificity, and accuracy and found out that KNN has the highest sensitivity while random forest has the highest of both specificity and accuracy. Second, we compared the ROC curve and AUC of the five models, which measured models' sensitivity and specificity at the same time. Random Forest, with an AUC of 0.821, was the best one among the five models. Then, we again compared the same two dimensions of the models without PCA. The improved method indicated Logistic Regression, KNN and Random Forest as the top three performers. Finally, we combined these three models using two classic ensemble classifications -- boosting and bagging, and both classifications performed well. Boosting increased the AUC of Logistic Regression to 0.986 and bagging increasing the AUC of Logistic Regression to 0.899.

Overall, we suggest finance companies collect more background information of their borrowers for an all-around understanding. In addition, Random Forest is quantitatively proved to be selected as a credit measurement model.

Key Words: Credit Risk, MLR, KNN, Random Forest

Contents

1. Introduction

Credit risk can be understood as a probability of a borrower to default or to fail to fulfill their contractual obligations. An efficient management of credit risk improves banks' performance (Wójcicka Wójtowicz, 2018). In the capital market, credit risk is one of the most concerned risks because the lenders themselves will have to take losses if default happens. However, lenders are not always able to know which borrower will default. They can only judge the possibility of default by assessing borrowers' personal background and loan information. Today, large financial institutions have accumulated thousands of loan data. It is important to make good use of these data in order to make more precise prediction on borrowers' default. Credit risk measurement has evolved so quickly with changes in loan market and different types of models showing up. In order to distinguish between creditworthy and non-creditworthy clients (or different credit-grade-level clients, such as A/B/C/D/E), credit risk assessment is done through the development of classification models. Such models can not only help these institutions identify customers, but also realize the effective allocation of money.

Credit risk measurement is essential to companies, making analysis models important. Whether the model is good or bad depends on its classifying accuracy. There is not a best model that fit for all data, so it is indispensable to apply various models and techniques to our data and make comparisons. In this report we will analyze the data we found in 2020 of credit risk to test some models that we think are effective to classify consumers according to their background and predict whether they will default or not. We then compare the results to select the most suitable model for predicting credit risk.

2. Literature Review

With the maturity of the loan market, several models have been designed to assist credit risk analysis. Henley and Hand applied the k-nearest-neighbourhood (KNN) method to the credit scoring problem and discussed the selection of optimal values of parameters k and D included in the method (Henley WE and Hand DJ, 1996). Farquad et al. proposed a PCA-SVM model which performs PCA for dimension reduction on dataset and SVM for classification. The PCA-SVM model performed well with a higher accuracy compared to SVM alone and PCA-Logistic Regression model (Farquad MAH et al., 2011). Lappas applied unsupervised machine learning combining expert knowledge with genetic algorithms in feature selection to credit risk assessment (Lappas Pantelis Z. et al., 2021). Ünvan used a quantile regression technique to help prevent credit risk (Ünvan Yüksel Akay, 2019). Some researchers empirically applied an extreme learning machine (ELM) for credit risk problems on the basis of a German credit risk dataset and compared it to naive Bayes, decision tree, and multi-layer perceptron (MLP). The simulation results of statistical measures of performance corroborated that the ELM outperforms other three models (Qasem, Mais Haj and Nemer, Loai, 2020)

Firm performance also have influence on consumer behavior and their default times. Some researchers use non-parametric linear programming methods to evaluate the performance of firms and logistic regression analysis methods to test the "importance of efficiency in predicting business failures". This paper offers non-financial factors in credit risk evaluation, giving a broader explanation of consumer behavior, which can help us give better suggestion to company.

In order to better manage credit risk, commercial banks need to understand the effectiveness of various risk management strategies and use them to minimize credit risk. An explanatory study analyzed the opinions of the employees of selected commercial banks about which strategies are useful for mitigating credit risk by multiple regression. The results identified four areas of impact on credit risk management (CRM): corporate governance, diversification, hedging and, the bank's Capital Adequacy Ratio. They highlighted that these four risk management strategies are critical for commercial banks to resolve their credit risk (Rehman, Zia Ur, Muhammad, Noor, Sarwar, Bilal, 2019).

3. Data Description

Our research is based on simulation credit bureau data from Kaggle.

There are overall 12 columns in the data, containing rich information about borrowers' background and some traits of their loans. We have 32573 data sets in total. We clean the data at first by ignoring the some extreme data sets that beyond the common sense, for example, there are three data set with ages above 140 years old be ignored. Some non-data features (like home ownership and loan intent) have been renamed by categorical data so that they will be more suitable to run in models.

Below is a table of the columns, introducing the meaning of each feature.

Table 1 Description of the Columns

Feature Name	Description	Meanings of Categorical Data
person_age	Age	
person_income	Annual Income	
person_home_ownership	Home ownership	1-RENT, 2-OWN, 3-MORTGAGE, 4-OTHER
person_emp_length	Employment length (in years)	
loan_intent	Loan intent	1-DEBTCONSOLIDATION, 2-EDUCATION, 3-HOMEIMPROVEMENT, 4-MEDICAL, 5-PERSONAL, 6-VENTURE
loan_grade	Loan grade	A/B/C/D/E/F (A is the highest grade)
loan_amnt	Loan amount	

loan_int_rate	interest rate	
loan_status	Loan status	0-non default, 1-default
loan_percent_income	Percent income	
cb_person_default_on_fil	Historical default	0-non default, 1-default
cb_person_cred_hist_length	Credit history length	

More description about the data are included in the Appendix.

4. Model Analysis

We preprocess the data before running the model. First we drop the rows with null values. Second, we use synthetic data generation to deal with the unbalanced data (22,430 non-default data, 6,202 default data), finally we get 14,411 non-default data and 14,221 default data. Third, with ten variables in the dataset, it is necessary to do correlation tests and principal component analysis. The correlation matrix, as shown in Figure 10, shows a strong correlation between loan amount/income percent and loan amount, default history and loan interest rate, besides, there exists certain correlation between income and home ownership, loan amount.

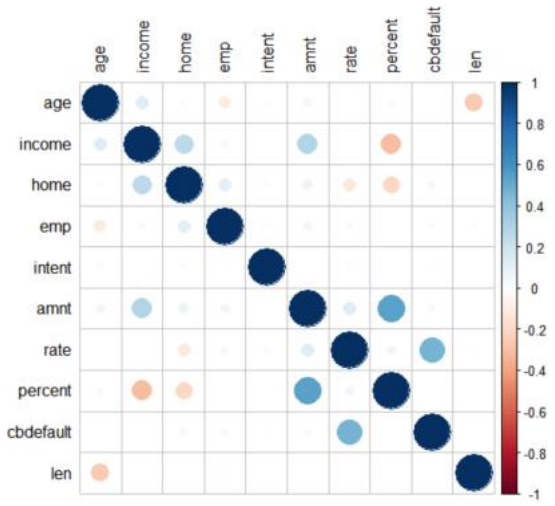

Figure 1 Correlation Matrix

Therefore, in order to avoid collinearity while preserving the data information as much as possible, we use principal component extraction. The explained variance of each principle component is shown in Figure 2, and as seen from Table 2, six principle components can explain the 78% variance, so we extract the six principle components. The composition of the first five principle components is shown in Figure 3.

For example, the first PC is mainly composed of loan/income percent and interest rate, and the second PC is mainly composed of income and loan amount.

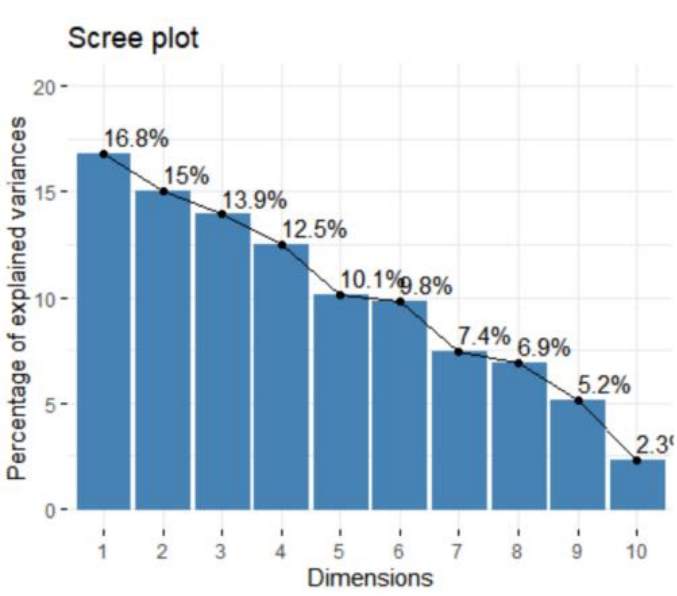

Figure 2 Percentage of explained variance

Figure 3 Composition of PC

Table 2 Information of PC

	Eigenvalue	Variance.percent	Cumulative.variance.percent
Dim 1	1.6817023	16.817023	16.81702
Dim 2	1.5033265	15.033265	31.85029
Dim 3	1.3922075	13.922075	45.77236
Dim 4	1.2489417	12.489417	58.26178
Dim 5	1.0132801	10.132801	68.39458
Dim 6	0.9830058	9.830058	78.22464
Dim 7	0.7427504	7.427504	85.65214
Dim 8	0.6876804	6.876804	92.53895

Dim 9	0.5166009	5.166009	97.69496
Dim 10	0.2305044	2.305044	100.00000

Finally, to avoid overfitting, we divide the datasets into training and test sets in a 4 : 1 scale.

4.1 Multiple Logistic Regression

The probability of default is postulated to be a function of a set of regressor variables $x_1, x_2, \cdots, x_k$. Multiple logistic regression can be explained in the following formula:

$$\ln\left(\frac{p}{1-p}\right) = X\beta = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix},$$

where p is probability, X is the matrix of all the regression variables, β is the vector of all the regression coefficients. That is ,

$$p = \frac{1}{1+e^{-X\beta}}.$$

Although logistic regression is quite simple, it is useful for predicting a certain type of probability and has a wide application. So we first choose logistic regression to predict the default probability based on our dataset. The result is,

$$\ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = -0.04 + 0.73\mathrm{R1} - 0.92\mathrm{R2} + 0.82\mathrm{R3} + 0.05\mathrm{R4} - 0.13\mathrm{R3} - 0.24\mathrm{R4}.$$

The mean square error on the training data is 0.176 and the mean square error on the testing data is 0.173. The ROC curve on testing data is shown in Figure 4, where AUC=0.820, indicating that the model is quite well.

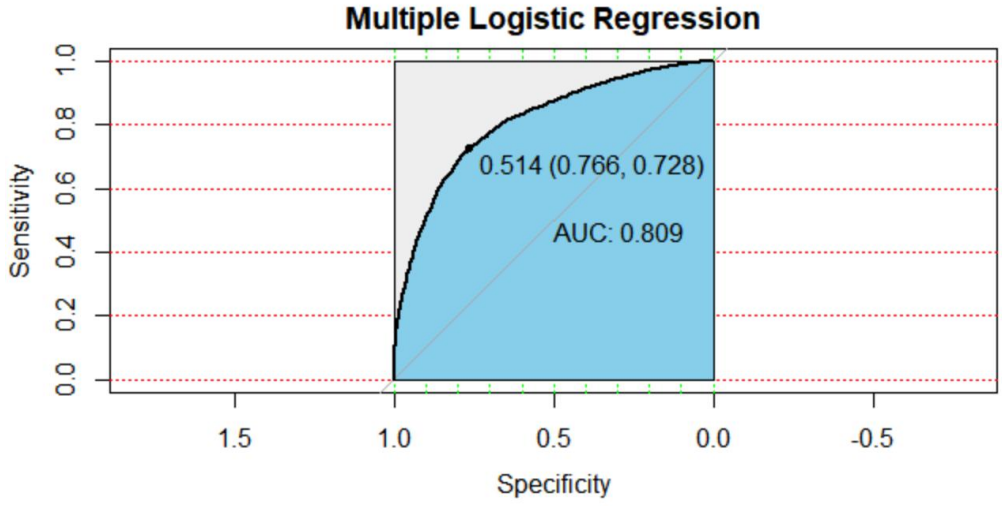

Figure 4 ROC of Multiple Logistic Regression

4.2 K-Nearest-Neighborhood (KNN)

A basic categorization is to label the testing data when its properties correspond to those of some training data whose class has been recorded. But it is almost impossible for every test data to match with a train data. Another problem is that one test sample may fit in with several train data with different classifications. Based on such obstacles, KNN method was originally proposed by Fix and Hodges (1952) and Cover and Hart (1967). In KNN, the distance between objects is calculated as an index of dissimilarity which avoids the matching problem. The distance is generally Euclidean distance or Manhattan distance. Besides, KNN make decisions on the basis of the dominant category of k objects rather than a single sample category decision. These two advantages make KNN perform well in classifications. In our KNN analysis, we will select different values of the parameters k and D to compare the fitting effect.

In the KNN model, if the k value is small, the training error is small (small deviation), and the generalization error will increase (large variance). In other words, small k value means a more complicated model and leads to overfitting. On the other hand, large k value means a simple model and leads to under-fitting. One extreme example is that k is equal to the number of samples and there is no classification at all.

In order to avoid overfitting and under-fitting, our parameters k and D are set as:

$$k = 15$$

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \text{ (Euclidean Distance)}$$

The confusion matrix is shown in the Table 3, the accuracy on the training data is 0.8879 while it is 0.8074 on the testing data:

Table 3 Confusion Matrix of KNN

train	**0**	**1**	**test**	**0**	**1**
0	10383	1157	**0**	2329	497
1	1411	9955	**1**	606	2294

The ROC curve on testing data is shown in Figure 5 where AUC=0.808.

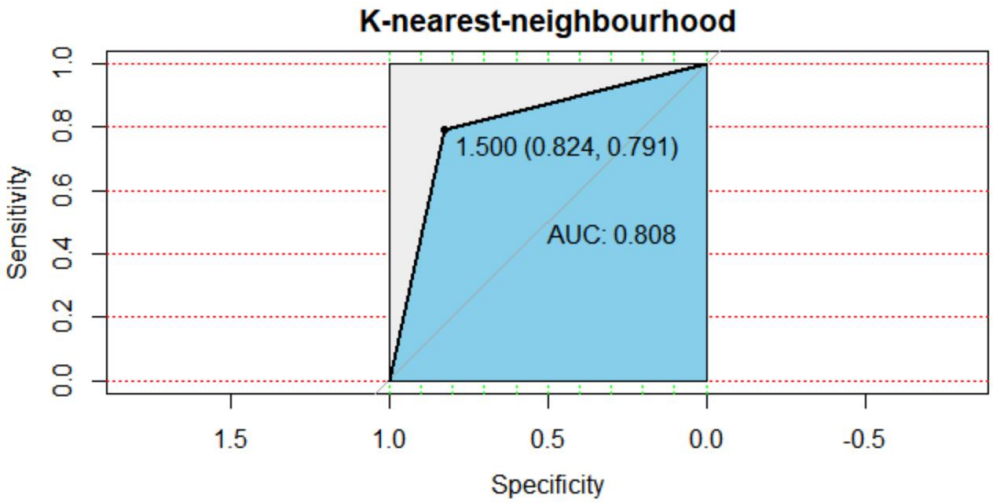

Figure 5 ROC of KNN

4.3 Support Vector Machine (SVM)

Support Vector Machine is a two-category model that maps the feature vector of an instance to some points in the space. The purpose of SVM is to draw a line to distinguish this "best" two types of points. SVM is suitable for small and medium-sized data samples, non-linear, high-dimensional classification problems. SVM was first proposed by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963, and the current version (soft margin) was proposed by Corinna Cortes and Vapnik in 1993 and published in 1995. There are four basic concepts, separating hyperplane, maximal margin hyperplane, soft margin and kernel function. To deal with a binary classification problem, we need a decision boundary, where samples are classified as A on one side and B on the other side._This decision boundary separates the two types of things, and the linear decision boundary is what we call separating hyperplane. Among all separating hyperplanes, we need to find the most suitable one called maximal margin hyperplane which should be kept as far as possible from both sides to leave enough margin, thus reducing generalization error, and ensuring robustness. However, data is not always linearly separable. Therefore, we have soft margin which allows some points in the wrong position and parameter "C" denotes to penalty factor for misclassification. Another solution is to use the kernel function to map the original data to the high-dimensional space. By selecting the appropriate kernel function, the data in the high-dimensional space is linearly separable._However, if the dimension is too high, overfitting problem will occur. As a result, how to select suitable kernel function and soft margin "C" is essential to get a good fit by SVM. Generally, the model will tend to overfit with a more complex kernel function and a larger "C". In order to find the optimal parameters, we use the caret package to construct the model, and use multiple cross validation to evaluate the model, and finally display the relationship between the parameters and the accuracy through graphics.

Here, we use 4 different kernel functions including radial, polynomial, linear and sigmoid function to build the SVM model, to find that the radial function performs far better than others after an initial simple analysis.

To evaluate the performance of the model, we need to see the confusion matrix, which is shown in Table 4. Besides, the accuracy on training data is 0.7811, and the accuracy on the testing data is 0.7745.

Table 4 Confusion Matrix of SVM

train	0	1	test	0	1
0	9275	2265	**0**	2274	739
1	2748	8618	**1**	552	2161

The ROC curve on testing data is shown in Figure 6, where AUC=0.775.

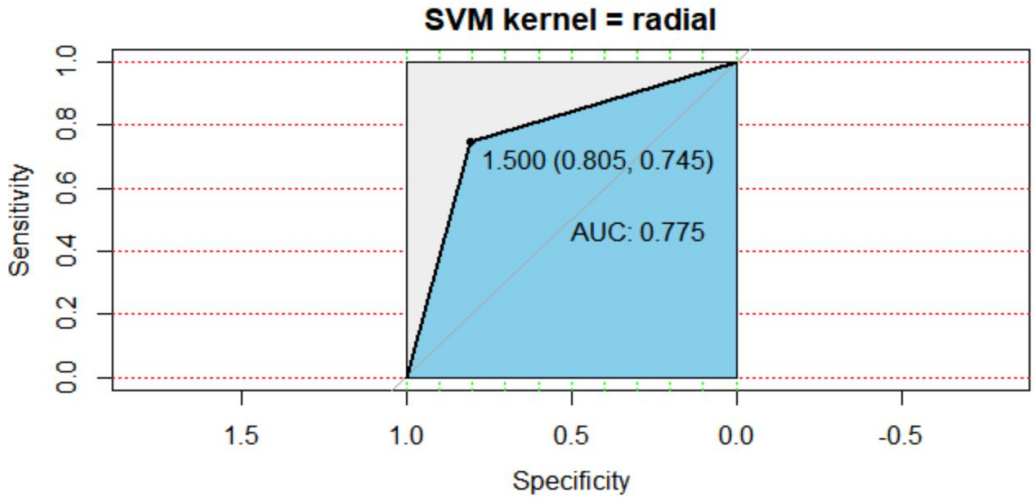

Figure 6 ROC of SVM

4.4 Decision Tree

Decision tree is a model for classification that uses recursive segmentation of nodes using criteria like "gini impurity" or "information gain" to create tree models. Decision trees look like a series of if-else statements arranged in tree form, so it is easy to understand.

In the decision tree algorithm, questions about data feature are asked and then data will be classified according to each answer. Specifically, we raise questions about each node and bifurcate the nodes according to the answer to achieve the purpose of classifying the data. Impurity is used as a criterion to assess the degree of data separation, and when dividing a node data into two sub-nodes, the best questions minimize the impurity of the sub-nodes. The more categories a node contain, the higher the impurity. On the contrary, with only one classification, the impurity is the lowest.

Because decision tree is easy to understand and is well explanatory, we consider using decision trees for classification. To avoid overfitting, we set max depth equals to 5, and the decision tree is shown in Figure 7.

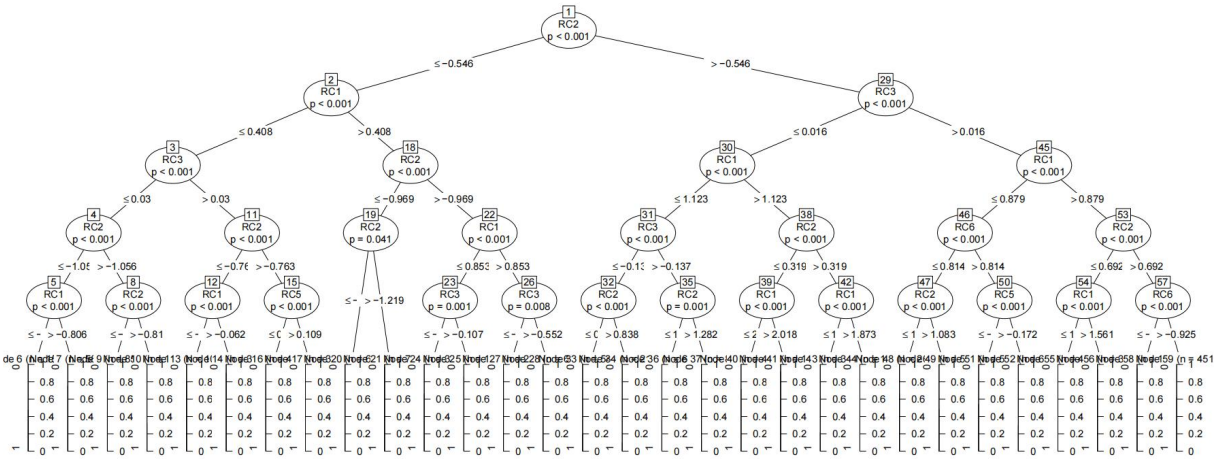

Figure 7 Decision Tree

To evaluate the performance of the model, we need to see the confusion matrix, which is shown in Table 5. Besides, the accuracy on training data is 0.7677, and the accuracy on the testing data is 0.7574.

Table 5 Confusion Matrix of Decision Tree

train	**0**	**1**	**test**	**0**	**1**
0	9202	2949	**0**	2188	751
1	2372	8383	**1**	638	2149

The ROC curve on testing data is shown in Figure 8, where AUC=0.758.

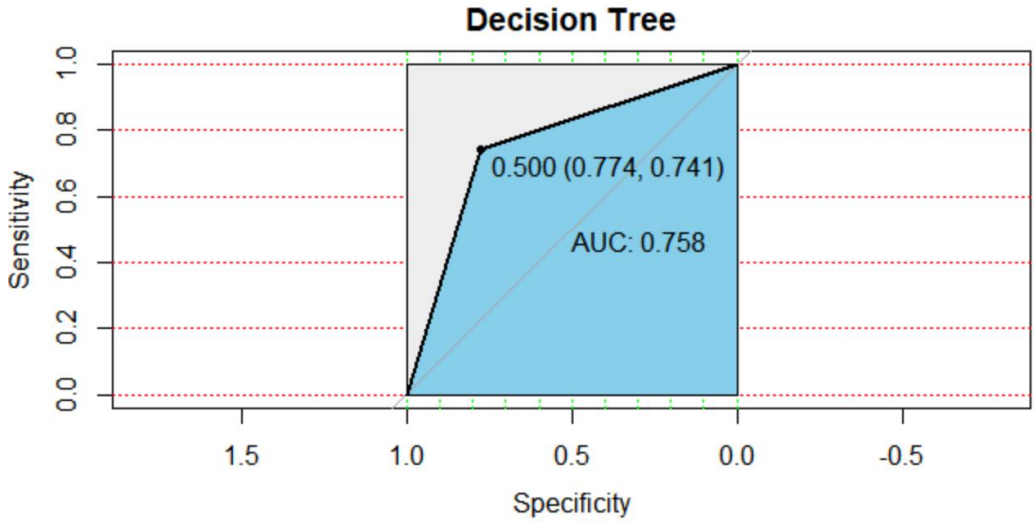

Figure 8 ROC of Decision Tree

4.5 Random Forest

Since decision trees is easy to overfit, and although the model can be simplified by pruning, a single tree is often not as effective as multiple trees, so the idea of random forests arise.

As the name "random forest" suggests, it is composed of many decision trees, and the term "random" is embodied in the following two aspects. First, if the training data is of size N, for each tree, we randomly take N training samples from the training data and put back (called the bootstrap sample method), as the training data of the tree. Therefore, each tree's training set is different and contains duplicate training samples. Second, if the feature dimension of each sample is M, we specify a constant m<<M, randomly select m subsets of features from the M features, and each time the tree divides, select the best from these m features. There is also noteworthy that each tree grows to its best without a pruning process. Finally, the output category of the random forest is the mode of all decision trees' output.

In our model, we use a forest consists of 100 trees (actually, after the test, increasing the number of trees does not improve the accuracy of the model significantly). The importance of the six principle components to the classification is as depicted in Figure 9, showing that RC3 is the most important variable in terms of the decrease of accuracy and RC2 is the most important with respect to the decrease of gini impurity.

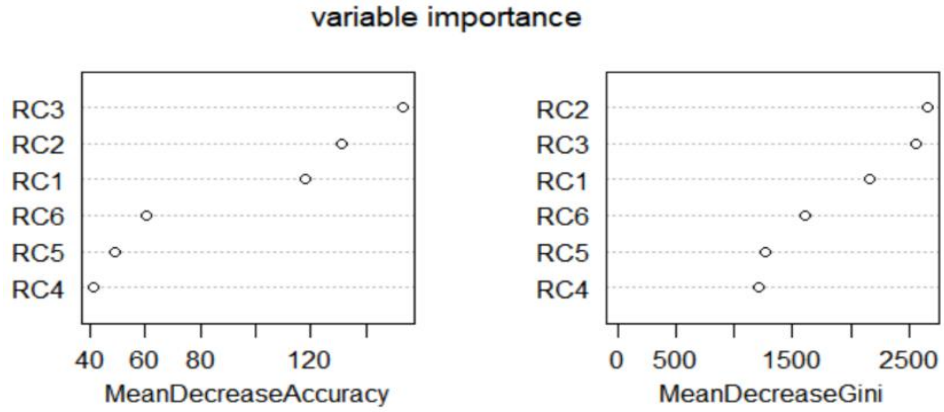

Figure 9 Variable Importance

Similarly, to evaluate the performance of the model, we need to see the confusion matrix, which is shown in Table 6. Besides, the accuracy on training data is 1, and the accuracy on the testing data is 0.8206, improving a lot compared to the decision tree model.

Table 6 Confusion Matrix of Random Forest

train	**0**	**1**	**test**	**0**	**1**
0	11574	0	**0**	2384	585
1	0	11332	**1**	442	2315

The ROC curve on testing data is shown in Figure 10, where AUC=0.821, showing that the model is much better than decision tree.

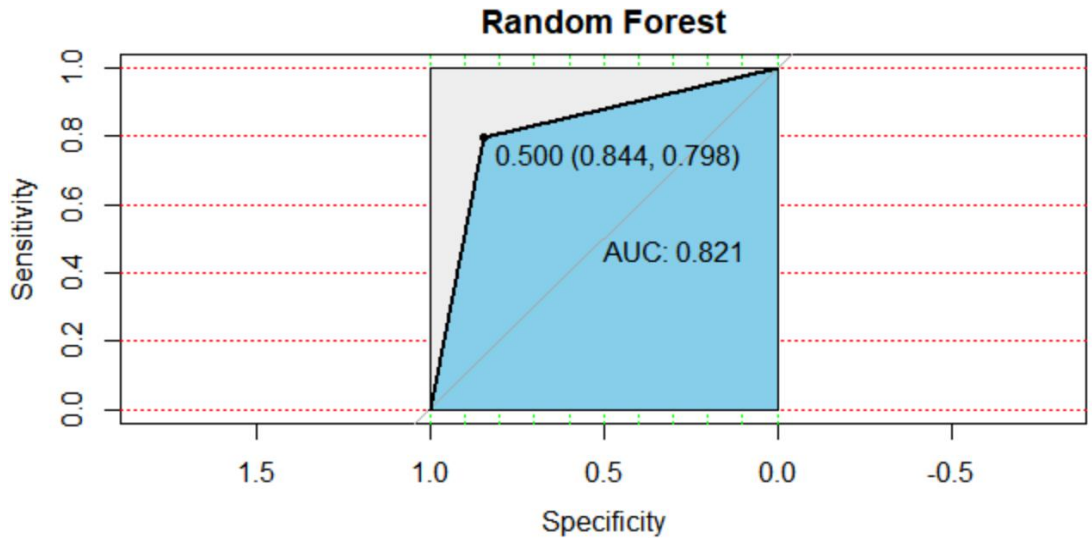

Figure 10 ROC of Random Forest

4.6 Model Comparison

In the original dataset, there is a column named 'loan_grade', which reflects certain credit risk based on the lender's model. In this section, we first assess the accuracy of the 'loan_grade' to see whether the default grade can predict the loan default rate well. Then the above five models are compared to find out the most suitable one for predicting credit risk.

Grade A and B are regarded as high credit so that chances are few for default on these loans. Other grades (Including grade C/D/E/F) are combined into a new tag--'Grade_other', and are believed to have high chances of default. After comparing the grade and the actual loan status, we get Table 7:

Table 7 Grade and Default

	Grade_A/B	**Grade_Other**
Default	2970	2774
Non- default	18254	8576

The grade in the original dataset is far from accurate. Overall only 21,028 of all the 32,574 samples (about 65%) have been correctly predicted. When analyzed separately, the original model is able to measure default at a 48% accuracy and can predict no default at a 68% accuracy. It can be concluded that the original grade is not strict enough to detect possible credit risk.

Besides, when taking a closer look at the grade, it can be found out that the grader tried to divide the grades evenly among different age groups. About 65% people among each age group could receive a high credit grade (A/B) while near 5% people among each age group were graded an especially low credit level. (Shown in Table 8) Whereas such grading method could avoid certain discrimination, banks may lose potential customers if making loans in accord to this grading method. Therefore, other models have to be applied to better the default prediction process, as well as to balance nondiscrimination and profit.

Table 8 Grading Level in Different Ages

	A/B	C/D	E/F/G
20-29	65%	31%	4%
30-39	64%	32%	4%
40-49	65%	31%	4%
⩾50	61%	33%	5%

Until now, we have exploited five models to classify the customers into two groups, non-default and default. What we should do now is to compare the five models according to some criteria, and see which model performs best based on our credit risk dataset.

Basically, we should compare the sensitivity, specificity and accuracy of the classification models based on the confusion matrix. In detail, sensitivity represents the correctly predicted rate of observed positive results, while the specificity indicates the ratio of observed negative results confused with the positive classification. Accuracy represents the correct proportion of the prediction. The three criteria can be shown in the following formula:

$$\text{sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = 1 - \text{FPR} = 1 - \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{TP} + \text{FN} + \text{TN}}$$

After calculation, the sensitivity, specificity and accuracy is shown in Table 9.

Table 9 Sensitivity, Specificity and Accuracy of the Models

	sensitivity	Specificity	accuracy
KNN	0.82	0.79	0.81
SVM	0.75	0.80	0.77
Decision Tree	0.74	0.77	0.76
Random Forest	0.80	0.84	0.82

To compare a model more intuitively, we draw a plot, which is depicted in Figure 11. As can be seen from the plot, KNN performs the best in terms of sensitivity, and in terms of specificity and accuracy, random forest performs the best. In general, random forest performance is quite superior.

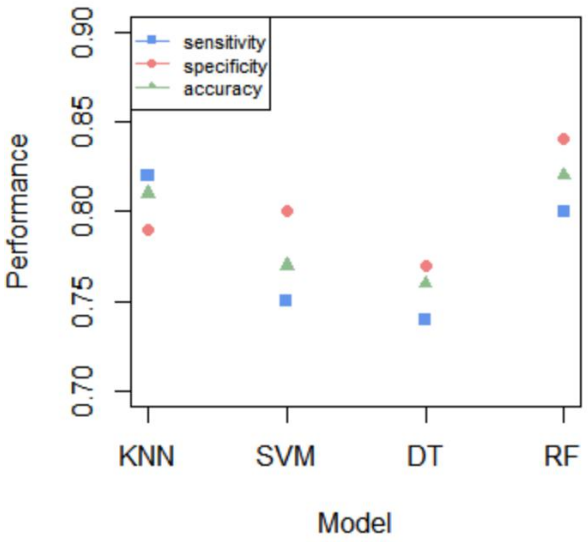

Figure 11 Sensitivity, Specificity and Accuracy

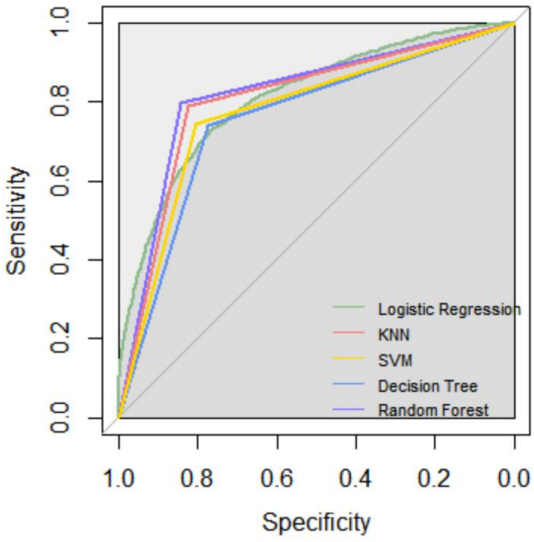

Figure 12 ROC Curve of the Models

Next, we will compare the ROC curve and AUC of the five models, which can consider both sensitivity and specificity. The receiver operating characteristic (ROC) curve is an important tool for the evaluation and comparison of predictive models when the outcome is binary. If the class membership of the outcomes are known, ROC can be constructed for a model, and the ROC with greater area under the curve (AUC) indicates better performance (Peizhou Liao, Hao Wu, and Tianwei Yu, 2017).

The ROC curves are shown in Figure 12, and the AUC of the five models are 0.809 (Multiple Logistic Regression), 0.808 (KNN), 0.775 (SVM), 0.758 (Decision Tree), and 0.821 (Random Forest), respectively. So Random Forest performs best, which is consistent with our comparison above.

4.7 Model Analysis without PCA

In order to compare which variables are most important and predictive, we decide to do the model analysis without PCA.

In the Multiple Logistic Regression Model, the result is:

$$\ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = -3.12 - 0.004\mathrm{age} - 0.000005\mathrm{income} - 0.35\mathrm{home} - 0.008\mathrm{emp} - 0.13\mathrm{intent} - 0.00003\mathrm{amnt} + 0.26\mathrm{rate} + 7.92\mathrm{percent} + 0.001\mathrm{len}.$$

The mean square error on the training data is 0.169 and the mean square error on the testing data is 0.168.

We notice that the variables “age”, “income”, “home ownership”, “employment”, “intent” and “loan amount” have negative correlation with default probability, while the variables “interest rate”, “loan amount/income percent” and “length” have positive correlation with default probability, which is consistent with our common sense.

Additionally, according to the summary of the regression result, the coefficients of all the variables except "age" and "length" are significant, indicating that the two variables have no significant influence on default probability.

In the KNN Model, we set k=15 and distance=2, as above. The confusion matrix is shown in Table 10, and AUC equals to 0.832.

Table 10 Confusion Matrix of KNN (without PCA)

train	**0**	**1**	**test**	**0**	**1**
0	10410	1177	**0**	2385	452
1	1164	10155	**1**	508	2381

In the SVM Model, we use radial as the kernel, which performs best as above. The confusion matrix is shown in Table 11, and AUC equals to 0.809. Besides, the accuracy on training data is 0.8106, and the accuracy on the testing data is 0.8082.

Table 11 Confusion Matrix of SVM (without PCA)

train	**0**	**1**	**test**	**0**	**1**
0	9813	2578	**0**	2401	662
1	1761	8754	**1**	436	2227

In the Decision Tree Model, the result is shown in Figure 13. We can see that the variable "percent (loan amount/ income)" is the most important factor, the variable "rate (interest rate)" and "home (home ownership)" follows.

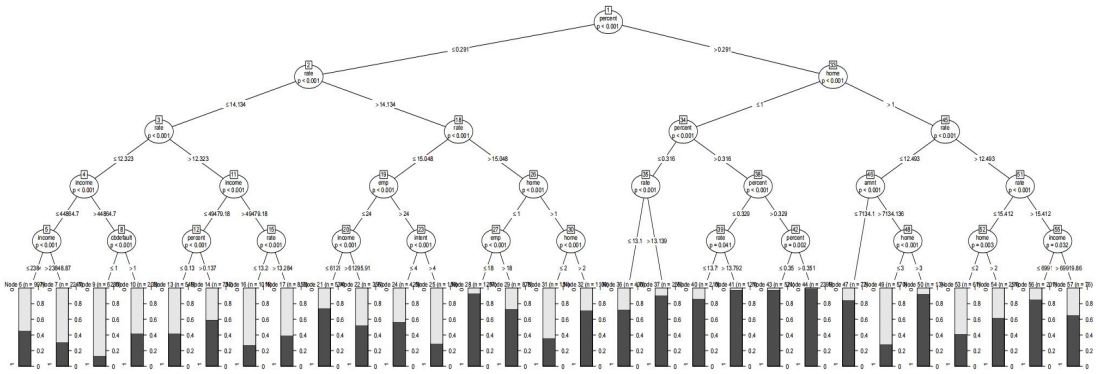

Figure 13 Decision Tree (without PCA)

The confusion matrix is shown in Table 12, and AUC equals to 0.779. Besides, the accuracy on training

data is 0.7814, and the accuracy on the testing data is 0.7784.

Table 12 Confusion Matrix of Decision Tree (without PCA)

train	**0**	**1**	**test**	**0**	**1**
0	9823	3257	**0**	2392	824
1	1751	8075	**1**	445	2065

In the Random Forest Model, the importance of the variables to the classification is as depicted in Figure 14, showing that "interest rate" are the most important variables in terms of the decrease of accuracy and "loan amount：income percent" is the most important with respect to the decrease of gini impurity. So for a bank, it should pay more attention to the two variables.

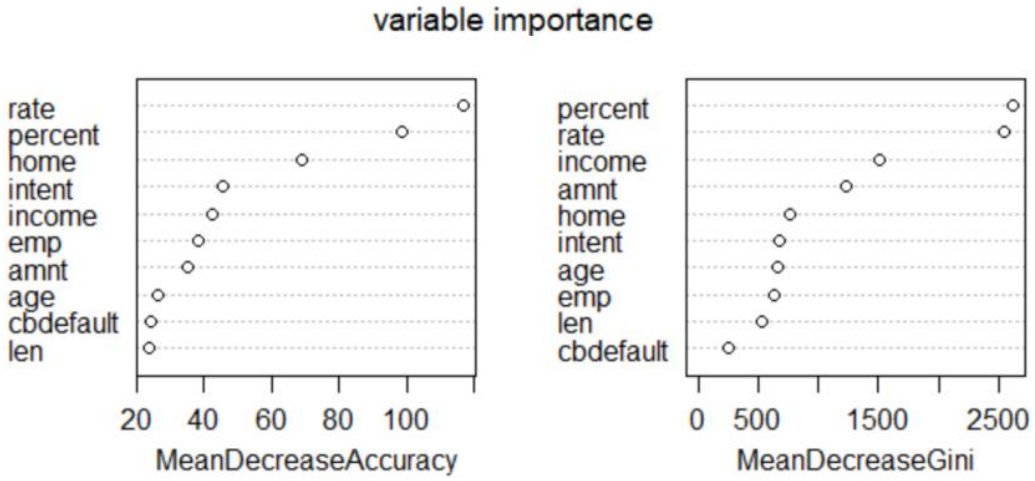

Figure 14 Variable Importance (without PCA)

The confusion matrix is shown in Table 13, and AUC equals to 0.831. Besides, the accuracy on training data is 1, and the accuracy on the testing data is 0.8308.

Table 13 Confusion Matrix of Decision Tree (without PCA)

train	**0**	**1**	**test**	**0**	**1**
0	11574	0	**0**	2460	592
1	0	11332	**1**	377	2297

Similarly, we are going to compare the five models without PCA, to see if the result is different from that with PCA.

After calculation, the sensitivity, specificity and accuracy is shown in Table 14.

Table 14 Sensitivity, Specificity and Accuracy of the Models (without PCA)

	sensitivity	specificity	accuracy
KNN	0.84	0.82	0.83
SVM	0.78	0.84	0.81
Decision Tree	0.74	0.82	0.78
Random Forest	0.81	0.86	0.83

To compare a model more intuitively, we draw a plot, which is depicted in Figure 15. As can be seen from the plot, KNN performs the best in terms of sensitivity and accuracy, and in terms of specificity and accuracy, random forest performs the best, which is similar to the models with PCA.

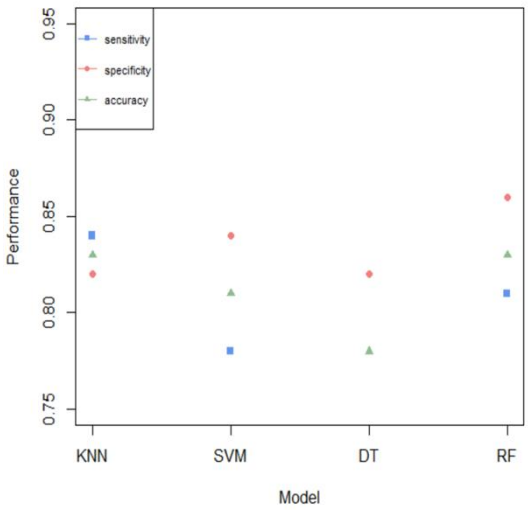

Figure 15 Sensitivity, Specificity and Accuracy

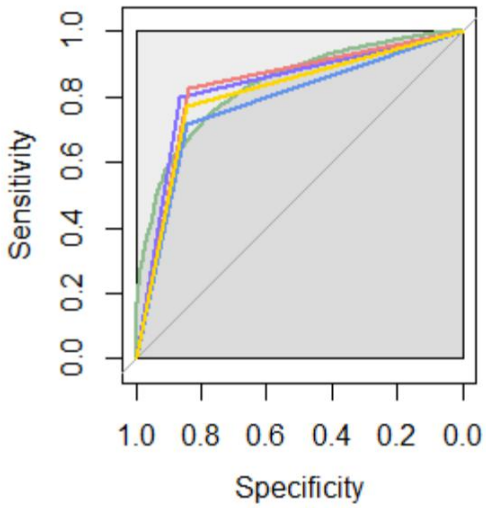

Figure 16 ROC Curve of the Models

Next, we will compare the ROC curve and AUC of the five models. The ROC curves are shown in Figure 16, and the AUC of the five models are 0.833 (Multiple Logistic Regression), 0.832 (KNN), 0.809 (SVM), 0.779 (Decision Tree), and 0.831 (Random Forest), respectively. So Multiple Logistic Regression performs best, KNN and Random Forest follows.

What should be noticed is that the performance of the models without PCA improves a lot than that with PCA since some information must be lost when reducing dimension. Besides, we can see from the results that the variable “loan amount/income percent” and “interest rate” are important and predictive, which deserve attention from banks.

5. Model Combination

In order to improve the accuracy of the models, we decide to combine three models with better performance together, which are Logistic Regression, KNN, and Random Forest. So we try two classic ensemble classifications, boosting and bagging.

5.1 Bagging

Bagging algorithm was proposed by Leo Breiman. It is a method to build a base classifier on each self-help training sample set and get the final category of test samples by voting. Specifically, samples are randomly selected from the returned data set to generate multiple self-help sample sets. The size of each self-help sample set is consistent with the original data set, so some samples may appear in the same self-help sample set multiple times. A base learner is trained for each self-help sample set, and the commonly used base learner is binary decision tree, because for problems with complex decision boundary, the performance of binary decision tree is unstable, which can be overcome by combining multiple decision tree models(what is called Random Forest). Finally, for the regression problem, the result is the mean of the base learner, and for the classification problem, the result is the probability or mean of each category derived from the percentage of the different categories.

Inspired by bagging algorithm, we select Logistic, KNN and Decision Tree as base learners and use bagging to solve the classification problem. The mean square of error of the bagging model is 0.13, decreasing a lot compared to Logistic Regression, and AUC equals to 0.899 (Figure 17), also improves a lot compared to all the previous single models.

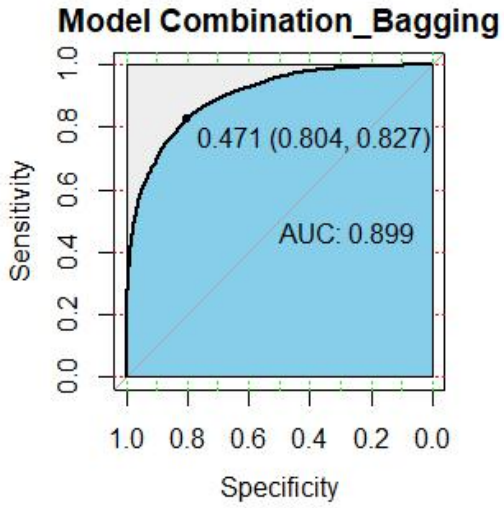

Figure 17 ROC Curve of Bagging Model

Table 15 Performance of Bagging Model

Performance of Bagging Model	
Sensitivity	0.83
Specificity	0.80
Accuracy	0.81

Similarly, we get the 0-1 result and calculate the sensitivity, specificity, and accuracy as shown in Table 15. We can see that the bagging model performs quite well.

5.2 Boosting

Boosting is an integrated learning algorithm that builds multiple weak classifiers to predict the datasets, and then integrates the predicted results of these classifiers with some strategy as the final prediction result.

Unlike bagging, there is a dependence between weak classifiers in the boosting algorithm.

Inspired by XGBoost, whose idea is to constantly add trees to fit the residuals of the previous step, our idea is to use Logistic Regression, Random Forest, and KNN in turn, to fit the residuals of the previous step respectively. The final score of a sample is to add the three scores predicted by the three models together. (Note that in this combined model, KNN and Random Forest are used to do regressions, so the final prediction of a sample is a numeric, not a factor.)

The mean square of error of the boosting model is 0.06, decreasing a lot compared to Logistic Regression, and AUC equals to 0.986 (Figure 18), also improves o lot compared to all the previous single models.

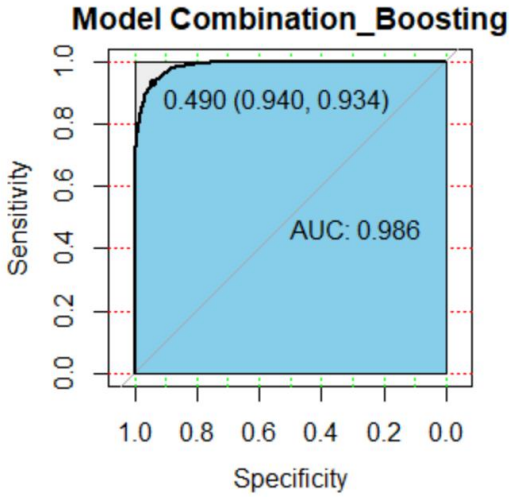

Figure 18 ROC Curve of Boosting Model

Table 16 Performance of Boosting Model

Performance of Boosting Model	
Sensitivity	0.93
Specificity	0.94
Accuracy	0.93

In order to transfer the numeric result into factors (0 or 1), we take the result which is larger than 0.5 as 1, which is less than 0.5 as 0, respectively. Subsequently, we get the 0-1 result and calculate the sensitivity, specificity, and accuracy as shown in Table 16. We can see that the boosting model performs quite well.

6. Further Discussion

6.1 Loan Rate Analysis

In this part, we will focus on the loan rate. From a narrow perspective, loan pricing refers to the price at which loan interest rates are determined. Reasonable loan rate can not only enable banks to obtain satisfactory profits and avoid risks, but also be acceptable to customers, thus achieving maximum efficiency and sustained profitability. High pricing is conducive to high profits after covering loan risks, but higher prices will increase the debt burden of customers. Under the background of interest rate marketization, a considerable number of customers will give up loan demand and turn to banks with low loan pricing or other financial channels. On the other hand, low pricing is difficult to ensure the realization of the bank's profit target, which exposes the bank to more loan risks in turn harming the bank's business development. Therefore, the scientific and reasonable pricing of loans is essential to the bank's loan system.

To have an overview of whether there exists bias, we do linear regression between age and rate. The

multiple R-squared is only 0.0001075 and adjusted R-squared is only 7.261e-05. The regression result is shown in Figure 19.

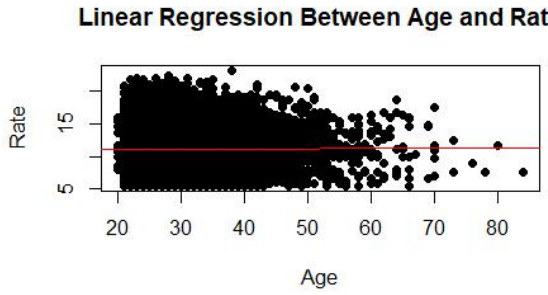

Figure 19 Linear Regression of age and rate

Then we do linear regression between grade and rate. The multiple R-squared is only 0.8714 and adjusted R-squared is also 0.8714. The result shows that there is a strong linear relationship between grade and rate:

$$rate = 2.575 \times grade + 5.303$$

The visualization is shown in Figure 20.

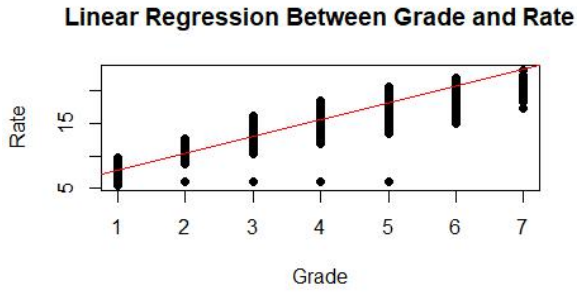

Figure 20 Linear Regression of Grade and Rate

The data we use is the original datasets after dropping the rows with blanks. It is obvious that there is a negative correlation between customers credit ratings and loan rates in general, that is, the higher the customer's credit rating, the lower the bank's pricing level of the rate, according to Figure 20. The situation demonstrates that customer credit grade is an important reference factor taken into account when pricing, and banks are more inclined to supporting customers with high credit level and give them lower rate. As for bias in the issued loans, for example, were higher rates given to younger people even though younger people had a lower default chance? Or the same question for older people. We label the grade A/B/C/D/E/F/G with digit 1/2/3/4/5/6/7, divide people into 5 groups according to their age (group1 is 20-25, group2 is 25-30, group3 is 30-40, group4 is 40-55 and group5 is over 55) and calculate the mean and standard deviation of each group's loan rates. The results is shown in Figure 21 and Figure 22 respectively about mean and standard deviation.

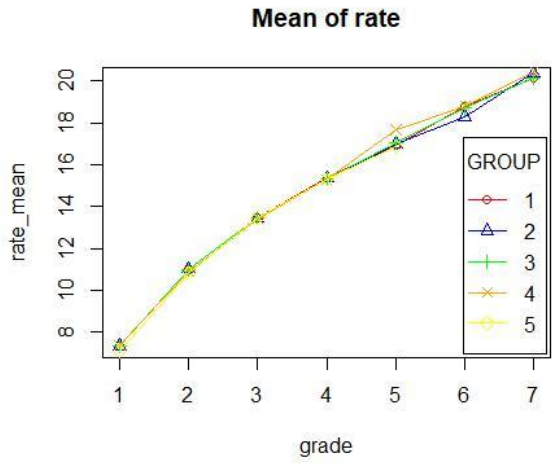

Figure 21 Mean of Rate

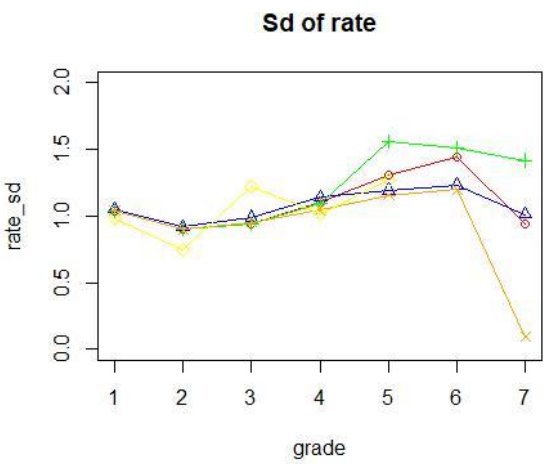

Figure 22 Standard deviation of Rate

Besides, to be specific, we also do linear regression between age and loan interest rate in each group and the result is shown in Figure 23. Because the samples are small in Group 7, the result is not so believable. The R-squared is far less than 0.001 in the former six groups, which shows that almost no linear relationship exists.

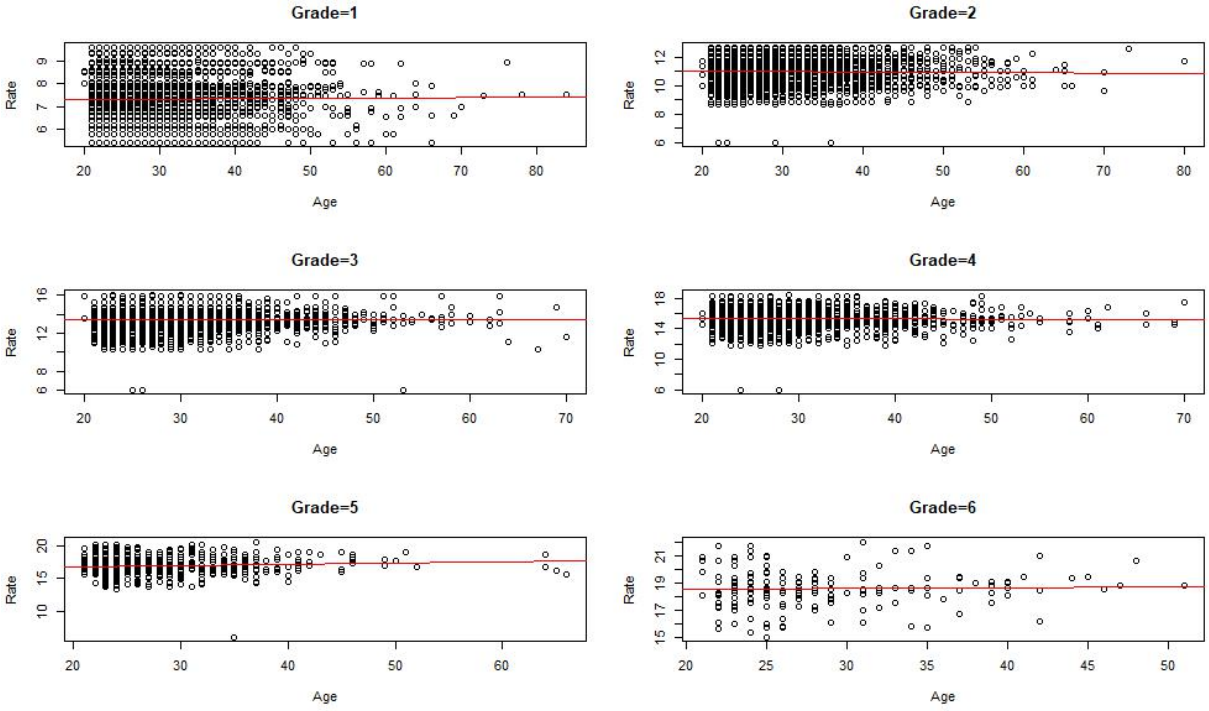

Figure 23 Linear Regression of Age and Rate in Group 1-6

In each age period or each group, the mean of the rates increase as the credit grade rises or in other words customers' default risk climbs and the standard deviation is less than about 1.75 which presents the mean is representative for the group's rates. This shows no matter in which period the person is, he or she will be given higher loan interest rates if he or she have a higher default chance, which proves the rate pricing is reasonable.

6.2 Limitation

Within this report, we also have some shortcomings. We have two major limitations, one is the number and the complexity of our models; the other is about the way we deal with the null value. For the model choosing, five models are a relatively small sample for model choosing we can discover more the more complex model to test the accuracy. Second, we directly ignore the null value which may influence the sample number and lead to data loss.

7. Conclusion & Suggestion

Overall, credit risk management is critical for commercial banks. Better understandings of borrowers' behaviors through data analysis can help banks lower the default risk from loans and improve the performance of their loans in order to protect their interest and maximize profits they can generate. Therefore, credit risk analysis is our main focus.

Through running the data sets from Kaggle we can detect some features about the borrowers and their relationship with default probability. Model choices and comparison are important ways for us to test the accuracy of the default risk prediction. We used five models in total. The receiver operating characteristic (ROC) curve is an essential curve we draw for each model, and the area under the curve (AUC) is the main value we focus on that estimates the accuracy of default risk prediction. The higher the value is the more accurate result is. Multiple logistic regression, K-Nearest-Neighborhood (KNN), Support Vector Machine (SVM), "decision tree" and Random Forest are the models we run.

As for the comparison of these models, we see the grade for each loan as one of our important indicators, which reflects certain credit risk based on the lender's model. Three aspects are needed to make the comparison of these five models ——sensitivity, specificity, and accuracy. According to the confusion matrix, KNN has the highest sensitivity and random forest has the highest of both specificity and accuracy. However, the grade in the original data set is far from accurate, which means the grade systems of this company may be problematic under our model analysis.

Based on all our analysis and data interpretation, our suggestion’s focal point is the grading system, since the grade, the system is an indicator for a company to classify their borrower and relates to the loan interest declaration as well as their profit earning it needs to be improved in terms of the accuracy. For example, the company can collect more background information of their borrowers for an all-around understanding or select a more suitable model the do the classification, under the models we run, Random Forest is the model type we suggest.

References

[1] Qasem, Mais Haj and Nemer, Loai. "Extreme Learning Machine for Credit Risk Analysis" Journal of Intelligent Systems, vol. 29, no. 1, 2020, pp. 640-652.

[2] Rehman, Zia Ur, Muhammad, Noor, Sarwar, Bilal.
"Impact of risk management strategies on the credit risk faced by commercial banks of Balochistan" FINANCIAL INNOVATION, vol. 5, no. 1, 2019.

[3] Psillaki, Maria, Ioannis E. Tsolas, and Dimitris Margaritis. "Evaluation of credit risk based on firm performance." European journal of operational research 201.3 (2010): 873-881.

[4] Henley WE, Hand DJ (1996) A k-nearest-neighbour classifier for assessing consumer credit risk. The Statistician 45(1):77

[5] Farquad MAH, Sriramjee VR, Praveen G (2011) Credit scoring using pca-svm hybrid model. In: Das VV, Stephen J, Chaba Y (eds), Computer networks and information technologies, Springer, Berlin

[6] Peizhou Liao, Hao Wu, and Tianwei Yu. ROC Curve Analysis in the Presence of Imperfect Reference Standards. Stat Biosci. 2017 June ; 9(1): 91–104. doi:10.1007/s12561-016-9159-7.

[7] Lappas P Z, Yannacopoulos A N. A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment[J]. Applied Soft Computing, 2021, 107: 107391.

[8] Masmoudi K, Abid L, Masmoudi A. Credit risk modeling using Bayesian network with a latent variable[J]. Expert Systems with Applications, 2019, 127: 157-166.

Appendix

Data Source: Kaggle *(https://www.kaggle.com/laotse/credit-risk-dataset)*.

The distribution of each column is shown below.

- person_age

Among our data for age, the maximum is 94 years old and the minimum is 20 years old , which has a 28 on average and 6.21 standard deviation. The major age group to borrow money is 20-30 years old.

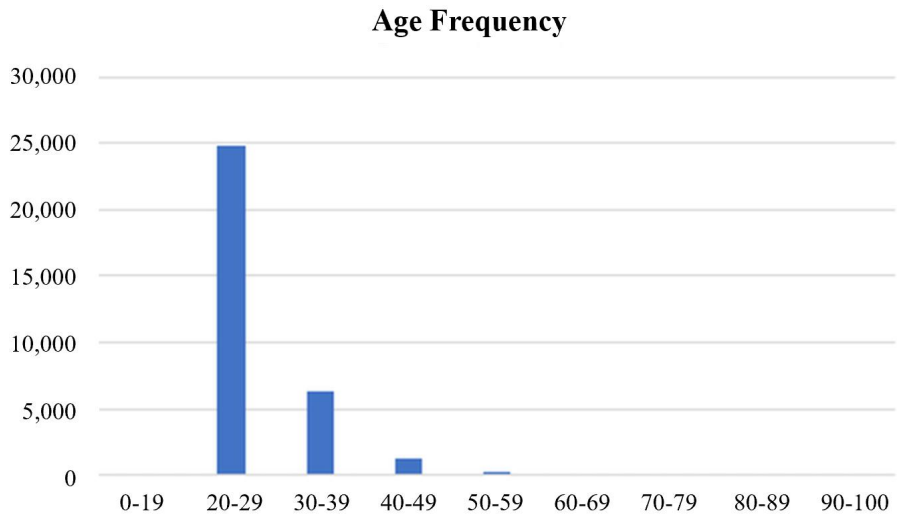

Figure Distribution of Age

- person_income

For income per person, the maximum is 2039784 dollar and the minimum is 4000 dollar. And the people in the dataset has an income of 65878.48 on average, with standard deviation equaling 52531.94, which indicates a wide gap of income. Most people's income is lower than 100000 dollars and the major income group to borrow is 40000-80000 dollars.

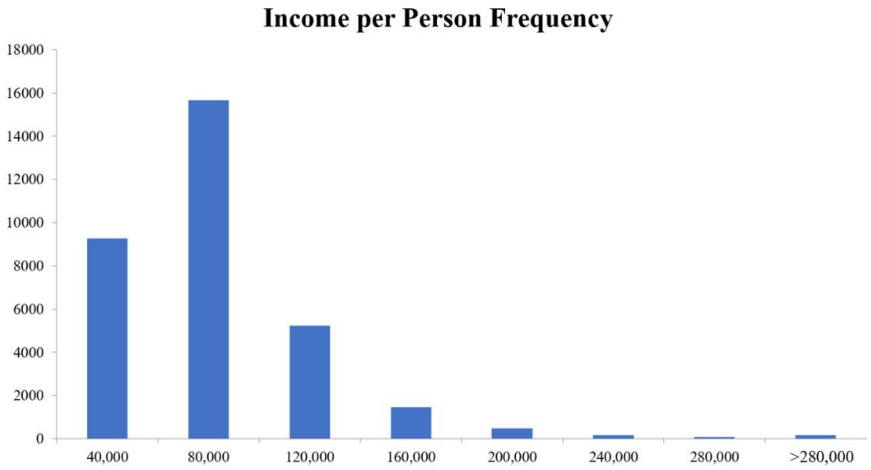

Figure Distribution of Income

- person_home_ownership

Rent, own and mortgage are three major types of ownership. As we can see though the distribution, people who rent house or do mortgage are more likely to borrow money from the financial company than those who own their homes.

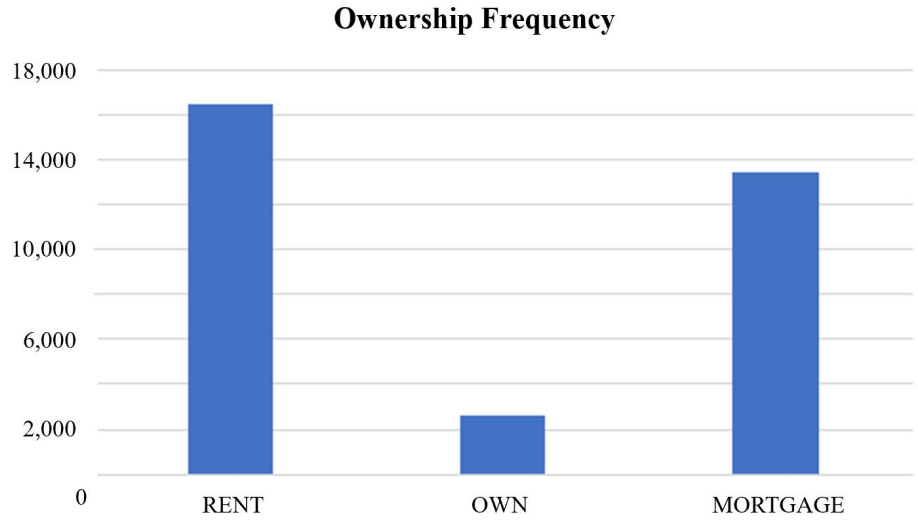

Figure Distribution of Home Ownership

- loan_amnt

For loan_amount, which describe the amount of money this person have loaned, the maximum is 35000 and the minimum is 500 which has a 9588.02 on average and 6320.25 standard deviation. Most loans are below 15,000 dollars from the distribution.

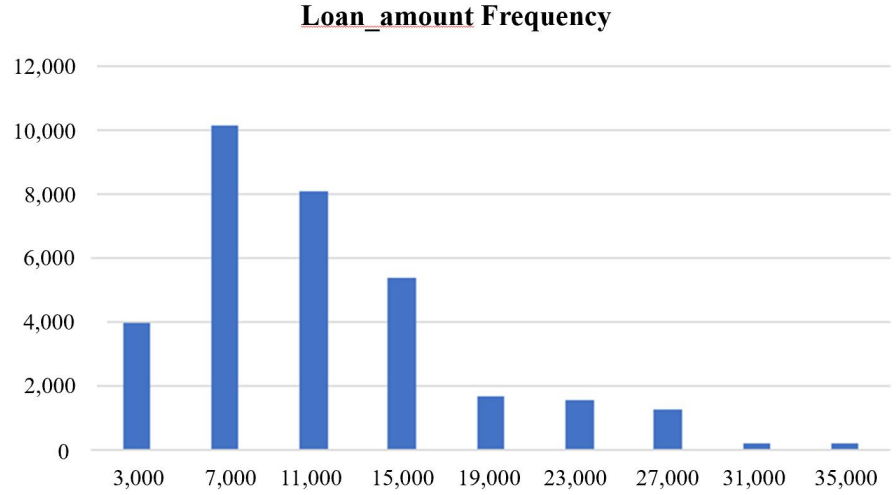

Figure Distribution of Loan amount

- person_emp_length

This data set describes the employment length. From the distribution we can see that the lower the employment length the more people will borrow. And the most frequent employment length is smaller than three years.

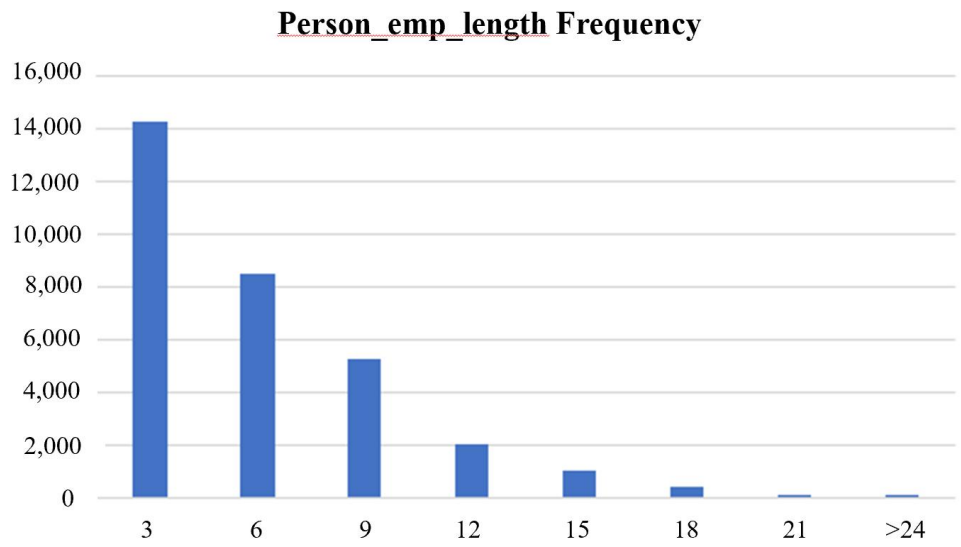

Figure Distribution of Employment length

- loan_intent

We have six loan intents the they have a relative even distribution except for Home improvement and Education is the most common reason for loan within this data set.

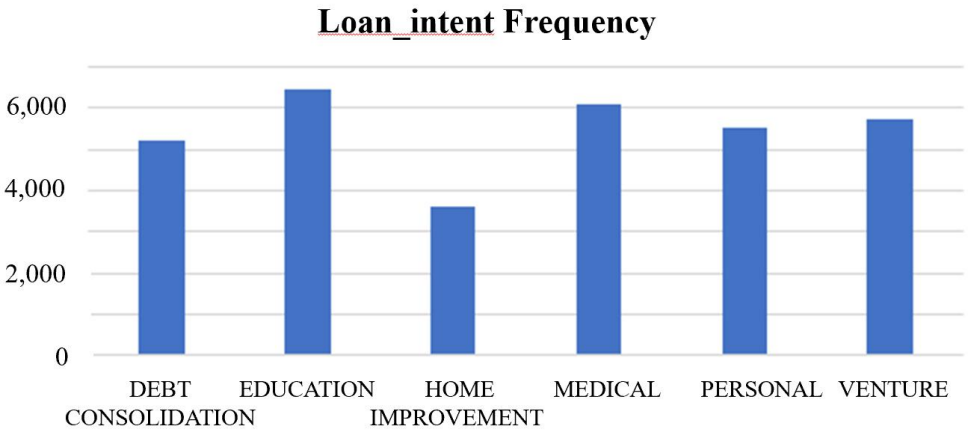

Figure Distribution of Loan intent

- loan_grade

Loan grade is the classification of borrower to assess their payback ability according to their default risk.From A-D, A is stands for those are most trustworthy people and D is the grade to whom are the easiest group to default.

- loan_int_rate

Loan interest rate may be different for different borrowers. For instance, people with bad credit record may receive a relatively high interest rate, which is often called risk premium. Sometimes, interest rate may also differ because of the floating risk-free interest rate. When economy is prospering, people can usually borrow money from banks with low capital cost.

In the dataset, we find out that the highest loan interest rate is 23.22% and the lowest one is only 5.42%. Except for the 3115 blanks in the column (account for 9.5% of all samples), the average loan interest rate is 11.01%, with standard deviation equals 3.24.

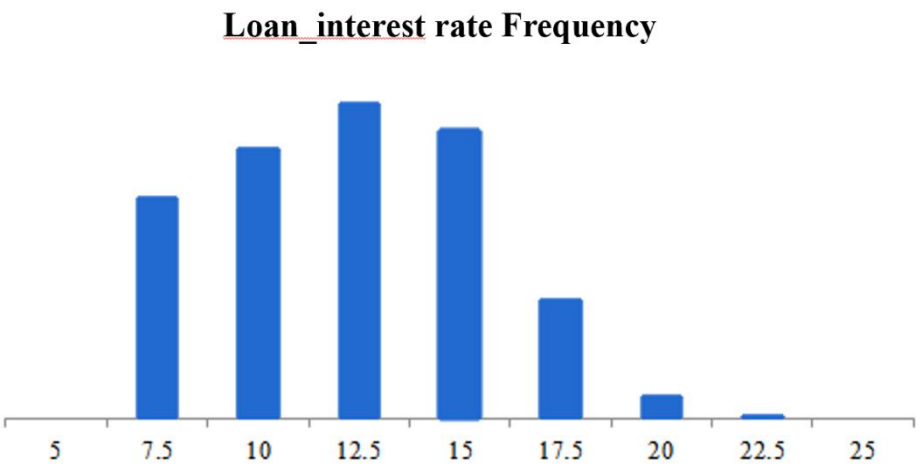

Figure Distribution of Loan interest rate

It can be concluded that most people in our sample borrowed their money at 7.5%~17.5% interest rate. It is higher than the present condition in China.

- loan_status

Loan status is one of the factors that the financial institutions like banks concerns most. If a loan default happens, the bank itself would not get its money back. Therefore, loan_status is the explained variable in our research. Various models are be applied to predicting whether a loan default would happen or not.

In the dataset, there are overall 7107 loan default, which account for 21.82% of all the sample loans. In reality, such default rate is quite high for banks and may cause severe risks in daily operation.

- loan_percent_income

The quotient between loan and income provides a potential reason for loan default. To be more specific, people with relatively large amount of loan but little income are always more possible to have fraudulent intent.

According to the dataset, people's loan divided by income equals 17% on average, which means that most are able to control their debt at a reasonable level. However, we also detect some extreme value in this column (the maximum is over 80%). This indicates that some people in our sample may be facing a poor financial condition and their income can hardly cover their loan.

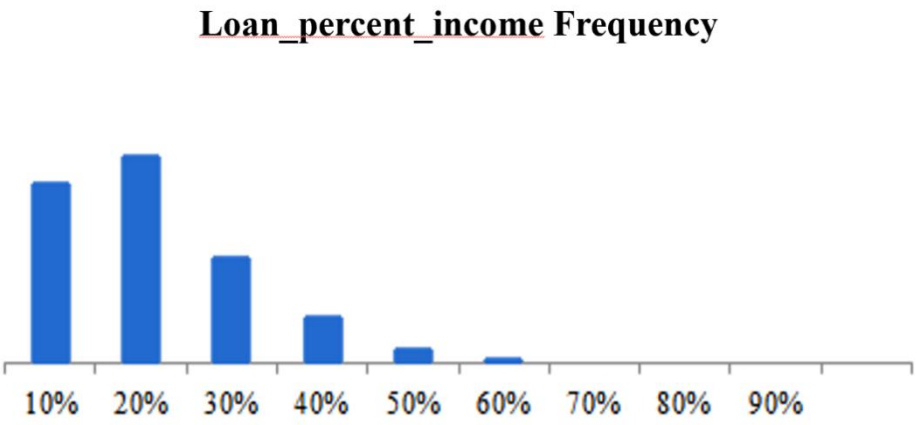

Figure Distribution of loan_percent_income

- cb_person_default_on_fil

As the old saying goes, it is hard to change one's nature. People with default record are always more probable to default twice. Based on our dataset, 5744 people had default on file, making up 17.6% of the whole sample. This ratio is comparative to the loan_status, making our sample data more reasonable to deal with.

- cb_person_cred_hist_length

Although in the cb_person_default_on_fil column, some do not have any default record. This does not necessarily mean that he/she never default. Another reason for this can be the short time of the credit length. Therefore, person's credit history length should also be taken into consideration in our model.

According to our dataset, people on average have 5-year credit record. The minimum credit length is only 2 years and the longest reaches 30 years. Since most people in the sample has a credit history for over 3 years, we regard the default record on file as believable.

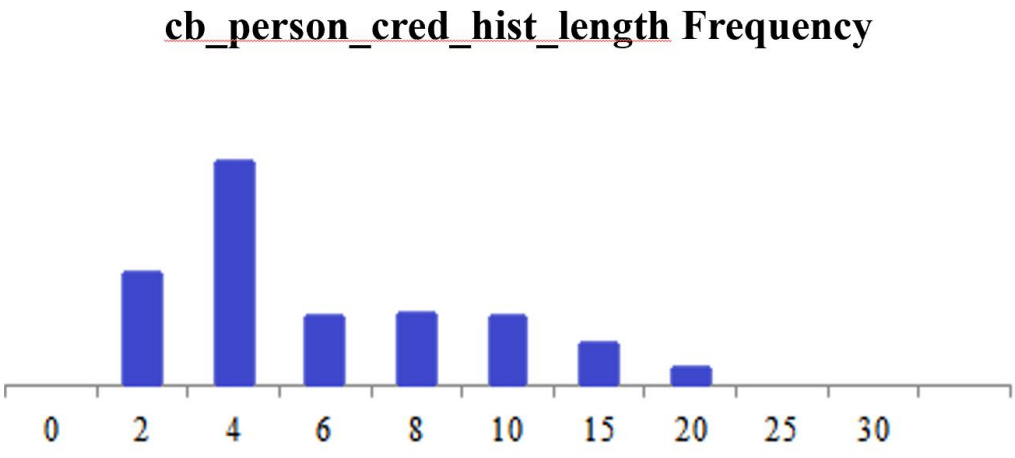

Figure Distribution of Credit history length