STATS 503: Final Project Report

A Comparative Analysis on Multiple Classification Models of Credit Scoring

Abstract

The need for controlling and effectively managing credit risk has led lending companies to excel in improving techniques designed for this purpose, contributing to the development of various quantitative models by financial institutions. Therefore, more and more academic studies about credit scoring are conducting a variety of classification methods to discriminate good and bad borrowers. Our project is intended to apply binary classification techniques for credit scoring financial analysis and to find the optimal classification model fitting the loan data of Lending Club on *Kaggle.*

1. Introduction

Nowadays, credit rating is becoming increasingly important as financial institutions have been experiencing serious competition and challenges. It is a crucial task for financial institutions to decide whether or not to grant credit to consumers who apply since a bad forecast will bring huge losses to the institutions. In fact, this kind of problem is a problem of classification, whose purpose is to make a distinction between "good" and "bad" customers. The financial institutions should extend credit to "good" ones in order to increase the revenue and reject "bad" ones to avoid economic losses. Our objective is to explore the recent loan data posted on *Kaggle*, which contains recent information up to 2016. We want to extract useful information from this dataset, compare different classification models in terms of error rates to see which classification method(s) are useful in handling this type of data, and analyze the impact of certain variables so that we can give some advice on selecting variables for building credit scoring system. In the following sections, we introduce some popular methods used in credit scoring analysis, perform classifications using these methods on our cleaned dataset and summarize our findings and results.

2. Classification Models

Based on the related papers on credit scoring models, the most commonly used methods for credit scoring are LDA, logistic regression, decision trees and SVM.

2.1 LDA

LDA (linear discriminate analysis) was first proposed by Fisher as a classification method. LDA uses linear discriminant function (LDF) which passes through the centroids of the two classes to classify the customers. The linear discriminate function is following:

$$LDF = a_0 + a_1x_1 + a_2x_2 + \ldots + a_\mathrm{n}x_\mathrm{n}$$

where $x_1 \ldots x_n$ represents feature variables of the customers and $a_1 \ldots a_n$ indicates discrimination coefficients for n variables. LDA is the most widely used statistical methods for credit scoring. However, it has also been criticized for its requirement of linear relationships between dependent variables and independent variables and the assumptions that the input variables must follow Normal distribution. To overcome these drawbacks, logistic regression, a model which not requires the normal distribution of variables, was thus introduced.

2.2 Logistic Regression

Logistic regression is a further development of linear regression. It has less restriction on hypothesis about the data and can deal with qualitative indicators. LDA analysis whether the user's characteristic variables are correlation, while Logistic regression has the ability to predict default probability of an applicant and identify the variables related to his behavior. The regression equation of logistic regression is:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

The probability p_i obtained by the equation above is a bound of classification. The customer is considered default if it is larger than 0.5 or not default on the contrary. Lin (2009) uses optimal cutoff point approach and cross-validation to construct a financial distress warning system and get a new cut point 0.314 for classification. Logistic regression is proved as effective and accurate as LDA, but does not require input variable to follow normal distribution.

2.3 Decision Tree

Decision tree method is also known as recursive partitioning. It works as follows. First, according to a certain standard, the customer data are divided into limited subsets to make the homogeneity of default risk in the subset higher than original sets. Then the division process continues until the new subsets meet the requirements of the end node. The construction of a decision tree process contains three elements: bifurcation rules, stopping rules and the rules deciding which class the end node belongs to. Bifurcation rules are used to divide new subsets. Stopping rules determine that the subset is whether or not an end node. C 4.5 and CART are the most two common methods of credit evaluation.

Random forest can also be conducted as an extension to decision tree. Random forest constructs multitude decision trees at training time and outputs the class that is the mode of the classes of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

2.4 Support Vector Machine (SVM)

This technique is a popular statistical classification method. Given a training set $\{(x_i, y_i)\}$ with $i = \{1, \dots, n\}$, where x_i is the explanatory variable vector, and y_i represents the binary category of interest, and n denotes the number of dimensions of input vectors. SVM attempts to find an optimal hyperplane, making it a non-probabilistic binary linear classifier. The optimal hyperplane could be written as follows:

$$\sum_{i=1}^{n} w_i x_i + b = 0$$

where $w = (w_1, \dots, w_n)$ is the normal of the hyperplane, and b is a scalar threshold. Considering the hyperplane separable with respect to $y_i \in \{-1, 1\}$ and with geometric distance $2/\| w \|^2$, the procedure maximizes this distance, subject to the constraint

$$y_i(\sum_{i=1}^{n} w_i x_i + b) \geq 1$$

Commonly, this maximization may be done through the Lagrange multipliers and using linear, polynomial, Gaussian or sigmoidal separations. There are many researches on credit scoring using SVM method. Huang, Chen & Hsu (2004) pointed out that SVM was superior to ANN in aspect of classification accuracy. Lee (2007) found that SVM was better than MDA, CBR and ANN models. Yang (2007) presented an adaptive scoring system based on SVM, which was adjusted according to an on-line update procedure. Kim & Sohn (2010) provides a SVM model to predict the default of funded SMEs. Possible particular methods of SVM are radial basis function SVM, least squares SVM.

3. Credit Data

3.1 Summary Statistics and Data Visualization

This data set is available at *Kaggle*. It contains complete loan data for all loans issued by Lending Club, a US peer-to-peer lending company, through 2007 to 2015. The file has about 887 thousand observations, each representing a single loan record, and 74 variables, including the credit grades rated by the company itself (A-G), loan amount, current loan status, purpose of the loan, number of finance inquiries, address including zip codes and state, and collections among others.

We select several important variables, namely loan status, loan amount, annual income and loan grade to produce summary statistics and data visualization. Loan status in this dataset is defined as “Current”, “Fully Paid”, “Charged Off”, “Issued”, “In Grace Period”, “Late” and “Default”.

Figure 1 shows the distribution by loan status for all loans issued. Only “Current”, “Issued” and “Fully Paid” are considered as good loan status. We can see from the figure that nearly 70% of the loan status is current, over 20% of the loans are fully paid, while the bad loan status like “late”, “charged off”, “in grace period” and “default” take up nearly 10% of the issued loans.

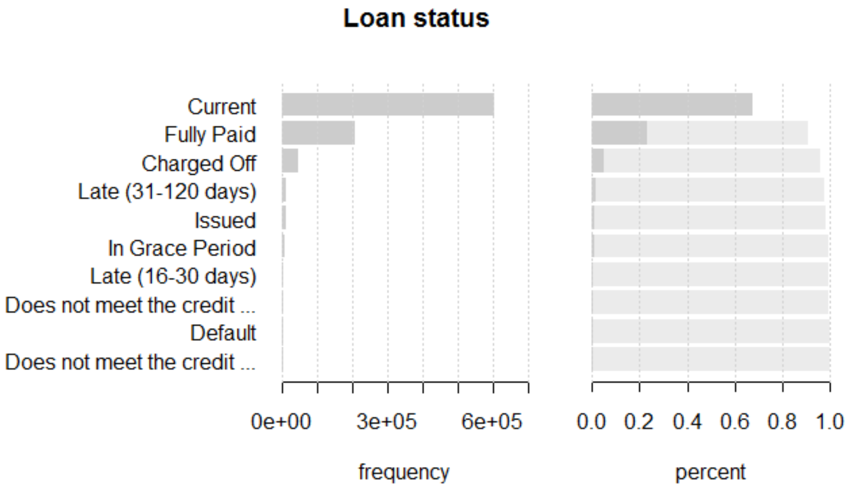

Figure 1: Summary chart of loan status.

Then we plot total loan amount from 2008 to 2016 with respect to loan status (Figure 2). As we can see, the total loan amount keeps increasing during these years. The portion of “current” increases, while the portion of “fully paid” and “charged off” decreases.

The boxplot in Figure 3 shows the summary of interest rate by different status. Bad status like “charged off”, “default”, “in grace period” and “late” has relatively higher interest rate than that of good status like “current”, “fully paid” and “issued”.

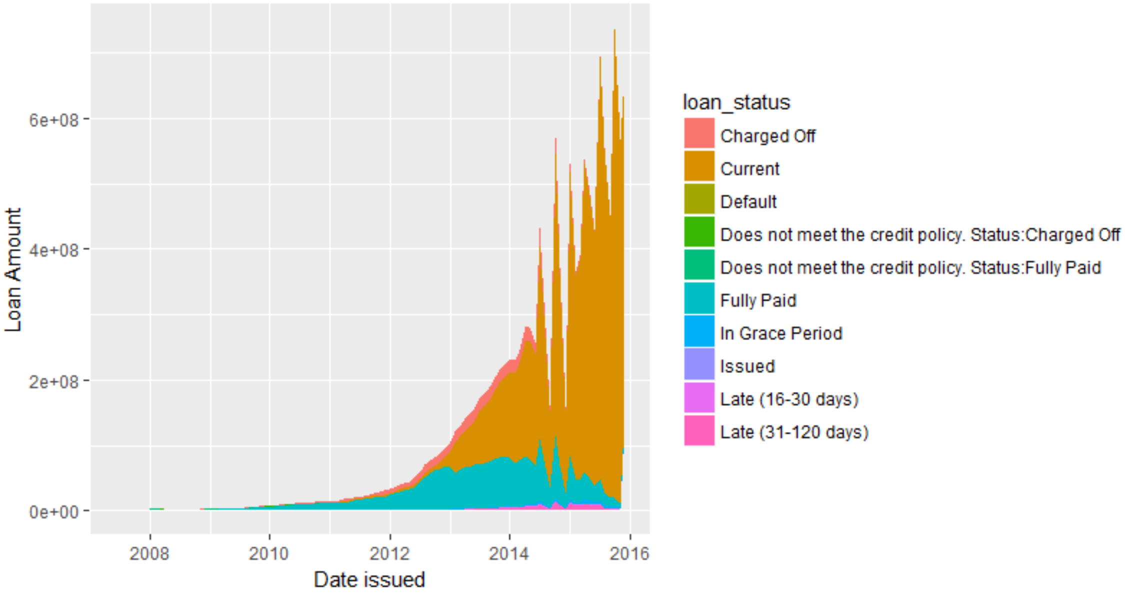

Figure 2: Total loan amount from 2008 to 2016.

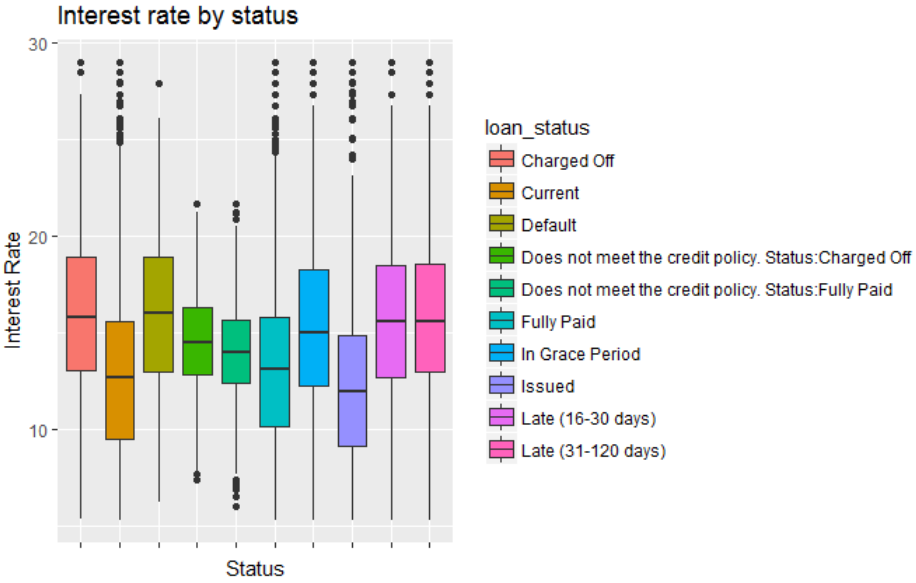

Figure 3: Boxplot of Interest rate by status.

We use "loan/income ratio" to analyze the grade variable. If this ratio is low, it means the loan only accounts for a relatively small portion of the borrower's annual income, making it easier for the borrower to pay back the loan. In general, the higher the ratio, the more difficult the loan can be fully paid. The left plot in Figure 4 indicates that the higher the loan/income ratio, the worse the borrower's grade would be, which complies with our theory. Moreover, the right plot in Figure 4 shows that the borrower with better grade will receive lower interest rate. The loan company would set higher interest rates for borrowers with higher tendency to default.

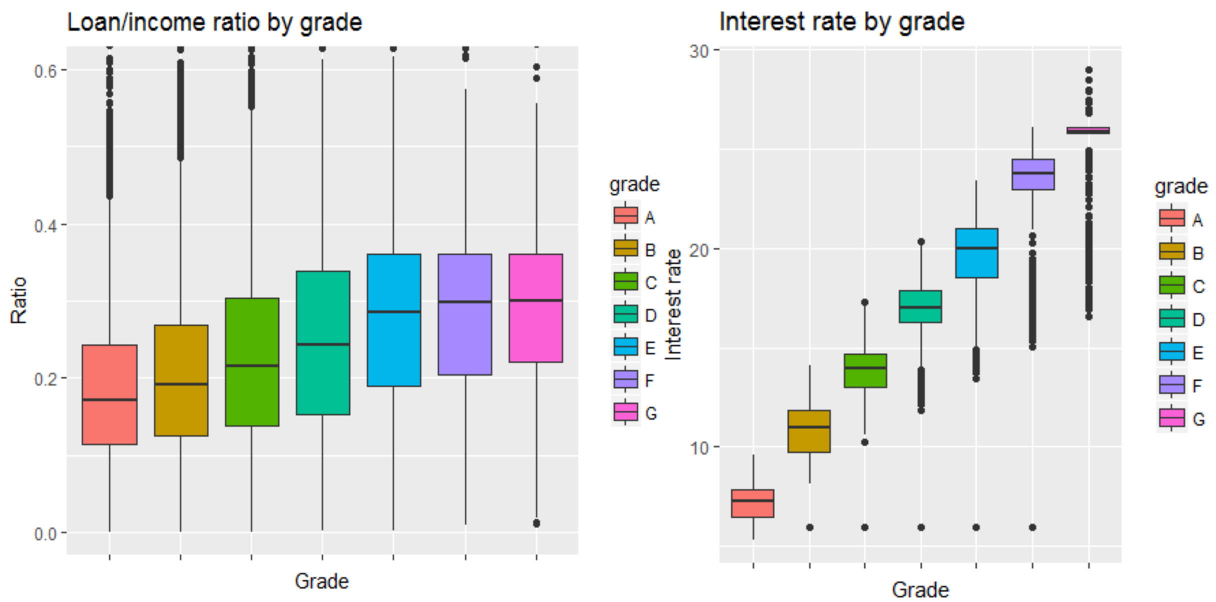

Figure 4: Boxplots of loan/income ratio and interest rate by grade.

Figure 5 presents the relationship between loan grade by loan status. We note that in "charged off" status, there are borrowers with Grade A, which means the grading scores are probably determined before the issuance of the loans. The loan company wouldn't give an "A" grade to a borrower whose loan has already been charged off, so we infer that in order to forecast the borrowers' credit score based on this dataset, the "grade" variable is not an optimal choice. Instead, we decide to use "loan status" as the variable we want to predict.

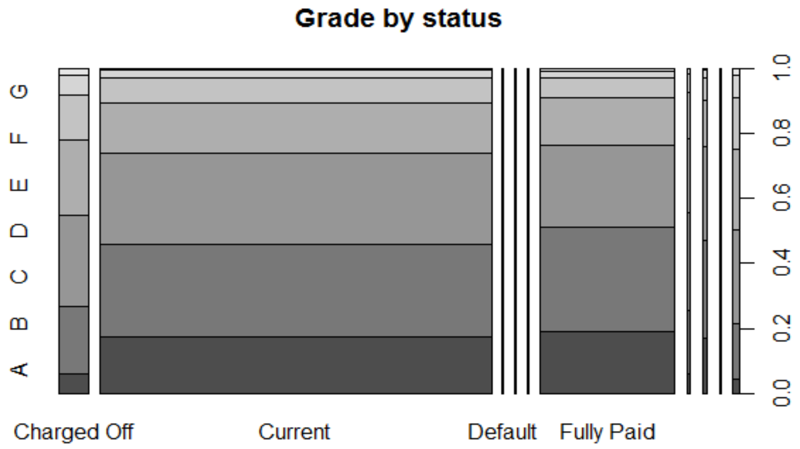

Figure 5: Relationship between loan grade by loan status.

Figure 6 presents the distribution of loan amount and the ratio of fully paid observations over total observation across the states. The east coast, California and Texas are borrowing larger amount of money compared with other states, however, the ratio of fully paid loans are higher on the west coast. Thus, we believe that the regional variable has an impact on the loan status of the borrowers.

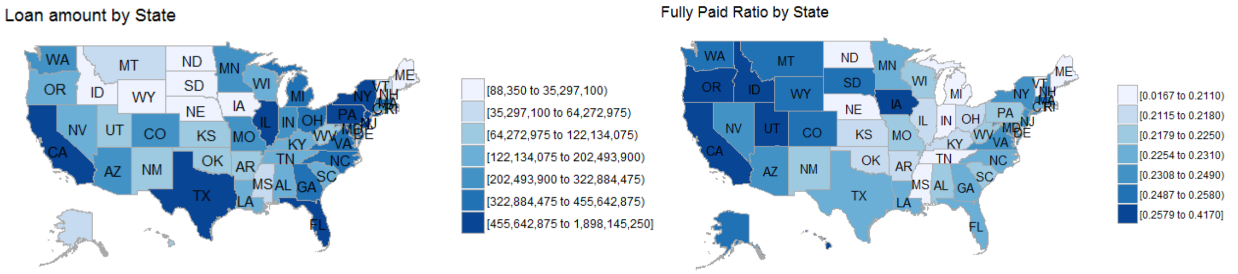
Loan amount by State
[88,350 to 35,297,100)
[35,297,100 to 64,272,975)
[64,272,975 to 122,134,075)
[122,134,075 to 202,493,900)
[202,493,900 to 322,884,475)
[322,884,475 to 455,642,875)
[455,642,875 to 1,898,145,250]
Fully Paid Ratio by State
[0.0167 to 0.2110)
[0.2115 to 0.2180)
[0.2179 to 0.2250)
[0.2254 to 0.2310)
[0.2308 to 0.2490)
[0.2487 to 0.2580)
[0.2579 to 0.4170]

Figure 6: Total loan amount by state (left); Average fully paid ratio by state (right).

4. Data Preprocessing

As described in the previous section, the original dataset contains 887,879 observations, each representing one member of the Lending Club, and there are 74 variables including two different IDs (variable description is available on Kaggle website). It is thus essential to clean our raw data, which is divided into two parts, variable selection and dealing with missing data.

4.1 Variable Selection

Since our goal is to construct a classification model to predict each member's final loan status, we use "loan_status" as our response variable and drop "grade" and "sub_grade" which are two grading systems used by the company itself. In addition, we split "loan_status" into two groups: "good" and "bad". Table 1 shows the definition of each status. According to the description, we group "Current", "Fully Paid" into "good" and "Charged Off", "Default", "Late" into "bad". We also delete observations with status "Issued", "In Grace Period" and "Does not meet the credit policy" in order to reduce noise, since loans in these status are hard to determine whether the outcomes are "good" or "bad".

Table 1: Loan Status Description

Loan Status	Description
Current	Current would appear to represent all loans that are making payments on time and as agreed.
Issued	Issued refers to the loans that are just issued and haven't made any payment yet.
In Grace Period	A grace period is the provision in most loan and insurance contracts that allows payment to be received for a certain period of time after the actual due date. During this period, no late fees are charged, and the late payment does not result in default or cancellation of the loan. A typical grace period is 15 days.
Late	Late status represents loans that are at least 16 days past due. Prosper moves loans that are 120 days past due (i.e. missing 4 payments) out of Late status to Charged Off Status.
Charged Off	After 4 missed payments (i.e. 120 days), a loan moves from Late to Charge Off. A loan designated as charged off is due in full immediately. The loan can also be sold to outside debt collectors. For investors, the entire balance moves into a charge off balance and is assumed to be lost.

Fully Paid	This is a loan that is fully paid off according to the schedule or prepaid. Most loans with this status have a principal balance of $0. Given the high growth of marketplace lending, it means that there are generally more outstanding loans than fully paid loans.
Default	Compared with Charged Off status, all loans with the defaulted designation should also have a loan default reason, such as Delinquency, Bankruptcy, Deceased, or Repurchased.

Among 887,879 observations, there are 886,868 individual loans and 511 joint loans with two co-borrowers. Due to huge difference in sample size between those two types of loan application, we only consider individual loans as our target observations, which results in variable reduction related to joint application (i.e. "application_type", "dti_joint", "annual_inc_joint", "verification_status_joint"). Also, there are only 10 observations with level "y" for the variable "pymnt_plan", thus we base our analysis on those with level "n".

First, we can reduce some of our quantitative variables by examining the correlation between variables. From Table 2, we can see that the three variables are highly correlated, thus it is reasonable to consider keeping only one of the three. Similarly, the correlation between "out_prncp_inv" and "out_prncp" is 0.9999972 and the correlation between "total_pymnt_inv" and "total_pymnt" is 0.9975923, so we only keep "loan_amnt", "out_prncp" and "total_pymnt" in the following analysis.

Table 2: Correlation matrix between "loan_amnt", "funded_amnt", and "funded_amnt_inv"

	loan_amnt	funded_amnt	funded_amnt_inv
loan_amnt	1.0000000	0.9992620	0.9971129
funded_amnt	0.9992620	1.0000000	0.9980235
funded_amnt_inv	0.9971129	0.9980235	1.0000000

Second, we drop some of our categorical variables by carefully evaluating and understanding the variables. We rule out "emp_title" from our model since it contains approximately 300,000 different job titles or companies which can be considered as every member's unique attribute. And also because its information can be reflected by annual income to some extent as well. By comparing "purpose" and "title", we find that they are almost the same thing, so we only include "purpose" in our analysis. Similarly, we abandon "zip_code" but keep "addr_state". For "desc" variable, we set it to be 1 if not missing and 0 otherwise.

There are 42 quantitative variables, 14 categorical variables and 886,858 observations left after getting rid of two ID variables "id" and "member_id", one "url" variable which is considered unnecessary in the following data analysis, "policy_code" which is same for all observations, and other variables dropped from previous steps.

4.2 Missing Data

By further examining the data, there are 14 out of 42 quantitative variables with over 97.5% missing values, 2 variables with over 75% missing values, 1 variable over 50% missing values, 3 variables with more than 5% missing values, 9 variables with under 5% missing values. Among the 15 categorical variables there are 5 variables contain missing data.

We note that the 2.5% observations without missing values are all from "good" loans, thus it will not only reduce our sample size but also generate huge bias if we simply deleted the observations with missing values. As a result, we apply our analysis on 97.5% of our data which contains same missing variables with high ratio of missing values. In this way, we can delete all 14 quantitative variables with high ratio of missing values without generating much bias. For remaining quantitative variables with missing ratio under 8%, we use predictive mean matching to impute the missing values. A nice feature of Predictive mean matching (PMM) is that PMM produces imputed values that are much more like real values which is useful for data that are not normally distributed. Also, if the original variable is skewed, the imputed values will also be skewed. If the original variable is bounded by 0 and 100, the imputed values will inherit the same range. And if the real values are integer, the imputed values will also be integers. We can see from Figure 7 that some of our variables are not normally distributed and some are integers. For categorical variables, we treat missing values as one group and others remain the same.

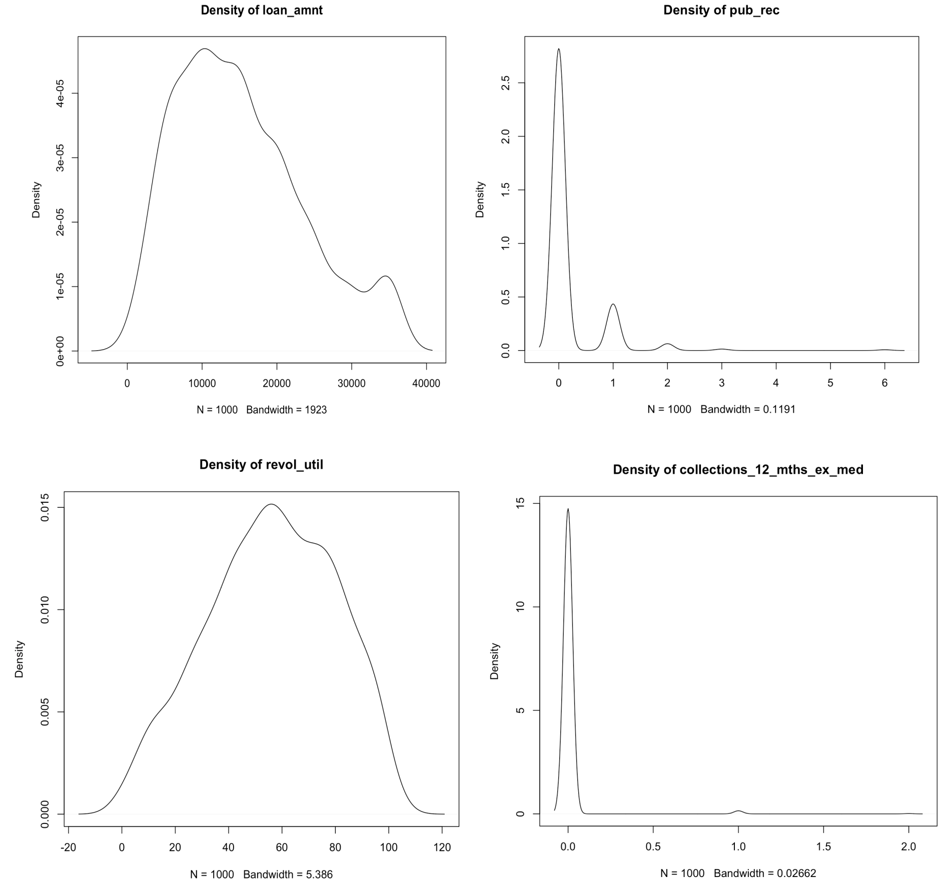

Figure 7: Density plots of four selected variables with 1000 random sampled observations.

4.3 Final Data and Models for Classification

After the first two steps of preprocessing and applying further transformation to some variables and observation selection (concentrate on issue year from after 2013, drop variables missing "emp_length", drop two states with only 2 observations in each), we now have 723,056 observations, 26 quantitative variables, 16 categorical variables, one class variable, among which contains 681,683 "good" loans and 41,373 "bad" loans. Then, we scale the quantitative variables in order to perform classification.

After investigating the methods used in credit scoring analysis in Section 3 and the credit dataset, we choose logistic regression, SVM with radial kernel, decision tree as our main methods based on the following reasons. First, our data is not a Gaussian which does not meet the assumptions of LDA. Second, our data contains a lot of categorical variables, among which "addr_state" contains 49 levels which makes it hard to implement neural network. Overall, the three selected methods are the most popular ones used in credit scoring analysis.

5. Results

In this section, we compare the accuracy of some classification methods, namely logistic regression, supporting vector machine and decision tree. Training and test data, which account for 70% and 30% of the total sample size respectively, are randomly selected by the computer, in which all the levels of each categorical variable can be found. After implementing these 3 methods, we found that the error within the class "bad" on the test data is really high, all of the three being larger than 25%, which is a serious problem because the chances of criminals' slipping away has to be controlled carefully. We think that the bad performance in predicting "bad" is a consequence of too few "bad" observations and we use the command `ubUnder` in the package `unbalanced`, which randomly excludes some instances from the "good" class and keeps all instances in the "bad" class in order to obtain a more balanced dataset. Due to under sampling, 68,168 "good" observations are sampled from the "good" pool of size 681,683. And the error rate of "bad" on test data is obviously slashed at the sacrifice of the increase in total error rate because of less sample size.

5.1 Logistic Regression, SVM and Decision Tree on Unbalanced Data

Table 3: Error report of unbalanced data

Classification Methods	Training			Test		
	Total	Bad	Good	Total	Bad	Good
Logistic	0.01625	0.27403	0.00062	0.01653	0.27576	0.00077
SVM	0.01613	0.27524	0.00042	0.01660	0.27954	0.00061
decision tree	0.01530	0.25748	0.00061	0.01620	0.26514	0.00106

Table 3 presents the training and test errors for each of the two classes separately. We can see that decision tree outperforms other methods in test errors no matter for "good" or "bad", but the leading-edge is not wide. Also, there are huge differences between the two types of errors. All of the three methods yield error greater than 25% in the "bad" class, which is definitely not an ideal result. We believe that the poor performance is resulted from the unbalanced numbers of observations between the two classes, therefore in the next section, we will further examine this issue.

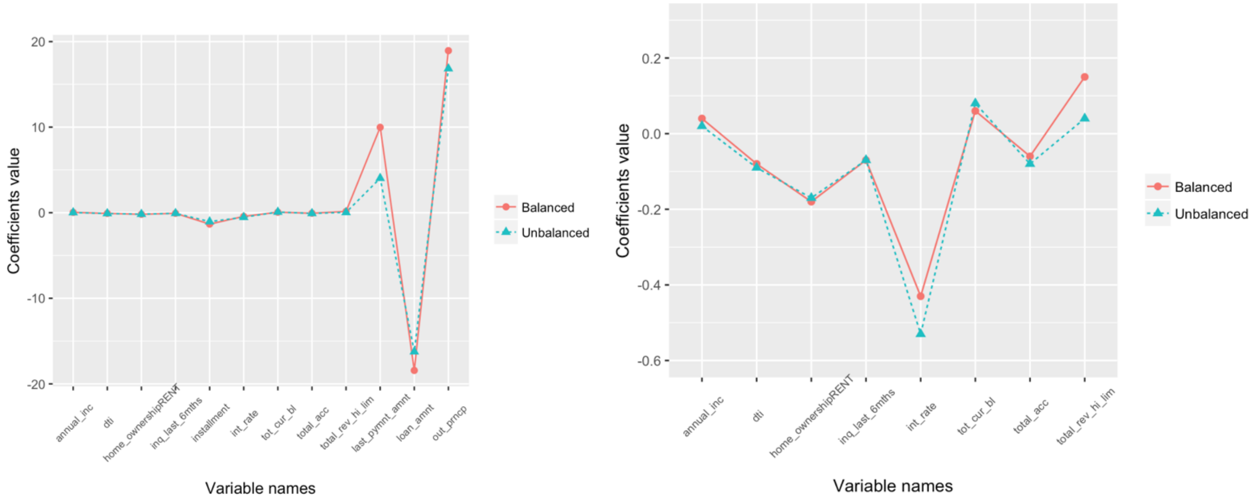

Figure 8: Coefficients of logistic regression on both balanced and unbalanced data.

Figure 8 shows the some of the significant coefficients of logistic regression. We can group the variables into two types: one defining the loan itself (loan_amnt, int_rate, installment), the other defining the borrower's personal status ("annual_inc", "dti", "home_ownershipRENT", "inq_last_6mths", "total_acc", "out_prncp", "last_pymnt_amnt", "tot_cur_bal", "total_rev_hi_lim"). We can see from the figure that the variables with positive coefficients are "annual_inc", "out_prncp", "last_pymnt_amnt", "tot_cur_bal", "total_rev_hi_lim" which signals us that people with higher annual income, remaining outstanding principal, previous payment amount, total current balance of all accounts, ratio of total revolving high credit and credit limit tend to increase the odds of good loans which is consistent with real world. In other hand, the negative coefficients show that people who had credit inquiries in the last 6 months, lived in a rented house (the variable 'home_ownershipRENT" is a factor, and the intercept correspond to the level "home_ownershipMORTGAGE", which is -80.39), and people who had higher dti (a ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income) tend to decrease the odds of good loans. Also, from the variables which characterize the loan, we can find that holding other variables constant, higher loan amount, interest rate, installment in each month will result in the decrease of the odds of good loans. Though those variables characterize whether the borrowers have the ability to pay their loans in the future in different ways, logistic regression can in fact capture these features quite well in terms of the signs.

Figure 9 displays higher branches of the splitting process carried out by decision tree. We can see from the first node that if recoveries are more than 2 in the original scale, the loan is probably a "bad" one, which totally makes sense since recoveries are required once your loan borrowed is charged off. Node 7 in the third level says that if late fees received to date is more than 15 converted back to original data in condition of last payment year being earlier than 2016, the loan is more likely a "good" one. If an agent can pay the late fee in time, the loan he/she borrowed can still be ranked as "good". One more check. We can tell from node 48 in the sixth level that if loan amount is larger than 7800 is more likely to become a "bad" loan given that recoveries are more than 2, last payment year is earlier than 2016, last payment amount is less than 1400, remaining outstanding principal is less than 180 and principal received to date is less than 2740. That larger loan amount gives rise to higher chances of being default matches common sense.

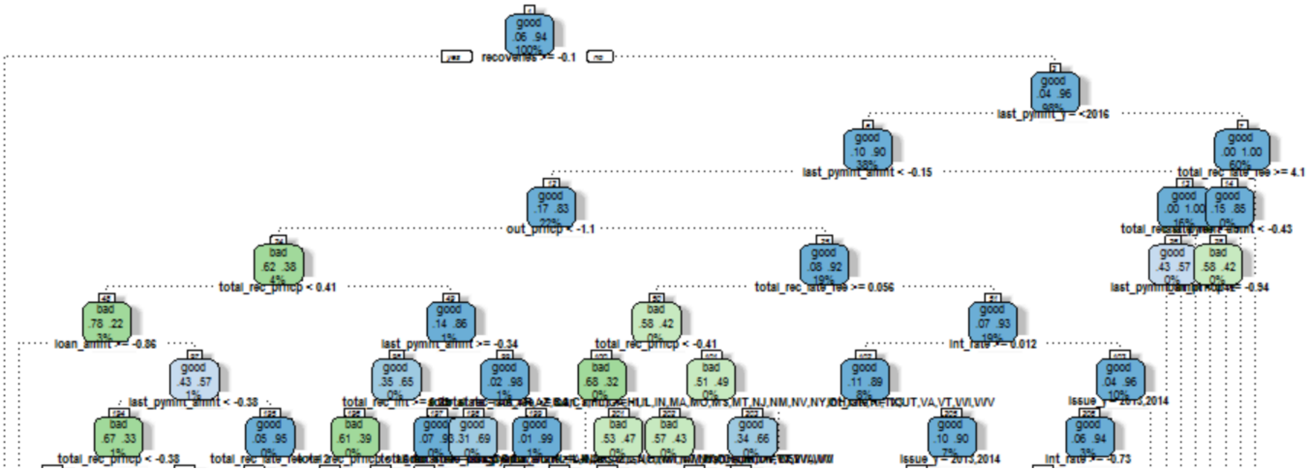

Figure 9: Decision tree plot.

In addition, we make a comparison between the grades ranked by lending club itself and our results to show our superiority over their grading system. Table 4 shows a cross table of the grades and the actual number of "bad" and "good" loans. We can see that most of the lower grades such as E, F and G are in fact "good" loans, and also the error rate of grade A, B, and C are 0.0135, 0.0365, 0.0583 respectively (if we consider "bad" to be error for grade A, B and C), which are larger than the error rates of our models.

Table 4: Whose performance is better? Ours!

	A	B	C	D	E	F	G
Bad	477	2242	3596	3090	1932	850	244
Good	34828	59127	58107	31089	15576	4663	1072

5.2 Logistic Regression, SVM, Decision Tree and Random Forest on Balanced Data

From Figure 8 in the previous section, we can see that the coefficients of logistic regression obtained from the balanced data is similar with the one obtained from unbalanced data, which proved that under sampling keeps overall distribution of the data. From Table 5, we note that this time SVM is the best considering total error and it nearly ties with decision tree in "bad" error.

Table 5: Error report of balanced data

Classification Methods	Training			Test		
	Total	Bad	Good	Total	Bad	Good
Logistic	0.07464	0.12208	0.04587	0.07512	0.12001	0.04783
SVM	0.07199	0.11842	0.04384	0.07393	0.11968	0.04612
Decision tree	0.06838	0.10747	0.04468	0.08027	0.11960	0.05635
Random Forest (mtry=2, ntree=500)	0.01377	0.02970	0.00411	0.07670	0.12701	0.04612

Random Forest with mtry=2 and ntree=500 is not our last trial. Actually, we made several attempts in searching for the best pair of control parameters. Please see table 6 and table 7 for details.

Table 6: Test error report of random forest with ntree=1000 fixed on balanced data

mtry	2	3	4	default
Total Error	0.07710	0.07871	0.08295	0.09153
Bad Error	0.13080	0.06459	0.04575	0.03455
Good Error	0.04446	0.08729	0.10556	0.12617

Table 7: Test error report of random forest with mtry=2 fixed on balanced data

ntree	200	500	1000	1500
Total Error	0.07865	0.07670	0.07710	0.07749
Bad Error	0.13813	0.12701	0.13080	0.13313
Good Error	0.04250	0.04612	0.04446	0.04367

We can see that increasing “ntree” will not improve the performance of random forests significantly, while changing “mtry” really matters. Keeping “ntree” fixed at 1000, increasing “mtry” lowers “bad” error but increases “good” and total error. This will lead to a trade-off balance between “bad” error and “good” error. The “good” error will decrease the revenue of financial institutions, since the loan company tends to issue fewer loans; while “bad” error will generate risk, since the loan company tends to trust more people. It is often believed that higher risk will result in higher revenue. Sometimes, in the real life it is a difficult task to increase revenue and control risk at the same time. Based on our analysis, we can also see that those two types of error are complementary. If financial institutes try to lower the risk, they have to sacrifice some of their revenue; and if financial institutes want to generate more profit, they have to take some more risk. According to our result, one simple way to reweight the balance between profit and risk is to change the ratio of "bad" / "good" observations. If we want to put more emphasis on the "bad" error, more observations of "bad" loans should be included to build the model, and vice versa. Thus, our results indicate that ensemble classifier combined with proper sampling method could achieve desired balance.

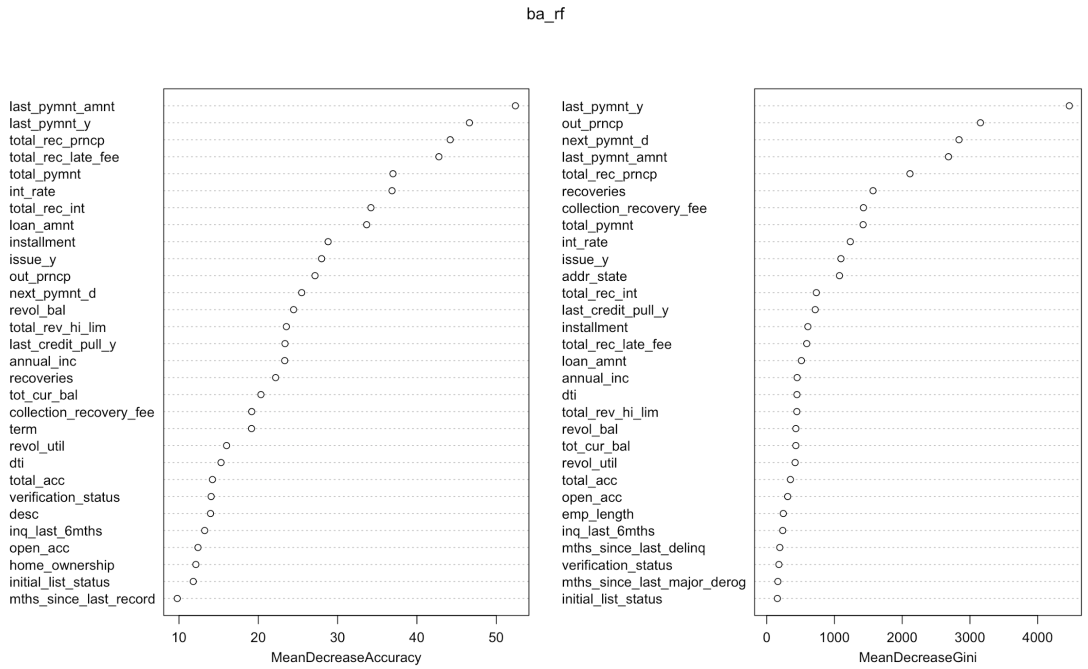

Figure 10: Variable importance in terms of mean decrease accuracy (left); Variable importance in terms of mean decrease Gini (right).

Figure 10 represents the importance of variables in terms of mean decrease accuracy and mean decrease Gini obtained by running random forest on balanced data. According to both impurity measurements, the variables which are considered important here are consistent with the higher branches of decision tree mentioned earlier. Also, we note that the posterior variables (i.e. data collected after issuing the loan) such as "last_pymnt_y", "next_pymnt_d", "last_pymnt_amnt" etc. play important roles in the model. This finding suggests that financial institutions should keep track of these information and update their credit scoring system in order to improve prediction accuracy.

6. Conclusion

We built up several classification models in previous sections, which turned out to be much more efficient than the grading system provided by the dataset itself. Our models have smaller errors in class "good", "bad" and throughout the entire test data. Among the classification models we used, decision tree model performs the best on the unbalanced data in terms of total error rates. Furthermore, SVM performs relatively better than the other models on the balanced data. We also noticed that we can use sampling methods to balance the sample size between two classes by inheriting the distribution from the original data, which inspired that we could further improve a certain type of error using proper sampling methods.

Our logistic model suggests that loan issuers should inspect carefully on borrowers' credit history and debt or obligations he/she is facing. And in general, lending companies have to pay more attention to home renters, since they are more likely to default than landlords. What's more, financial institutions should realize the higher chances of bad loan when the market interest rate is high.

The decision tree and random forest model implies that loan companies are supposed to update credit scoring system according to the latest conditions associated with their borrowers as time passes by and take precautions to minimize their credit risk.

References

[1] https://www.kaggle.com/wendykan/lending-club-loan-data

[2] Z. Huang, H. Chen, C. J. Hsu, W. H. Chen and S. Wu, "Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study," Decision Support System, Vol. 37, No. 4, 2004, pp.543-558.

[3] Y. X. Yang, "Adaptive Credit Scoring with Kernel Learning Methods," European Journal of Operational Resear*ch*, Vol. 183, No. 3, 2007, pp. 1521-1536.

[4] H. S. Kim and S. Y. Sohn, "Support Vector Machines for Default Prediction of SMEs Based on Technology Credit," E*uropean Journal of Operational Research*, Vol.201, No. 3, 2010, pp. 838-846.

[5] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, Vol. 2, No.7, 1936, pp. 179-188

[6] S. L. Lin, "A New Two-Stage Hybrid Approach of Credit Risk in Banking Industry," *Expert Systems with* Applications, Vol. 36, No. 4, 2009, pp. 8333-8341.

[7] Louzada, Francisco, Anderson Ara, and Guilherme B. Fernandes. "Classification methods applied to credit scoring: Systematic review and overall comparison." *Surveys in Operations Research and Management Science* (2016).

[8] Yeh, I-Cheng, and Che-hui Lien. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." *Expert Systems with Applications* 36.2 (2009): 2473-2480.

[9] Twala, Bhekisipho. "Multiple classifier application to credit risk assessment." *Expert Systems with Applications* 37.4 (2010): 3326-3336.

[10] Li, Xiao-Lin, and Yu Zhong. "An overview of personal credit scoring: techniques and future work." (2012).

[11] Emel, Ahmet Burak, et al. "A credit scoring approach for the commercial banking sector." *Socio-Economic Planning Sciences* 37.2 (2003): 103-123.

[12] Morris, Tim P., Ian R. White, and Patrick Royston. "Tuning multiple imputation by predictive mean matching and local residual draws." *BMC Medical Research Methodology* 14 (2014): 75-87.