Default Rate Prediction using Lending Club's Loan Data

By Qiulan Zeng

Oct. 2017

Objective

To predict borrower's default rate (default definition is loan with any late payments.)

1. Describe the data available and identify which features to use or create your own features based on the given attributes.

In the "LoanStats3a.csv" file, there are totally 42539 rows by 137 columns (129 features), which makes this data set very high-dimensional. Therefore, feature selections can be used to reduce overfitting,improve the generalization of models and help us gain a better understanding of the features and their relationship to default rate.

Generally, there are several popular approaches for feature selection, including univariate feature selection, recursive feature elimination, removing features with low variance, and feature selection using SelectFromModel [1]. In this case project, Tree-based feature selection, which belongs to feature selection using SelectFromModel, will be employed.

1.1 Data Description

In this subsection, I will give a brief description and pretreatment of the available data. To start with, some features which have no or only one unique value, such as *pymnt_plan*, *url* and *policy_code*, etc., will be excluded as the input variables. In addition, for simplicity, those features that have many missing values such as *mths_since_last_major_derog* and *mths_since_last_delinq* will not be considered in the first round of analysis. Until now, the total number of initial features is 78.

There are four types of variables: continuous or discrete numerical variables, and ordinal or non-ordinal categorical variables. The types of these variables in Lending Club's Loan Data sheet are summarized in Table 1. Before the prediction, we take the following ways to deal with different types of data.

For ordinal categorical variable, we map the values to discrete numerical values. For example, the variable *Grade* which ranges from "A-G" is mapped from 1 to 7. On the other hand, the non-ordinal categorical variables, such as *home_ownership*, *verification_status* and *addr_state*, are converted into dummy variables. To be more specifically, we create individual columns for each value of the non-ordinal categorical variables, and assign each column a Boolean values 0 or 1. Regarding the numerical variables, we convert the type to "float" if they are of "object" type.

Note that the 153 NA rowsonly accounts for 0.15% of the total 105453 rows. Thus, we drop the rows that contain NA. Next, by plotting Annual Income against Funded Amount, one can

observe that there are two obvious outliers, as shown in Fig. 1. The maximum Annual Income is $ 6, 000, 000. Hence, the Annual Income is limited up to $ 6, 000, 000. Figure 2 demonstrates the Number of Loans against Annual Income with (a) and without (b) the maximum clip in a bar chart. The details of the Number of Loans versus Annual Income are revealed after the clip of maximum of Annual Income. In addition, Figure 3 displays Loan versus Funded Amount with different income maximum without (a) and with (b) the clip. No obvious difference is observed. Therefore, it is reasonable to limit the Annual Income up to $ 6, 000, 000, and the following analysis is based on these filtered data.

The default rate is defined as the rate of borrowers who fail to remain current on their loans. *loan_status* has five category values as shown in Fig. 4(a): Current, Fully Paid, Late(31-120 days), Late(16-30 days), In Grace Period, and Charged off. First, I will remove the Current data, and then split the data into Fully Paid and the remaining as Not Fully Paid. Next, we group them together and convert them to Boolean values (Fully Paid 1 and Not Fully Paid 0), as demonstrated in Fig. 4(b).

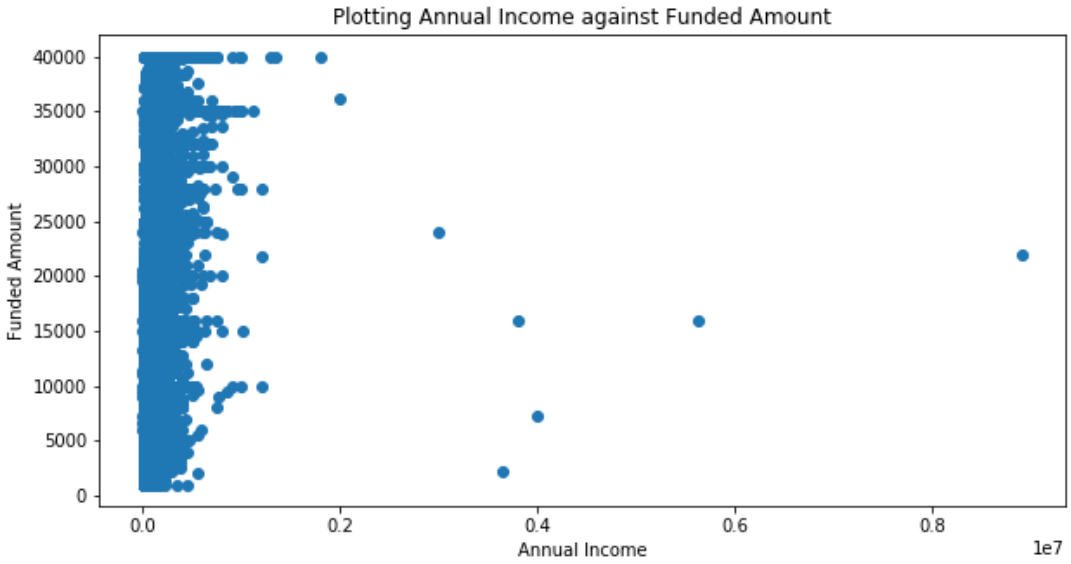
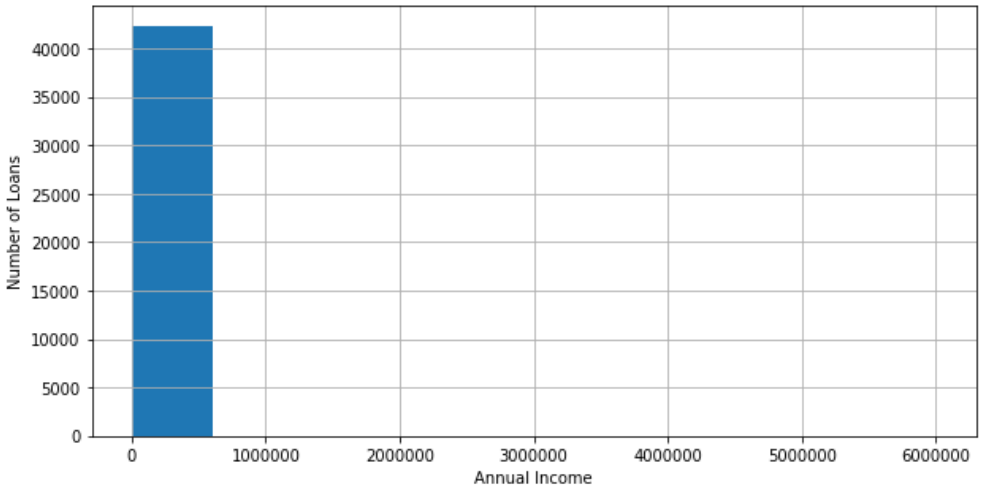
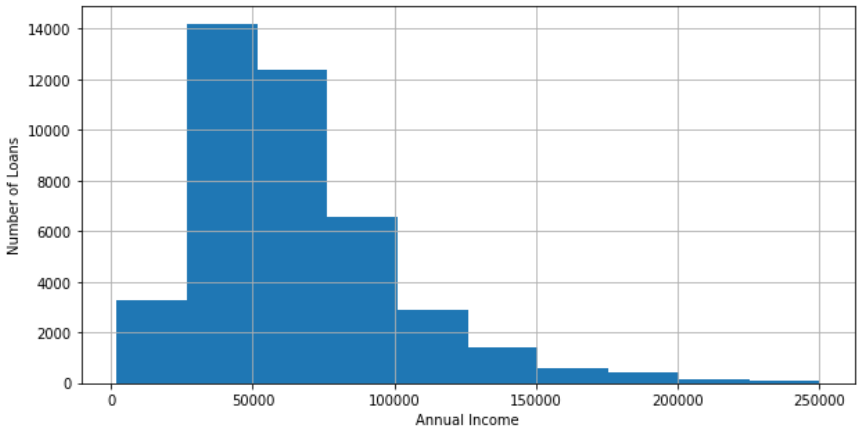


Fig. 1. Funded Amount v.s. Annual Income .

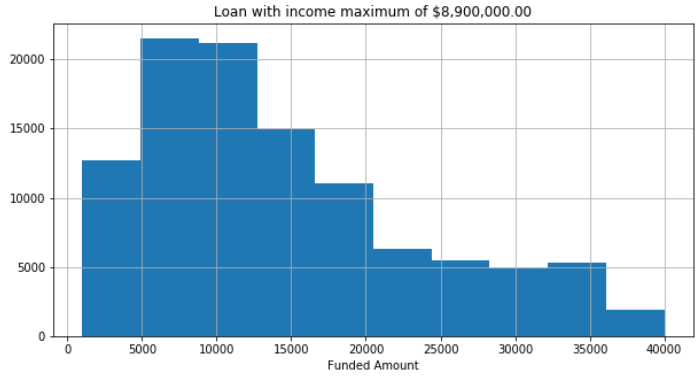


(a) Number of Loans versus Annual Income without clip

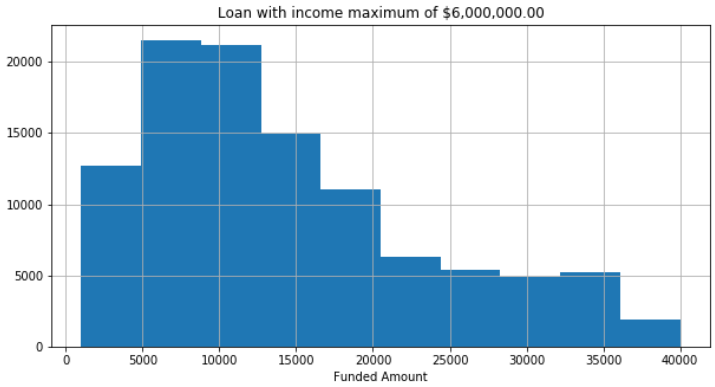


(b) Number of Loans versus Annual Income with clip

Fig. 2. Number of Loans versus Annual Income.

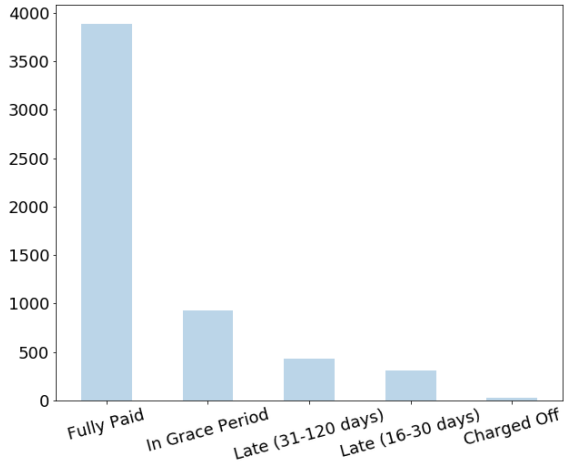


(a) Annual Income without clip

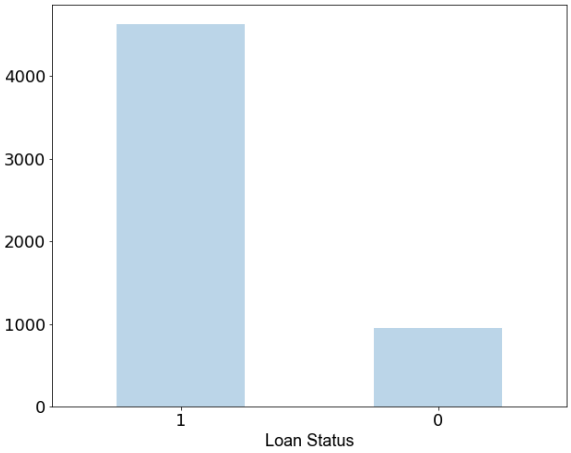


(b) Annual Income with clip

Fig. 3. Loan versus Funded Amount with different income maximum.



(a) Loan Status before grouping.



(b) Loan Status after grouping

Fig. 4. Loan Status.

1.2 Feature Selection

Tree-based feature selection method is used to inform the relative importance of each feature. This method uses ensembles of decision trees such as Extra Trees. “An extra-trees classifier implements a meta estimator that fits a number of randomized decision trees on various sub-samples of the dataset, and use averaging to improve the predictive accuracy and control over-fitting.” [2]. Figure 5 displays the relative importance of the features in our data set. It can be seen that the most ten important features from high level to relative low level are *total_rec_prncp*, *recoveries*, *total_pymnt, collection_recovery_fee, total_pymnt_inv, installment, funded_amnt_inv, loan_amnt, funded_amnt,* and *last_pymnt_amnt*. The other features seem relative small compared to the above-mentioned ten features. Therefore, these 10 features will be employed at the first step for the prediction modelling. In Section 2, I will explore the accuracy difference using 10, 20, and 30 most important features.

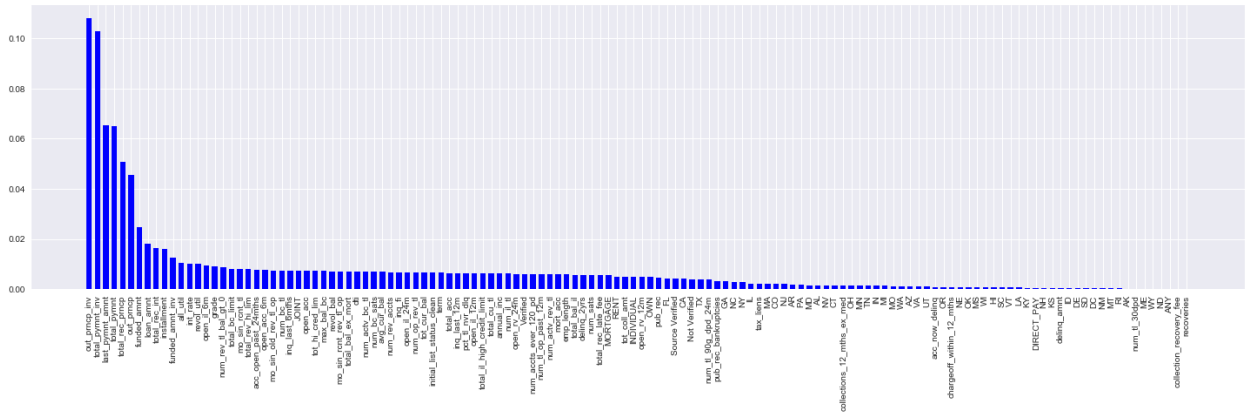


Fig. 5. Feature importance.

2. Pick 2 of the algorithms and implement them in Python/R. (Possible algorithms include but not limited to: Logistic Regression, SVM, Naive Bayes, Linear Regression, Random Forest, Ada Boost, etc.). Compare the two models.

In this case project, Logistic Regression and Random Forest models will be employed. The main implementation process in python is introduced as follows:

Step1. After pretreatment of the variables as described in Section 1, we split the data into train and test dataset using the scikit-learn method *train_test_split* in PYTHON. The selected features data introduced in Subsection 1.2 are regarded as input variables, and *loan_status* as output variable. We impose the test_size to be 0.6 which means that out of the all observations, 60% randomly selected samples are used for training and the remaining 40% for testing. The *train_test_split* provides four outputs: *train_x, test_x, train_y,* and *test_y.*

Step2. Logistic Regression and Random Forest models are applied on the training dataset, *train_x* and *train_y*. After the training, *.score* is used to calculate the mean accuracy with the train data and test data.

Step3. A for-loop is executed to repeatedly perform the above first and second procedure to obtain the mean accuracy of the models. The iteration number is set to be 20.

Step4. Finally, the performance of the Logistic Regression and Random Forest models is compared by classification report and confusion matrix.

The model accuracy results are given in Table 1. From Table 1, one can see that the mean accuracy with both models is high, greater than 85%. In addition, there seem no obvious difference using 10, 20, and 30 important features. Hence, 10 most important features can be employed for prediction. Furthermore, Table 2 reveals that the mean accuracy of both the train and test dataset with Random Forest model is higher that with Logistic Regression model. However, the rate of positive examples is about 83%. Thus, accuracy does not indicate the true performance of the models. Therefore, I will further check other metrics such as precision and recall provided by Table.

Tables 2-4 shows the classification report and confusion matrix. One can see that the recall value 0.83 obtained with the Random Forest model is much higher than 0.47 which is obtained with Logistic Regression model. In addition, the precision with Random Forest model is also higher than that of the Logistic Regression model. Therefore, Random Forest model returns more desirable performance on this dataset.

Table 1. Mean accuracy for Logistic Regression and Random Forest models

models	First 10 important Features		First 20 important Features		First 30 important Features	
	Train dataset	Test dataset	Train dataset	Test dataset	Train dataset	Test dataset
Logistic Regression	0.8745	0.8710	0.8717	0.8664	0.8700	0.8626
Random Forest	0.9138	0.8925	0.9159	0.8912	0.9170	0.8891

Table 2. Classification report for Logistic Regression model and Random Forest model

		precision	recall	F1-score	support
Logistic Regression model	0	0.65	0.47	0.54	956
	1	0.90	0.95	0.92	4632
	Avg/total	0.91	0.90	0.90	5588
Random Forest model	0	0.67	0.83	0.74	956
	1	0.96	0.91	0.94	4632
	Avg/total	0.91	0.90	0.90	5588

Table 3. Confusion matrix for Logistic Regression model

446	510
243	4389

Table 4. Confusion matrix for Random Forest model

790	166
396	4236

3. What are the most important features according to your models?

Table 4 lists the coefficients of the features we selected for Logistic Regression model. The following features have positive values: *total_pymnt_inv*, *out_prncp*, *out_prncp_inv*, *installment* and *last_pymnt_amnt*; whereas the coefficients of *total_rec_int*, *total_rec_prncp*, *funded_amnt* and *loan_amnt* are negative. Positive coefficients mean that the change of the loan being fully paid will increase with the corresponding features. On the contrary, negative coefficients indicate that the probability of being charged off will increase with the corresponding features. Furthermore, we should especially pay attention to the negative coefficients with high absolute value. As shown in Table 4, *total_rec_int* has the smallest negative coefficient. In other words, the larger the total interested received to date to that loan at that point in time, the higher risk of default. On the other hand, *total_pymnt_inv* has the largest positive coefficients, meaning with larger payments received to date for portion of total amount funded by investors, the loans tend to be much less riskier. To sum up, *total_rec_int* and *total_pymnt_inv* are two most important features for Logistic Regression model.

Table 4. Coefficients with Logistic Regression model

out_prncp_inv	0.002795
total_pymnt_inv	0.031947
last_pymnt_amnt	0.000930
total_pymnt	0.031540
total_rec_prncp	-0.034656
out_prncp	0.025685
funded_amnt	-0.014243
loan_amnt	-0.014243
total_rec_int	-0.065985
installment	0.000617

Table 5 shows the importance order of the features we selected for Random Forest model. It can be seen that the most important features for this model in this case are *out_prncp* and *total_rec_prncp*.

Table 5. Coefficients with Random Forest model

```
out_prncp_inv     0.119655
total_pymnt_inv   0.069954
last_pymnt_amnt   0.176330
total_pymnt       0.108585
total_rec_prncp   0.209569
out_prncp         0.236415
funded_amnt       0.007389
loan_amnt         0.010461
total_rec_int     0.037596
installment       0.024046
```

4. If you were to predict default rate for the next 5 years, describe what information (data) you will need and which methodology you will use.

To predict default rate for the next 5 years, Random Forest model can be used based on the above analysis. To obtain desirable prediction performance, I have to acquire enough historical samples of the important features illustrated in Section 3. For example, I can train the Random Forest model using the data of year 2000 as one input sample to predict 2005's default rate, and then using 2001's data to predict 2006's default rate, and then repeat the process until 2012. Then I will get the coefficients for the Random Forest model. What I have to do is to apply different size of sample data to plot the learn curve for testing the Bias vs Variance. By doing this, I can determine the size of dataset and features to optimize my Random Forest model to predict default rate for the next 5 years. An alternative is to first predict the input variables for the next 5 years, and then use these estimated input features to predict the default rate by the adoption of the Random Forest model in Section 3.

5. What business insights could you offer based on the "Declined Loan " datasets made available on the site.

5.1 Feature statistic distribution analysis

The statistic description of Risk Score is given in Table 6. Figs. 6-8 display the density distribution of the Amount Requested, Risk Score, and Debt-to-Income, respectively. All of these features are right skewed. Fig 9. shows the Employment Length histogram. It can be seen that among all declined loan, those with one year Employment Length, instead of with zero Employment Length, gets the highest chance to be declined. Interestingly, the second highest rejected Employment Length is five years.

Table 6. Statistic description of Amount Requested, Risk Score, Debt-to-Income Ratio, and Employment Length

	count	mean	std	min	25%	50%	75%	max
Amount Requested	1,665,309	12,858	14.882	0	40,000	10,000	20,000	30,000
Risk Score	643425	633.07	66.73	300.00	590.00	633.00	673.00	990.00
Debt-to-Income Ratio	1,665,309	13,716.64	2.07	-0.01	0.069	0.19	0.35	11,435.9
Employment Length	1,665,309	1.92	1.91	0.00	1.00	1.00	1.00	10.00

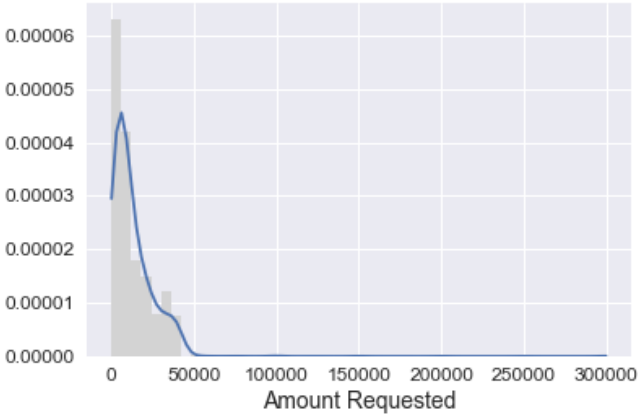


Fig. 6. Amount Requested distribution.

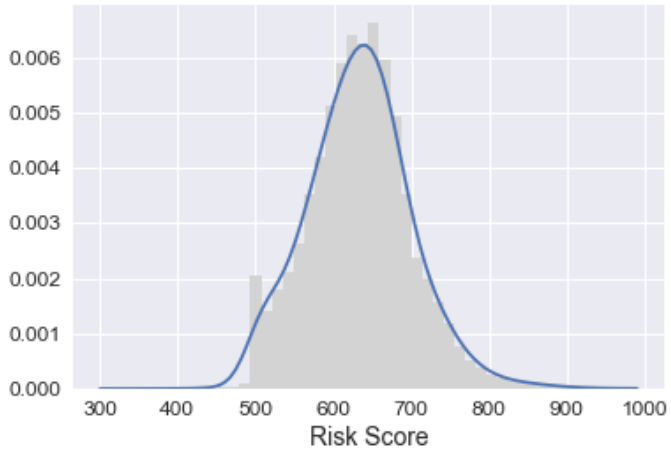


Fig. 7. Risk_Score distribution.

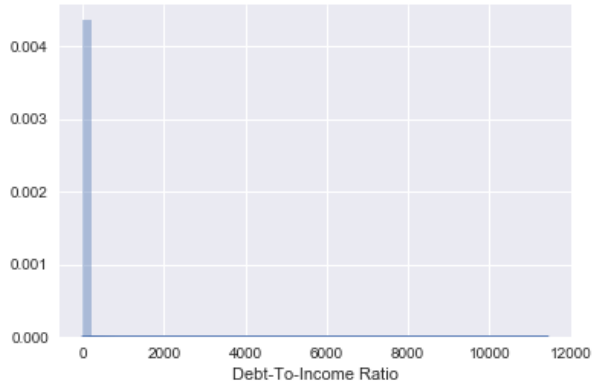


Fig. 8. Debt to Income Ratio distribution.

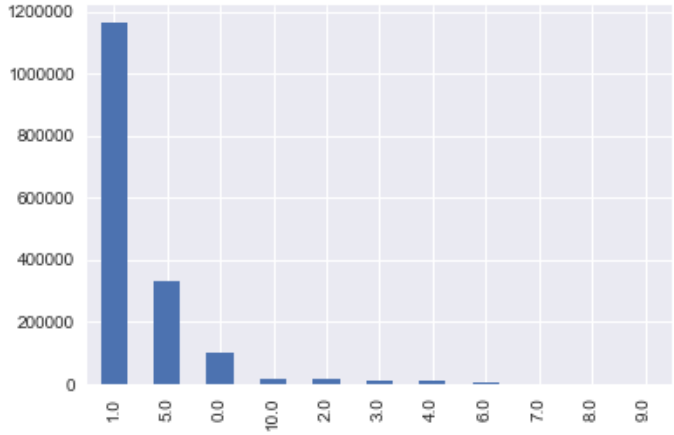
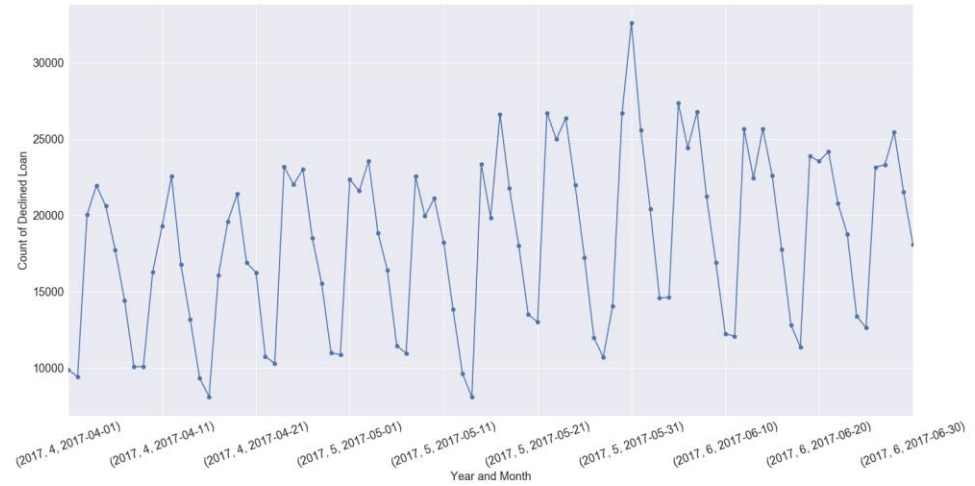


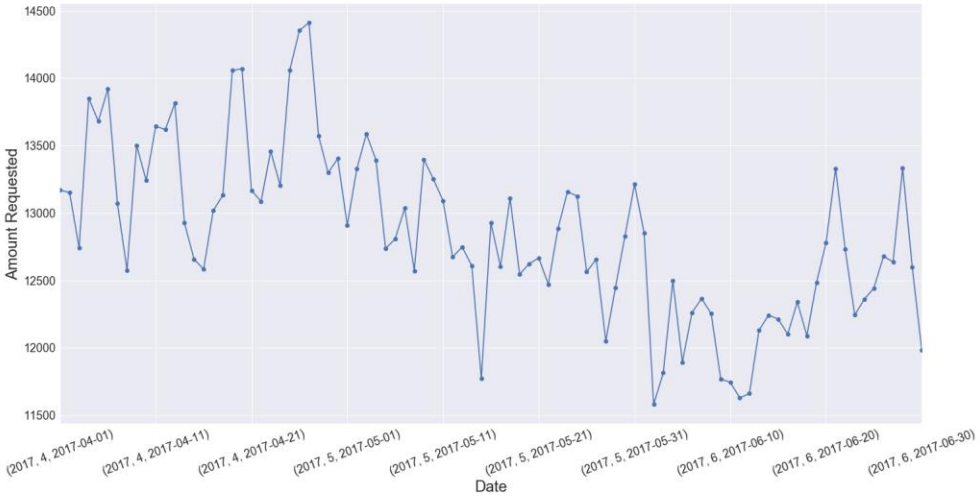
Fig. 9 Employment Length distribution.

5.2 Features over time

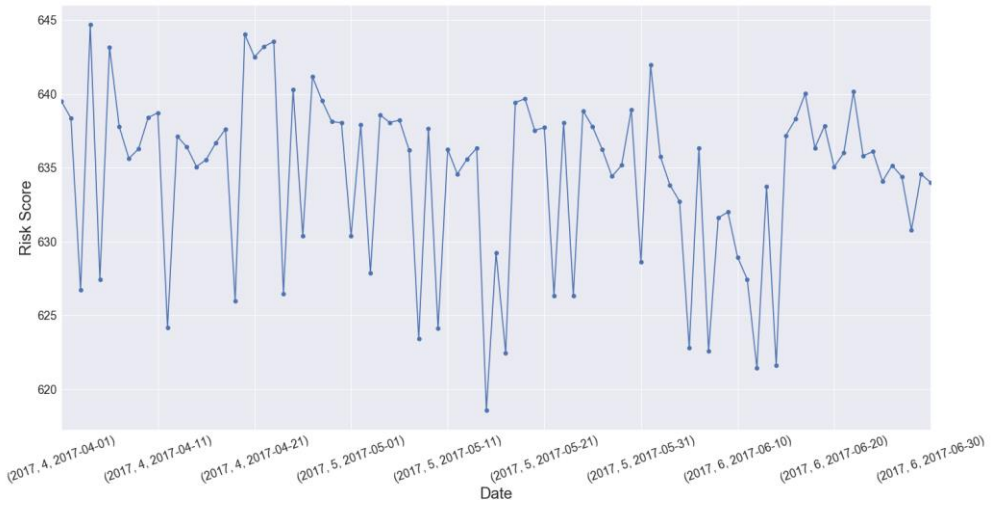
Figs. 10 (a)- (e) show the mean value of Declined Loan number, Amount Requested, Risk Score, Debt-to-Income, and Employment Length, respectively, every day from April to June in 2017. From Fig 10. (a), it can be seen that the declined loan number varies approximately on a 7 day cycle; whereas there seem no obvious rhythm for the other features. Therefore, I take the average of the above-mentioned features' values on a month unit, and the results are provided in Table. One can observe that from April to June, the Amount Requested is decreased, while the Declined Loan Count and Debt-to-Income Ratio is increased.



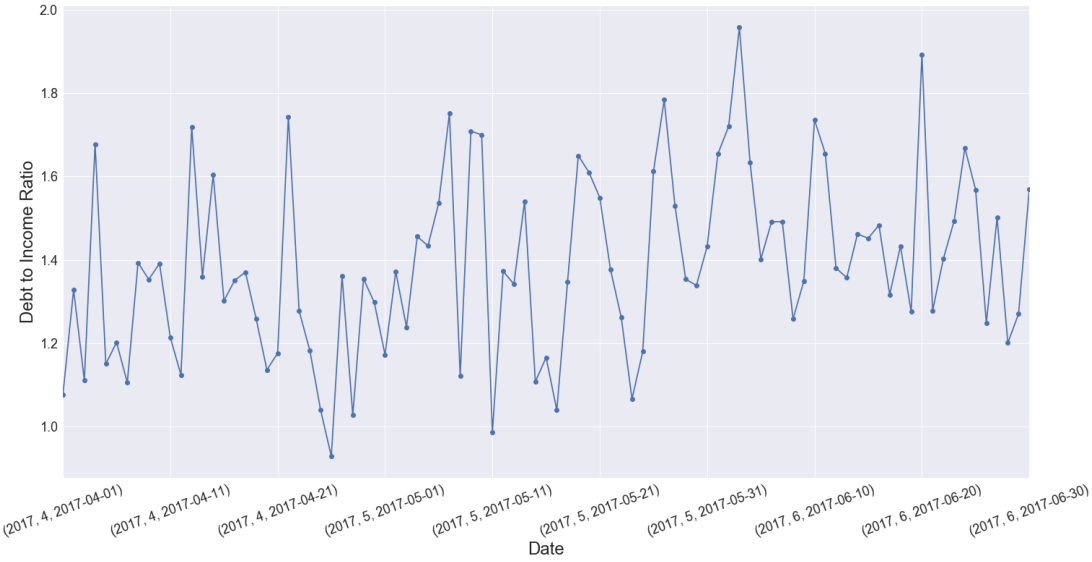
(a) Count of Declined Loan.



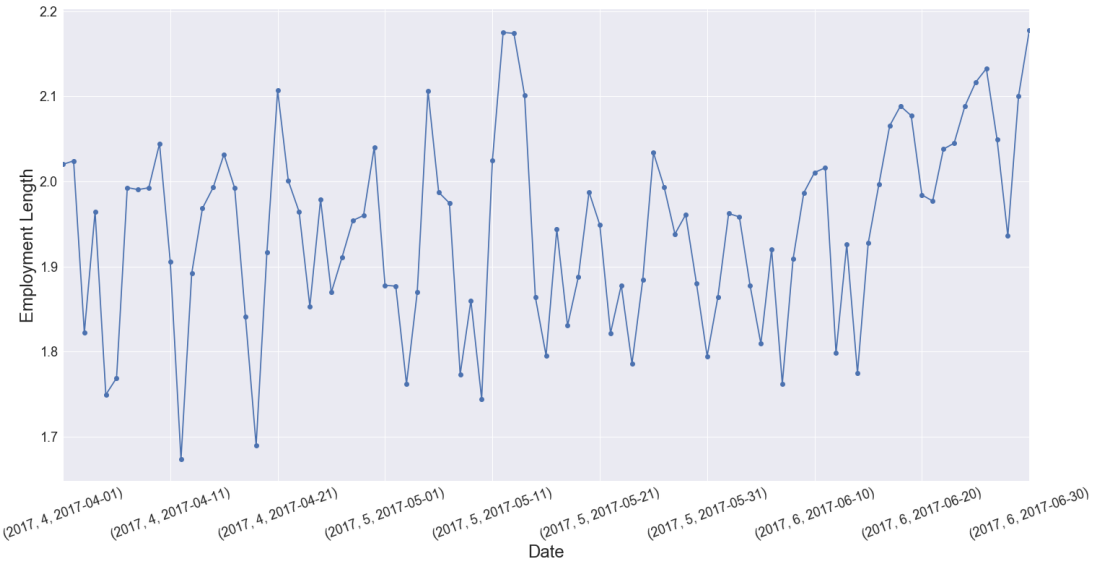
(b) Amount Requested.



(c) Risk Score.



(c) Debt-to-Income Ratio



(d) Employment Length Ratio

Fig. 10. Date time development of Amount Requested (a), Risk Score (b), Debt-to-Income (c), and Employment (d) from April to June in 2017.

Table 7. Quantitative description of Declined Loan count, arithmetic averaged Amount Requested, Risk Score, Debt-to-Income, and Employment from April to June in 2017.

	Declined Loan	Amount Requested ($)	Risk Score	Debt-to-Income Ratio	Employment Length (year)
April	474768	13502.87	635.23	1.27	1.91
May	587641	12867.08	631.94	1.36	1.89
June	602900	12342.21	632.47	1.47	1.96

5.3 Feature Analysis for States

Figure 11 shows the Declined Loan count for all the states. It is demonstrated that CA, TX, FL, and NY have the largest declined loan number; whereas the total number of loan declined for WV is extremely low. Table 8 summarizes the quantitative description of Declined Loan count for the above-mentioned states. The relative number denotes the Declined Loan number of a state relative to the average Declined Loan number in a certain month. From Table 8, we can see that the relative values of Declined Loan in all the three months are below 0.2%.

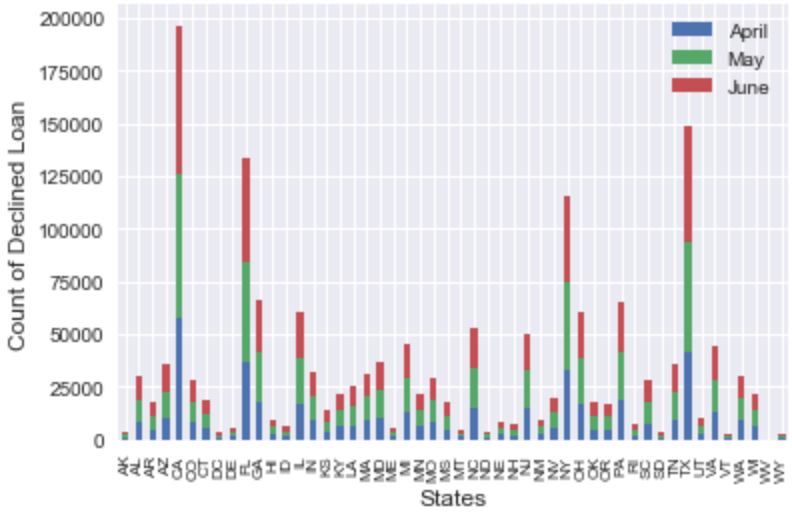


Fig. 11. Count of Declined Loan against States from April to June, 2017.

Table. 8. Declined Loan count for CA, TX, FL, NY, and WV

	Absolute value of Declined Loan			Relative value of Declined Loan		
	April	May	June	April	May	June
Mean	909	998	1037	1	1	1
CA	57745	68617	70230	6.08	5.84	5.82
TX	41265	52413	55280	4.35	4.46	4.58
FL	37305	47231	49293	3.93	4.02	4.09
NY	33387	41520	40275	3.52	3.53	3.34
WV	15	16	14	0.0016	0.0014	0.0012

Figure 12 shows the Amount requested by different states. It is obvious that WV has the highest Amount Requested in all these three months, about 2 to 4 times as much as that of other states; whereas the Amount Requested by other states is on the same order.

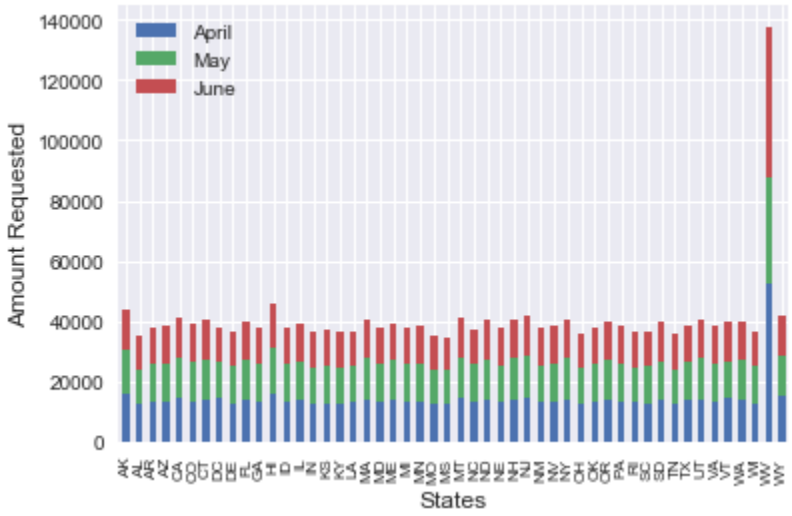


Fig. 12. Amount Requested against States from April to June, 2017.

Figure 13 presents the Average Employment Length reported by different States from April to June in 2017. It can be seen that the Average Employment Length is on the same level among the states in these three months, except for WV, whose value is at a half of that of other states.

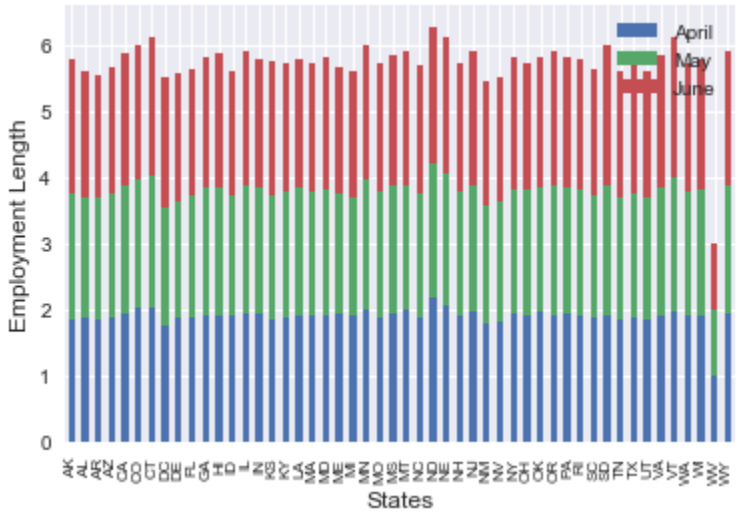


Fig. 13. Average Employment Length reported by States from April to June, 2017.

Figure 14 displays the Debt-to-Income ratio of the states from April to June in 2017. It can be seen that this feature is on the same level in April for all the states except for RI, WY, and WV. In April, RI and WY has very high Debt-to-Income ratio relative to the other states. Particularly, this feature in April of RI is around 6 times larger than that of other states. However, this value goes down in the following two months. It should be noted that the Debt-to-Income ratio of WV is negative from April to June in 2017, as shown in Table 9.

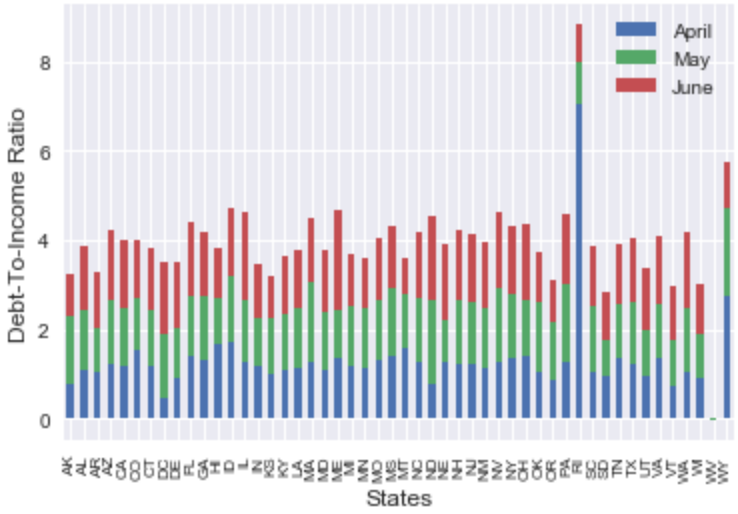


Fig. 14. Debt-to-Income Ratio against States from April to June, 2017.

Table. 9. Debt-to-Income ratio of WV

Month	April	May	June
Debt-to-Income ratio	-0.01	-0.01	-0.01

It is shown in Fig. 15. that the Risk Score is similar for all the states. Note that Risk Score values are missing for WV state.

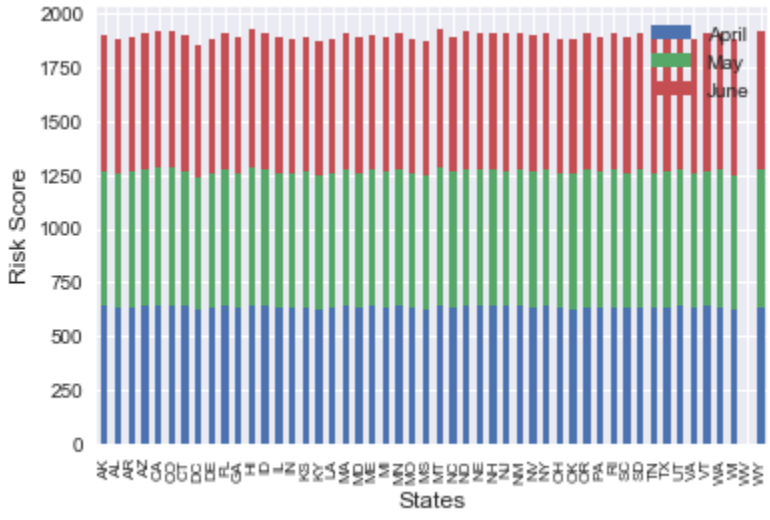


Fig. 15. Risk Score against States from April to June, 2017.

Figure 16 demonstrates the Reported Annual Income during registration from April to June, 2017. It can be seen that in April, the Reported Annual Income is getting more heterogeneous among States as time moves on.

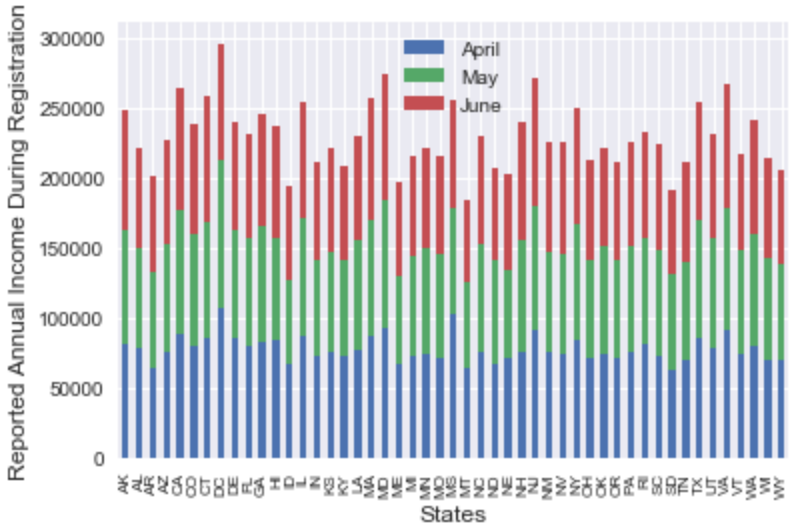


Fig. 16. Reported Annual Income during registration against States from April to June, 2017.

5.3 Declined Loan Count versus Loan Title

It can be seen from Fig. 17 that from April to June in 2017, loans with Loan Title of “Debt consolidation” have the largest declined count, and with “Other” the second largest. Hence one can conclude that it is easy to be declined with the loan title “Debt consolidation” or “Other”. On the contrary, loans with title “Green loan” and “Renewable energy” have the least declined frequency. This indicates that the Lending Club is likely to finance toward green entrepreneurship or energy improvement.

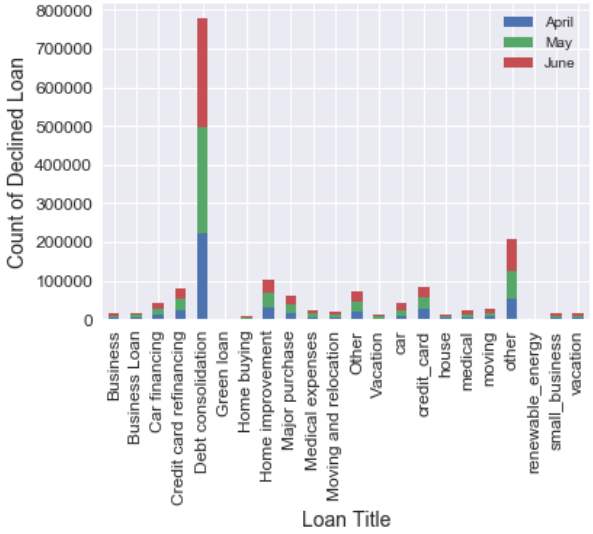


Fig. 17 Count of Declined Loan against Loan Title from April to June, 2017.

6. Conclusions and discussion

This case project uses Tree-based feature selection method to identify the importance of the features, and selects the most important 10, 20, and 30 features for the simulation. After that, Logistic Regression model and Random Forest model were employed to predict the default status with the selected features as input variables. To perform the simulation, the sample data were filtered and split into training and test datasets. Next, a for-loop is conducted with the train-test process to estimate the coefficients of the two models. Then, Logistic Regression model and Random Forest model are compared by analyzing the metrics such as accuracy, precision, and recall. After that, I gave a brief introduction of the information and methodology I would use for the prediction of default rate for the next 5 years. Finally, the "Declined Loan" datasets were analyzed from four aspects, including features' statistical distribution, features' variation with time, features' characteristics for the states, and the Declined Loan Count versus Loan Title. Based on the analysis, we can offer the following business insights:

(1) loans with one year Employment Length and loans with Loan Title of "Debt consolidation" or "Other" has very high chance to be declined;
(2) Loans from West Virginia (WV) state are least likely to be rejected;
(3) The Lending Club is likely to finance toward green entrepreneurship or energy improvement.

Reference

[1] http://scikit-learn.org/stable/modules/feature_selection.html

[2] http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html#sklearn.ensemble.ExtraTreesClassifier

Appendix

Data Type	Variables	Description
Continuous numerical	loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
	funded_amnt	The total amount committed to that loan at that point in time.
	funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
	int_rate	Interest Rate on the loan.
	installment	Months since oldest bank installment account opened.
	annual_inc	The self-reported annual income provided by the borrower during registration.
	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.
	revol_bal	Total credit revolving balance.
	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
	total_pymnt	Payments received to date for total amount funded.
	total_pymnt_inv	Payments received to date for portion of total amount funded by investors.
	total_rec_prncp	Principal received to date.
	total_rec_int	Interest received to date.
	total_rec_late_fee	Late fees received to date.
	recoveries	post charge off gross recovery.
	collection_recovery_fee	post charge off collection fee.
	last_pymnt_amnt	Last total payment amount received.
Discrete numerical	Term	The number of payments on the loan. Values are in months and can be either 36 or 60.
	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
	delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
	inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries).
	open_acc	The number of open credit lines in the borrower's credit file.
	pub_rec	Number of derogatory public records.
	total_acc	The total number of credit lines currently in the borrower's credit file.
Ordinal categorical	Grade	LC assigned loan grade.
Non-ordinal categorical	home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER.
	verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified.
	loan_status	Current status of the loan.
	addr_state	The state provided by the borrower in the loan application.
	application_type	
	initial_list_status	

Table 1. Types of variables.