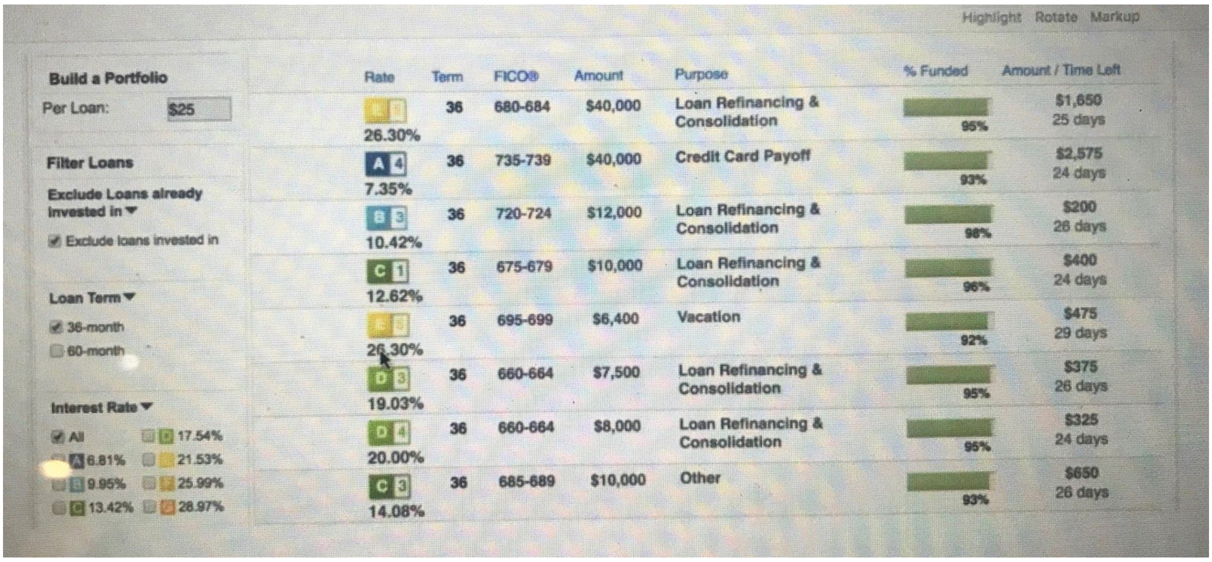
FinTech Smart Investment Prediction Project Report

—— Xingwen Niu

1 Definition

1.1 Project Background

Lending Club is the trailblazer in peer-to-peer lending, it has evolved into America’s largest online marketplace that allows borrowers to apply for personal loans, auto refinancing, business loans, and elective medical procedures. Lending Club provides two ways of investing: manually investing and automatically investing. In Figure 1 and Figure 2 we show the interface of the two ways.



Build a Portfolio

Per Loan: $25

Filter Loans

Exclude Loans already invested in

☑ Exclude loans invested in

Loan Term

☑ 36-month

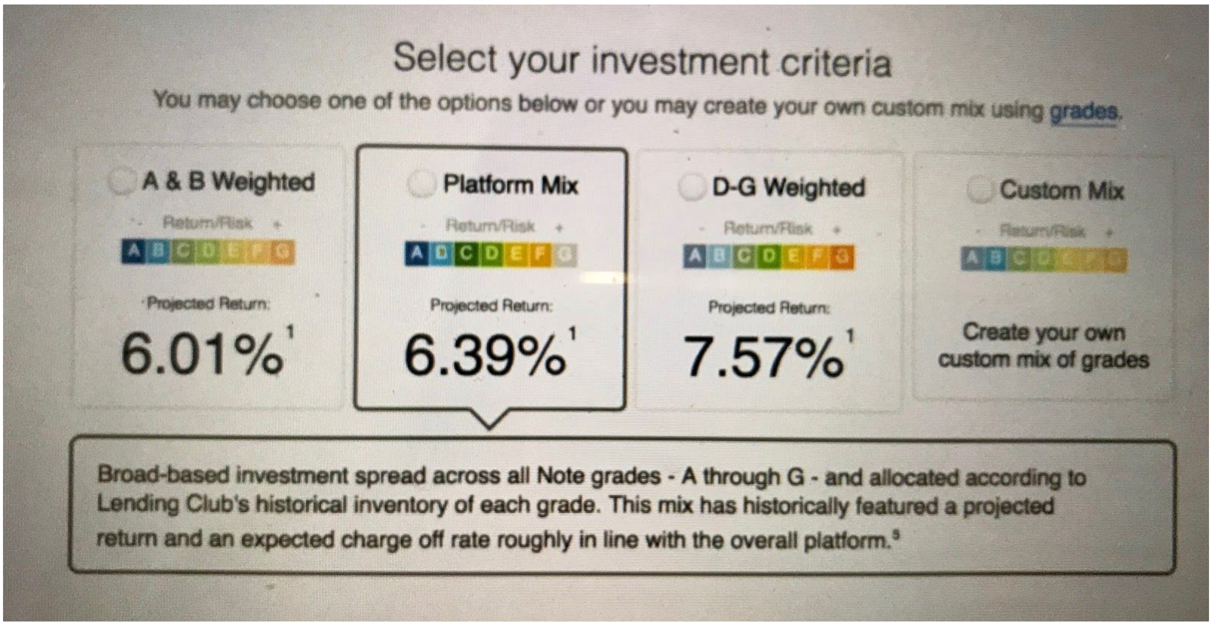
□ 60-month

Interest Rate

☑ All	□ D 17.54%
□ A 6.81%	□ 21.53%
□ B 9.95%	□ 25.99%
□ C 13.42%	□ 28.97%

Rate	Term	FICO®	Amount	Purpose	% Funded	Amount / Time Left
26.30%	36	680-684	$40,000	Loan Refinancing & Consolidation	95%	$1,650 25 days
A 4 7.35%	36	735-739	$40,000	Credit Card Payoff	93%	$2,575 24 days
B 3 10.42%	36	720-724	$12,000	Loan Refinancing & Consolidation	98%	$200 26 days
C 1 12.62%	36	675-679	$10,000	Loan Refinancing & Consolidation	96%	$400 24 days
26.30%	36	695-699	$6,400	Vacation	92%	$475 29 days
D 3 19.03%	36	660-664	$7,500	Loan Refinancing & Consolidation	95%	$375 26 days
D 4 20.00%	36	660-664	$8,000	Loan Refinancing & Consolidation	95%	$325 24 days
C 3 14.08%	36	685-689	$10,000	Other	93%	$650 26 days

Figure 1. Current Loans



Select your investment criteria

You may choose one of the options below or you may create your own custom mix using grades.

○ A & B Weighted	○ Platform Mix	○ D-G Weighted	○ Custom Mix
- Return/Risk +	- Return/Risk +	- Return/Risk +	- Return/Risk +
A B C D E F G	A B C D E F G	A B C D E F G	A B C D E F G
Projected Return: 6.01%[1]	Projected Return: 6.39%[1]	Projected Return: 7.57%[1]	Create your own custom mix of grades

Broad-based investment spread across all Note grades - A through G - and allocated according to Lending Club's historical inventory of each grade. This mix has historically featured a projected return and an expected charge off rate roughly in line with the overall platform.[5]

Figure 2. Automated Investing

Lending Club contains hundreds of loans which makes it difficult for investors to choose a profitable one. In this project, I design a system with a machine learning engine as a smart investment prediction tool, helping investors identify the values of different loans in Lending Club, to determine the optimal loans to invest in.

A simple web page is also designed to implement the dynamic interaction between investors and our system and investors. When an investor views the loans on the platform, he/she only needs to input the key features, the system can analyze the loan's parameters and screen out the best ones for the investor accurately and quickly.

web page

1.2 Problem Statement

The centerpiece of the system is using machine learning algorithms to choose loans to invest in. There are many ways to evaluate the performance of a loan, for example, ROI, the return on investment or the risk level of the investment. Regression models can be applied to predict ROI based on the historical data and classification algorithms are able to predict whether the status of a loan will be charged-off or default. With a reliable prediction, it becomes possible for investors to pick up the good loans to invest.

In this project, we will predict the status of the loans. That is, it is handled as a binary classification problem. We are more interested in predicting the probability of charged-off or default. The algorithms used are **RandomForest** and **XGBoost**, which is a common example of boosting methods. We have seen in many Kaggle competitions that XGBoost outperforms with so many advantages, for example, its nature of handling of data with heterogeneous features, strong predictive power, robustness to outliers in output space, ability to figure out important features and so on. After training and evaluating an optimal model, we will save it using pickle and develop a web interface using **Flask**.

Meanwhile, XGBoost can generate importance score for all training columns. We can select top 10 important features which are meaningful and easier to explain to use as input features. It not only greatly reduces the work of investors but also provides accurately, efficiently predictions.

Framework

1.3 Metrics

For classification problems, accuracy is an important metric to evaluate the performance of models. However, for our problem, it is not enough. On one hand, if a “bad” loan is predicted to be a “good" loan, the investors may lose a lot of money and it will affect the credibility of our product. On the other hand, if we are too conservative to predict a loan to be a good one, then this will reduce the amount of transitions on our platform. So considering the two aspects together, metrics like AUC and F-ß score are more appropriate to be applied to evaluate the performance of our models.

For readers who are interested in the mathematical definitions of those metrics, we show them here. Accuracy is defined to be the percentage of the correct predictions among all predictions. Precision is the percentage of the correct predictions among all cases which are predicted to be positive (charged-off). Recall is the percentage of correct prediction among all cases whose true labels are positive (charged-off). We also use the **F1 score** (also F-score or F-measure) to evaluate the performance. F1 score considers both the precision and the recall. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0:

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2 * P * R}{P + R}$$

2 Modeling Process

In the following sections, we will demonstrate the modeling process. This process is put into 5 separate experiments as shown in Figure 3.

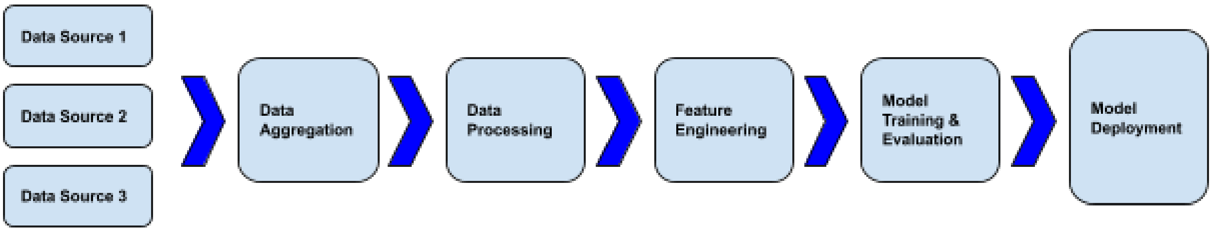


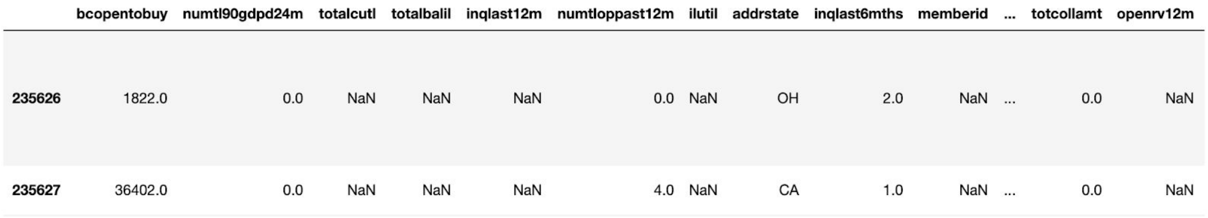
Figure 3: Modeling Process

2.1 Data Sources

The datasets used in this project are from the official website of Lending Club. We downloaded the data of loans launched in 2014(37646 kb) and obtained the current data of 7/13/2017 in **JSON** format via API (A personal account has to be created).

The data in 2014 contains complete data for all loans issued through the year, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. There are 143 features and 235,629 rows. The definitions for all the data attributes included in the historical data file can be found in the Data Dictionary. The reason why we chose the data in 2014 is that there are two types of loans: ‘term 36 months’ and ‘term 60 months’. For this project, we only build the model based on the loans with a term of 36 months. Most of the loans issued in 2014 are to be mature for the moment, and we have more information about them.

The current data is about the loans that currently are to be funded. There are 103 features inside. Here is the sample data.



	bcopentobuy	numtl90gdpd24m	totalcutl	totalbalil	inqlast12m	numtloppast12m	ilutil	addrstate	inqlast6mths	memberid	...	totcollamt	openrv12m
235626	1822.0	0.0	NaN	NaN	NaN	0.0	NaN	OH	2.0	NaN	...	0.0	NaN
235627	36402.0	0.0	NaN	NaN	NaN	4.0	NaN	CA	1.0	NaN	...	0.0	NaN

Figure 4: Sample Data

2.2 Data Preprocessing

Our goal is to make predictions for the current loan notes based on the model trained using the historical dataset. So it is necessary that the current dataset and the historical dataset have the same set of features. From the names and the meanings of the features, it turns out that there are 95 common features in both datasets.
Although the features “issued” and “loan status” are not available in the current dataset, they are useful in the later process, we will keep them. Feature “issued” indicates the month in which the loans were published, so we will use it to split the training and testing dataset. Feature “loan status”, as the name suggests, is the status of the loans when they are mature, which acts as our prediction target.
Among the 95 columns, there are 19 object features, 76 numeric features. But 18 columns of them are Nulls (which miss all the values through all observations), and 17 columns have Nulls. Remove the all-null features and the features with only one value, such as term, application_type. Drop duplicated feature fundedamnt (same to loanamnt). There are 59 numerical features and 18 categorical features remained. Figure 5 illustrates the information of the features.

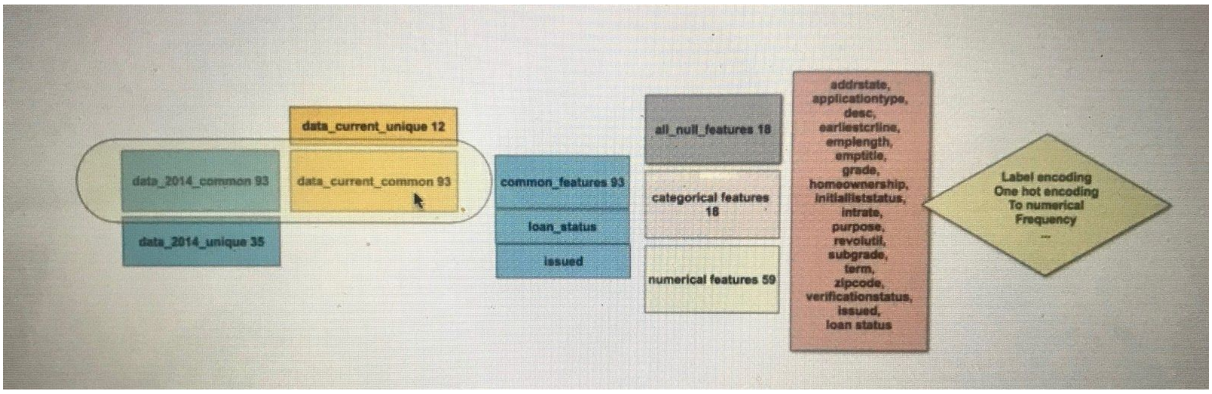


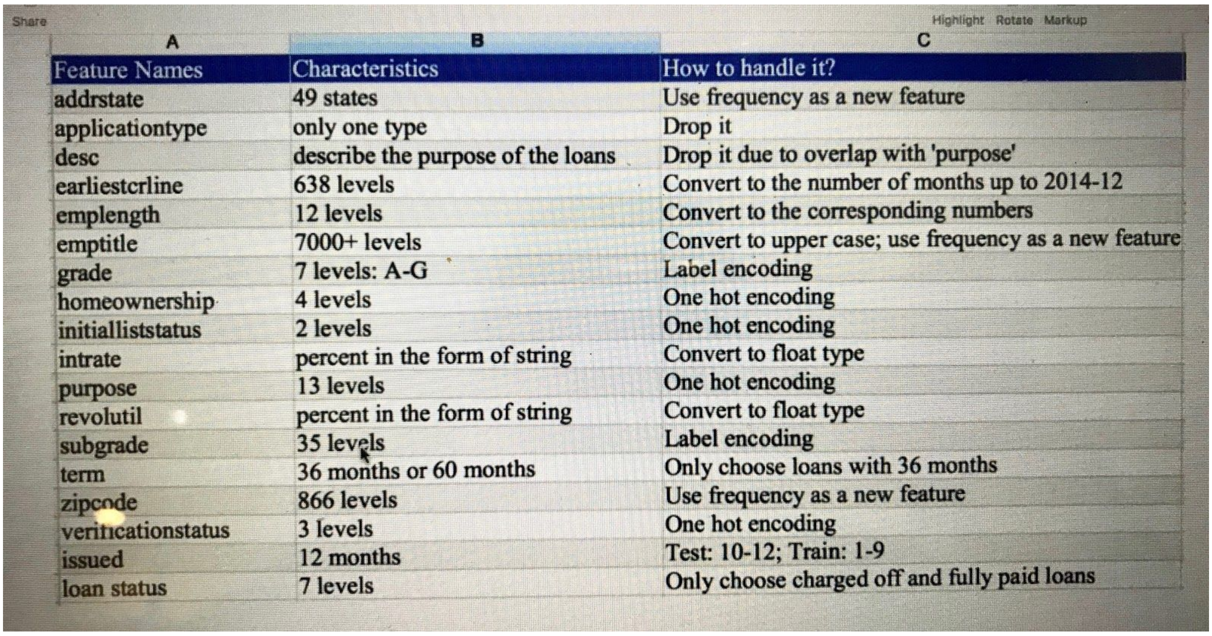
Figure 5: Features

2.2.1 Categorical Features

The 18 categorical features are handled in the way shown in Figure 6. We make some further remarks here.
For employment length, there are 12 levels. For levels like “n years” where n is from 1 to 9, we convert them to n. For level “< 1 year”, it is represented by 0 and for level “> 10 years”, it is converted to be 10. There are also some rows with “NA”, it is very likely that they don’t have jobs. This increases the probability that the loan will be charged off. So we set the employment length for such rows to be -9999 to distinguish from the other levels.

For the feature “zipcode”, there are 866 levels. One way is to use the frequency of the 866 levels to represent the levels, as what we did in this project. Another possible way to deal with zipcode is to clustering the areas according to the loan status. It should work well if we can take care of the risk of information leakage.

Aslo for the feature “desc”, natural language processing can be considered to extract more information. But since in the current dataset, all rows miss the value of “desc”, we just drop this feature.


Share
Highlight
Rotate
Markup
A
B
C

Feature Names	Characteristics	How to handle it?
addrstate	49 states	Use frequency as a new feature
applicationtype	only one type	Drop it
desc	describe the purpose of the loans	Drop it due to overlap with 'purpose'
earliestcrline	638 levels	Convert to the number of months up to 2014-12
emplength	12 levels	Convert to the corresponding numbers
emptitle	7000+ levels	Convert to upper case; use frequency as a new feature
grade	7 levels: A-G	Label encoding
homeownership	4 levels	One hot encoding
initialliststatus	2 levels	One hot encoding
intrate	percent in the form of string	Convert to float type
purpose	13 levels	One hot encoding
revolutil	percent in the form of string	Convert to float type
subgrade	35 levels	Label encoding
term	36 months or 60 months	Only choose loans with 36 months
zipcode	866 levels	Use frequency as a new feature
verificationstatus	3 levels	One hot encoding
issued	12 months	Test: 10-12; Train: 1-9
loan status	7 levels	Only choose charged off and fully paid loans

Figure 6. Categorical features

2.2.2 Numerical Features

There are some numerical features having missing values. But since the XGBoost has the nature of dealing with the missing values, we won’t fill the missing values or drop them. The details of the missing values are shown in Figure 7 and Figure 8.

Some features have outliers, which will be considered in the future work.

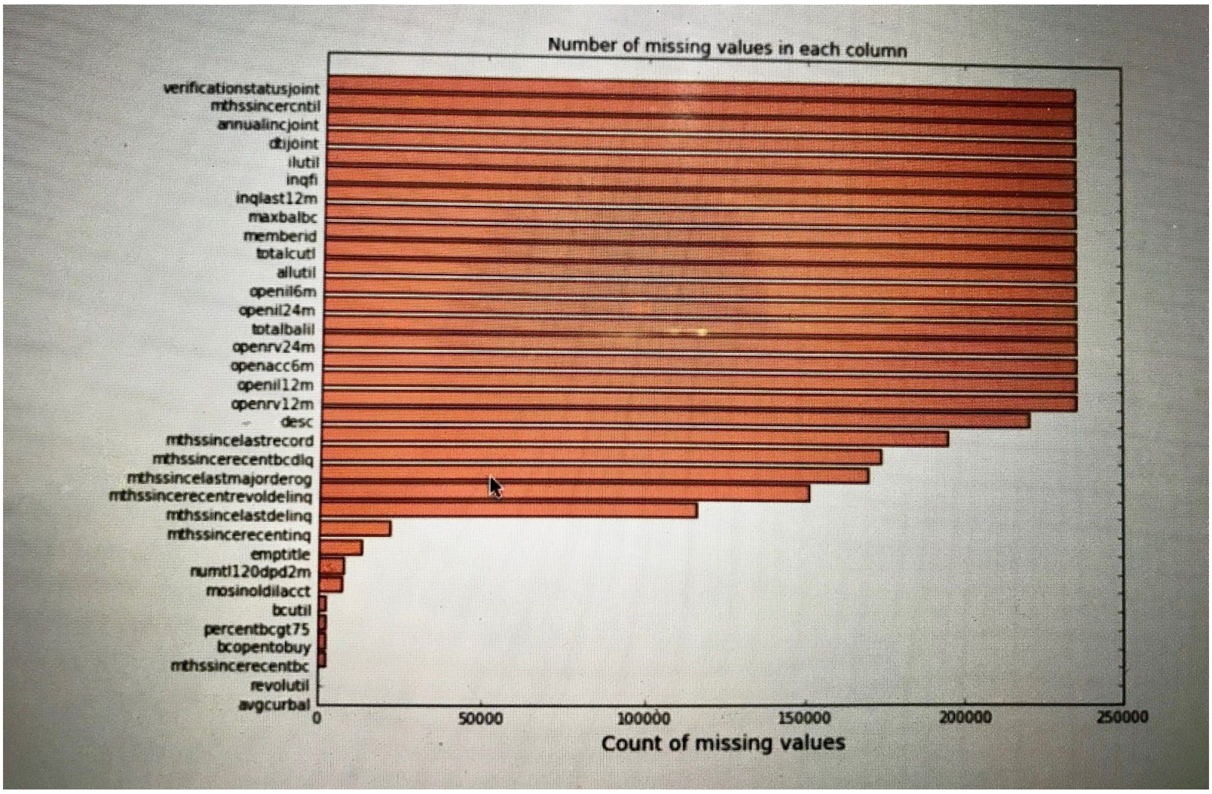


Figure 7: Features with missing values in data of 2014

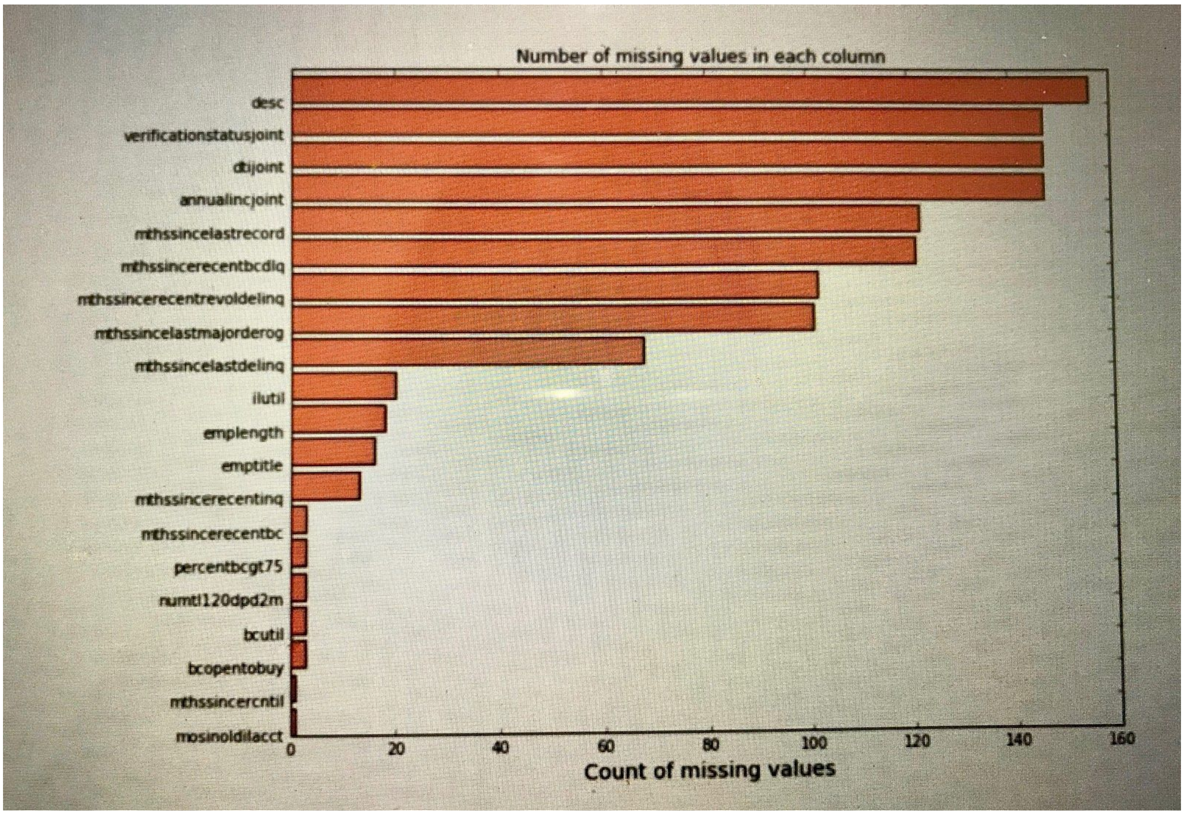
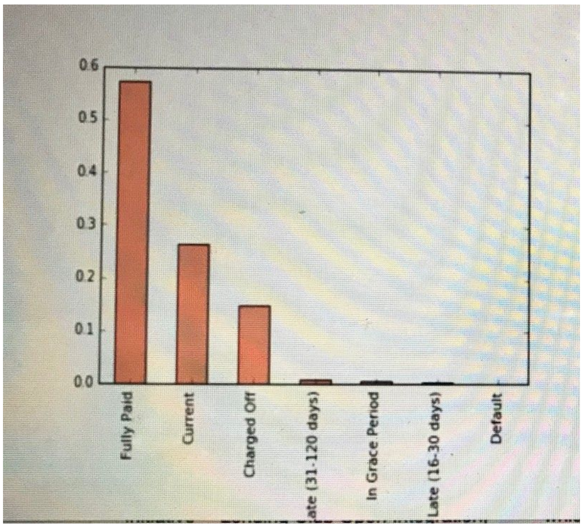


Figure 8: Features with missing values in the current dataset

2.3 Data Exploration and Visualization

In this section, we show some interesting discoveries when we explore the data.
In Figure 9, one can find that nearly 60% of loans are fully paid when we downloaded the data of 2014. There are about 15% loans that are ‘charged off’.
As for the purpose of the loans, one can find in Figure 10 that more than 70% of them are used for debt consolidation and credit cards. Common sense tells us that people who borrow money to consolidate debt or to pay their credit cards are less likely to leave their loans unpaid.



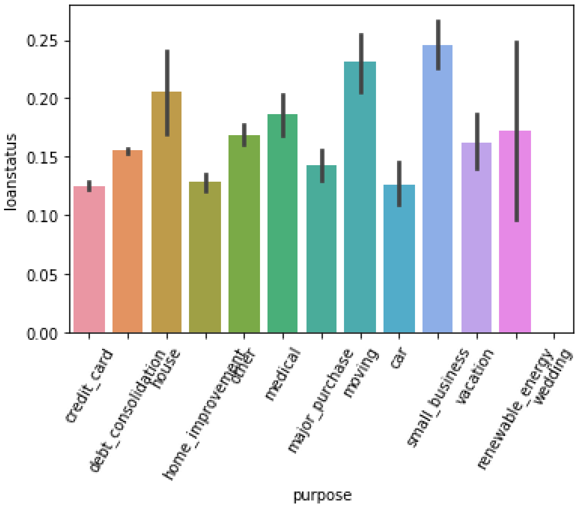
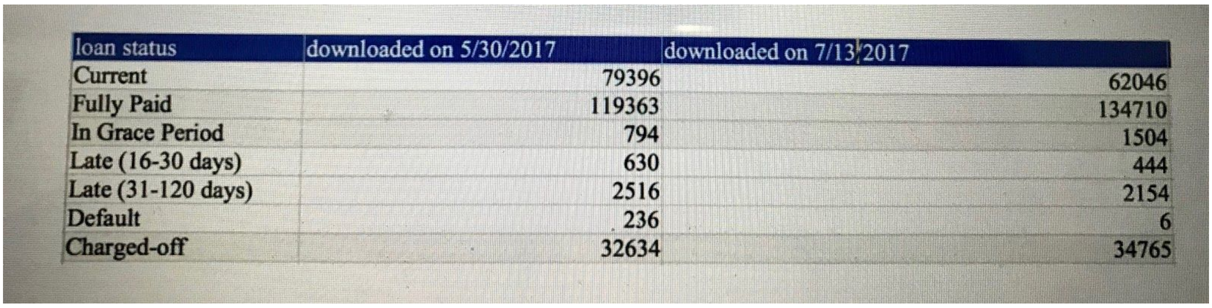


Figure 9: Distributions of loan status

Figure 10: Distributions of loan purpose

The distribution of the loans changes as time goes on as shown in Figure 11. We marked the increasing trend using grey.



loan status	downloaded on 5/30/2017	downloaded on 7/13/2017
Current	79396	62046
Fully Paid	119363	134710
In Grace Period	794	1504
Late (16-30 days)	630	444
Late (31-120 days)	2516	2154
Default	236	6
Charged-off	32634	34765

Figure 11: Trend of loan status

We also explore the relationship between some categorical features with the loan status. For example, in Figure 12, the distribution of the loan status across the features “grade” and “home-ownership” are shown. One can find that loans with grade B and C are more than the other levels. As grade decreases, the percent of charged-off loans increases. People with mortgages are more likely to borrow money from Lending Club. But for people renting, the percent of charged-off is higher than that of the mortgage class.

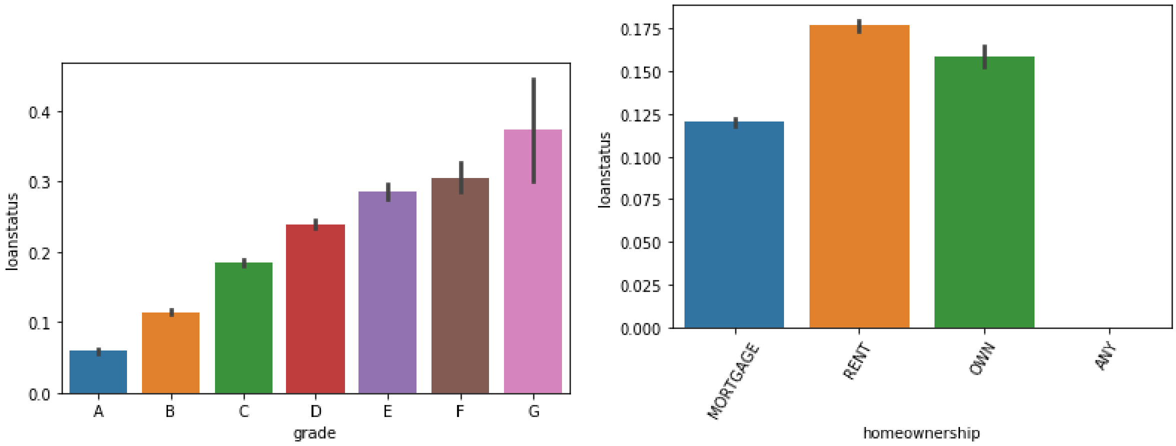


Figure 12. The relationship between grade, home-ownership and the loan status

From Figure 13, we can see that the distributions of loan amount show similar patterns for different loan status. Figure 14 shows the distribution of the total loan amount in different states. California, Texas and New York rank the first three places. Here we have not considered the fact that 100 dollars' real values are different across the 50 states.

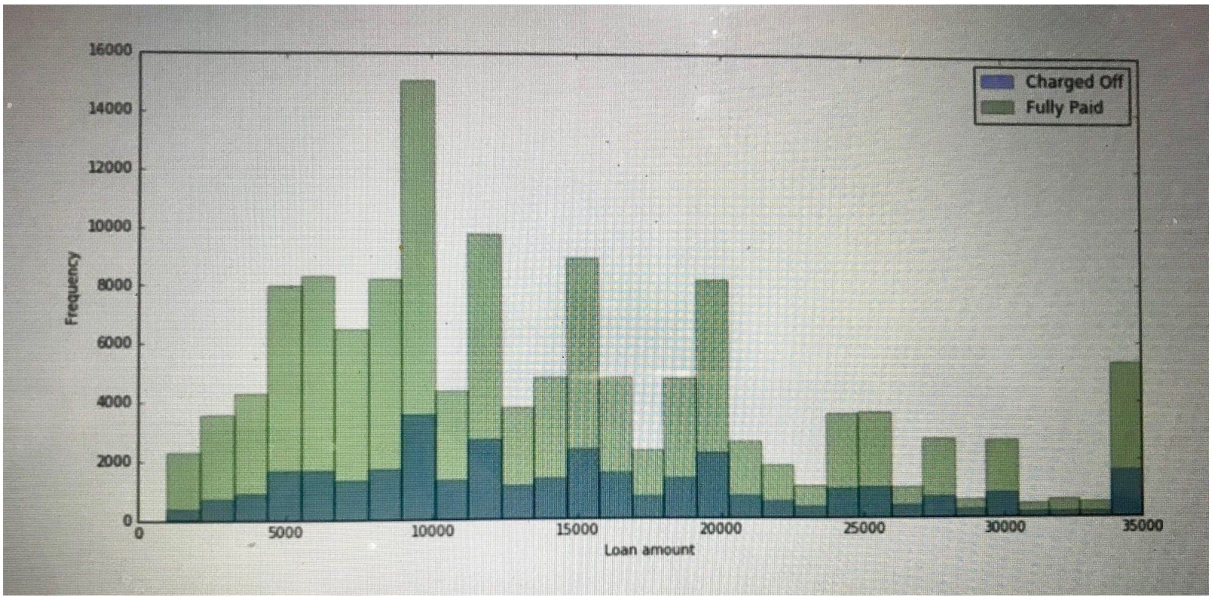


Figure 13. The loan amount distribution and the loan status

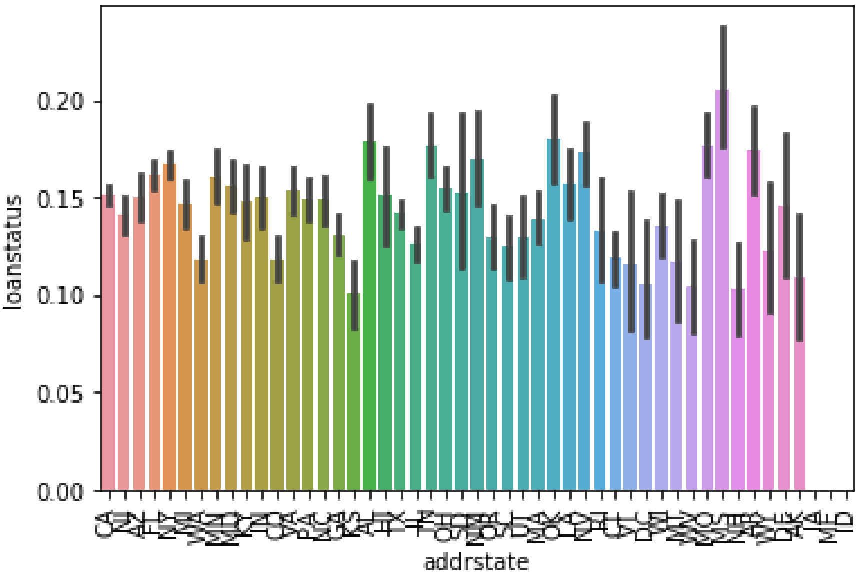


Figure 14. The total loan amount distribution in different states

The interests of the loan from different grades are shown in Figure 15.

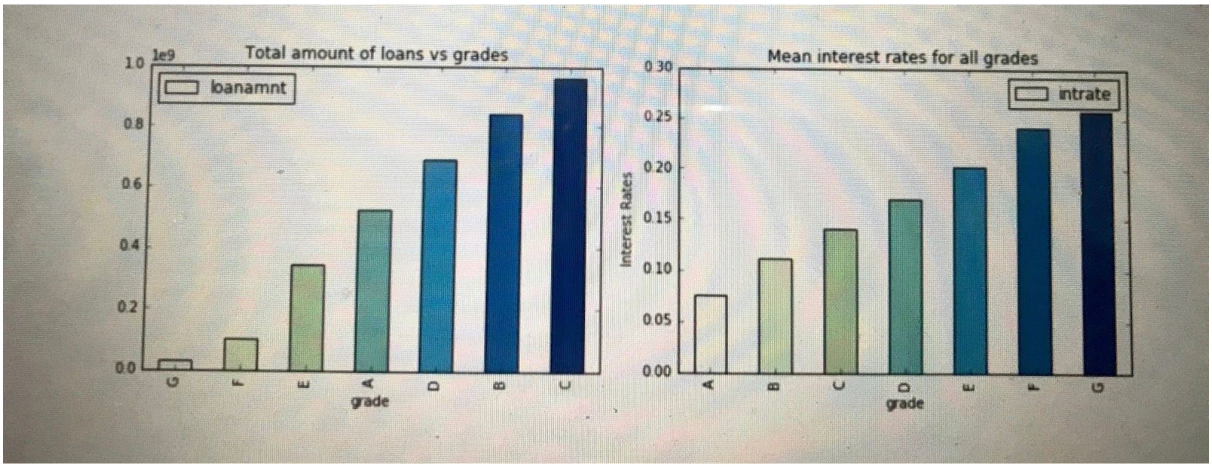


Figure 15. The total amount of loans and mean interest rate vs grades

We are also interested in the relationship between the amount of loans and the annual incomes. From Figure 16, one can see there is a linear relationship when the income is under $50000. The highest amount of loans is $35000. It may be related to the policy of Lending Club. The scatter plot is very dense when the income is under $200000. They are the main target customers of Lending Club.

There is still a lot of information waiting to be explored. We will not go further for the moment.

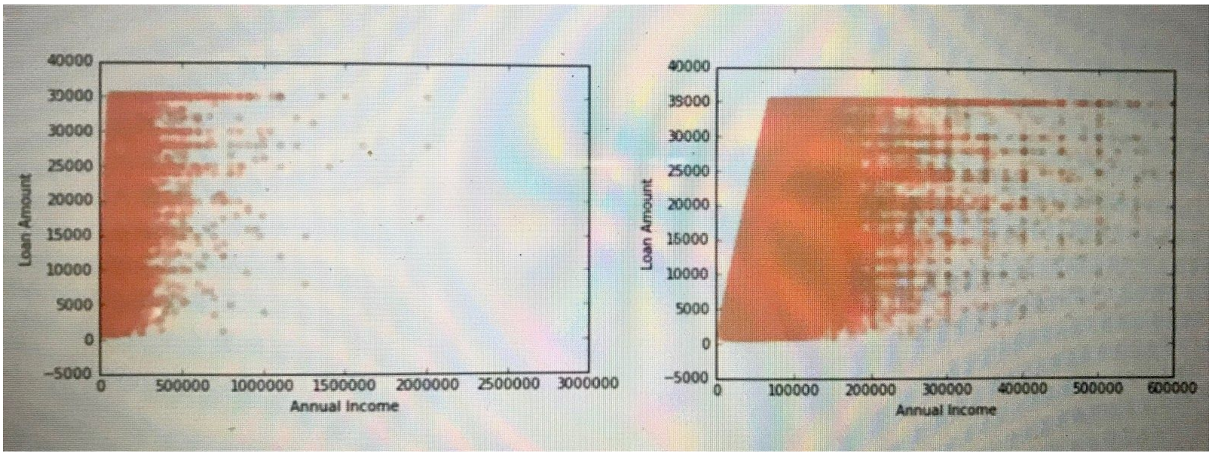


Figure 16. The amount of loans vs incomes

2.4 Benchmark

As we discussed before, accuracy is not the only measure to evaluate the performance of the model. So when we train our model, not only the accuracy should be greater than 0.1599 (the percent of the loans charged off among all loans charged off and fully paid is 15.99%), but also the AUC should be larger than 0.5. We will also build a Random Forest model and set it to be the benchmark.

3 Methodology

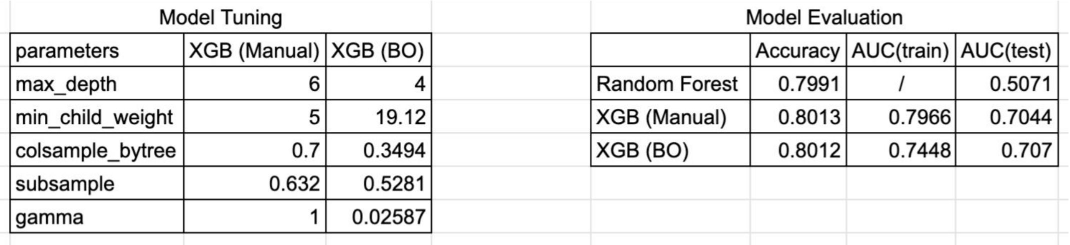
3.1 Training and Testing Datasets

We use the data of loans issued in the first 9 months as the training dataset. It will be used to train the model via cross validation The data in the last 3 months will be used as the testing dataset. For the current dataset, since we have no labels, there is no way to use it to evaluate the models. It will only be used to make predictions.

3.2 Model Training and Tuning

We first choose four features “dti”, “annualinc”, “totalilhighcreditlimit”, “revolutil” to train a Random Forest model. Unlike XGBoost algorithm, which can handle the missing values by itself, Adaboost Tree in sklearn can not deal with missing values. So we use the imputer in sklearn to fill the missing values by the most frequency strategy. Without tuning parameter, the accuracy is 0.8528 and the AUC score is 0.395. So it is far from being satisfying.
We choose XGBoost models by tuning various parameters manually and automatically.
The following parameters are tuned in sequence: max depth, min child weight, colsample bytree, subsample and gamma. It turns out that the best max depth is 3, the best min child weight is 1, the best colsample bytree is 1, the best subsample is 0.2 and the best gamma is 0.2. With the best parameters obtained manually, the AUC score is 0.69 on the test dataset.
With the best parameters from manually tuning in mind, we choose finer grids and use Bayesian Optimization to tune the parameters. The best parameters vary and they are as

follows: subsample: 1.0; max depth: 2; eta: 0.1, gamma: 2.0, min child weight: 35, colsample bytree: 0.126.



Model Tuning

parameters	XGB (Manual)	XGB (BO)
max_depth	6	4
min_child_weight	5	19.12
colsample_bytree	0.7	0.3494
subsample	0.632	0.5281
gamma	1	0.02587

Model Evaluation

	Accuracy	AUC(train)	AUC(test)
Random Forest	0.7991	/	0.5071
XGB (Manual)	0.8013	0.7966	0.7044
XGB (BO)	0.8012	0.7448	0.707

Figure 17. Model Tuning

3.3 Results

In Figure 18, we show the results of the two models. The left one is the XGBoost model with the parameters tuned manually; and the right one is the model with the parameters tuned automatically. The second one is more stable with less variability on the training and testing datasets.

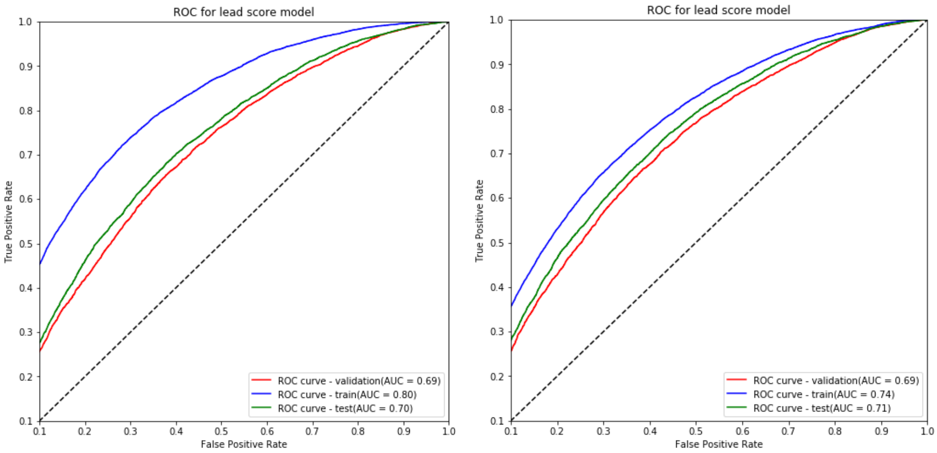


Figure 18. The performance of the XGBoost models

One advantage of tree models is that they can tell the feature importance. In Figure 19, we can get the most important features corresponding the two models. the following features “dti”, “intrate”, “annualinc”, “avgcurbal”, “bcutil”, “earliestcrline month” and “mosinoldrevlop” are important in both models.

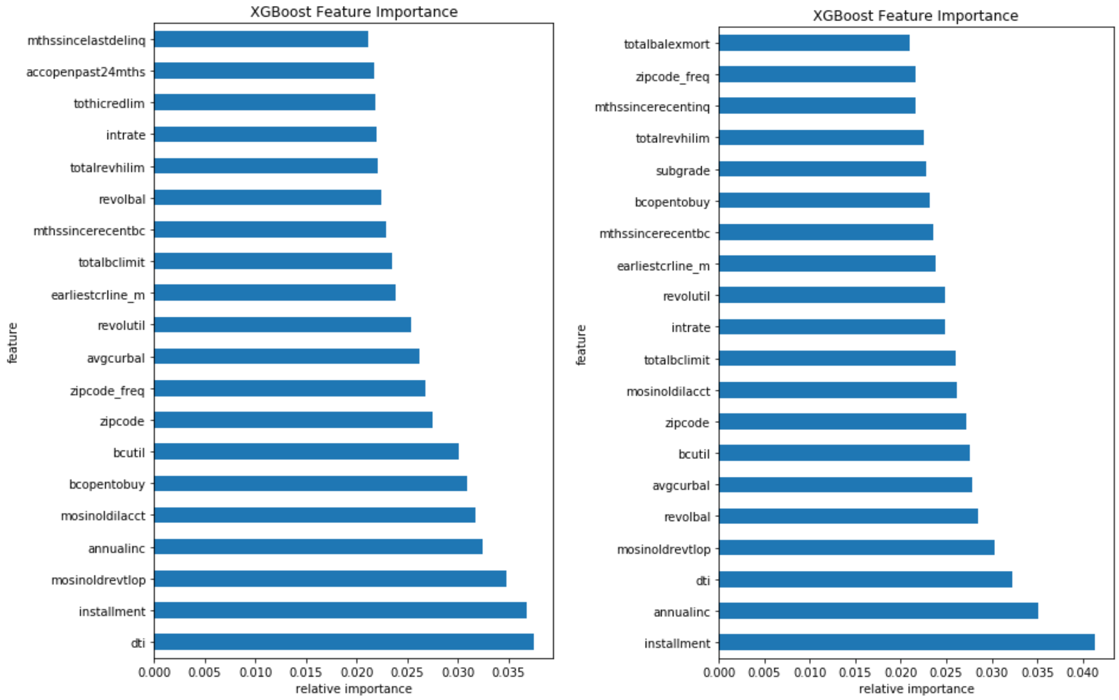


Figure 19. Feature Importance

Now let’s have a closer look at those important features to understand why they have strong predictive power in Figure 20.

The feature “dti” is a ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested Lending Club loan, divided by the borrower’s self-reported monthly income. So it is not a surprise that it is important.

The feature “intrate” is the interest rate on the loan, which is positively related to the grade of the loans. From this, we can see that the grades given by Lending Club are relatively reliable.

Feature “annualinc” is the self-reported annual income provided by the borrower during registration. It indicates the borrowers’ ability of returning the money they borrowed.

Feature “avgcurbal” is the average current balance of all accounts. Feature “bcutil” is the ratio of total current balance to high credit/credit limit for all bankcard accounts. They show the financial situations that the borrowers are in.

Feature “earliestcrline month” is the month the borrowers’ earliest reported credit line was opened up to 2014-12. Feature “mosinoldrevtlop” is Months since oldest revolving account opened. They describe borrowers’ historical information. They are also related to the ages and employment lengths of borrowers. So it makes sense that those features are important when considering whether a loan will be charged off or fully paid.

The relative important features from xgb(BO) make more sense and are easier to explain. The created features like earliestcrline_m, zipcode_freq are in the top 20.

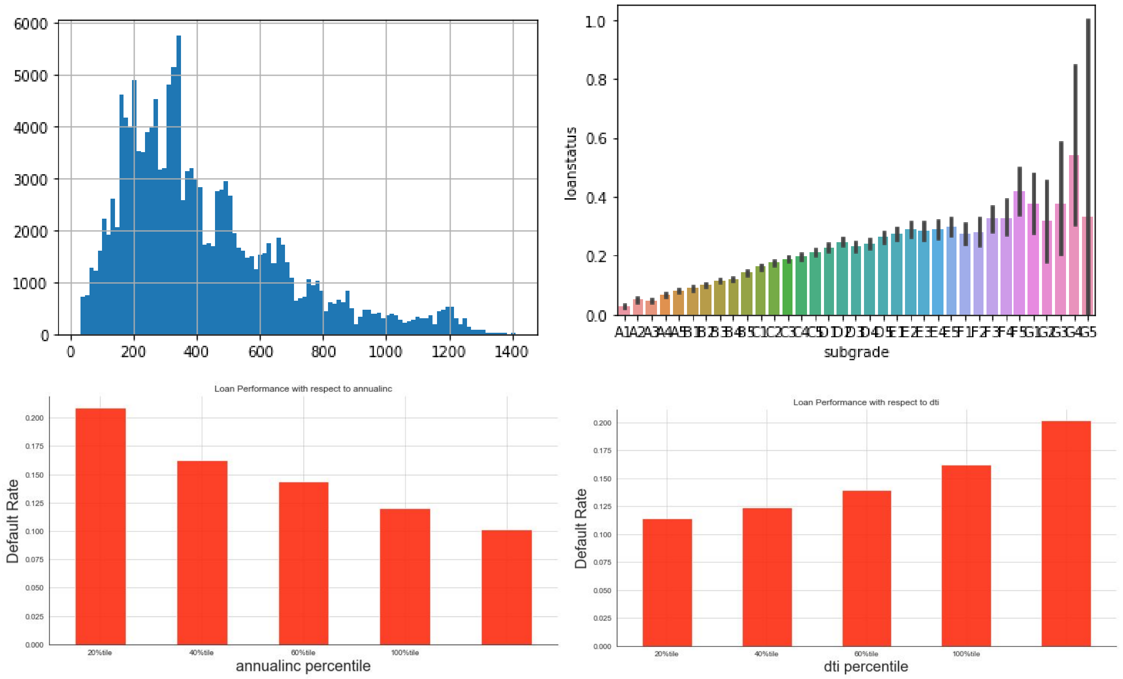


Figure 20. Closer Look at the Important Features

4 Conclusion

After an exploration of data analysis and feature engineering, we trained RandomForest, XGBoost models to predict the status of loans and get a satisfying AUC. Actually, we can do more work in the future to improve it, such as summarize emptitle, analysis ‘desc’ with NLP methods, create relationship between annualinc and addrstates, handle missing values and outliers more carefully, train the model with more sample data, etc.