Bad Loan Predection on P2P Lending Platform

—— LendingClub loan data analysis

Xinyi LI, Fangbei CHENG, Deokkil PARK, Qiu JIN

MSc Fintech

Hong Kong University of Science and Technology

ABSTRACT

Technology, regulation, innovation spires and changes to the macrocosmic and financial landscape cause to supply an innovative financial product to the market. In addition, due to high demand of the financial innovation, FinTech become more popular such as Equity crowdfunding, Peer to Peer Lending Planform, Marketplace lending. (Schindler, 2017)

Peer to Peer Lending Planform become popular and increase market value in financial lending market in 2006. However, the market value of Peer-to-Peer has been decreased since 2017. The reason of disadvantaged of Peer to Peer for example mismatch, excessive finance, low barrier to entry, high default risk, lack of regulation and lack of supervisor.

Therefore, this research used the existing customer information to establish a credit evaluation model to predict future loan default which is one of the main drawbacks in Peer to Peer Lending Platform. We want to find the main factors that cause a loan to be considered as a “Bad Loan”, and predict whether the loan will end up in "Good" or "Bad" status. In the end, we will provide suggestions to help the Lending Club platform avoid default risk.

BACKGROUDS

Lending Club

The history of Lending Club is of great significance to fully understand company. As a premier player in the peer-to-peer(P2P) lending space, Lending Club was founded in 2006, headquartered in San Francisco. Its main business is to provide intermediary services for P2P loans, that brings lenders and investors together and changes the way people get credit. When a loan applicant applies for a loan from the Lending Club platform, the Lending Club platform allows customers to fill out loan application forms online or offline, collects basic information about customers, and usually uses third-party platforms such as credit agencies or FICO to decide whether to lend or not and the level of interest rate. In 2012, the total amount of platform loans reached 1 billion US dollars. In December 2014, Lending Club was listed on the New York Stock Exchange, becoming the largest technology stock IPO of the year and the only listed company in the P2P network lending industry. After 2014, the company began to provide loans not only for individuals, but for small businesses as well. In addition, Lending Club only services the loans with 36 months and 64 months. New loans reached $8.36 billion in 2015. In the first half of 2016, the scandal of illegal lending broke out and the stock price began to fall

The remainder of the paper is organized as follows. Section 1 presents Data Preprocessing. Section 2 presents Feature selection & under sampling. Section 3 presents Data Modeling, Training and Testing. Finally, conclusions are presented.

INTRODUCTION

• Raw Data

The data used in this study is downloaded from Kaggle (Lending Club Loan Data). The time range is from 2007 to 2015. This Lending club data includes 2260668 raw data with 145 columns. The dataset includes Loan characteristics such as loan purpose; Borrower characteristics like annual income, current housing situation, credit history etc.

• Data Mining Process and Technology Tool

The Table 1 shows that Data Mining Process of the project. Data Preprocessing is a preparation state before the actual data mining tasks are performed including Data Cleaning, Data integration, Data Transformation and Data Reduction. Feature selection & under sampling is a find actual relative variables for data mining. This research uses Recursive feature elimination, Pearson correlation, Random Forest using by Python. Data Modeling, Training and Testing stage this paper used two technology tool one is Python and Azure.

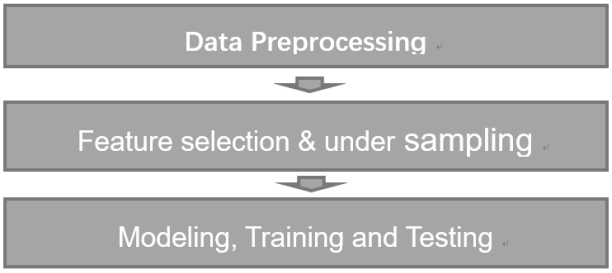

Figure 1: Data Mining Process

KEYWORDS

Lending Club, Data Preprocessing, Feature selection & under sampling, Data Mining

1. Data Preprocessing

Before preprocessing, it is necessary to have a comprehensive understanding of the data. Statistics shows that:

(1) there are more than 2260668 loan applications, 144 feature variables and 1 target variable (loan status) in the data set.
(2) among the 145 variables, there are 36 categorical variables and 109 numeric variables.

First, we do data preprocessing to have the clean data for later analysis.

1.1 Data Cleaning

First we have a general understanding of the missing data. For the processing of missing values, generally speaking, the first step is to determine whether the missing data is meaningful to the purpose of the research.

After sorting the features by the missing value, it can be found that the attributes with more missing values in this data set have little significance in predicting our model, such as id, member_id and url. Further, we use the missingno module to visualize the location of missing values, in which, white space represents the field of null value. Then we delete these meaningless features with a high proportion of missing values . In addition, if missing values are meaningful to attributes, it is necessary to subdivide the attributes into numeric and categorical variables.

	count	max	missing
id	0.0	NaN	1.000000
member_id	0.0	NaN	1.000000
url	0.0	NaN	1.000000
orig_projected_additional_accrued_interest	8426.0	2680.89	0.996273
hardship_last_payment_amount	10613.0	1407.86	0.995305
hardship_payoff_balance_amount	10613.0	40306.41	0.995305
hardship_dpd	10613.0	37.00	0.995305
hardship_length	10613.0	3.00	0.995305
hardship_amount	10613.0	943.94	0.995305

Figure 2: The proportion of missing values of different feature variables, sorted in descending order

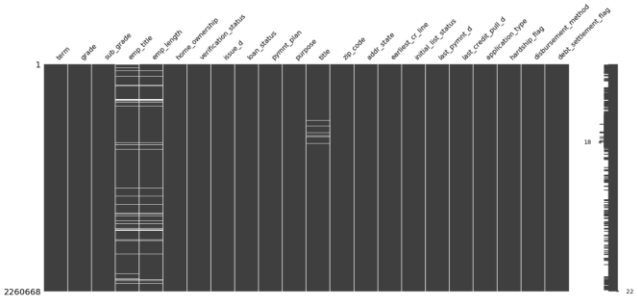

Figure 3: The heatmap visualization of the location of missing values.

When the proportion of missing value of a feature is large, the information of that feature cannot be effectively reflected, so we deleted those variables with more than 40% missing values.

There are 46 feature variables with more than 40% missing values. Among them, 24 variables have more than 90% missing value. After data cleaning, we have 99 features.

1.2 Data Integration

Due to the different purposes of data mining, for the same data set, researchers will select different features relevant to their purposes. We do not need to put all the features while training the models, so the next step is to select features according to the meaning of the features variables and remove the meaningless feature variables.

The criteria of selection are as follows:

1. The feature variable which is not relevant to the project
2. The feature variable which is redundant to another.
3. The feature variable with more than 90% same value

Therefore, feature variables are selected according to the 3 principles above.

1. Removing meaningless and redundant features

After looking up the meaning of each feature variable, we choose the following 24 features that are irrelevant or redundant:

'funded_amnt','emp_title','funded_amnt_inv','sub_grade', 'addr_state','verification_status','last_credit_pull_d','issue_d', 'pymnt_plan','title','zip_code','earliest_cr_line','revol_util', 'out_prncp_inv','total_pymnt_inv','total_rec_late_fee', 'recoveries','collection_recovery_fee','last_pymnt_d', 'collections_12_mths_ex_med','policy_code','hardship_flag', 'debt_settlement_flag','inq_last_12m'

The reasons are as below:

'Sub_grade': Repeat the information with Grade;
'Emp_title': too many classifications, and can not reflect the real situation of the borrower's income or assets;
'last_credit_pull_d': LendingClub Platform's most recent time to provide loans, not related to the purpose of the study
'Zip_code': Address Zip code, the zip code display is incomplete, meaningless;
'Addr_state': The state where the application address belongs does not reflect the borrower's solvency.
'Title': The information of 'Title' and 'Purpose' is repeated, and the classification information of title is more discrete.
'***Polic_code***': Not related to the purpose of the study;
'Earliest_cr_line': Records the time when the borrower made the first loan;
'Issue_d': The time of loan issuance, which leaks critic information to the prediction in advance.
'last_pymnt_d','Issue_d', 'collection_recovery_fee': Predictive loan default model is a means of risk control before loan, and the information after loan will affect the accuracy of our models.

'Funed_amnt', 'funded_amnt_inv' and 'loan_amnt' : Data are identical, removing 'Funed_amnt' and 'funded_amnt_inv'
'Out_prncp' and 'out_prncp_inv': data are identical，removing 'out_prncp_inv'
'Tot_pymnt' and 'total_pymnt_inv' : data are identical，removing 'total_pymnt_inv'

Figure 4: Diagonal Correlation Matrix between Features

After removing meaningless and redundant features according to the meaning of variables, categorical variables are reduced from 28 columns to 9 columns. There are 75 feature variables in total.

	term	grade	emp_length	home_ownership	loan_status	purpose	initial_list_status	application_type	disbursement_method
0	36 months	C	10+ years	RENT	Current	debt_consolidation	w	Individual	Cash
1	60 months	D	10+ years	MORTGAGE	Current	debt_consolidation	w	Individual	Cash
2	36 months	D	6 years	MORTGAGE	Current	debt_consolidation	w	Individual	Cash
3	36 months	D	10+ years	MORTGAGE	Current	debt_consolidation	w	Individual	Cash
4	60 months	C	10+ years	MORTGAGE	Current	debt_consolidation	w	Individual	Cash

Figure 5: Categorical attributes after removing meaningless and redundant attributes

2. Removing features with more than 90% same values

After calculating the proportion of the highest frequent occurrences of the variables in each column, we delete the following 9 features with more than 90% same values:

'delinq_amnt','acc_now_delinq','chargeoff_within_12_mths','tax_liens','num_tl_30dpd','disbursement_method','application_type','num_tl_120dpd_2m','num_tl_90g_dpd_24m'

	simi
delinq_amnt	0.996814
acc_now_delinq	0.996102
chargeoff_within_12_mths	0.992335
tax_liens	0.971365
num_tl_30dpd	0.966334
disbursement_method	0.965443
application_type	0.946604
num_tl_120dpd_2m	0.931467
num_tl_90g_dpd_24m	0.917012
pub_rec_bankruptcies	0.879113
pub_rec	0.841680
tot_coll_amt	0.821053
delinq_2yrs	0.813524

Figure 6: The proportion of the highest frequent occurrences of the variable

After preliminary feature selection, the initial 145 feature variables become 66. In a good credit model, the feature variables are usually fewer, so we will further select the feature variables.

1.3 Data Transformation

1.3.1 Data Type Transformation

1. Divide Categorical and Numeric Variables

We divide the feature variable into two classes according to the type of the data. The first type of feature variable is categorical, and the second type is numeric.

Seven object features have missing values, the proportion is shown below. Given the number of missing values is not big, we will fill the null values in categorical variables with “Unknown”. There are still some missing values in the numeric columns which will affect later analyzing steps, so we drop the rows with missing values.

	missing
emp_title	7.385826e-02
emp_length	6.498389e-02
title	1.031775e-02
last_pymnt_d	1.073134e-03
last_credit_pull_d	3.229134e-05
earliest_cr_line	1.282807e-05
zip_code	4.423471e-07
disbursement_method	0.000000e+00
hardship_flag	0.000000e+00
application_type	0.000000e+00
initial_list_status	0.000000e+00

Figure 7: The proportion of missing values in the categorical attributes in descending order

2. Transform Categorical Attributes into Numbers

There are two types of categorical variables, i.e. ordinal and nominal attributes. We use different methods to transform them into numbers.

- **Ordinal Attributes**

For the ordinal attributes, they provide enough information to order objects. Ordered multivalued variables are non-numeric data of an ordered category. For example, products are classified into first-class products, second-class products, third-class products, inferior products and so on. We learned that Lending Club classifies loan applicants' credit grades A to G, and matches loan interest rates according to different credit grades. Customers whose credit rating is A are better than those whose credit rating is B.

A < B < C < D < E < F < G: credit risk ranks from low to high

So we use the function mapping_dict in python to transfer the ordinal attributes: *'emp_length','grade'*

	emp_length	grade
0	10	3
1	10	4
2	6	4
3	10	4
4	10	3

Figure 8: The ordinal attributes after being applied mapping_dict View data information

- **Nominal Attributes**

For nominal attributes with no significance of the order between the values, they are the result of categorizing. Data is categorized and expressed in words. For example, borrowers are categorized by gender into two categories: male and female; the categorization in the categorized data is disorderly, which means that we can't sort the nominal attributes ("purpose") like the ordinal attributes.

Car < wedding < Education < moving < house: this sort does not make common sense, nor is it meaningful.

So for nominal attributes, we use get_dummies to transfer them. get_dummies function can convert categorical variable into dummy or indicator variables. The nominal attributes include: *'term', 'home_ownership', 'purpose', 'initial_list_status'*

	home_ownership_ANY	home_ownership_MORTGAGE	home_ownership_NONE	home_ownership_OWN	home_ownership_RENT
0	0	0	0	0	1
1	0	1	0	0	0
2	0	1	0	0	0
3	0	1	0	0	0
4	0	1	0	0	0

Figure 9: The norminal attributes after being applied get_dummies

1.3.2 Feature Abstraction

Feature abstraction refers to the transformation of data into data that can be understood by algorithms.

There are several statuses which indicate whether the loans are performing well or not. The criteria of different loan status given by the Lending Club are taken into consideration when we are defining the classes of the loan status and we created a column called 'loan condition' with a value of 0 or 1 to represent good loan and bad loan repectively.

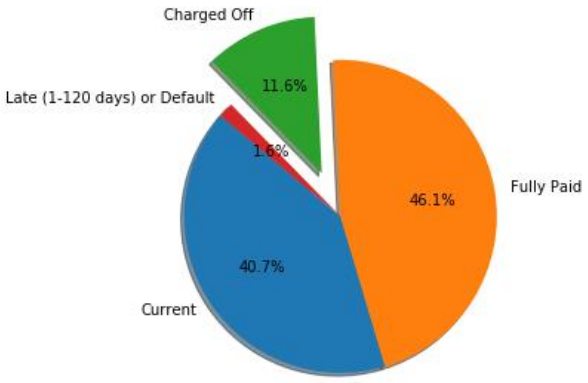

Figure 10: Breakdown of Loan Status

The bad loan includes status of "Charged Off", "Default", "Does not meet the credit policy. Status:Charged Off", "In Grace Period", "Late (16-30 days)", "Late (31-120 days)", which imply the loan is past due, default or charged off.

The good loan includes the rest of other loan status, which include *'Current', 'Fully Paid' and 'Does not meet the credit policy. Status: Fully Paid',* which imply the loan is up to date on all outstanding payments, or has been payed off.

The end result is a data set with 13.2 % of them being bad loans. It is noteworthy to point out a potential sampling bias in this dataset.

1.3.3 Feature scaling

Feature scaling refers to the restriction of variable data to a certain range during preprocessing. Feature scaling is essentially a de-dimensioning process, and it can speed up the convergence of the algorithm.

We adopt the standardization method which enables the data contains the useful information in the outliers, and makes the model less affected by the outliers. In the feature scaling step, we call the StandardScaler module to standerdize the 'int64', 'float64' and 'object' data.

Finally, we deploy the method of pandas info () to check the dataset after preprocessing. It is found that all categorical variables have been transformed and all data types meet the requirements of later requirement. After preprocessing, we have 118540 records and 66 feature variables in total.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1185400 entries, 0 to 2260666
Data columns (total 85 columns):
dtypes: float64(61), int64(1), uint8(23)
memory usage: 595.8 MB
```

Figure 11: The data information after preprocessing

1.4 Dimensionality Reduction

After standardization, we need to further preprocess the data. We already dropped columns which are irrelevant, redundant and meaningless in the last several data cleaning steps. However, there are still 84 columns, in other words, features remained. So dimensionality reduction is required and here are the behind motivations:

1. Avoid the curse of dimensionality

The curse of dimensionality often arises when analyzing data in high-dimensional spaces. The common scene of this problem is that when the dimensionality increases, the volume of the whole space equally increases, in a fast speed, so that valuable data become sparse. This problematic sparsity will prevent our model from being efficient.

2. Reduce noise, time and space

Even though we have already dropped several irrelevant features, but these features are those obviously can't affect the loan status, and we choose these features barely by definition. So dimensionality reduction can help us eliminate other subtle irrelevant features, which will reduce the noise and time in machine learning. Also, less data means less storage space required.

In a nutshell, dimensionality reduction will improve model's accuracy while using less computing and less storage space. So how to apply specific technique to realize it? In our project, we choose feature selection method.

1.4.1 Feature Selection

Feature selection is the process of identifying and selecting the import features. And it can be divided into 3 mainly categories:

- ***Wrapper Method***
- ***Filter Method***
- ***Embedded Method***

We follow these methods one by one, and get an array of significant features.

1. Wrapper Method

In this method, machine learning algorithm is required, and the selection of features is based on the algorithm's performance, which can be regarded as evaluation criteria. Normal wrapper methods are backward elimination, forward selection, and recursive feature elimination. Here we apply RFE technique to select feature.

- **Definition**

Recursive feature elimination (RFE) removes the weakest features until the number of remained features reaches the specified number. And the features are ranked by their importance. The input of RFE method is a selected model and the number of required feature. And in the output, RFE give all the features a ranking, smaller number means more important. It also shows a boolean result, true stands for relevant feature and false stands for irrelevant.

- **Implementation**

In our project, we use linear regression model with 30 features, and here are the 30 most important features selected by RFE:

'loan_amnt', 'int_rate', 'installment', 'grade', 'out_prncp',
'total_rec_prncp', 'total_rec_int', 'last_pymnt_amnt',
'open_rv_24m',
'all_util', 'bc_open_to_buy', 'mths_since_recent_inq',
'num_actv_rev_tl', 'num_rev_tl_bal_gt_0',
'percent_bc_gt_75',
'total_bc_limit', 'term_ 36 months', 'term_ 60 months',
'home_ownership_NONE', 'home_ownership_OWN',
'home_ownership_RENT',
'purpose_debt_consolidation', 'purpose_educational',
'purpose_home_improvement', 'purpose_medical',
'purpose_moving',
'purpose_renewable_energy', 'purpose_small_business',
'purpose_wedding',
'initial_list_status_f'

2. Filter Method

After RFE, the number of feature goes from 84 to only 30. And we can still filter the features to a smaller scale. Filtering can be realized by building correlation matrix. In our project, we apply Pearson correlation to see the relevance of these 30 features.

- **Definition**

A Pearson correlation is a number between -1 and 1 that indicates the extent to which two variables are linearly related. A Pearson correlation between variables X and Y is calculated by:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

The result r can range from -1 to 1. And an r of 1 indicates that X and Y have a perfect positive linear relationship. -1 stands for a perfect negative linear relationship. 0 indicates that there is no relationship between these two variables. The relationships are showed as below:

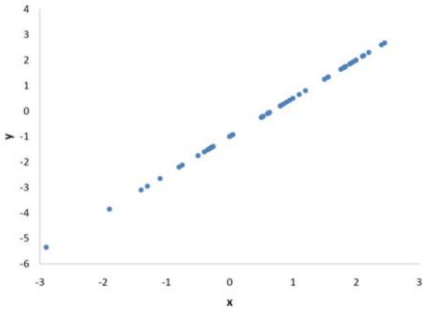

Figure12: Pearson r=1

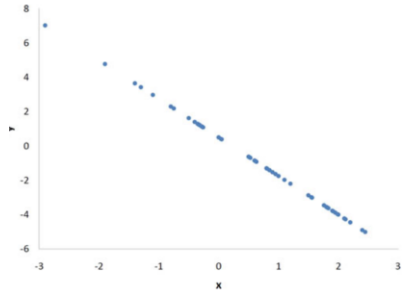

Figure13: Pearson r=-1

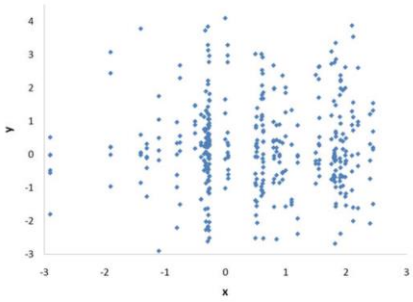

Figure14: Pearson r=0

- **Implementation**

In our project, firstly we plot the Pearson correlation heatmap. The graph is showed below. Then we can subtract features which are highly correlated to the other. And here are the features we choose to drop:

'installment': The Pearson r of 'installment' and 'loan amount' is 0.95, highly closed to 1.0, which means they are highly correlated. From the definition, we know that installment is the monthly payment owed by the borrower if the loan originates. So it's a redundant feature, we can drop it.

'grade': The Pearson r of 'grade' and 'interest rate' is also highly close to 1.0. And after doing research on LendingClub's product information, we know that, higher grade indicates lower credit quality and default risk, so interest rate would be higher in case to compensate the high risk. Therefore, we can drop this redundant feature.

'term_ 36 months': In our heatmap, we found a perfect negative correlation between ***'term_ 36 months'*** and ***'term_ 60 months'***. And the reason is that, LendingClub platform only provide these 2 different loan products. The product information can be detected by only one feature.

Based on same logic, we further dropped ***'num_rev_tl_bal_gt_0' and 'bc_open_to_buy'***. Now we have 25 features left.

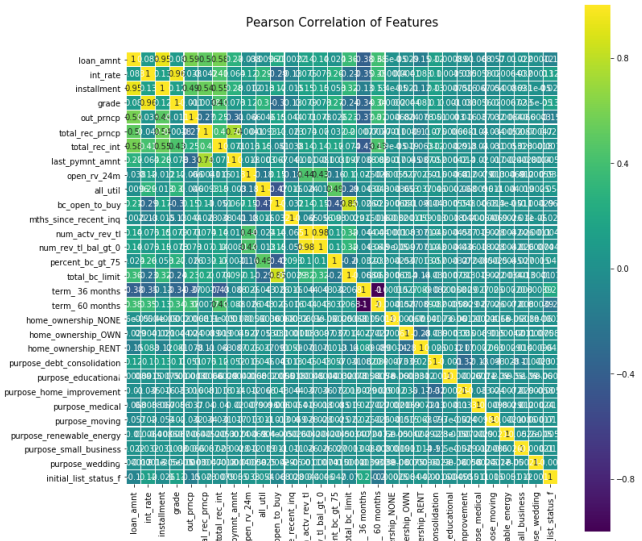

Figure15: Pearson Correlation of 30 Features

3. Embedded Method

Embedded method is similar to wrapper method; they both require a learning algorithm. But in embedded method, the feature selection algorithm is integrated as part of the learning algorithm. In our project, we apply random forest technique.

- **Definition**

Random forest consists of several hundred decision trees, each of them build on a random extraction of the features and labels. In decision tree, the more a feature decreased the impurity the more important the feature is. So in random forest, the impurity decrease from feature will be averaged across trees to determine the final importance of the feature.

- **Implementation**

The output of random forest straightly shows the importance of each features on the decision. And here is the result after applying this method to the remained 25 features:

('loan_amnt', 0.08245493292607543)
('int_rate', 0.04380088519570329)
('out_prncp', 0.30826837097276)
('total_rec_prncp', 0.20310375973382735)
('total_rec_int', 0.060779619249482914)
('last_pymnt_amnt', 0.16899717258610833)
('open_rv_24m', 0.013351819483143512)
('all_util', 0.023293047063843036)
('mths_since_recent_inq', 0.01564651266281928)

('num_actv_rev_tl', 0.0152381317142664 2)
('percent_bc_gt_75', 0.012966753056025756)
('total_bc_limit', 0.0274083447531297)
('term_ 60 months', 0.0076223635855243105)
('home_ownership_NONE', 2.6522606831236165e-10)
('home_ownership_OWN', 0.002274681805166684)
('home_ownership_RENT', 0.003557415486084212)
('purpose_debt_consolidation', 0.004282563341139682)
('purpose_educational', 0.0)
('purpose_home_improvement', 0.0017023179409134577)
('purpose_medical', 0.0007361831106278751)
('purpose_moving', 0.0005188056431232 11)
('purpose_renewable_energy', 8.762529427179857e-05)
('purpose_small_business', 0.0007850313251661223)
('purpose_wedding', 1.3886975093733716e-07)
('initial_list_status_f', 0.0031235239358207404)

From the result we can get these highly important features:

'out_prncp': Remaining outstanding principal for total amount funded. This feature has the highest importance score, which means that remaining principal can highly contribute to the decision of final loan status.

'total_rec_prncp': Principal received to date. This feature also has a strong effect on the final loan status.

And these two features are both related to principal, which indicates that LendingClub platform should be alert to borrower's principal data.

Besides, we can also further drop insignificant features, which are ***'home_ownership_NONE','purpose_educational','purpose_renewable_energy','purpose_wedding'.***

1.4.2 Conclusion

Among these three feature selection methods, wrapper method measures the usefulness of features, and filter method focus on the correlation between features, and last, embedded method combines the characteristics of the other 2 methods. So the embedded method can provide a more accurate and detailed result. But it can also be computationally more expensive than the other 2 methods, especially filter method. So we incorporate these 3 methods to achieve a complete feature selection.

So far, after feature selection, our selected features go from 84 to 21. And the whole dataset is close to fully prepared for model training and testing.

1.5 Undersampling

1.5.1 Definition

There are total 2 techniques can be used for fixing class imbalance, oversampling and undersampling.
Undersampling is the process that randomly deleting samples from majority class in order to match the minority class. Oversampling is the process that randomly generating samples in minority class in order to match the majority one.

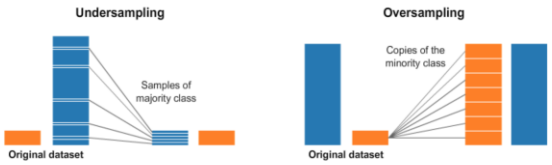

Figure16: Undersampling and Oversamping

Many machine-learning techniques, such as neural network, can get more reliable prediction from being trained with balanced data.

1.5.2 Implementation

In our data set, good loan accounts for 89.13%, and bad loan accounts for 10.87%.

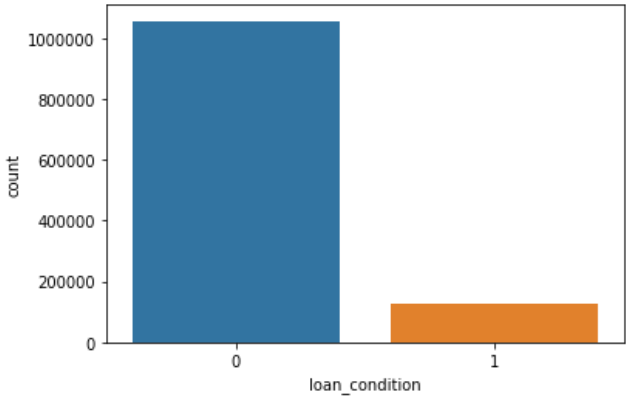

Figure17: Class distribution

So in order to get a balanced data for a better model performance, we employ undersampling method to realize it. The graph below shows the evenly distributed class after undersampling.

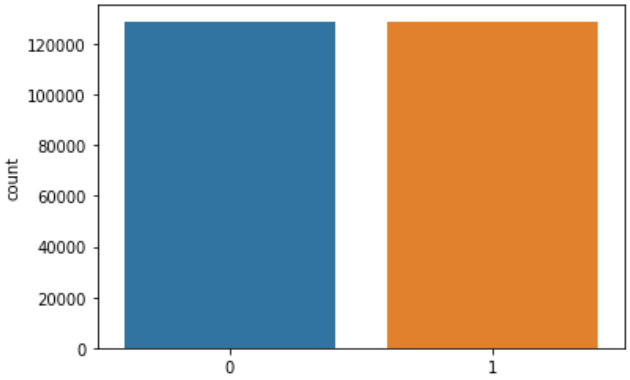

Figure18: Class distribution after undersampling

At this point, the whole preprocessing job is over, and we get a cleaned, balanced and meaningful data which will help our following models perform better in learning features.

2. Model

2.1 Logistic regression

2.1.1 Concept of logistic regression

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc... Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

As we want to use the data to do default prediction, the outcome of the model will contain two values: good loan and bad loan. So logistic regression is also a good method to help meet our goal. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables[2]. For our data, we use 0 to represent good loan and 1 to represent bad loan. Then we can transfer our data to these two values by using the sigmoid function.

A sigmoid function is a bounded, differentiable, real function that is defined for all real input values and has a non-negative derivative at each point. Asigmoid "function" and a sigmoid "curve" refer to the same object. The function is:

$$g(x)=1/(1+e-x) \quad (4)$$

The graph of this function is as follow:

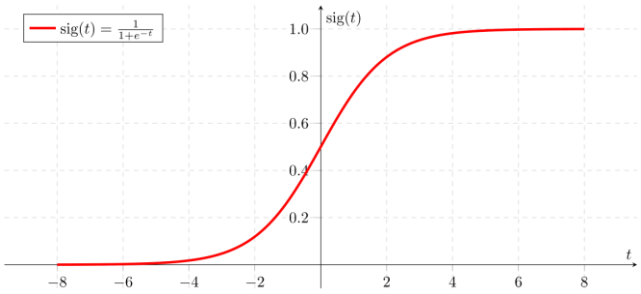

Figure 19: the graph of sigmoid function

After using this function, the data can be defined into two groups: bad loan and good loan.

2.1.2 Result of logistic regression

The outcome by using Gini index is shown below

	Precision	Recall	F1 score	Accuracy
0	0.83	0.99	0.90	0.89
1	0.98	0.80	0.88	

It means that for the whole charged off companies, default companies, and companies which don't meet the credit policy, 83% of them will be detected by the model. And for those companies defined as bad companies by the model, 99% of them are correctly defined. The accuracy of the model is 89%.

2.2 Classification

Classification is a machine learning method that uses data to determine the category, type, or class of an item or row of data. For example, the classification method is able to identify sentiments “good” or “bad”; to categorize customers’ credit grade.

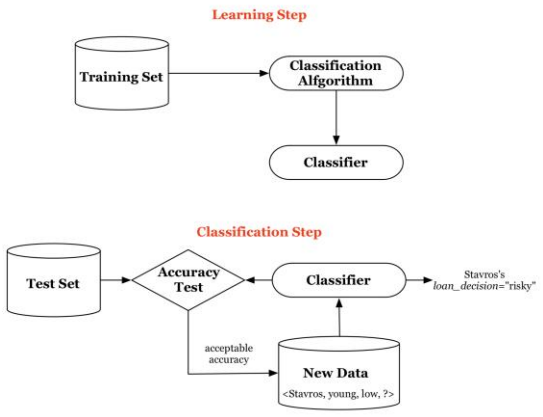

Figure 20: Classification Process with Training and Test Set

The process of the classification consists of four different stages

i. Gathering a set of previous data such as Lending Club Dataset. This Dataset is called Training Set. The Training Set must have more than one attribute. Moreover at least one of the attributes' type should be a categorical type therefore the dataset can classify by through Classification. The reason the categorical attribute will be used as values the desired labels in the process of classification method. This attribute is defined as Class Attribute.

ii. Analyzing the training set and creating model. In this stage, a Class Value will be returned with input the non-class attribute values via the model. During the Classification Data Mining method there are several different models have been developed, for example, Decision Trees, Rule-based Classification, Bayesian Classification, Neural Network-based Classification, Support Vector Machines, K-Nearest Neighbor Classification.

iii. Estimating the accuracy of the model. Find another set of previous data, with the same attributes as the Training Set which is called Test Set. Attribute value of the Test Set should be disjoint from those in the Training Set.

iv. Finally, if the estimated accuracy of the classifier is acceptable, use this model to classify new objects.

This research generated data with two classification machine learning method called Neural Network Model and Perceptron model.

2.2.1 Decision tree

1. concept of decision tree

A decision tree is a series of decisions undertaken through some form of information heuristic and stored in a tree like hierarchical structure[1].

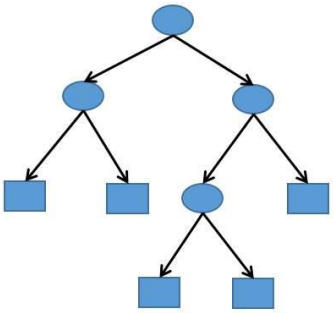

Figure 21:Flow chart of Decision tree

A more intuitive explanation can be shown through playing a game. Now the player has twenty chances to guess a number written down on a paper by a person in advance. The range that the number belongs to is shown before the starting of the game. Every time after the player give an answer, the person who wrote down the number will tell him/her that his/her answer is higher or lower than the number on the paper. So the player will adjust his/her strategy each time he/her gets a feedback. Since we have limited number of questions, we have to determine the value of each question asked so we are able to narrow down the space of possible subjects. If we draw the series of questions the resulting graph represents a tree with binary splits at each node. Each question is carefully crafted to provide the most information regarding the secret subject, and this is the intuition behind decision trees. There are several approaches to build decision trees in literature. As we used PCA during the data processing step, all the features are numeric. So we need to find out the appropriate splitting attribute. Then we choose Information Gain and Gini index to select the splitting attribute with the best score.

2. Process

1.2.1 Gini index: The first step is to load the data and then calculate the Gini index as the criterion. And then Gini index is used to find out the best splitting attribute and build up the decision tree. Finally, test data is used to evaluate the decision tree model.

1.2.2 Information Gain: For this part, Information Gain is used as a substitute of Gini index. First, the etropy is calculated and then it is used to find out the largest Information Gain. Finally the best splitting attribute will be found out.

3. result

Accuracy, the percentage number of correctly classifies predictions, is one of the most intuitive ways to evaluate learner. However, the metric can be shown to be flawed when in the use of highly skewed data. For example, if the minority class was only 10 percent of the dataset, a learner could simply attribute all the dato to the majority class and would be able to achieve an accuracy score of 90%, which is meaningless. Although, on its surface, an accuracy of 90% may show that we have a good learner, but that is further from the truth considering in certain cases a misclassification of the minority class is not acceptable. So accuracy may not be the only factor that we used to evaluate our final model. Thus it is imperative to add some other factors while evaluating our model. Under this condition, precision, recall, accuracy score, and F-1 score are all introduced. These metrics are defined below.

		Model prediction	
Actual loan status		No default	Default
	No default	TN	FP
	Default	FN	TP

Table 1 which sections of the confusion matrix are labeled as TN, FP, FN, TP

Precision is used to judge how many p2p companies indeed default in the pool of companies labeled as default company by the model. It can be calculated as follow:

$$Precision=TP/(TP+FP) \quad (1)$$

This formula tells the probability that the prediction is right when the model judges the company as a default one. So this factor shows how much we can trust our model when it gives us an answer: default.

Recall is another factor showing the probability that the model will successfully identify the default company. It can be calculated as follow:

$$Recall=TP/(TP+FN) \quad (2)$$

From this formula, it is easy to see that the factor, recall, tells the capability of a model to find out all the default companies.

In statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a factor to judge the model's accuracy. It considers both the precision p and the recall r of the test to compute the score: p(precision) is the number of correct positive results divided by the number of all positive results returned by the classifier, and r(recall) is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic mean of the precision and recall, where an F1score reaches its best value at 1 (perfect precision and recall) and worst at 0. F1 can be calculated as follow:

$$F1\ score=2*precision*recall/(precision+recall) \quad (3)$$

Now, let's get back to the former example: for a highly skewed data(10% default and 90% not default). If the model overfits the date by telling that all the companies do not default, the recall equals to zero, showing that the model doesn't perform very well.

The outcome by using Gini index is shown below:

	Precision	Recall	F1 score	Accuracy
0	0.86	0.85	0.86	0.86
1	0.85	0.86	0.86	

Table 2 outcome of decision tree(Gini index)

It means that for the whole charged off companies, default companies, and companies which don't meet the credit policy, 86% of them will be detected by the model. And for those companies defined as bad companies by the model, 85% of them are correctly defined. The accuracy of the model is 86%.

The outcome by using Information Gain is shown below:

	Precision	Recall	F1 score	Accuracy
0	0.86	0.86	0.86	0.86
1	0.86	0.86	0.86	

Table 3 outcome of decision tree(Information Gain)

It means that for the whole charged off companies, default companies, and companies which don't meet the credit policy, 86% of them will be detected by the model. And for those companies defined as bad companies by the model, 86% of them are correctly defined. The accuracy of the model is 86%.

2.2.2 Neural network

1. Concept of Neural Network Model

A Neural Network is a series of algorithms that use complex networks of simple computing elements as mathematical models to mimic the functions of the brain. Artificial neurons were first proposed in 1943 by Warren McCulloch, a neurophysiologist, and Walter Pitts, a logician, who first collaborated at the University of Chicago. (McCulloch, Warren; Pitts, Walter 1943)

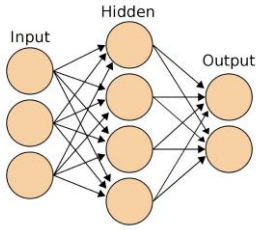

Figure 22: Simple Neural Networks

A neural network is a set of interconnected layers. The inputs are the first layer, and are connected to an output layer by an acyclic graph comprised of weighted edges and nodes (Figure22). Between the input and output layers you can insert multiple hidden layers. Most predictive tasks can be accomplished easily with only one or a few hidden layers.

2. Concept of Perceptron Model

The Perceptron Algorithm is one of the classification algorithms that help to find method for tagged dataset including a label column. In other words, the Perceptron Model is only capable to lean simple functions that are linearly separable. Microsoft Office pointed that the Perceptron Model is still powerful model to predict binary outcomes in neural network model. For example, such as where or not a loan is likely to fail or not, or whether a machine is likely to be broken or not, or a patient can have a disease or not.

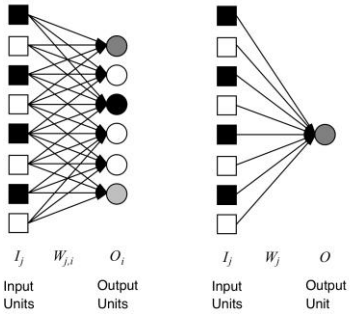

Figure 23: Perceptron Network & Single Perceptron

As Figure 23, a feed-forward network with only one layer of adjustable and weights connected to one or more threshold units as output units.

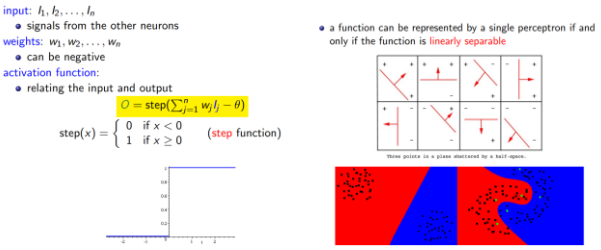

Figure 24: Perceptron Network & Single Perceptron

Output value is computed by the activation function with input and weights. And If the activation function can divide or classify the output value then this function can be defined as a single perceptron as Figure 24.

3. Concept of Deep Neural Network Model

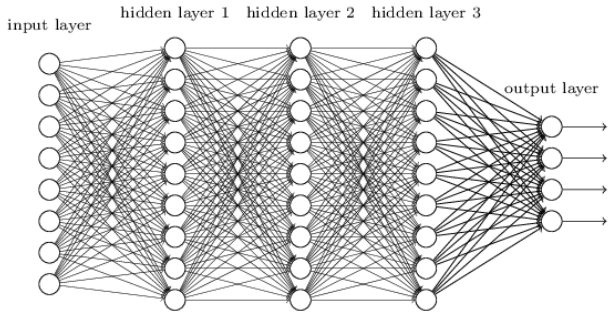

Figure 25: Deep Neural Network

Including Dai, W., & Berleant, D. (2019) and recent research have shown that deep neural networks with many layers can be very effective in complex tasks such as image or speech recognition. The successive layers are used to model increasing levels of semantic depth.

The relationship between inputs and outputs is learned from training the neural network on the input data. The direction of the data proceeds from the inputs through the hidden layer and to the output layer to nodes in the next layer.

To compute the output of the network for a particular input, a value is calculated at each node in the hidden layers and in the output layer. The value is set by calculating the weighted sum of the values of the nodes from the previous layer. An activation function is then applied to that weighted sum.

4. Result of Perceptron Network and Neural Networks Model

The outcome by using Perceptron Network and Multi-Layer Neural Networks Model is relatively shown as below Figure 26. The result shown that the accuracy of Neural Networks Model is 90.3% with all variables. However, with our selected variables the accuracy has been increased to 97.6%. This means the Model could predict which loan could be “good” or “bad” with accuracy level 97.6%.

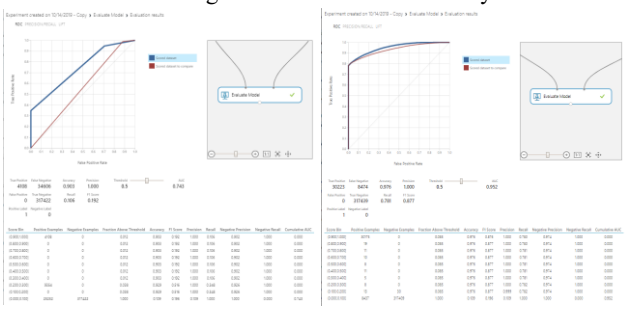

Figure 26: the result from Neural Networks Model (Blue Line)

In Perceptron Networks Model, the result with all attributes shows that the accuracy of the Perceptron Model is relatively much lower shown as 21.2%, However, with our selected attributes the accuracy extremely increases to 97.6% (Figure 27)

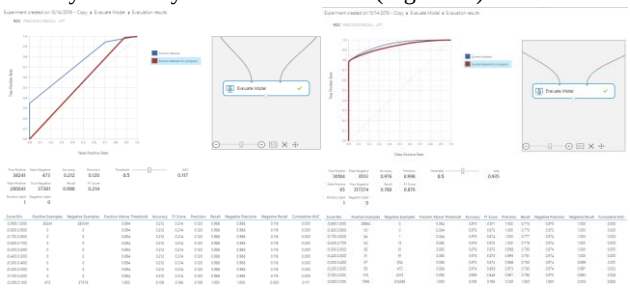

Figure 27: the result from Perceptron Network (Red Line)

Based on this evaluation result between two different selected features, both models shown that our selected variables can help to predict loan condition better.

3. Conclusion

3.1 Best model

Four methods: decision tree, logistic regression, Perceptron model and Multi-Layer Neural Networks Model, are used to predict whether a loan is 'good' or 'bad'. The best-performing model is Neural network with the highest accuracy(97.6%). The precision and recall of other methods are all more than 80%. And another important conclusion is that the accuracy can be improved sharply by using the selected attributes.

3.2 Strength

Compared with other analysis based on the LendingClub loan dataset, our group put a lot effort on preprocessing, especially in dimentionality reduction and class rebalance. So our models all got a over-average performance

3.3 Future improvement

1. When dealing with the columns with significant missing values, we remove the columns with over 40% missing values. But for future work, we still need to specify the exact threshold to remove the feature.

2. In wrapper method, we set the input with linear regression with 30 features due to the limitation of computer power. The ideal way is, firstly, figuring out the optimal feature number than applying logistic regression model to eliminate features.

3.4 Suggestions to platform

From our project, we find that data related to principal contributed the most to the final loan status, so LendingClub platform should be alert to borrower's principal amount. And with the help of the neural network model, P2P lending platform can judge whether current loan will turn into bad loan, so as to make a risk warning in advance and avoid liquidity risk.

REFERENCES

[1] Peter Renton, Renaud Laplanche (2012), The Lending Club Story ISBN 978-1-48113-173-5

[2] McCulloch, Warren; Pitts, Walter (1943). "A Logical Calculus of Ideas Immanent in Nervous Activity". Bulletin of Mathematical Biophysics. 5 (4): 115–133. doi:10.1007/BF02478259.

[3] Peter Renton, Renaud Laplanche (2012), The Lending Club Story ISBN 978-1-48113-173-5

[4] Alvarez, Daniel & Cerezo-Hernández, Ana & López-Muñiz, Graciela & Castro, Tania & Albi, Tomas & Hornero, Roberto & del Campo, Felix. (2017). Usefulness of Artificial Neural Networks in the Diagnosis and Treatment of Sleep Apnea-Hypopnea Syndrome. 10.5772/66570.

[5] Data Mining: Concepts and Techniques by Jiawei Han and Micheline Kamber, Morgan Kaufmann Publishers, 3rd Edition, 2011 (QA769.D343 H36)

[6] American Journal of Mining and Metallurgy. 2015, 3(3), 58-62 doi:10.12691/ajmm-3-3-1

[7] Dai, W., & Berleant, D. (2019). Benchmarking Deep Learning Hardware and Frameworks: Qualitative Metrics. ArXiv, abs/1907.03626.

[8] Michael A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015.

[9] Schindler, J., 2017. "FinTech and Financial Innovation: Drivers and Depth", Finance and Economics Discussion Series, Federal Reserve Board of Governors. Peer-to-peer lending: Advantages and disadvantages for loan customers. (n.d.). Retrieved from https://www.lendingworks.co.uk/finance-guides/p2p-lending/peer-to-peer-lending-advantages-disadvantages-borrowers. Peter Michael，2019

[10] Shen, Y and C Li (2018): "网☒借☒☒☒☒机制研究" (research on the risk relief mechanism of internet financing), Institute of Digital Finance, Beijing University.