7/11/2022

Predicting default for lending club data

Rashmi Kalra

STA 536

Contents

[**1.** **Introduction** 1](#_Toc108453842)

[**1.1.** **Background Information:** 1](#_Toc108453843)

[**1.2.** **Data Source:** 1](#_Toc108453844)

[**1.3.** **Project Scope** 2](#_Toc108453845)

[**1.4.** **Response variable definition** 2](#_Toc108453846)

[**2.** **Explanatory Data Analysis** 2](#_Toc108453847)

[**2.1.** **Lending Club Loan Amount:** 3](#_Toc108453848)

[**2.2.** **Lending Club Installment:** 4](#_Toc108453849)

[**2.3.** **Home Ownership** 5](#_Toc108453850)

[**2.4.** **Employment length** 6](#_Toc108453851)

[**2.5.** **Scatterplot of loan status against Annual income** 8](#_Toc108453852)

[**2.6.** **Annual Income and Loan Amount** 9](#_Toc108453853)

[**2.7.** **Analyzing loans by its purpose:** 9](#_Toc108453854)

[**3.** **Data Cleaning and Missing Imputation (Preprocessing)** 10](#_Toc108453855)

[**3.1.** **Data Cleaning:** 10](#_Toc108453856)

[**3.2.** **Missing value Imputation:** 12](#_Toc108453857)

[**4.** **Unsupervised Learning:** 12](#_Toc108453858)

[**4.1.** **4.1 PCA and Clustering Analysis:** 12](#_Toc108453859)

[**4.2.** **K-MEAN Clustering:** 15](#_Toc108453860)

[**5.** **Supervised learning:** 16](#_Toc108453861)

[**5.1.** **Training and Test Data:** 17](#_Toc108453862)

[**5.2.** **5.1 Logistic Regression:** 17](#_Toc108453863)

[**5.3.** **K- Nearest Neighbors:** 19](#_Toc108453864)

[**5.4.** **Single Classification tree:** 22](#_Toc108453865)

[**5.5.** **Random Forest** 23](#_Toc108453866)

[Important Features: Variable Importance 23](#_Toc108453867)

[**6.** **Summary** 24](#_Toc108453868)

[**7.** **Future Work** 25](#_Toc108453869)

**Lending Club Loan Data Mining Project**

### **Introduction**

### **Background Information:**

Lending Club is a United States peer-to-peer lending company that was founded in 2006. It is the world’s largest marketplace connecting borrowers and investors, where consumers and small business owners lower the cost of their credit and enjoy a better experience than traditional bank lending, and investors earn attractive risk-adjusted returns. Lenders do not lend directly to borrowers, as is the case with some other peer-to-peer lenders. Instead, borrowers apply for loans online and Lending Club reviews the loan application based on the borrower’s credit score, credit history, debt-to-income ratio, and other factors. From there, Lending Club rejects approximately 90% of the loan applications as the company has decided to only focus on high creditworthy borrowers. The loans can be from $500 to $40,000 for a 3- or 5-year term and if accepted, they are then placed on the Lending Club website for investors to browse. Investors then can choose to buy “notes” from Lending Club that are graded A to G according to the risk of default. The money that the investors provide is the money that the borrowers receive for their loan amount. These loans provide lower rates than a typical bank because of Lending Club’s low-cost structure with operations fully online and no branch infrastructure. Lending Club makes money off these loans by charging borrowers an origination fee and charging lenders a service fee.

### **Data Source:**

The data comes from Lending Club (<https://www.lendingclub.com/info/statistics.action>), which provides a very large, open set of data on the people who received loans through their platform. It is also available on Kaggle ([All Lending Club loan data | Kaggle](https://www.kaggle.com/datasets/wordsforthewise/lending-club)) and data dictionary can be found at  [Lending Club Data Dictionary | Kaggle](https://www.kaggle.com/datasets/jonchan2003/lending-club-data-dictionary?select=Lending+Club+Data+Dictionary+Notes.csv)

The lending club loan data contains loan data issued through 2015 to 2018. It also contains the loan status for the customer i.e., Current, Charged Off, Late Payments, Fully Paid, Etc. The dataset has 151 variables and 2.26 million records. Other features that the dataset includes are credit scores, address including zip codes, states, interest rate, loan amount, annual income, etc. The dataset also describes what is the purpose of the loan, i.e., for debt consolidation, credit card, home improvement, vacation, education, or wedding.

The response variable is Loan Status (Default or Not Default).

### **Project Scope**

Our project scope is to run the exploratory data analysis to find the business insights from our loan data, and to build a learning model using data mining that will use the historic loan data to learn and helps to identify loans/borrowers who are likely to default.

As per the recent studies, 3-4% of the total loans defaults every year. This is a huge risk for the investors who is funding the loans. Investors require more comprehensive assessment of these borrowers than what is presented by Lending Club to make a smart business decision. Data mining techniques/analysis could help predicting the loan default likelihood which may allow investors to avoid loan defaults thus limiting the risk of their investments.

### **Response variable definition**

Charged Off is assigned label 1 and Fully Paid is assigned label 0.

### **Explanatory Data Analysis**

Before the modeling process and EDA could start, there were some major challenges that had to be solved.

* First, the dataset contains information about the loans that originated on the Lending Club platform from 2015 to 2018.Since the initial size of the dataset was over 1.6GB. After parsing through the 151 columns and creating a data dictionary, we decided which fields to keep and which fields to reject.

A thorough understanding of the domain and all the variables is necessary to remove irrelevant variables, so I used the lending club’s data dictionary to identify and remove these variables.

Links:

LC Dictionary - <https://help.lendingclub.com/hc/en-us/articles/216127307-Data-Dictionaries>

* Second, our response variable was loan\_status (Charge off, Fully Paid and others). We want to predict if a given person will default or not so. This variable is Categorical in nature. Moreover, there are some more values present in the loan\_status but we are only interested in seeing Fully Paid and Charged off accounts. For this reason, rest of the values were deleted from the column.

Now we looked at the various plots to examine the relationship between the variables.

### **Lending Club Loan Amount:**

Firstly, we want to check the distribution of the loan amount. It is one of the most important and primary features in our dataset. The distribution is right-skewed as we would normally expect for this kind of variable. The maximum loan amount is $40,000 and the minimum is $1,000. The distribution is not normal, has multiple peaks, and there are some outliers at the right tail.

Chart, histogram

Description automatically generated

### **Lending Club Installment:**

The range of installments is wide, from $200 to $1,319 per month. Obviously, one would expect the installments to be highly correlated to the loan amount.

Chart, histogram

Description automatically generated

### **Home Ownership**

Next the home ownership variable is another crucial predictor of the default. Intuition would say that homeowners would be good borrowers and people on rent or mortgage would be bad investment. We deleted the ANY from this variable because of the few values.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

### **Employment length**

We converted the variable (emp\_length) from character to numeric. The variable had values like “1 Year”, “2 Years” …”10 years”. We changed the values to numeric 1-10 so that they can be better represented on the plot. Also we have replaced any value less than 1 year with 1.

Chart, bar chart, histogram

Description automatically generatedWe can see that majority of clients have 10+ years of experience and has highest number of defaulted loans.

### **Scatterplot of loan status against Annual income**

Chart

Description automatically generated

The figure above shows association between loan status and the annual income, the higher the annual income the higher the likelihood of the loan being fully paid. However, if the annual income is low, there is chances of the loan to be Charged Off(default).

### **Annual Income and Loan Amount**

Chart

Description automatically generated

The scatterplot above suggests the presence of outliers.

### **Analyzing loans by its purpose:**

Loans are taken mostly for debt consolidation and credit card payment. Whereas the debt consolidation has highest fully paid loan and we also see that the most defaulted loans are for credit cards and debt consolidation.

Chart, bar chart

Description automatically generated

### **Data Cleaning and Missing Imputation (Preprocessing)**

### **Data Cleaning:**

As already explained, we removed all irrelevant fields which do not contribute to data analysis or model building. In the next step, we looked for the outliers.

Outliers are observations that are abnormally out of range of the other values for a random sample from the population. To find out outliers I looked at the summary of the consolidated lending dataset. This helped understand that mostly none of the features had such abnormal observations, except for a couple of important ones like annual\_inc and DTI.

From the above graphics we can see that one of the observations for annual\_inc is outlier.

Here is the plot of Annual Income before and after the outlier removal.

Chart

Description automatically generated

Figure1-Before Outlier Removal

Chart, box and whisker chart

Description automatically generated

Figure2-After Outlier removal

We looked at the Debt-to-Income graphs which is shown below. We can see the presence of outliers. Next step is to remove those outliers. We removed the DTI values with value 0 and greater than 100.

Chart, box and whisker chart

Description automatically generated Figure3-After Outlier Removal

### **Missing value Imputation:**

Next step is to identify the missing values and impute them with the mean. We checked the columns and we saw that:

* revol\_util had 8 missing values
* Multiple missing values are also present in the employment length.

We removed the missing values by imputing those with the median values. At this step we are done with the data cleaning step as all the missing values and outliers are removed.

### **Unsupervised Learning:**

### **4.1 PCA and Clustering Analysis:**

Principal Component Analysis is a method of Dimensionality Reduction. Here we reduce a higher dimension to a lower one, retaining the variance of the original one. It involves the use of Eigen Vectors and a Covariance Matrix.

With the given dataset, it is imperative to reduce the dataset to smaller set of variables to derive a conclusion. With Multi-collinearity, two or more variables can share the same dimension. Each dimension can be viewed as 23-dimensional graph when the data is projected as orthonormal. Hence, PCA procedure is used to reduce the dimension of the data and it gives us a direction that approximately how many variables can be used to explain the data

Our first attempt at running PCA shows that the first PC accounts for about 29.63% of the variation in the data set. The first two PCs account for over 43.03% of the variation and 5 PCs explain about the 66.08% of the variation.

A picture containing text

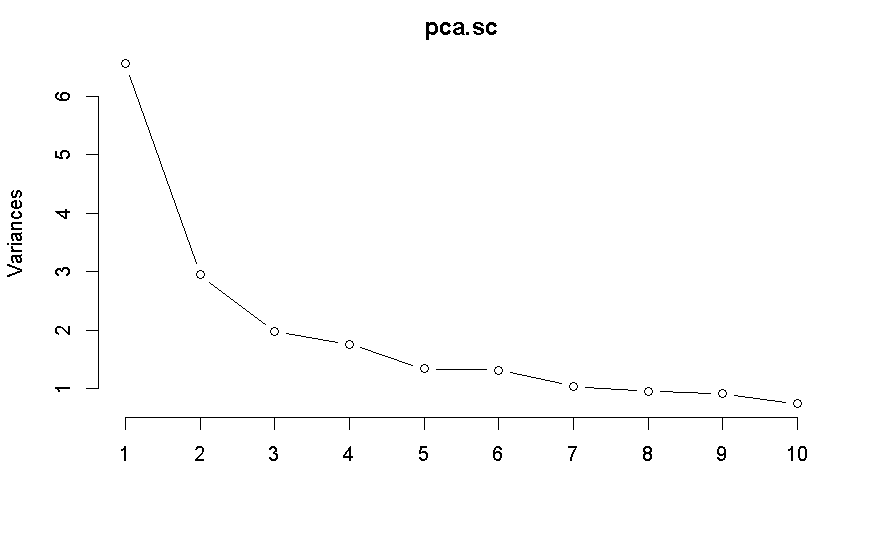
Description automatically generated

We also looked at it using the Biplot. A \*biplot\* plots the observations in the first two dimensions (PCs) of the new coordinate system. Effectively, PCA rotates the observations such the first dimension explains the most variation, the second variation explains the second most variation, etc. and we know that each dimension in this new coordinate system is perpendicular to each other (i.e., uncorrelated).

Chart

Description automatically generated

From the scree Plot we can see that the 5 PC explain nearly about 66.08% of the variation. Also, a scree plot shows the drop off in explained variance from one PC to the next. We typically choose the number of PCs based on the elbow of the plot. At that point we are not explaining enough additional variation to warrant including another dimension. Hence, for now we can ignore principal components greater than 3.



### **K-MEAN Clustering:**

After performing principal component analysis, the next step is using clustering and identify large clusters. Few iterations were tried with different cluster sizes.

We are running the K-means with K = 2 using the projected observations onto the first 2 PCs from the previous part.

In the next step, we did the clustering for K=3,4,5,6 and selected the optimal cluster using the silhouette plots. From the silhouette scores is highest is 0.23 at k=2 so the two clusters are optimal. In the below figure we are plotting the projected observation in two-dimension space and two different clusters are shown.

Chart, scatter chart

Description automatically generated

So, to summarize we have represented the 28 variables into two-dimensional space where PC1 is a new variable that explain the most variation in our dataset. Similarly, PC2 is another new variable that explain the second most variation in your data. Then we did the clustering which is grouping observations based on the 3 PC. That is, we find the observations which has most similar values in the principal components.

### **Supervised learning:**

Our classification goal is to predict which class the loan belongs to: either Default or Non-default. In the following sections, we will share and discuss our experiments using Logistic Regression, KNN and Random Forest for classification problem.

### **Training and Test Data:**

We will be building our model on the train data and making predictions on the train data to check how our model is performing. In order to do that we will be doing 10 folds Cross validation.

It consists in dividing the original set of observations into k subset of more or less same size. Then, we will use one of the subsets as test set and the remaining subsets will be used to form our training set. We will repeat this 10 times, where each time the subset used as test set will change. As an example, if we use 3-fold CV, your original set will be divided into k1, k2, k3. First, k1 will form the test set, k2 and k3 will form the training set. Then, k2 will form the test set, k1 and k3 will form the training set. Finally, k3 will form the test set, k1 and k3 will form the training set. For each fold, we output the results and then aggregate these to obtain the final result.

### **Logistic Regression:**

Logistic regression is useful for (discrete) qualitative responses referred to as categorical. It represents the probability that the response belongs to a category rather than telling about the outcome directly as default or not.

Below figure is the output from logistic regression model.

Graphical user interface, application

Description automatically generated

Looking at the output of the logistic regression we see that significant variables marked with asterisk. From the summary of our model, we can see that, judging by the significance parameter asterisk, there are some insignificant variables. Therefore, we can get rid of these insignificant variables and only include the significant variables. We modify our model as:

Text

Description automatically generated

Interpretation of logistic regression:

The regression coefficients tell us about the change in the log odds in the outcome for one unit increase in the predictor variable.

The estimates for the coefficients of the variables Interest Rate, DTI, delinquency in 2 years, inquiries in last two months, open accounts, public records and revolving util are positive meaning that an increase in one of these variables would reflect in an increasing chance of defaulting.

Example: For every one percent change in Interest Rate, the log odds of default increases by 4.485e-02.

### **K- Nearest Neighbors:**

This is the non-parametric regression approach. From the KNN approach we got the Accuracy of 91.96%. The final value used for the model was k = 17 which can be interpreted as: For each of the 10-folds, we can compare which class the KNN model from the training set would predict for the held-out test set to their actual category. Then, we can average over the 10-folds. This is the Accuracy. In other words, we correctly predicted the response of our test values about 91.96% of the time.

Then we predicted the probability of default or non-default for each of the observation and draw the confusion matrix with the actual and predicted good and bad.

The confusion matrix summarizes, for a given threshold, the number of cases in which:

* The model predicted a default and the borrower defaulted — **True Positive.**
* The model predicted a default, and the borrower did not default — **False Positive.**
* The model predicted no default, and the borrower did not default — **True Negative.**
* The model predicted no default and the borrower defaulted — **False Negative**

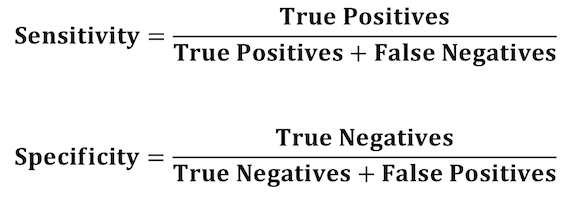
When we used a threshold of 0.50, we get a training accuracy of 0.9264323Notice that this is higher than the test accuracy. We get a sensitivity of 0.9415557 and specificity of 0.9187259. This means that we correctly predict the success(default)about 95% of the time and we correctly predict the failures(non-default) about 92% of the time.

Suppose we want to increase the sensitivity. Then, we can make it easier to predict defaults responses by decreasing the threshold. This decreases the overall accuracy to 0.9235464 but it increases the sensitivity to 0.988

For example, let us use the probability of 0.33 as our threshold, i.e., borrows with a probability greater than 0.33, are predicted to default, and borrowers with a probability below 0.33 are predicted not to default. Confusion matrix is given below:

* 

Senstivity and specificity is defined as below:



Thus, for out model we have:

Sensitivity = 7051/( 84 + 7051) = 0.988227

Specificity = 12470/(12470+1532)= 0. 8905871

We get a sensitivity of 0. 988 and specificity of 0. 890. This means that we correctly predict the success(default)about 99% of the time and we correctly predict the failures(non-default) about 89% of the time.

### **Single Classification tree:**

Tree models where the target variable can take a discrete set of values are called [classification](https://en.wikipedia.org/wiki/Classification) [trees](https://en.wikipedia.org/wiki/Decision_tree); in these tree structures, [leaves](https://en.wikipedia.org/wiki/Leaf_node) represent class labels and branches represent [conjunctions](https://en.wikipedia.org/wiki/Logical_conjunction) of features that lead to those class labels. Here our target variable is categorical in nature, and it used GINI Index for making the splits.

The tree looks like this:

Text

Description automatically generated with medium confidence

The first cut is at the variable last\_pymnt\_amnt with a value 1921. The first split is here because of the low GINI Index for this variable and value. If no split is made, then we got the value of GINI as 0.2235976. If the split is made, then the value is 0.04452728 which has improved as we can see. The Accuracy from the classification tree is 0.9751183. In other words, we correctly predicted the response of our test values about 97.51% of the time.

### **Random Forest**

The downside of using a single decision tree is that it tends to suffer from [high variance](https://www.statology.org/bias-variance-tradeoff/). That is, if we split the dataset into two halves and apply the decision tree to both halves, the results could be quite different.

One method that we can use to reduce the variance of a single decision tree is to build a [random forest model](https://www.statology.org/random-forests/), which works as follows:

**1.** Take *b* bootstrapped samples from the original dataset.

**2.** Build a decision tree for each bootstrapped sample.

**3.** Average the predictions of each tree to come up with a final model.

For our model, the Accuracy was used to select the optimal model using the largest value of 0.9792309. The final value used for the model was mtry = 10.

## Important Features: Variable Importance

Random forests can be used to rank the importance of variables in a regression or classification problem.

The most important variable is last\_pymnt\_amnt and least important variable is pub\_Rec.

**Predictions using Random Forest:**

Based on the above model we can now predict if any account will default in future or not. Example, we passed the below values to our model, and it came out that this account will default which was actually true.

Graphical user interface, text, application

Description automatically generated

### **Summary**

We ran the following models to predict if the account will default or not and it comes out that Random Forest method is the best because 98% of the time it is predicting accurately.

|  |  |
| --- | --- |
|  | Accuracy (%) |
| Logistic Regression | 94.659 |
| KNN | 91.37 |
| Classification Trees | 97.51 |
| Random Forest | 97.92 |

### **Future Work**

It would be great if lending club can provide the more recent samples to validate the model performance and stability. There have been a lot of macro-economic changes since 2018 like COVID and other changes like interest rates, new products, employment rate changes and inflation etc. that have led to significant changes in individual credit history. For example, payment plan/payment holidays, new products like buy now pay later have changed the behavior of how people shop for credit for their personal needs.