Matiullah Hasher

AIT 580

12/2/2019

Technical Report

Lending Club Loans Analysis

1. **Abstract**

**2.0 Introduction**

This aim of this project is to predict whether a borrower will **default** on a loan or not using LendingClub data. The project will extract the key features that have a direct relation to defaulting on a loan. With these key features, I will be using machine learning models to see how accurately I can predict whether the loan will be defaulted.

I am choosing to explore what factors are related to loans defaulting due to the astonishing amount of debt that Americans are in. According to Lending Tree, about 19 million individuals currently have a personal loan. **Personal loans** make up just under 1% of all outstanding consumer debt. The average annual percentage rate (APR) on personal loans taken in the first quarter of 2019 was 33.38% (please note the dataset being explored in this project is from 2008 t0 2018). Generation X has the highest loan amount, members of Gen X have an average personal loan amount of $9,722. [1].

**3.0 Objectives**

This aim of this project is to predict whether a borrower will **default** on a loan or not using LendingClub data. The project will extract the key features that have a direct relation to defaulting on a loan. With these key features, I will be using machine learning models to see how accurately I can predict whether the loan will be defaulted.

I am choosing to explore what factors are related to loans defaulting due to the astonishing amount of debt that Americans are in. According to Lending Tree, about 19 million individuals currently have a personal loan. Personal loans make up just under 1% of all outstanding consumer debt. The average annual percentage rate (APR) on personal loans taken in the first quarter of 2019 was 33.38% (please note the dataset being explored in this project is from 2008 t0 2018). Generation X has the highest loan amount, members of Gen X have an average personal loan amount of $9,722. [1].

**4.0 Dataset**

**4.1 Selection**

This project will be analyzing the **LendingClub** loan dataset. The data has been captured from LendingClub’s dataset on Kaggle. The dataset can be found [here](https://www.kaggle.com/wordsforthewise/lending-club).

**4.2 Description**

**4.2.1 Size**

This dataset contains over 2,260,668 records and a total of 100 **features**. data is over 2.6 GB when contained in a CSV file.

**4.2.2 Who**

This data was collected by LendingClub and posted to **Kaggle**. Lending Club is a service that connects borrowers and investors. LendingClub classifies as a loan trading company on the secondary market. It is the world’s largest peer-to-peer lending platform. The dataset used in this project collected from 2007 until the end of 2015. By the end of 2015, Lending Club had claimed to have originated $15.98 billion in loans. LendingClub allows borrowers to create unsecured personal loans between $1,000 and $40,000. LendingClub also does traditional loans such as auto-loans and mortgages. These loans are not funded by investors, instead they are held by other financial institutions.

**4.2.3 Need**

**4.2.3.1 Why was this data collected?**

This data can be used to figure out what are the common causes of defaulting on loans, one can improve decisions on who to lend to.

**4.2.3.2 Why questions can be answered by studying this data?**

1. What is the ratio of loans paid off to loans charged off?
2. Is there a relationship between loan amount, interest rate, and loan status? Do loans that are charged off tend to be higher in interest rate and loan amount?
3. Is there a difference in distribution of number of open credit accounts for customers of loans that are fully paid as opposed to customers of loans that are charged off?
4. Is there a difference in distribution of loans that are income verified or not for fully paid loans compared to charged off loans?
5. What is the correlation between the plethora of features available in this dataset for the loan/applicants from whom the data was collected?
6. How accurate of a model can be created using a logistic regression?
7. Are there 3 well defined clusters within the data when analyzing interest rate, and number of credit inquiries in the past year for charged off loans?
8. How well can the loan status (charged off or paid off) be predicted using a Gradient Boosting Classifier model?

**4.2.3.3 What are the privacy and ethical issues with this data?**

Although there are no privacy concerns with this data, there is an ethical dilemma as accurately predicting who should be able to know who should be able to pay their loans back on as planned may limit others from having access to loans as well.

**4.3 Dataset Schema**

|  |  |  |
| --- | --- | --- |
| **LoanStatNew** | **Description** | **Data Type** |
| acc\_open\_past\_24mths | Number of trades opened in past 24 months. | float64 |
| all\_util | Balance to credit limit on all trades | float64 |
| annual\_inc | The self-reported annual income | float64 |
| application\_type | individual application or joint application | object |
| avg\_cur\_bal | Average current balance of all accounts | float64 |
| bc\_util | Ratio of total current balance to credit limit for bank cards | float64 |
| chargeoff\_within\_12\_mths | Number of charge-offs within 12 months | float64 |
| collections\_12\_mths\_ex\_med | Number of collections in 12 months excl medical | float64 |
| delinq\_2yrs | No. past-due incidences of delinq for the past 2 years | float64 |
| dti | debt-to-income | float64 |
| emp\_length | Employment length in years. | object |
| emp\_title | Job Title | object |
| funded\_amnt | The total amount committed | int64 |
| grade | LC assigned loan grade | object |
| home\_ownership | home ownership status | object |
| inq\_last\_12m | Number of credit inquiries in past 12 months | float64 |
| installment | The monthly payment | float64 |
| int\_rate | Interest Rate on the loan | float64 |
| loan\_amnt | The listed amount of the loan applied for by the borrower. | int64 |
| loan\_status | Current status of the loan | int64 |
| num\_accts\_ever\_120\_pd | Number of accounts ever 120 or more days past due | float64 |
| num\_tl\_90g\_dpd\_24m | No. of accs 90 or more days past due in last 24 months | float64 |
| num\_tl\_op\_past\_12m | Number of accounts opened in past 12 months | float64 |
| open\_acc | The number of open credit lines in the borrower's credit file. | float64 |
| pct\_tl\_nvr\_dlq | Percent of trades never delinquent | float64 |
| percent\_bc\_gt\_75 | Percentage of all bankcard accounts > 75% of limit. | float64 |
| pub\_rec | Number of derogatory public records | float64 |
| pub\_rec\_bankruptcies | Number of public record bankruptcies | float64 |
| purpose | A category provided by the borrower for the loan request. | object |
| revol\_bal | Total credit revolving balance | int64 |
| revol\_util | Revolving line utilization rate | float64 |
| tot\_cur\_bal | Total current balance of all accounts | float64 |
| verification\_status | Was income was verified by LC | object |

**4.4 Data Pre-Processing**

The following Pre-Processing techniques were used in preparing the data to be visualized/analyzed:

* + 1. Dropping unnecessary features – after realized that the columns: ‘out\_prncp’ and ‘out\_prncp\_inv’ ended up not being needed, pd.drop was used to remove them.



* + 1. Splitting the dataset based on loan status to further analyze.



* + 1. Converting the ‘loan\_status’ column which contained categorical data in object form to numeric form replacing the “Fully Paid” loans with “0” and the “Charged Off” loans with “1”. This was necessary to run regression analysis, correlation analysis, and for the prediction model using Gradient Boosting Classification.

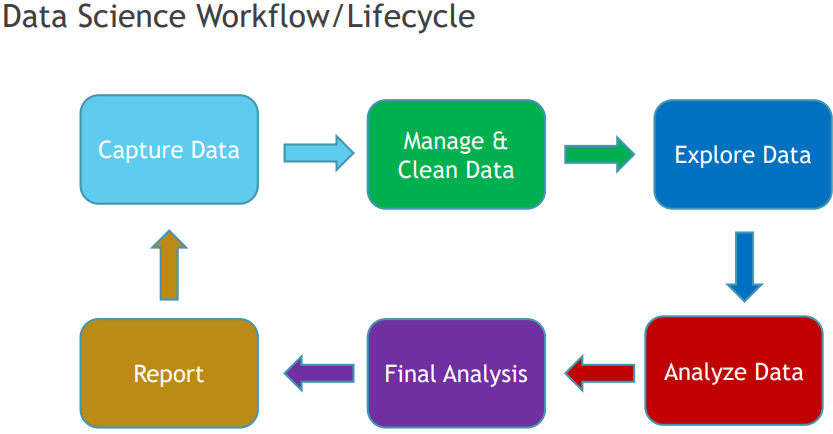


* + 1. There were rows of data that had infinite values and null values. This was an issue when applying machine learning techniques such as regression and classification. To remove them, the infinite values were converted to null and then all null values were dropped.



**5.0 System**

**5.1 Architecture**



The following steps occurred within the Data Science Workflow/Lifecycle:

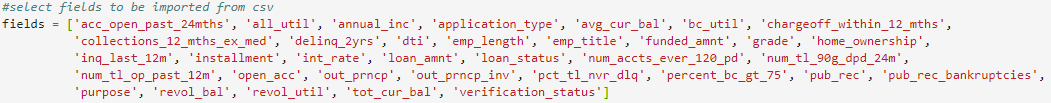
**5.1.1 Capture Data**

The data has been captured from LendingClub’s dataset on Kaggle. The dataset can be found [here](https://www.kaggle.com/wordsforthewise/lending-club). The dataset is a csv file which was read in using Pandas’ pd.read\_csv feature.

**5.1.2 Manage & Clean Data**

The data dictionary contained all the columns in the dataset along with an explanation of what those columns contained. This allowed me to select which fields to use and only bring those columns in as opposed to bringing in the entire dataset which would be a lot less efficient. The dataset imported contained 35 columns as opposed to the 100 contained in the original dataset. The steps for reading in only the chosen features are below:

**5.1.2.1** Select the fields to be imported by placing the column names in a list.



Read in the data using read\_csv’s usecols attribute to specify the fields selected in the list.



**5.1.3 Data Exploration**

The following data exploration techniques were used:

**5.1.3.1** Viewing all the columns when viewing the first five rows to get a better sense of the dataset.



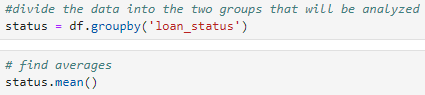
**5.1.3.2** Explore the number of rows and columns.



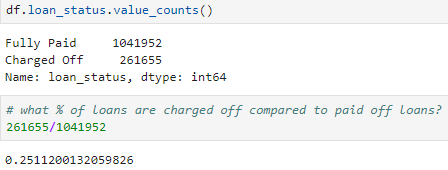
**5.1.3.3** Explore the data types of the features (output can be seen in Excel table above in section 4.3 Data Schema).



**5.1.3.4** Explore the mean by type of loan status.



**5.1.3.5** Determine what percentage of loans are charged off compared to paid off.



**5.1.3.6** **Data Visualization**

Data visualization techniques were used to explore the data. Bar charts, grouped bar charts, distribution plots, violin plots, scatter plots, and heat maps were all used to visualize the dataset. The visualizations were created by the following Python libraries: Pandas, Matplotlib, and Seaborn.

**5.1.4 Data Analysis**

The two primary methods of analyzing the data was correlation analysis and K Mean Clustering analysis. These two methods provided information necessary to use in the final analysis portion below which is the predictive portion of the analysis.

**5.1.4.1** Finding the correlation between the features of the dataset.



A correlation over 0.7 is considered strong. Below is a table of correlations within the features that are considered strong.

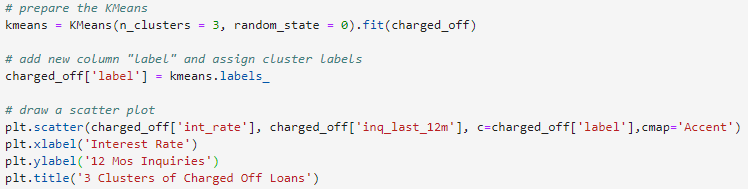
|  |  |  |
| --- | --- | --- |
| **Feature 1** | **Feature 2** | **Correlation** |
| loan\_amnt | installment | 0.95 |
| bc\_util | percent\_bc\_gt\_75 | 0.85 |
| bc\_util | revol\_bal | 0.84 |
| acc\_open\_past\_24mths | num\_tl\_op\_past\_12mn | 0.76 |
| revol\_util | percent\_bc\_gt\_75 | 0.84 |

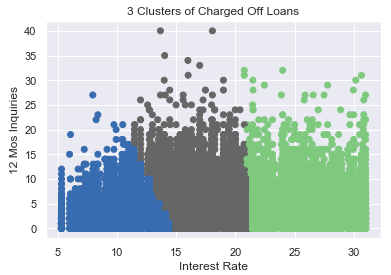
The correlation is strong (.95) with **loan amount** and **installment** as the installment represents the monthly payment which will be obviously higher the larger the loan is. The reason this is not directly a 1.0 correlation is due to the fact that customers have the option of choosing between a 3 year loan or a 5 year loan, which makes a difference in the installment.

The correlation is strong (0.85) between **bank card utilization** and **percentage of bank cards with over 75% utilization** as they are almost measuring the same thing. If an individual has high overall bank card utilization, then it is very likely that a good portion of their bank cards is utilized over 75% leading to this strong correlation. This same assumption can be made for **revolving utilization**, which measures the utilization of all revolving credit, and **percent of bank cards utilized greater than 75%**  as there is a strong correlation (0.84) there as well.

**5.1.4.2** Cluster Analysis

Cluster analysis was conducted on charged off loans. The two features that were analyzed were **interest rate** and **number of credit inquiries in the past 12 months**. These features were selected as they tend to be red flags in different perspectives. A credit inquiry is an indicator of uncertainty, and that equates to possible increase in risk for a lender. Higher interest rates are given to borrower’s who have lower credit worthiness compared to lower interest rates which are given to borrower’s who have proved to be able to make their payments on time. The code for the analysis along with the result of this cluster analysis is below.





**5.1.5 Final Analysis**

There were two portions to the final analysis for this project- finding the solution to the hypothesis test and predicting whether a loan will be fully paid or charged off. The solution to the hypothesis test was found by conducting a one-sample t test. There were two types of machine learning models used to predict whether a loan will be fully paid or charged off - binary logistic regression and Gradient Boosting Classifier. Section 5.1.5.2 and Section 5.1.5.3 will describe these models and go over the algorithms in detail.

**5.1.5.1 Hypothesis Testing**

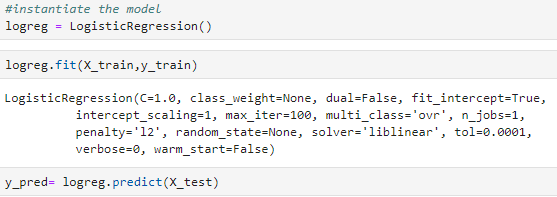
The hypothesis that will be tested is the following: For loans that have defaulted, the mean interest rate is 16.00%. This is the chosen hypothesis as the mean interest rate of the charged off loans was 16.03%. This hypothesis will validate whether the population sample being tested with t test will reflect the overall mean. T t-test conducted resulted in a p-value of 0.0549 which is greater than 0.05 so the hypothesis was not rejected.

**5.1.5.2 Binary Logistic Regression**

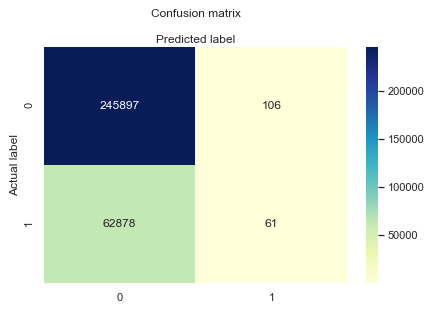
The regression analysis was conducted to determine whether a loan would be charged off or fully paid. The dataset used was filtered down to only include loans with these two statuses. The data was split into training and testing data as shown in the code below.



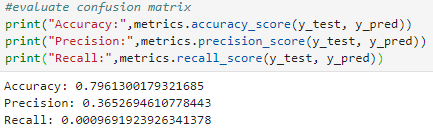
75% of the data will be used to train the model, the remaining 75% will be used to test the model’s accuracy, precision, and recall. The model variables above were fit to the logistic regression model and the predictions were made.



The confusion matrix below was created using Seaborn to visualize the results of the model when inputting the test data.



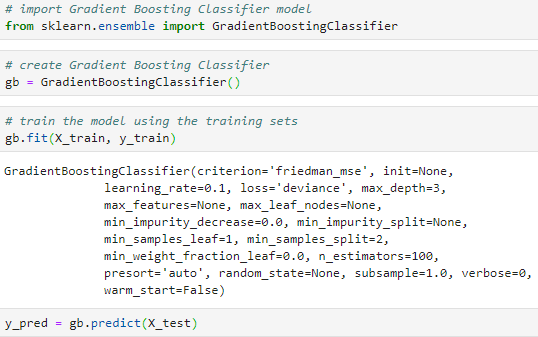
The confusion matrix reveals that there were 245,897 correct predictions of loans being fully paid and 61 correct predictions of loans being charged off. There were 62,878 incorrect predictions of loans being fully paid and 106 incorrect predictions of loans being fully paid. This model heavily predicted that loans will be fully paid – over 99.9% of the predictions were that the loans would be fully paid. This model did not do well in that aspect at all. Below are the metrics when evaluating the confusion matrix.



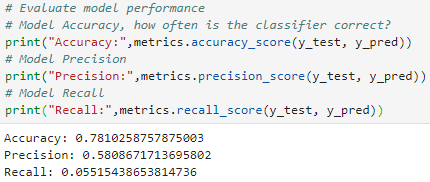
The accuracy of the model is 79.6% which is considered good accuracy. However, this accuracy can be misleading as the precision and recall scores are examined. Precision measures how accurate a prediction is out of those predicted positive. Since almost all the loans were predicted to be fully paid, the precision score of 0.36 is very low as expected. The recall score is a similar metric, it calculates how many of the actual positives the model captures through labeling it as positive. The recall score was extremely low, 0.0009, since the model only predicted that 167 loans were charged off which was very far off from the actual figure which should be around 25% (test data ratio) of 261,655 (entire charged off loans in dataset). Evaluating these metrics proves this model is useless for determining whether a loan is fully paid or charged off.

**5.1.5.3 Gradient Boosting Classifier**

The data was prepared like the logistic regression when conducting the gradient boosting classifier method. The code below shows how the model was used to make the predictions.



Below are the evaluation metrics of the model.



This model is slightly less accurate, 78.1%, compared to the logistic regression model which had an accuracy of 79.6%. However, the model is a far superior when evaluating the precision and recall scores. The recall score is significantly better at 0.05 compared to 0.0009 for the logistic regression model, but still not nearly an acceptable score. The recall score is why this model is not recommended either for determining whether or not a loan will be fully paid or charged off.

**5.2 Software Development Platforms**

This project was conducted primarily in a notebook in Jupyter Lab. JupyterLab is a web-based user interface which enables the user to write and execute code from notebooks.

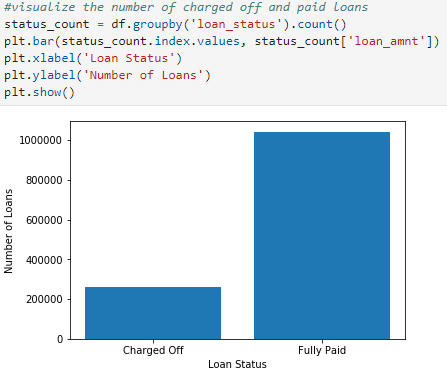
The code used in the Jupyter notebooks was Python. The Python libraries used were pandas, numpy, matplotlib, seaborn, sklearn, scipy, and IPython. Pandas was used for descriptive statistics, data manipulation and data analysis. Numpy was used for data manipulation and statistics. Matplotlib and seaborn were used to visualize the data. Sklearn was used in the predictive models. Scipy and IPython were used in the hypothesis testing.

**6.0 Experimental Results and Analysis**

Below is a summary of the data visualization aspect of this project.

**6.1 Bar Chart**

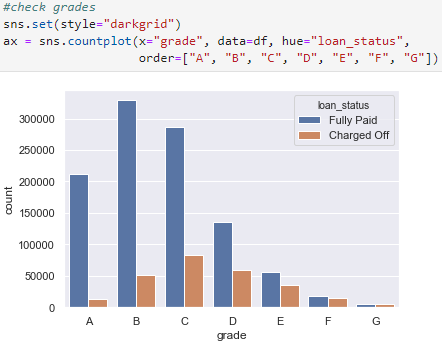
The code below visualizes the number of charged off loans and paid off loans with a bar chart using matplotlib.



The bar chart illustrates that a lot more loans are fully paid than are charged off. This is verified in section 5.1.3.5 of this report which states that the ratio of fully paid loans to charged off loans is roughly 3:1.

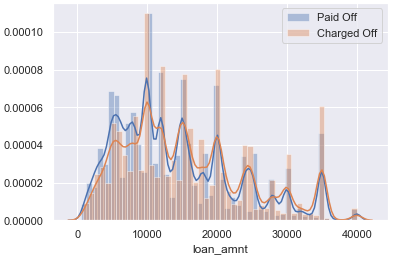
**6.2 Count Plot**

The count plot below visualizes how the distribution of the loans is within the **grade** feature. The grade is given by the Loan Counselor upon approval of the loan. Grades are on ranked in alphabetical order with “A” being the best grade and “G” being the worst grade. The count plot below shoes that fully paid loans are mostly in the A-D range. Most fully paid loans have a B grade while the majority of charged off loans have a C grade.



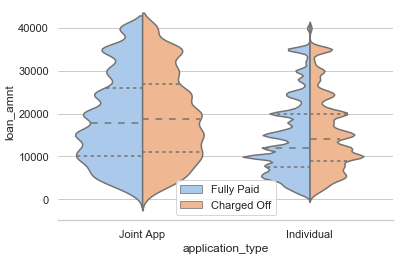
**6.3 Dist Plot**

The dist plot below was created using the Seaborn library. It shows the distribution of paid off loans and charged off loans based on loan amount and interest rate on a normalized scale. This reveals that lower loans tend to be paid off more often that they are charged off. As the loan amount increases, so does the likelyhood that the loan will be charged off.



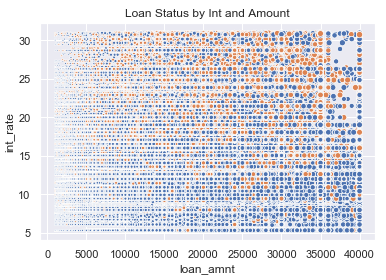
**6.4 Violin Plot**

The violin plot below was created using the Seaborn library. It visualizes the relationship between loan amount and application type (individual or joint) based on whether the loan is fully paid or charged off. The width of the violin plot is based on the amount of data. This violin plot reveals that loans with higher amounts tend to have joint applicants while it was very rare for an individual to get approved for loans on the higher end.



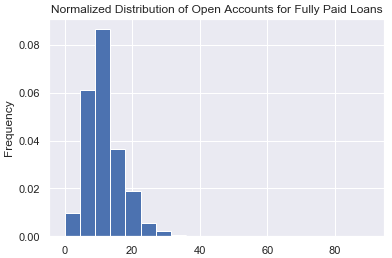
**6.5 Scatter Plot**

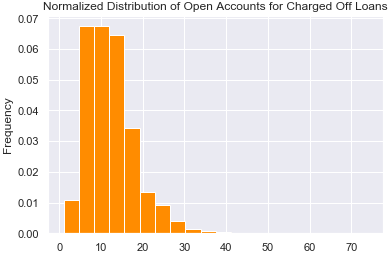
The scatter plot below shows the relationship between fully paid loans, represented by blue circles, and charged off loans, represented by orange circles, when visualizing their loan amount and interest rates. This scatterplot shows that the majority of loans that are charged off are higher than 13%. The higher the loan amount is the more likely the loan is to be charged off as well. The scatter plot also reveals that Lending Club does not have many loans in lower amounts or lower interest rates. The lower and left portion of the plot has a lot less avalues.



**6.6 Histogram**

The two histograms below show the distribution of how many open accounts the borrowers have. The visualizations reveal that borrower’s who have more open accounts are more likely to charge off a loan compared to borrower’s who have fewer open accounts.





**Conclusions**

Correlation Analysis

Loan Amount and installment – **.9533**

Bc\_util (bank card utilization) and revol\_bal - **.8542**

Bc\_util and percent\_bc\_gt\_75 -

Acc\_open\_past\_24mths and num\_tl\_op\_past\_12m - **.7579**

Revol\_util and percent\_bc\_gt\_75 - .84

Pub\_rec and pub\_bankruptcies - .6493