Lending Club Case Study

Ashish L Parmar and Eeshan Gupta

Problem Statement

Problem Statement

Lending Club is an online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.

Like most other lenders, Lending Club too wants to **minimize their risk**. The risk in their case are customers who took loans and now refuse to pay or have run away with the money. Such applicants are called **defaulters** and their loan is then **charged-off**. Identifying such risky application before issuing loans will reduce the credit loss.

Lending Club wants to understand the **driving factors** behind loan defaults and utilize this knowledge for its portfolio and risk assessment

Data & Analysis Approach

Data & Analysis Approach

- ❑ Data is provided to us by the Lending Club
- ❑ The data is a subset starting from **2007 till 2011**
- ❑ The data is also accompanied by data dictionary which provides a description of the columns present in the data
- ❑ The data has approximately **40000 rows and 111 columns**
- ❑ Among these 111 columns the target column is '**loan_status**'

Our approach to tackle this problem is as follows:

- ➔ Load and Clean the data
- ➔ Conduct an univariate analysis followed by segmented univariate analysis
- ➔ Derive metrics from the selected columns
- ➔ Based on selected and derived columns and conduct bivariate analysis

Cleaning the data

Cleaning the data

We loaded the complete data. Then proceeded based on missing value analysis

- Initially we had 39717 rows and 111 columns
- Missing value analysis yielded the following
  - More than 50 columns have almost no values
  - We choose drop them
  - Furthermore there were more columns where majority of rows were empty
  - We chose to drop any column where the more 60% of values are missing
- Finally after conducting missing value analysis, we choose to remove 57 columns based %age of missing values

Cleaning the data

- After removing the mostly empty columns, next we observed columns which were populated with only 1 values
- We found that 9 columns were populated only with one value
- Since there is no information stored in the column, therefore we chose to drop them
- Finally after dropping these columns, we are left with 44 columns
- We then clean the remaining columns by changing the columns to appropriate data type
- We extracted integers and floating point numbers amid the string
    - For example, extracting interest rate from a columns containing floating point number and a % sign
- We also extracted month and year from many date columns present
- We observed a text column which had short description of loan application, but had no structure, so no standard technique can be used to extract information. Therefore, we chose to drop this column

Univariate and Segmented Univariate Analysis

Univariate Analysis

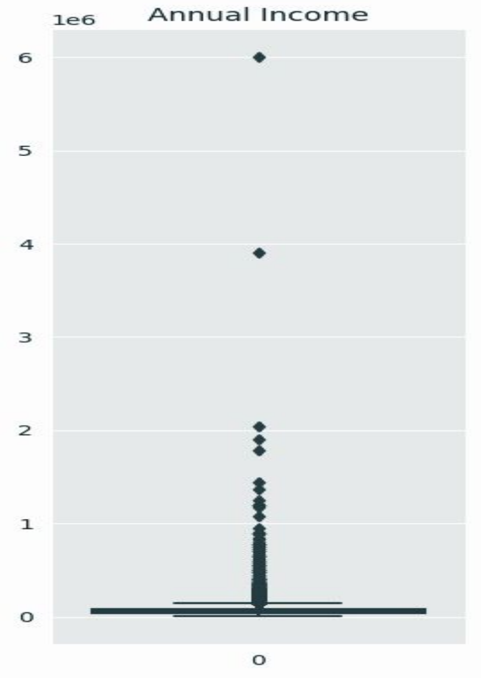
- We studied the distribution of all the remaining columns
- Some columns had outliers, which made their analysis skewed. We choose to drop those rows which had outliers based on the a threshold value

Annual Income

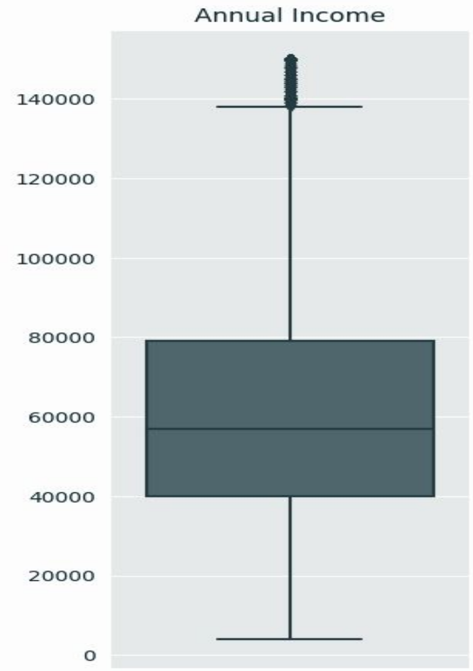
Column ‘**annual_inc**’ contains the self reports income of the borrowers

Originally, the 75th percentile was only **$82,300**, but the maximum value was **$6,000,000**. The difference was huge and therefore we chose an upper limit of **$150,000**

Doing reduced the number of rows by ~**1500**



Before removing outliers



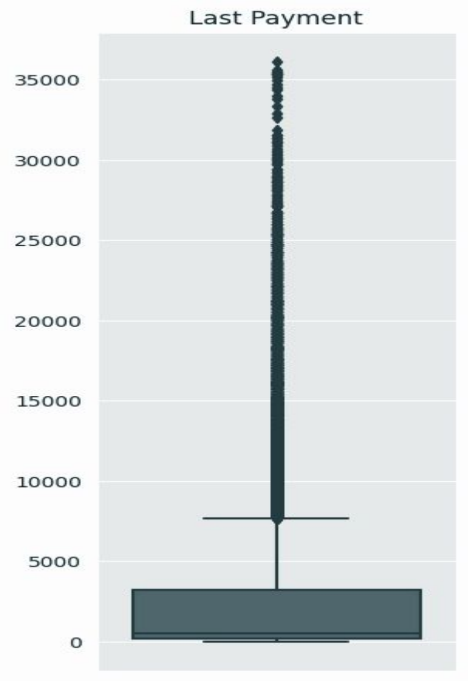
After removing outliers

Last Payment Amount

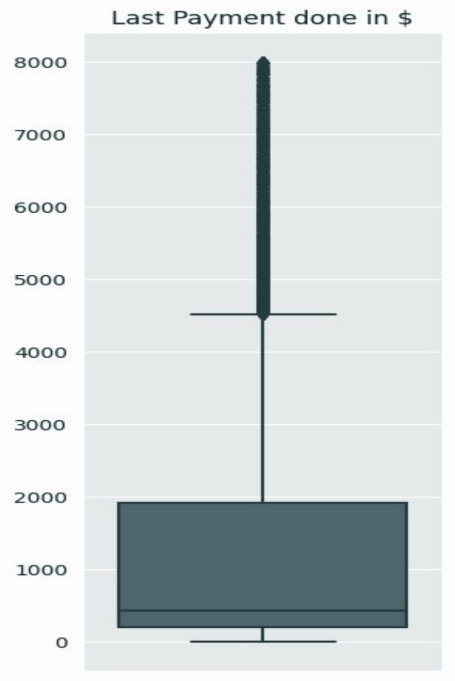
Column ‘**last_pymnt_amnt**’ contains the value of last total payment amount received

Originally, the 75th percentile was only **$3,201**, but the maximum value was approx. **$32,000**. The difference was huge and therefore we chose an upper limit of **$8,000**

Doing reduced the number of rows by **~3500**



Before removing outliers



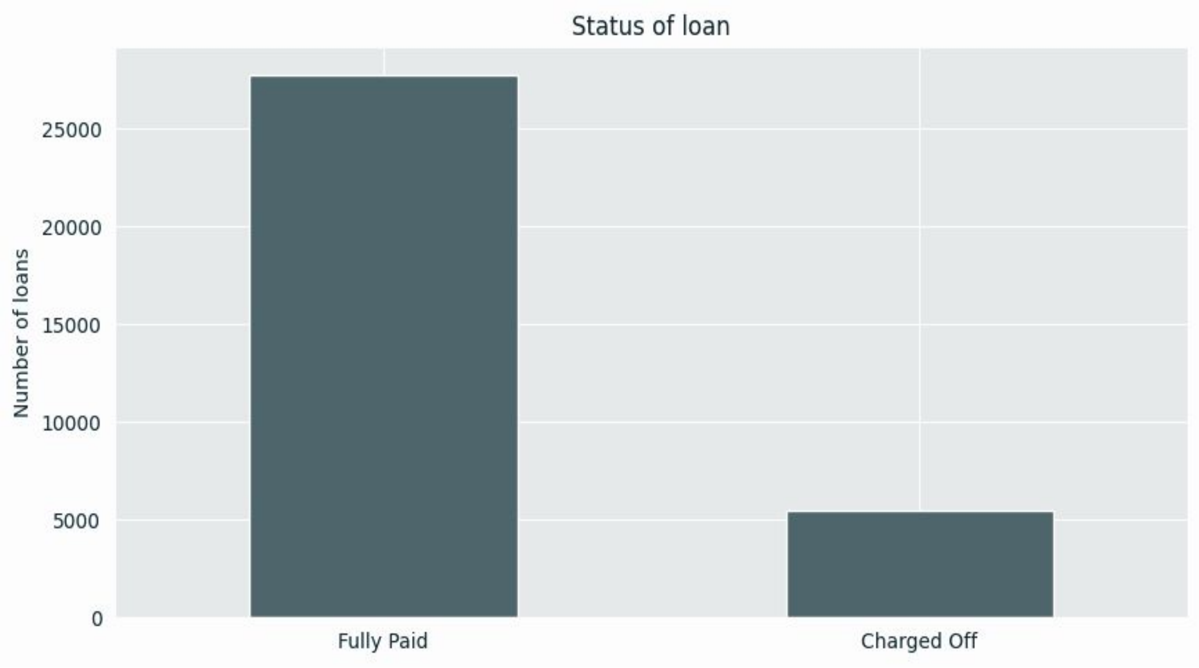
After removing outliers

Loan Status

Column ‘loan_status’ is our target column. It contain the current status of the loan, whether it is **fully paid**, or **charged-off**, or is it still ongoing (**current**)

According to problem statement, the focus is only on the risky loans. We choose to find patterns for the statuses of completed loans only by dropping the loans which are still ongoing

This further reduced ~1000 more rows



Insights from Univariate Analysis

- 50% of the applicant only wanted to borrow $8000, and only 25% wanted more than $12800
- Most of the loan amount is funded by investors, only a small percentage of the loan is funded by the company
- There are only 2 terms of loan, 36 month and 60 month. Number of short term loans are high as compared to long term loans
- Interest rate of the loans vary from 5% to 24%, but 50% of the loans have rate in between 8.8% and 14.1%
- Median monthly installments is ~$250 per month and 75% of the borrowers had to pay less than $375 per month
- Most of the loans are of grade A, B and C
- US Army personnel borrow the most from Lending Club
- A large chunk of borrowers have more than 10 years of experience
- The most common purpose for applying for loan is Debt Consolidation
- Most loans are issued for the applicants residing in the state of California
- Median Debt-to-Income ratio is 13.3, which is quite good
- Since 2007, number of loans have steadily increased

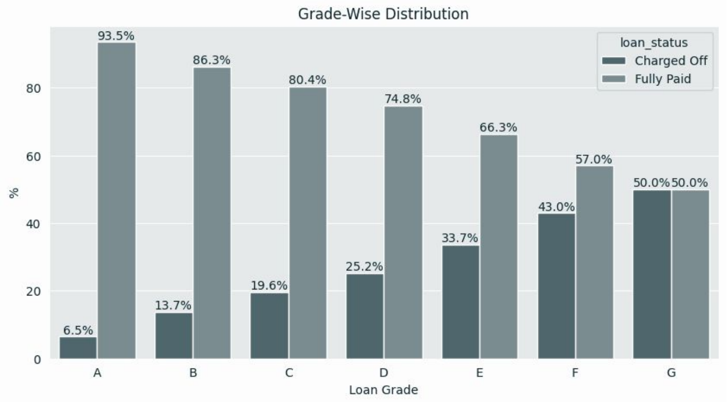
Segmented Univariate Analysis

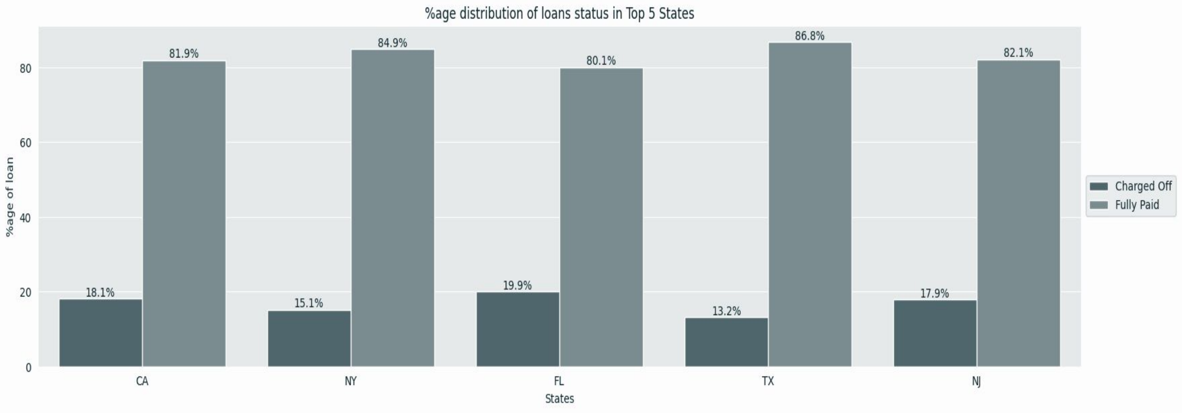
Based on categories, we compared the some aggregate values of columns

This yielded some interesting insights and helped us select some columns which are drivers

Loan Status

- **~84%** of loans are **Fully Paid** or good or not risky. **~16%** are risky loan labelled as **Charged-Off**
- The median value loan amount, funded amount and amount funded by investors are usually **higher for Charged-Off** loan than for Fully Paid ones by **~$2000**
- The mean and median interest rate for charged-off loan is higher by ~2.5%
- There is a difference of ~$40 in the median monthly installments for charged-off and fully paid loans
- Charged-off loans have a higher median revolving utilization as compared to Fully Paid loans
- Loans issued for the purpose of small business have a higher chance to default than any other purpose

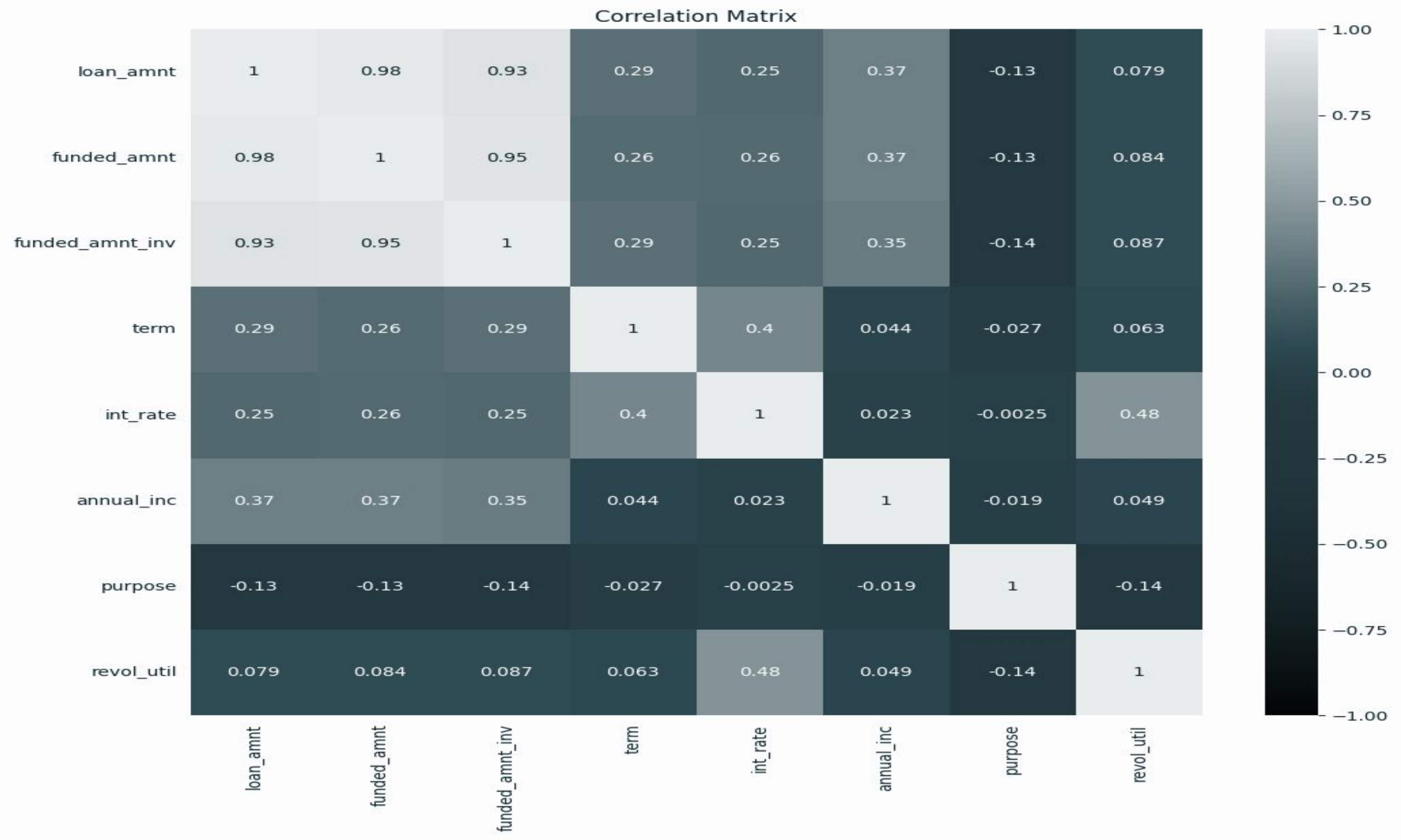




Bivariate Analysis

Bivariate Analysis

- We selected a handful of columns which drive the risk of a loan
- Some of them were numerical while others were categorical
- We chose to replace the categories with numbers
- We then calculated the correlation among each other



	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	annual_inc	purpose	revol_util
loan_amnt	1	0.98	0.93	0.29	0.25	0.37	-0.13	0.079
funded_amnt	0.98	1	0.95	0.26	0.26	0.37	-0.13	0.084
funded_amnt_inv	0.93	0.95	1	0.29	0.25	0.35	-0.14	0.087
term	0.29	0.26	0.29	1	0.4	0.044	-0.027	0.063
int_rate	0.25	0.26	0.25	0.4	1	0.023	-0.0025	0.48
annual_inc	0.37	0.37	0.35	0.044	0.023	1	-0.019	0.049
purpose	-0.13	-0.13	-0.14	-0.027	-0.0025	-0.019	1	-0.14
revol_util	0.079	0.084	0.087	0.063	0.48	0.049	-0.14	1

Correlation Matrix

Conclusion

CONCLUSION

We conducted our analysis on the loan dataset provided by Lending Club to identify the factors that drive the risk of loans. We followed multiple steps and eliminated variables as when necessary. At last we conducted the bivariate analysis, comparing each variable to another and we conclude the following:

- **Loan Amount** is an important driving factor. It is highly correlated to **Funded Amount** and **Investor Funded Amount**
    - This is obvious as the loan application if approved will only approve the amount asked by the borrower and investors will only invest accordingly
    - Therefore we are choosing to keep only **Loan Amount** as the factor
- The factors '**Revolving Utilisation of the credit accounts'** and '**Interest Rates'** also show a slight positive correlation
    - The reason why they show a positive correlation is that, the more a person uses the Revolving credit, he/she is more likely to be in more debt thus more likely to charge off or default in the EMI payments.
- '**Interest Rate**' also show a slight positive correlation to '**Term**' of the loan
- '**Annual Income**' of the borrower and '**Purpose**' of the loan are also the factors we chose, which drive the risk of a loan