Lending Club Case Study

Submitted by:
Mayank Joshi

The basis idea for the analysis is to understand the driving factor leading to loan default.

The analysis is divided into four main parts:

- Data understanding
- Data cleaning (cleaning missing values, removing columns which are not used for analysis etc.)
- Data Analysis
- Recommendations

Data Understanding

- In the given dataset, there are 111 columns with total rows as 39717
- The data set is from past loan applications which show the status of Default, Current or Fully paid.
- We have 2 types of variables
  - One which gives information about the customer behavior while paying the loan. We will ignore those columns since those will not be available for new loan applications
  - Second types of variables are which impact the loan decision for the company like employment length, home ownership, verification status. We will use these data columns for our analysis.

Data Cleaning

- We started by finding columns which have all values as null. Those columns were removed from the data set.
- Few other columns had more than 90% of values as blank those were also removed.
- Some columns like Employee title which mostly had 1 unique value were also removed.
- Some rows were having employment length as unavailable so we removed those rows also as we had ample amount of data to analyze further

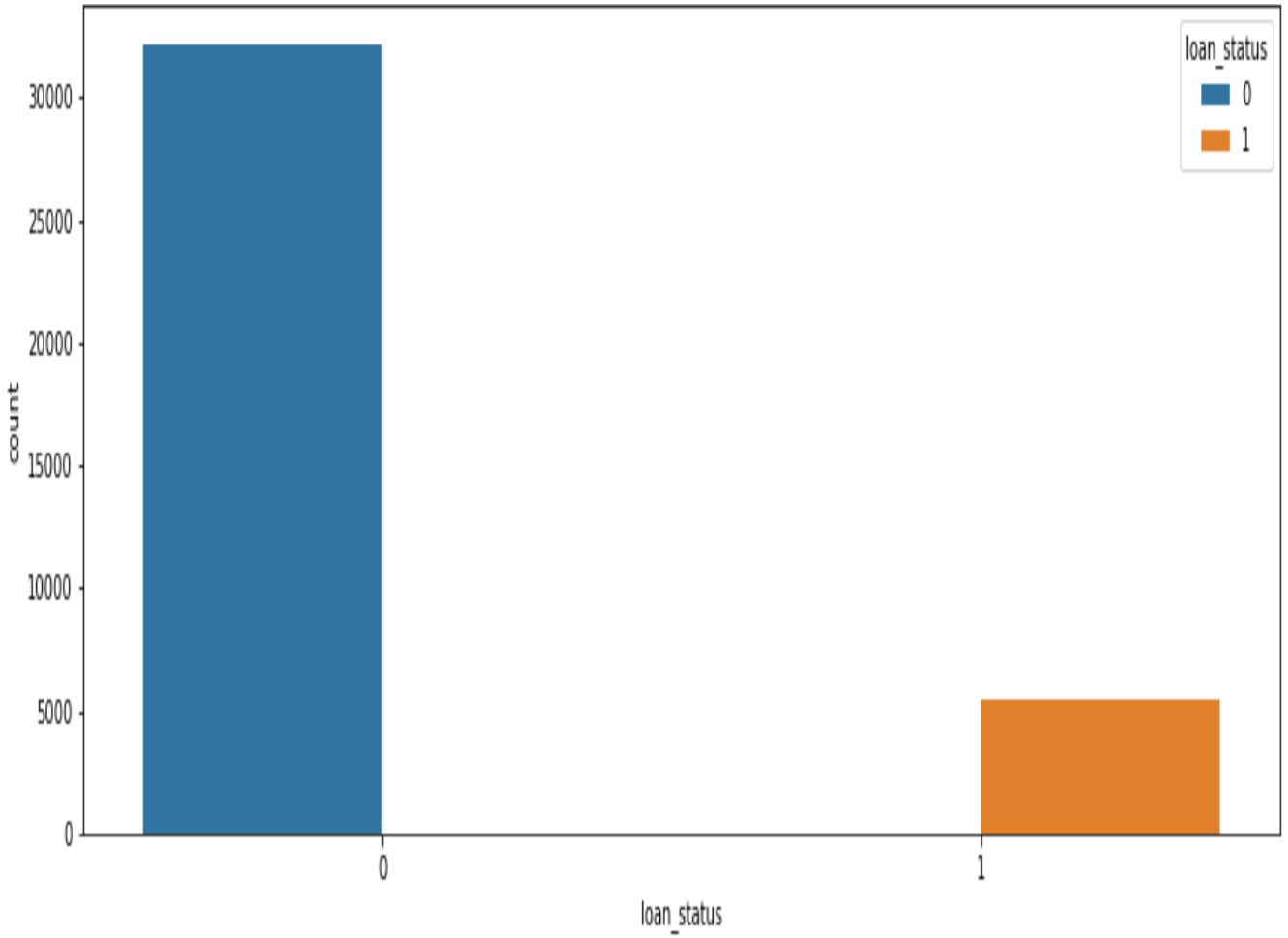
- From loan status value as “Current” we couldn’t determine if it will be defaulted or paid off so we removed rows with loan status value as Current.
- From issue_d column we derived month and year so that those can also be used for our analysis.
- Data types for different columns was also changed from object type to more logical dtype e.g Int_rate, loan_amnt, annual_inc

Data Analysis

- Loan data distribution

The analysis showed that 14% of the total loan applications were defaulted(fig is a countplot of loan status).

We will further analyze various factors effecting the loan to get default or paid off.



- Univariate analysis:

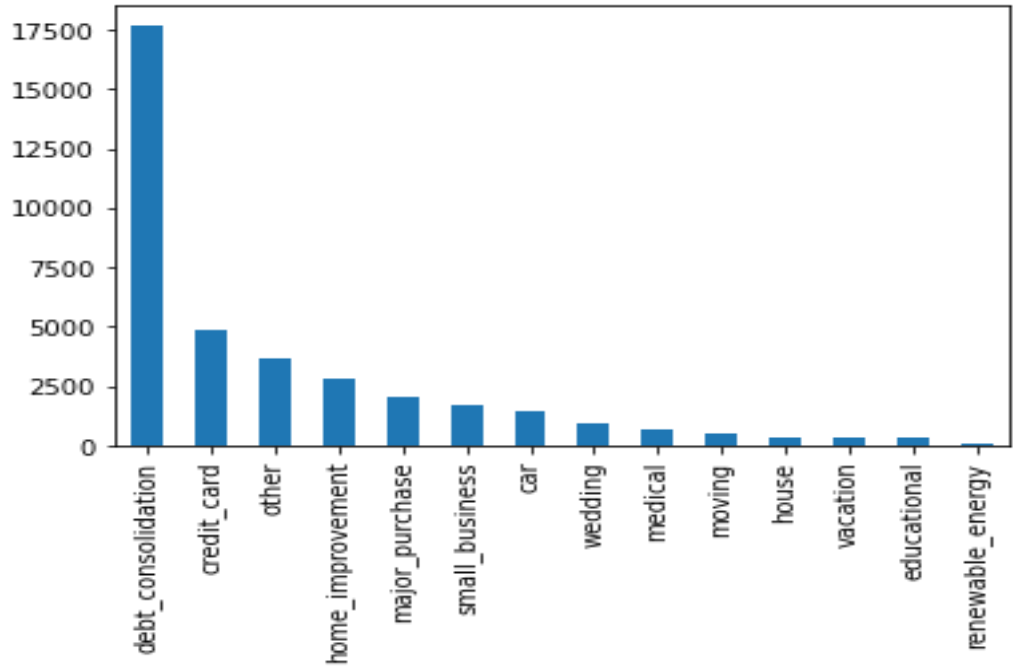
We analyzed different columns to see the pattern of data from excel. Below are some of the observations:

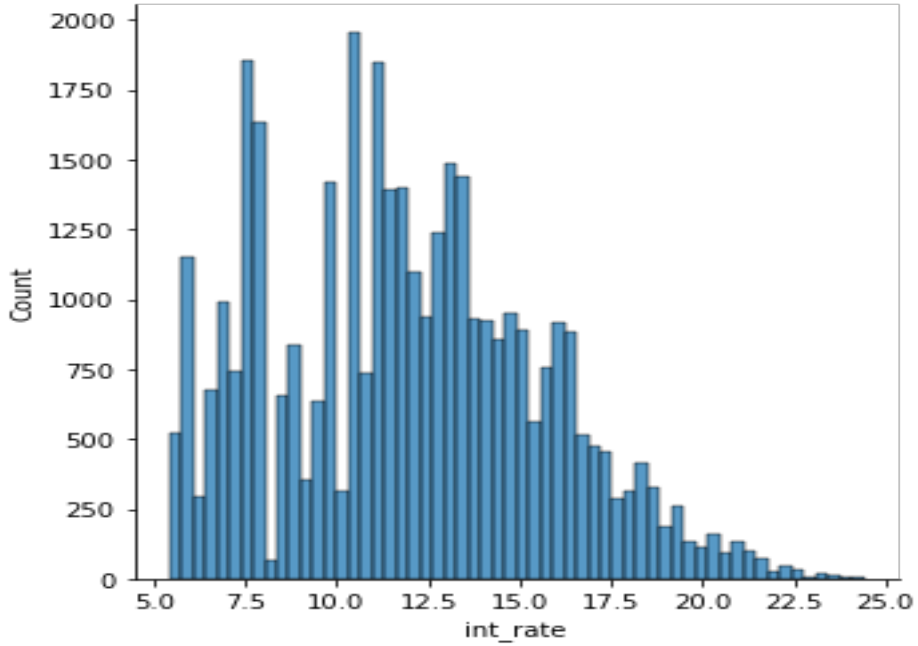
1. Most of the loan applications were for debt consolidation.

2. Most of the applications were approved for int rate of 7.5 or 10%

3. Most of the applications were from CA state.

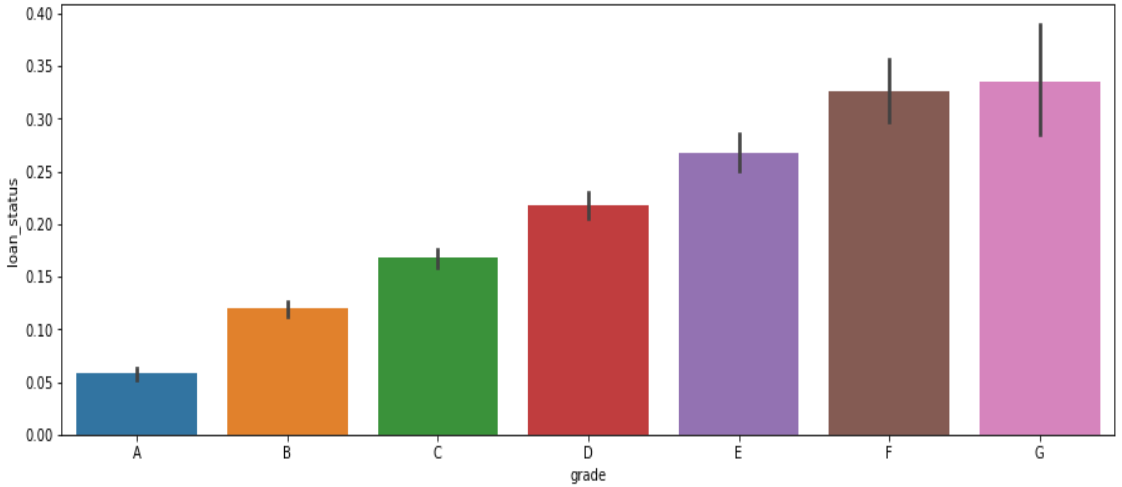
4. No of applications were more in the later part of the year



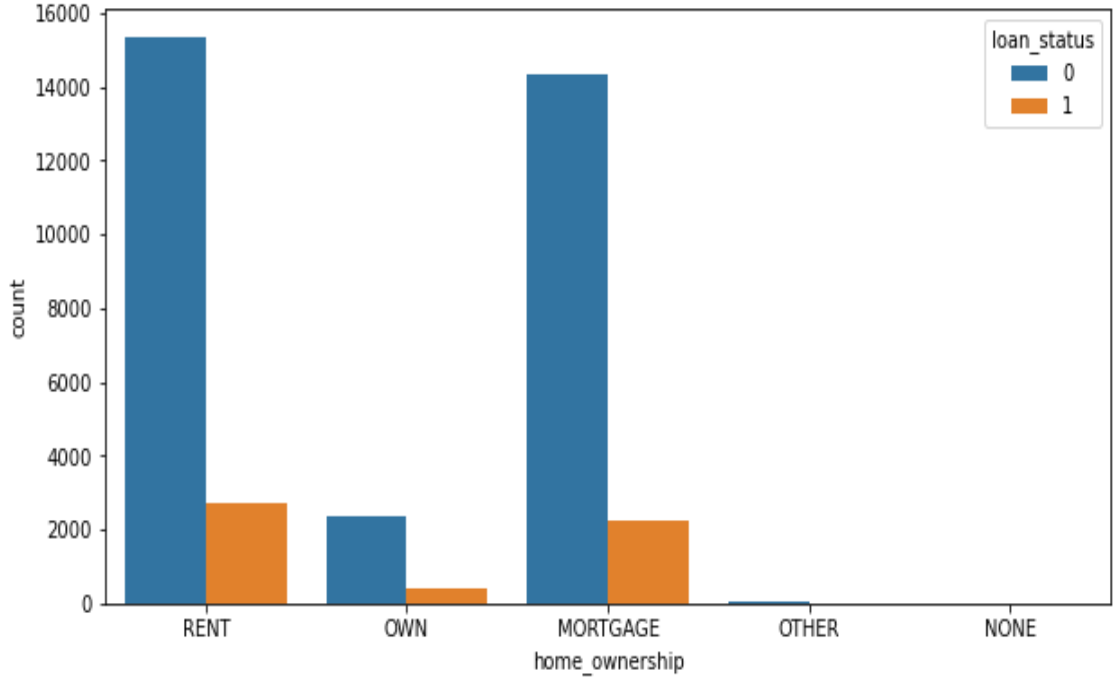


• Bivariate/Segmented Analysis

Bivariate and Segmented analysis was performed to see patterns of lean default w.r.t to different factors and below was observed



1. Grades E,F,G had higher default rate. So lending club should be careful while approving loan applications for those grades.
2. People with Home Ownership as Rent or Mortgage are more prone to get defaulted for the loan.
3. Default rate for loans with ‘Term’ as 60 months was higher than 36 months.
4. From the analysis it was found that percentage of defaulters is higher for purposes as “Small Business”, “Renewable Energy” & “Other”.
5. The state of NE, NV, SD has highest defaulters. We should be more cautious while approving loans from these state.



Conclusions

Based on the analysis conducted across various variables and factors, below are found to be impacting factors for loan default:

- Address State
- Home ownership
- Grades/Sub Grades
- Purpose of loan and Loan amount

Thank You