Lending Club Case Study

…

Mallikarjunappa S Sangannavar

Sriramsubramaniyan Nadarajan

Problem statement and analysis approach

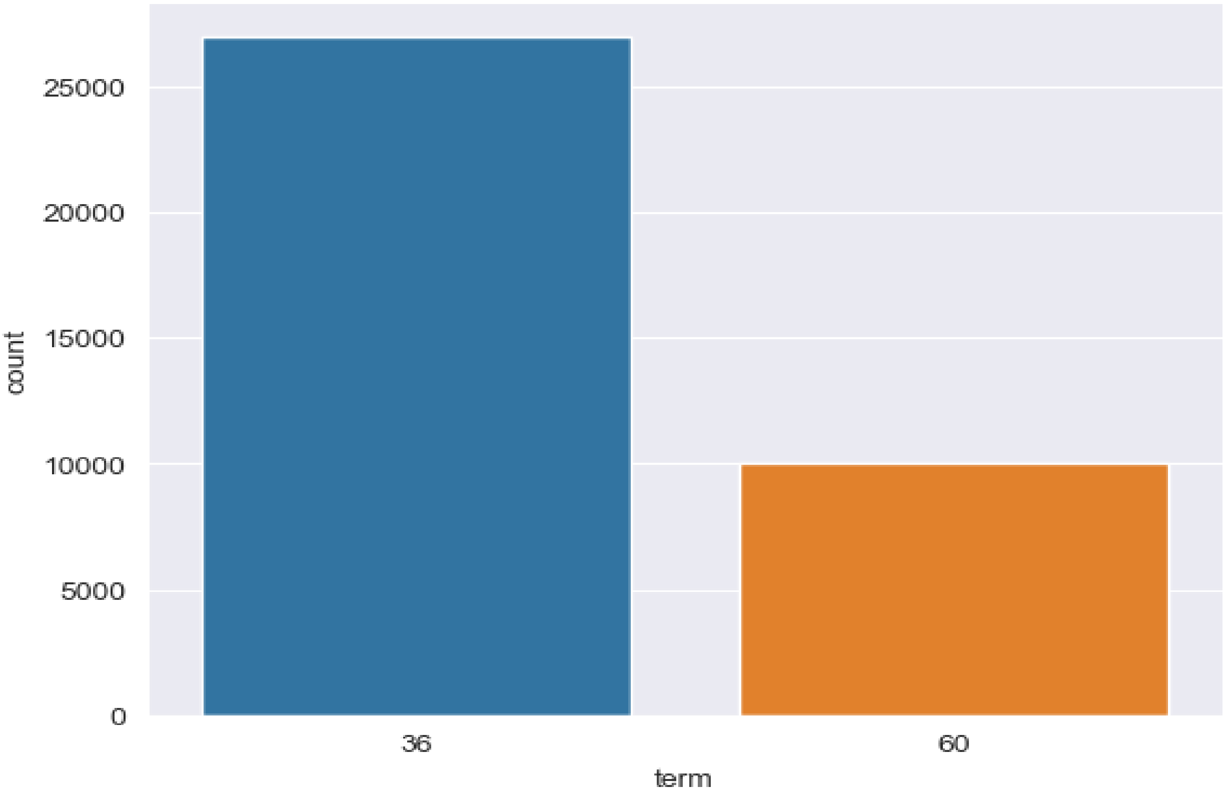
- Problem statement : There is a loan data set from Consumer Finance Company. Need to analyze the customers who defaulted the loan. Come up with consumer and loan attributes influencing the loan defaults. These can be used in deciding the loan for news customer based on these recommendations.
- Two types of risks associated with bank’s decision
  - If loan is rejected for applicant who is likely to repay the loan, results in business loss to the financing compuny
  - If loan is accepted for an applicant who is not likely repay the loan, may result in loan default hence financial loss to the company.
- Analysis Approach :
  - Understand the data present in data set
  - Check quick view of missing values and decide if any columns or rows can be dropped
  - Analyse missing values and impute missing values with appropriate values
  - Analyse if any constant values in columns or rows. Drop the columns with same values
  - Check outlier in the data and address the outlier
  - Start univariate, bivariate variable analysis to come with factors influencing the loan default.
- Based on analysis output, two types of decision are possible
  - Loan Accepted : In applicant meets all requirements to provide the loan, then loan application will be accepted
  - Loan Rejected : Based on factors influencing default, finance company can reject the loan application

Understand the data and Analysing the missing values

- Original data set size is : Shape of dataframe : (39717, 111)
- Print the number of missing values per column. Based on the output saw many columns having all values missing. Hence decided to drop columns with more than 90% missing values
- Based on further analysis saw few columns with all zeros and same values like ‘f’ and ‘n’. Drop these columns
- Now again print missing values per columns. There are some columns with very few missing values, hence decided to drop rows where columns have few missing values

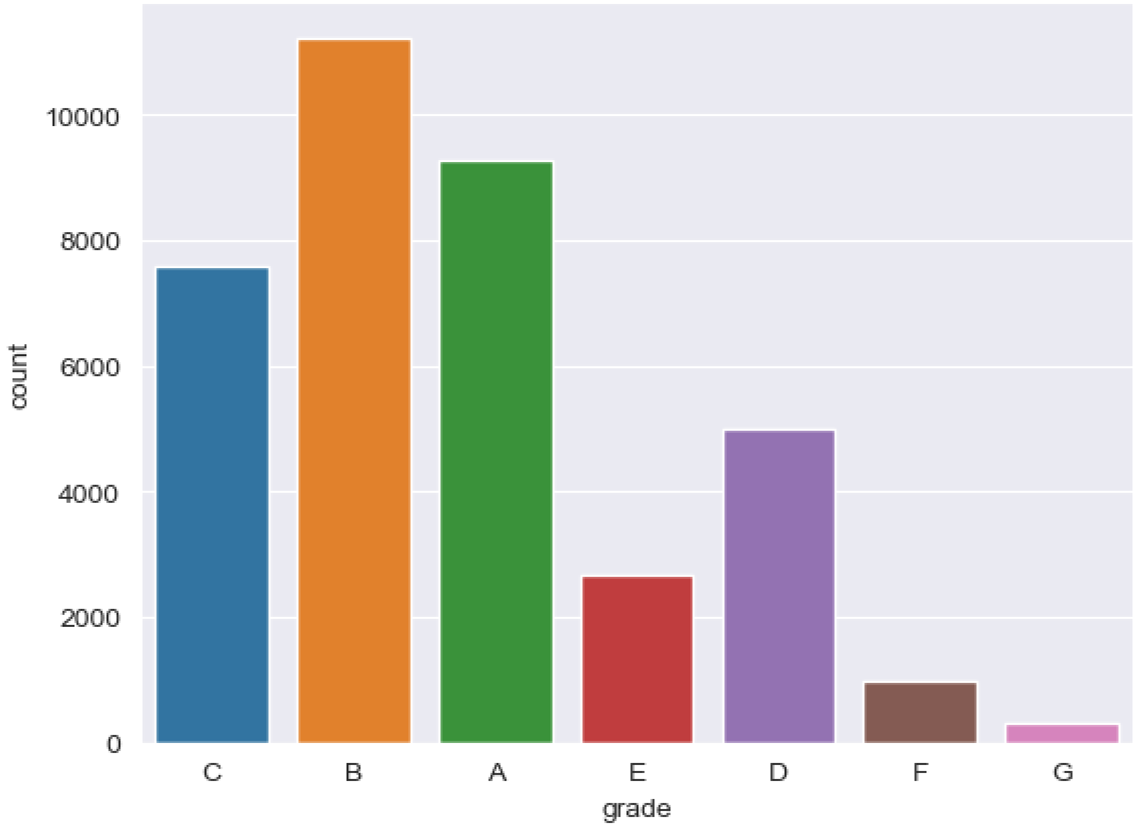
Univariate variable analysis -- term

- Out of close 39000 accepted loans
- Close 28000 has 36 months as loan term which is 3 years
- Close to 10000 has 60 months as loan term, which is 5 years



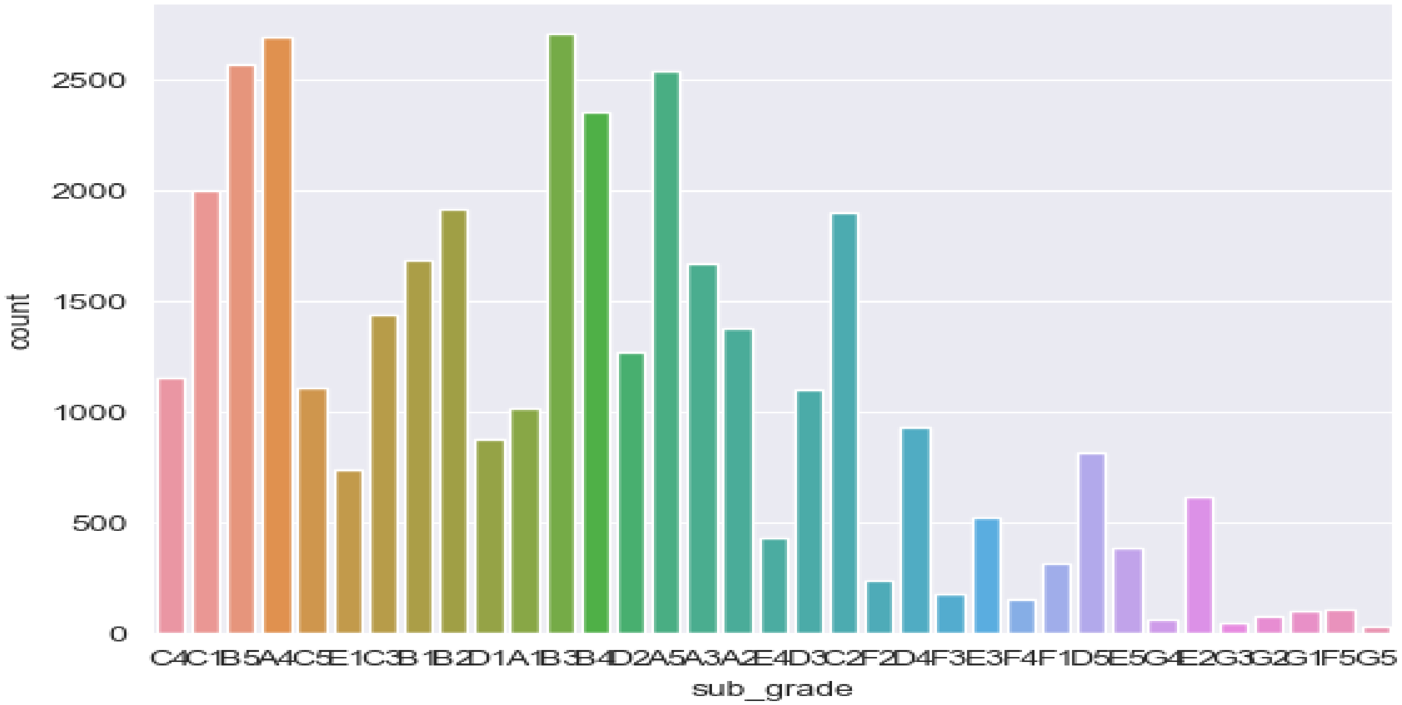
Univariate variable analysis -- Grade

- Highest loan acceptance is from Grade - B
- Second acceptance is from Grade – A followed by Grade C and D
- Lowest acceptance is from Grade G



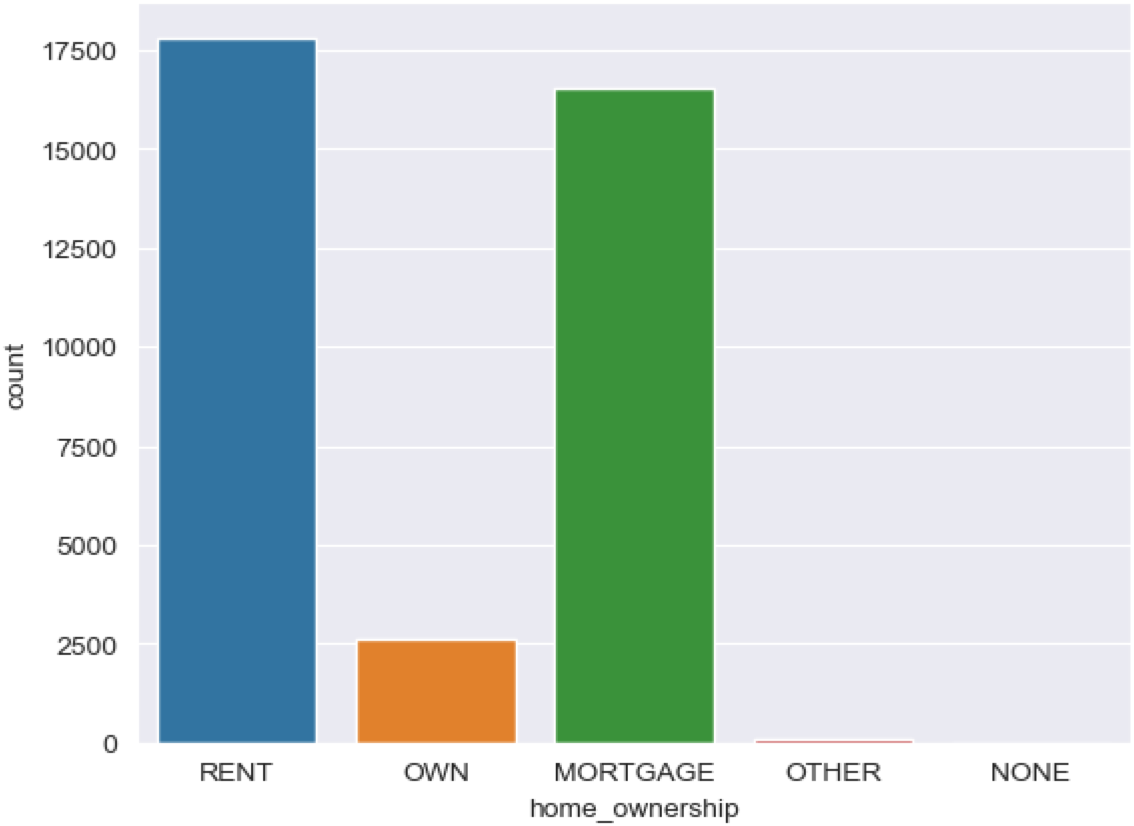
Univariate variable analysis – Sub-Grade

- From Sub-Grade analysis also, highest loan acceptance is from sub-grades of Grade – B and A
- Lowest acceptance is from sub-grades of Grade G



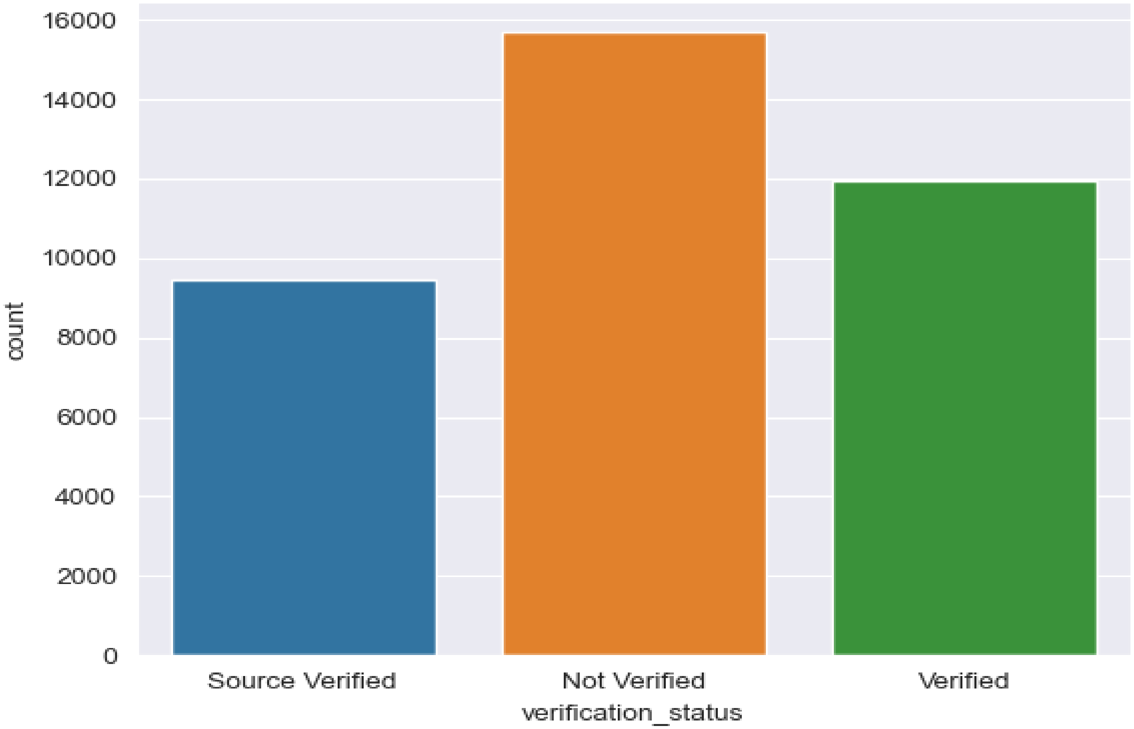
Univariate variable analysis – Home ownership

- Maximum loan acceptance is for people staying in rented house
- Second Maximum loan acceptance is for people staying in mortgage house
- Lowest acceptance is for people staying in other category house
- So more importancce should be given for loan defaults from rented house and mortgage house



Univariate variable analysis – Verification status

- Maximum loans are given without verification
- Hence need to check if more loan defaults from not verified



Bivariate variable analysis

Recommendation from analysis