Lending club case study

09-Feb-2020

-Varsha R, Shreyas P

Problem statement:

To understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default which would help in better management of company portfolio and risk assessment

Analysis approach:

We carried out our analysis using the following approach:

- Initial data understanding: Go through the data and
  - get an understanding of the variables
  - level of data
  - identifying categorical and numerical values
  - distinct values present in the categorical columns and the range of values in the numerical variables

- Data cleaning
  - Removing columns with >50% of null values
  - Removing columns which have information about customer behavior and which are irrelevant
  - Removing highly correlated columns
  - Standardizing the values of columns
  - Creating box plots to remove the outliers in numerical columns

- Univariate analysis for unordered categorical variables
  - Creating log frequency rank plots

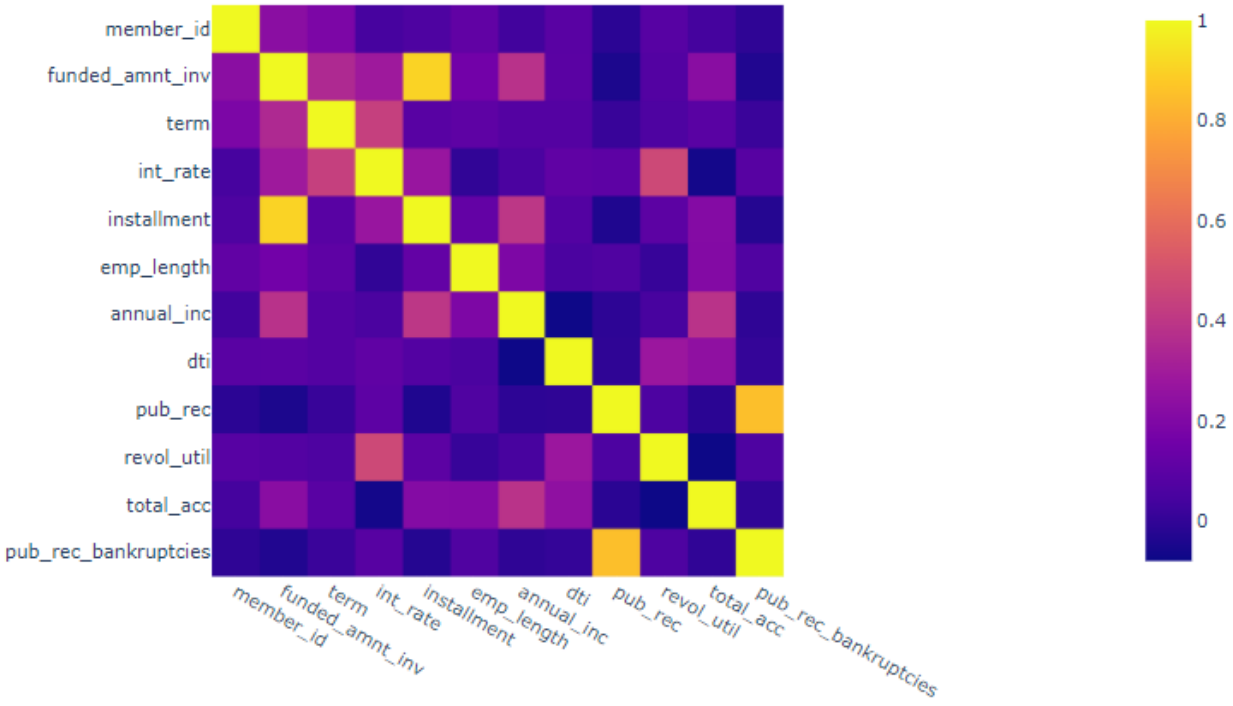
- Univariate analysis for ordered categorical variables
  - Creating bar plots across variable – one bar plot for charged off members and the other one for fully paid members
  - Creating histograms by considering bins for numerical variables

- Bivariate analysis
  - Creating bar plots by plotting 2 variables against each other and also creating labels for charge offs and fully paid for comparison

Results

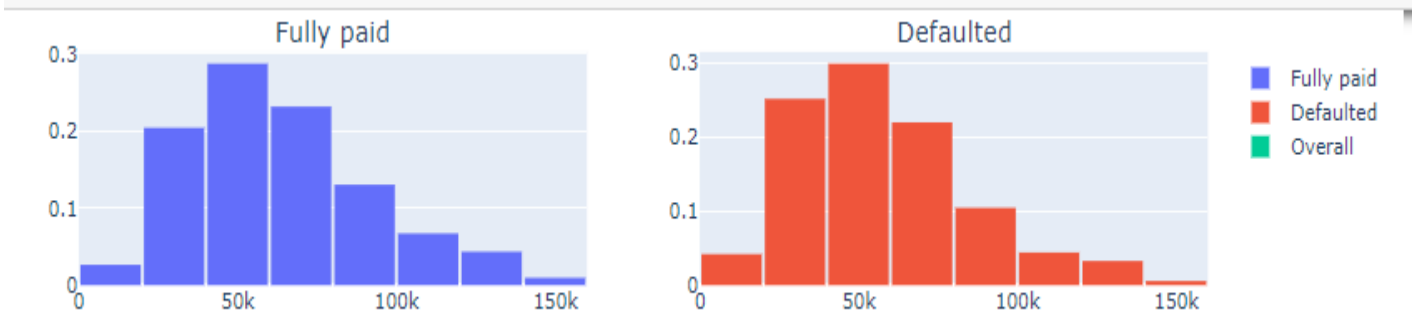
- **Univariate analysis on unordered categorical variables**
  - Based on the log frequency rank plot for loan purpose and state, it doesn't give us clear insights on whether charge offs would happen or not

- **Univariate analysis on ordered categorical variables**
  - Except for correlation between funded amount inv and installment, pub_rec and pub_rec_bankrutcies, no significant positive or negative correlation found between other variables

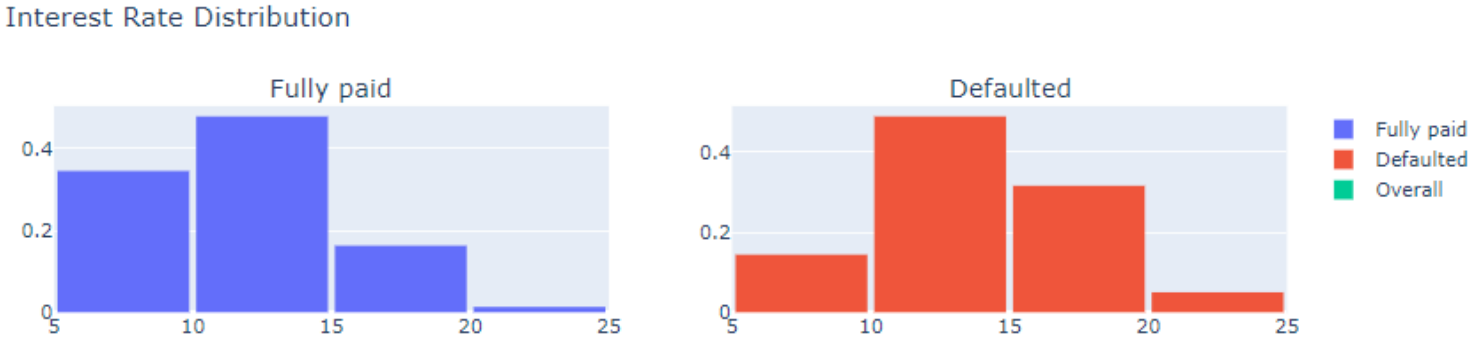


Univariate analysis on ordered categorical variables

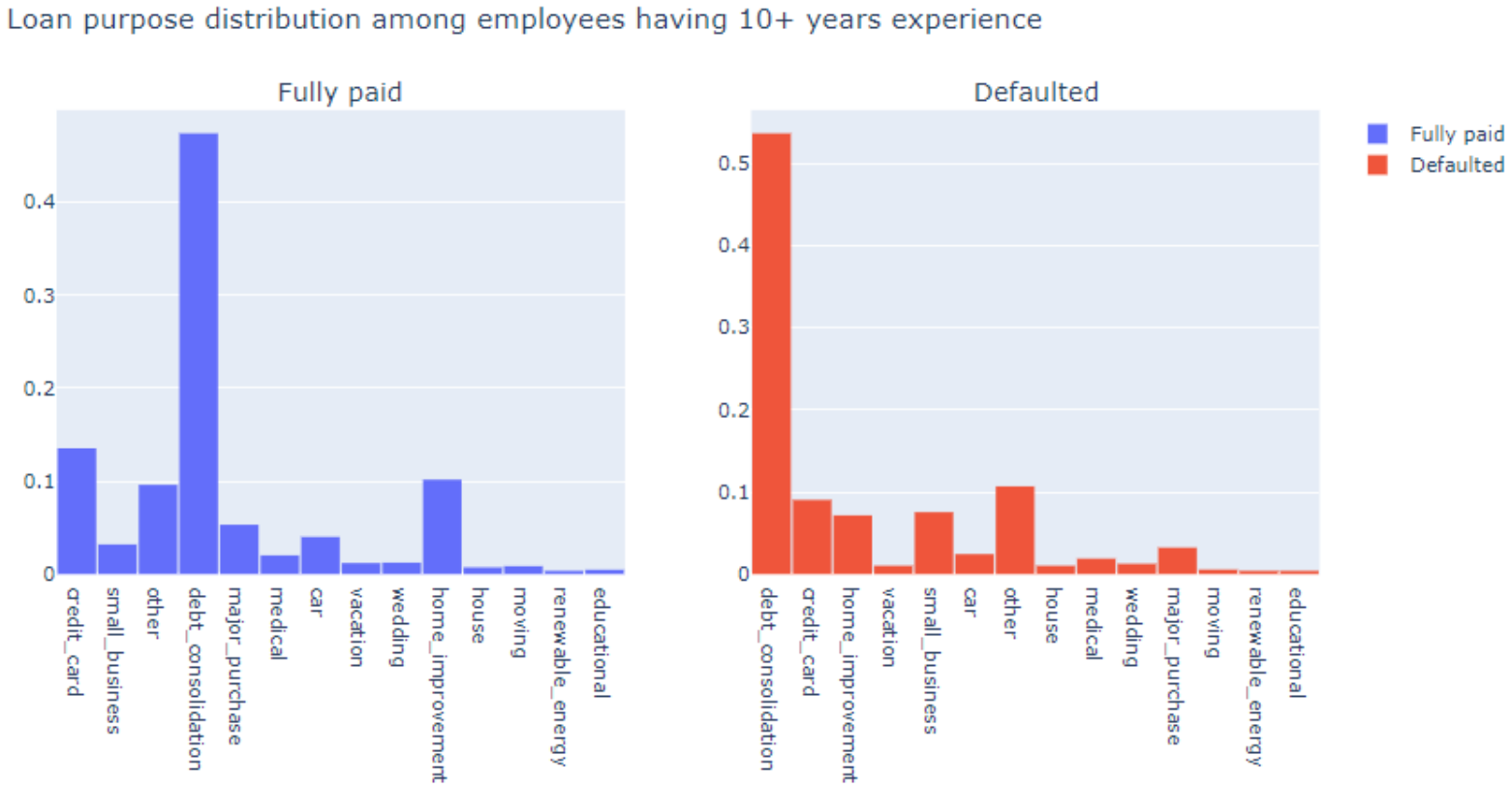
- Income distribution: T*he probability of defaulters in lower income range (<50K) is higher*



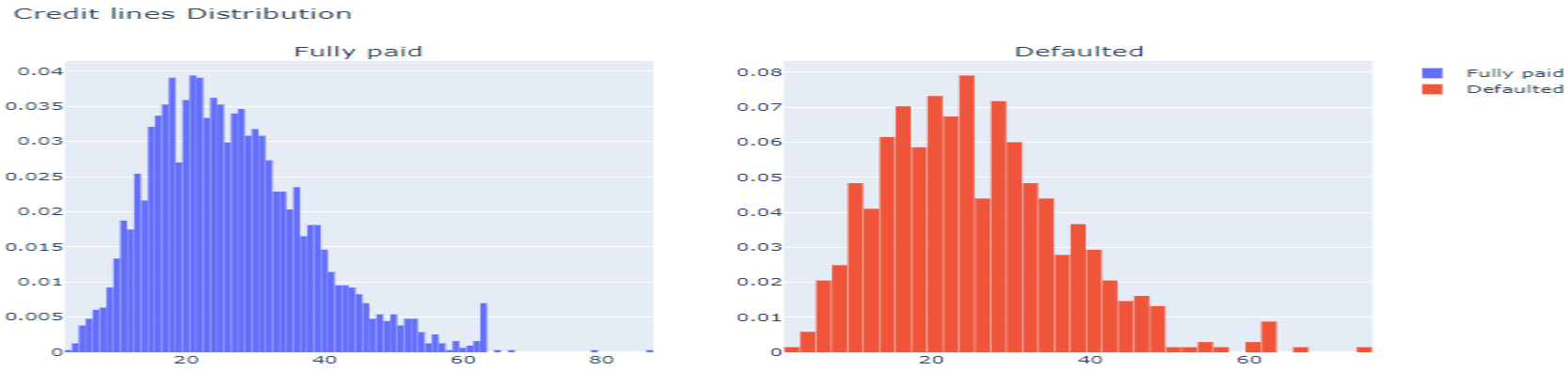
- Interest rate*: For lower interest rates, loans are more likely to be fully paid whereas higher interest rates (>15%) saw higher number of defaulters*



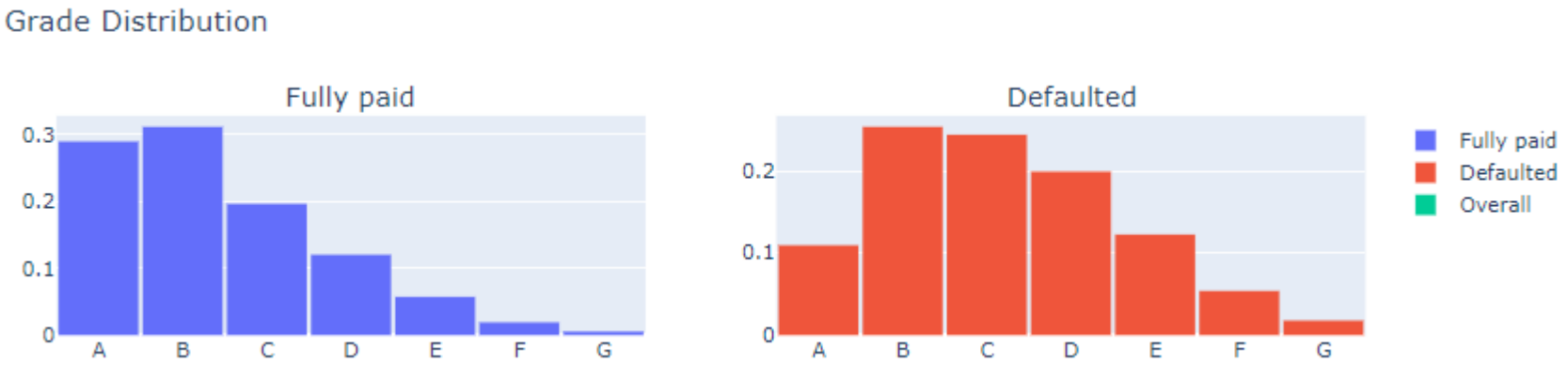
- Employment length distribution: *The number of employment years do not differ much for fully paid and defaulters. Max probability of defaulters are seen in 10+ years of experience and deep diving into loan purpose among these members, it was observed that small business have a slightly higher probability of defaulting*



- Credit line distribution among members having 10+ years of employment and have debt consolidation as loan purpose: *Members having 10+ years of emp length who have taken the loan for debt consolidation and having higher credit lines are more likely to get a charge off*



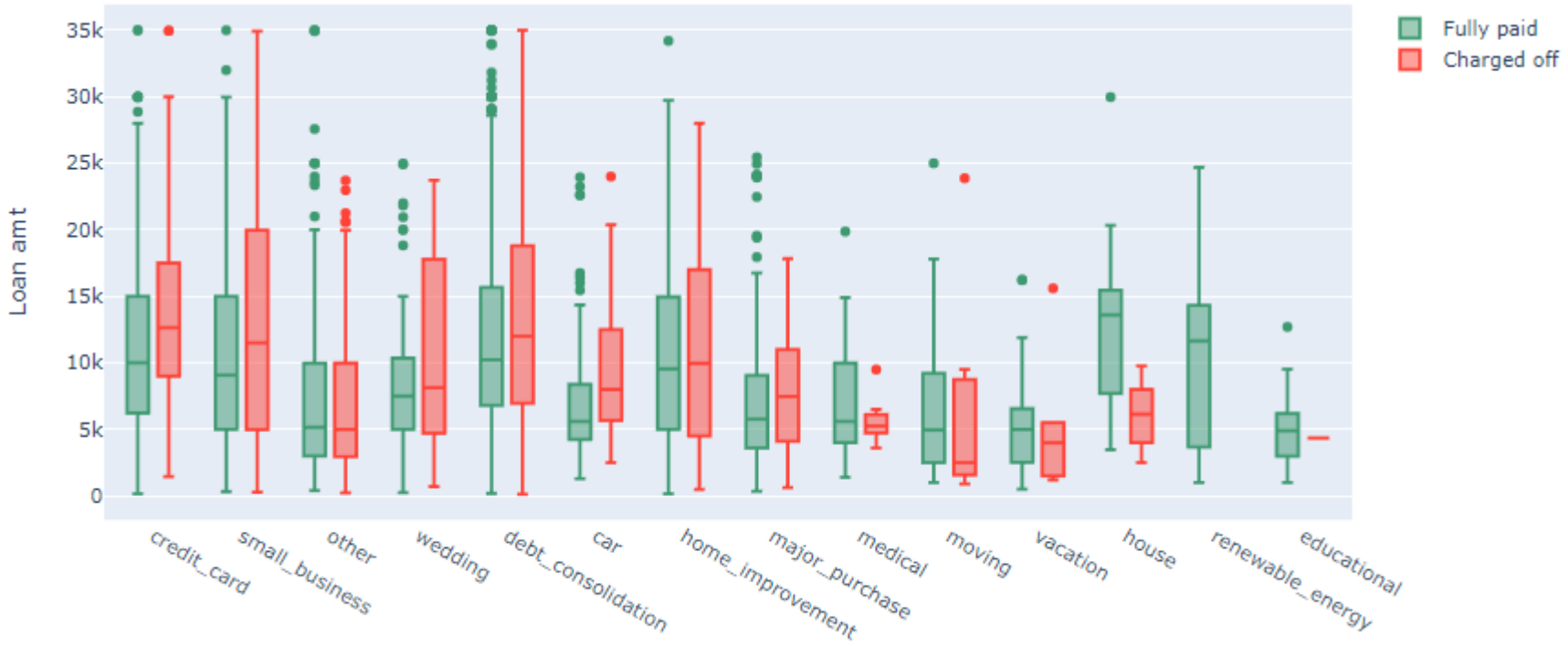
- **Grade distribution:** Higher grades (>C) are likely to have more number of charge offs¶



Bivariate analysis

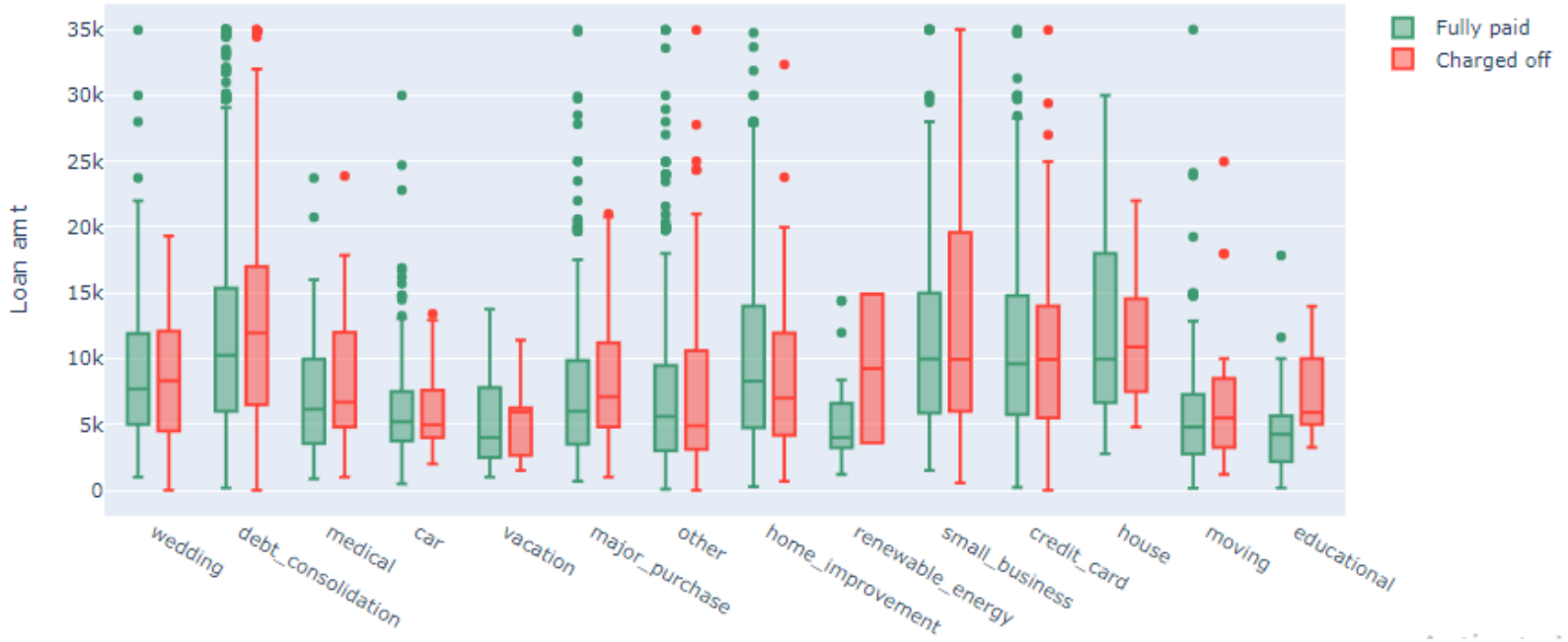
- Funded amount across months – *box plot comparison between charged off and fully paid members*

- Funded amount across across loan purpose for the month of december: *box plot comparison between charged off and fully paid members*



- *December being the holiday period might see an increase in marriage loans which might get charged off*
- *End of year might observe higher defaulters in small business*

- Funded amount across loan purpose for the month of august-september: *box plot comparison between charged off and fully paid members*



- *Schools and colleges usually start in the month of AUG / SEPT, so educational loans taken are higher and so is the defaulters*
- *Loans taken on renewable energy is also higher in aug-sep*

- Funded amount across grade– *box plot comparison between charged off and fully paid members*