Lending Club Case Study

Exploratory Data Analysis

Palak Sharma

- Problem Statement
- Data Summary
- Data Cleaning
- Data conversions vs Derived Columns
- Dropping/Imputing the Rows
- Outliers
- Univariate Analysis
- Bivariate Analysis
- Correlations
- Conclusions

Problem Statement

Problem:

- You work for a consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
    - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
    - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Objective:

- Use EDA to understand how consumer attributes and loan attributes influence the tendency of default

Constraints:

- When a person applies for a loan, there are two types of decisions that could be taken by the company:
    - **Loan accepted**: If the company approves the loan, there are 3 possible scenarios described below:
        - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
        - **Current:** Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
        - **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan
    - **Loan rejected:** The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the
        - loan was rejected, there is no transactional history of those applicants with the company and so this data is notavailable with the company (and thus in this dataset)

Data Summary

- Loan.csv file contains 39717 rows and 111 columns.
- There are two types of attributes Loan Attribute and Customer attributes.

Data Cleaning

- There were no header, footers, summary or Total rows found.
- There were no duplicates rows found.
- There were 1140 rows present of loan_status=‘current’ which has been deleted as loan_status =‘current’ does n’t participate in analysis.
- There were 55 columns which is having all the rows values as null/blank and doesn’t participate in analyse has been removed.
- ‘url’ and ‘member_id’ is unique in nature and has been deleted.Have kept ‘id’ for future purpose analyse.
- ‘desc’ and ‘title’ text/description values and doesn’t participate has been dropped from analysis.
- Limiting our analysis till ‘Group’ level only hence sub group has been dropped.
- Using domain knowledge, behavioural data is captured and hence will not available during the loan approval and doesn’t participate in analysis. 21 behavioural data columns has deleted.
- 8 columns whose values were 1, and is uniqueness in nature has been dropped from analysis.
- There were two columns which is having more that 50% of data as na has been removed.
- After all the Data cleaning process we are left with 38577 rows and 20 columns.

Data Conversions vs Derived Columns

- Additional string value has been trimmed from ‘term’ column and has been converted to int data types.
- ‘int_rate’ has been converted from string to int. Additional ‘%’ has been trimmed.
- Column ‘loan_funded_amnt’ and ‘funded_amnt’ converted to float.
- loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', ‘dti' columns valued rounded off to two decimal points.
- issue_d has been converted to datatype.
- Creating a derived columns for ‘issue_year’ and ‘issue_month ’ from ‘issue_d’ which will be using for further analysis.
- ‘loan_amnt_b’, ‘annual_inc_b’, ‘int_rate_b, and ‘dti_b’ derived columns(multiple bucket kind of data from continuous data) has been created for better analysis.

Dropping/Inputing the rows

- ‘emp_lenght’ and pub_rec_bankruptcies contains 2.67% and 1.80% of rows as null, which is very small percetnage of data which we can drop it.
- Total % of rows deleted: 4.48%,
- Outliers exits for numeric data 'loan_amnt', 'funded_amnt', ‘funded_amnt_inv','int_rate', 'installment', ‘annual_inc'.
- Outliers treatment has been done for above fields using quantile mechanism.

Conclusions

- Income range between 0-20000 has high chances of charged off.
- Interest rate more than 16% has good chances of charged offas compared to other category interest rates.
- Those who are not owning the home is having high chances of loan defaulter.
- Those applicants having loan for small business is having high chances for loan defaults.
- High DTI value having high risk of defaults.
- Higher the Bankruptcies record higher the chance of loan defaults.
- DE States is holding highest number of loan defaults.
- The Loan applicants with loan Grade G is having highest Loan Defaults.