Lending Club Case Study

Krishna Moorthy Rohit Gopal

Problem Statement

- In a typical financial based company, there will be various types of loans given to customers. However, when the company receives a loan application, the company has to decide if loan can be given based on applicant's profile. If we reject the loan and if the applicant is likely to repay the loan, then there will be loss of business to the company. If the loan is approved and in case the loan is not repaid, then it will be a loss for the company.
- We need to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default which will help the company use this and assess the risk

Objective

- We need to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default which will help the company use this and assess the risk

Approach Taken

Data Cleaning

First, we will need to clean the data before we proceed with further analysis. Below are the steps taken for the cleanup:

- Duplicate Rows:
  - There are no duplicate rows based on id hence nothing is removed
- Missing Values(null/NA) on Columns:
  - Dropped all columns which has all missing values. Ex: dti_joint, verification_status_joint and few other columns
- Missing Values(null/NA) on Rows:
  - There are no rows will all missing values
- Remove unnecessary columns:
  - Removed columns like pymnt_plan, initial_list_status, collections_12_mths_ex_med, policy_code and few other columns

Data Enrichment

- Derived Metrics:
  - Split issue_d column to month and year
  - Remove the % in emp_length, int_rate, revol_util which can be used for plotting graphs
- Imputation of data for some columns based on % of missing values:
  - Replace the missing values of emp_length and filled with 0 as it is a small number and wont impact much. This is with assumption that data is not updated for 0 emp_length
- Converting columns to numeric data:
  - Convert to numeric data type for loan_amnt, funded_amnt, int_rate, funded_amnt_inv, annual_inc, emp_length etc.

Data Analysis – Univariate / Bivariate

Plotted a graph on the following Ordered Categorical Variables:

1. emp_length - This is to analyze on the employees length
2. int_rate - Is there an higher interest rate hence the defaulting of loan
3. revol_util - Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit
4. total_pymnt - Total Amount which is repaid as part of the loan
5. loan_amnt - Total Amount of the loan taken
6. term - Term of the loan
7. open_account - Number of accounts open
8. annual_inc - Annual Income of the applicant
9. enq_last_6mnths - Enquiries made during last 6 months
10. pub_rec - Public Record available
11. installment - Installment Amount
12. dti
13. loan_funded_inv - Based on investor funding

Data Analysis – Univariate / Bivariate

Plotted graph on following Unordered Categorical Variables:

1. loan_status - Status of the loan
2. purpose - What purpose was the loan taken
3. grade - Loan Grade
4. sub_grade - Loan Subgrade
5. Home Ownership - Ownership of the home
6. verification_status - If information about applicant is being verified or not

Probability of Defaulting Loan using Charged Off Status

Based on the graph analysis related to the charged off loans, there is more probability of defaulting on the below scenarios:

1. Loan applicants who has house ownership as RENT

2. Loan taken for Debt Consolidation. Probably to clear other debts

3. Grade of B

4. Sub Grade of B5

5. Those who are paying interest rate of 13% to 17%

6. Employment length of 10+ years

7. Who have open account of 2 - 10

8. Those who have annual income of 31k to 58k

9. Whenever the loan status is not verified

10. Loan term of 36 months

11. If there are 0 enquiries in the last 6 months

12. When public record is 0

13. When monthly installment is between 145-274

14. When DTI is between 12-18

15. Whoever has taken loan amount of 5k to 10k

16. When funded amount by investor is between 5k to 10k

Most importantly, loans taken in 2011 year and during December have more defaulting of loan. This could be because of some crisis or loss of jobs during that period which could have triggered

Probability of Defaulting Loan using Charged Off Status / Fully Paid / Current

The above analysis with respect to the entire loan data which includes of Fully Paid, Charged Off and Current. There is a more probability of defaulting when:

1. Loan applicants who has taken loan for home improvement and have income of 60k to 80k
2. Loan taken as MORTGAGE and have income of 60k to 80k
3. Applicants who have taken loan amount of 15k to 25k and annual income of 112k to 140k
4. Applicants who have taken loan amount of 28k+ and charged interest of 15% to 17.5%
5. Loan amount of 12500 to 17500 and taken for MORTGAGE
6. Applicants who took loan for small business and the loan amount of 15k+
7. Loan Amount of 18K and loan taken in the month of December
8. Loan Amount of 12K and loan taken in the year of 2011
9. Loan Amount of 18K and grade of G
10. Loan Amount of 25K and grade of G1
11. When employment length is of 10 years and loan amount is 12k-15k
12. When the loan is verified and loan amount is above 16k+