Lending Club Case Study

Team : Gnanaprakash S and Vikas Sharma

Date : 06-09-2023

upGrad

ASSIGNMENT

Name: Lending Club Case Study

Problem Statement

Consumer finance company 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

Assignment Objective

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

PART I : Data Cleaning

Analyzing Defaulted Applicants

There are 39717 rows and 111 columns present in the dataset.

Fix rows and columns:

- Dropped the columns which are having more than 80% of invalid data.
- Dropped the rows which are having more than 80% of invalid data.
- Dropped the columns when it has less than 2 unique values.

Fix missing values:

- Filling employee title with 'NaN' as 'Unknown'
- Filling Homeownership with 'NONE' as 'OTHER'.

Results

- Post cleaning there are 39717 rows and 44 columns are present

PART II: Data Analysis

Analyzing Defaulted Applicants

Variable under consideration:

The 'Charged off' value in Column 'loan_status' refers to defaulted applicants.

Analyzing the driving factor for defaulting the loan applicants:

- Based on analyzing the Meta data, removing the columns which are considered to be non-driving factor for defaulting the loan applicants

- Out of **44 columns**, these columns are considered to be non-driving factor [id", "member_id", "url", "title", "zip_code", "addr_state", "emp_title", "delinq_2yrs", "revol_bal", "out_prncp", "out_prncp_inv", "total_pymnt", "total_pymnt_inv", "total_rec_prncp", "total_rec_int", "last_pymnt_d", "last_pymnt_amnt", "last_credit_pull_d"]

- Post droping out, we are having **26 columns** [loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, installment, grade, sub_grade, emp_length, home_ownership, annual_inc, verification_status, issue_d, loan_status, purpose, dti, earliest_cr_line, inq_last_6mths, open_acc, pub_rec, revol_util, total_acc, total_rec_late_fee, recoveries, collection_recovery_fee, pub_rec_bankruptcies

PART III: Univariate Analysis

Analyzing Defaulted Applicants

Variable under consideration:

- Dropping the columns which do not affect the analysis - "id", "member_id", "url", "title", "zip_code", "addr_state", "delinq_2yrs", "emp_title", "revol_bal", "out_prncp", "out_prncp_inv", "total_pymnt", "total_pymnt_inv", "total_rec_prncp", "total_rec_int", "last_pymnt_d", "last_pymnt_amnt", "last_credit_pull_d“.
- Convert all percentages to float type.
- Filling the ‘revol_util’ column NaN values with Median.
- Convert all time-series columns to Date-Time format.
- ‘sub_grade’ refers to Grade + Sub division value from grade, Updating the value with Sub division of it.

Categorizing the column for the Univariant Analyzing:

- **categorical_columns** = ['term', 'emp_length', 'home_ownership', 'verification_status', 'purpose', 'inq_last_6mths', 'pub_rec', 'pub_rec_bankruptcies', 'grade']
- **numerical_columns** = ['loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'installment', 'annual_inc', 'dti', 'open_acc', 'total_acc', 'total_rec_late_fee’, 'recoveries', 'collection_recovery_fee', 'int_rate_p’, 'revol_util_p']
- **date_time_columns** = ['issue_d' , 'earliest_cr_line']

PART II: Univariate Analysis

Analyzing Defaulted Applicants

Variable under consideration:

- Get the defaulted applicant.
- Creating quartile bins for numerical columns to make them categorical.

PART III: Univariate Analysis (continue..)

Observations on Univariate Analysis:

- Lower the loan_amnt increases the chance of defaulted - Range (500 - 7720.0)
- Lower the funded_amnt increases the chance of defaulted - Range (500 - 7720.0)
- Lower the installment increases the chance of defaulted - Range (15.69 - 279.27)
- Lower the annual_inc increases the chance of defaulted - Range (4000.0 - 253264.0)
- Between 40%-60% of the dti increases the chance of defaulted - Range (11.94 - 17.91)
- Lower the open_acc increases the chance of defaulted - Range (2 - 9.2)
- Between 20%-40% of the total_acc increases the chance of defaulted - Range (16.4 - 30.8)
- Lower the total_rec_late_fee increases the chance of defaulted - Range (0.0 - 36.04)
- Lower the recoveries increases the chance of defaulted - Range (0.0 - 5924.67)
- Lower the collection_recovery_fee increases the chance of defaulted - Range (0.0 - 1400.43)
- Between 40%-60% of the int_rate_p increases the chance of defaulted - Range (13.01 - 16.80)
- Between 60%-80% of the revol_util_p increases the chance of defaulted - Range (59.94 - 79.92)
- Higher chance of defaulted for " 36 months" in term column
- Higher chance of defaulted for " 10+ years" in emp_length column
- Higher chance of defaulted for " RENT" in home_ownership column
- Higher chance of defaulted for " Not Verified" in verification_status column
- Higher chance of defaulted for " debt_consolidation" in purpose column
- Higher chance of defaulted for " 0" in inq_last_6mths column
- Higher chance of defaulted for " 0" in pub_rec column
- Higher chance of defaulted for " 0.0" in pub_rec_bankruptcies column
- Higher chance of defaulted for " B" in grade column
- Higher chance of defaulted for B grade of sub_grade " 5" column
- Higher chance of defaulted on month 12 of year 1911 on issue_d column
- Higher chance of defaulted on month 10 of year 1900 on earliest_cr_line column

PART IV : Bivariate Analysis

Analyzing Defaulted Applicants

Variables under consideration:

- categorical_columns = ['term', 'emp_length', 'home_ownership', 'verification_status', 'purpose', 'inq_last_6mths’, 'pub_rec', 'pub_rec_bankruptcies', 'grade']
- numerical_columns = ['loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'installment', 'annual_inc’, 'dti', 'open_acc', 'total_acc', 'total_rec_late_fee', 'recoveries’, 'collection_recovery_fee', 'int_rate_p', 'revol_util_p’].
- Analyse ‘loan_amount’ with Interest rates.
- Analyse ‘loan_amount’ with Home ownership.
- Analyse ‘loan_amount’ with purpose.

PART IV : Bivariate Analysis

Analyzing Defaulted Applicants

Variables under consideration:

- Analyse ‘annual_income’ with loan amount.
- Analyse ‘annual_income’ with home ownership.
- Analyse ‘annual_income’ with employment length.
- Analyse ‘annual_income’ with installment.
- Analyse ‘annual_income’ with grades/sub-grades.

PART IV : Bivariate Analysis

Analyzing Defaulted Applicants

Observations on Univariate Analysis:

Observations on Loan Amount:

- Loan amount increases on Interest rate increses where more chance of Applicant movng under Defaults category.
- High loan amount with applicant in MORTAGE have higher chance of Defaulted.
- Loan amount with installment increases where the chance of defaults are higher.
- Loan amount for 'Small business' purspose have high chance of defaults.

Observations on Loan Amount

- Applicant with Annual income between (1203200.0 - 2402400.0) with higher loan_amnt have chance of defaulted.
- Applicant with Annual income with home_ownership as MORTAGE has higher chance of defaulted.
- Applicant with Annual income and having 10+ years of experience has higher chance of defaulted.
- Applicant with Annual income between (1203200.0 - 2402400.0) with more installment have chance of defaulted.
- Applicant with Annual income between (4080.0 - 1203200.0) with grade B and subgrade of 3-3.5 have chance of defaulted

PART IV : Bivariate Analysis

Analyzing Defaulted Applicants