Lending Club Case Study

Upgrad lending club case study project

Table of Contents

- The Problem
  - Domain, Context and Problem Statement
  - Decision Matrix
- Dataset understanding
- Assumptions
- Challenges deep dive
- Approach
  - Key Columns
  - Excluded Columns
  - Data preprocessing
  - Analysis (Univariate, Bivariate, Multivariate)
- Inferences
- Result
- Recommendations
- Team

The problem

Context

The *lending company* faces largest amount of ***financial loss*** due to ***loan defaults***. The amount lost by the lender when a borrower defaults is called ***credit loss***. A loan is called to be defaulted when borrower does not return the amount fully/partially. These loans are also called to be Charged Off. Borrowers can easily access lower interest rate loans through a fast online interface. Company approves loan based on its current risk assesment of tendency of the loan getting default.

Company wants to reduce the loss due to loan defaults by rejecting loans which are likely to be default. However risk associated with this rejection decision is to lose a loan which otherwise would have been fully paid. Company is looking improve their ability to more accurately understand the lieklyhood of a loan application getting default if approved.

Problem statement

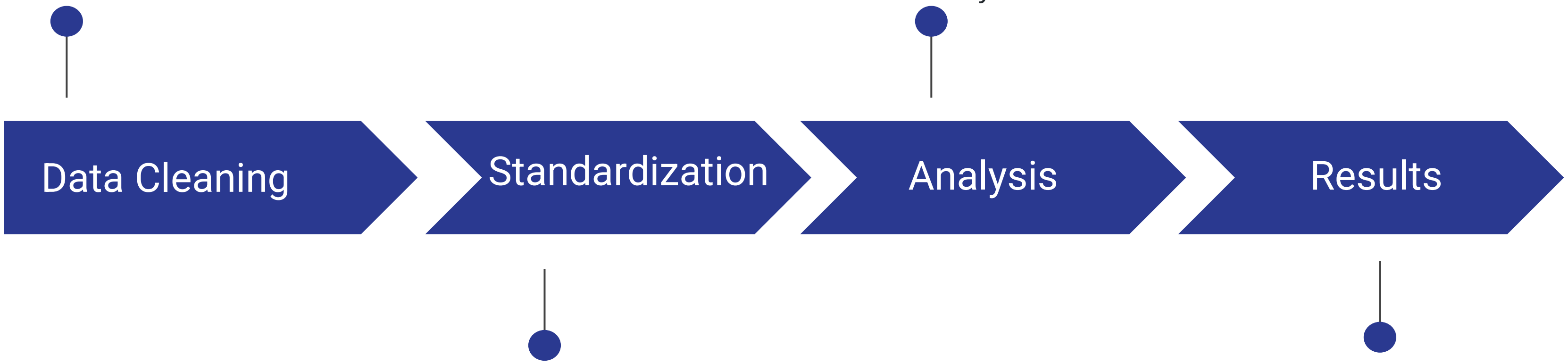
The company has provided few years of data which consists of various attributes of loans which where approved and are etiher fully paid or charged off (defaulted) or is currently running.
We need to understand the influences of loan and consumer atributes towards tendency to default and indentify variables which are strong indicators of default.
The company can utilise this knowledge for its risk assessment to reduce the loss.

Assumptions

- loan_status is “Current” is not useful.
- Binning with uniform intervals
- Monotonous column, ids (such as id, member_id) columns are not useful.
- Free form columns like description , title are not influensive to result.
- Atributes created after loan approval are not considered for analysis

Approach

Stages



- Drop Null Columns
- Drop Current loan_status
- Handle Missing

Data Cleaning

Standardization

Standarized Column values:

- Percentage
- Date
- Parsing Dates where years going back to 1946.

Analysis

Analysis of the dataset

- Univariate
- Segmented Univariate
- Bivariate
- Multivariate analysis

Results

Conclusions,Inferences and Recommendations

Key Column Selection

Customer Demographics

- Annual Income (annual_inc)
- Home Ownership (home_ownership)
- Employment Length (emp_length)
- Debt to Income (dti)
- State (addr_state)

Derived Attributes

- Range bins for selected continuous data types
- Year and Month columns for selected date types.

Loan Attributes

- Loan Amount (loan_amt)
- Grade (grade)
- Term (term)
- Loan Date (issue_date)
- Purpose of Loan (purpose)
- Verification Status (verification_status)
- Interest Rate (int_rate)
- Installment (installment)
- Public Records (public_rec)
- Public Records Bankruptcy (public_rec_bankruptcy)

- Selected above column based on the market research indicating their influences on loan default.

Exclusion Made

Columns

- Columns containing NA values only
- Columns containing single unique value
- Non business columns like ids, title, desc, url

Rows

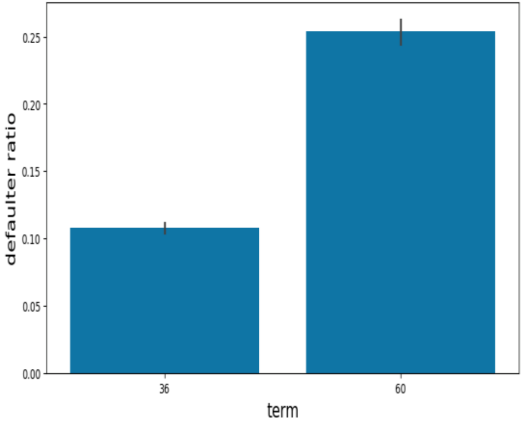
- “Current” loan_status
- Rows for low caridnality dimensions
- Rows corresponding to outliers for various continous attributes.

Selection of Analysis Type

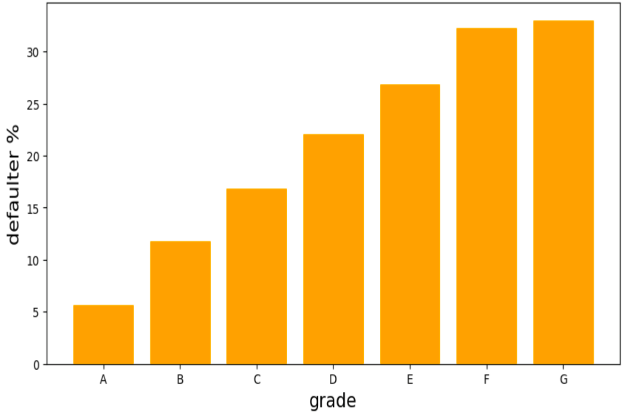
- Frequency bar plot for categorial data
- Summary analysis of continuous data
- Scatter Plot between two continuous data type
- Bar Plot for analysis for a categorial data segmented by loan_status
- Bar Plot for Bivariate Analysis between categorial data
- Box Plot for bivariate or segmented univariate with continuous data
- Heatmap for multivariate analysis

Univariate Analysis

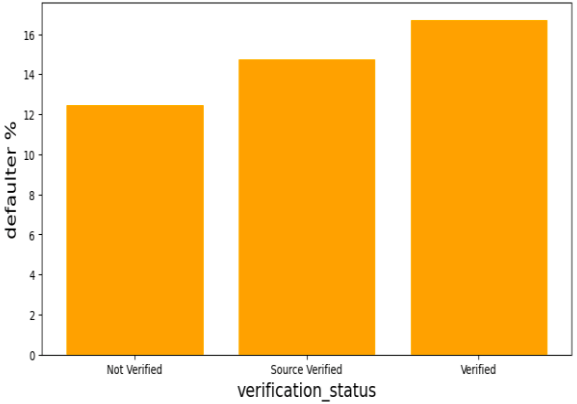
Loan Defaults wrt term - countplot



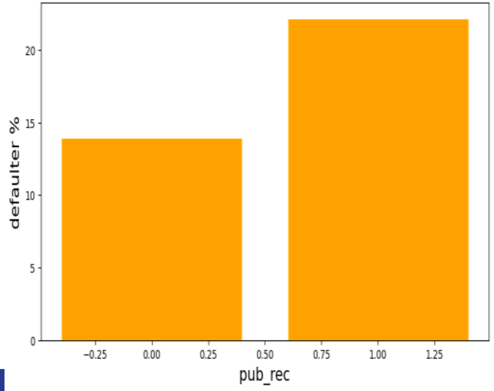
Loan Defaults wrt grade feature - countplot



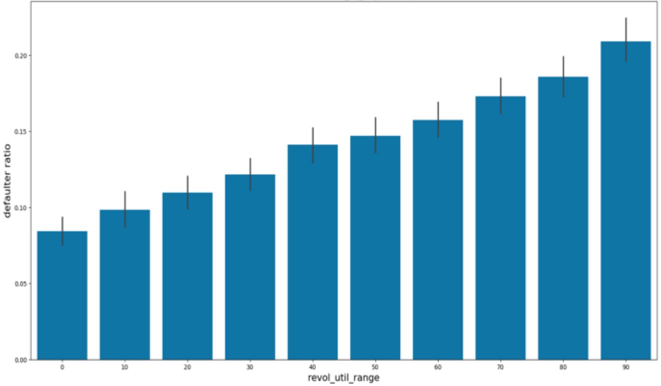
Loan Defaults wrt verification_status feature - countplot



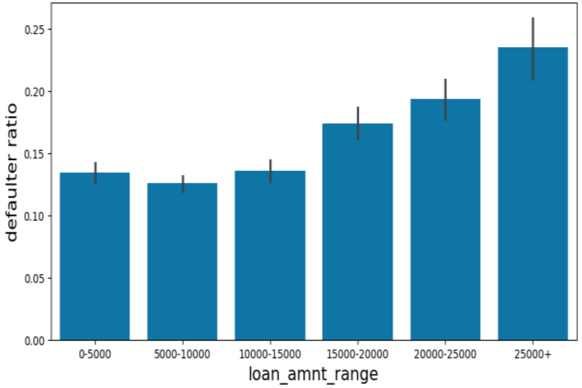
Loan Defaults wrt pub_rec feature - countplot



Loan Defaults wrt revol_util_range - countplot

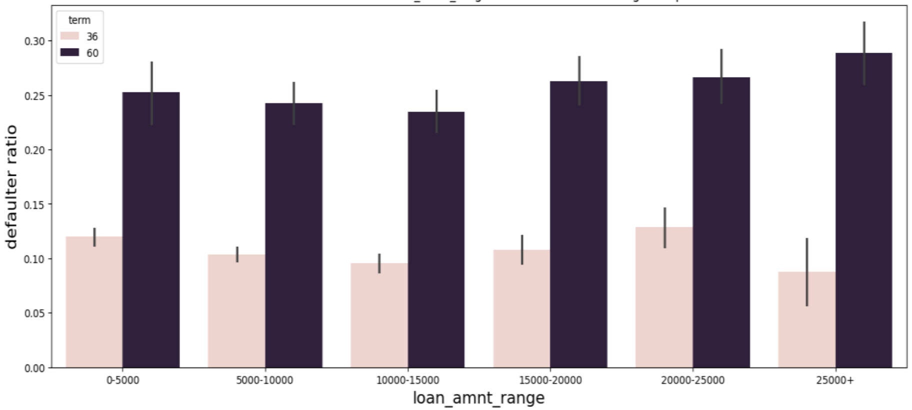


Loan Defaults wrt loan_amnt_range - countplot

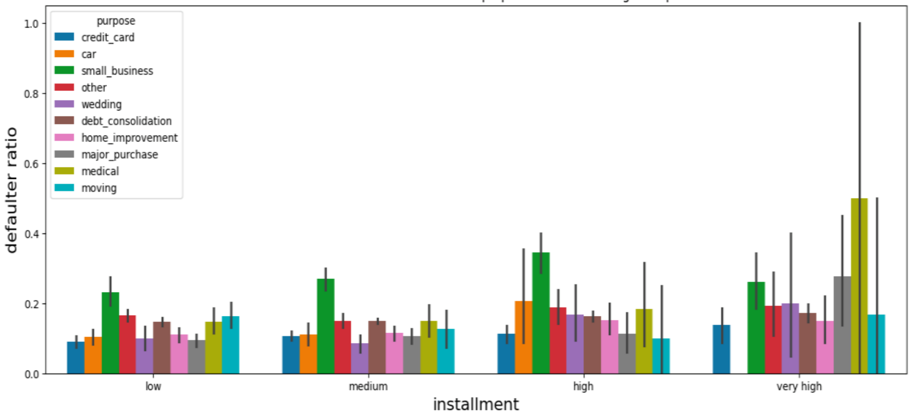


Bivariate Analysis

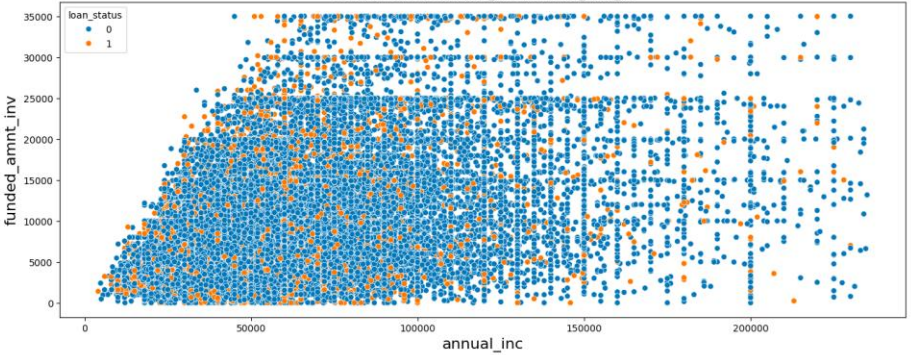
Loan Default ratio wrt loan_amnt_range for term in the data using countplot



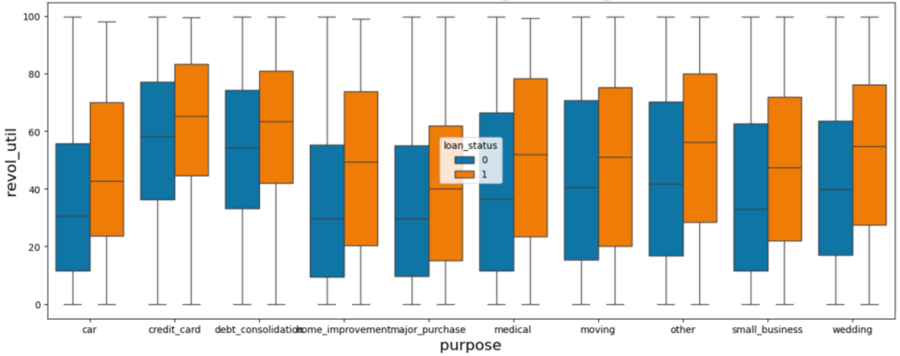
Loan Default ratio wrt installment for purpose in the data using countplot



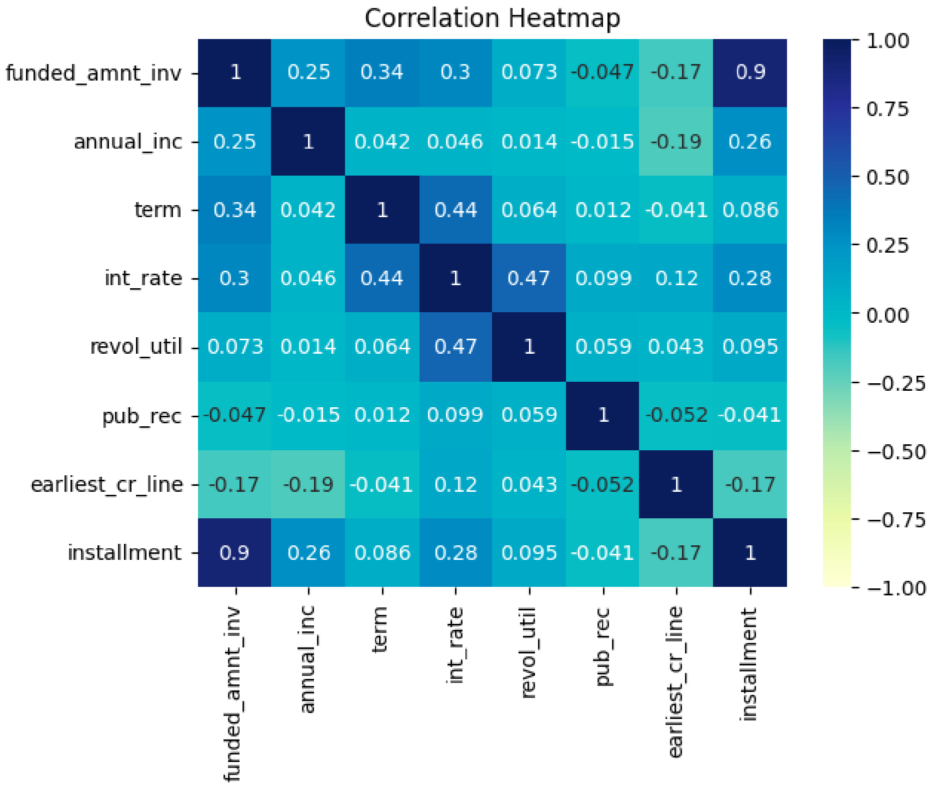
Scatter plot between annual_inc and funded_amnt_inv



Box plot between purpose and revol_util for each loan_status



Multivariate Analysis



Correlation Heatmap

	funded_amnt_inv	annual_inc	term	int_rate	revol_util	pub_rec	earliest_cr_line	installment
funded_amnt_inv	1	0.25	0.34	0.3	0.073	-0.047	-0.17	0.9
annual_inc	0.25	1	0.042	0.046	0.014	-0.015	-0.19	0.26
term	0.34	0.042	1	0.44	0.064	0.012	-0.041	0.086
int_rate	0.3	0.046	0.44	1	0.47	0.099	0.12	0.28
revol_util	0.073	0.014	0.064	0.47	1	0.059	0.043	0.095
pub_rec	-0.047	-0.015	0.012	0.099	0.059	1	-0.052	-0.041
earliest_cr_line	-0.17	-0.19	-0.041	0.12	0.043	-0.052	1	-0.17
installment	0.9	0.26	0.086	0.28	0.095	-0.041	-0.17	1

Inferences

Univariate Analysis

- Defaulter rate is increasing with respect to term, grade, sub_grade, verification_status, pub_rec, pub_rec_bankruptcies, revol_util, loan_amnt, int_rate, installment.
- Defaulter rate is decreasing with annual_inc.
- We have seen that grade D, E, F, G has ~23% total loan however chances of being defaulted is also high here and cover 40% of total defaulters. Hence it makes sense to reject loans for borrowers in these three categories. Hence `grade` can become a deciding variable for loan approval.
- Small_business has higher defaulter rate among other purposes.

Bivariate Analysis

- medical purpose has highest tendency to default with installments > 800.

Multivariate Analysis

- There is no stronger correlation with columns which greatly affects loan_status.

Result

As per dataset analysis, below variables can be driving factors for loan approval process

- Term
- Grade
- Purpose
- Revolving line utilization rate (revol_util)
- Interest rate
- Installment
- Annual income
- Public Record
- Public Record for bankruptcies
- Funded amount by investors

Recommendations

As per dataset analysis, below points might be considered for loan approval process

- tendency to default is very high with borrowers having 2 bankruptcies.
- Implement Stricter Criteria for Grades E, F and G, they increase tendecy to default.
- Evaluate and Limit 60-Month Loans
- Review Verification Process as from dataset it is observed that verified loans are more defaulted which should be otherwise.
- Adjust Interest Rates Based on DTI Ratios
- Consider Annual Income Levels for Affordability

Team

Raman Pandey
Rajeev Ranjan