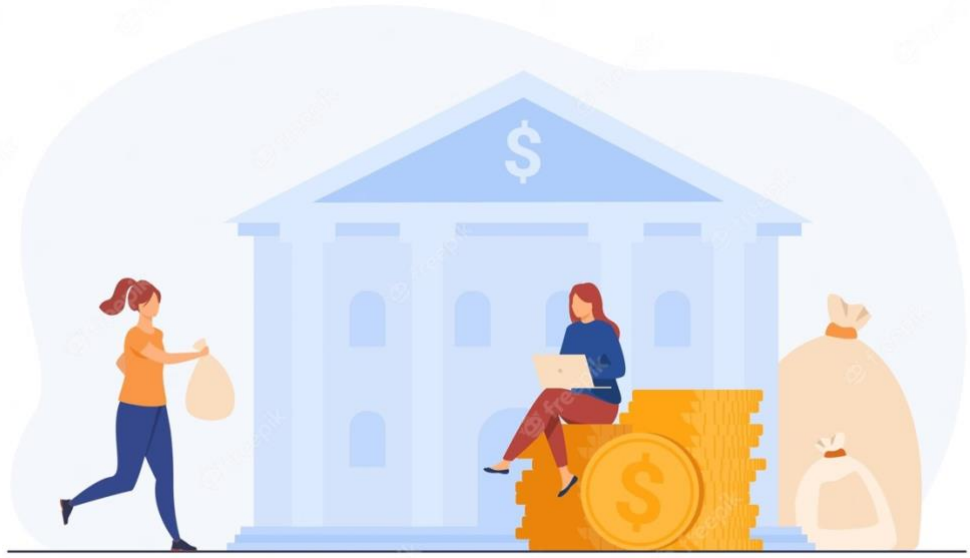
Assignment # 1

Lending Club Case Study

Saad Hashmi & Krishna Kishore



Objective of the Project:

A consumer finance company gives loans to urban consumers. Their main goal is to provide loans to customers who will pay it back and NOT to people who will default. The main objective of this case study is to:

- identify patterns which indicate if a person is likely to default
- What necessary actions to take for those identified patterns

Approach

The following steps were taken for reaching from raw data to the final output:

1. Data Understanding
2. Data Cleaning & Manipulation
3. Data Analysis

4. Final Outcome

1. Data Understanding

In this phase, we understand the data available to us – all the different types of columns, their meaning, potential impact based on business sense.

After that level of understanding, we're able to take a decision on the important data points, the data type, and how to organize it.

We import all python libraries (pandas, seaborn, matplotlib, etc.) and see the info and describe statements for the tables.

2. Data Cleaning & Manipulation

This step is the most important one before moving on to the analysis part. Firstly, we reduced the number of columns to make the next steps understandable. This is how we identified columns to drop from our analysis:

a. Removing columns having only NULL values
b. Removing columns having a high correlation with each other
c. Removing columns that have only 1 value
d. Removing columns that are abstract and beyond score of EDA, etc.

Next part was data cleaning:

a. Converting columns into relevant data types
b. Filling/Removing NULL values from columns

Data Manipulation:

a. Converting numerical columns to categorical variables for easier analysis
b. Removing Outliers for numerical variables

Once the data is clean, we move to Data Analysis.

3. Data Analysis

This is the most important part of the assignment. We broke down the analysis into following parts:

a. Univariate Analysis (Quantitative Variables, Ordered and Unordered Categorical Variables)
b. Segmented Univariate Analysis

c. Bi-variate Analysis

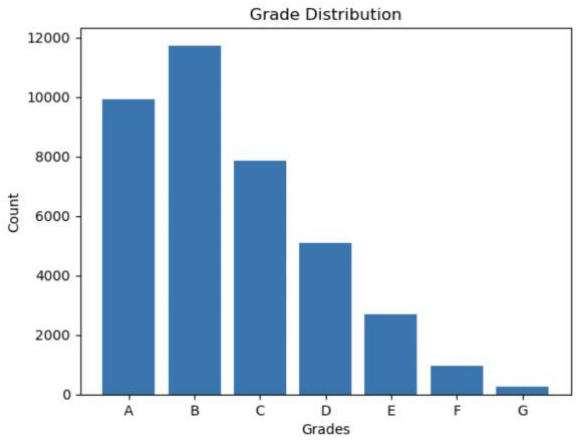
a. Univariate Analysis:

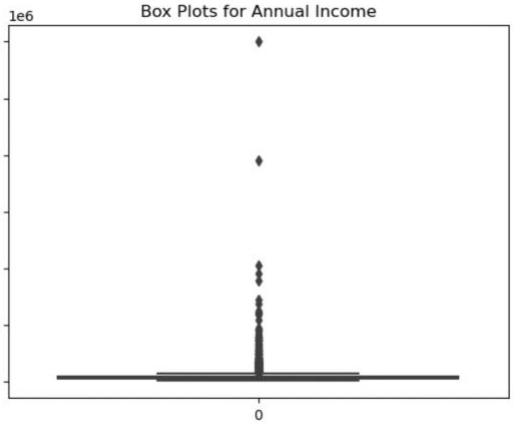
Here we used box plots to study the spread of values over individual numerical parameters. This was useful to find out the outliers, see max/min and 75th percentile values in a visually appealing way.

We also used histograms/bar plots to see the spread of information for categorical variables. This helped us to quickly visualize the peaks and troughs.

We were able to find out where and how the loans were spread across various categories by end of univariate analysis.

Top findings: There are certain outliers for 'Income', 'Open Credit Lines', 'Revolving Balance'. Also, most of the loans are given for Grade A, B, C thereby progressively reducing till G. This area helped to identify and treat outliers.



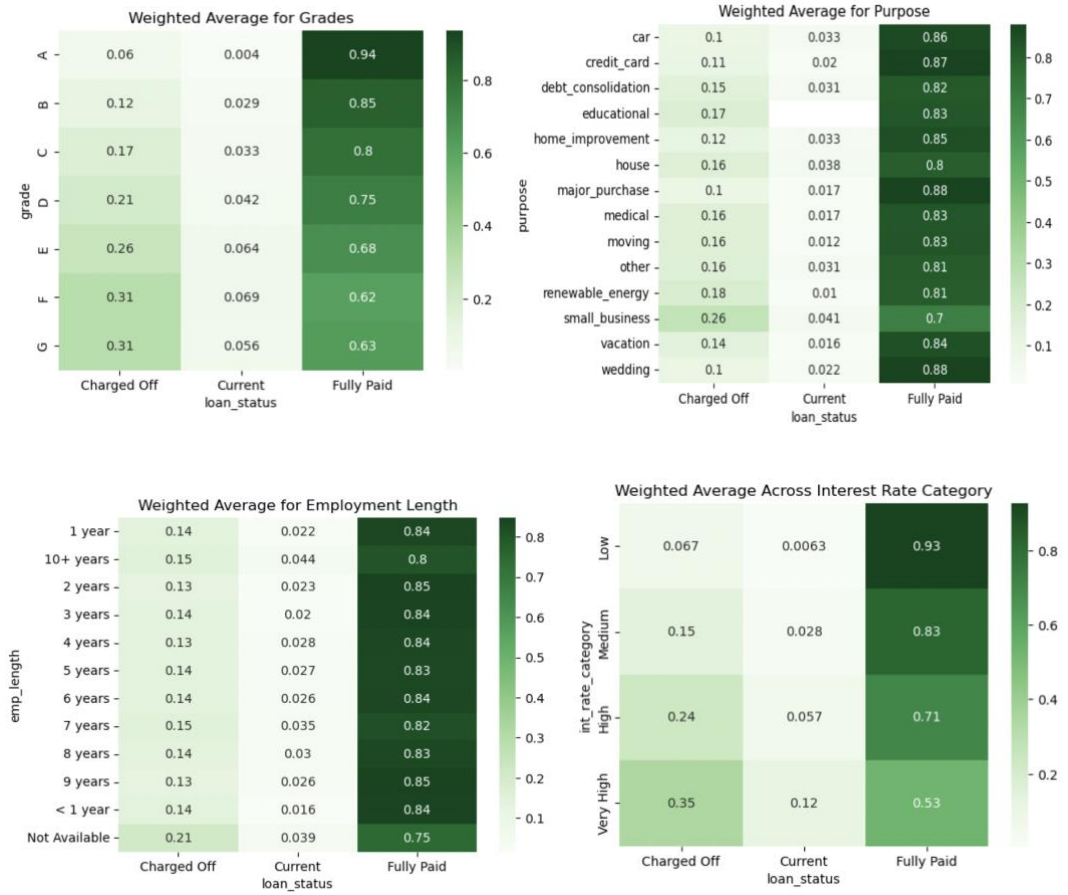


b. Segmented Univariate:

Here, we studied the variation of loan status across variables. We found the following variables having the most impact:

i. Grade
ii. Interest Rate
iii. Employment Duration
iv. Purpose
v. Loan Amount
vi. Revolving Line Utilization

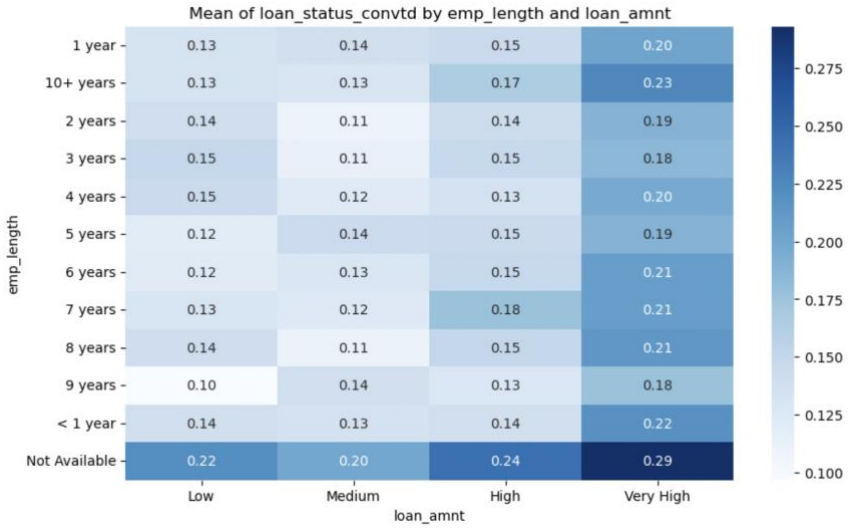
We also found interest rate has a strong correlation with Grade and rev_util and removed it from further analysis.

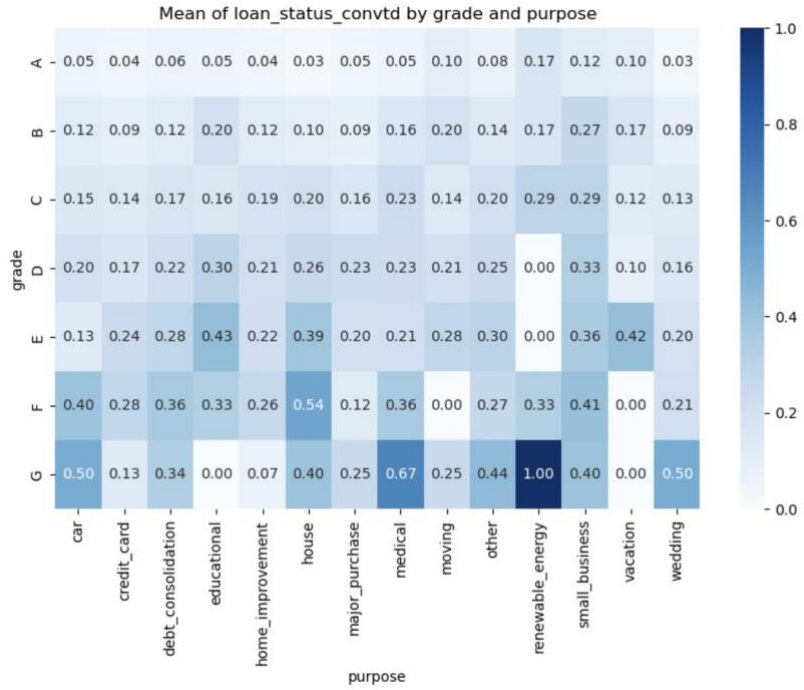


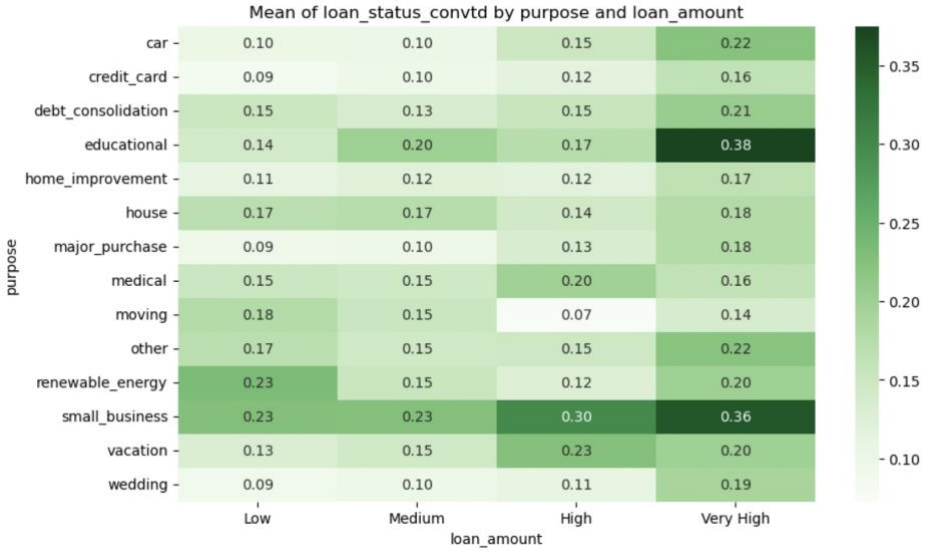
c. Bi-variate Analysis:

Additionally, we also converted the Fully Paid to 0 and Charged Off values to 1 for easier analysis.

Here, we mapped the top 5 reasons for defaulters in a 2x2 matrix with average default rate as values to get additional insights for our data.







Some important findings here are highlighted in the next section (Outcome).

4. Outcome

Some of the most important findings are:

a. Grades, Purpose of Loan, Employment Duration, Loan Amount and revolving line utilization, interest rate are most important parameters that impact loan_status
b. The bank should be cautious of customers who do not enter employment duration – they have the highest default rates of any other value

c. Grade is the easiest way to identify defaulting customers (it is also highly corelated with interest rate)
d. Revolving line utilization is directly proportional with higher default rates.
e. **Loan Amount:** For most reasons of purchase, the loan default rate increases sharply for 'Very High' loam amount (>20K $)

5. Recommendations:

1. **Considering grades**, since customers in grade E and F default the most, their profiles must very well studied before offering a loan. Even if they are offered a loan, it should be for 'safer' purposes (credit card, education, home improvement, vacation) or it should be at a higher rate of interest (already being followed)
2. Bank should identify the reason for 'NULL' **employment duration.** If this means that the customer is unemployed, then the customer should be verified carefully because these set of customers default the most in this segment. Purposes like education, moving, small business should be avoided. Similary, these set of customers are more likely to default with a higher loan amount, so a low to mid amount of loan (<15K) is fine
3. **Small Business is the purpose** of maximum default. Higher interest rates should be charged for this (irrespective of any other factor). Currently, it is only marginally higher than other reasons for loan. This purpose is followed by 'Renewable Energy'
4. **Loan Amount:** For educational and wedding purpose, it is much safer to loan less than $20K as default rates spike after that.