**Assignment 1**

CSCI 4155

By Shadman Mahmood

B00780608

# Objective

From the “loan-data” dataset we will be considering 10000 Records each to create the Credit check Dataset, which further must go through the cleaning process and data analysis. The missing data from the dataset needed to be filled in using appropriate statistical methods. Redundant fields could be dropped from the dataset and some fields could be encoded to binary values or a complete numerical format. The data is finally sorted using all the relevant information and is provided for good business practices.

# Data Summary

Our analysis is based on the "Loan-data" dataset, which contains information about the numerous loans obtained by customers from different states. The dataset spans the months of January 2022 through December 2022 and contains 10,000 records. The following table lists the characteristics of the data set:

|  |  |
| --- | --- |
| **ColumnName** | **Description** |
| id | A unique LC assigned ID for the loan listing. |
| issue\_d | The month which the loan was funded |
| loan\_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| loan\_status | Current status of the loan |
| funded\_amnt | The total amount committed to that loan at that point in time. |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| int\_rate | Interest Rate on the loan |
| installment | The monthly payment owed by the borrower if the loan originates. |
| grade | LC assigned loan grade |
| sub\_grade | LC assigned loan subgrade |
| verification\_status | Indicates if the borrower’s income was verified by LC, not verified, or if the income source was verified |
| url | URL for the LC page with listing data. |
| addr\_state | The state provided by the borrower in the loan application |
| total\_pymnt | Payments received to date for total amount funded |

# Detailed Process

Initially, the excel file needed to be formatted as there were several special characters in integer/float fields so after formatting by US-ASCII I managed to remove the special characters and for the “???” in the int\_rates column I used the replace function in excel to change the “???” to empty spaces.

After importing the dataset, I chose to separate into the data into two dataframes. X\_independant variable contained all columns from ID to verification and Y\_dependant contained addr\_state and total\_payment columns. Although technically the only dependant variable I could identify here was the total\_payment. Another notable mention of a dependable variable observed was interest rate, grade and sub\_grade. I performed a check using pd.isnull to find if there were any missing values in both my dataframes. After identifying the missing data, I proceeded to remove the year from the issue data by parsing it using str.split then I performed the imputer function with mean strategy on both the month and day columns. Finally, I concatenated them back into 1 column and saved it in the dataset. Since loan amount and funded amount fields were the same, I dropped the funded field after using it to fill all the nan rows of the loan amount. I created a dictionary for loan status and set current, fully paid and issued as 1 since they are good criteria and the others as 0 for bad criteria. After creating the dictionary, I simply used the map function to map the data. In the interest rate and installment fields I used the impute with mean strategy to fill out their respective nan values however, I do recognize interest rate could be better processed by using knn or another machine learning algorithm. Unfortunately, I am from the engineering department and I do not know yet how to apply a machine learning model. I used the grades to fill subgrades and then used a dictionary to fill the nan and subgrades which only had letters as worst case scenario as the bank expects the worst in case of missing data as mentioned in the pdf. Verification column was encoded to 0 and 1 depending on the information being verified or not. Address was mapped using a dictionary to 4 regions: East, West, MidWest, South. Total payment was filled based on impute with mean strategy.