**Credit Score Analysis Using Machine Learning Methods**

Group 4

Team members: Patrick Wolff, Safwaan Karim, Joseph Bokan, Peter Warzecha, Akshita Khandelwal

Spring 2023 Big Data Analytics (BYGB-7990-001)

**Executive Summary**

This research uses a credit score classification dataset with 100,000 records to estimate the most important determinants of an individual’s credit score and models to assist in future classification problems. 27 attributes pertaining to the personal, financial, and credit histories of the clients make up the dataset. The data was preprocessed and visualized using Python and Jupyter Notebook, and different classification techniques were utilized to create a prediction model.

After an extensive data cleaning and preprocessing process, we used a logistic regression model to find the most and least important predictors of an acceptable or unacceptable credit score by incorporating a penalty system when it came to including different features in the dataset. By adjusting the penalty level in various ways, a pattern emerged which is that how a person manages his/her credit card debt rather than his/her earning ability or usage of the card determines the ultimate credit score.

As for our credit score prediction model, we found the random forest model to be the best performing, but it is worth noting that the Knn model also performed quite well overall and it was able to generate a result much more quickly than the Random Forest model, so if time and computing capacity is limited, the Knn model could potentially act as a backup to support a future credit score classification project.

The accuracy of the models could be improved through further investigation by remembering to collect additional data regarding the clients' employment status, living situation etc. In general, this assignment provides insights regarding expected credit status and may be helpful to financial institutions and credit offices.

**Business Problem**

**Background information**

The American financial system of today is not complete without credit scores. Lenders, landlords, and employers use them to assess a person's creditworthiness. A person's credit report is used to establish their three-digit credit score, which represents their chance of making timely payments on their debts. While a low credit score implies a high chance of default, a high credit score suggests solid creditworthiness. Due to the widespread usage of credit scores by businesses, it is essential for people to keep a decent credit score in order to fully participate in contemporary society.

In some circumstances, getting a loan, renting a house, or even getting a job depends on maintaining a solid credit score. However, a lot of people don't know how to raise their credit scores or what factors affect them. People who lack knowledge and understanding may pass up opportunities or be subject to higher interest rates, both of which have an impact on their financial wellbeing. The purpose of this research is to determine the key indicators of a credit score that is considered acceptable and to offer advice on how people can raise or maintain their credit scores. Individuals can improve their financial situation by making educated judgments and taking the appropriate steps by understanding the important aspects that affect credit ratings.

**Problem Statement**

What factors most strongly affect a person's credit score and what are the main determinants of creditworthiness in the United States?

How can people increase their creditworthiness and keep a high credit score to increase the likelihood that they will be accepted for loans, rents, and employment opportunities?

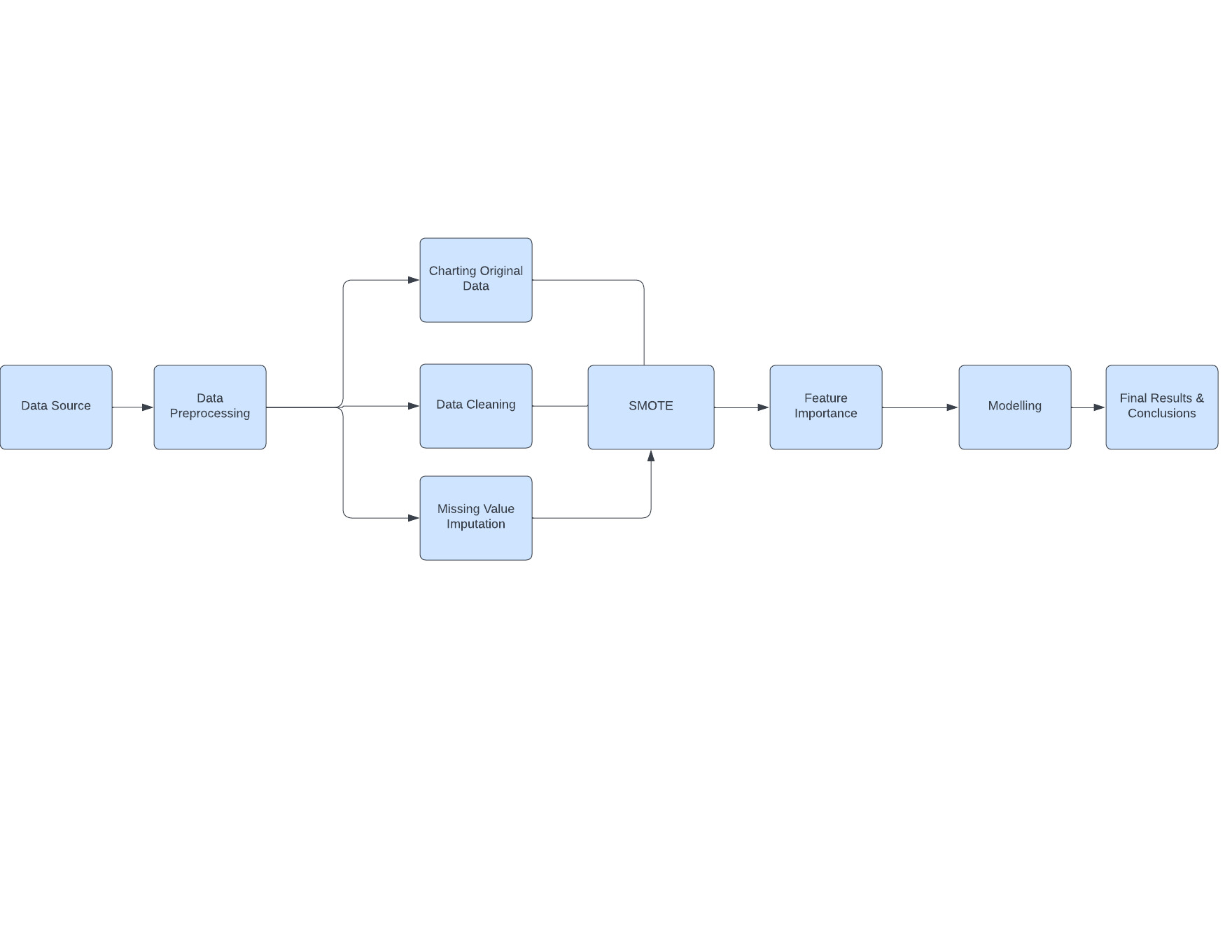
Which factors have little to no bearing on a person's creditworthiness, and what are the common myths and misconceptions about credit scores?

**Dataset Description**

The dataset we used comes from Kaggle and has credit related information for 100,000 individuals with 27 features. The features are a mix of categorical and numerical variables. Below is the breakdown of the original dataset prior to any data cleaning and preprocessing.

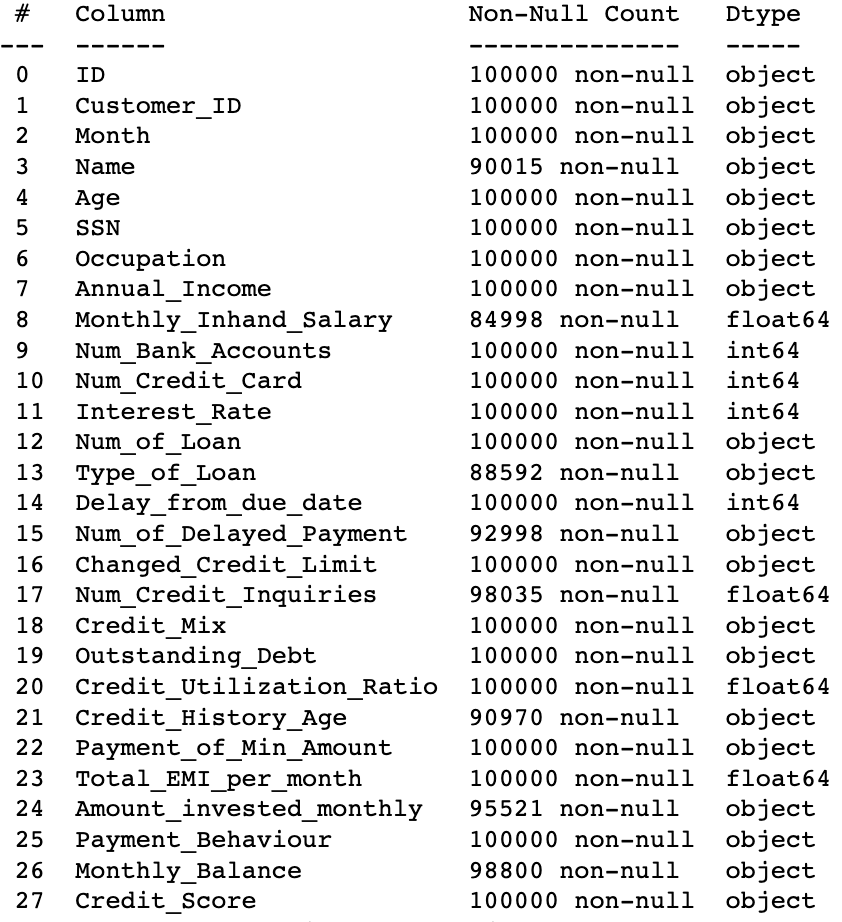
| Initial Variables List | | | | | |
| --- | --- | --- | --- | --- | --- |
| ID | Customer\_ID | Month | Name | SSN | Occupation |
| Num\_of\_Loan | Annual\_Income | Monthly\_Inhand\_Salary | Num\_Bank\_Accounts | Num\_Credit\_Card | Interest\_Rate |
| Changed\_Credit\_Limit | Num\_Credit\_Inquiries | Type\_of\_Loan | Delay\_from\_due\_date | Num\_of\_Delayed\_Payment | Credit\_History\_Age |
| Outstanding\_Debt | Credit\_Utilization\_Ratio | Credit\_Mix | Payment\_of\_Min\_Amount | Total\_EMI\_per\_month | Amount\_invested\_monthly |
| Payment\_Behaviour | Credit\_Score | Age |  |  |  |

**System Design**

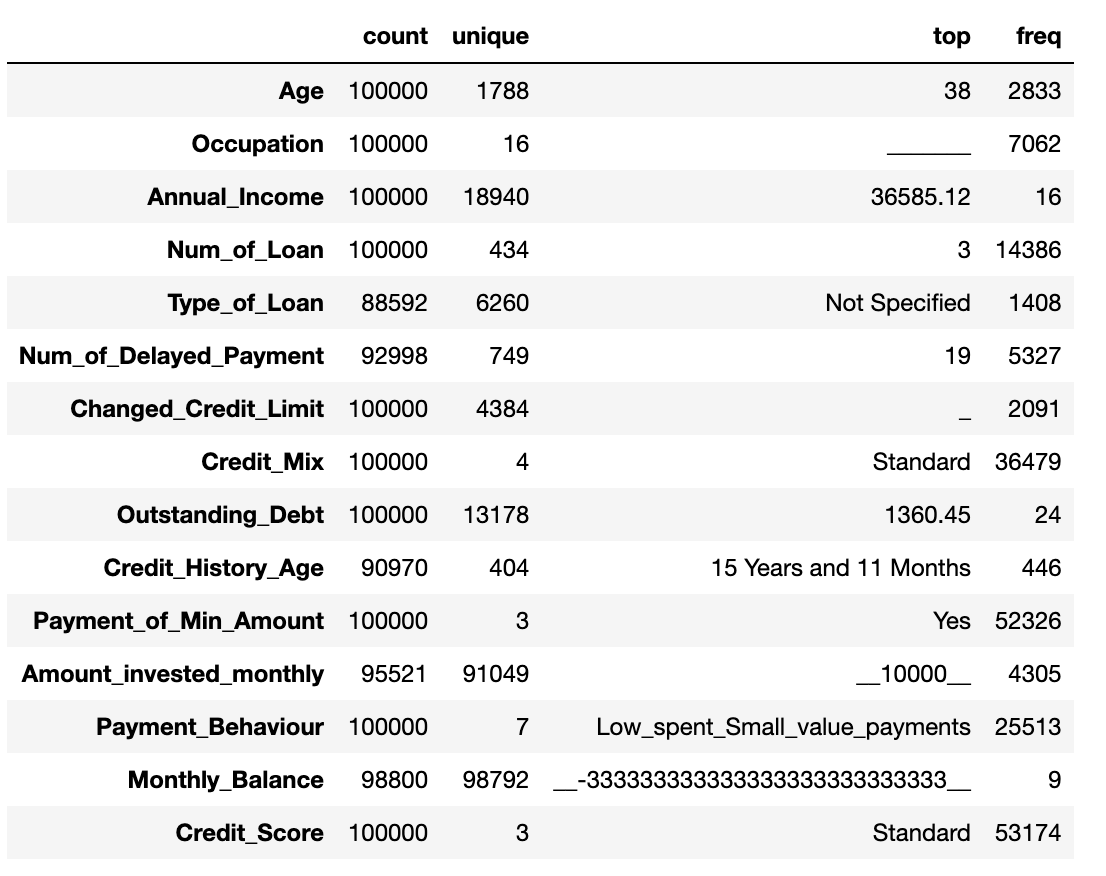


**Exploratory Data Analysis**

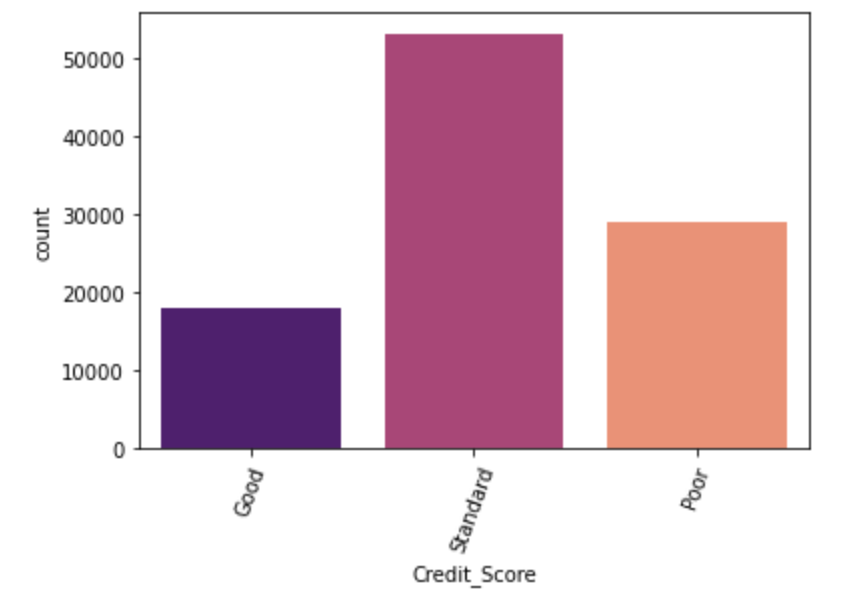
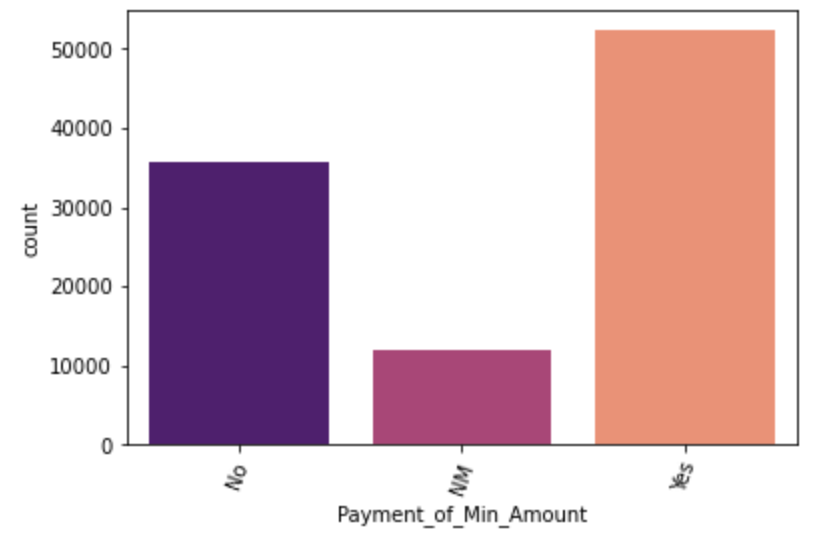
Before we started the data cleaning and preprocessing we decided to analyze our dataset and the following table shows us the name of each column, an idea of the no. of missing values and the data types we are working with. We came to see that a lot of the columns which should be a numerical data type were incorrectly classified as categorical. For example, the Annual\_Income and Monthly\_Balance column among others.

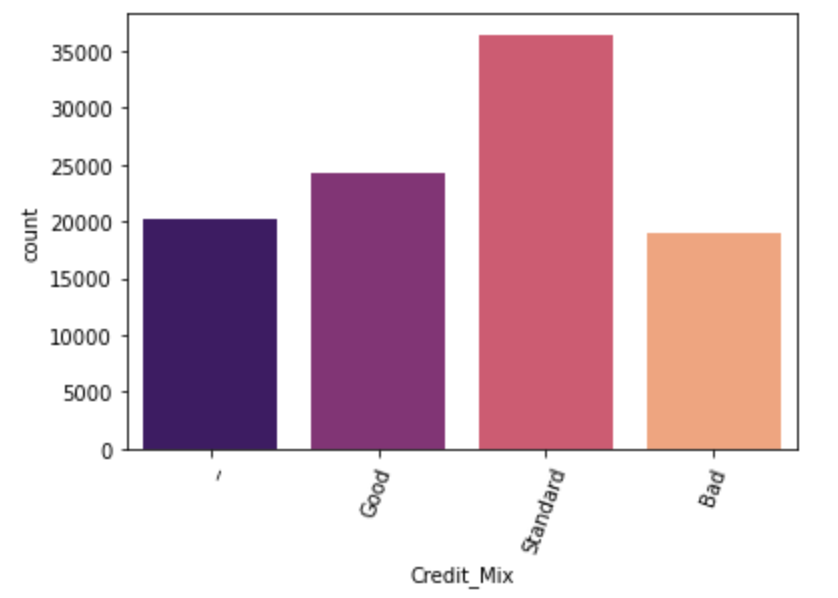
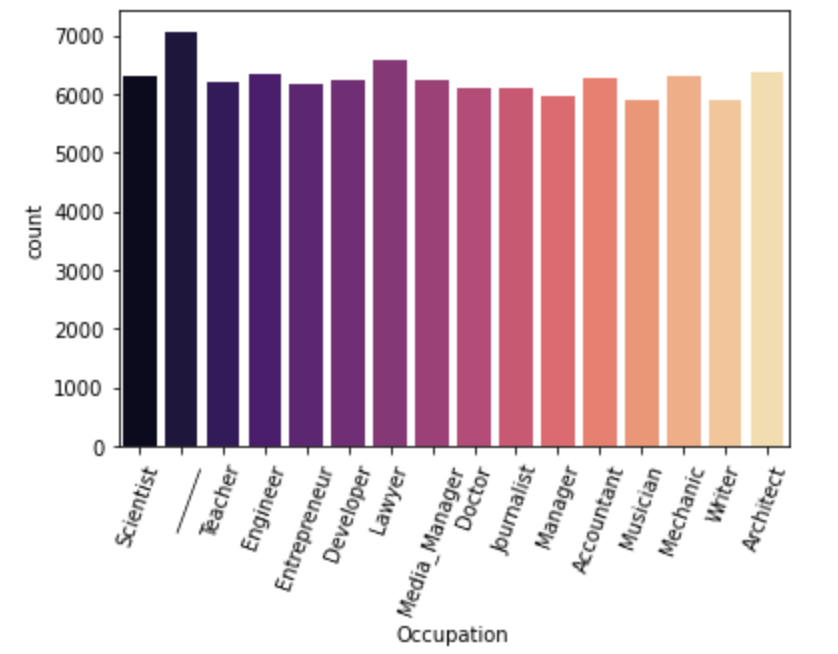
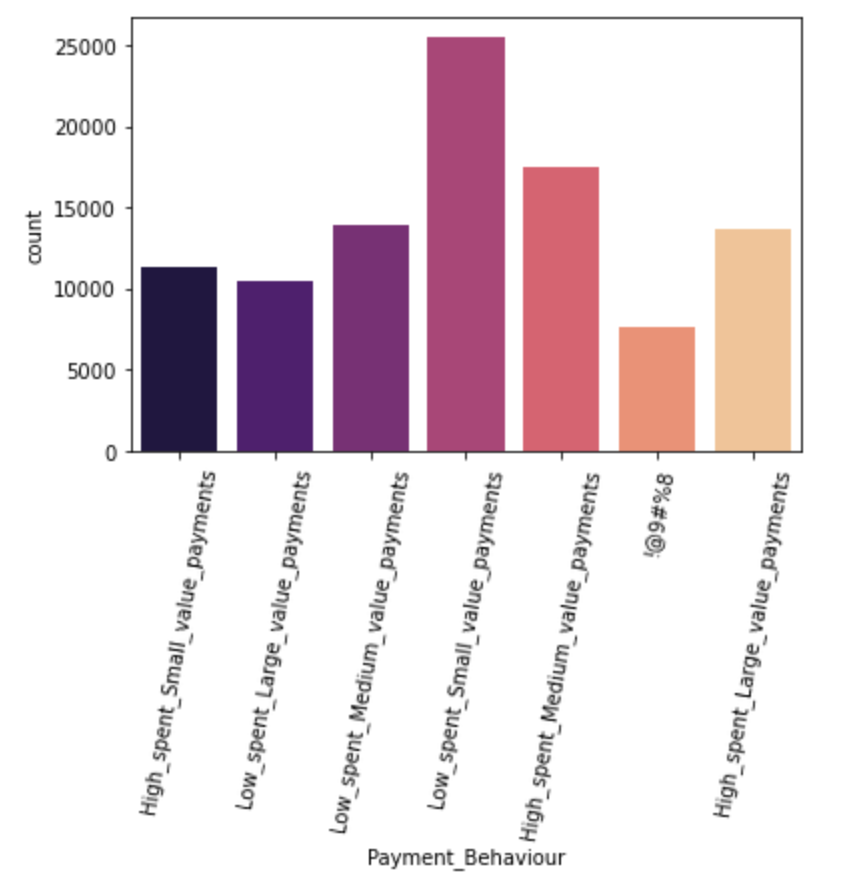


The table below gives us a different type of breakdown where we can see that the columns misclassified as objects/categorical variables have strings as values here and there. This will need to be dealt with at the data cleaning and preprocessing stage.



We also charted out the data for various columns that needed a deeper look and found that our target variable, Credit\_Score, is unequally distributed, Payment\_of\_Min\_Amount has an irregular value, Payment\_Behaviour has a several columns with a combination of special characters while the Occupation and Credit\_Mix columns have several rows filled by ‘\_\_\_\_\_\_\_’ and ‘\_’ respectively. These are all issues we tackled in the data preprocessing stage.





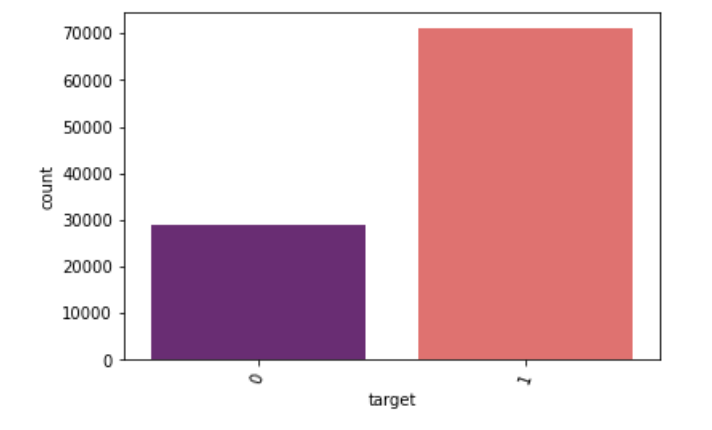
**Data Cleaning & Preprocessing**

First, we tackled the issue of numerical columns having string values. We started by creating a user defined function to get rid of all the special characters from every row from the columns concerned and putting in a NaN value where there was a missing value.

Credit\_History\_Age had its rows in the form of ‘13 years and 2 months’, so we also wrote a function to turn it into the total no. of months instead for each row of data. We then proceeded to replace all the missing values for the categorical variables with the mode of that column and then the numerical variables with the median for that column after we got rid of the erroneous negative values. Note, for occupation, we used the second most frequent value for missing value imputation as the irregular value was the most frequent value before the preprocessing.

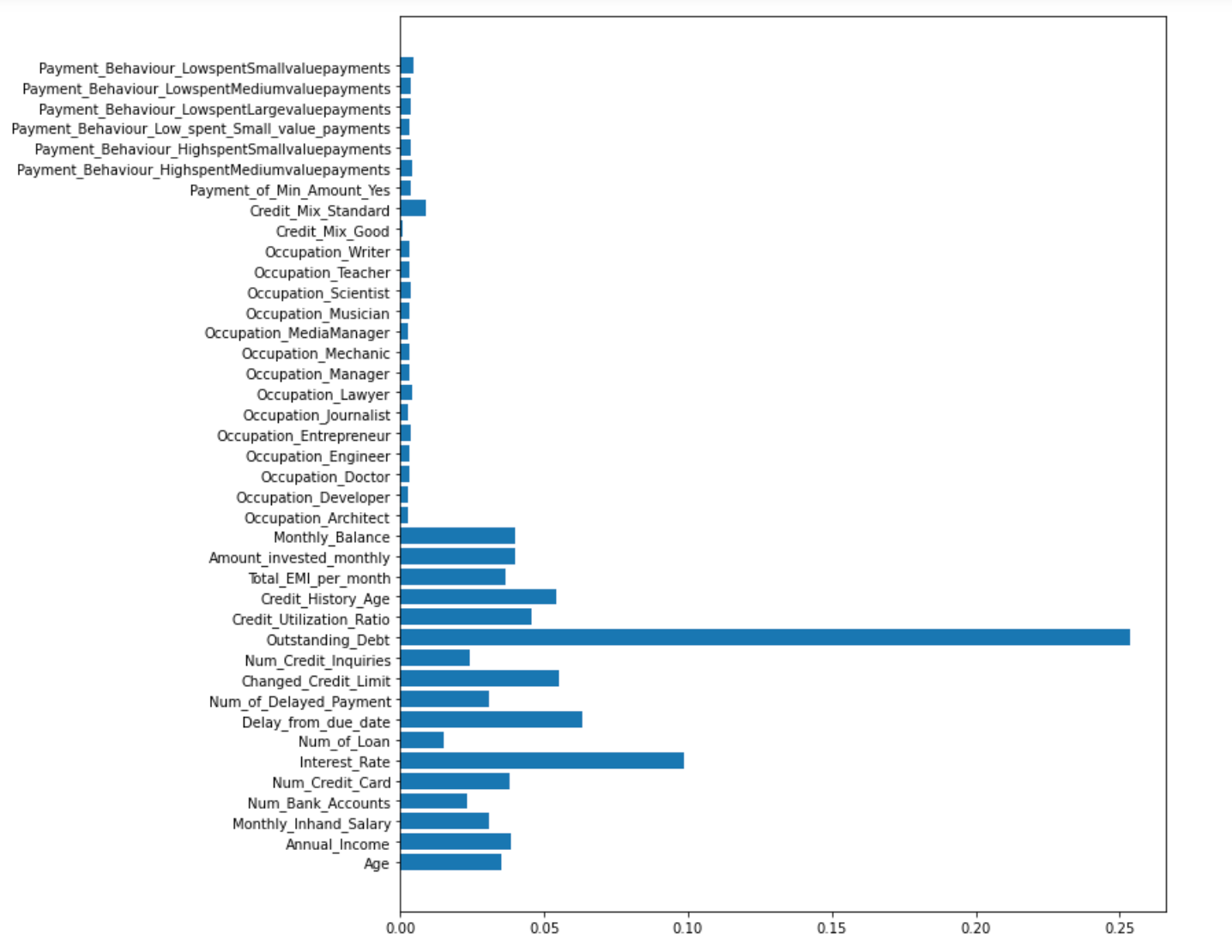
Our target variable, Credit\_Score, was initially a multiclass variable with the categories ‘Good’, ‘Standard’ and ‘Poor’, but as a part of our modeling, we wanted to use logistic regression, so we turned this column into a binary variable with “Good” and “Standard” taking on the value of 1 and “Poor” taking on the value 0. We created a new column called ‘target’ with that mapping and then dropped the original column.

After the mapping, we can see below that the dataset is still very unbalanced, so we applied SMOTE (Synthetic Minority Oversampling Technique) to balance our dataset and created dummy variables for the next step of the preprocessing.



Lastly, to maintain a parsimonious structure for our model, we decided to find the most and least relevant features in the dataset and then drop the less important features. We used a method that employs the use of a Random Forest model to chart the relative importance of each and every feature.

From the table on the next page, it can been seen that Payment\_Behaviour, Credit\_Mix, Occupation, 'Num\_Credit\_Card', 'Num\_Bank\_Accounts' and 'Num\_of\_Loan' are not as important in this dataset, so we decide to drop the corresponding columns and then we moved on to building our algorithm. We also dropped the interest\_rate column as it can be assumed the interest rate a person is charged for their credit card is, in a way, a proxy for their credit score i.e. The higher the interest rate someone is charged on their credit card, the higher the risk and, therefore, the more likely it is that the person has a bad credit score.



| Final Variables List | | | | |
| --- | --- | --- | --- | --- |
| Interest\_Rate | Annual\_Income | Monthly\_Inhand\_Salary | Amount\_invested\_monthly | Monthly\_Balance |
| Changed\_Credit\_Limit | Num\_Credit\_Inquiries | Type\_of\_Loan | Delay\_from\_due\_date | Num\_of\_Delayed\_Payment |
| Credit\_History\_Age | Payment\_of\_Min\_Amount | Credit\_Mix | Outstanding\_Debt | Credit\_Utilization\_Ratio |
| Total\_EMI\_per\_month | Credit\_Score |  |  |  |

**Algorithm Development, Results and Evaluation**

Models

For our analysis, we decided to use a logistic regression model to help with finding the most and least important predictors of an acceptable or unacceptable credit score and then we created a k-nearest neighbor and random forest model to serve as a predictive model that could be utilized in a more practical setting where new data could be fed into the model to assist with real world situations like loan applications and other financial services.

Logistic Regression

Given that our target variable is now a binary variable with 1 meaning that the individual has either a “Standard” or a “Good” score and 0 meaning that the individual has a “Poor” score, we can use a logistic regression model with a pre-specified alpha to find the features from our dataset that are the most important determinants of what value someone’s credit score ends up being.

For the logistic regression model we created, as we kept increasing the penalty for including a feature on the model, the first features to drop out were Monthly\_Balance, Credit\_Utilization\_Ratio and Monthly\_Inhand\_Salary, in that order, which means that those are the least important features out of all the features that we did not drop earlier.

When testing alphas beyond 15000, The last 3 features to drop out were Delay\_from\_due\_date, Num\_Credit\_Inquiries and Outstanding\_Debt in that order, which tells us that those 3 features are the most important and they determine what kind of credit score a person ends up having based on our dataset.

K-Nearest Neighbors

The K-nearest neighbors model had an accuracy of 79% and also reported an impressive precision, recall and F1-score across both the classes that we were investigating as a part of the research

| Credit Score Type | Precision | Recall | F1-score |
| --- | --- | --- | --- |
| Poor | 0.60 | 0.80 | 0.68 |
| Standard/Good | 0.91 | 0.78 | 0.84 |

Random Forest

The random forest model had an accuracy of 87%. This is the better performing model out of the 2 that we are looking at and managed to beat out the Knn model on every metric except for the recall on the poor credit score class where it manages to just equal the metric, so we will go with this model for any use case moving forward

| Credit Score Type | Precision | Recall | F1-score |
| --- | --- | --- | --- |
| Poor | 0.75 | 0.80 | 0.78 |
| Standard/Good | 0.92 | 0.89 | 0.90 |

**Conclusion, Limitation and Future Research**

In conclusion, from the results above, we can surmise that rather than how much a person uses a credit card (Credit\_Utilization\_Ratio), how much a person makes or has in his/her account on a month to month basis (Monthly\_Balance & Monthly\_Inhand\_Salary), the factors that appear to influence a credit score classification appears to be more tied to how a person manages his/her debt in terms of size (Outstanding\_Debt), punctuality when it comes to meeting payment deadlines (Delay\_from\_due\_date) and how many additional credit cards a person requests (Num\_Credit\_Inquiries).

A limitation of our research is that given the size of the dataset and the number of features provided, we had to do relevance importance and trim the number of features to create our final model as we were restricted from a computational capacity standpoint. Since processing of big data can take up significant CPU load, it is worth mentioning that perhaps with more powerful systems to support us, we may have ended up with slightly different results compared to our findings above.

Additionally, for future research, it would be helpful to build out this model on a periodic basis with changing economic conditions as the importance of a certain feature may increase and decrease over time. E.g. Now, we are currently in a cycle of tightening monetary policy with rising interest rates, strong labor market, but with an expectation of a recession later this year. These economic factors may play a role in how financial institutions choose to prioritize features that make up the credit score of an individual.

**References**

Rohan Paris. Jun 22, 2022. *Credit score classification.* Version 1. Retrieved from: <https://www.kaggle.com/datasets/parisrohan/credit-score-classification?resource=download>

Terence Shin. Feb 26, 2021. *Understanding Feature Importance and How to Implement it in Python.*

Retrieved from: <https://towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285>