Name: Ravindra Neralla

DSC630 – Predictive Analytics

**Executive Summary:**

**Introduction:**

In financial industry, deciding on a loan application, whether to approve or not is a two-edged sword. Bank should not lose business by denying a legitimate customer, who can repay. Also, it should not approve loan to in-eligible customer. Banks are playing important role in challenging times like now, with COVID pandemic across the globe.

**Objective:**

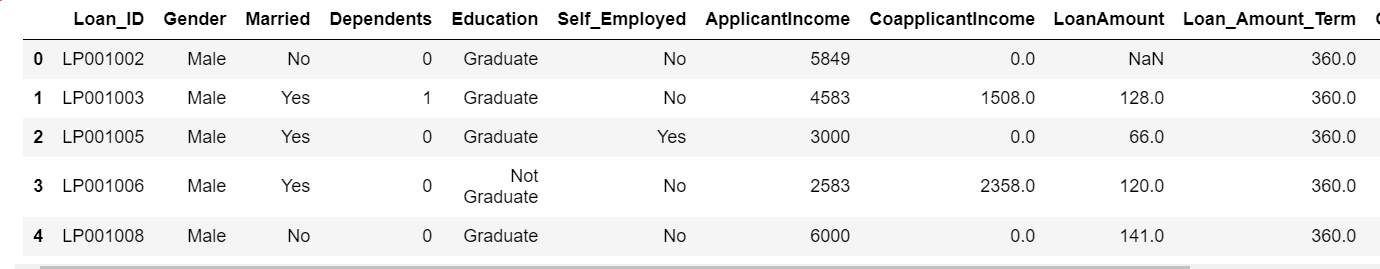
The major aim of this project is to predict whether a loan application would be approved or not. Also, it is important to know the key features in deciding the loan application, this will help relationship managers to communicate with the customers.

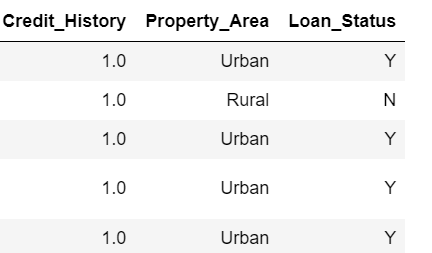
**Approach:**

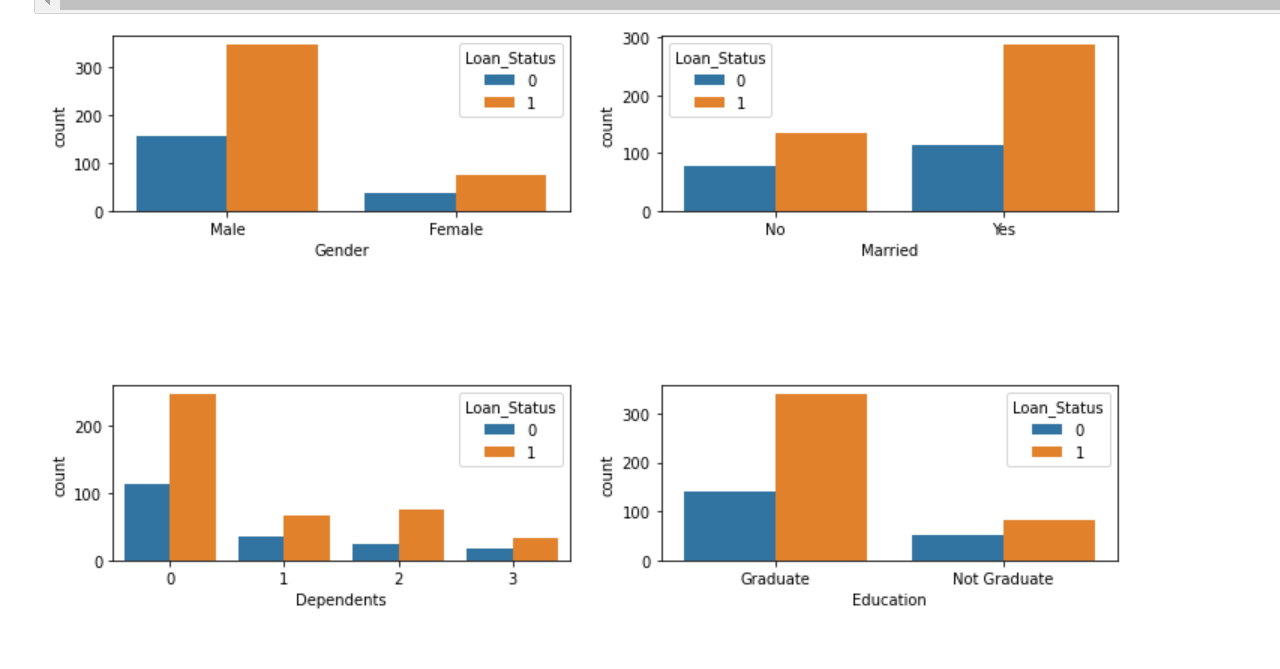
Technical approach involves understanding the data by doing exploratory data analysis. It involves drawing multiple bar charts and histograms to observe the independent variables with respect to target variable.

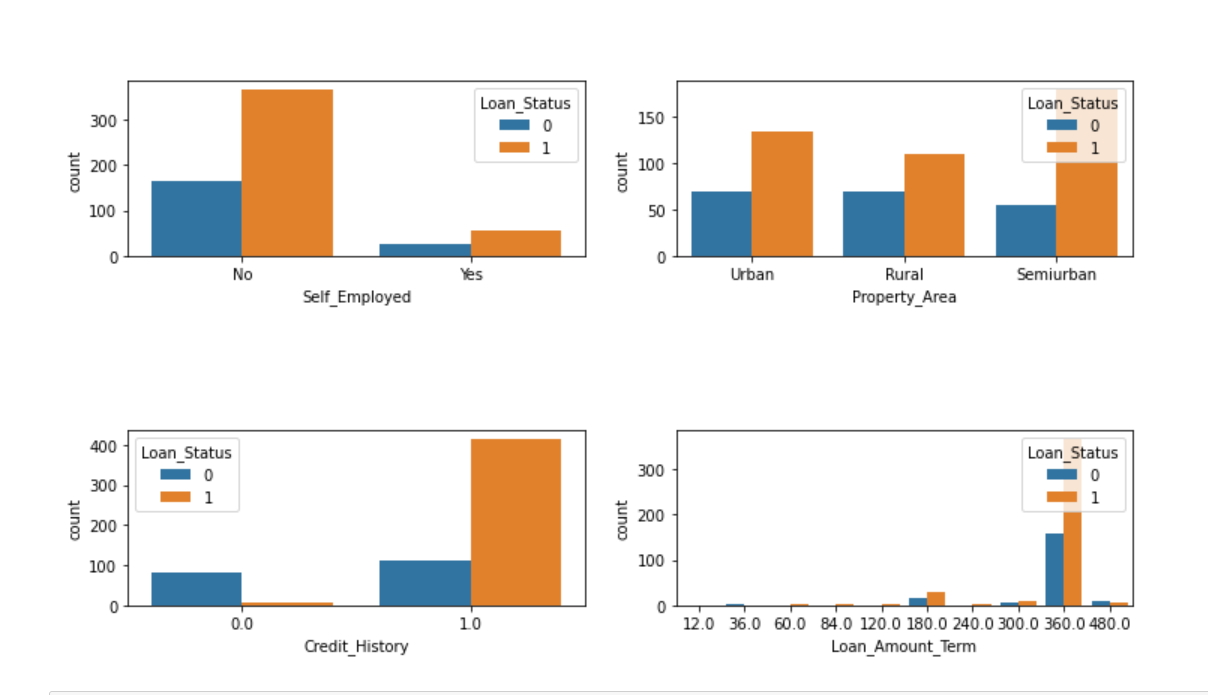
After completing the preliminary data analysis and cleansing, build model using the different machine learning algorithms and evaluate the accuracy of the model through confusion matrix.

Sample Data:









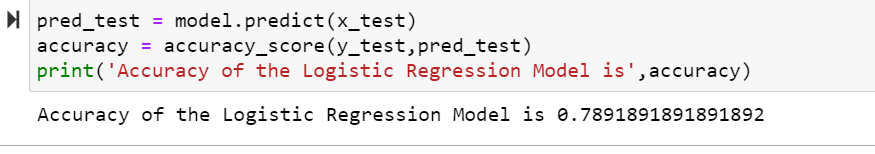
* **Gender:** There are around 300 Men and 100+ Women in the dataset.
* **Marital Status:** Around 65% of the population in the dataset is Marred. Also, married applicants are more likely to be granted loans.
* **Dependents:** Most of the loan applicants have zero dependents and are also likely to accepted for loan.
* **Education:** About 80% of the population is Graduate and graduates have higher probability of loan approval
* **Employment**: 80% of population are not self-employed.
* **Property Area:** Majority applicants are from Semi-urban and also likely to be granted loans.
* **Credit\_History:** Applicants with credit history are far more likely to be accepted.
* **Loan Amount Term:** Majority of the loans taken are for 360 Months (30 years).

**Model Building:**

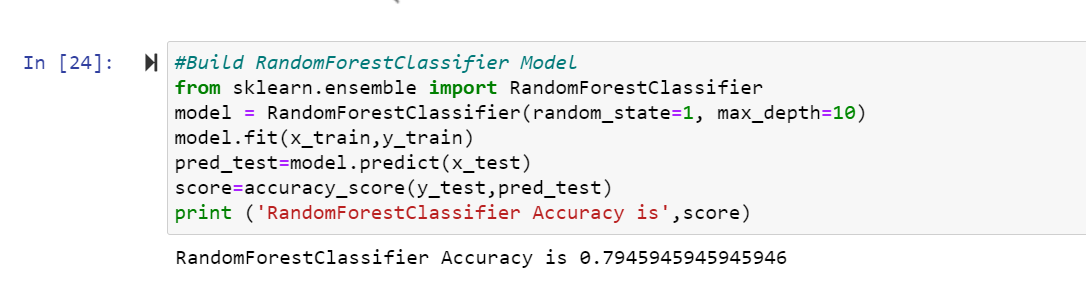
As part of the current project, four models were developed after data preparation steps. Data is split in the ratio of 70:30 for train and test, i.e. 70% of the data is fed to the model to understand the patterns and remembering the outcome, later 30% of the data is used to validate the performance of the model.

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 0.76 |
| Random Forest | 0.76 |
| Naïve Bayes | 0.77 |
| SVM | 0.65 |

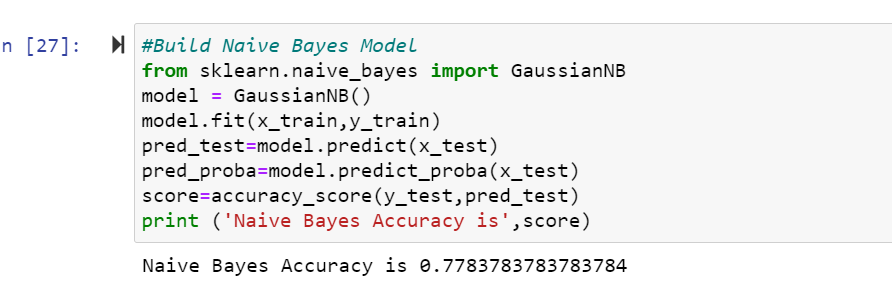
* Logistic Regression



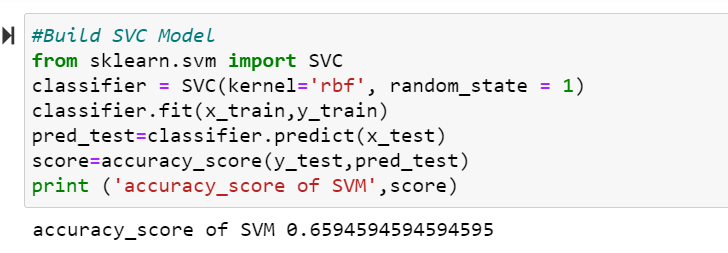
* Random Forest



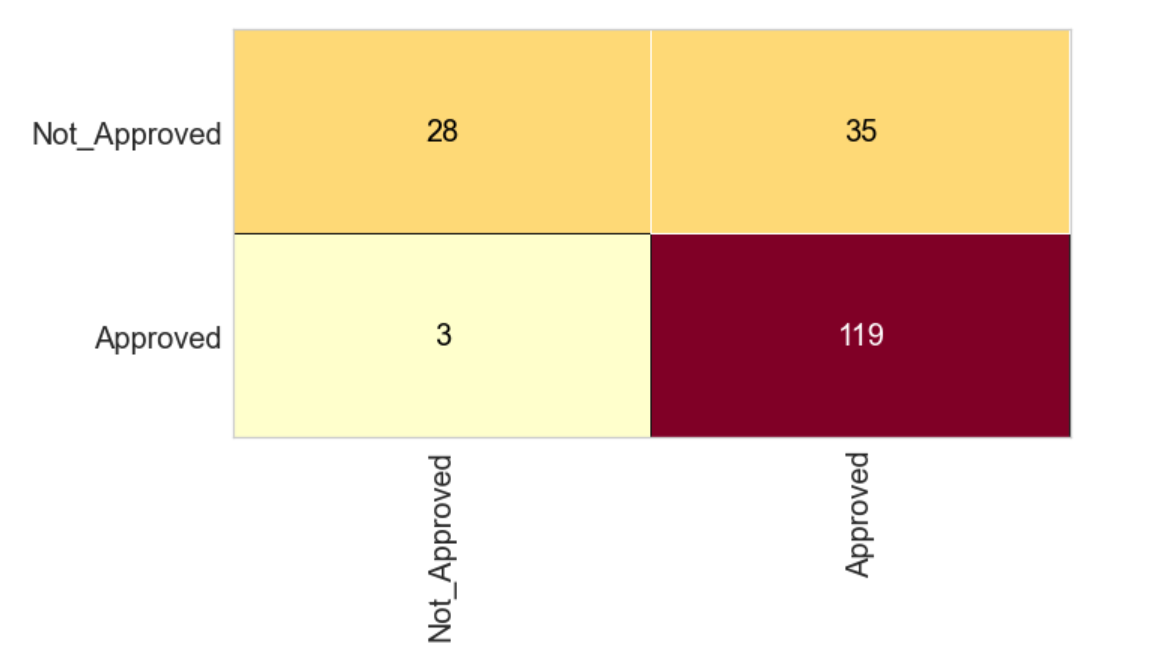
* Naïve Bayes



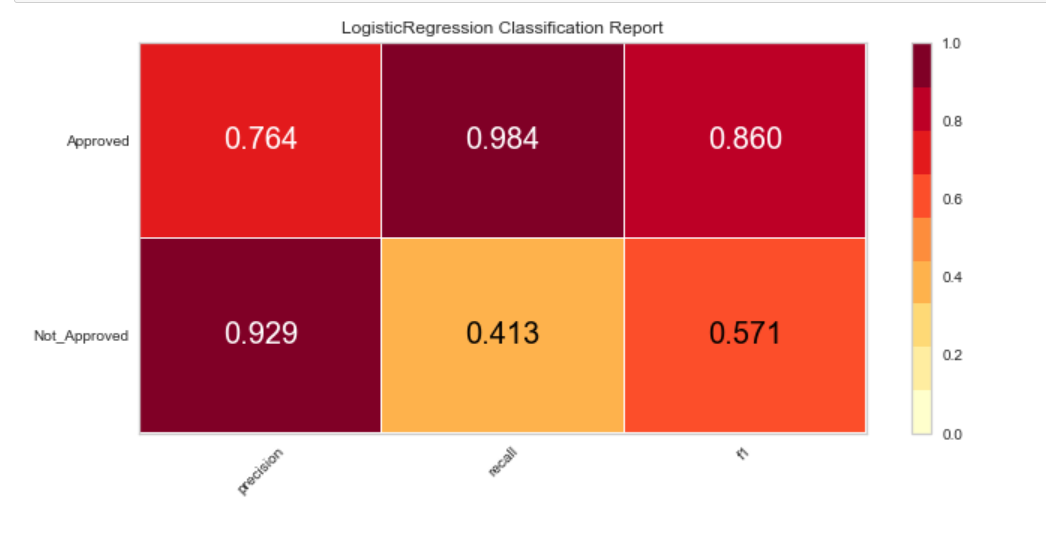
* SVM



**Confusion Matrix:**

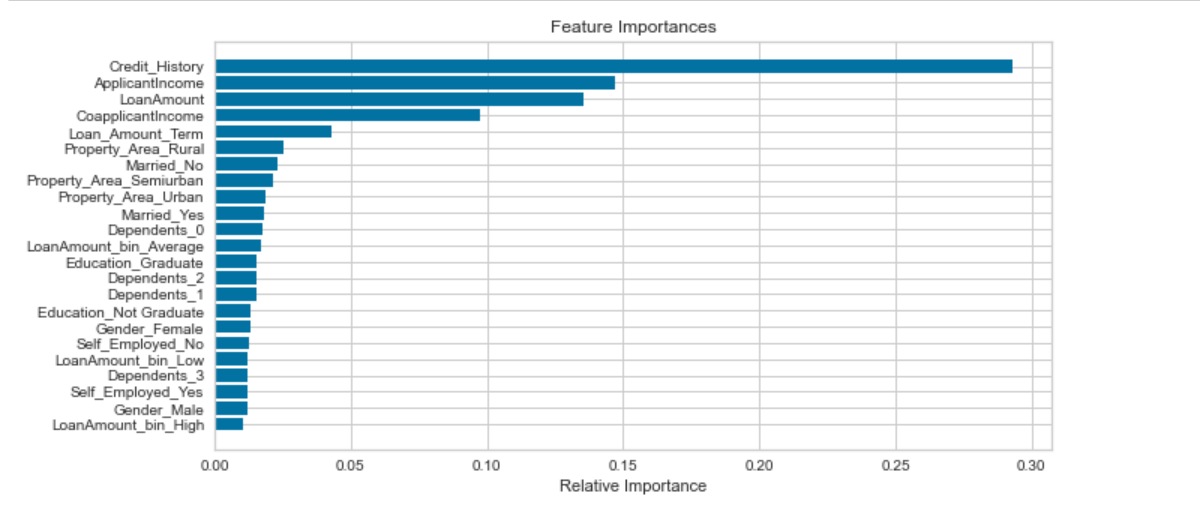
Below is the confusion matrix on logistic regression it shows that, the model is predicting with 76 percent accuracy.

**Classification Report:**



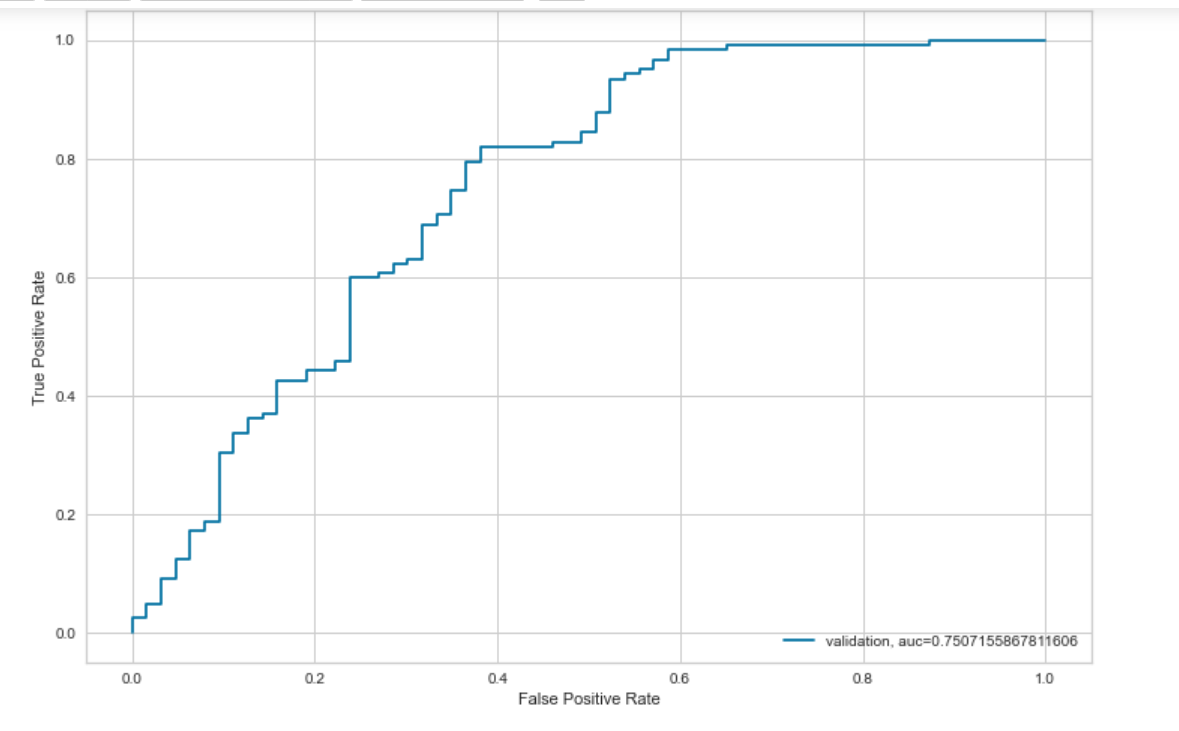
Above diagram shows precision, recall and f1 values. f1 value 0.86 indicates that the model is a good model.

**Feature Importance:**



Above chart shows that credit history is most important factor in deciding the loan application followed by applicant income and loan amount.

**Area Under the Curve:**



Area Under the Curve is 0.75, it indicates model is performing with good accuracy.

**Conclusion:**

As per the steps outlined in the project plan, I have loaded data into a data frame and performed exploratory data analysis and cleansed data. I have built four models namely, logistic regression, random forest, naïve bayes and SVM. I have selected random forest model, as it has highest accuracy of 79%. Also, I found that credit history, applicant income and loan amount are top three key contributors in deciding whether a loan application has to be approved or not.

**References:**

1. <https://www.kaggle.com/premptk/loan-approval-prediction->model/data?select=LoanApprovalPrediction.csv
2. <https://medium.com/@rahulshuklawork/prediction-of-loan-approval-with-machine-learning-539cbd2aad31>