Loan Approval Prediction

Data

The dataset for this project is downloaded from Kaggle. The training dataset and test dataset were provided separately. Both the datasets were downloaded as csv files.

Training dataset: It consists of details of the customer (like Gender, Marital Status, Number of Dependents, Education, Income, Loan Amount and others) along with the target binary column which is "Loan_Status" with values:
1. 'Y': If Loan is approved
2. 'N': If Loan is not approved

Test Dataset: It consists of similar details (like Gender, Marital Status, Number of Dependents, Education, Income, Loan Amount and others) as the training dataset but for new customers. However, the target column, that is Loan_Status is not present in this dataset.

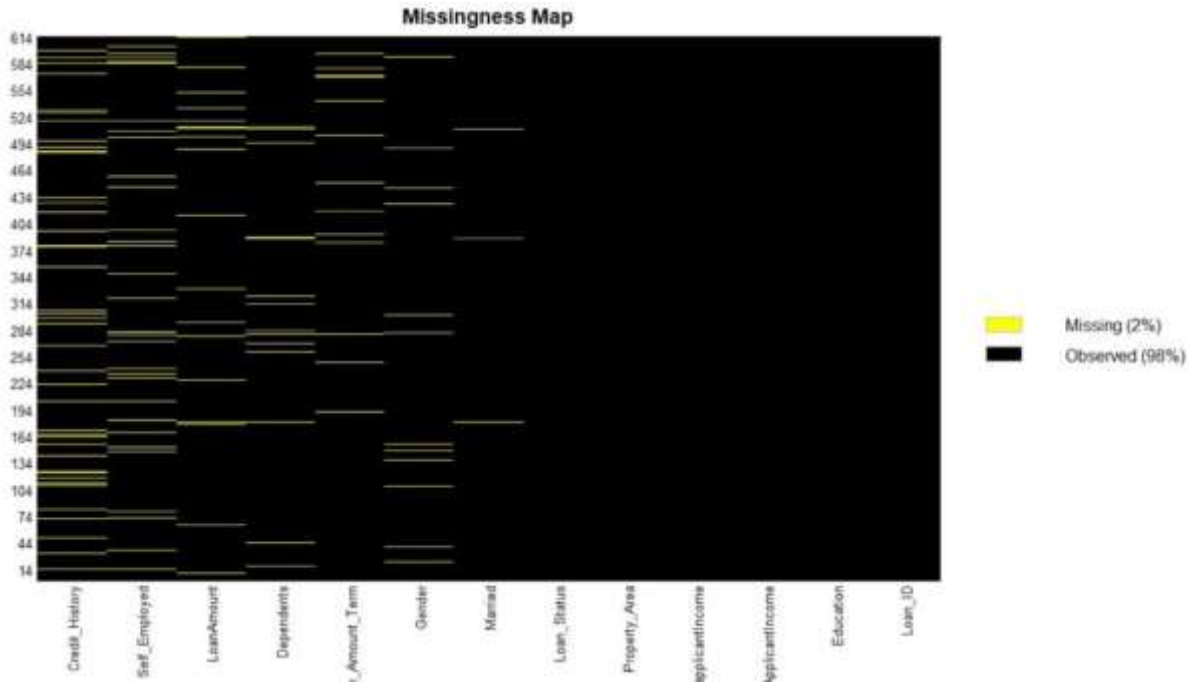
Using the training dataset, a Machine Learning model is built which will predict the "Loan_Status" for the customers in the test dataset.

Objective

The aim of this project is to predict which customers will get their loan approved (Yes/No). Therefore, this supervised classification problem which will be trained with algorithms like Logistic Regression, Random Forest, Support Vector Machines (SVM), Naive Bayes Algorithm.

Data Preprocessing

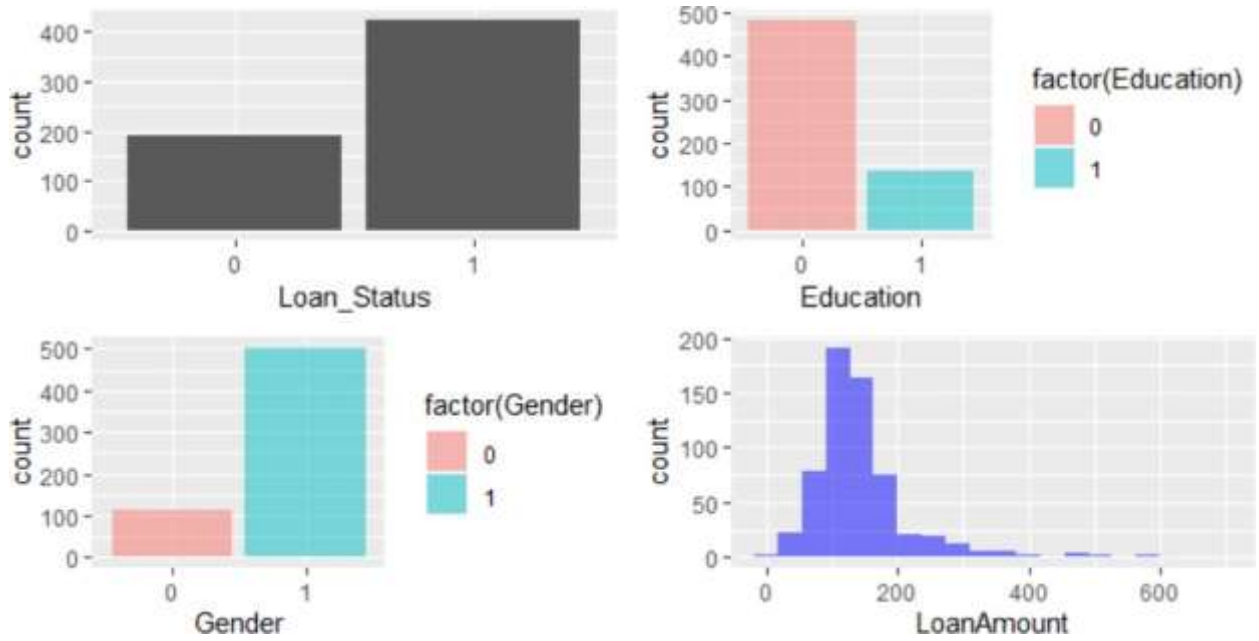
Many entries in the data had "NA" values as depicted in the graph below



It can be seen that there were 2% missing values in the complete data. The appropriate changes were made to fix these values.

Further, the “Loan_ID” column was removed from both training as well as test dataset since it was not adding any value. Also, few of the variables were converted from categorical to factor variables. Lastly, scaling was performed so that the model can easily learn and understand the problem.

Data Visualization



It can be inferred from the above plots that in our observed data,

- Approximately 80% of loan applicants are male in the training dataset.
- There are about 75% of loan applicants that are graduates.
- Also, the loan has been approved for more than 65% of applicants.

Modelling

The training set was split into 80:20 ratio so that 80% of the data can be used to train the model and remaining 20% of the data can be used to test it. Here, the dependent variable is the status of the loan (Loan_Status), whether it will get approved or not.

The *logistic regression* which is a supervised learning algorithm can be applied since this algorithm is appropriate when the dependent variable is binary.
The model accuracy obtained using this algorithm is approximately 82%.

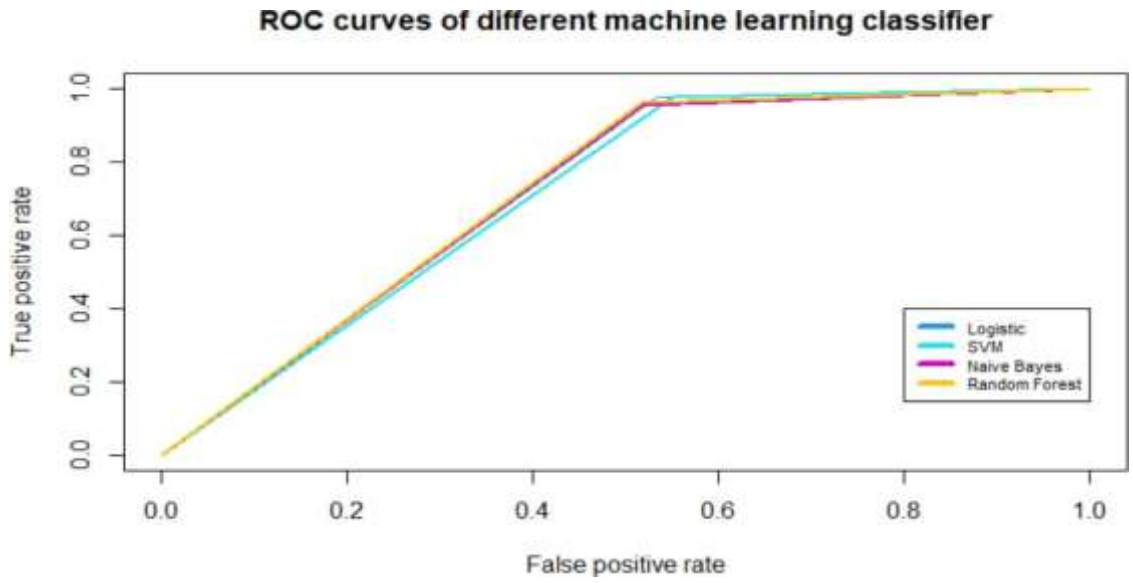
Support Vector Machine (SVM) is also a supervised learning algorithm that can be used for classification problems. The accuracy obtained using this algorithm is approximately 81%.

The *Naïve Bayes classification* algorithm is a probabilistic classifier. It is based on the models that have strong independence assumptions.
The accuracy received using this algorithm is approximately 80%.

The *Random Forest* algorithm selects the observations and features randomly to build decision trees. It then averages the results.
The accuracy received using this algorithm is approximately 81%.

Conclusion

The ROC curves obtained using all the four algorithms is shown in the graph below.



Going by the accuracies and ROC curves for all the models, the algorithms have performed equally well. Although, it can be seen from the evaluation of all the four models that Logistic Regression performed better than others followed by Random Forest and Support Vector Machine.