

Machine Learning

CHURN MODELLING ON BANK DATASET

Introduction



DEFINING THE
BUSINESS PROBLEM



EXPLORATORY
FINDINGS

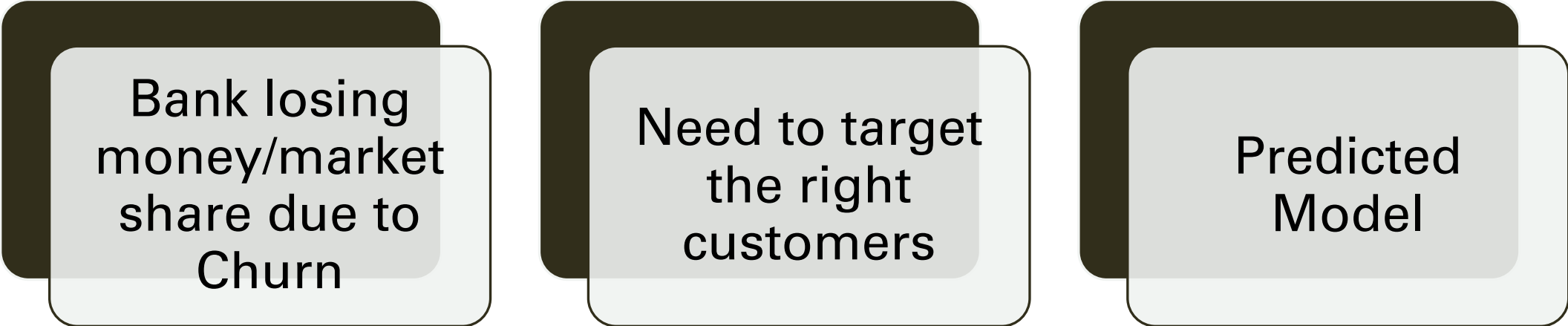


MACHINE LEARNING
MODELS AND
EVALUATION



CONCLUSION

Problem Definition



Bank losing
money/market
share due to
Churn

Need to target
the right
customers

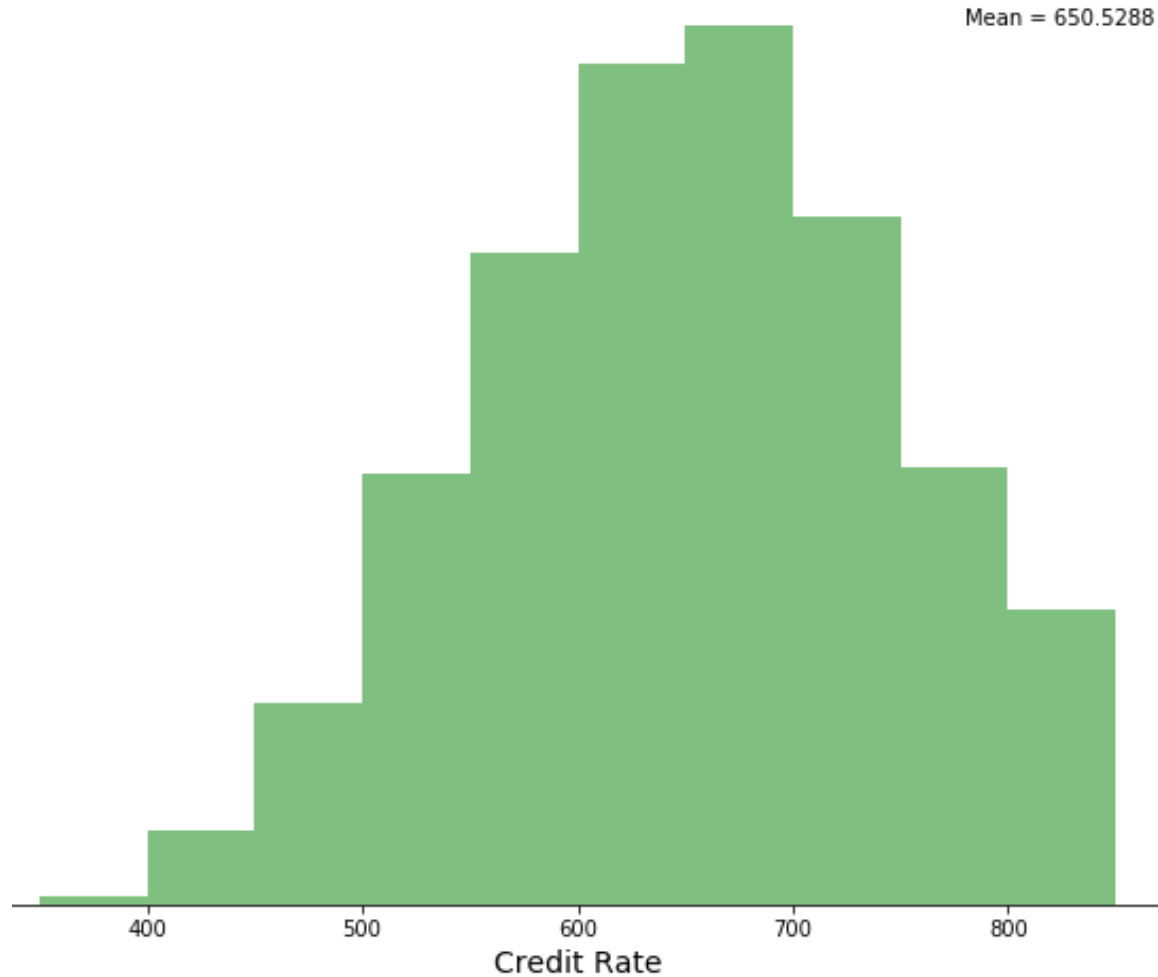
Predicted
Model



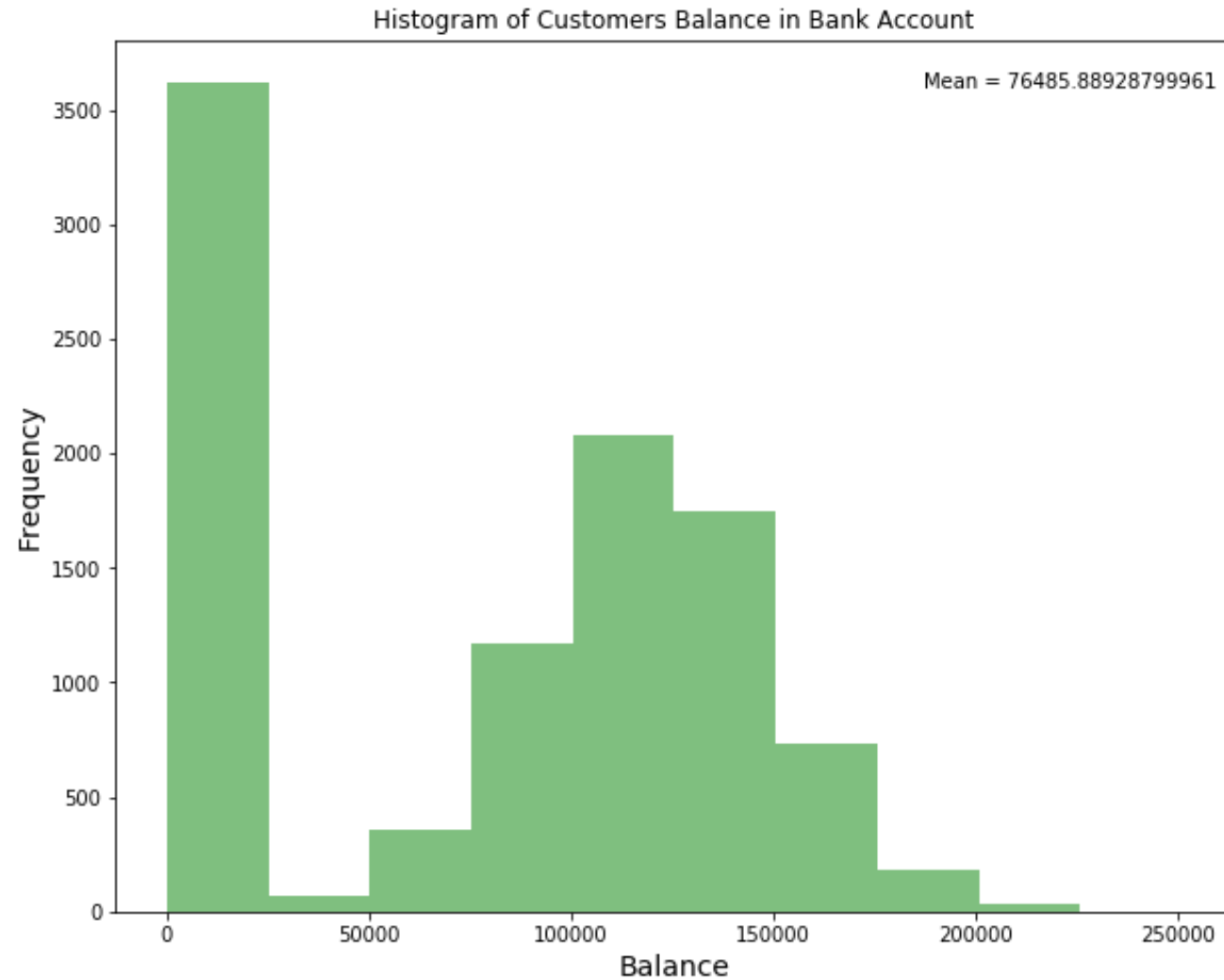
Data Exploratory

- It shows that there are **10K** entries and **12** columns in the dataset.
- Gender column has 4 null values, Age column has 6 and EstimatedSalary column has 4 null values.
- There are no null value for the other columns.

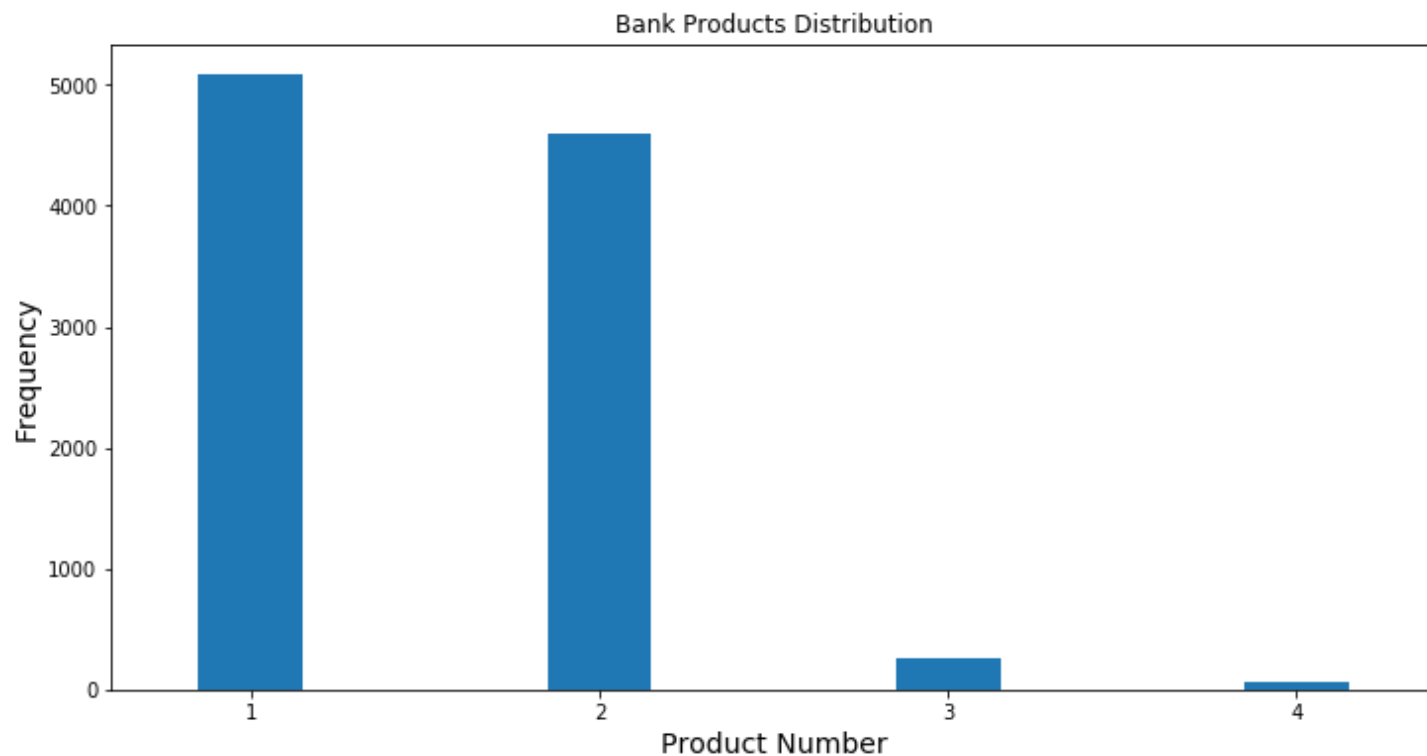
Histogram of Customers Credit Rate in Bank Account



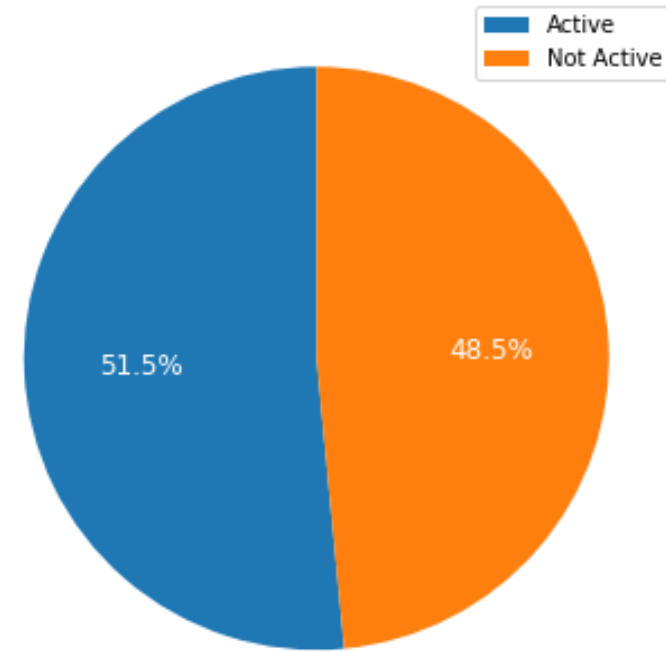
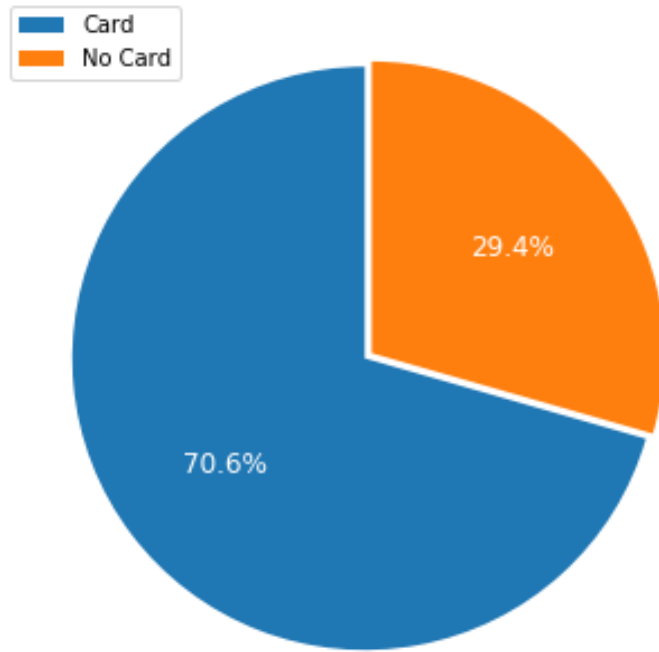
Credit Score of customers



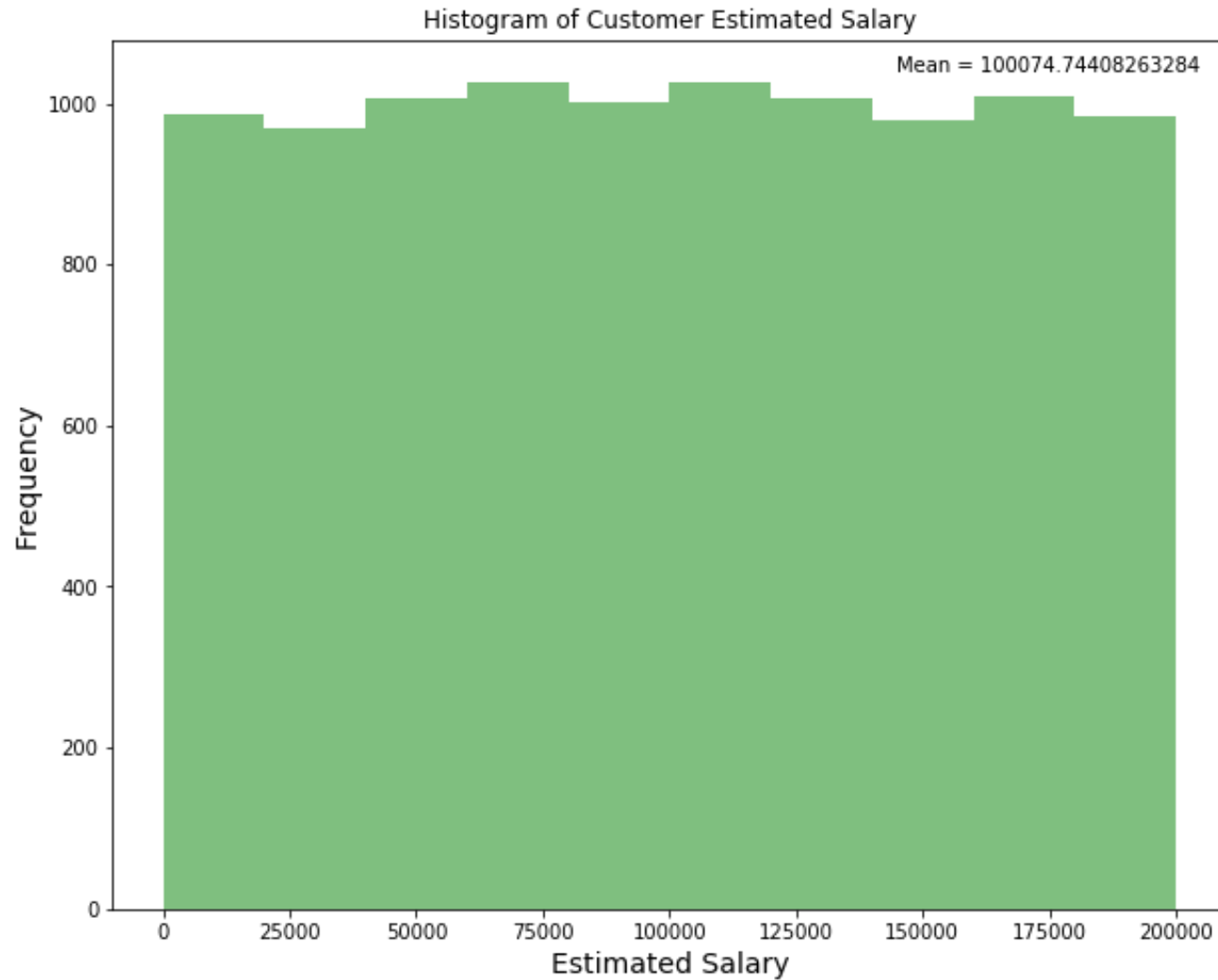
Balance in Account



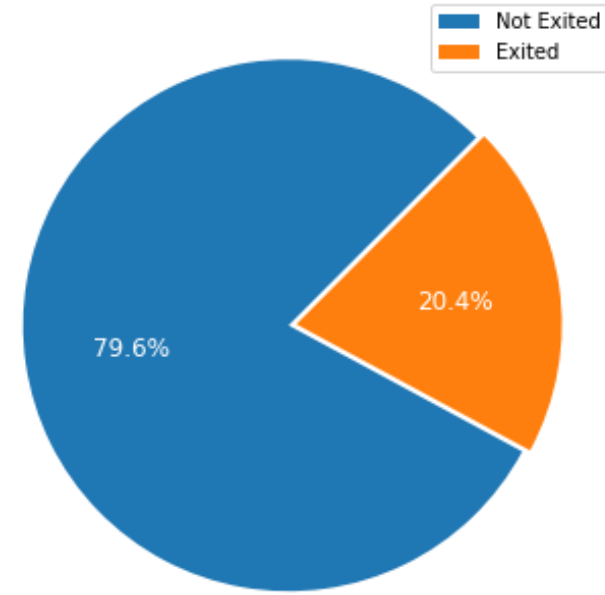
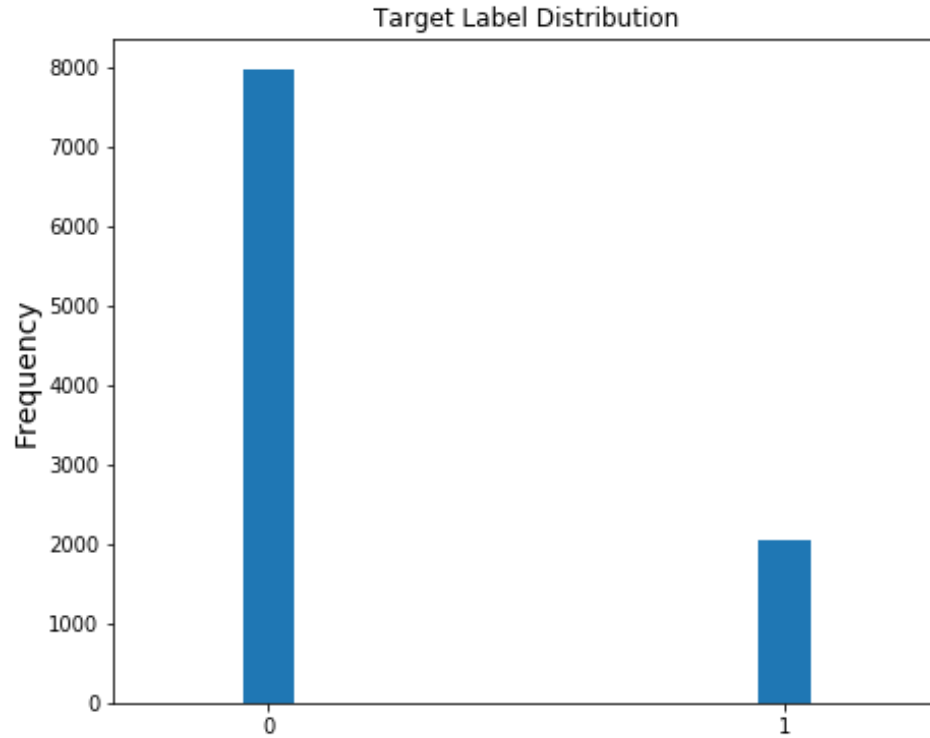
Bank Product Distribution



Credit Card and Member Distribution



Annual Estimated Salary



Class Distribution of the Target

Machine Learning Model

Naive Bayes

Decision Tree

Random Forest

Support Vector Machine (SVM)

It is a probabilistic machine learning model which is used for classification task. It is based on ****Bayes Theorem**** with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

It is easy and fast to predict the class of test data set.

It performs well in case of categorical input variables as in our case we have two categorical input variables (Geography and Gender).

Naive Bayes

Decision tree is one of the supervised machine learning algorithms. This algorithm can be used for regression and classification problems. It is mostly used for classification problems.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Decision Tree

This is an extension to the decision tree. Random forest builds multiple decision trees each based on a random sample of the training data and merges them together to get a more accurate and stable prediction by means of voting.

It can handle thousands of input variables without variable deletion which in our case will also help as we have multiple input attributes.

Random Forest

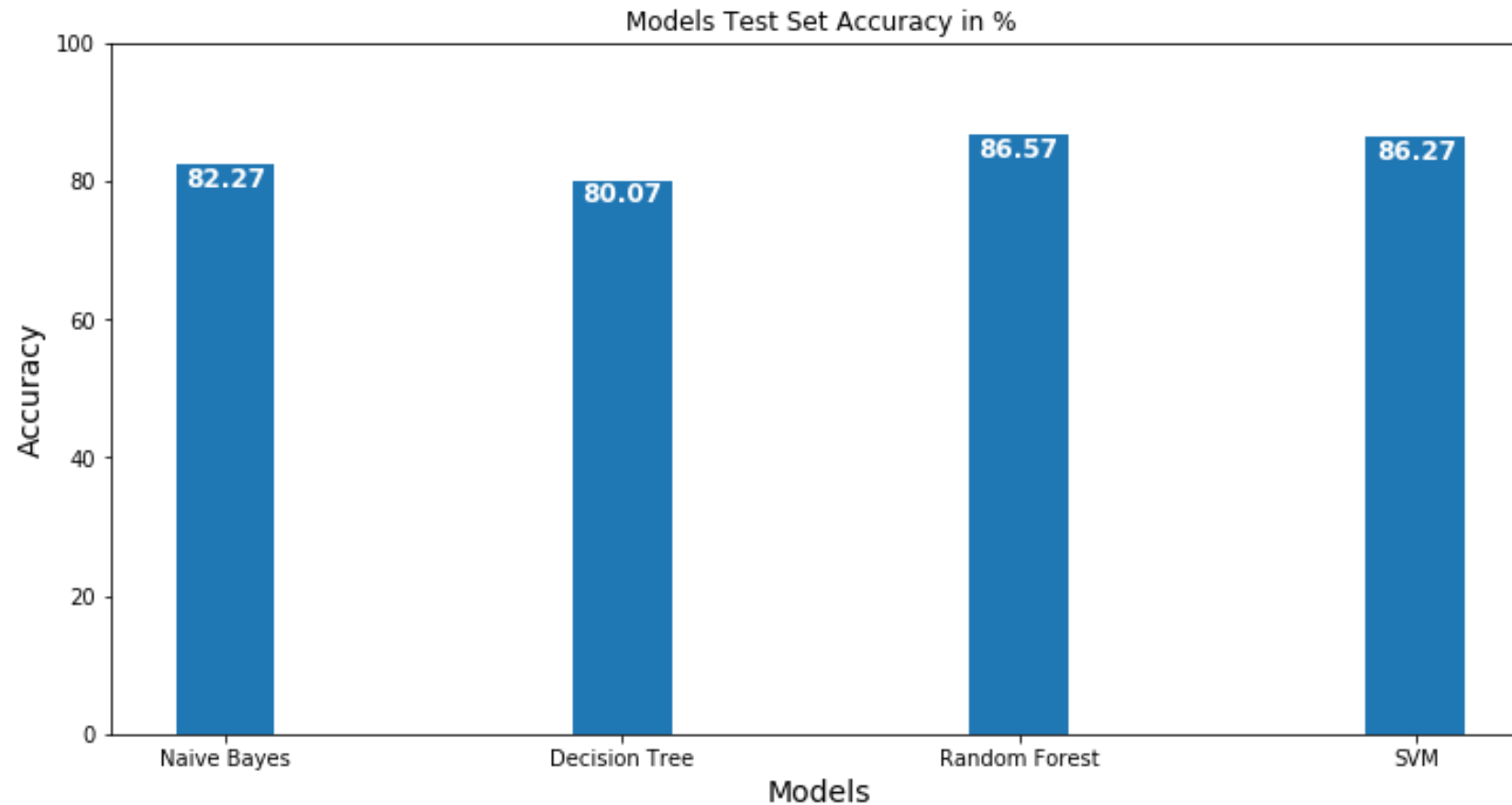
In the SVM algorithm, each data item is plot as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

Then, perform classification by finding the hyper-plane that differentiates the two classes very well. It works really well with a clear margin of separation. It is effective in high dimensional spaces which in our case will also help as we have multiple input attributes.

SVM has different types of kernels like linear, radial basis function (rbf) and poly. I have used radial basis function (rbf) kernel as our problem is non-linear.

Support Vector Machine (SVM)

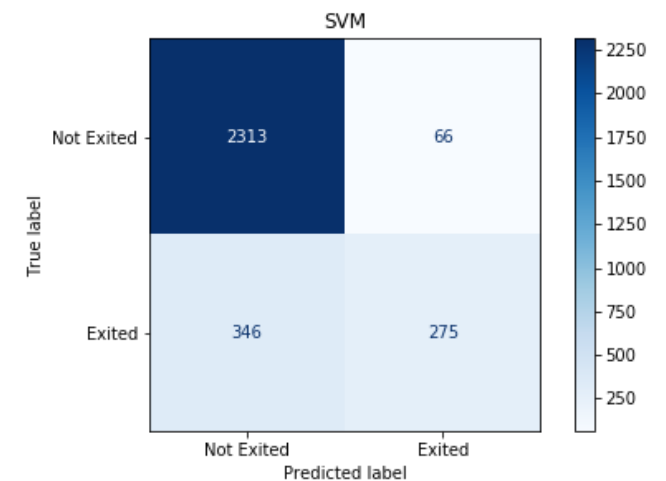
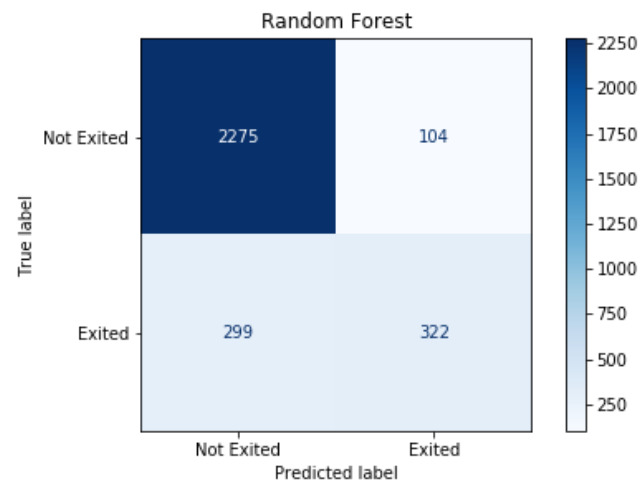
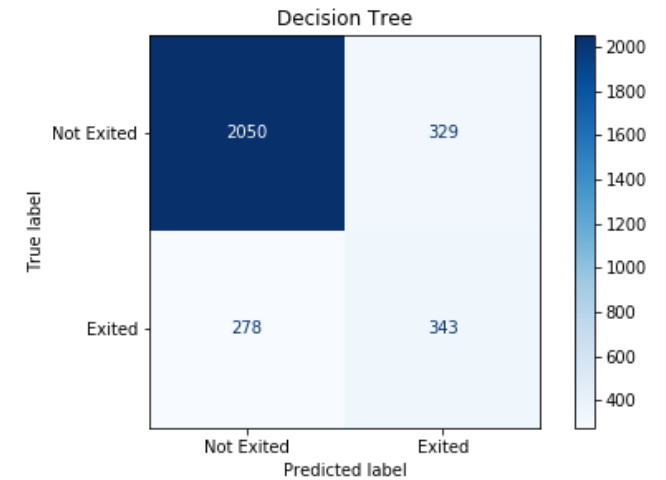
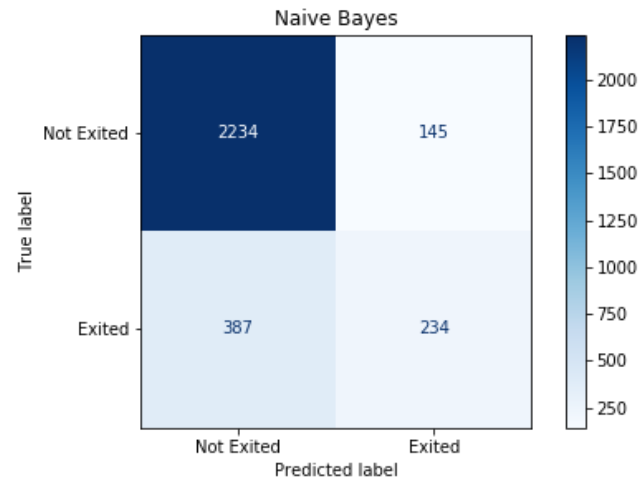
Model Evaluation – Accuracy Score



Model Evaluation – Confusion Matrix

Random Forest is the model which has more correctly classified the 'not exited' and 'exited' labels

Decision Tree is the model which has more correctly classified the 'exited' label than the others whereas SVM is the model which has more correctly classified the 'not exited' label



Thank You

A thin vertical line is positioned to the right of the text "Thank You", extending from the top of the text to the bottom of the text.