*** Project 1 ***

Introduction:-

Being in Santander, our mission is to help people and businesses prosper. We have been constantly working upon customers' data to help them in their financial health, product and services.

In this Project we will deeply analyze the customers' transaction and will predict which customers will make a specific transaction in the future, irrespective of the amount of money transacted by them.

Method:-

As we have the track of previously made transactions of our customer in the last few months, We will use the same to predict the forthcoming transaction through different algorithms in both R and Python.

We have to predict if the transaction will take place or not for a particular person hence this falls under **Classification Model** in **Supervised Learning**.

Firstly, We will bring the data which contains customer information into environment with the help of libraries and will name the file as "data".

To start with, we have cleaned our data with the help of Missing Value Analysis, Feature Scaling and others but the data doesn't **have any missing value** and **neither any value was bias with respect to data.**

Now when the data is fit for use in the algorithms, We will divide it into **Train(80%) & Test(20%)** data so that we can use our algorithms on major part of it to make a model and then will test it to find the accuracy of the model.

In this project I have used 4 algorithms namely:-

- Decision Tree.
- Logistic Regression.
- K-Nearest Neighbor (KNN).
- Naïve Bayes.

After working on each algorithm I have taken the Accuracy and False Negative Rate with the help of Confusion Matrix which is listed below with a reason of Acceptance or Rejection.

R Programming :-

After running the codes in the R environment I have collected the following data of Accuracy and False Negative Rate to choose the best model.

Decision Tree (R)	Perc (%)
Accuracy	89.6269%
False Negative	94.4048%

Logistic Regression (R)	Perc (%)
Accuracy	91.4282%
False Negative	72.8291%

KNN (R)	K=1	K=3	K=5	K=7
Accuracy	82.9874%	85.7365%	87.3256%	88.2761%
False Negative	91.9374%	93.9738%	96.4389%	96.5942%

Naïve Bayes (R)	Perc (%)
Accuracy	92.2493%
False Negative	74.8754%

After going through all the 4 algorithms' result I have chosen **Logistic Regression** in R as minimum False Negative Rate is telecasted by it (app 70%) and it also holds the accuracy above 90% which is very much positive for a model.

Python Programming:-

Now when I have run the provided codes in Jupyter. The following observations were collected.

Decision Tree (Python)	Perc (%)
Accuracy	83.3725%
False Negative	80.7398%

Logistic Regression (Python)	Perc (%)
Accuracy	91.4948%
False Negative	72.6523%

KNN (Python)	K=1	K=3	K=5	K=7
Accuracy	84.7951%	88.8438%	89.6780%	89.8275%
False Negative	91.9702%	98.3890%	99.5559%	99.9008%

Naïve Bayes (Python)	Perc (%)
Accuracy	90.4875%
False Negative	93.0855%

Here also after working on 4 algorithms' the result is pretty much same and I will go with **Logistic Regression** again as False Negative Rate is lowest when compare with other algorithms and it does have a decent Accuracy rate(>90%).

Result:-

The outcome which came out with test data after running the model on it has been listed below.

Test Data	0	1
R	193904	6096
Python	193845	6155

where:-

0 = Transaction will not happen,

1 = Transaction will happen.

Conclusion :-

- In **R** Out of 2 Lakh odd customers provided in the test data, **6096** customer are predicted by our model who are going to make the transaction in the upcoming time
- Whereas in **Python** model **6155** customers are expected to make the transaction in future.

Now considering the Model we have roughly 6000 odd customers who are likely to make the transaction in the upcoming future hence we can provide them with some good plans or we can increase their credit limit(if data is of credit card) so that the customer help us to generate more revenue while making transaction.