Project Description and Requirement

1. PROJECT DESCRIPTION

The goal of this project is to build a deployable predictive model which will help the marketing strategy of a business enterprise in classifying and targeting customers. These classes of customers are those who purchase few expensive products and those customers who purchase cheaper products. The set of these customers will be analyzed, classified, and studied through this project.

This project will be able to deliver the following solutions:

- Predictive model for classifying customers
- Actionable insights into customer behaviors and transaction activities
- Smarter marketing strategies for targeting possible customers
- Feasibility of implementing machine learning in customer classification.

2. TYPE OF PREDICTIVE ALGORITHM

Based on the project description, this project requires a classification machine learning algorithm. Considering that there are two classes of customers to be predicted, this becomes a binary classification problem (Yes/No, True/False, etc.) and the best machine learning algorithm suitable for handling this problem is a classification algorithm.

To determine the particular classification algorithm, that will have the best accuracy based on the performance matrices that will be implemented, the following supervised learning algorithms that will be experimented on are:

- Decision Tree Algorithm
- The K-Nearest Neighbor Algorithm
- The Support Vector Machine Classifier Algorithm
- Random Forest Classifier Algorithm

3. DATA REQUIREMENTS

Building a robust data-driven solution requires certain specific data. This is to ensure that the algorithm is being fed with quality data to efficiently train it for building a reliable predictive model. These data include;

1. **Customer Demographics dataset**: These are those datasets describing the basic information about the customers, for example:
    - Gender
    - Occupation
    - Age
    - Marital status
    - Education level
    - Income level etc.
2. **Customer Geographic dataset:** These include the following;
    - Country
    - State
    - City of residence
    - Specific towns or cities
3. **Customer Behavior Dataset:** These include the following:
    - actions or inactions
    - Spending/consumption habits or rates
    - Average order value etc.
4. **Products and Transaction Dataset:** These are the description of the goods and purchasing rates of customers;
    - Product type
    - Cost of Product
    - Average transaction per product etc.

5. PERFORMANCE MATRICES

Performance matrices are measuring marks used to ascertain the quality or performance of a developed model. However, it varies from one supervised learning technique to the other.

For this project, we will be making use of the following matrices:

1. **The classification accuracy score:** It is the ratio of the number of correct predictions to the total number of input samples. It shows the number of values that were predicted correctly when given validation values.
2. **Confusion Matrix:** This gives a matrix as output and describes the complete performance of the model. This matrix will compare the number of YES predictions that were meant to be NO and the number of NO predictions that were meant to be YES.
3. **AUC Curve:** This is used for the binary classification problem, like in this project. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.
4. **F1, Recall and Precision:** F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances) [1]. The **Precision** and **Recall** are matrices used in comparison with the **F1** result when positive and negative results are to be evaluated respectively.

Reference

[1] https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[2] https://www.analyticsvidhya.com/blog/2021/06/how-to-solve-customer-segmentation-problem-with-machine-learning/

[3] https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html

[4] https://www.altexsoft.com/blog/datascience/preparing-your-dataset-for-machine-learning-8-basic-techniques-that-make-your-data-better/