MACHINE LEARNING ENGINEER NANODEGREE
CAPSTONE PROPOSAL

BY WILFREDO VILLALOBOS

October 2021

Domain Background

Currently, Data Science is a necessary tool in almost every enterprise in the world. The extent in which data is being analyzed in order to achieve different outputs is huge. As a result, it is important for the leaders in the organizations to know the different problems data could solve, and get the job done. In this project, it is expected to run a model for Starbucks, one of the most famous branches in the world, to make an analysis of which offers could be better to be sent to customers to make them acquire their products, based on simulated data. Personally, I expect to learn more about how to predict the best output for distinct groups of customers.

Problem Statement

Starbucks needs to develop a model to identify clusters of customers depending on their personal information and on their historic activities. The main goal is to define if a customer has to receive an offer or not, and, if the customer has to, to which type of offer they are more responsive.

Datasets and Inputs

Starbucks and Udacity provide the main sources in order to execute the project successfully. The information is simulated based on what Starbucks actually get when working on their Rewards Program. The sources of information are the three following datasets:

1. profile.json – This dataset contains the users of rewards programs for Starbucks. It contains information like the age, gender, income, and date of becoming a member of the program.
2. portfolio.json – This dataset contains the data of offers sent during a 30-day test period. It contains information of rewards, channels, difficulty, duration, and offer type.
3. transcript.json – This dataset contains the information of the events. In this case, the event is referred to what happened to the offers. The offers could be received, viewed, a transaction could be made, and the offer could be completed.

Solution Statement

The solution proposed is to apply different Machine Learning techniques to prove which one helps the model to be more effective in terms of the evaluation metrics that will be applied for this project. The different techniques will be evaluated and just a couple of them will be executed and tested to prove which one is better.

More specifically, distinct techniques will be evaluated to complete the main goal of the project. Those techniques will be Regressions, Random Forests, Neural Networks, etc. These techniques will not necessarily be the final technique implemented, as in the practice it could happen that another model turns more useful for achieving our purpose.

Benchmark Model

As mentioned before, the benchmark models could be:

- Linear Regression.
- Logistic Regression.
- Random Forest Classifiers.
- RNN (Recurrent Neural Network).
- Any other model that could be useful for our purpose.

Note: Listing all these models does not mean that they will all be executed, but they will all be evaluated in order to just execute the more adapted to our goals.

Evaluation Metrics

In order to achieve this project, we will need to measure if it is going well. For that, we will need some metrics:

- Accuracy – It will be used to measure the performance of our models.
- Recall – It will be used to measure how much of the positive outputs were correctly classified.
- Any other metric that could be useful to achieve our goals.

Project Design

The design of the project will be the following:

1. Loading and exploring the data
2. Cleaning the data, and pre-processing the data
3. Merging the clean data
4. Splitting the data into train, test, and validation data
5. Training the model/models chosen
6. Evaluating and comparing between models
7. Analyzing the results
8. Presenting the results, conclusions, and recommendations