STARBUCKS CAPSTONE PROPOSAL

Introduction

This project aims to fulfill the requirement to finalize the nanodegree Machine Learning Engineer at Udacity and also to build a decision making tool based on Machine learning algorithms fed with data containing offers, customer demographics and transactions provided by Starbucks.

DOMAIN BACKGROUND

Starbucks is a world reference in the coffee business that is known by almost everyone. They have created a data set that simulates customer activity in their app. This information should help Starbucks to optimize the interaction with its customers, optimize marketing campaigns and increase sales.

I consider this project to be very important because it proposes finding data-based solutions to some questions that many retail businesses are interested in solving, such as: How do I get my offers to the right destination? Can I predict if an offer will become a purchase?

PROBLEM STATEMENT.

The objective of this project is to create a model, based on demographic, offers and transactional information that allows Starbucks to predict whether or not a customer belonging to a demographic group will complete or not a received offer.

DATASETS

The data set has been provided by Starbucks and contains three data domains:

- Descriptive information on the offers created and sent to customers
  file: **portfolio.json**
  information details: Basically offer metadata. Include type of offer and also reward, duration and difficulty.

- Customer demographics. Information of all customers that use the app
  file: **profile.json**
  information details: age, income, gender, date when customer created an app account.

- Transactional information on customer activity. This includes offers received, offers viewed and offers completed
 file: **transcript.json**
information details: Event type (offer received, offer viewed, etc) , customer id, amount and time in hours since start of test

Since the objective is to create a binary classification model that allows Starbucks to predict whether or not a customer belonging to a demographic group will complete or not a received offer , it is important to mention that for this type of model to work well, there must be a balance in the number of positive and negative results (in this case, offer completed or not). In this case an exploratory analysis of the data shows that there is a reasonable balance in the number of transactional data for completed and uncompleted offers.

SOLUTION

Whereas this is a classification problem, several models like K-nearest neighbors, decision trees and others will be created. The accuracy of the models will be evaluated by applying historical test data.

BENCHMARK

The results of the three algorithms will be compared. Applying Hyperparameter tuning during the refinement of the models. The model with the best accuracy will be chosen.

EVALUATION METRICS

The accuracy of the model will be evaluated. Accuracy is a good metric for binary classification models, as this one, where we are predicting whether the offer will be completed or not. Accuracy is a good metric when the data is balanced. A review of the data shows that there is a reasonable balance in the number of transactional data for completed and uncompleted offers, especially for men.

PROJECT DESIGN

The project will be implemented in the following phases

Data exploration: The idea is to understand and become familiar with the data set provided. Generate graphs and determine the correlation of characteristics.

Data preparation and cleaning: Identify and create new data sets suitable for the models, based on the existing ones. The data will also be cleaned

Creation and training of the models:
The models will be created and adjusted with the best hyperparameters.
You will train the models using the data sets.

Benchmarking
The results of all the models are compared and the best one is chosen.