# Starbucks Capstone Project Proposal Abhay Mahajan | July 28, 2020 Udacity – Machine Learning Engineer Nanodegree

# Background:

Starbucks is a passionate purveyor of coffee and other beverages, headquartered in Seattle, Washington. The corporation is ranked  $121_{st}$  in the list of 2019 Fortune 500 companies. They have a mobile application where registered users can use it to order coffee for pickup while mobile, pay in-store directly using the app, and collect rewards points. This app also offers promotions for bonus points to these users. The promotional offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). This project is focused on tailoring the promotional offers for customers based on their responses to the previous offers and find out which of them are most likely to respond to an offer.

#### **Problem Statement:**

The goal that I have to achieve here is to best determine which kind of offer to send to each user based on their response to the previously sent offers. Not all users receive the same offer, and that is the challenge to solve using the data set that is provided by Starbucks, which was captured over 30 days. I'll also build a machine learning model that will predict the response of a customer to an offer.

# Datasets and Inputs:

This data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. The data set is provided in form of three JSON files:

- portfolio.json containing offer ids and meta data about each offer (duration, type, etc.)
- profile.json demographic data for each customer
- transcript.json records for transactions, offers received, offers viewed, and offers completed

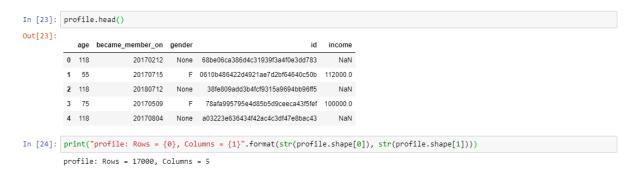
Here is the schema and explanation of each variable in the files: portfolio.json

```
id (string) - offer id
offer_type (string) - type of offer ie BOGO, discount, informational
difficulty (int) - minimum required spend to complete an offer
reward (int) - reward given for completing an offer
duration (int) - time for offer to be open, in days
Channels (list of strings)
```

```
In [8]: portfolio.head(10)
Out[8]:
                       channels difficulty duration
                                                                         id offer_type reward
        1 [web. email. mobile. social]
                                            5 4d5c57ea9a6940dd891ad53e9dbe8da0
        2 [web, email, mobile] 0 4 3f207df678b143eea3cee63160fa8bed informational
                                            7 9b98b8c7a33c4b65b9aebfe6a799e6d9
               [web, email, mobile]
                   [web, email] 20 10 0b1e1539f2cc45b7b9fa7c272da2e1d7 discount 5
         5 [web, email, mobile, social]
                                          7 2298d6c36e964ae4a3e7e9706d1fb8c2 discount
          6 \quad \text{[web, email, mobile, social]} \qquad \qquad 10 \qquad \qquad 10 \qquad \qquad \text{fafdcd668e3743c1bb461111dcafc2a4} \qquad \qquad \text{discount} \qquad \qquad 2
              [email, mobile, social] 0 3 5a8bc65990b245e5a138643cd4eb9837 informational
         8 [web, email, mobile, social] 5 5 f19421c1d4aa40978ebb69ca19b0e20d bogo 5
                [web, email, mobile] 10 7 2906b810c7d4411798c6938adc9daaa5 discount
In [9]: print("portfolio: Rows = {0}, Columns = {1}".format(str(portfolio.shape[0]), str(portfolio.shape[1])))
        portfolio: Rows = 10, Columns = 6
```

#### profile.json

- age (int) age of the customer
- became\_member\_on (int) date when customer created an app account
- $\bullet$  gender (str) gender of the customer (note some entries contain '0' for other rather than M or F)
- id (str) customer id
- income (float) customer's income



#### transcript.json

- event (str) record description (ie transaction, offer received, offer viewed, etc.)
- person (str) customer id
- time (int) time in hours since start of test. The data begins at time t=0
- value (dict of strings) either an offer id or transaction amount depending on the record



The portfolio.json contains offer\_type column, which describes the types of offers that Starbucks is looking to potentially send its customers:

- 1) BOGO (Buy-One-Get-One): This offer enables a customer to receive an extra and equal product at no additional cost. The customer must spend a certain threshold in order to make this reward available.
- 2) Informational: This offer doesn't necessarily include a reward, but rather an opportunity for a customer to purchase a certain object given a requisite amount of money.
- 3) Discount: With this offer, a customer is given a reward that knocks a certain percentage off the original cost of the product they're choosing to purchase, subject to limitations.

## **Solution Statement**

To find out which offers are to be sent to the customers, I'll find out the offers that interests them the most, and consider Exploratory Data Analysis to cover a few points like:

- 1) most responded offer
- 2) response to an offer
- 3) age & gender groups which are greatly interested in offers

These points will be discussed for the combined population, and for the individual personalized level as well.

To find out the appropriate response of a customer to an offer, I'll be leveraging models like RandomForestClassifier and DecisionTreeClassifier, to determine which model best represents our data on hand.

## Benchmark Model

A quick and fairly accurate model can be considered as a benchmark. I will use the KNeighborsClassifier to build the benchmark, as it is a fast and standard method for binary classification machine learning problems and evaluate the model result using F1 score as the evaluation metric.

### Evaluation Metrics

I will consider the F1 score as the model metric to assess the quality of the approach and determine which model gives the best results. It can be interpreted as the weighted average of the precision and recall. The traditional or balanced F-score (F1 score) is the harmonic mean of precision and recall, where an F1 score reaches its best value at 1 and worst at 0.

# Project Design

Here is the general flow for how I will be conducting this project:

- 1) Establishing the workspace in a Jupyter environment
- 2) Cleaning up the data as needed for the modeling purposes.
- 3) Performing a deep-dive exploratory analysis on the data
- 4) Building different models to determine the most appropriate one for the data
- 5) Using the evaluation metrics for predicting the accuracy and the viability of the model.