Udacity Capstone Project Proposal

UDACITY

STARBUCKS®

Starbucks Capstone Project

Presented by: Syed Muhammad Haider Jafri

1. Domain Background

Starbucks is a coffee company based out of Seattle, WA, with a global presence in over 75 countries and operates over 27,000 stores worldwide.

Since Starbucks does not rely on the franchise model for expansion, it needs to innovate at very high rates to ensure, maintain and increase their profit margins in the cut-throat business of coffee. Starbucks wants to be your “third” place - not just your coffee supplier. (The concept of third place is that there is your home, office, and the “third” place where you could hang out with your family and/or friends. “Third place” used to be the shopping malls in the US, but as they are in decline, Starbucks wants to fill the vacuum and become your “third place”).

Starbucks, therefore, must upsell to existing and loyal customers, and create attractive offerings to non-frequent customers at the same time to increase profit margins.

2. Problem Statement

The problem we have to solve is that given a dataset, we have to craft product offerings to our customers, both loyal and otherwise, to increase their spending in the Starbucks ecosystem. No customer receives more than 1 offer at any given time.

We will assume the offerings that we were able to craft are;

A. BOGO (buy one, get one free).
B. Discount code.
C. Simple advertisement.
D. No offer for the week.

3. Datasets and inputs

We have 3 datasets in JSON for this project, provided by Starbucks.

A. Profile.json

Rewards program users (17000 users x 5 fields)

I. gender: (cateforical) M,F,O or null
II. age: (numeric) missong value encoded as 118
III. id: (string/hash)
IV. became_member_on:(date) format YYYYMMDD
V. income: (numeric)

B. Portfolio.json

Offers sent during the 30-day test period (10 offers x 6 fields)

a. Rewards: (numeric) money awarded for the amount spent
b. Channels: (list) web, email, mobile, social
c. Difficulty: (numeric) money required to be spent to receive rewards
d. Duration: (numeric) time for the offer to be open, in days
e. Offer_type: (string) BOGO, discount, informational
f. id: (string/hash)

C. Transcript.json

Event log (306648 events x 4 fields)

a. Person: (string/hash)
b. Event: (string) offer received, offer viewed, transaction, offer completed
c. Value: (dictionary) different values depending on event type
    i. Offer id: (string/hash) not associated with andy “transaction”
    ii. Amount: (numeric) money spent in “transaction”
    iii. Id: (string/hash)

4. Solution Statement

Our solution statement is simple. Since we do not have labels for our data, we will primarily rely on clustering analysis of our dataset using unsupervised learning techniques. Unsupervised learning methods will help us in crafting our strategy better and will give us enough leverage to come up with multiple solutions.

5. Benchmark Model

Since our dataset is not labeled, we do not have an objective benchmark model to compare our solutions to. We will, however, determine the performance of our model relative to the number of clusters (i.e. the size of k) to be selected, vs the output of the “elbow method ” k-means clustering. If at k=4 of below, the sum of squared errors (SSE) keeps increasing instead of decreasing, then it is a bad sign and we have to further tune our model.
If the error keeps decreasing, then we can select k=4 but can go up to the value of k after which SSE starts to increase. This value of k will give us the maximum number of clusters we can get from our data, which translates to the number of possible offers we can generate.

As mentioned above, we have to cluster our dataset into 4 segments (3 with offers, and 1 without any offer).

6. Evaluation metrics

We will use the “silhouette coefficient” method along with the “elbow method” to find the ideal number of clusters for our project. The elbow method will give us a visual representation of the sum of squared errors, which will help us in selecting the optimal number of clusters for our problem.

7. Project Design

Since we are using unsupervised learning techniques, we will start from one of the most popular techniques of the domain, the K-Means clustering method. We have to group all customers into 3 groups, so we will start with k=3. These groups would then be provided with their custom Starbucks offers.

As we have done in the labs already, we will use the “elbow method” to evaluate our project.

According to the book “Hands-on machine learning with Scikit-learn, Keras and Tensorflow”, there are several other unsupervised machine learning algorithms like DBSCAN and Hierarchical Cluster Analysis (HCA). We will use those models as well and figure out which ones work the best for our problem.