Name: Andrew Kalil
Subject: Capstone Project Proposal

Starbucks Capstone Challenge

Project's domain background

Due to the recent advancements in technology, the world is causing us to adapt extremely fast to the changes. It is thanks to the technology that I am able to write this proposal and email it to a mentor to be revised. With this being said, it is no secret that large enterprises and multi millionaire companies use this amazing invention called “the internet” to reach a larger audience and grow their business.

Starbucks is one of these many businesses that take advantage of the internet to obtain new customers daily. The following are some of the ways that Starbucks uses technology to reach new customers.

1) Free WiFi: What a better way to attract buyers and tempt them to buy your product than giving them free access to WiFi almost any time of the day?
2) Mobile or online purchase: Studies show that 11% of starbucks sales come from mobile purchases.
3) Discounts and Offers: Like any other store, Starbucks sends their customers offers which can be very tempting but rewarding. This therefore helps to create loyal customers and increase sales.

Therefore, Starbucks is one of the first companies to take advantage of technology as a marketing strategy. “[…] In many technology-related industries, competition is intense and can often be the reason why a startup is not
able to be profitable. However, competition cuts across the board and can contribute to potential business failure, regardless of the sector."[1] With that, half of small businesses survive beyond five years. Additionally, one third of these businesses make it to 10 years.

With that being said, marketing is crucial for understanding how to meet customer’s needs. This of course keeping in mind that customers vary in characteristics like flavor preference, behaviors, traditions, economic situation, etc. This is where artificial intelligence and machine learning comes in. With the recent developments in the area, machine learning has been able to use marketing information to detect behavior patterns in customers and segmenting groups of people based on demographic information.

For this reason, this project will study the behaviors and reactions of different Starbucks customers. We will determine which customers are actually influenced by offers and which are not. This also tells us which customers are considered “loyal”. Finally, the end goal is to predict how a demographic group will react to an offer.

Problem statement:

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. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). Some users might not receive any offer during certain weeks.

Not all users receive the same offer, and that is the challenge to solve with this data set.

Keep in mind as well that someone using the app might make a purchase through the app without having received an offer or seen an offer.

The input for the experiment will be the starbucks datasets which are thoroughly described in the next section. It contains important information like users gender, incomes, age, offers they received, offers they completed, etc.

The estimated output for the experiment is to determine whether or not a customer will buy an offer based on tendencies and behaviors. For this, a thorough data study is required.

The machine learning task responsible for this project is K-means and its variations. The idea is to use different models and algorithms and see which one best performs the task to obtain best results.

Datasets and inputs:

The datasets used for this project are the following:
profile.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)

Portfolio.json

- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer (note some entries contain 'O' 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

Proposed Solution

The proposed solution to the problem is determining the most appropriate unsupervised machine learning model by running a test to see which one outputs the best score. The intention is to try K-means, MiniBatch K-means, Hierarchical Clustering, DBSCAN, and other related models.

Once this is done, the idea is to create a report or blog post explaining the step by step procedures that lead to the results. The intention is to try to complete all of the proposed tasks in this document. Also, it is important to make the blog very understandable as if it was being used to explain it to someone who has little knowledge of the topic.

Benchmark model

The benchmark model that I will utilize for this project is a very similar project with the same task but a slightly different approach than mine. I will be comparing my project against this one from a previous Udacity student:

https://github.com/dkhundley/starbucks-ml-capstone/blob/master/Hundley-Starbucks-Project.ipynb

However, like I previously mentioned, I will use a slightly different approach and compare against the benchmark results. The budget algorithms which will be used in comparison to the benchmark model are:

1. K-means
2. MiniBatch K-Means

3. Hierarchical Clustering
4. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
5. Gaussian Mixture Modelling (GMM)
6. MeanShift

Evaluation metrics

The planned metrics for this project are:

1. Elbow method: For determining the optimal number k-means clusters by plotting the value of the cost function produced by k values.
2. Silhouette value: Measures similarity between a point and its own cluster (cohesion) compared to other clusters (separation).
3. Davies Bouldin metric: Average similarity measure of each cluster to its most similar cluster. In this case similarity is the ratio of distances within the cluster to distances between clusters. The minimum score is zero, and lower values indicate better clustering.

Project design

The project for the problem is as following:

1. Establish a workspace in a jupyter environment
2. Download the data provided into my jupyter notebook
3. Initial data cleaning
4. Perform exploratory analysis on the data
5. Cleaning up the data as needed for modeling
6. Experimenting to determine most appropriate unsupervised learning model to use for
   the data, whether that be k-means clustering, DBSCAN, or some other model
7. Leveraging our benchmark model and evaluation metric(s) to ensure sanity
8. Summarizing our findings writing a blog post

References

[1] Otar, C., 2018. What Percentage Of Small Businesses Fail -- And How Can You Avoid Being One Of Them?. [online] Forbes. Available at: <https://www.forbes.com/sites/forbesfinancecouncil/2018/10/25/what-percentage-of-smallbusinesses-fail-and-how-can-you-avoid-being-one-of-them/?sh=2cf9ecba43b5>