

Starbucks Capstone Project Proposal

Deepthi M | January 21, 2022

Udacity – AWS Machine Learning Engineer Nanodegree Program

Background

Starbucks founded by *Jerry Baldwin, Zev Siegl, and Gordon Bowker* in 1971 is now the largest coffeehouse chain in the world. Through their constant improvisation and attention to detail, Starbucks have increased their customers immensely. Starbucks is further enhancing its customer experience with the introduction of a new Starbucks Rewards program.

Through this program, there are three types of offers that can be sent: buy-one-get-one (BOGO), discount, and informational. In a BOGO offer, a user needs to spend a certain amount to get a reward equal to that threshold amount. In a discount, a user gains a reward equal to a fraction of the amount spent. In an informational offer, there is no reward, but neither is there a requisite amount that the user is expected to spend. Offers can be delivered via multiple channels.

With this information, it would be useful to perform Customer Segmentation. Customer targeting and segmentation is a popular tactic to re-engage old customers, invite new ones and drive sales. Customer Segmentation is the process of dividing the customers into groups or clusters based on certain similarities they have so that we can market or sell to each of these groups properly.

It was in the 1950s that, customer segmentation became a formal part of modern system. The definition of customer segmentation as given by *Smith* in 1956 is “Market segmentation involves viewing a heterogeneous market as a number of smaller homogeneous markets, in response to differing preferences, attributable to the desires of customers for more precise satisfactions of their varying wants”. *Wedel and Kamakura* also stated “Market Segmentation is an essential element of marketing in industrialized countries. Goods can no longer be produced and sold without considering customer needs and recognizing the heterogeneity of those needs” in their book [*‘Market Segmentation’*](#).

Therefore, customer segmentation mainly depends on three assumptions:

- The group of customers is heterogeneous
- But these heterogeneous groupings should have characteristics that can be recognized and examined.
- Exclusive promotions to fulfil the demands of customers

Problem Statement

A data set contains simulated data that mimics customer behaviour on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of its 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.

The problem statement is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type.

Since this problem is based on customer segmentation, it is a clustering problem. Cluster analysis is using a mathematical and statistical model to detect populations of similar customers. Since this is a clustering problem (unsupervised learning), there is no target variable but instead looks at the relationships between the input variables. Like, the relationships between the offers and demographics which respond to them.

Dataset

profile.json

Rewards program users (17000 users x 5 fields)

- gender: (categorical) M, F, O, or null
- age: (numeric) missing value encoded as 118
- id: (string/hash)
- became_member_on: (date) format YYYYMMDD
- income: (numeric)

```
profile.head()
```

	age	became_member_on	gender	id	income
0	118	20170212	None	68be06ca386d4c31939f3a4f0e3dd783	NaN
1	55	20170715	F	0610b486422d4921ae7d2bf64640c50b	112000.0
2	118	20180712	None	38fe809add3b4fcf9315a9694bb96ff5	NaN
3	75	20170509	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0
4	118	20170804	None	a03223e636434f42ac4c3df47e8bac43	NaN

First five rows of the profile file

portfolio.json

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

- reward: (numeric) money awarded for the amount spent
- channels: (list) web, email, mobile, social
- difficulty: (numeric) money required to be spent to receive reward
- duration: (numeric) time for offer to be open, in days
- offer_type: (string) bogo, discount, informational
- id: (string/hash)

```
portfolio.head()
```

	channels	difficulty	duration	id	offer_type	reward
0	[email, mobile, social]	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10
1	[web, email, mobile, social]	10	5	4d5c57ea9a6940dd891ad53e9dbe8da0	bogo	10
2	[web, email, mobile]	0	4	3f207df678b143eea3cee63160fa8bed	informational	0
3	[web, email, mobile]	5	7	9b98b8c7a33c4b65b9aebfe6a799e6d9	bogo	5
4	[web, email]	20	10	0b1e1539f2cc45b7b9fa7c272da2e1d7	discount	5

First five rows of the portfolio file

transcript.json

Event log (306648 events x 4 fields)

- person: (string/hash)
- event: (string) offer received, offer viewed, transaction, offer completed
- value: (dictionary) different values depending on event type
 - offer id: (string/hash) not associated with any "transaction"
 - amount: (numeric) money spent in "transaction"
 - reward: (numeric) money gained from "offer completed"
- time: (numeric) hours after start of test

```
transcript.head()
```

	event	person	time	value
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}
1	offer received	a03223e636434f42ac4c3df47e8bac43	0	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}
2	offer received	e2127556f4f64592b11af22de27a7932	0	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}
3	offer received	8ec6ce2a7e7949b1bf142def7d0e0586	0	{'offer id': 'fafdc668e3743c1bb461111dcafc2a4'}
4	offer received	68617ca6246f4fbc85e91a2a49552598	0	{'offer id': '4d5c57ea9a6940dd891ad53e9dbe8da0'}

First five rows of the transcript file

Solution

The first step would be to perform Exploratory Data Analysis (EDA) and describe the dataset. This EDA would expose various trends, patterns and relationships that are not visibly apparent. It will also help to identify obvious errors as well as detect possible outliers.

The second step would be data visualization to understand the dataset better and get a hold of the quantity of each demographic, each type of offer or event.

To identify which groups of people are most responsive to each type of offer, and how best to present each type of offer, various models including the benchmark model are analyzed to determine which model helps us to solve the problem efficiently.

Apart from the benchmark model, the other model used to evaluate, will be the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) model. This will give us a better result as it will be able to work well with outliers and noise.

Benchmark Model

A baseline model provides a point of comparison for the more advanced models that will be evaluated later on. There are few requirements for a good baseline model:

- Baseline model should be simple.
- Simple models are less likely to overfit
- Baseline model should be interpretable.

The baseline model would be a K-Means model. K-Means is one of the widely used model for clustering-based problems. This looks for a pre-defined number of clusters with an unlabeled multi-dimensional dataset.

Evaluation Metrics

Since this is a clustering problem (unsupervised algorithm) the metrics will help us to evaluate our model without the help of any labelled data. The metrics I will be using are the *Silhouette coefficient*, *Elbow Method* and the *Dunn's Index*. These metrics allow us to evaluate the models based on similarity or dissimilarity. The SSE (Sum of Squared Errors) will also be evaluated for each model. All these together will help us finalize the number of clusters in our ultimate model.

Project Flow

1. The entire Jupyter notebook will be run using an Amazon Sagemaker instance.
 - a. Appropriate instance will be chosen to ensure the notebook is run effectively.
 - b. Obtain dataset and domain knowledge
2. The input files are uploaded to Amazon S3.
 - a. A S3 bucket is created to hold the input dataset.
 - b. The input data is uploaded to S3 via the Jupyter notebook.
3. Exploratory Data Analysis is performed on the dataset.
 - a. The 3 data files (profile, portfolio and transcript) are described.
 - b. Their columns and rows are displayed
 - c. Measures of central tendency are calculated
 - d. Confirming the data types
4. The dataset is visualized using various graphs and charts.
 - a. Matplotlib, pyplot are used to plot and visualize the input data.
 - b. Various demographic data are plotted.
5. The dataset is then cleaned.
 - a. Any null or empty values are filtered out.
 - b. These are replaced with the mean values or the column is dropped.
6. Various models are trained including one baseline or benchmark model.
 - a. The baseline model is trained (KMeans)
 - b. The DBSCAN model is trained.
7. Evaluating the models and then choosing the best one.
 - a. Both the models are evaluated and the findings are analyzed
8. Publish project work in a detailed blog post.