

Background Knowledge and Problem Statement:

The capstone project represents an experiment in Starbucks that was carried out for a period of one month. The insight that we are after is: *what next?* More specifically, we want to have a proper **customer segmentation**. Certain kinds of people will represent to certain offers in different ways.

Some of the questions that are of interest to us include:

- What are the various types of customers in the data?
- How attractive is each individual offer, and to which type of customers?
- Is there an offer that we are better off without? Hopefully not.

In fact, Peter Fader, in his book, **Customer Centricity: Focus on the Right Customers for Strategic Advantage**, distinguishes between two approaches with respect to customers, those that are customer-friendly, and those that are customer-centric. The decided factor for customer centric companies is their conscious effort to try to take advantage of the deeply ingrained buying habits that is established over the years. Doing so not only is directly linked to the bottom-line, but also it will create more loyal customers in the long-run, which creates a positive feedback loop, leading to cementing the company's strategic positioning in the market.

Dataset:

An important feature of this dataset is that, this is **a simulated data for the sake of testing the algorithms**, and not a simulated data for the purpose of mimicking real people/customers. What we want to know is to understand the cause and effect and to find a way to frame and think about the problem so that it is consistent with what the data is telling.

The dataset is composed of 3 different sources:

1. A **portfolio** of various offers and their characteristics, including their duration, the channel that offer was promoted etc.
2. A **profile** of the customers: their age, income, their tenure etc.
3. And the **transcript** data which includes about 300K events: offer received, viewed, transaction, and completed.

There are three different offers – informational, discount, and buy-one-get-one – that could be offered via various channels.

Solution Statement:

This is an **unsupervised machine learning** task. Here we are looking for groups of customers that are as similar as possible within each group, but different – in terms of responding to the promotions – when compared to other groups.

Benchmark Model and Evaluation Metrics:

A naïve model would be to simply use the profile data – information of customers – to assign customers to various clusters. A hyper-parameter of interest would be the number of clusters/groups.

Project Design:

As is typical for any data science and machine learning problem, I'll start by digging deep into the various sources of data at hand. Performing exploratory data analysis should hopefully shed some lights on the problem and new avenues of interest perhaps reveal themselves.

Aside from the profile data, the transcript data seems to be a promising source of information for feature engineering; depending on the number of features generated perhaps PCA could also be an interesting algorithm to enhance customer segmentation.

References:

Fader P. *Customer Centricity: Focus o the Right Customers for Strategic Advantage*. 2012, Wharton School Press