Starbucks Project Capstone Proposal

David Hundley | July 2019
Udacity - ML Engineer Nanodegree

Introduction

As I write this, I am sipping on a cup of Starbucks coffee. I visit Starbucks perhaps two to three times per week. A daddy to two little girls under two years old, I often leverage one of our local Starbucks locations as a getaway means to study or focus on projects… like this one! I would normally love to choose my own project that would be relevant to my machine learning work at State Farm, but due to the proprietary nature of that work, I'm unable to do so. Instead, we'll be leveraging Udacity's suggested project here, which I'm looking forward to as an avid customer of Starbucks myself!

Domain Background

A desire to glean insights about customer behavior is one of the hottest uses of machine learning today, and rightfully so. It is often difficult to understand necessarily how certain behaviors influence others, so leveraging things like unsupervised predictive learning models go a long way to better understand customer behavior.

An employee of State Farm, I have had experience working directly with customers in things like focus groups, if not utilizing predictive modeling. (At least, not yet.) Fortunately, I am also a student of Udacity's Data Science nanodegree program and recently completed the Arvato project on determining customer segments. Coincidentally, that was another offering I could have selected for this particular capstone, but I wanted to focus my sights on something else entirely. Still, the work from that project will serve me well as I seek to pivot the unsupervised learning knowledge applied there to this particular project. (Link to Hundley - Arvato Customer Segments GitHub)

Problem Statement

The problem we are looking to solve here is conceptually easy to understand, albeit difficult to answer. **We are looking to best determine which kind of offer to send to each customer segment based on their purchasing decisions.** We'll touch more on what these offers are and data we'll be utilizing down in the next section. We will leverage traditional evaluation metrics to determine which model is most appropriate for our dataset. These evaluation metrics will be discussed in an upcoming section.

Datasets and Inputs

For this project, we will be leveraging the data graciously provided to us by Udacity. This is given to us in the form of three JSON files. Before delving into those individual files, let us first understand the three types of offers that Starbucks is looking to potentially send its customers:

- Buy-One-Get-One (BOGO): In this particular offer, a customer is given a reward that enables them to receive an extra, equal product at no cost. The customer must spend a certain threshold in order to make this reward available.
- Discount: With this offer, a customer is given a reward that knocks a certain percentage off the original cost of the product they are choosing to purchase, subject to limitations.
- Informational: With this final offer, there isn't necessarily a reward but rather an opportunity for a customer to purchase a certain object given a requisite amount of money. (This might be something like letting customers know that Pumpkin Spice Latte is coming available again toward the beginning of autumn.)

With that understanding established, let's now look at the three provided JSON files and their respective elements:

1. profile.json

This file contains dummy information about Rewards program users. This will serve as the basis for basic customer information.
(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)

2. portfolio.json

This file contains offers sent during a 30-day test period. This will serve as the basis to understand our customers' purchasing patterns.
(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)

3. transcript.json

This file contains event log information. Complementing the file above, this file will serve as a more granular look into customer behavior.
(306648 events x 4 fields)

- **person**: (string / hash)
- **event**: (string) offer received, offer viewed, transaction, offer completed
- **value**: (dictionary) different values depending on the 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

Solution Statement

Given that we do not have any labels or ground truth that would enable us to leverage supervised learning models, **we will be leveraging unsupervised learning methods amongst our data to determine appropriate offers to share based on prior customer purchasing patterns**. Specifically, we will be leveraging models like k-means clustering, DBSCAN, and more to determine which model best represents our data on hand.

Benchmark Model

For this project, I will be leveraging insights from a similar project I did as part of the Udacity Data Scientist nanodegree program. That project was the Bertelsmann-Arvato customer segmentation project. (Link to Hundley - Arvato Customer Segments GitHub) Given that that project was also very focused on customer segmentation and already reviewed by Udacity, I believe the insights gleaned there will serve well as a benchmark for this project.

Evaluation Metrics

Given that we will be leveraging unsupervised clustering models for our project, we will be using some metrics that enable us to validate our clusters without having labelled data. Namely, we will be leveraging the **silhouette coefficient**. Because we don't have labelled data, the silhouette coefficient is appropriate since it produces a score between the range of -1 and 1 based on internal indices. It also happens to be easy to calculate with help from sci-kit learn. (I'm less familiar with it today, but I may also explore leveraging the Calinski-Harabasz score as well.)

Project Design

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

1. Establishing the workspace in a Jupyter environment
2. Migrating the data from Udacity's provided environment into my local Jupyter environment
3. Initial cleansing of the data
4. Performing a deep-dive exploratory analysis on the data
5. Cleaning up the data as needed for modeling purposes
6. Conducting experiments to determine most appropriate unsupervised learning model 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 and work in a detailed blog post