Todd Cunningham

01 January 2020

Machine Learning Capstone — Proposal

Domain Background

Starbucks, a retail coffee establishment, currently has a loyalty app which allows users to track their purchases and receive occasional rewards after meeting various personal milestones. Through the app, Starbucks also has the ability to provide both generic and individualized time-sensitive offers to each customer registered through the app. In tandem with these offers, Starbucks also has information about the users which has been provided by them and can be associated with their purchase history.

As a result of this coinciding user information, this data is a useful metric for attempting to create distinct purchasing patterns based on common demographic attributes or purchasing patterns and create generalized user profiles. These demographics are a key indicator for differentiating consumer segments and how likely they are to respond to each offer within a set of potential promotions. This is supported by academic research, showing that "demographics and user-characteristics may have a significant impact on click-through rates on (research paper) recommender systems" (Beel, Joeran, et al. 399).

Problem Statement

Since their rise to popularity in the 20th century, advertisements have become a widely accepted part of daily life in modern society as a mechanism for creating economic value. Along with the advancement of technology, consumers have also grown in complexity and the market is now more demographically segmented than it has ever been.

Due to these developments, it has become more and more difficult and expensive to develop effective strategies when it comes to developing targeted marketing campaigns in the current digital landscape. Fortunately, technology has also enabled the ability to learn more about individual consumers needs and wants than ever before. The problem that this project aims to solve is improving the user experience and relevance for consumer loyalty mobile applications by personalizing the offer recommendations to match a users' actual interests while maintaining a reasonable profitability level for Starbucks.

Datasets and Inputs

Three datasets have been provided by Starbucks and Udacity for the purposes of generating a model for this project. The scope of this data covers three primary areas.

1. Offer data - Information about the possible offers included in the rewards app
2. User data - Demographic data about the users of the rewards app
3. Interaction data - This data contains information about various interactions between the user and the offers that they receive within the app

Within these datasets, 17,000 users are represented along with their corresponding gender, age, and income information — Unfortunately, one potentially valuable metric missing is a corresponding

geographical location for the provided users. For these users, we have access to the 10 different possible offers which can be included along with the various offer attributes (duration, type, completion reward, along with a few others). The final dataset containing the set of events containing information about the user interaction with these offers consists of about 306,000 records this will amount to about 18 events per user on average.

Solution Statement

The solution for the problem of better understanding demographic segments which exist within the customer base can be represented in the form of a population segmentation model as well as a set of discrete facts which can be interpreted from the outputs to represent each customer segment. The outcome of this intermediary model will consist of three components:

1. Text description which defines a user archetype for each segment of the population cluster (i.e. age groups, time of purchase, product loyalists, etc.)

2. List of the top 3 offers which are most commonly completed for each of these groups based on the estimates from the model. These will be derived from statistical analysis which will be performed on the profiles included in each cluster, specifically, the offers which were most successful in generating the most revenue for the company. Pending the results, further analysis may also be done in order to determine the most profitable offers for each segment.

3. List of characteristics from the datasets which have very little predictive effect when attempting to determine which offers to send to which users.

Benchmark Model

For the benchmark model, I have performed a very simple prototype of the PCA-driven k-means cluster model described in the project design, with one caveat: my calculations were based only on offers which were completed: all incomplete offer data was dropped from the principal component analysis and consequently, the clustering algorithm. As a result, much of the data relating to transactions which were related to partially-completed offers was not able to be used for analysis. From my initial results, I have been able to establish a few base results which can be assessed using the evaluation metrics defined below — Silhouette score and Dunn Index. Using my baseline model, I was able to establish a peak silhouette score of `0.724` and a Dunn Index of `0.549` for m=6 clusters.

Evaluation Metrics

The Starbucks customer segmentation model can be evaluated by its ability to capture the variance which is included within the dataset including the minimal number of features. The goal will be to capture at least 80% of the data variance while keeping the number of principal components as low as possible.

In addition to the ability of the Principal Components to capture the variance, the Silhouette index can be used to evaluate the success of the k-means clustering. represented by the mathematical definition below for each object in any given cluster.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The Silhouette index value will be high for a provided object when it is represented well by the cluster it is within and poorly by neighboring clusters.

Another metric which will be used is the Dunn Index, which is defined as:

$$DI_m = \frac{\min\limits_{1 \leqslant i < j \leqslant m} \delta(C_i, C_j)}{\max\limits_{1 \leqslant k \leqslant m} \Delta_k}$$

For the Dunn Index, the number of clusters is provided. The Dunn Index use used as a validity metric for the provided set of clusters and determine whether the clusters are well defined and well separated - for *m* clusters, a higher DI_m indicates that the clusters are truly representative of divisions within the provided components.

Project Design

In order to train and deploy the discussed classification models, there are several components required during the process. Each step has been highlighted below with a brief description of the relevant work and outcome.

1) **Exploration, Pre-processing and Data cleaning** - Evaluation of the data will be done to glean any interesting patterns or potential features, along with the organization and cleaning required to create a dataset ready for modelling and PCA.

2) **Data modelling and Principal Component Analysis** - Our data is nearly ready. A few visualizations will be generated and make a few notes which may be of useful reference at later stages. Data will be normalized and dimension reduction will be performed with PCA.

3) **Transformation and k-means Preparation** - Once PCA has completed, the data variance will be calculated and used to determine an optimal component count to best capture the customer segments identified within the dataset. The component count will be used for our k-means model. A feature makeup breakdown for each of these components will also be done at this stage.

4) **Population Segmentation and Evaluation** - At this stage, the number customer segments characterized by the data will have been identified. Additionally, we'll be able to highlight the features and components which best identify these segments. Using the traditional metrics highlighted above, we'll also determine the mathematical buoyancy of our findings and most importantly the validity of the customer segments (clusters) we've identified.

Once these stages have been completed, we will be provided with the outcomes which have been discussed in the previous Solution Statement. Several clear strategies which can be taken to market will

be highlighted. In addition to these offer strategies, a list of offer recommendations for each of the customer segments will also be provided. As mentioned above, this information will be calculated based on statistical analysis for customers contained within each cluster maximizing revenue generated by each offer (Pending the results of this analysis, information about the offers which are most effective profit drivers may also be highlighted). In the future, these recommendation lists could also be evaluated using a supervised learning algorithm driven by continued customer segment classification combined with live user/offer interaction data.

Works Cited

Beel, Joeran, et al. “The Impact of Demographics (Age and Gender) and Other User-Characteristics on Evaluating Recommender Systems.” *Research and Advanced Technology for Digital Libraries Lecture Notes in Computer Science*, 2013, pp. 396–400., doi:10.1007/978-3-642-40501-3_45.

“Dunn Index.” *Wikipedia*, Wikimedia Foundation, 30 Nov. 2019, en.wikipedia.org/wiki/Dunn_index.

“Silhouette (Clustering).” *Wikipedia*, Wikimedia Foundation, 18 Dec. 2019, en.wikipedia.org/wiki/Silhouette_(clustering).