Todd Cunningham

11 January 2020

Machine Learning Capstone — Report

I. Definition

Project Overview

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 can be 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). The primary purpose of this project is to provide useful marketing advice to Starbucks derived from the provided data.

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.

In order to solve this problem the following sequential approach will be taken. Each of these steps will be broken down in further detail later on in the report:

- **Data acquisition and Preprocessing -** For this project, data acquisition was simply to download the datasets provided by the client. From there, a few steps were performed to massage the data.

- **Principal Component Analysis** - After performing all necessary preprocessing on the data, we'll perform PCA to determine the most meaningful sources to be used for clustering.
- **Clustering by k-means** - Here, the different customer segments will be identified by using the principal components to classify each customer.
- **Results and Evaluation** - At this point, we'll identify each of the customer segments which is highlighted by the k-means model and determine the validity of each cluster. Some market analysis will also be done and recommendations for capturing each market segment will be made.

Once these stages have been completed, the evaluation will provide a very clear set of market segments as well as customer information which characterizes each of these segments. Included with these segments, several clear strategies which can be taken to market will be highlighted, as well as a list of offer recommendations for each of the customer segments will also be provided.

In the future, the success of these recommendations could also be evaluated using a supervised learning algorithm driven by continued customer segment classification combined with live user/offer interaction data.

Metrics

The metrics which will be used to measure the performance of the model which has been developed will be two common indices which are often employed for evaluation and validation of data clustering algorithms — Dunn Index and Silhouette Index.

The Silhouette Index value is used to determine how well a given datapoint is represented within its cluster. The Silhouette index can be calculated by the following formula:

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

The Dunn Index is used to determine the validity for a set of given clusters. Specifically, a high Dunn Index indicates that the clusters are well defined and well separated. The higher the output (DI_m), the more representative the cluster divisions are of the provided data components. The Dunn Index is defined by the following equation:

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

Using these metrics, I can analyze the performance of the output population clusters for the different customer segments which are identified in the data. The higher the value output by these indices, the more confident I can be about the segments reflecting real-life customer segments which may or may not have been previously identified.

II. Analysis

Data Exploration

As I began my work on the project, the first key step for building a successful model is to process the data in order to highlight the most important factors. In my case, the most important factor was the maximized revenue for Starbucks, and the other attributes were treated as a function of this number. First, I'll provide a very brief summary of the data which was provided and then dive into the ways that I processed this data to obtain a few meaningful metrics which would identify several different customer segments.

Provided Data

1. Offer information

channels	difficulty	duration	id	offer_type	reward
[email, mobile, social]	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10

2. Profile information

age	became_member_on	gender	id	income
118	20170212	None	68be06ca386d4c31939f3a4f0e3dd783	NaN

3. Event information

event	person	time	value
offer received	78afa995795e4d85b5d9ceeca43f5fef	0	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}

The first task of the project was to import this data and to get a better understanding of what's represented. One of the first things that I observed were a few anomalies around null and invalid data. In total, there were 17,000 user profiles included in the data. Of these 17 thousand, 14,835 users reported both their age and their income and 2175 did not. In addition to this information, 2175 users also failed to report their gender. Of the remaining 14,835 gender datapoints, 6129 were women, 8484 were men, and 212 other (unspecified).

In addition to the information calculated for each user profile, some basic information was also gathered on the event transcript data. In total, there were 306,534 records available in this set. Of these records, I found the types to be broken down by each of the following counts:

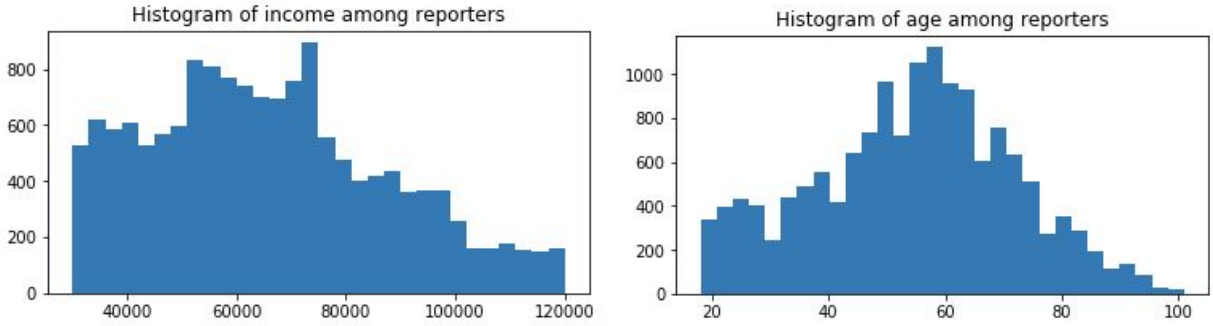
Offer Received	76277
Offer Viewed	57725
Offer Transaction	138953
Offer Completion	33579

Just by performing some basic calculations, there was roughly a 75% viewing rate (57725 / 76277) and a completion rate of about 58% (33579 / 57725) for the offers that were viewed.

As an initial dataset, there seems to be enough information that will be useful for further processing. I'll re-address this later on in section III, when we dive into the preprocessing that I decided to perform to attempt to gain further value from the already provided information.

Exploratory Visualization

In order to gain a better understanding of the distributions of the age and income amongst the provided users, the following plots were created. The y-axis represents the number or records for each of the income and age brackets on the histograms below.



Based on the above histograms, we can see that the most common age is about 58 and the most common income is roughly 72,000. Further analysis done in the notebook reveals that the average age and income are 55~ and $67.8k which reveals that the data follows a relatively normal distribution.

Algorithms and Techniques

The Primary algorithms which were used during the process of this capstone were Principal Component Analysis (PCA) and k-means clustering. Prior to performing PCA, much effort (see *Data Exploration*) was put into making sure that the process included some valuable information about individual users as well as their interaction with the offers which were included in the mobile application.

One of the primary motivations for the usage of PCA was to take a data-driven approach to classification of the different customer segments in the market. Often, marketing segments are obtained from external agencies or indirect assumptions, which was one of the key pitfalls that I was trying to avoid during the course of this project. I wanted to approach this problem without pre-defining different segments by age, gender, or any other typical pattern, and rather let these patterns emerge naturally. By using the principal components as an input to the k-means algorithm, the variance can be appropriately captured and used to generate realistic customer segments.

Once PCA was used to determine the primary source of variance within the provided data, I chose to perform the population segmentation using the k-means clustering algorithm. This is a common clustering algorithm, and one that I had experience with using already as part of previous course projects.

Benchmark

Due to the lack of pre-existing models available for the particular dataset which has been used for the Starbucks Capstone project, I used an early iteration of my own model to set a baseline level of performance which could be compared against after further data refinement.

The performance of this benchmark model was measured using identical metrics to the final model — Silhouette score and Dunn Index. Using my baseline model, I was able to establish a peak silhouette score of `0.0724` and a Dunn Index of `0.00549` for m=6 clusters[1]. In my benchmark model, one factor which was not thoroughly tested was varying the number of clusters as the main purpose was to obtain some initial values.

III. Methodology

Data Preprocessing

Some basic preprocessing was initially required in order to begin enhancing the quality of the features which would be used for the remainder of the modelling process. After some basic analysis, I found that there were 2175 users who failed to report both age and income. I really wanted to include these features in the PCA process, so consequently I decided to drop these users as I felt they were significant features that could be utilized.

After dropping these users, the next cleanup that I felt would help improve the quality of the outcome of the results was removing users which failed to meet a few other simple criteria — I only wanted to look at users which had completed at least 1 offer, performed at least 3 transactions, and spent at least $20 in offer transactions. I spent some time tuning these values and found that these expectations applied to quite a large percentage of the remaining 14,825 users. After these users were filtered out, I was left with 10,913 users who met all my criteria. Upon completion of this step, we can proceed to PCA.

As Principal Component Analysis will be used for the processed data to determine which features are responsible for capturing the data variance, the primary goal of this step is to maximize the number of useful features which are available for the PCA model.

Additionally, the primary goal of this project is to identify unique customer segments to which marketing strategies can be applied. As such, the primary focus of the preprocessing efforts will be the profile data which as we want each datapoint in my final k-means model to represent a customer.

With that in mind, there are still some useful ways the data from the other sets can be applied to each customer. The first way I thought to gain additional information for each customer was to perform some processing for each event which applied to that customer. Specifically, I created 4 new columns which were derived from the offer event transcript set: received, viewed, transactions, and completed. Each of

[1] Originally in my project proposal, I reported Silhouette and Dunn index scores of `0.724` and `0.549` respectively. However, during the process of further developing my model I found a miscalculation was performed and the correct values for my benchmark were actually `0.0724` and `0.00549`

these columns represent the corresponding number of events for each type which are associated with that user id. Below is a brief preview of a few of the records from the preprocessed table:

age	became_member_on	gender	id	income	gender_val	received	viewed	transactions	completed	revenue
118	20170212	None	68be06ca386d4c31939f3a4f0e3dd783	NaN	0	5	5	9	2	20.40
55	20170715	F	0610b486422d4921ae7d2bf64640c50b	112000.0	3	2	0	3	1	77.01
118	20180712	None	38fe809add3b4fcf9315a9694bb96ff5	NaN	0	2	2	6	0	14.30
75	20170509	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0	3	4	4	7	3	159.27
118	20170804	None	a03223e636434f42ac4c3df47e8bac43	NaN	0	5	3	3	0	4.65
68	20180426	M	e2127556f4f64592b11af22de27a7932	70000.0	2	4	3	3	2	57.73
118	20170925	None	8ec6ce2a7e7949b1bf142def7d0e0586	NaN	0	5	5	0	0	0.00
118	20171002	None	68617ca6246f4fbc85e91a2a49552598	NaN	0	5	4	2	0	0.24

In addition to the event type columns above, there are a couple additional features which were also added in later instances of iterating across my model - gender_val and revenue. These two values represent a gender value which can be normalized and used for PCA, and the total revenue generated by all transactions for each individual user. At various points, I experimented with several different metrics, such as average timing of different events but found that they were often insignificant in the resulting PCA components and consequently removed.

Implementation

In this section, I will break down each step of processing which was performed on the data as well as each of the metrics and algorithms which were used to evaluate the results at different stages. We've touched on these in other sections of this report but implementation details are detailed below:

1. **Metrics** - As discussed in the Definition section, the two primary metrics were *Silhouette Score* and *Dunn Index*. During the benchmark stage, I implemented my own versions of these calculations based on their mathematical definitions. However, as previously mentioned I found issues with my calculations and as a result I decided to take advantage of some existing implementations which I had seen used in several other cases. I used the *Silhouette Score* from the scikit-learn metrics library. For the *Dunn Index*, I used a library call jqm_cvi.

2. **Algorithms** - Two main algorithms were used during the development of my project: PCA and KMeans. For both of these algorithms, I used the default sagemaker implementation of these models. Fortunately, these two algorithms were ones that we had gained some experience with during earlier projects in the course. They key inputs for the PCA which was done were the normalized, preprocessed data, and the number of components which were used as input into the PCA model (7 in my case). After completion of the PCA, the top 4 components were then fed directly into the KMeans model. The number of top components was selected as 4 to capture roughly 80% of the data variance (78% captured in my case). The number of clusters was also defined in the KMeans model which was chosen to be 7 as well. This number was chosen based on the evaluation which was done to determine the optimal number of clusters to maximize the *Silhouette Score* and *Dunn Index*. Once KMeans was completed, the remainder of the outstanding work was to interpret and evaluate the results.

Refinement

During the process of the development of the model, the primary source of iteration in my case was the data preprocessing step. The other sources for refinement were primarily adjustments which were made to the number of clusters and number of components which were used during the PCA process depending on the proportion of variance captured.

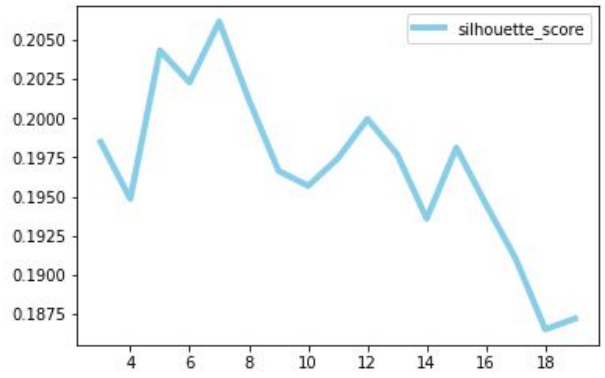
My initial steps for refinement were implementing the event counts for each of the event types included in the offer event information. Following these inclusions, I decided to also include the revenue provided from the offers for each user.

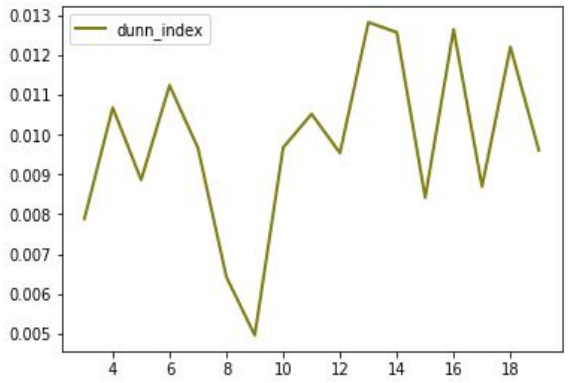
At various points during the model iteration stage I also experimented with additional fields, such as accounting for various durations about each user's offer, including time to view, time to completion, and other various metrics. However, I did not see much improvement during the inclusion of these features and left them out of later iterations of the model.

IV. Results

Model Evaluation and Validation

As discussed in the metrics at the start of the report, the two sources I've chosen to use to measure the performance of my model is the silhouette score and Dunn Index, two common metrics used for measuring cluster validity in segmentation models. In my notebook, I included calculation for these metrics and also did so for a range of cluster sizes. These calculations are based on the top 4 components which capture 78.0% of the variance in the processed data. Graphics which show the results are below:





Based on the graphs above we can see that the optimal number of clusters for the data we've processed using k-means is 7. The `silhouette_score` is `0.206` and `dunn_index` is `0.0097` for m=7 clusters. Further detailed values for each of the other cluster findings are included in the notebook in this repository.

Justification

Based on the outcome results for both the benchmark and the final version of the iterated model which was used to obtain the final outcomes, the increase `silhouette_score` from `0.072` to `0.206` and `dunn_index` from `0.0054` to `0.0097` indicates an increase in validity for the clusters of 186% and 79% respectively. On a positive note, there is a clear indication that my model has improved quite substantially in segmenting the clusters based on these clustering metrics. However, these scores are still very low and in my opinion indicate that they do not successfully represent valid clusters.

As a result of these findings, I am disappointed to say that the expected outcomes of identifying successful market strategies is not possible. I could theoretically perform some marketing analysis on the final data which has been found. However, based on the metrics the data does not capture the market well enough to provide a basis on which to make any valid marketing assertions.

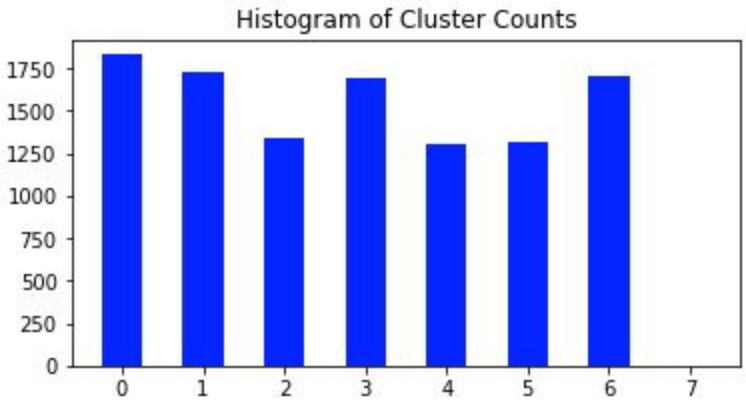
My recommendation for the continued progress of this project would be to re-evaluate the preprocessed features in addition to the creation of many new features. These new features would naturally be eliminated by the PCA process as they are found to be insignificant. Additionally, PCA could possibly be eliminated from the process altogether and the data could be evaluated directly by k-means. There are quite a few possibilities for improvement of these results and these are only a few suggestions.

V. Conclusion

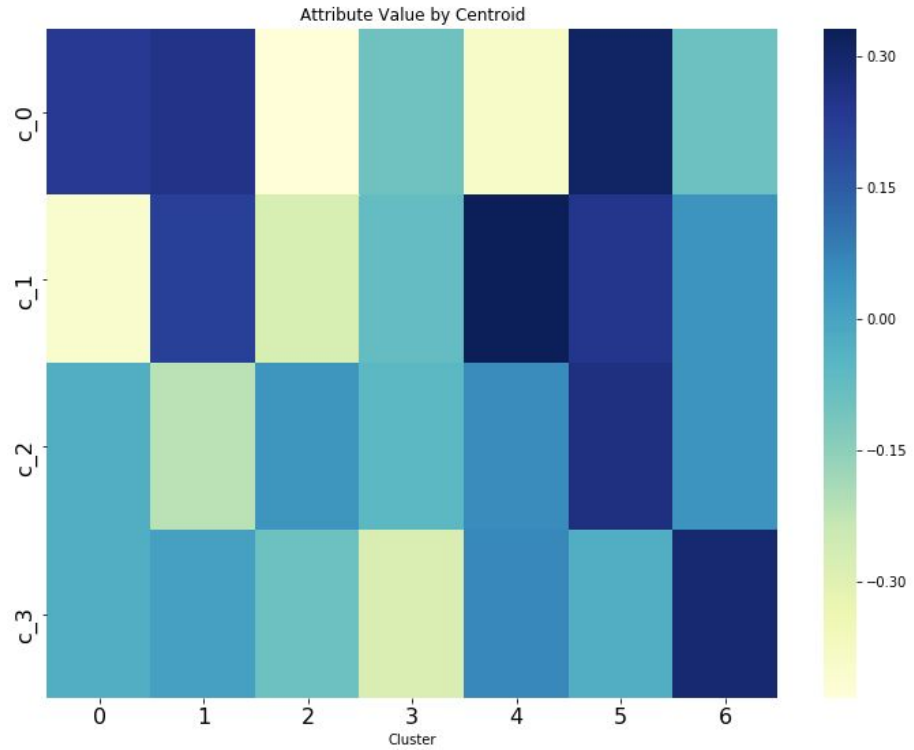
Free-Form Visualization

In addition to the final outcomes which are shown above by the dunn index and the silhouette scores, There were several steps along the way which exhibited a few interesting graphics which I believe are worth discussing and interpreting as part of the discussion of this report.

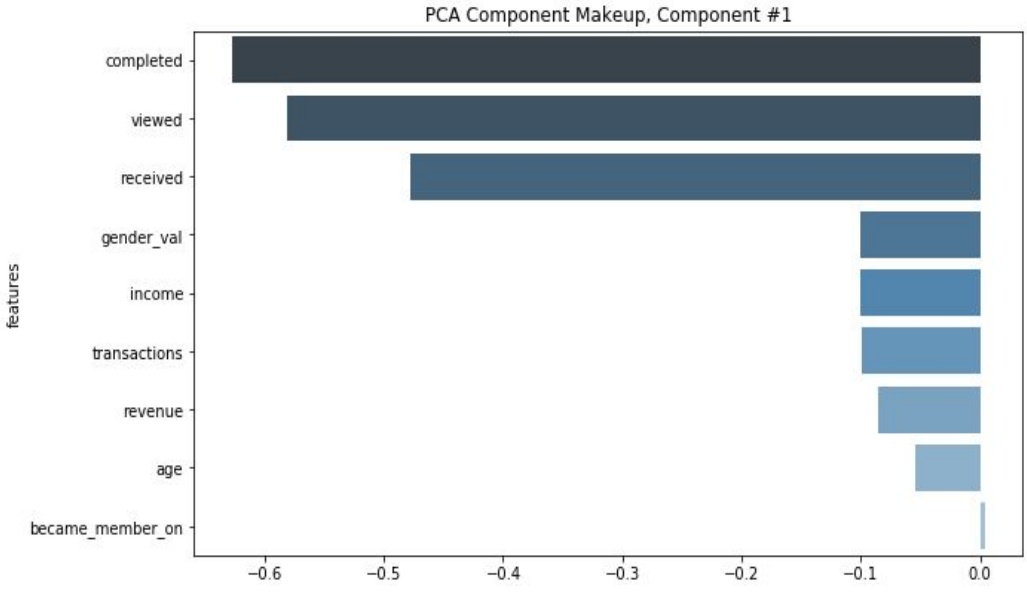
The first graphic which will help get an idea of the final cluster sizes for each of the clusters Identified by the k-means model:



As we can see, the clusters are relatively evenly distributed. Following the breakdown of the size of each cluster is a heatmap for each and their components. This allows us to see which characteristics define each of the clusters.

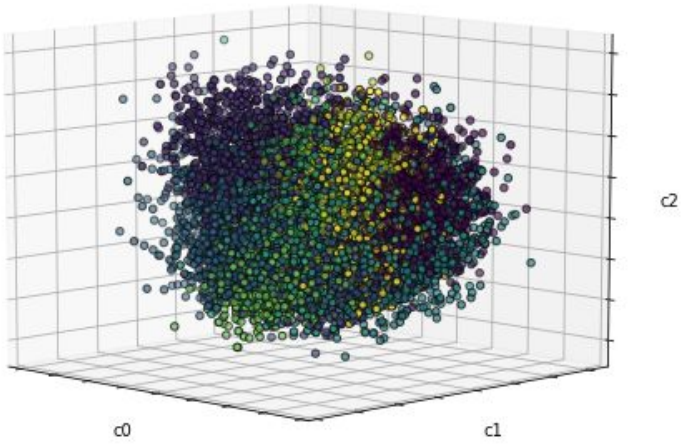


The next graphic which allows us to further interpret the output of the clustering are the component makeup diagrams. Each of these shows how the component is built up for each of the features which were included in the Principal Component Analysis. For a detailed breakdown for each of the components, please see the Jupyter Notebook included in the GitHub repository.



The above graph shows the breakdown for the first component represented by c_0 in the heatmap. We can see that this component is primarily defined by the transaction events - completion, viewed, and received. This means that the centroid for cluster 0 was primarily influenced by these factors.

One last final graphic which I was able to visualize during the process of my capstone project is displayed below and represents a 3-component graph of a set of clusters which really helped me understand part of the issues that I've observed during the course of this project and realize how some of the improvements that are discussed later in this report could influence future outcomes.



As it is easily observable in the diagram above, we can see quite a convolution between the different clusters in the diagram. This can likely be attributed to several factors, however one thing that I never understood until seeing this diagram was the close relationship between all of the data points for each of the user profiles. To me, this signalled that I had a long road of improvements ahead in order to successfully attempt to break these customers apart into more distinct groups.

Reflection

Upon completion of this project, I found that I was able to take away quite a bit of understanding of common problems that arise during the process of using machine learning tools and attempting to use models to provide tangible outcomes using these models. In an attempt to breakdown the reflection process, I will discuss my personal thoughts and findings for each set of the steps which were broken down in the problem statement.

Data acquisition and preprocessing - Data acquisition didn't really apply as the data was provided by Starbucks and I did not require any external data based on my proposal. As for the preprocessing, I found this to be the opportunity that I viewed as the most interesting spot to create some valuable insight into the provided data. Unfortunately, I feel that I was unable to highlight the features which were able to allow PCA to highlight the conclusion that I was trying to reach and resulted in some of my later shortcomings and as a result I believe this to be the most difficult step to get right in the process.

Principal Component Analysis - PCA was one of my personal favorite procedures that was used in this course which I never realized was a tool which could be utilized in the machine learning process. I find

that it has the possibility to take many uninteresting values and create exciting results if you can provide the right inputs. I didn't find this part to be particularly difficult as I had previous experience using PCA via sagemaker.

Clustering by k-means - Similarly to PCA, I was intrigued by the applications of k-means and particularly so with respect to the starbucks data. I see that there are some big opportunities if this can be applied and trained correctly.

Results and Evaluation - Unfortunately, I was quite disappointed with the results of the clustering which my model produced. I feel that in some ways I made no progress, but in others I feel that I have succeeded in finding a way that does not successfully segment the users, but rather identifies users which are similar to one another in ways that would otherwise never be realized.

Overall, I've found the process of performing this capstone project to be very informative and like to think that I have absorbed a very brief glimpse into common problems of professional machine learning engineers. It has also opened my eyes on how modelling can be tricky in general and ways of taking a better approach to this even outside of a machine learning environment.

Improvement

As mentioned in the previous sections, there are quite a few improvements which could be made on the project implementation based on the outcome results. A detailed breakdown of each of these improvements is as follows:

1. **Improve the Solution Statement -** I feel that this is the most significant shortcoming of my project as a whole. After the continued evaluation and reflection of my findings, I feel that under any circumstances it would be difficult to make a successful assertion about various segments of the customer base which is measurable. Additionally, there are some aspects of the data which are not included which may have a great impact on the effectiveness of my supposed recommendations (i.e. geographical impact, seasonal impact, etc.)

2. **Improved Preprocessing -** I believe that there are still quite a few improvements which could be made just by improving the preprocessing step of the overall procedure which was exercised for my project. I'm unsure whether it's possible to completely resolve all issues, however. I believe that the issues highlighted in items 1 and 3 would likely become much more obvious as the value highlighted by the preprocessing stage improved.

3. **Changes to the segmentation procedure -** The segmentation procedure which was followed for the course of the project was successfully completed in terms of input and output. However, I feel that there are some changes which could possible make the segmentation within the customer base more obvious. One primary example that I can think of is to remove the PCA which was done and simply run the k-means clustering after preprocessing the data. This would remove the normalization step which was performed and as a result, I may have been able to more easily observe interesting segments within the 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).

Viegas, Joaquim, jqm_cvi, (2016), GitHub repository, https://github.com/jqmviegas/jqm_cvi