**Starbucks Targeted Marketing Model**

**Business Context:**

Starbucks’ holiday gift card, seasonal drinks and holiday special branded goods have positioned themselves as few of the “go-to” presents during the holiday season. Especially, Starbucks’ marketing campaign that connects the gift cards to its seasonal menus and products have been bullish and generates around $11 billion of revenue for Starbucks annually. As the end of the year comes near and with holiday spirits rebounding after a not so celebratory holiday season due to the pandemic last year, Starbucks is aiming to recover its lost customers and revenue now that the pandemic is coming to an end.

**Business Purpose:**

In order to assist Starbucks’ in doing so, our team has come up with a machine learning model that analyzes customer data and clusters customers together based on their response to purchase offers sent via the Starbucks app. From this model, Starbucks would be able to gain valuable insight on what marketing tactics and offers they can deploy towards what kind of customers. Our model would assist Starbucks in not only maximizing their marketing efforts, but also enable them to make more effective marketing decisions to ultimately lead to the customers purchasing the offered product.

**Data wrangling**

We obtained the Starbucks dataset from [Kaggle](https://www.kaggle.com/blacktile/starbucks-app-customer-reward-program-data). It is made by Starbucks for Udacity scholars, and it contains three files: portfolio.json, profile.json, transcript.json. Portfolio.json introduces each offer (duration, type, etc.); profile.json contains the customer demographic data; transcript.json contains records for transactions, offers received, offers viewed, and offers completed. The detailed feature explanation table is attached in exhibit 1.

For portfolio.json, there are 10 observations and 6 features. The summary statistics of the numeric variables are shown in Exhibit 2: Table 1. For channels, there are only 4 types - (web, email, mobile, social), (web, email, mobile), (email, mobile, social), and (web, email), arranged in descending counts. For offer types, there are only three types: “bogo” and “discount” each have 4 counts, while “informational” has 2 counts.

Profile.json has 17000 observations and 5 features. The summary of the numeric variables are in table 2. For gender, 8484 customers identified as male, 6129 were female, and 212 chose others. For transcript.json, the summary statistics are shown in table 3 and table 4. For the event variable, there were 138953 transactions, 76277 ‘offer received’, 57725 ‘offer viewed’, 33579 ‘offer completed’.

For data wrangling, we conducted feature engineering. In our combined data (where we joined the three files), we added several dummy variables for each level of all the categorical variables, to prepare for the clustering next step. We also calculated the average amount each customer paid per visit, the visiting frequency, the conversion rates (from receive to view, and from view to completed), and the total offers each customer received, viewed or completed.

The challenges and limitations of the data include not having enough features or data such as education, location, purchasing preference, etc. Additionally, results might not be helpful for all the stores: different regions will not have the same clusters, so the results could be too general for each store/neighborhood/city.

**Model Free Insights:**

After cleaning and prepping the data, we used queries to gain insights on our dataset (Exhibit 3). We decided to gain the bulk of our insights from grouping by these event types: offer completed, offer received, offer viewed and transaction. We observe that Starbucks customers all have an average income in the 60k range but those who complete the offer have the highest average income ($69.4k) and those who simply complete a transaction have the lowest average income ($61.8k). Additionally, the average age of customers are in the early to late 50s as well with not much fluctuation within each event type. Next, we evaluate the number of customers based on gender and how they engage with Starbucks offers. Overall, we can observe that the majority of Starbucks’ customers are male.

**Reason of choosing clustering analysis and Spark**

To help Starbucks achieve more effective customer marketing, we used clustering analysis (KMeans analysis) to accurately segment customers. Specifically, with the dataset we have, we segmented customers into several groups with similar characteristics and buying habits. By doing the clustering analysis, we are able to allocate marketing budgets, target customers with more personalized offers, and maximize the cross or up-selling opportunities.

For making the clustering analysis, we need the historical transactions data, customer data, and data about our previous offers. These kinds of data are stored on a very large scale. For now, we know that in their last annual reporting period, Starbucks likely used between 2.916 and 2.946 billion cups at their stores, or an average of 8,070,428 per day. Therefore, we need to rely on big data processing and analysis tools to run the model and gain insights. In our project, we used Spark.

Compared to Spark, Pandas is a great tool for analyzing data on a local machine, but it is naturally well adapted to relatively small datasets. To avoid downsampling which compromise the dataset, we could use tools such as Spark to aggregate and process data. Especially in our case, when training a machine-learning model, it would be better if we have as much data as possible. Spark can work with data in a distributed storage system, such as HDFS, to parallelize the process and accelerate computations. With the ability to compute in real-time, Spark can enable faster marketing decisions.

**Clustering Analysis and insights from the results.**

After the feature engineering, we got our final dataset for the clustering analysis. We used KMeans Clustering in PySpark to segment the customers. In order to choose the optimal number of clusters, we calculated the silhouette scores, aka the separation distance, between the resulting clusters. From the line plot of silhouette scores against the number of clusters, we could observe that when we use 4 clusters, the silhouette score is the largest. The clustering results are visualized as the bar charts in Exhibit 5 of the appendix. The bar charts show the characteristics of these 4 clusters. We can develop different marketing strategies for these different kinds of customers.

Cluster 1: These customers received many offers but they seldom view them or complete them. They seldom visit our stores and they don’t spend a lot in our stores. From demographic results, we found that most of them are male customers with the lowest average income of $57.5k, and the lowest average age of 51.

Business action: due to the low conversion rate, we should avoid targeting them as main customers in the marketing.

Cluster 2: These customers received many offers but they seldom view them or complete them. They don’t spend a lot of money in our stores. However, they visit our stores the most frequently. We picture them as regular customers of Starbucks. They might just get a cup of coffee everyday before work, but do not necessarily spend a lot. The demographic of these customers are majority male at the average age of 52. They make an average income of around $57.6k.

Business action: We should reduce the frequency of sending them offers. As regular customers, they will still visit our stores. We can use other incentive plans to keep these customers.

Cluster 3: They received the most amount of offers and they often view or complete the offers than the other clusters. They often visit our stores, but they do not spend so much money in our stores. These customers might also be our regular customers, but they are more sensitive to offers they received and would like to view and use them. These customers are also majority male with an average income of $65.2k with an average age of 54.

Business action: We should continue sending them offers.

Cluster 4: These customers receive offers the least often, but they view the offer and complete the offer most of the times when they receive them. They often visit our stores and spend the most amount of money per visit in our stores. They have the highest average income of $70.8k and are also the oldest with an average age of 56. This is the only cluster that is mostly women.

Business action: We should send more offers to these customers, and they are the customers we should target to seek potential sales increase.

**SWOT Analysis:**

*Strengths*: Deploying our targeted marketing model would be beneficial to Starbucks in that it will be more effective than Starbucks sending out uniform marketing offers to its whole customer population. As our model would offer Starbucks a better understanding of what types of customers respond to their offers based on their age, gender, and income.

*Weaknesses*: One of the weaknesses of our targeted marketing model would be that because we are “segmenting” customers into groups to send out targeted marketing campaigns, it may be possible that there may be some people who are wrongly grouped and there is a potential of our marketing model not being so effective against those people. Additionally, because of the limited information we have on each customer profile, there may be more variables that further distinguish each cluster’s consumer base.

*Opportunities*: By saving up on time and resources on automatically segmenting our customers, it would provide Starbucks to redirect these resources to other parts of the company that need growth or development. Doing so would enable Starbucks not only to improve their current business structure, but also have more resources to spare to expand into new areas.

*Threats*: One risk of utilizing our model too much may be inaccuracy and leak of customer data. Similar to what happened to Zillow’s algorithm model, there is a possibility that our targeted marketing model would not be suitable when there are changes in certain factors. Additionally, the more we use customer data from our databases, it increases the risk of the customer data being breached by external parties. Both of the threats above can leave an impact on Starbucks’ brand, which is one of its core competencies.

**Appendix**

**Exhibit 1**

| **Portfolio.json File** |  |
| --- | --- |
| **Feature** | **Explanation** |
| id (string) | offer id |
| offer\_type (string) | type of offer i.e. BOGO, discount, informational. |
| difficulty (int) | minimum required spend to complete an offer. |
| reward (int) | reward given for completing an offer. |
| duration (int) | time for the offer to be open, in days. |
| channels (list of strings) | "email"; "mobile"; "social"; "webl" |
|  |  |
| **Profile.json File** |  |
| **Feature** | **Explanation** |
| age (int) | age of the customer. |
| became\_member\_on (int) | date when the customer created an app account. |
| gender (string) | gender of the customer (note some entries contain 'O' for other rather than M or F) |
| id (string) | customer id. |
| income (float) | customer's income. |
|  |  |
| **Transcript.json File** |  |
| **Feature** | **Explanation** |
| event (str) | record description (ie transaction, offer received, offer viewed, etc.) |
| person (str) | customer id |
| time (int) | time in hours since the start of the test. The data begins at time t=0. |
| value (dict of strings) | either an offer id or transaction amount depending on the record. |

**Exhibit 2:**

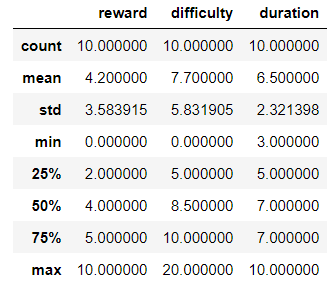
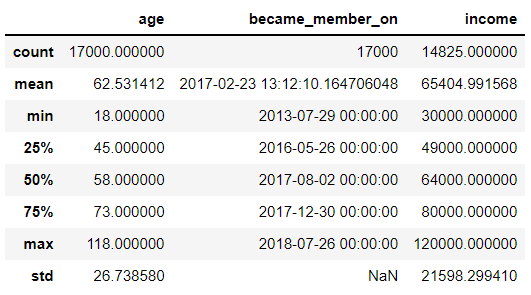
 

Table 1: Portfolio.json Numeric Variables Table 2: Profile.json Numeric Variables

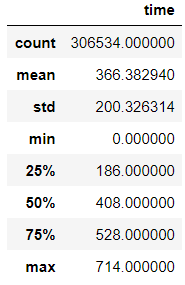
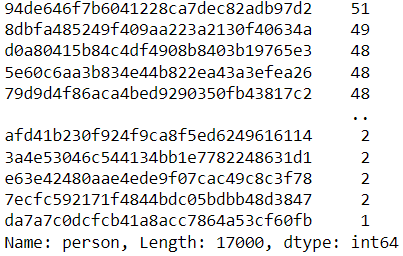
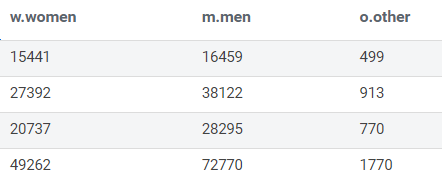
 

Table 3: Transcript.json Time Variable Table 4: Transcript.json Customer ID Offer Counts

**Exhibit 3: Query Table of Cleaned Raw Data Pre Modeling**

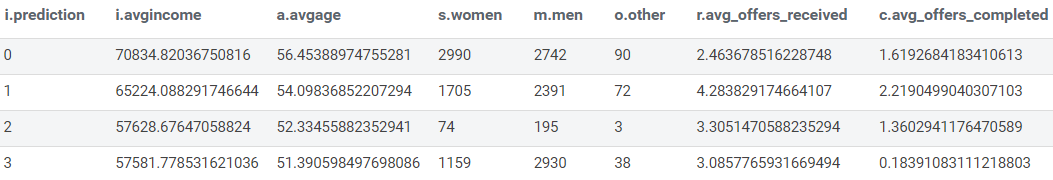


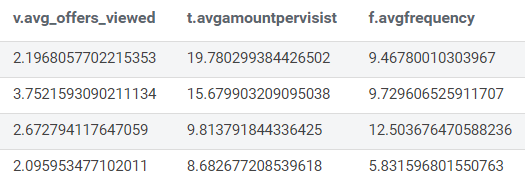


**Exhibit 4: Query Tables of Clusters**

Prediction Column:

0 is green, aka cluster 4; 1 is yellow, aka cluster 3; 2 is brown, aka cluster 2; 3 is orange, aka cluster 1





**Exhibit 5: Clustering Results**

