Using Customer Segmentation to Improve Starbucks' Marketing Campaign

Dariia Dragunova, Sameera Rachakonda

Introduction Section

In marketing, customer segmentation is the process responsible for dividing a heterogeneous market into several more similar segments based on specific geographic, demographic, psychological, and behavioral parameters. The goal of market segmentation is to determine which products are most likely to reach the target market share and find the best way to introduce the products to the market. The data collected based on customer segmentation enables companies to obtain information about their customer preferences. In addition, companies can find strategies that can help them maximize profits, formulate marketing strategies more effectively, and reduce investment risks. Market segmentation is an important process for any campaign because it allows you to better understand the audience of your product and identify your ideal customers.

In data science, customer segmentation is an unsupervised learning application that uses clustering techniques to help organizations group their customers based on some common characteristics (such as gender, age, interests, or different consumption habits). A common cluster analysis method is a mathematical algorithm known as k-means cluster analysis, sometimes referred to as scientific segmentation. The clusters that result assist in better customer modeling and predictive analytics, and are also used to target customers with offers and incentives personalized to their wants, needs and preferences.The data itself reveals the customer prototypes that inherently exist within the population of customers.

For our project, we decided to focus on developing a customer segmentation model to improve Starbucks' Marketing Campaign. Starbucks is a multimillion dollar company that has customers all over the world. Unlike smaller companies, Starbucks has millions of customers that have their own preferences and thus need a specified marketing approach. Their new mobile application "Starbucks Rewards App" aims to grant access to customers through their fingertips while giving them exclusive offers and notifying them about most recent rewards. Thus, the motivation behind our project is to segment customers into different groups that respond best to a particular marketing approach and automate an effective outreach campaign for the Starbucks Rewards mobile app.

Our contribution to the project was that we were able to cluster customers into 5 categories, rather than 4 that can help cater marketing campaigns to be most profitable. We found that the higher number of clusters would help make the marketing approaches more specific and thus, more effective. We explain our approach and results in the following report.

Approach

The main question that we were aiming to answer was "How to use cluster analysis to target marketing outreach?". For our project, we explored customer transaction and marketing offer datasets provided by Starbucks. Starbucks publishes sets of simulated data that mimics customer behavior on the Starbucks Rewards mobile app. In the mobile app, Starbucks sends out an offer to its users once every couple days. An offer can be merely an informational advertisement or an actual offer such as a discount or buy-one-get-one-free. The published dataset was a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.

The approach that we decided to implement was to use the Principal Component Analysis (PCA) method and the K-Means clustering unsupervised Machine Learning algorithm. The data that we worked with combined three datasets that included information about the profiles of customers, portfolio of offers

that were sent to the customers, and a transcript of offer interactions. The profiles dataset contained data of customers, specifically their age, gender, income, and date of becoming a Starbucks member. The portfolio dataset consisted of the portfolio of offers that were sent to the customers. These marketing offers were sent over a period of 30 day testing and they were attached via web, email, mobile and social media. The offers have varying levels of difficulty that are based on the minimum amount of money and reward and they fall into three categories, such as Discount, Buy-one-get-one (BOGO), and Informational. Finally, the transcript dataset contained a list of offer interactions that could be one of the three types: receive, view, or complete, and all other transactions during the test period.

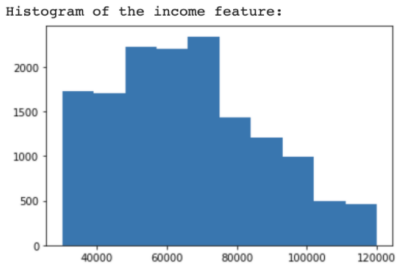
Here is the basic information of our datasets.

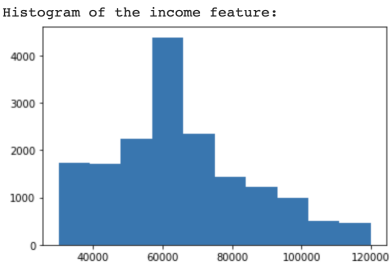
```
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   reward      10 non-null     int64 
 1   channels    10 non-null     object
 2   difficulty  10 non-null     int64 
 3   duration    10 non-null     int64 
 4   offer_type  10 non-null     object
 5   id          10 non-null     object
dtypes: int64(3), object(3)
```

```
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype         
---  ------            --------------  -----         
 0   gender            17000 non-null  int64         
 1   age               17000 non-null  int64         
 2   id                17000 non-null  object        
 3   became_member_on  17000 non-null  datetime64[ns
 4   income            17000 non-null  float64       
dtypes: datetime64[ns](1), float64(1), int64(2), obj
memory usage: 664.2+ KB
```

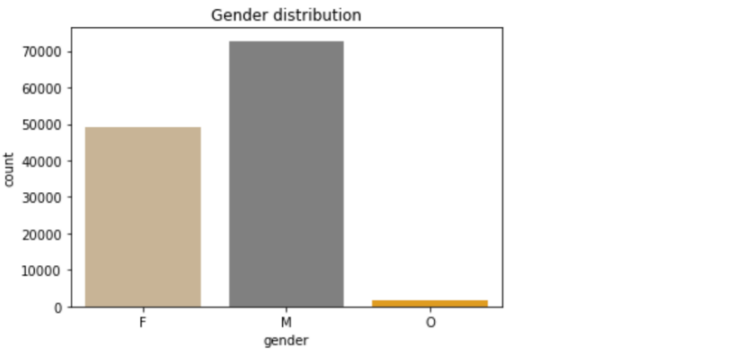
As our first step, we pre-processed the Customer Profile dataset. This dataset included 17 thousands samples with features such as their age, gender, income, and date of becoming a member. To pre-process data, we intend to find the missing values, convert categorical values to numerical features and scale features. The missing values in this data set were only found under categories of income and gender. To replace missing values in the income feature, we first graphed the data to identify if it had longtail or non-longtail distribution. Since the graphs showed non-longtail distribution, we replaced the missing values with a mean income value. For the transcript data set, we split the transactions data into transactions and offers.

Here is the distribution of the income feature before and after replacing missing values:





The graph below is a histogram from our code that plots the gender distribution of the dataset, where there is female, male, and other. From this dataset we were able to learn that most customers are male.




```
Descriptive stats for age and income:
                age         income
count  17000.000000   17000.000000
mean      62.531412   65404.991568
std       26.738580   20169.288288
min       18.000000   30000.000000
25%       45.000000   51000.000000
50%       58.000000   65404.991568
75%       73.000000   76000.000000
max      118.000000  120000.000000
```

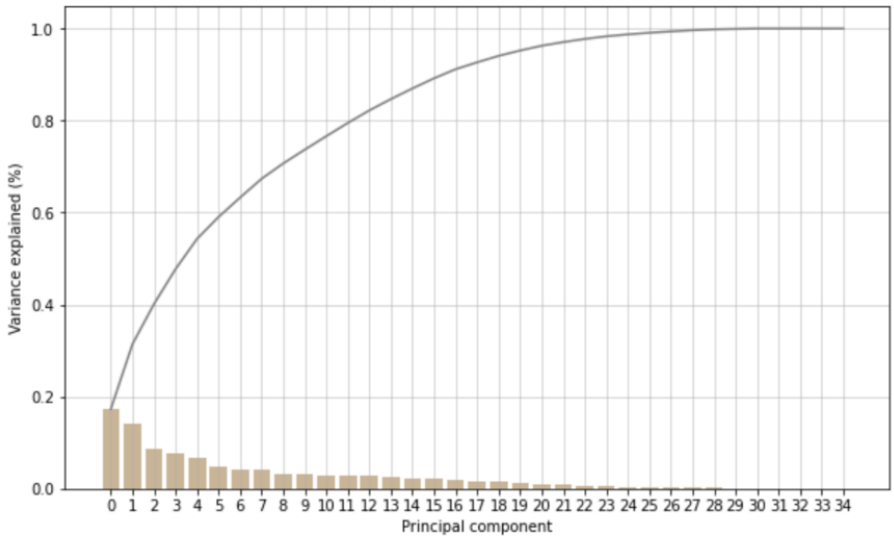
Here, we can see some descriptive information about the data that shows that the mean average age is 62 years old and their income is around $65k which helps us better understand the market that we are working with.

We also plotted age, income, and spending by gender on a scatter plot.

Additionally, we graphed the three different types of offers: BOGO, Discount, and Informational. We used a bar graph to show how many of each were received, viewed, and completed.

Next, one of the biggest steps was to merge all the datasets into one. For this step we decided to use Dummy Datasets. In order to connect the datasets, we merged it by the offer_id feature. We also added receipt, view and completion time of the transactions in order to better understand the relationship between the users and the offers. We then grouped the offers by person, offer_id and the event they were associated with.

Since the datasets that we worked with contained a large amount of data, we used Principal Component Analysis (PCA) to reduce the dimensionality of it. The following graph shows the percentage of explained variance for principal components.



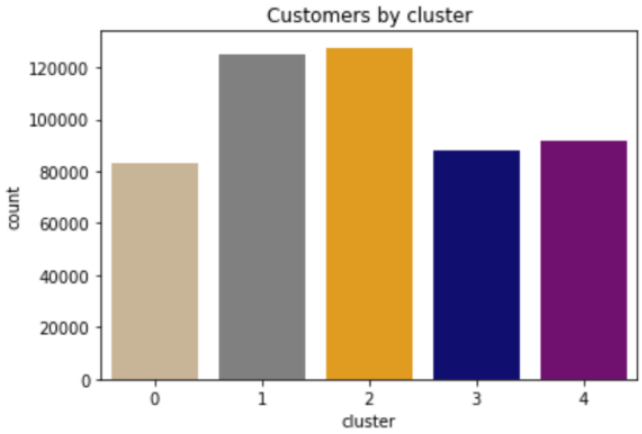
Lastly, we used K-means clustering to obtain the customer segments we were looking for. We decided to go with five clusters because it was more effective to target higher number of different types of customers to get better marketing outreach.

Results

As a result of our project, we created a model that identified five market segments with different buying habits and responses to marketing offers. The first segment contains customers that receive regular

Buy One Get One offers, and almost no discount offers. The second segment describes customers that receive a higher than average number of offers, and convert really well for both Buy One Get Ones and discounts. Customers in the third segment receive no Buy One Get One offers but they do get occasional discount offers, to which they convert well. The fourth segment identifies customers that receive regular offers, which they open, but never convert. The fifth segment consisted of customers that received discount offers and almost none of the other types of offers.

The following bar graph shows the five clusters we created, and the customers in each one. As can be seen from the graph, most customers belong to cluster #3 (2 on the graph) that mostly enjoy discount offers. Following it is cluster #2 (1 on the graph) that shows users that convert well to BOGO and discount offers.



Conclusion

In summary, we began this project by preprocessing and cleaning our data. Majority of our project dealt with preprocessing the datasets that we had. As our end result, we generated five segments using K-Means clustering technique to highlight the different purchasing habits of consumers and reactions to marketing offers with the ultimate goal of creating an effective outreach campaign for Starbucks Rewards Users. The motivation behind our project was to create a campaign to target Starbucks Rewards Users, which we achieved by creating different segments of BOGO, Discount, and Informational.

Acknowledgements

We would like to acknowledge the following web resources for helping us finish our project. We used the datasets provided by the following source.
https://seifip.medium.com/starbucks-offers-advanced-customer-segmentation-with-python-737f22e245a4