**Capstone Project Report**

Udacity Machine Learning Engineer Nanodegree

Starbucks Project

By Jing Li

August 21, 2020

**Table of Contents**

1. Definition
   1. Project Overview
   2. Problem Statement
   3. Metrics
2. Methodology
   1. Analysis Components Overview
   2. Data Preprocessing Steps
   3. Engineered Feature Summary
3. Analysis
   1. Descriptive Analysis
   2. Customer Segmentation
      1. RFM Analysis
      2. K-Pototype Clustering
   3. Supervised Learning
      1. Data Labeling Logic
      2. Baseline Model and Potential Improvements
      3. Final Model Assessment
   4. Cost Simulations
4. Results
   1. Descriptive Analysis Findings
   2. Customer Segmentation Findings
   3. Supervised Learning Results
   4. Cost Simulation Results
5. Conclusion
   1. Problem Statements Review
   2. Conclusions to each Stated Problems

**Definition**

**Project Overview**

Starbucks is the largest international coffeehouse chain that based in Seattle, Washington. Introduced in April 2008, Starbucks launched its first ever rewards program. Over the course of past several years, the company had replaced its one-star-per-transaction model in favor of a revenue-based loyalty program. Early last year, 2019, Starbucks redesigned the program to include tiers, opening more ways to redeem rewards.

Closely related, this project provides actionable insights that help drive the company’s marketing campaign efforts to a rather optimal state which in return, more effectively reaching out the right customers with individually tailored rewards offerings.

Intuitively, the benefit of this analytical process, which described below, is to enable healthy marketing promotions that could be more smoothly embedded within multiple channels such as email, mobile, and web, etc... Thus, achieving higher revenue.

**Problem Statement**

This report is consisted of three components that collectively provides business users an end-to-end solution with the goal of answering the following key business questions:

1. Which type of offers (informational, discount or bogo offers), a given customer is most likely to complete?
2. What improvements can be made in regard to modeling performance, so that a minimized cost is achieved after applying the cost assumptions?

In addition, the stated key questions are further decomposed into smaller questions that are answered throughout each solution components. These decomposed questions help us gain better understanding of our sampled rewards program customers. These questions are listed in the following:

1. What are the success rates for each offer?
2. What statements can we make when comparing the bogo and discount offers?
3. When people make certain decisions about the offers, are there any common characteristics that these people share? In addition, what customer characteristics would be considered as ideal or valuable to the business?
4. How does the customer groups labeled as “ideal” and “valuable” react to each type of offer?

**Metrics**

For supervised learning task, the performance metric that are being used in this analysis is average weighted F1 score. This metric has a range of 0 to 1 and the higher the number is, the better performance of a model. Furthermore, false positive rate (Type I error) and false negative rate (Type II error) are being used during the composing of a cost function in order to analyze expected cost per customer related to various modeling error rates.

For unsupervised learning task, the Hopkins statistic was used to determine how well the data can be clustered prior to the clustering work. This statistic tests the spatial randomness of the data and indicates the cluster tendency. Ranged from 0 to 1, a number that is 0.5 or below indicates the data is uniformly distributed and hence it is unlikely to have significant results. However, If the statistic is above 0.5, then the data has high tendency to cluster and therefore likely to have statistically significant results.

**Methodology**

**Analysis Component Overview**

1. Descriptive analysis:

This component aims to provide business users with a soft understanding of how customers can be viewed as a certain group based on some common characteristics. Upon the successful execution of this analysis, a user would be able to make initial conclusions, through descriptive analysis findings, about where do the current rewards offerings stand in terms of success rate, influence rate, etc.

1. Customer segmentation:

The goal of this component is to form groupings from the current customers base based on hidden characteristics, if there is any. In addition, the following analysis will provide detailed explanations on the grouping results as well as the methodologies being employed based on customers’ purchasing frequency, monetary value spent, age and gender, etc.

1. Model training & Cost simulations

Now that we have had a basic understanding of the collected customer population, and some common characteristics that can be employed to segment customers into multiple groups. We would like to all that previously generated results as our newly engineered features to let a machine learning model to learn from. Furthermore, a simulation for the cost of sending the wrong offers is included to determine under what circumstances, or modeling errors, a trained model would produce the lowest cost for the business. Upon successful execution of this component, a trained model will be prepared for further predicting that which existing rewards program customers will likely to complete which type of offers next.

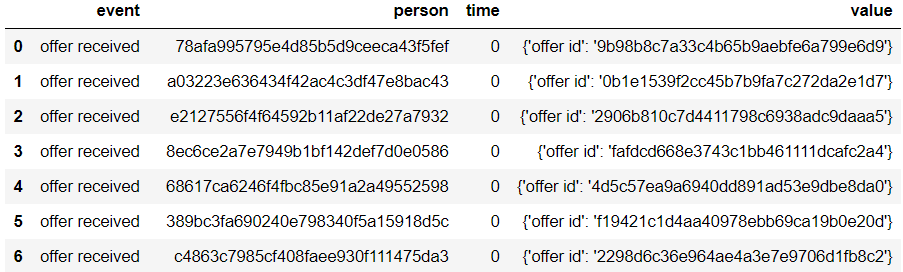
**Data Preprocessing Steps**

1. Profile dataset:

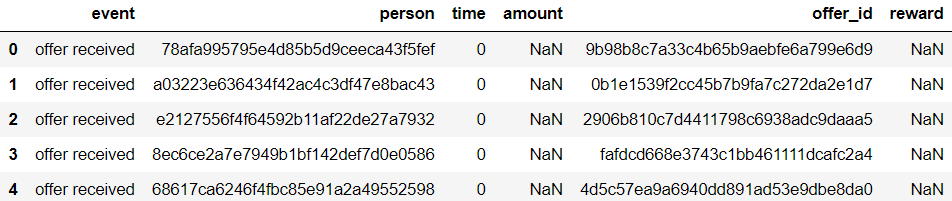
* A new age variable “age\_group” was created to bin customers age into 10 groups. These groups are 0-9, 10-19, 20-29, …, 50-59, etc.
* The variable “became\_member\_on” was used to extract its detailed information such as, “year”, “month”, “day” and “weekday”. As a result, this extracted information are created as separate columns.

1. Transcript dataset:

* Use the transcript dataset to join with itself after converting the structured data contained in the “value” column into dataframe.
* Updating the newly created “offer\_id” with the unpacked “offer id” value (previously contained in the “value” column) and its associated information about transaction and rewards amount. The following shows before and after processing:







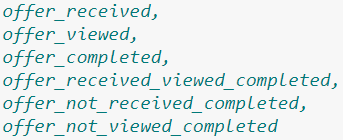
1. Portfolio dataset joined with processed transcript dataset:

* The portfolio dataset itself only contains information about offers (total of 10 offers). After joined with the transcript dataset, we obtained cumulated data that describes how well each offer performed in terms of number of times being completed, received, and viewed. As a result, variables “offer\_completed”, “offer\_received” and “offer\_viewed” were created.
* In addition, a new variable “influence” was created in order to determine the percentage of number of times each offer being viewed out of total number of times each offers are received.

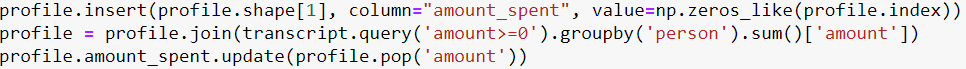


1. Profile dataset joined with processed transcript dataset:

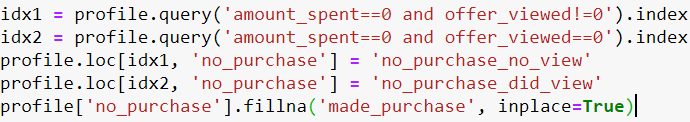
* In this part, we aim to bring in accumulated data, from the transcript dataset, for each customer that describes the following scenarios:



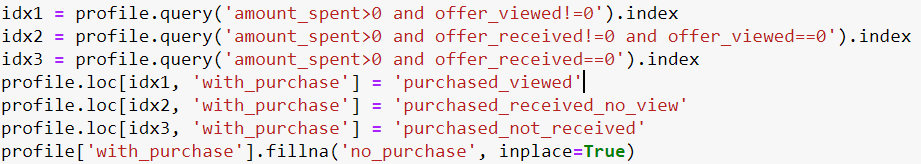
* Next, the cumulated spent amount was calculated and appended to the profile dataset as well.



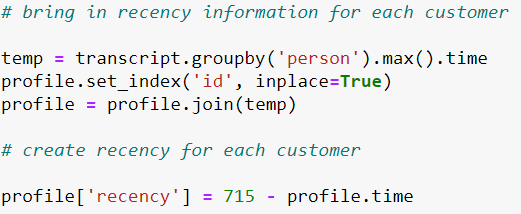
* Third, series of steps were made in order to identify customers who made no purchase and customers who made at least one purchase. As a result, two extra columns “no\_purchase” and “with\_purchase” were created.
  + Within the “no\_purchase” column, customers were further grouped into two sub-groups, one indicates no purchase because never received an offer and the other indicates no purchase after receiving the offers, regardless whether the offers are viewed or not.



* + Within the “with\_purchase” column, customers were further grouped into three sub-groups. The first sub-group represents customers who made single/multiple purchase after receiving the offers and viewed. The second sub-group represents customers who made single/multiple purchase after receiving the offers, but not viewed. The third sub-group represents customers who made single/multiple purchase even when no offers are being sent to them.



* Fourth, a new column “recency” was created by bringing in the time, from the transcript dataset, when a customer last interacts (receive, view, transaction, or complete) with an offer. It’s observed that the the maximum value of the "time" column is 714 which means the number of hours has passed, for that person to initiate certain action. Therefore, a decision was made to use 715 to represents the starting hour for a RFM analysis.



* Lastly, results from an RFM analysis and K-Prototype clustering were appended to the profile dataset named as “rfm\_score”, “rfm\_score\_sum”, “rfm\_segments”, and “clusters”. These variables will serve as engineered features to be included in model training which will be performed later.

**Engineered Feature Summary**

1. Profile dataset:

* “age\_group”
* “became\_member\_year”, “became\_member\_month”, “became\_member\_day”, “became\_member\_weekday”

1. Transcript dataset:

* “amount”, “offer\_id” and “reward”

1. Portfolio dataset joined with transcript dataset:

* “offer\_received”, “offer\_viewed”, “offer\_completed”
* “influence”

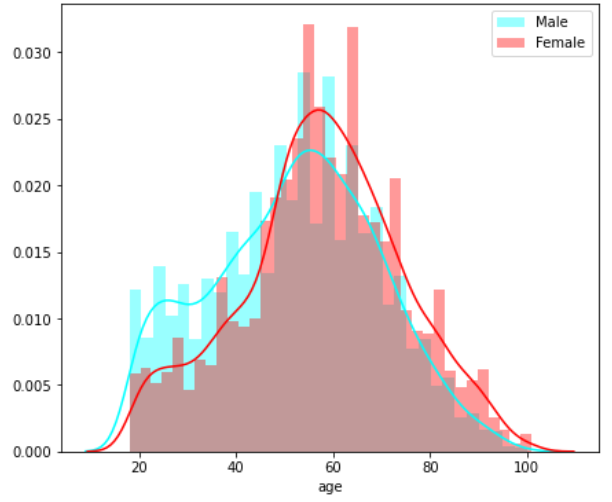
1. Profile dataset joined with transcript dataset:

* “offer\_received”, “offer\_viewed”, “offer\_completed”, “offer\_received\_viewed\_completed”, “offer\_not\_received\_completed”, “offer\_not\_viewed\_completed”
* “amount\_spent”, “reward\_received”, “no\_purchase”, “with\_purchase”
* “recency”, “rfm\_score”, “rfm\_score\_sum”, “rfm\_segments”, “clusters”

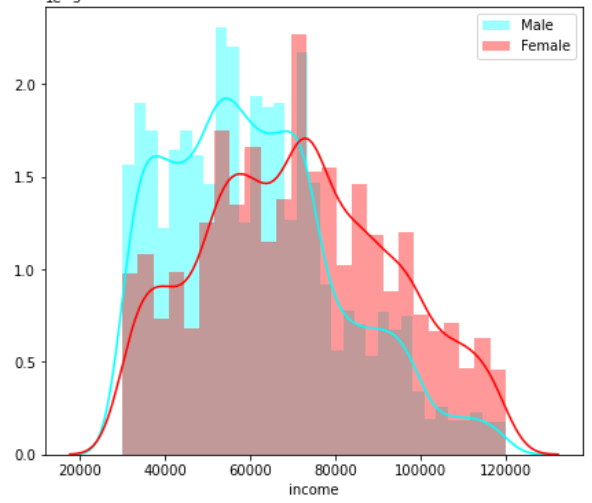
**Analysis**

**Descriptive Analysis**

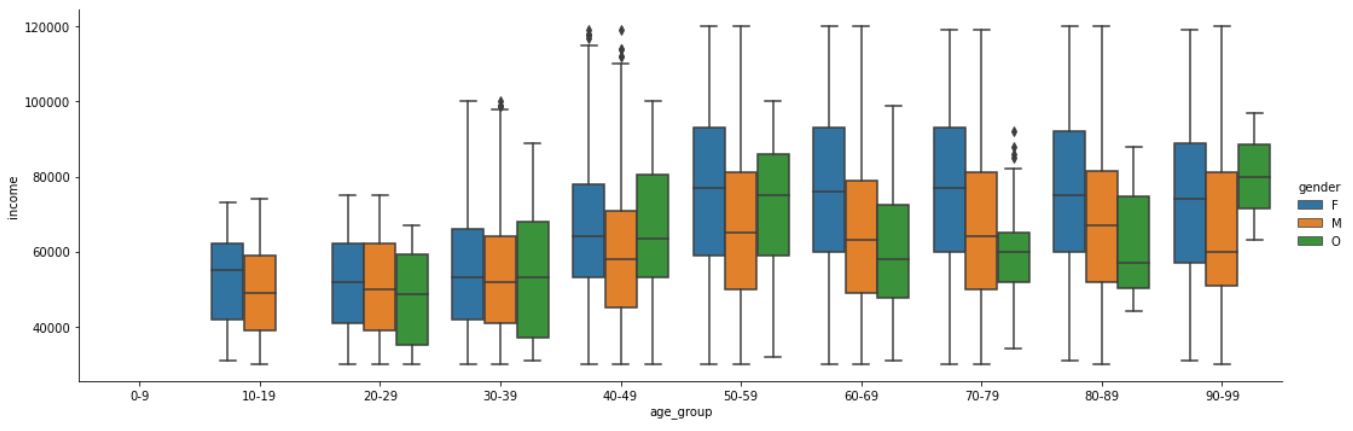
We will begin the analysis by taking a first glance of the profile dataset, we observe that the male population is 49.91%, and the female population is 36.05%. The rest 12.79% customers have missing gender information thus they are labeled as “O”. From customers’ age perspective, male and female population all followed normal bell curve that centers around the age of 60. It's observed that blow the age of 50, the male population is larger than the female population.



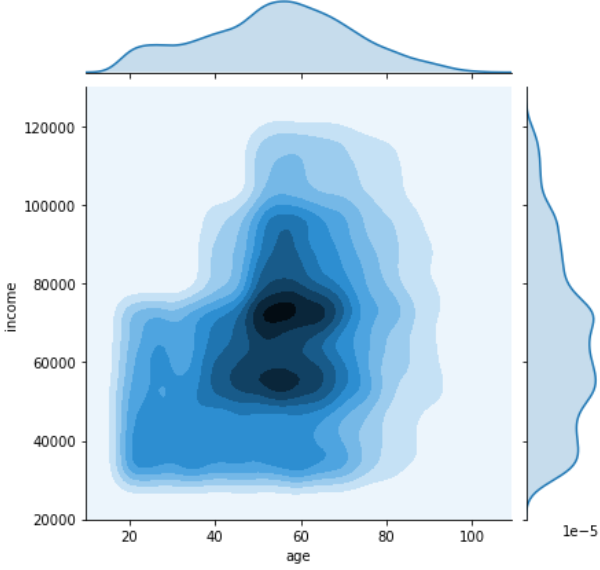
Looking at customers’ income distribution, there are a couple local peaks, which are about $58,000, and $75,000. It's observed that the male population is larger than the female population when the income level is below $70,000.



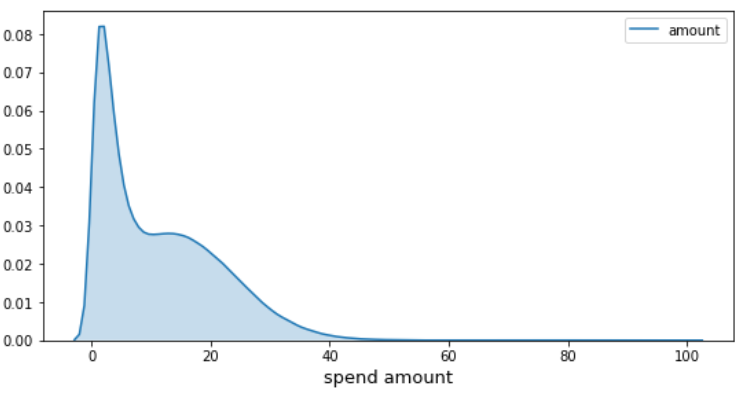
For the female population, the median income is higher than that of the male customers across all age groups.



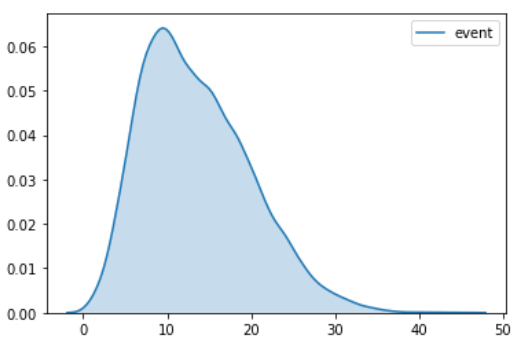
For both female and male, the most population has their age centers around age 50-70, and income centers just under $60,000 and slightly above $70,000. This is in sync with the previous observations.



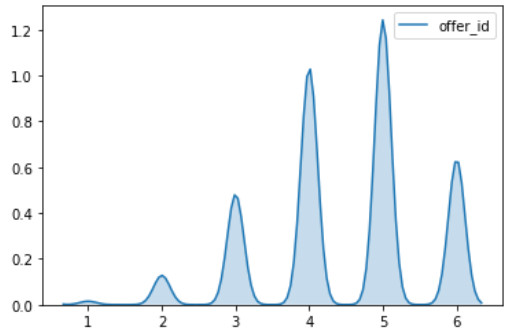
For a single transaction, there are more customers who spent less than about $10 dollars, and less customers who spend more than about $20 dollars. Please note that these numbers are the results of only looking at transactions one at a time. If a customer who spent $10 dollars today and $20 dollars tomorrow, these transactions will be viewed as two different transactions.



In terms of the type of events (viewed, purchased, and completed) that could happen to a customer, there are customers who initiated as many as 45 such events, and there are customers who initiated as less as 1 such event. However, the majority of the population initiate about 10 events, all types of events are considered except "offer received" since this is a passive action.

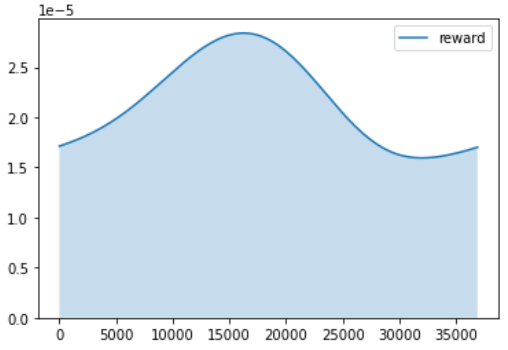


In terms of the number of offers received, the majority of the population receives either 4 or 5 offers during the data gathering period.

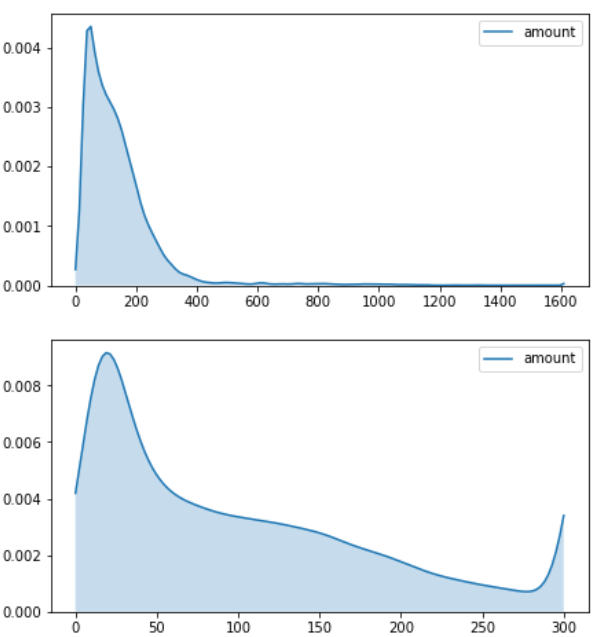


Offers that got completed the most times are "fafdcd668e3743c1bb461111dcafc2a4", and "2298d6c36e964ae4a3e7e9706d1fb8c2", with 5317 completions and 5156 completions respectively.

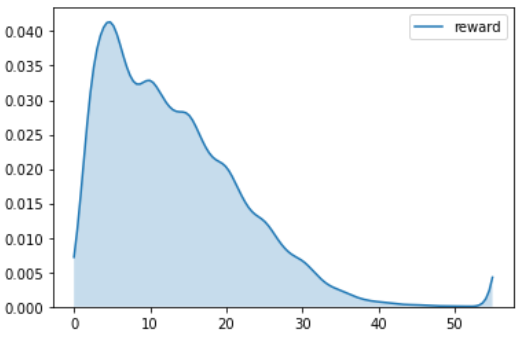
The two most rewarding offers are "ae264e3637204a6fb9bb56bc8210ddfd", and "4d5c57ea9a6940dd891ad53e9dbe8da0". They each rewards total of $36880 and $33310 to all participants. When it comes to total rewarding amount, most offers generate total rewards from around $15,000 to $20,000 to all customers.



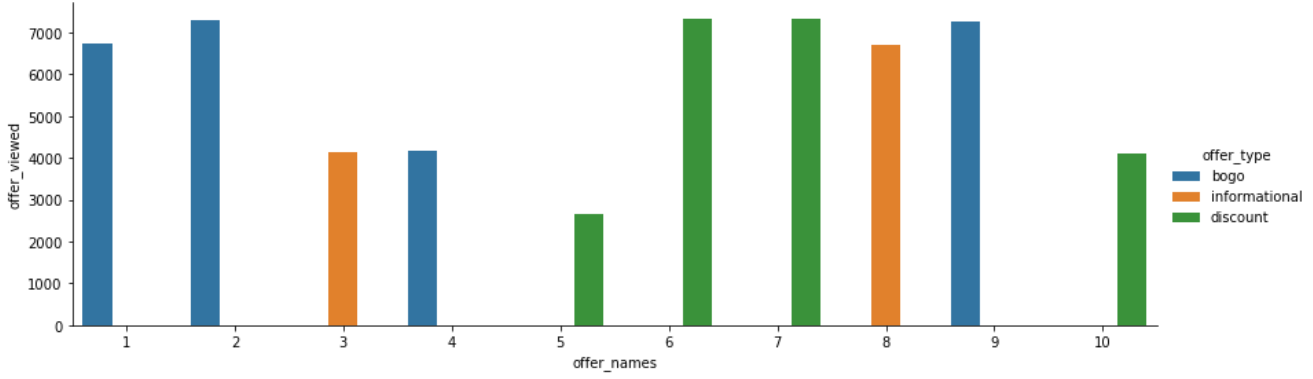
Looking at top 10 customers in terms of who made the largest total amount of purchase, their spent amount ranges from $1200 to $1600. However, the majority of the population spent less than about $50 dollars during the testing period. Once we pass the $50 benchmark, we see a clear decreasing trend in terms of purchasing amount that the participating customers made.



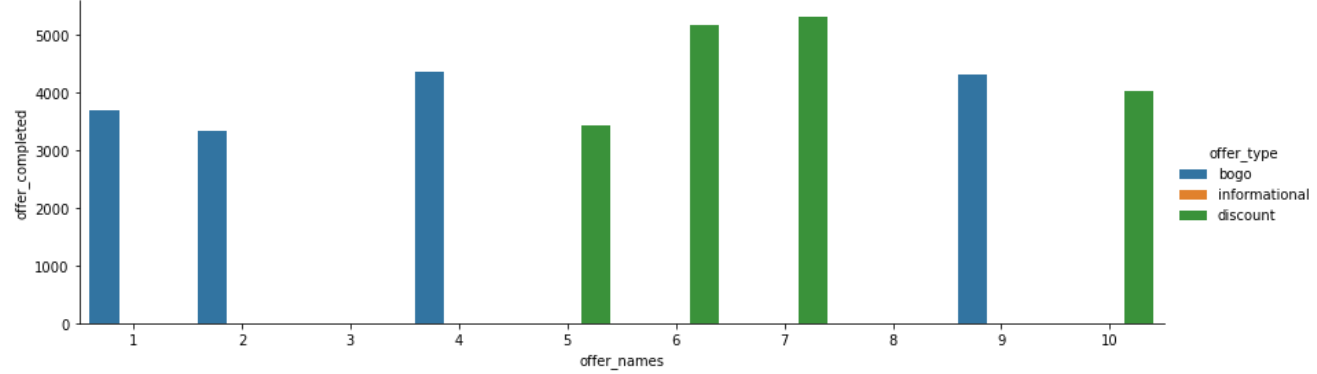
Looking at top 10 customers in terms of who receives the highest rewarding amount in total, most of them receives about $50 dollars. However, the majority of the population receives about less than $10 dollars. Once we pass the $10 benchmark, we see a clear decreasing trend in terms of rewarding amount given to the participating customers.

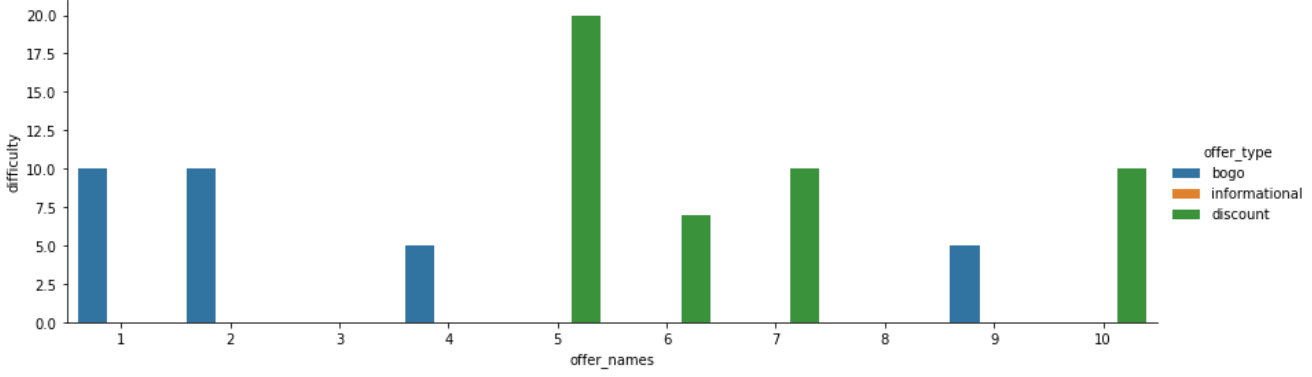


Looking at the joined dataset of portfolio and transcript, we ware able to determine that “bogo” offer type has the highest average number of views.

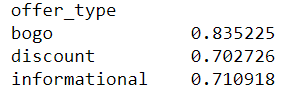


“discount” offers have the highest average number of completions, despite it has the highest difficulty level (averaged) to complete.



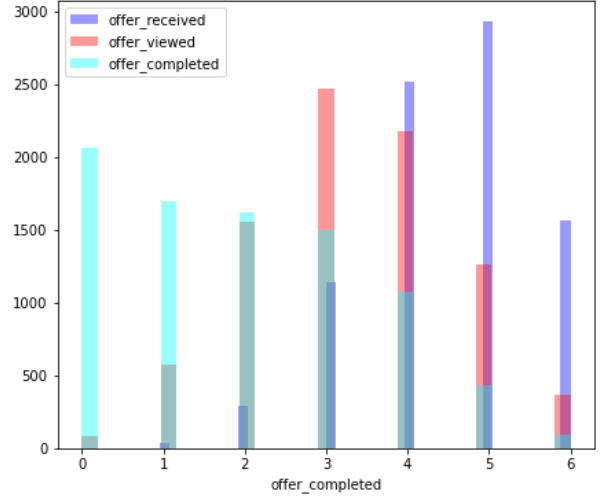


Using the number of times an offer is viewed divided by the number of times the same offer is received, we were able to calculate its level of influence. “bogo” offers has the highest level of influence among all offers.

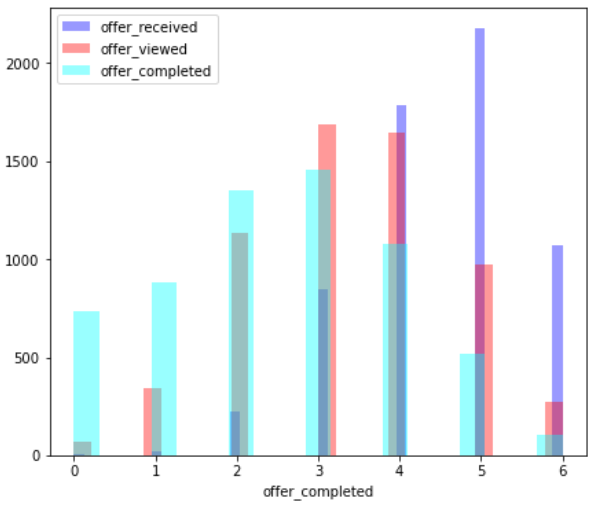


Next, we wanted to compare female and male customers purchase habits by join the profile dataset with transcript dataset, we observed that most of the male customers, they received about 4-6 offers while only view about 3-4 ones and complete from 0-3 offers. The female population has better statistics when comparing to the male population. Most female customers, they receive about 4-6 offers, and likely to view about 3-4 ones. But most female customers complete about 2-4 offers.

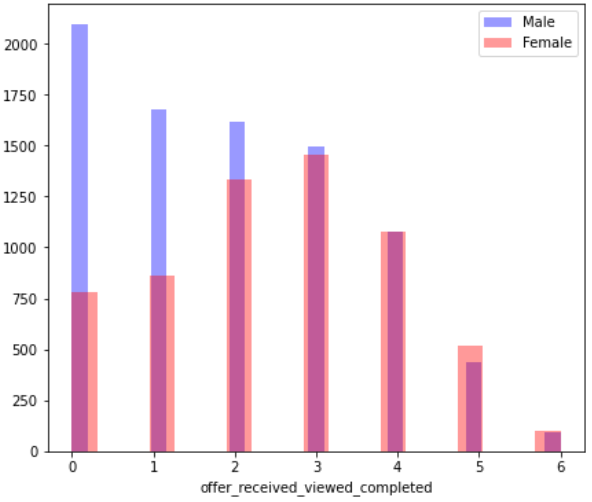
* Male customers:



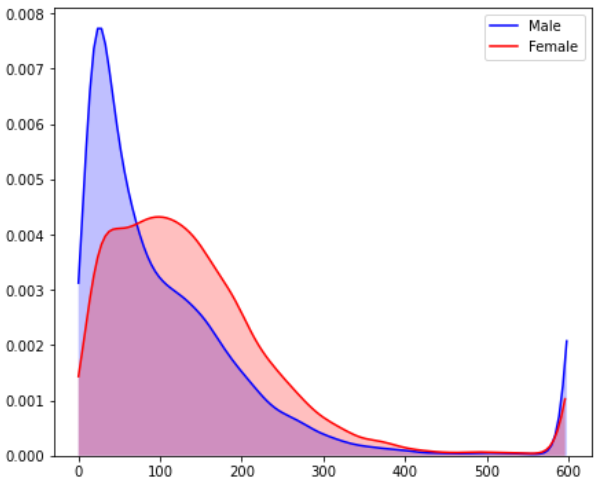
* Female customers:



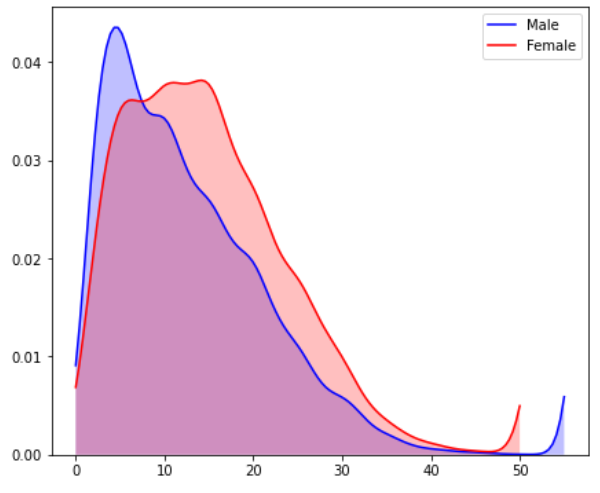
When considering the offers that got completed only after being received and viewed (thus followed all stages), male customers tend to complete about 0-3 such offers. While female customers, on the other hand, tend to complete from 2-4 of such offers.



It's observed that most of the female population has much higher total spent amount that centers around at $100 dollars, whereas for the male population, most of the total spent amount centers around at less than $50 dollars.



As a result, most of the female population receives more rewards than the male population does.



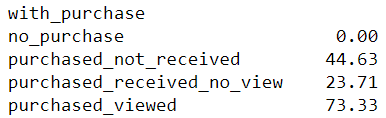
We also did customer profiling based whether a purchase is made or not. Based on the previously engineered features, a customer can be grouped into the following cases. Please note that customer profiling used in the descriptive analysis is not a formal attempt to segment customers. It is a method to identify whether there are rare purchase histories. For example, the number of customers who completed offers without viewing them.

* Customers with no purchase:
  + Customers who made no purchase because they never viewed any offer.
  + Customers who made no purchase after received and viewed the offers.
* Customers with purchase:
  + customers who made single/multiple purchase after receiving the offers and viewed.
  + Customers who made single/multiple purchase after receiving the offers, but not viewed.
  + Customers who made single/multiple purchase even when no offers are being sent to them.

Following the profiling logic listed above, we were able to determine that:

* There are 422 customers who made no purchases during the testing period.
  + 0 customers did not receive any offers.
  + 412 customers did not view any offers.
  + 10 customers viewed the offers but still did not make purchase.
* There are 16578 customers who made purchases during the testing period.
  + 16422 customers received offers and viewed them.
  + 150 customers received offers but did not view them.
  + 6 customers did not receive any offers.

Lastly, we examined how does "with\_purchase" column affect customers’ total spent amount.



Clearly, we see that the median purchase amount for those customers who received and viewed the offers, $73.33 dollars, is higher than the rest of the categories. So, customer who follow all stages, offer received, offer viewed and offer completed, make them valuable.

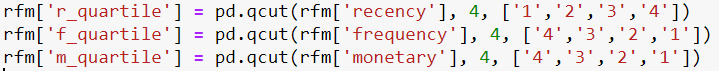
**Customer Segmentation**

There are two techniques being used in this component, one is RFM analysis and the other one is principle component analysis with K-Prototype clustering. We will proceed with the RFM analysis first.

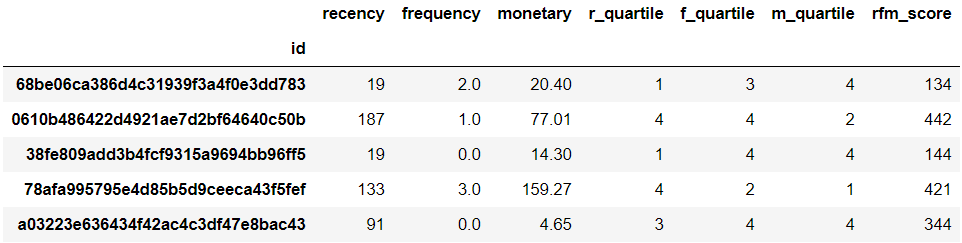
To begin with the RFM analysis, there are three required features we would need to obtain first. They are “monetary value”, “purchase recency”, and “purchase frequency”. For this reason, we take “amount\_spent”, “offer\_completed”, and “recency” features that were previously created to meet the requirements.

Please note that the “recency” feature was created based on the last time that a customer interacts (offer received, offer viewed, transaction, offer completed) with the offers. This time value is interpreted as the hours that have been passed since the beginning of the testing period and all rewards program customers start with “0”. We also observed that the maximum value of the "time" feature is 714 which means, for that person, that last action initiated was at the 714th hour since the testing period began. Therefore, an assumption is made that the starting hour for the RFM analysis to begin is at the 715th hour. This indicates that all time values will be subtracted from 715 and the results represent customers’ recency with certain offers.

Next, we will calculate the quantile of these RFM features. Customers with lowest recency, highest frequency and spent amount are considered as top customers. The following code as used to meet this description:



To calculate RFM results, we would simply concatenate the RFM quantile into one number to obtain the following:

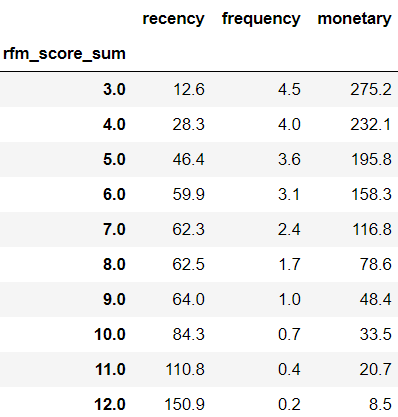


According to this result, the lower recency, frequency, and monetary quantile, the better score a customer will have. In other words, RFM score of 111 is better than 222, and score 222 is better than 224. The following lists top 5 RFM group representation, in terms of grouping size:

* 444, customers with high recency, low frequency and low monetary value.
  + Customers who don't interact with the offers they received, and they tend to complete less offers that worth low total purchase amount.
* 344, customers with relatively high recency, low frequency and low monetary value.
  + Customers who rarely interact with the offers they received, and they tend to complete less offers that worth low total purchase amount.
* 244, customers with medium recency, low frequency and low monetary value.
  + Customers who sometimes interact with the offers they received, but they tend to complete less offers that worth low total purchase amount.
* 111, customers with low recency, high frequency and high monetary value.
  + Customers who interact with the offers they receive often, and they tend to complete many offers that worth high total purchase amount.
* 144, customers with low recency, low frequency and low monetary value.
  + Customers who interact with the offers they receive often, but they tend to complete less offers that worth low total worth of purchase amount.

However, there is a problem with the stated method to calculate RFM results, and that is there are 60 of such resulted groupings. It is hard to interpret them all, so to reduce the high cardinality, we can simply sum up all three quantiles and get a single result that ranges from 3-12. This way, we obtain 10 unique grouping results, rather than 60. For example, our top customers with previous segmentation label "111", is now a 3, and customers with previous segmentation label "444" is a 12 now.

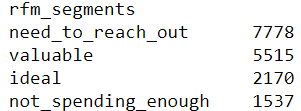




Using this new grouping method, we observed the following:

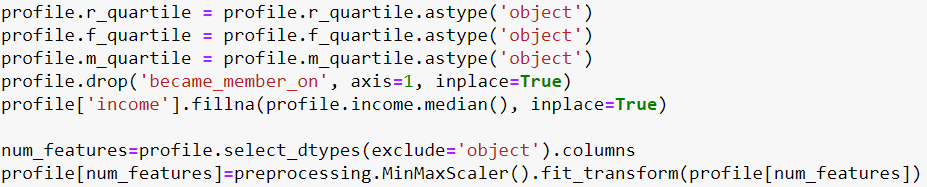
* Customers with new grouping score of 3, and 4 have the highest average monetary value, these customers should be labeled as “ideal”.
* Customers with new grouping score of 5-7 should be interpreted together because their average monetary value are all above $100 dollars. These customers should be labeled as “valuable”.
* Customers with new grouping score of 8-11 should be interpreted together because their average monetary value are all less $100 but still maintains 2-digit average spent. These customers should be labeled as “need to reach out”.
* Customer with new grouping score being 12 should be interpreted alone because their average monetary value is only $8.5 dollars. Thus, they are labeled as “not spending enough”.

In addition, using the table below we determined that the "ideal" customer group counted as 12.7% of the population, ranked as the third largest group among all FRM segmentations. The "valuable" customer group counted as 32.4% of the population and ranked the second largest group of the all RFM segmentations.



Now that an RFM analysis is conducted, we like to move on to the next part which is to use an unsupervised learning algorithm to automatically determine customer groupings for us. The learning method is called K-Prototype clustering. This clustering algorithm is used when the dataset contains categorical features. But first, we would need to reduce the number of dimensions that numerical features represent. This can be done through Principle Component Analysis.

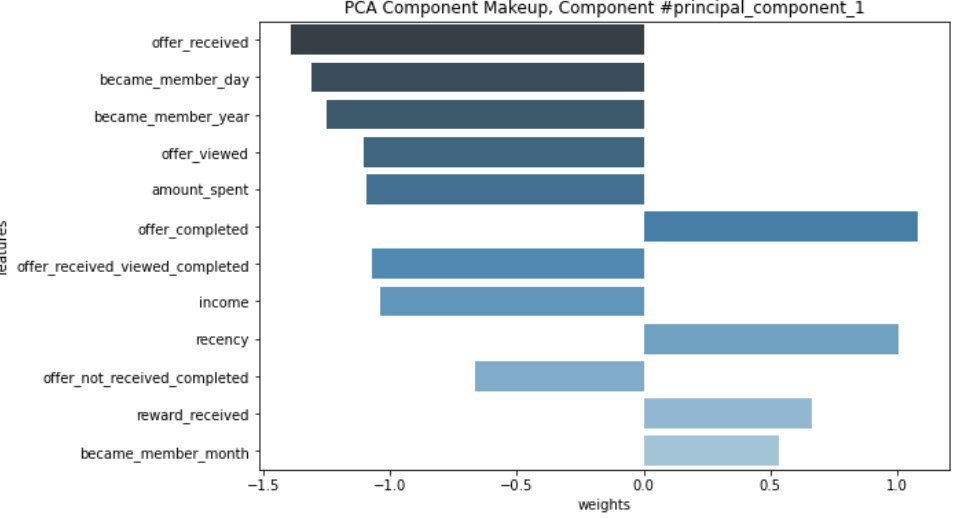
Before proceeding with PCA, numerical features first need to be scaled and filled in the missing values. The following code was used:



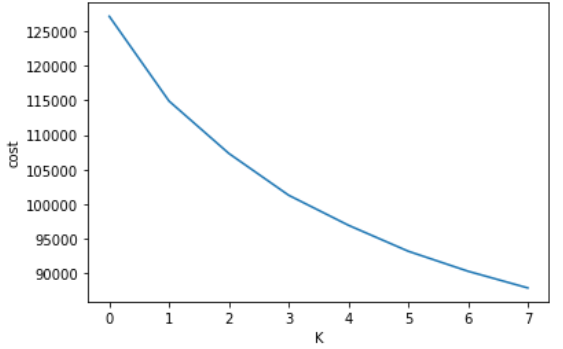
We also like to use the Hopkin Statistic to determine whether the dataset can be clustered. The Hopkins statistic tests the spatial randomness of the data and indicates the cluster tendency or how well the data can be clustered. If the value is around 0.5 or below, the data is uniformly distributed and hence it is unlikely to have significant results. If the value is above 0.5, it has a high tendency to cluster and therefore likely to have statistically significant results.

As a result, we obtained a Hopkins statistic of 0.78. This is a medium score regarding the potential grouping results that K-Prototype will yield. Due to the Hopkins statistic obtain, I remain conservative about whether there will be distinct groupings found by the K-Prototype algorithm. Nonetheless, I still think that the unsupervised learning algorithm will bring value because its ability to find attribute contribution to some of the clusters defined by the algorithm.

After PCA, we observed that the first 7 components contribute about 88% of the total variance explained. Thus, we will only include the first 7 components to the K-Prototype clustering, along with the rest of the categorical features. Taking the first principle component as an example, we observed that features such as "offer\_received", "became\_member\_day", "offer\_viewed" and “amount\_spent” have the highest weight. In other words, this component intends to transform each customer's record based on emphasizing the features described above.



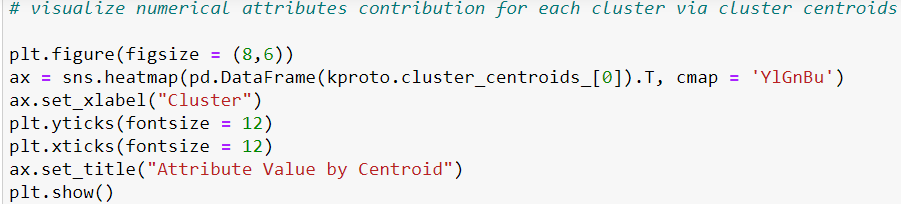
Now, we are ready to perform K-Prototype clustering. Using the “elbow” method, it is determined that there are no natural distinct groupings found in the data. Thus, to save computation resources, 4 is chosen as the number of clusters to form.

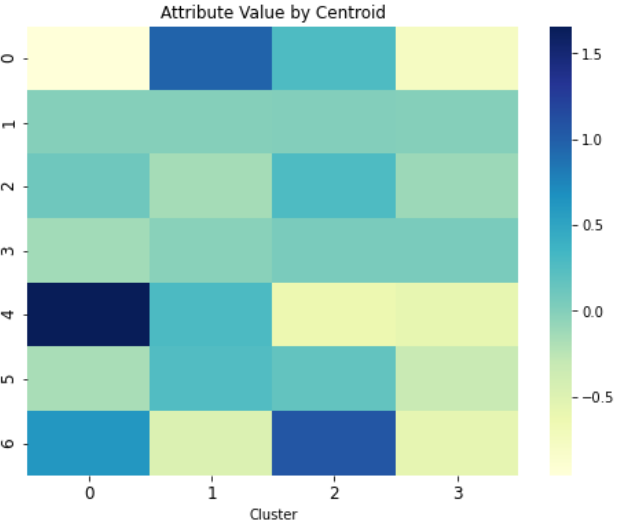


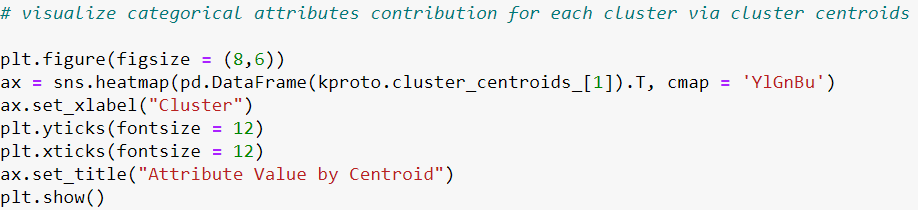
Upon conducting the clustering, the following observations were made:

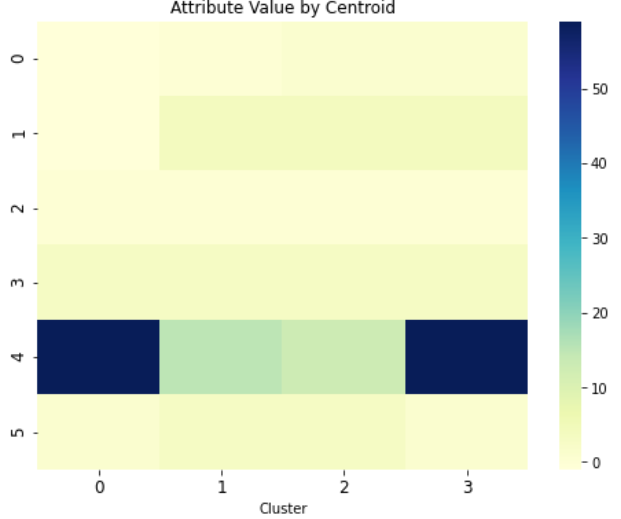
* Cluster “0” centroid is placed mostly within principle component 5, 7 and categorical feature "rfm\_score".
* Cluster “1” centroid is placed mostly within principle component 1.
* Cluster “2” centroid is placed mostly within principle component 7.
* Cluster “3” centroid is placed mostly within categorical feature "rfm\_score".











**Supervised Learning**

Now that we have had a basic understanding of the customers through descriptive analysis and have determined the characteristics of different grouped customers via RFM analysis as well as K-Prototype clustering, we are ready to answer first key question stated in the previous sections of this report, “Which type of offers (informational, discount or bogo offers), a given customer is most likely to complete?”. To proceed, we will need to train a machine learning model. The specific steps involved in model training are data labeling, baseline model training, model improvements, and final model assessment.

Based on the nature of our stated question, it is decided that our model learning objective would be predicting multi-class labels. Thus, the following are used in the process of data labeling:

* A customer with a "info\_or\_promo" 0 indicates the following scenarios:
  + This customer has not received any offers at all, hence labeled as unsuccessful.
  + This customer has received informational offers (at least one) but failed to view them. Hence labeled as unsuccessful.
  + This customer has received promotion offers (at least one) but failed to complete them. Hence labeled as unsuccessful.
* A customer with a "info\_or\_promo" 1 indicates this customer has received and viewed at least one informational offers.
* A customer with a "info\_or\_promo" 2 indicates this customer has received and completed at least one discount offers.
* A customer with a "info\_or\_promo" 3 indicates this customer has received and completed at least one bogo offers.

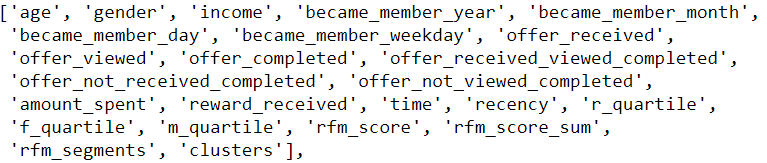
Please note that success, for an informational offer, means viewing the offer; for a promotion offer, success means complete the offer. In addition, some customers may receive the same offer multiple times and they may or may not complete all of them. For this case, we only want to count when the customer completes such offer at least once. This is a modeling decision because when customer complete the same offer multiple times, it introduces duplicates to the dataset. And when the customer does not complete all of the same offers, it introduces noise to the dataset.

One last step before model training is final feature preparation. Since both “no\_purchase” and “with\_purchase” (as results of customer profiling from the descriptive analysis) represent cumulated results that describes individual customers. If we were to include these columns, then it wouldn't make sense because our label column is based on each purchase register. So, the “no\_purchase” and “with\_purchase” columns would contain values that are conflicts with the associated labeling values.

We also need to drop the columns that describe characteristics of each offers because we want to intentionally train a model that learn off of customers' characteristics and not to rely on the information about offer themselves. That said, we will not be joining with the portfolio dataset.

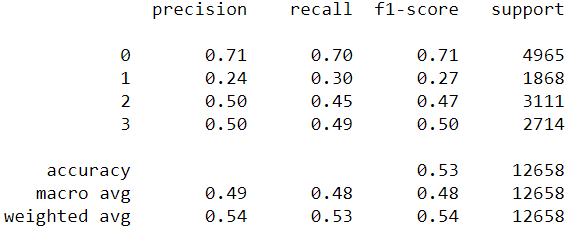
“rfm\_segments” and clustering results will be kept because we assume that offers will be sending to the customers who are already members of the rewards program with some purchasing history. In other words, we assume that we wouldn't send to those who just joined our program with little or no purchasing history. The same engineering steps (rfm analysis, k-prototype clustering) for the unseen customers can be replicated in the same way that the trained model can utilize.

As a result, the training set will include the following features:



And the training target will be, “info\_or\_promo”.

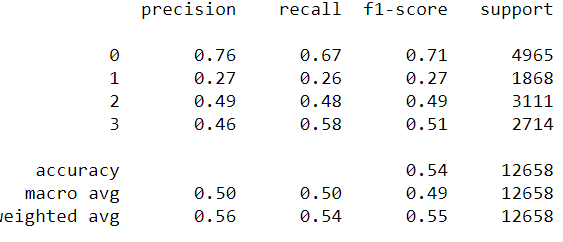
As the model training begin, we first train a logistic regression model that serves as the baseline model. The “class\_weight” option was set as “balanced” in order to counter class imbalance. As a result, it’s observed that the weighted average f1 score as 0.54 which is only slightly better than a dummy classifier.



It’s also observed that the logistic regression does not converge.



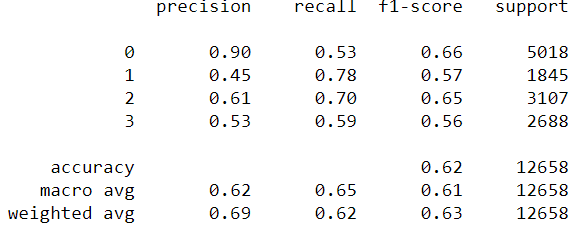
For this reason, I will train a lightGBM model with the option of “early\_stopping”. However, the results did not show any improvements even the lightGBM model is considered as a potentially better model in many cases.



The lightGBM model reached its best iteration before overfit. We can conclude that the current train set does not contain meaningful information that allows a model to rely on in order to make correction predictions. To solve this problem, I will add additional information, about characteristics of offers themselves, from the portfolio dataset.

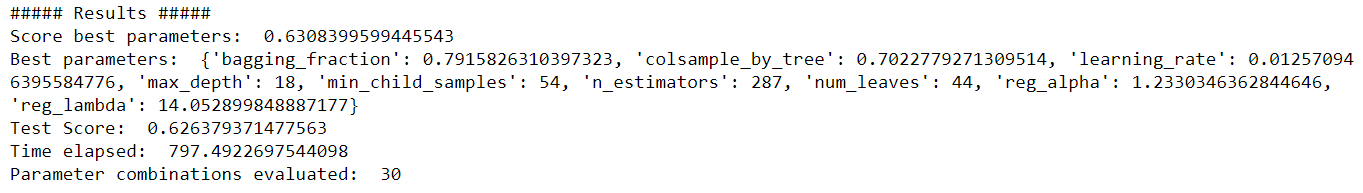
Please note, that this move certainly will improve the modeling score, but we also take the risk of information leakage since different types of offers have fix set of channels, difficulty, and duration combinations. So, it is my responsibility to find the right balance between reveal too much information and not including enough of information. For this reason, I will only select the channels feature from the portfolio dataset.

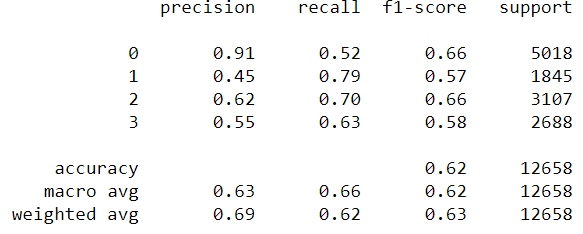
After joined the training set with the portfolio dataset, we added one additional field which is “channels”. The following results were obtained:



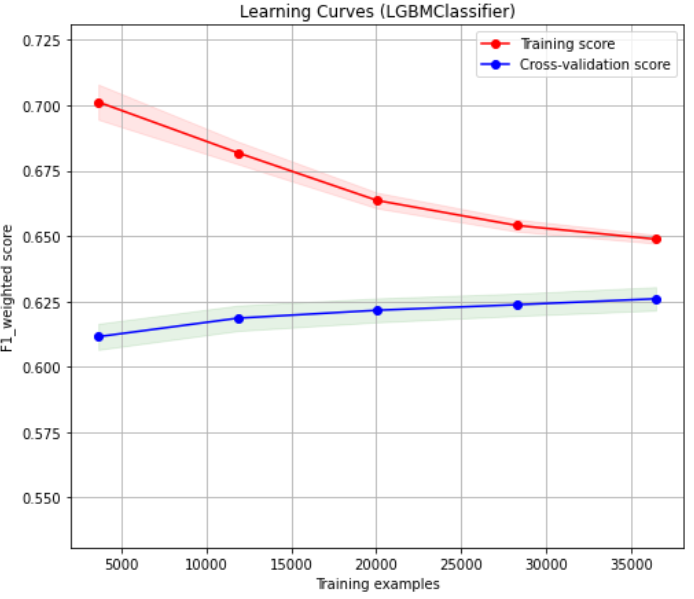
It's observed that the modeling score, in terms of weighted avg f1, has been improved from 0.55 to 0.63. Including the new feature "channels" from the portfolio dataset is the maximum level of tolerance I have towards training features. Next, I will tune the model and use its learning curve to determine whether the model has reached its potential.

Based on Bayesian Optimization method, the following tuning results were produced:





Hyperparameter did not improve model's performance. This indicates that there is a fix amount of information can be learned by the model and that information mainly comes from the “channels” column. Next, we will take a final assessment of the tuned model using learning curve.



We observe that when the training size is small, the gap test learning curve and train learning curve is large. This is expected as the model could just memorize all the answers given that there are only a few training samples. However, as the training size became larger and larger, the gap between train learning curve and test learning curve narrows down. This trend indicates that the model became more and more stable as more training samples are introduced. In addition, there is no sign of overfitting, and it is very likely that the training error and test error came across at around 0.63.

**Cost Simulation**

Now that the first key question (which type of offers to send) is answered, we can move onto the second key question, "which particular offer would help business achieve highest revenue and lowest cost?". A common approach to solve this problem would be starting from a train model and analyze its error rates such as recall, and precision. By adjusting the probability threshold (only applicable in a binary problem), we would be able to observe how recall and precisions behaves accordingly. Then by applying the cost assumptions to various recall/precision, we would be able to obtain a cost function that allows us to determine at what probability threshold and recall/precision level, the cost of modeling errors could be minimized.

However, since our modeling objective is multi-class prediction by design, the concept of adjusting the probability threshold will not be applicable in our case. Thus, I purposely choose to do a cost simulation instead of a cost analysis. The mechanism that drives the simulation is that by adjusting predicted probabilities of each class, we would obtain various level of simulated modeling error rates for which we could then apply to the cost assumptions. As a result, different cost will be shown in addition to their associated error rates and other modeling performance metrics. We would then be able to pick a cost that's acceptable. In the subsequent model improvement works, we can aim to achieve that observed modeling performance, knowing that the cost would be minimized as a result of this.

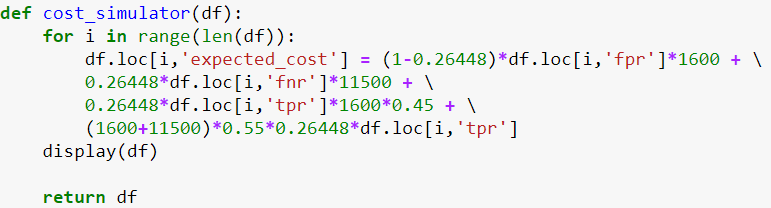
Because of the decision made above, I will rephrase the second question which is stated in the beginning of this analysis. Thus, instead of answering "which particular offer would help business achieve highest revenue and lowest cost?", I will use the following cost simulation to answer "given the current state of the model, what improvements can be made in regards to modeling performance for each class label, to achieve a minimized cost after applying the cost assumptions?".

Please note that it's because predicted class probabilities are the main driver for various level of prevision and recall, we choose it as the sole input of the simulator function. The current model performance is rather poor. This is another reason I decided to do simulations because it represents hypothetical cases to show a minimized cost that can be potentially achieved if model performance is improved to a certain level.

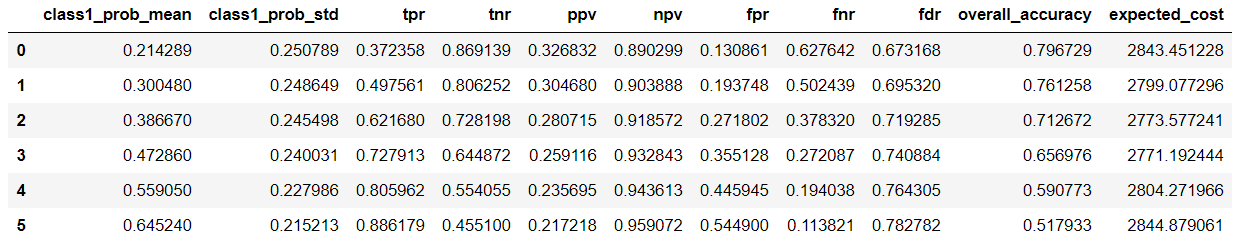
The following cost assumptions were used in creating the cost function:

* Based on the entire customer population, which is unknown to us, the percentage of rewards program members who remain to be in-active (at-risk) is about 26.448%.
* The “bogo” and “discount” offers cost the same, $1600 dollars per customer (one-time).
* If a potentially interested customer is lost, then it will cost the business $11500 dollars per customer (one-time).
* The “bogo” and “discount” offers are assumed to be 45% effective on average. That is 45% of the rewards program members who are identified as potential participants and they have received the “bogo” and “discount” offers, actually decide to participate and make purchasing decisions. The remaining 55% will not participate anyway.

The cost function is created as the following:

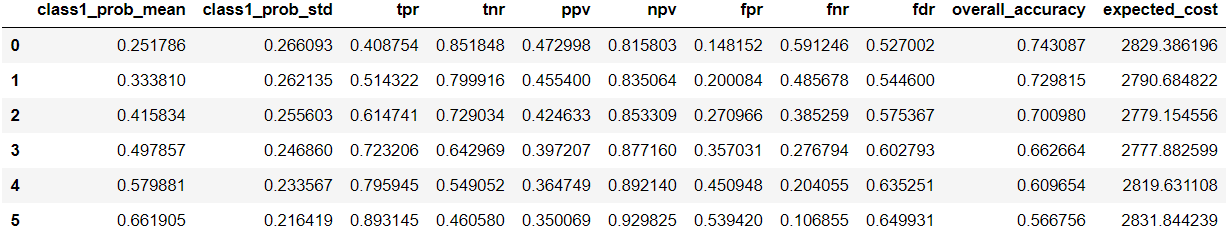


Simulation results for class 1 predictions:



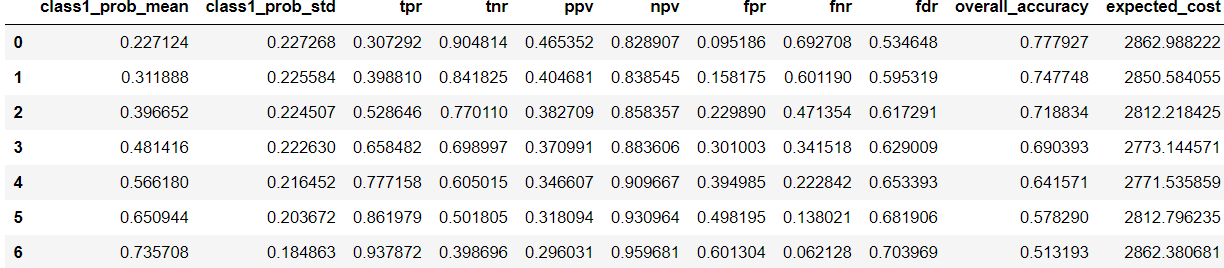
The final lightGBM classifer after tuned has a class 1 recall rate at 0.79. We observe from the simulation results, an optimal cost is achieved when class 1 recall rate is at 0.72. At first, this observation might seem contradicting to what our goal would be, which is to improve recall rate for all class predictions. However, this observation can be interpreted as a potentially better model could sacrifice its performance in class 1 predictions in order for it to gain performance in other class predictions. This is a business decision to be made. As the simulated results suggest, a recall at 0.72 would achieve an expected cost of $2771 dollars per customer.

Simulation results for class 2 predictions:



As the simulated results show, an optimal cost is achieved when the recall for class 2 predictions is at 0.72. Our final model's recall, for class 2 predictions, is 0.70. So, we interpret this as the current final model is already at its optimal state in regard to class 2 predictions. Using the stated cost assumption, it is reasonable to conclude that our final model has a expected cost of $2777 per customer.

Simulation results for class 3 predictions:



Our final currently has a recall of 0.60 for class 3 predictions. However, as the simulation results suggest, modeling performance for class 3 predictions would our priority in the next iterations of modeling improvement works. If we were able to achieve 0.77 for the class 3 recall, then the minimized expected cost would be achieved at $2771 dollars per customer.

**Results**

**Descriptive Analysis Findings**

Upon analyzing the provided three datasets (profile, transcript, portfolio), we were able to conduct the following findings:

1. Gender and age

There are 17000 unique customers. While half of them are male, about 36% are female and the rest are labeled as NAN in the gender column. Both male and female population are around 55-60 years old. However, the male population is moderately younger than the female population due to the fact that there are more males under the age of 50 than that of the females.

1. Income

It's observed that there are more males who make less than about $70,000 than that of females, while the opposite is true when considering people who make more than $70,000. In addition, the median income of the female group is higher than that of the male's across all age groups. When pass the age of 50, it is shown that female tend to have a higher income than male do by a large margin.

It's concluded that the female group while may be moderately older than the male group, they are very likely to have a higher income than the male customers.

1. Customer profiling

For customers who made no purchase (412 people), the majority did receive offers, but never viewed them.

For customer who made at least one purchase, the majority followed all stages that include offer received, viewed, then made the purchase. This sequence of actions is really valuable because the median purchase amount for profiled customers is higher than the rest of the profiled groups. There is only a small portion of the customers (150 people) who made purchase but did not view the offers they received.

1. Offer completion rate

All offers got received about 7000 times for the entire customer population. Offer completion rates (informational offers are excluded since they don't generate revenue):

* 9b98b8c7a33c4b65b9aebfe6a799e6d9: 56%
* 0b1e1539f2cc45b7b9fa7c272da2e1d7: 44%
* ae264e3637204a6fb9bb56bc8210ddfd: 48%
* 2298d6c36e964ae4a3e7e9706d1fb8c2: 67%
* 2906b810c7d4411798c6938adc9daaa5: 52%
* fafdcd668e3743c1bb461111dcafc2a4: 70%
* 4d5c57ea9a6940dd891ad53e9dbe8da0: 44%
* f19421c1d4aa40978ebb69ca19b0e20d: 57%

1. Offers nature

Offers "fafdcd668e3743c1bb461111dcafc2a4" and "2298d6c36e964ae4a3e7e9706d1fb8c2" are the most successful offers in terms of having the greatest number of completions. However, they are also the ones with lower rewarding amounts generated to all participants. This could potentially due to that these offers have lower difficulty level required to complete, thus they are less rewarding by design.

Conversely, offer "4d5c57ea9a6940dd891ad53e9dbe8da0", and "ae264e3637204a6fb9bb56bc8210ddfd" got completed with the fewest number of times, but they generated the highest amount of rewards. Intuitively thinking, these offers might have a higher difficulty level to complete, thus highly rewarding.

The "bogo" offers have the highest average number of views, so as a result, this offer type also has the highest degree of influence. The "discount" offers took the longest to complete, but their averaged number of completions is also the highest.

1. Total transaction made/rewards received by person

The majority of the population spent less than about $50 dollars during the testing period. The majority of the population receives about less than $10 dollars.

1. Female customers

It's observed that most of the female population has much higher total spent amount that centers around at $100 dollars, whereas for the male population, most of the total spent amount centers around at less than $50 dollars. As a result, most of the female population receives more rewards than the male population does. We also observe that female customers receive and view about the same amount of offers, but they tend to complete more offers than the male customers do.

**Customer Segmentation Findings**

1. RFM analysis

We grouped customers into four distinct groups based on their spent amount, offer interaction recency, and purchase frequency. These four groups are, “ideal”, “valuable”, “need\_to\_reach\_out” and “not\_spending\_enough”. The "ideal" customer group made up 12.7% of the entire population, while the "valuable" customer group counts 32.4% of the population.

Among these two preferred groups, customers are expected to have higher total purchase amount, high number of completed offers, and their last interaction with the offers they receive is very close to the time that this analysis is performed.

1. K-Prototype clustering

Features that are heavily relied on by all the clusters are: principle component 1, 5, 7 and along with a categorical feature "rfm\_score" (the rfm segmentation results obtained from previous section).

Principle component 1 major features:

* “offer\_received”, “become\_member\_year”, “become\_member\_day”, “offer\_viewed”

Principle component 5 major features:

* “offer\_received”, “age”, “time” and “offer\_not\_viewed\_completed”

Principle component 7 major features:

* “time”, “income”, “offer\_not\_viewed\_completed”, “rfm\_score\_sum”

**Supervised Learning Results**

We conclude that a trained lightGBM classifier model eventually will achieve 0.63 as the final testing f1 weighted score. This score has been improved from 0.55 when the channels column was not included. Since we only added the channels column as an extra feature and no sign of improvements are observed during parameter tuning, we conclude that the testing score improvements is mainly caused by the channels column, and there is only a fixed amount of information that a model can learn off of. This conclusion can be proven by observing from model's curve.

This scoring result is rather poor, and it implies that the model should not be used in production until either more meaningful features are added, or data labeling logics are refined.

**Cost Simulation Results**

Our current final model has a weighted avg f1 score of 0.63. By conducting simulations of various probabilities for each class, we were able to conclude that the area that future improvements is most needed is class 3 results. Specifically, we would like to achieve a 0.77 for class 3 recall.

**Conclusion**

**Problem Statements Review**

1. What are the success rates for each offer?
2. What statements can we make when comparing the “bogo” and “discount” offers?
3. When customers make certain decisions about the offers, are there any common characteristics that these customers share?
4. How does the customer groups labeled as ideal and valuable react to each offer types?
5. Should the marketing team send offer to a particular rewards program customer?
6. Given the current state of the model, what improvements can be made in regard to modeling performance for each class label, to achieve a minimized cost after applying the cost assumptions?

**Conclusions to each Stated Problems**

1. All offers got received about 7000 times for the entire customer population. However, offer "fafdcd668e3743c1bb461111dcafc2a4" and "2298d6c36e964ae4a3e7e9706d1fb8c2" are the most successful ones in terms of offer completion rates with 70% and 67% respectively. They are both “discount” offers
2. The "bogo" offers have the highest average number of views, so as a result, this offer type also has the highest degree of influence. The "discount" offers took the longest to complete, but their averaged number of completions is the highest.
3. Female customers receive and view about the same amount of offers, but they tend to complete more offers than the male customers do. We also observe that the female population has much higher individual total spent amount that centers around at $100 dollars; whereas, the male population's most frequent total spent amount centers at less than $40-$50 dollars.
   1. However, we also observe that within the customers who made no purchases, the majority comes from female. Female customers have higher median income than male customers do. If the offers we sent are mainly for informational purposes, that's fine. But, we should be focusing on sending promotion offers such as "bogo" and "discounts” because we've observed that this particular group of customers has higher total spent amount and they are likely to complete more offers than the male customers across all ages and income levels.
4. For customers who did not make any purchase, most of them did receive the offers, but never viewed them. We should design offers that are likely to be viewed.
   1. For customers who made at least one purchase, most of them followed all stages which include receive and view the offer first, then complete the offer. This sequence of actions is desired because the median purchase amount, for customers who followed all stages, is higher than the rest of the scenarios. So, customer that follow all stages make them valuable.
5. Using the RFM segmentation method, it is observed that customers with the segment score of 3 and 4 (these are the lowest possible scores) have the highest average monetary value, these customers should be labeled as most ideal to the business. Customers from this group commonly spend above $200 dollars on average during the entire testing period.
   1. The ideal customer group counted as 12.7% of the population, ranked as the third largest group among all customers.
6. Customers with RFM score from 5 to 7 should be interpreted together because their average monetary value are all above $100 dollars. Thus, these customers are labeled as "valuable" to the business.
   1. The "valuable" customer group counted as 32.4% of the population and ranked the second largest group of the all segmentations.
7. Among these two preferred groups, customers are expected to have higher average purchase amount, high number of completed offers, and their last interaction with the offers they receive is very close to the time that this analysis is performed.
8. Customers with RFM score from 8 to 11 should be interpreted together because their average monetary value are all less than $100 dollars but still maintaining 2-digit average spent amount. We assume that the business like to encourage these customers, therefore, this segment is labeled as "need to reach out".
9. Discount offers are the most popular ones among customers who are labeled as "ideal" and "valuable". In addition, the same groups of customers most likely come from cluster 1.
10. To the second last question, our original intent was to train/tune a machine learning model and let it predict which offer types a given customer would most likely to complete. Then the business could plan to use these prediction results. However, even after taking the risk of information leakage by adding the "channels" column, our final model still produces inferior results.
11. Based on this condition and the learning objective (multi-class) of our model, I switched from conducting a cost analysis to a cost simulation, in order to answer question 3. As a result, we found the area that future improvements are most needed is class 3 results. Specifically, we would like to achieve a 0.77 for class 3 recall in order to minimize the cost, given all other class predictions does not change.