Capstone Project
Data Science Nanodegree

Gerard Sho
August 18, 2022

Definition

Project Overview

This project is to mainly assess the effectiveness of a marketing campaign deployed by Starbucks towards different demographic groups of customers. Starbucks sends out offers of different natures to users of their mobile app, Starbucks rewards. These offers could be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). To begin, the strategy for now is to send randomly to different users at different time. Some users might not receive any offer and some may be receiving the same offer more than once. A customer might get one of the following:

- Informational offer (mere advertisement)
- Discount offer
- Buy one get one free (BOGO) offer

Discount and BOGO offers have a challenge, that is, the customer must make a minimum purchase before it can redeem the offer. Additionally, each offer has an expiration date. In the case of the informational offers, the expiration date is when the customer stop feeling the influence of the advertisement. As the campaign will eventually have to be refined; for instance it will be a waste of resources to also send the same promotional offers to groups who are likely to buy from Starbucks anyway, as opposed to those who are constantly looking for deals before making purchases. The main goal of the project is hence, not just to determine the segment of customers most likely to have viewed an offer but also to have completed the purchase.

Problem Statement

Data are collected from the Starbucks customers test groups who have received the offers over a period of time. These data includes the characteristics of the offer types (portfolio.json file) activity logs and transaction data (transcript.json file as well as some of the demographic details of the customers - such as gender, age, income and the date when a person became a member of the rewards system - captured in the profile.json file. The data are anonymized and each customer is only

identified via a unique id. The activity logs of the customers would tell if an offer has been received; viewed and completed and the transactions completed by each user id. The goal of sending advertisement and offers to customers is to increase the customer overall purchases. However, it would be wasteful to send all offers to all customers at the same time. Looking at the past transactions and demographics data, that may help Starbucks to target the offers more narrowly towards different groups of customers.

Therefore, to refine the marketing campaign from the data, here are the main questions that would like to be answered via the project:

- Which segments of customers are more likely to respond to an offer ?
- And, is there any incremental impact from the marketing campaign ?

A machine learning model i.e. a classifier, would be trained on the cleaned and wrangled data set. The trained model could inform of the features of customers who are likely to contribute to sales.

Metrics

As this will be treated as a classification problem, the common use of classification metrics such as the follows applies to evaluate the models and to decide on the outcomes:

- Accuracy - The proportion of correct predictions out of all the observations.
- Precision - The proportion of positive cases that were correctly identified.
- Recall - The proportion of actual positive cases which are correctly identified.
- F1 - score, that combines the two previous measures.

Mainly, the accuracy score is found to be sufficient for use here as the binary outcomes from result is not thought “positive” or “negative” but rather each outcome would contribute to answering who should the offers be sent to.

Analysis

Data Exploration

The data is contained in three files:

* portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
* profile.json - demographic data for each customer
* transcript.json - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

portfolio.json
* 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 offer to be open, in days
* channels (list of strings)

profile.json
* age (int) - age of the customer
* became_member_on (int) - date when customer created an app account
* gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
* id (str) - customer id
* income (float) - customer's income

transcript.json
* event (str) - record description (i.e. transaction, offer received, offer viewed, etc.)
* person (str) - customer id
* time (int) - time in hours since start of test. The data begins at time t=0
* value - (dict of strings) - either an offer id or transaction amount depending on the record

```
reward        0      gender              2175      person    0
channels      0      age                    0      event     0
difficulty    0      id                     0      value     0
duration      0      became_member_on       0      time      0
offer_type    0      income              2175      dtype: int64
id            0      dtype: int64
dtype: int64
```

Figure 1 : null values found in each of the portfolio.json, profile.json, transcript.json

Upon checking for null values in each data sets, it is found that there are missing information on the gender and income for 2,175 customers. In addition, the distribution of some of demographic data of the customers can also be better understood in visualisation form.

Data Visualisation

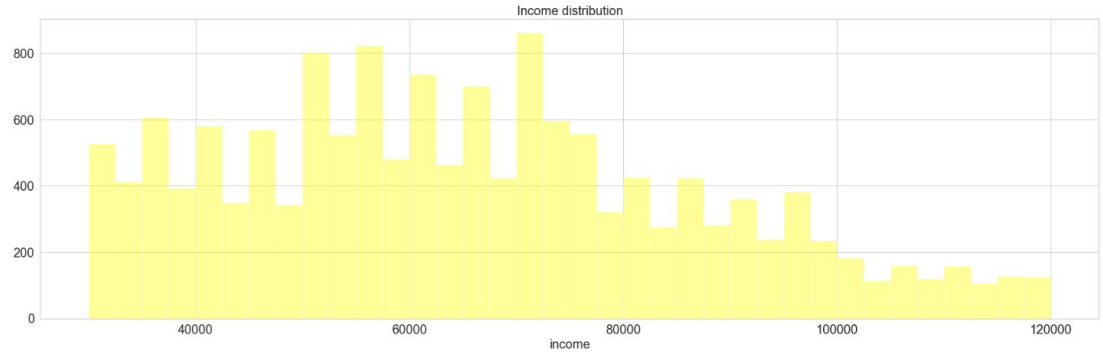

Figure 2 : income distribution of customers

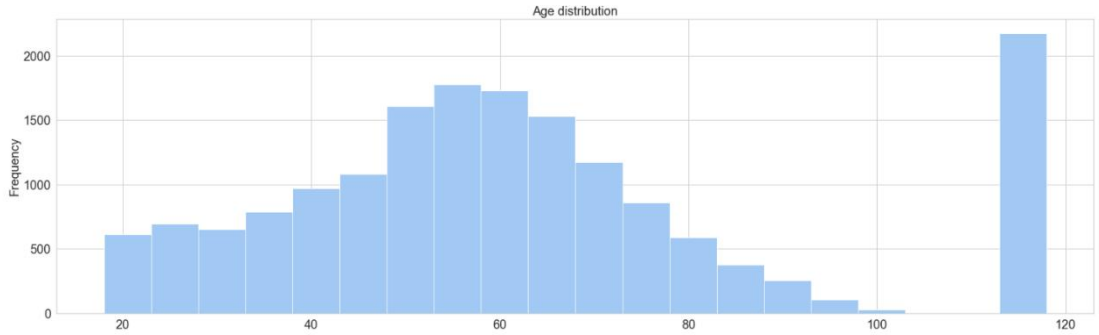

Figure 3 : age distribution of customers

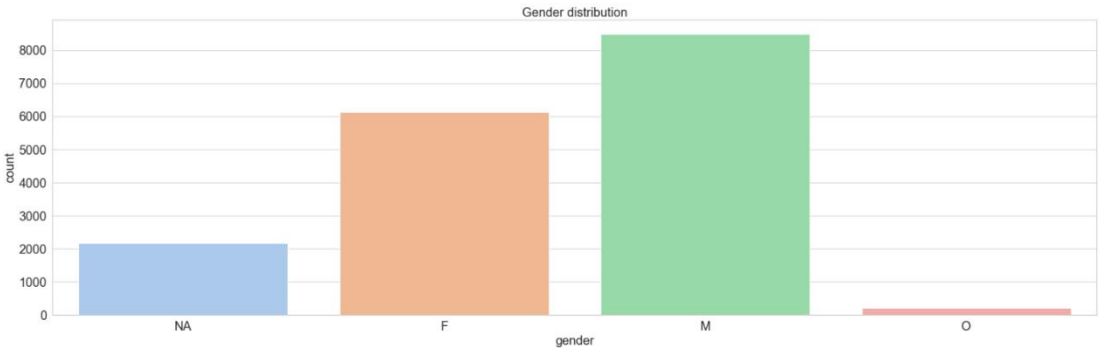

Figure 4 : gender distribution of customers

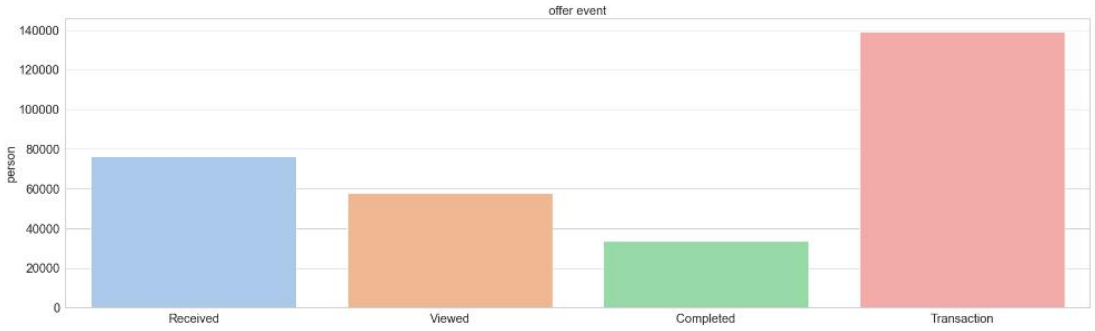

Figure 5 : composition of offer events in transactions

The income distribution is within expectation, however there seems to be more than 2,000 customers with the age above 100, likely due to wrong data entry. Meanwhile in the gender distribution, apart from the null values, there is another category, “O”, catered for customers who are neither “F” (females) or “M” (males). The distribution of the offer events does shed some light on the recording of each transaction data. Primarily, a diminishing count from received offers, viewed offers, and completed suggests a sequence of events where customers are being sent an offer and would either view the offer or not to later completing the offer within the validity period of the offer. Given the exceedingly high bar of transaction events in Figure 5, its highly suggestive that the data set includes transactions of drink purchases that might come from customers using their offers, customers not using their offers and also customers who have not been sent any offers.

Methodology

Data Preprocessing

The age group outliers can first be discarded as there is no way to accurately assign an age to these rows of data. As the purpose of the model is to identify the contributing demographic data to the effective offer redemption, data without gender and/or income is also not too helpful and hence can be omitted. As for the gender category classified as ‘O’, since the count is only 212 out of 17,000 complete customers, it may not be sufficiently enough for any model training. Secondly, as this group may consist of different sub-varieties, modelling as a main category “O” is also not too desirable.

	person	event	value	time
0	78afa995795e4d85b5d9ceeca43f5fef	offer received	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}	0
15561	78afa995795e4d85b5d9ceeca43f5fef	offer viewed	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}	6
47582	78afa995795e4d85b5d9ceeca43f5fef	transaction	{'amount': 19.89}	132
47583	78afa995795e4d85b5d9ceeca43f5fef	offer completed	{'offer_id': '9b98b8c7a33c4b65b9aebfe6a799e6d9...	132
49502	78afa995795e4d85b5d9ceeca43f5fef	transaction	{'amount': 17.78}	144

Figure 6 : snippet of transcript.json

The original transcript table (Figure 6) captures the offer id that each customer either receives, views or completes in the “value” column. For any transaction, an amount spent is also being stored as a dictionary in the same “value” column.

	person	event	time	offer_id	amount_spent	reward_claimed
0	78afa995795e4d85b5d9ceeca43f5fef	offer received	0	9b98b8c7a33c4b65b9aebfe6a799e6d9	0.00	0
15561	78afa995795e4d85b5d9ceeca43f5fef	offer viewed	6	9b98b8c7a33c4b65b9aebfe6a799e6d9	0.00	0
47582	78afa995795e4d85b5d9ceeca43f5fef	transaction	132	transaction	19.89	0
47583	78afa995795e4d85b5d9ceeca43f5fef	offer completed	132	9b98b8c7a33c4b65b9aebfe6a799e6d9	0.00	5
49502	78afa995795e4d85b5d9ceeca43f5fef	transaction	144	transaction	17.78	0

Figure 7 : snippet of transcript.json after splitting the value column into offer_id, amount_spent, and reward_claimed

For the purpose of modelling, it would be desirable to have the data set formatted as row-wise offers that includes its eventual outcomes (offer received, viewed etc). However, as each offer may be repeated sent to a customer at different times, there is a need to identify which times is a customer receiving an offer for. Finally, the transcript dataset can then be merged with the portfolio and profile datasets.

Strictly speaking an informational offer is not an “offer” per se. Understandably, customers would not need to disclose if they are purchasing under the influence of any promotional offer unlike the other two offers where customers would need to perform some actions (like a click) on the offers in their Rewards apps. The above statement is supported by looking at the composition of the completed offers; only ids belonging to the discount and BOGO offers are recorded, but none for informational offers.

Figure 8 : composition of offer type in the all the offer completed events

However, it would be more meaningful to actually include informational campaign in the picture as well. As the datasets is only applicable to one product, this makes it easier for the simplification that if a customer, views AND purchases a drink within the validity period (of the informational offer), it can be counted that the effect is felt from the informational offer sent. In contrast, as with the other offers, if the offers are not viewed but purchases are completed within the valid

periods, these would not be considered as the effect of the promotional offers. Secondly, the “offer viewed” and “offer completed” status are merely the action event captured by the app of a customer. A customer may still complete an offer without viewing an offer, or completes an offer but only views the offer after that.

The data wrangling process must present better clarity for these two mentioned effects.

	143532	143533
index	0	15561
person	78afa995795e4d85b5d9ceeca43f5fef	78afa995795e4d85b5d9ceeca43f5fef
event	offer received	offer viewed
time	0	6
offer_id	9b98b8c7a33c4b65b9aebfe6a799e6d9	9b98b8c7a33c4b65b9aebfe6a799e6d9
amount_spent	0.0	0.0
reward_claimed	0	0
cum_count	1	1
gender	F	F
age	75	75
became_member_on	20170509	20170509
income	100000.0	100000.0
Transaction_From_Offer	False	False
Which_Offer	1.0	1.0
Related_Offer_Id	9b98b8c7a33c4b65b9aebfe6a799e6d9	9b98b8c7a33c4b65b9aebfe6a799e6d9
reward	5.0	5.0
channels	[web, email, mobile]	[web, email, mobile]
difficulty	5.0	5.0
duration	7.0	7.0
offer_type	bogo	bogo
membership_terms(y)	5	5
personxoffer_id	78afa995795e4d85b5d9ceeca43f5fef9b98b8c7a33c4b...	78afa995795e4d85b5d9ceeca43f5fef9b98b8c7a33c4b...
time_of_offer	0.0	0.0
hasViewed_on_time	True	True
hasViewed	True	True
hasCompleted	True	True
transaction_completed	True	True
hasCompleted | hasViewed	1	1

Figure 9 : a snippet of transcript.json after successive steps of data preprocessing

The data wrangling have achieved the effect as in Figure 9. What is almost now a ready datasets for modelling has the transaction details merged with the customer details and the offer details. The new features of the merged data set includes :

* Which_Offer (int) - the nth time a same offer has been received by a customer
* Transaction_From_Offer (bool) - Does the culmination of transaction deduced to be from any offer
* Related_Offer_Id (str) -The offer id from which a transaction is deduced to be derived

* membership_terms(y) (int) - Number of years a customer has been a member of Starbucks rewards (engineered from “became_member_on”)
* time_of_offer (int) - The relative time measure when the Related_Offer_id was sent
* hasViewed (bool) - Has a specific offer of the nth time has been viewed by the customer
* hasViewed_on_time (bool) - Has a specific offer of the nth time has been viewed by the customer, by the validity period
* hasCompleted (bool) - Has a specific offer of the nth time has been completed by the customer, by the validity period
* transaction_completed (bool) - Does the row culminate in a completed transaction?
* hasCompleted|hasViewed (binary) - Has a specific offer of the nth time has been viewed AND completed by the customer, by the validity period (the pipe symbol denotes the conditional of a customer having completed an offer conditional on viewing the offer; not the OR in programming logic)

The data set is now ready for analysis. At the same time, let's remove some of the columns not vital for model training. While usually the marketing channels would be of a great variable to model, the marketing channels are pretty much the same in all offers. Hence let's assume the effect of marketing channel as negligible and remove the "channels" column. For now the "amount_spent" is also left out as it is the outcome of the transaction, not to be a predictor of it. Note that along with stripping “event, “time” etc, there will now be plenty of duplicated rows. This is the case of say, an offer was received, the same offer was viewed, and then completed, and eventually an transaction was based on the offer, there would then be four similar rows present in the data. It is safe to remove duplicates for this case, as the event category has not been pivoted as column heading, resulting in no loss of important.

The key here is to model for every offer/transaction per user basis. Therefore, transactions without any accompanying offers shall be excluded in the model training. Finally, what is also important is to remember the objective to identify which customer segments that Starbucks can actually target for the offers. Having age and income in the continuous may not be effective as that. Hence, the continuous data can be binned accordingly to be converted as categorical data.

	gender	age_group	income_group	membership_terms_group	Which_Offer	reward	difficulty	duration	offer_type	hasCompleted | hasViewed
0	M	30-40	70-80k	2-5	1.0	0.0	0.0	3.0	informational	0
3	M	30-40	70-80k	2-5	1.0	0.0	0.0	4.0	informational	0
5	M	30-40	70-80k	2-5	1.0	5.0	5.0	5.0	bogo	0
9	M	30-40	70-80k	2-5	1.0	2.0	10.0	10.0	discount	0
14	M	30-40	70-80k	2-5	1.0	2.0	10.0	7.0	discount	0
...	...	...	...	...	...	...	...	...	...	...
306496	M	30-40	30-40k	2-5	1.0	5.0	5.0	7.0	bogo	0
306501	F	40-50	60-70k	5-8	1.0	2.0	10.0	10.0	discount	1
306507	F	40-50	60-70k	5-8	1.0	5.0	20.0	10.0	discount	1
306515	F	40-50	60-70k	5-8	1.0	2.0	10.0	7.0	discount	1
306524	F	40-50	60-70k	5-8	1.0	5.0	5.0	7.0	bogo	0

Figure 10 : processed data set ready for modelling

Implementation

Four models are explored on the data set, namely Random Forest Classifier, Gradient Boosting, Light Gradient Boosting, Extreme Gradient Boosting. The result is as follows :

- Random Forest model train accuracy: 0.7840 variance: 0.0022 test_score: 0.7837
- Gradient Boosting model train accuracy: 0.7978 variance: 0.0021 test_score: 0.7934
- Extreme Gradient Boosting train accuracy: 0.7930 variance: 0.0012 test_score: 0.7908
- Light Gradient Boosting train accuracy: 0.7968 variance: 0.0011 test_score: 0.7930

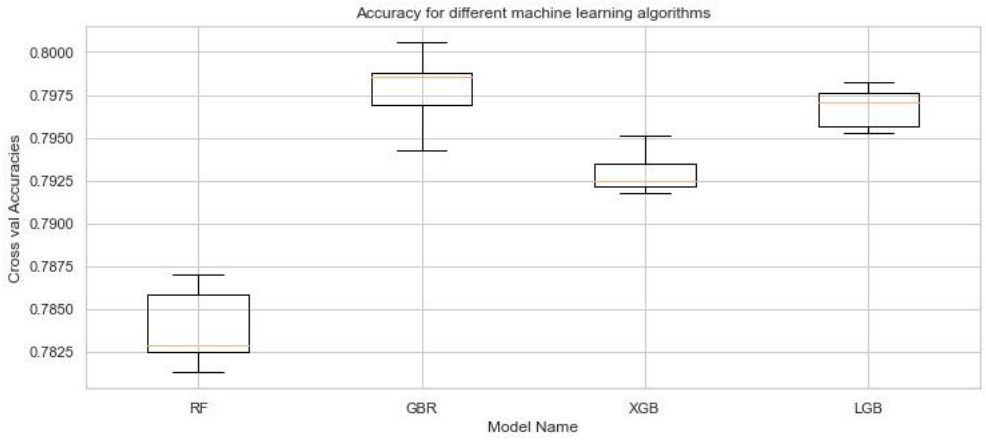

Figure 11 : accuracy and variance of different machine learning algorithms

The best scoring model is Gradient Boosting, which scores 79.78 % on training set with 79.34 % on test set.

	precision	recall	f1-score	support
0	0.80	0.98	0.88	10118
1	0.61	0.10	0.17	2754
accuracy			0.79	12872
macro avg	0.70	0.54	0.53	12872
weighted avg	0.76	0.79	0.73	12872

Figure 12 : classification matrix for the Gradient Boosting

Refinement

The selected model is then grid searched for the n_estimators, min_samples_split, learning_rate parameters to improve the model. Given the subspace, it is found that with n_estimators = 200, min_samples_split=10 and learning_rate = 0.05 give a more optimal model. The improved model scores 79.94 % on training set with 79.37 % on test set.

Results

Model Evaluation and Validation

When the data set is studied for correlation patterns among the features. It could be sent the top ranked features that are positively correlated to the outcome of successful transactions.

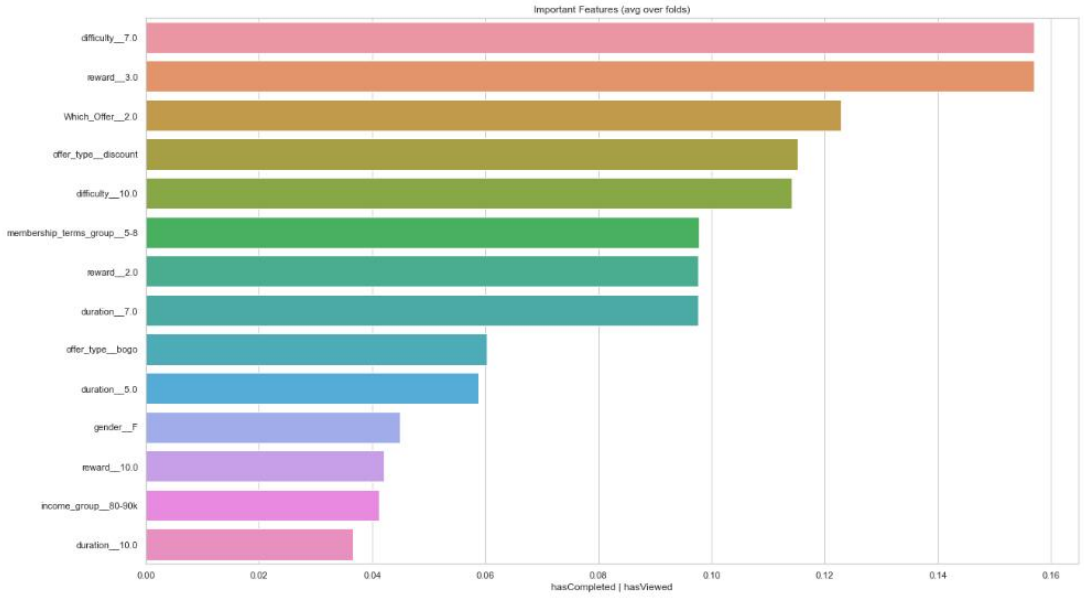

Figure 13 : top positively correlated features with successful transaction

Looking at the overall summary of correlation, the demographic properties of consumers from the highest to lowest would suggest that female customers having a membership term between 5-8 years and with the income level of 80 k to 90 k are probably a good target. It could be hypothesized that these segments customers are likely to respond favorably to offers.

When tested on the test set using the fitted model; using the metric hit_rate (defined as *amount of targeted customers who have viewed and completed an offer /amount of targeted customers who have been sent an offer*), the female customers having a membership term between 5-8 years and with the income level of 80 k to

90 k scored the highest at 0.196. Given the proximity, it is insightful to also include the income range of 70 k to 80 k and 60 k to 70 k.

	gender	membership_terms_group	income_group	hit_rate
20	F	5-8	80-90k	0.196078
32	F	5-8	70-80k	0.194611
2	F	5-8	60-70k	0.193050
8	F	5-8	50-60k	0.174041
39	M	5-8	90-100k	0.158654
21	M	5-8	80-90k	0.115556
38	F	5-8	90-100k	0.080972
47	M	8-10	30-40k	0.058824
16	F	8-10	40-50k	0.055556
9	M	5-8	50-60k	0.043825

Figure 14 : top ranked customer segment of varied gender, income_group, and membership_terms

Justification

While it has now been identified the segment of customers whom Starbucks can target, there is still some justification needed before Starbucks could act on the information. This is due to the fact that every offer would usually have some direct or indirect costs associated with it. The reward can be one of it, and even informational and entails some dissemination cost. What should be identified is that does the promotion bring incremental benefits to Starbucks from the targeted groups.

Let's try to model the effectiveness of such campaign over these segments of customers namely the female members, with membership terms between 5-8 years, and with the income level of 60k to 90 k (will be known as *target customers*, henceforth). We will select the metric that will help to distinguish the difference to be Average Net Amount Spent (*Average Amount Spent - Average Rebate claimed*) on the transaction. To do that, the transaction data of females customers within 5 to 8 years into their membership and earns between 60 k to 90 k but *were* not sent any offers can be used as the control group, denoted as A.

It must be noted that the number of customers in group A and customers in group B are different as the offers were randomly sent out and purchases occur on a non orderly fashion and each sub segment of customers would also vary. However, as the main purpose is to measure the difference within the target segment and the percentage of the segment to the population of A and B, it is felt that the two populations should be taken into consideration entirely as opposed to ensuring that both population have the same customers.

For the control group A, it is found that there is a **net** spent of $18.93/transaction and the number of transactions constitutes 6.38 % of the total non-offer receiving purchases. For purchases who have received an offer and completed the offer after viewing, the same targeted customers have spent a **net** of $17.35/transaction, while making up 10.09% of the total purchases.

The total target customers' transaction is at 6.38 % out of total non-offer receiving purchases and the percentage of target customers who received offer is 7.47% of all the offer receiving customers. However, when it comes to completion of offers, the percentage grew from 7.47 % to 10.09%. At the very least, it can be said that the target customers are found to have a higher than average utilisation of the offers.

Using t-test on the average net spent for A and B :
Test statistic : 0.45
p-value for two tailed test : 0.65

Since p-value(=0.65) > alpha(=0.05) We do not reject the null hypothesis H_0. There is no statistical evidence that there is a decrease in net spent in the same target group who were being sent offers.

In short, this suggests Starbucks could target this segment of customers as they is found to be most responsive, without hurting the bottom line. Using the fitted model, Starbucks could also use it as a sales forecaster engine before sending out the offers.

Conclusion

Given the findings, it would be a good try to test on the hypothesis that offers sent to female customers between 5-8 years and earning 60k to 90 k would help to induce more purchases. However, it is advisable to have the experiment planned that will minimise the bias from the finding. That would include identifying the necessary test statistics to use and ensure enough data is collected.

There are opportunities to further develop the analysis by extending it to each type of offer. It may be insightful to identify the different segments that are responsive towards each offer type. This should be even improve the model accuracy in giving a better prediction.

Speaking of the models, the different algorithms indeed have scored quite close results to one another and are able to demonstrate little over-fitting with accuracy scores that are close between the train set and test set. This fitted model can also serve to be used to forecast the likely performance before the offers are tested. From a commercial sense, it is more appropriate to be able to forecast the sales projection and to give a layer of certainty.

Finally, let's address some shortcomings and oversimplification of the project as a whole. While the count of missing data and erroneous entries are not too many, it could be great to have them prevented. For the issue of those who lists as "O" for their genders; given their minority, it may be hard to collect enough data to provide a good classifier, let alone each subcategory of the other gender. If the sub-population size is still significant small, that can be ignored for now.

Secondly, there should also be some tracking mechanism for the use of informational offers. If in the event that it is not possible, perhaps it is more desirable to exclude the informational offer from the study. Else, if the informational offer is to announce for some new items coming out and sold as introductory prices, there should be a valid number of permissible redemption which should come with tracking mechanism. The assumption that if a customer purchased a drink within the validity period of the informational offer, they are purchasing under the influence of the said offer is too simplistic. What if a customers purchases five drinks separately within the validity period, would that all count to be purchases under the offer's influence ?

Overall, the data set collected is fit for use. The challenge has been on data wrangling and data cleaning to extra the useful features. Once the strategy of turning the data into row wise transaction-offer has been implemented, it eases the other processes. The evidence does indeed prove that some segments show more favorable responses when received offers. Also, by bringing in the amount spent into play, it can be said that the project was quite substantiated.