Starbucks Customer Rewards Mobile App Data Analysis

[Internship Project Report]

By
Rakshita Arora
A525117720010
Msc Data Science
Amity Institute of Integrative Sciences and Health
Amity University Haryana

Internship project under the guidance of

Alok Srivastava
Assistant Professor
Amity Institute of Integrative Sciences and Health
Amity University Haryana

DECLARATION

I, Rakshita Arora, hereby declare that the presented report of internship titled "*Starbucks Customer Rewards App Data Analysis Project*" is uniquely prepared by me after the completion of '2 months' work under Dr Alok Srivastava.

I also confirm that the report is only prepared for my academic requirement, not for any other purpose.

CERTIFICATE

This is to certify that **Rakshita Arora** has done her 2 month internship in Data Science under Dr Alok Srivastava, Assistant Professor, Amity Institute of Integrative Sciences and Health, Amity University Haryana.

She has worked on a project titled '*Starbucks Customer Rewards App Data Analysis Project'.*

During her internship she has demonstrated her skills with self-motivation to learn new skills. Her performance exceeded our expectations and she was able to complete the project on time.

We wish her all the best for her upcoming career.

Alok Srivastava
Assistant Professor
Amity Institute of Integrative Sciences and Health
Amity University Haryana

ACKNOWLEDGEMENT

I express my sincere gratitude to Dr Alok Srivastava for the help and support at various stages during the project work and for their invaluable guidance, cooperation and suggestions in planning and execution of the project. Without him this project would not have been possible.

I am grateful and really thankful to all of my peers for their ideas and support throughout the project.

Rakshita Arora
Msc Data Science
Amity Institute of Integrative Sciences & Health
Amity University Haryana

ABSTRACT

Starbucks is by far one of the most popular and recognizable brands across the world. Their brand recognition is something many small businesses can only dream of. It is important to remember that their success and their marketing started on a smaller, less grand scale at one time too. Starbucks has implemented a stellar marketing strategy that has worked for them

Starbucks sends an offer to their users over the given period of time and they have collected each event happened as a log in a file, an offer can be Informational, Discount or Bogo, some user might receive the same offer again, some might not receive the same offer. The aim of this project is to build a machine model that could help predict whether a customer will actually use the offer or not.

Table of Contents

Introduction

The dataset chosen for this project contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). Some users might not receive any offer during certain weeks. Not all users receive the same offer, and that is the challenge to solve with this data set.

The task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. The data set used is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.

Every offer has a validity period before the offer expires. As an example, a BOGO offer might be valid for only 5 days. In the data set informational offers have a validity period even though these ads are merely providing information about a product; for example, if an informational offer has 7 days of validity, you can assume the customer is feeling the influence of the offer for 7 days after receiving the advertisement.

The transactional data shows user purchases made on the app including the timestamp of purchase and the amount of money spent on a purchase. This transactional data also has a record for each offer that a user receives as well as a record for when a user actually views the offer. There are also records for when a user completes an offer.

I also kept in mind that some customers do not opt into the offers that they receive; in other words, a user can receive an offer, never actually view the offer, and still complete the offer. For example, a user might receive the "buy 10 dollars get 2 dollars off offer", but the user never opens the offer during the 10 day validity period. The customer spends 15 dollars during those ten days. There will be an offer completion record in the data set; however, the customer was not influenced by the offer because the customer never viewed the offer.

Background

Starbucks Corporation is an American multinational chain of coffeehouses and roastery reserves headquartered in Seattle, Washington. The largest coffeehouse company in the world, with 32,938 retail locations as of the first quarter of 2021, followed distantly by coffee shop chains such as Dunkin Donuts with about 10,000 restaurants, Tim Hortons with 4,300 outlets, and Costa Coffee with nearly 1,700 stores worldwide. Starbucks expanded to **221** stores across India in financial year 2021. Starbucks sends out an offer to users of the mobile app to attract new customers and try to bring the inactive customers back to their stores. The dataset for the Starbucks rewards mobile app is available on Kaggle. The data set can be downloaded from the following link:
https://www.kaggle.com/blacktile/starbucks-app-customer-reward-program-data

I researched on the work that has been already done with this dataset and in most of the works performed by various data analysts, I observed that most of the model accuracies were below 75%. Some of the analysts failed to explore the data to the extent that this dataset holds the potential. Only a few features were explored and very few business questions were answered.

In most of the work already done I observed that the analysts failed to observe that the offer could be completed though the customers never received or viewed the offer. Due to the fact that the offer was completed coincidently and not through Starbucks promotional offers, such kind of entries is actually not usable for our analysis. To overcome this problem I created a new column ‘wasted’ which shows the customers who actually wasted the offer and who didn’t.

Another issue I found with the work already done was that the data wasn’t sorted on the basis of time, which led to taking into consideration those offers which were received or viewed after the offer completion. These entries are not usable because these customers completed the offer due to coincidence and then viewed the offer on the app. To the best of my ability I tried to overcome these limitations through my analysis.

Objective

The goal of this project to complete an exploratory analysis to determine which demographic group responds best to which offer type and what type of customers are likely to waste an offer. The aim is to build a machine model that could help predict whether a customer will actually use the offer or not.

Description of the data

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) - the type of offer ie BOGO, discount, informational
difficulty (int) - the minimum required to spend to complete an offer
reward (int) - the reward is given for completing an offer
duration (int) - time for the offer to be open, in days
channels (list of strings)

```
portfolio.head()
```

	reward	channels	difficulty	duration	offer_type	id
0	10	[email, mobile, social]	10	7	bogo	ae264e3637204a6fb9bb56bc8210ddfd
1	10	[web, email, mobile, social]	10	5	bogo	4d5c57ea9a6940dd891ad53e9dbe8da0
2	0	[web, email, mobile]	0	4	informational	3f207df678b143eea3cee63160fa8bed
3	5	[web, email, mobile]	5	7	bogo	9b98b8c7a33c4b65b9aebfe6a799e6d9
4	5	[web, email]	20	10	discount	0b1e1539f2cc45b7b9fa7c272da2e1d7

```
portfolio.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
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)
memory usage: 608.0+ bytes
```

Portfolio Dataset:

- This dataset has no missing values.
- The ‘difficulty’ column unit is in dollars, which does not reflect how difficult to be rewarded. Re-scaling this feature is a useful step to do.
- The ‘channels’ column contains a list of channels through which the offers are sent. I think it would be useful to separate the elements of this column.

profile.json

age (int) - age of the customer
became_member_on (int) - the 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

```
profile.head()
```

	gender	age	id	became_member_on	income
0	None	118	68be06ca386d4c31939f3a4f0e3dd783	20170212	NaN
1	F	55	0610b486422d4921ae7d2bf64640c50b	20170715	112000.0
2	None	118	38fe809add3b4fcf9315a9694bb96ff5	20180712	NaN
3	F	75	78afa995795e4d85b5d9ceeca43f5fef	20170509	100000.0
4	None	118	a03223e636434f42ac4c3df47e8bac43	20170804	NaN

```
profile.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17000 entries, 0 to 16999
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   gender            14825 non-null  object 
 1   age               17000 non-null  int64  
 2   id                17000 non-null  object 
 3   became_member_on  17000 non-null  int64  
 4   income            14825 non-null  float64
dtypes: float64(1), int64(2), object(2)
memory usage: 664.2+ KB
```

Profile Dataset:

- The dataset has 2175 missing values on each of: ‘gender’, ’income’ variables.
- The customers’ ages range from 18 to 101. Although those 2175 customers were registered at age 118, I still considered this specific age an outlier because it appears clear that there is something wrong related to these 2175 rows in the dataset. The missing values in ‘gender’ and ‘income’ variables which are related solely and specifically with the 2175 customers registered at age 118. In other words, customers at age 118 have no registered ‘gender’ and ‘income’.

transcript.json

event (str) - record description (i.e. ‘transaction’, ’ offer received’, ‘offer viewed’, and ‘offer completed’.)

person (str) - customer id

time (int) - time in hours since the start of the test. The data begins at time t=0

value - (dict of strings) - either an offer id or transaction amount depending on the record

```
transcript.tail()
```

	person	event	value	time
306529	b3a1272bc9904337b331bf348c3e8c17	transaction	{'amount': 1.5899999999999999}	714
306530	68213b08d99a4ae1b0dcb72aebd9aa35	transaction	{'amount': 9.53}	714
306531	a00058cf10334a308c68e7631c529907	transaction	{'amount': 3.61}	714
306532	76ddbd6576844afe811f1a3c0fbb5bec	transaction	{'amount': 3.5300000000000002}	714
306533	c02b10e8752c4d8e9b73f918558531f7	transaction	{'amount': 4.05}	714

```
transcript.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306534 entries, 0 to 306533
Data columns (total 4 columns):
 #   Column  Non-Null Count   Dtype 
---  ------  --------------   ----- 
 0   person  306534 non-null  object
 1   event   306534 non-null  object
 2   value   306534 non-null  object
 3   time    306534 non-null  int64 
dtypes: int64(1), object(3)
memory usage: 9.4+ MB
```

Transcript Dataset :

- The dataset has no missing values.
- The ‘value’ column is a dictionary in which we can apply some kind of Feature Engineering to extract useful data.

Data Processing

After cleaning the data, I converted the columns with categorical values to numerical and in some cases performed one hot encoding and label encoding, and then finally merged the three datasets.

To see which customers might waste the offer, I have added a new column into the dataset named ‘waste’. I am assigning the value 0 to the column for the customers who have received and viewed the offer and then completed it. 1 is assigned to the customers who have not received, viewed and completed the offer. 1 is also assignment to the customers who have completed the offer but they never received or never viewed the offer.

person	event / offer_id	offer completed	offer received	offer viewed	wasted
0009655768c64bdeb2e877511632db8f	2906b810c7d4411798c6938adc9daaa5	1	1	0	1
	3f207df678b143eea3cee63160fa8bed	0	1	1	1
	5a8bc65990b245e5a138643cd4eb9837	0	1	1	1
	f19421c1d4aa40978ebb69ca19b0e20d	1	1	1	0
	fafdcd668e3743c1bb461111dcafc2a4	1	1	1	0
...	...	...	...	...	...
fffad4f4828548d1b5583907f2e9906b	f19421c1d4aa40978ebb69ca19b0e20d	1	1	1	0
ffff82501cea40309d5fdd7edcca4a07	0b1e1539f2cc45b7b9fa7c272da2e1d7	1	1	1	0
	2906b810c7d4411798c6938adc9daaa5	1	1	1	0
	9b98b8c7a33c4b65b9aebfe6a799e6d9	1	1	1	0
	fafdcd668e3743c1bb461111dcafc2a4	1	1	1	0

55222 rows × 4 columns

My final merged dataset:

person	wasted	age	income	reward	difficulty	duration	web	email	mobile	social	bogo	discount	informati	gender_id	year_id	event_id	start_year	offer_typ	offer_id_id
000965576	1	33	72000	2	10	168	1	1	1	0	0	1	0	1	4	0	2017	0	0
000965576	1	33	72000	2	10	168	1	1	1	0	0	1	0	1	4	2	2017	0	0
000965576	1	33	72000	0	0	96	1	1	1	0	0	0	1	1	4	0	2017	1	1
000965576	1	33	72000	0	0	96	1	1	1	0	0	0	1	1	4	1	2017	1	1
000965576	1	33	72000	0	0	72	0	1	1	1	0	0	1	1	4	0	2017	1	2
000965576	1	33	72000	0	0	72	0	1	1	1	0	0	1	1	4	1	2017	1	2
000965576	0	33	72000	5	5	120	1	1	1	1	1	0	0	1	4	0	2017	2	3
000965576	0	33	72000	5	5	120	1	1	1	1	1	0	0	1	4	2	2017	2	3
000965576	0	33	72000	5	5	120	1	1	1	1	1	0	0	1	4	1	2017	2	3
000965576	0	33	72000	2	10	240	1	1	1	1	0	1	0	1	4	0	2017	0	4
000965576	0	33	72000	2	10	240	1	1	1	1	0	1	0	1	4	2	2017	0	4
000965576	0	33	72000	2	10	240	1	1	1	1	0	1	0	1	4	1	2017	0	4
0011e0d4e	0	40	57000	5	20	240	1	1	0	0	0	1	0	2	5	0	2018	0	5
0011e0d4e	0	40	57000	5	20	240	1	1	0	0	0	1	0	2	5	1	2018	0	5
0011e0d4e	0	40	57000	5	20	240	1	1	0	0	0	1	0	2	5	2	2018	0	5
0011e0d4e	0	40	57000	3	7	168	1	1	1	1	0	1	0	2	5	0	2018	0	6
0011e0d4e	0	40	57000	3	7	168	1	1	1	1	0	1	0	2	5	1	2018	0	6

EXPLORATORY DATA ANALYSIS

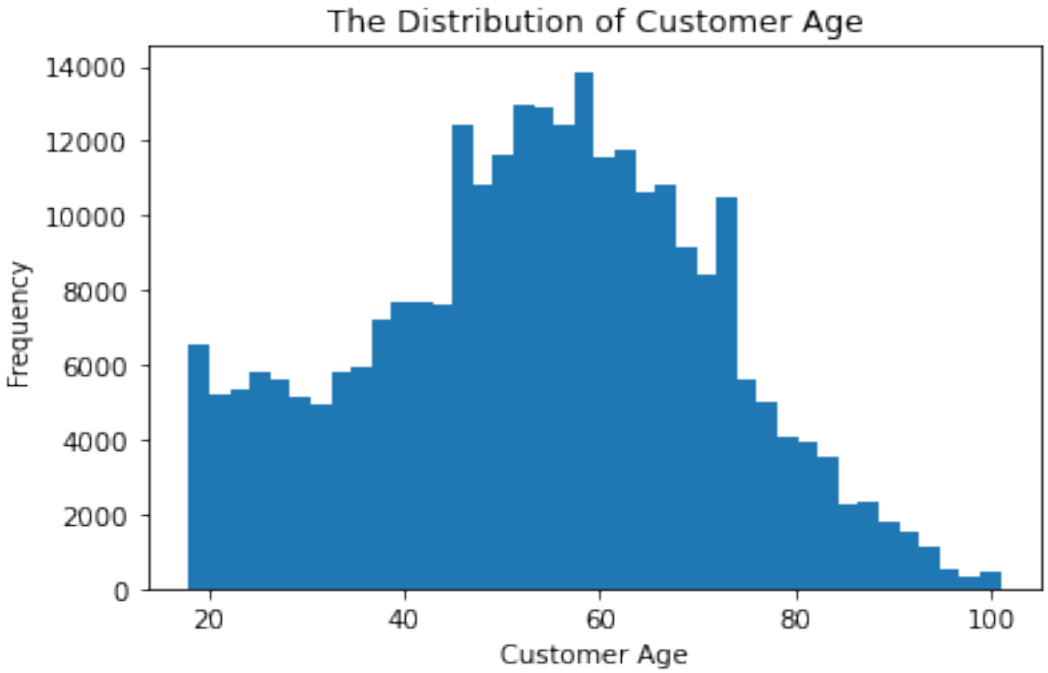

Figure1: This figure shows the distribution of Customer's age. Seems that the age group 45-60 is the most common in customers

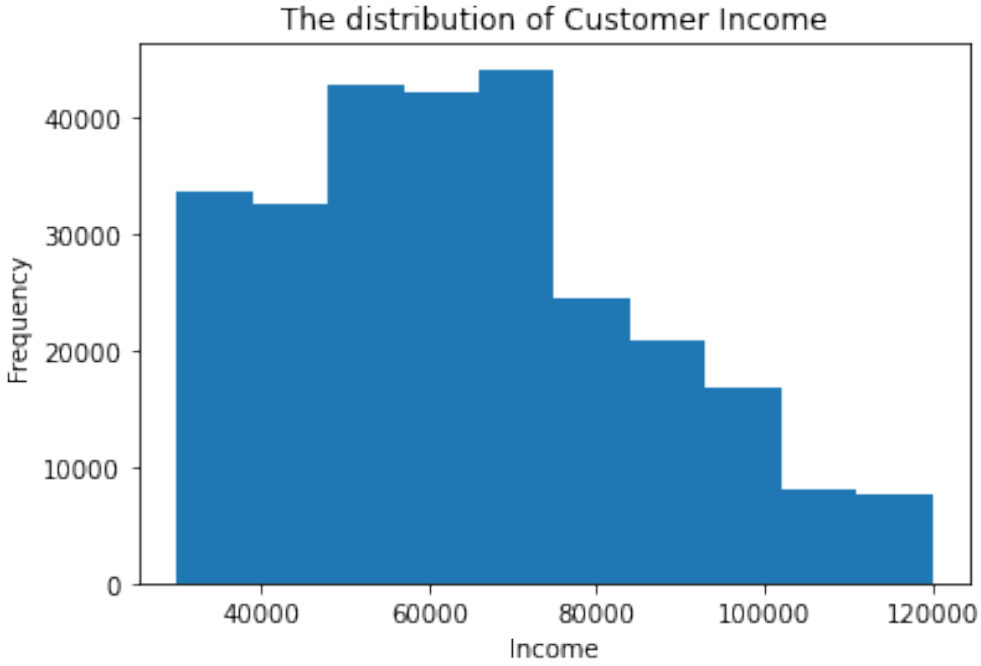

Figure 2: This figure represents the customer Income. Most Customers have income ranging between 50000-70000.


```
Female success percentage of discount offer completion:  72.88677693502696
Male success percentage of discount offer completion:  49.39505523408732

Female success of buy one get one offer completion:  68.34624145785877
Male success of buy one get one offer completion:  49.39505523408732

Discount success rate:  64.45394539453946
Buy one get one success rate:  57.497079549308516
```

Figure 3: This figure shows the female and male offer completion rates for discount and BOGO offers. Discount offer is more popular because not only the absolute number of 'offer completed' is slightly higher than BOGO offer, its overall completed/received rate is also about 7% higher. Females are more likely to complete the bogo or discount offer and males have a low offer completion rate

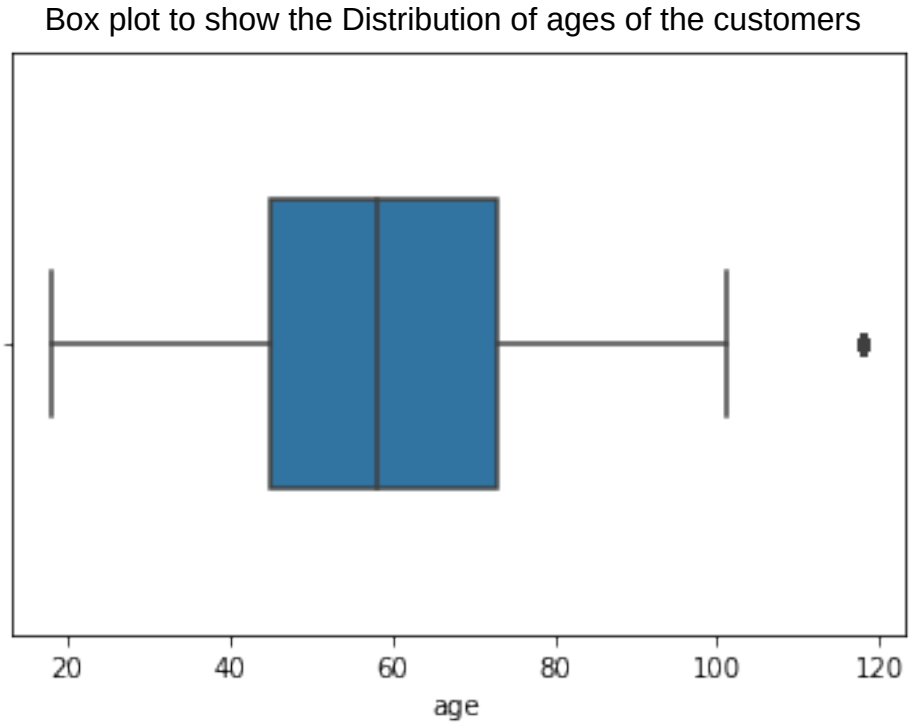

Figure 4: This figure shows the distribution of ages of customers through a boxplot to observe outliers (if any). It seems to be that people with age greater than 80 don't use the app much or they may not drink many beverages. There is one entry for age 118. I will be considering it an outlier.

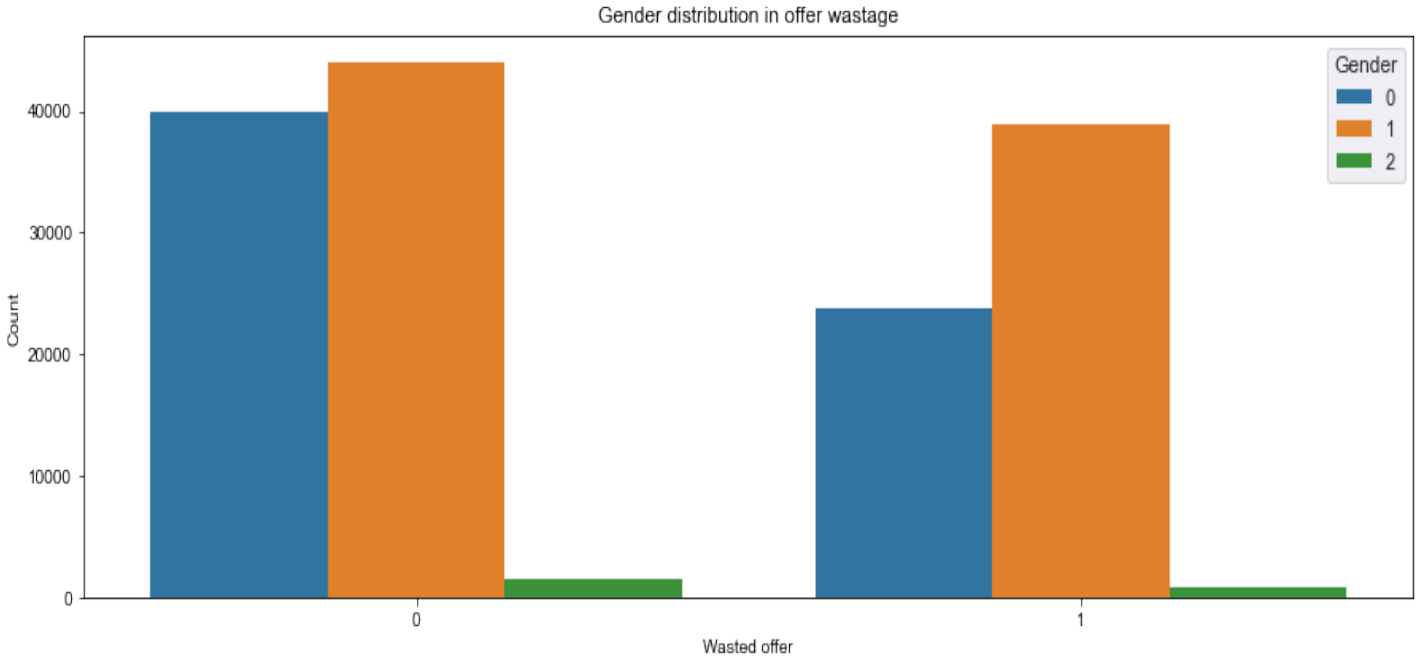

Figure 5: This figure shows the gender distribution of customers vs the offer wastage. 0 stands for offer not wastage and 1 shows offer wasted. The blue bars show females, orange are for males and green for others.

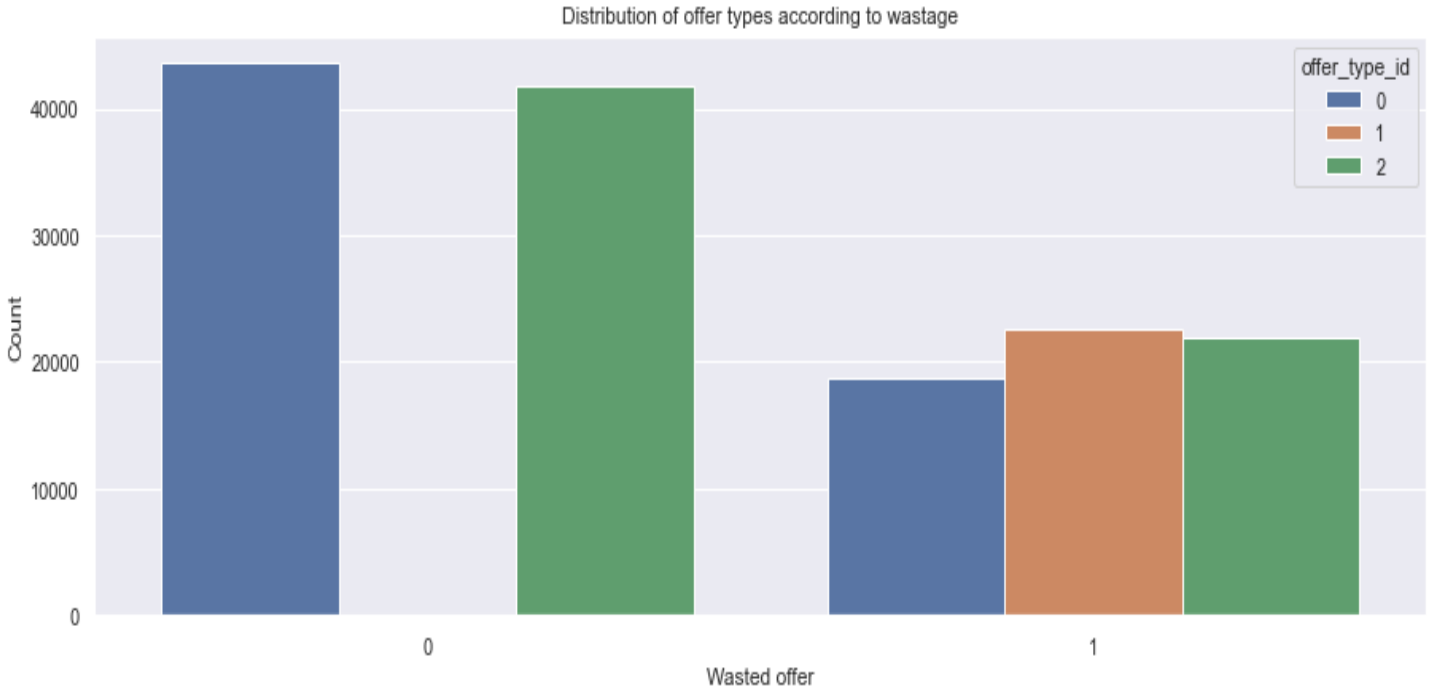

Figure 6: The blue bars show BOGO offers and green are the discount offer. The brown shows the informational offers. This chart shows the distribution of offer types according to wastage. Buy one get one free offer is less likely to get wasted and most of the informational offers are most likely to get wasted

Figure 7: This graph visualizes the distribution of year of membership of customers and wastage of offers. Year_id 0-5 shows the years from 2013 to 2018. Customers who joined in 2017 are less likely to waste an offer.

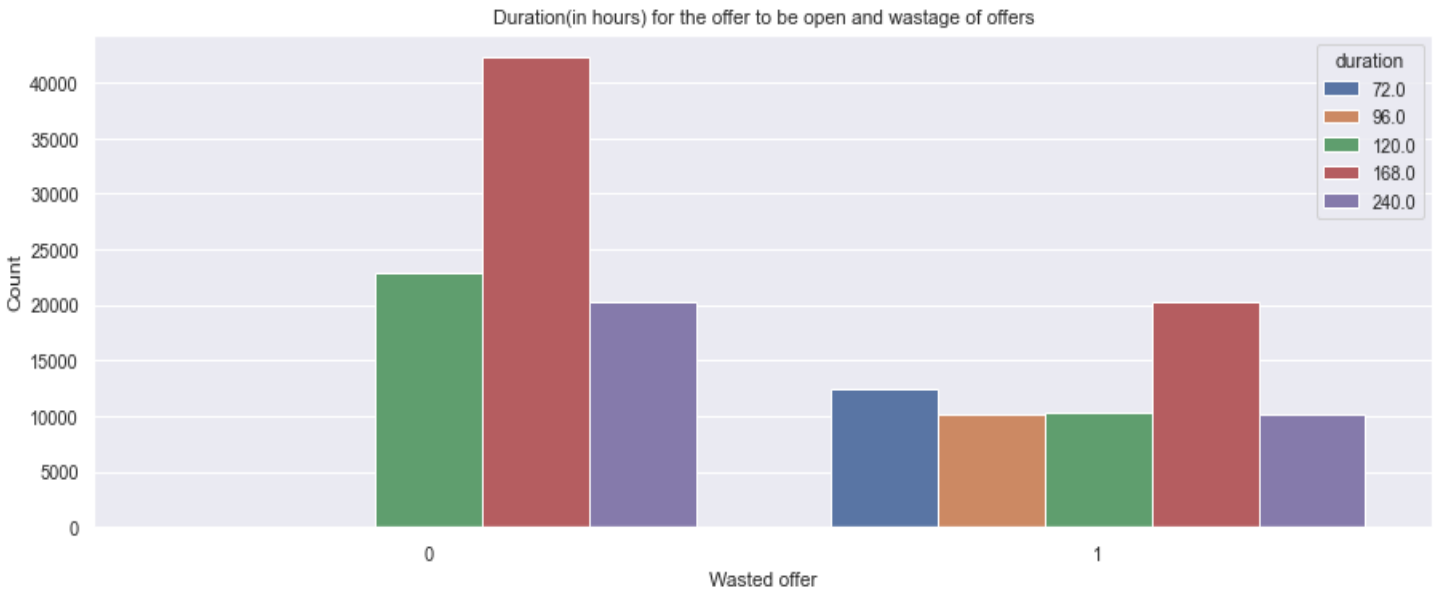

Figure 8: This chart show the offer validity (in hours) vs the offer wastage rates. The ideal time for the offer to be open should be 168 hours ie. 7 days

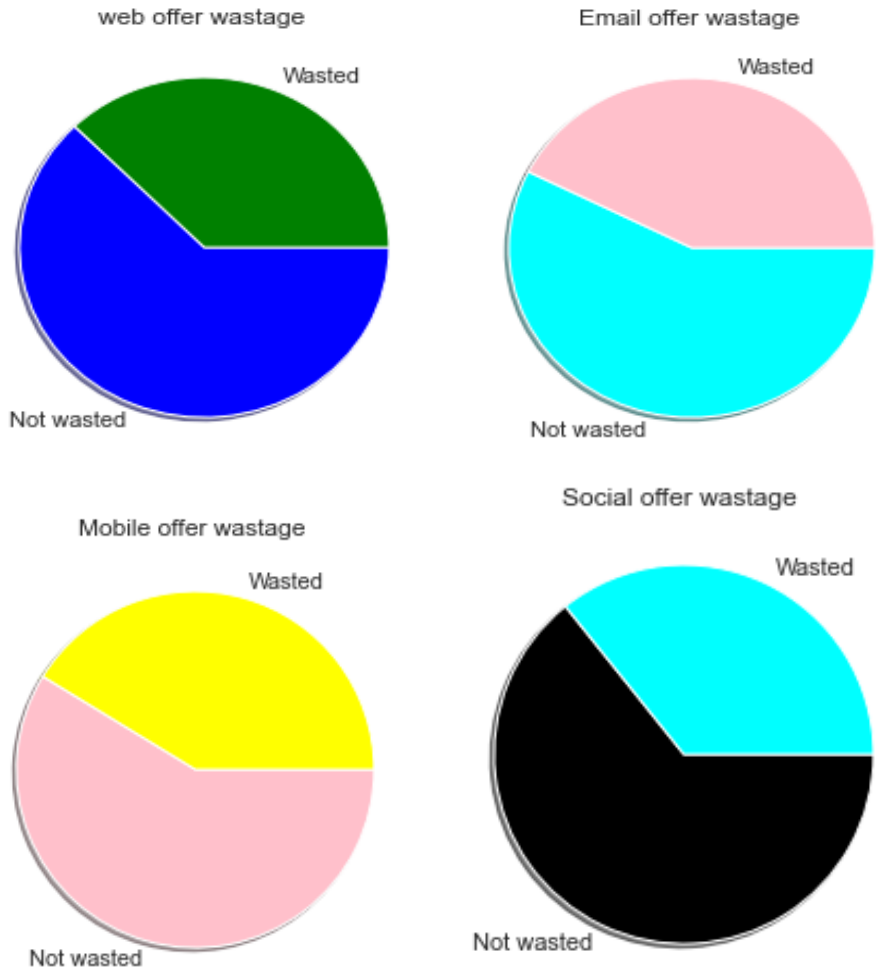

Figure 9: These pie charts shows the offer wastages through various channels. Less number of offers is wasted if offers are sent through social media or web.

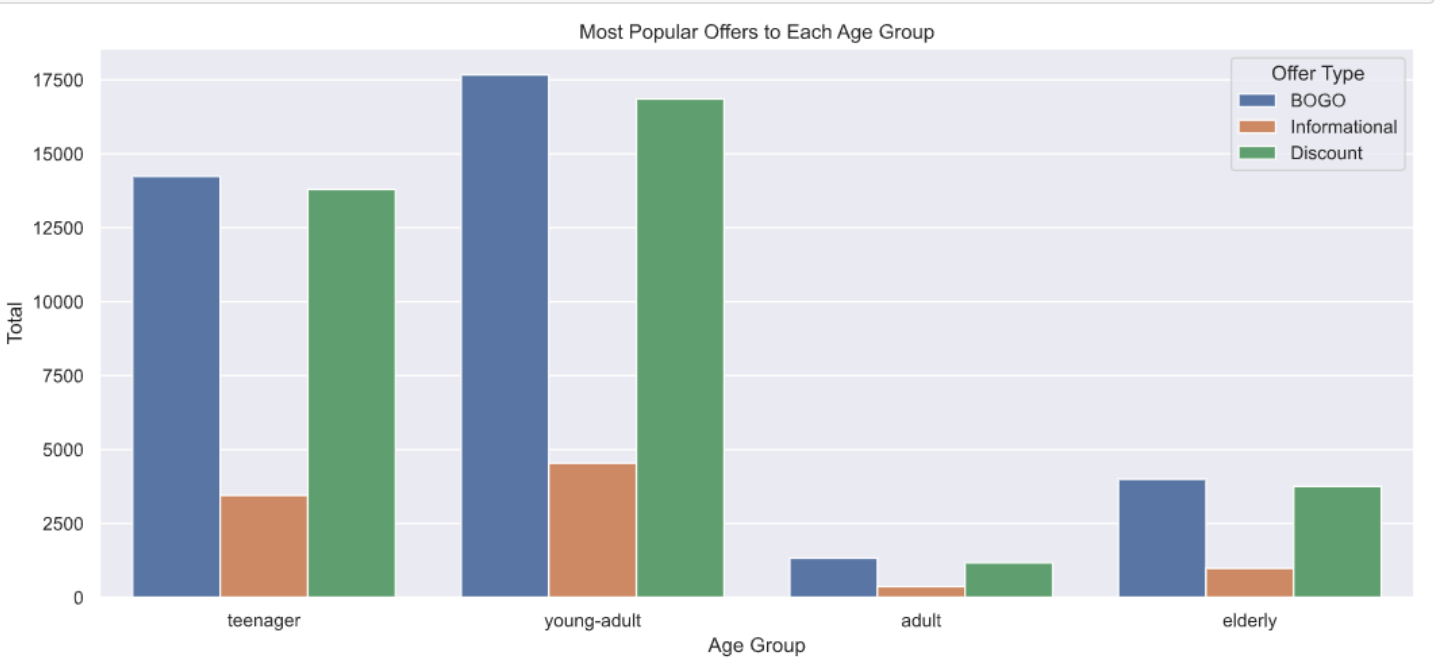

Figure 10: This chart shows the most popular offers to each age group. The most common offer type among all age groups is the BOGO followed by the Discount Offers. Whereas, the least common offer to be sent is the informational offer. I believe that BOGO offers are more attractive compared to other offers provided by Starbucks.

Predictive Modeling

I tried to build a model that can predict whether a customer is likely to waste an offer. I considered only those features that I believe are important. The target column is 'wasted'. 1 means the customer will waste the offer and 0 means they won't. I built a model on 4 classification algorithms and then chose the one with the maximum accuracy. I trained my model on SVM, RandomForest, Decision Tree, Logistic Regression. I also performed hyperparameter tuning to improve my accuracies.

Hyper parameter tuning

While building a Machine learning model we always define two things that are model parameters and model hyperparameters of a predictive algorithm. Model parameters are the ones that are an internal part of the model and their value is computed automatically by the model referring to the data like support vectors in a support vector machine, the weights in an artificial neural network or the coefficients in a linear regression or logistic regression.

But hyperparameters are the ones that can be manipulated by the programmer to improve the performance of the model like the learning rate of a deep learning model, the C and sigma hyperparameters for support vector machines and the k in k-nearest neighbors.

They are the one that commands over the algorithm and are initialized in the form of a tuple.

GridSearchCV and RandomizedSearchCV

Model Hyperparameter tuning is very useful to enhance the performance of a machine learning model. The two approaches to do the tuning is GridSearchCV and RandomizedSeachCV.

The only difference between both the approaches is in grid search we define the combinations and do training of the model whereas in RandomizedSearchCV the model selects the combinations randomly. Both are very effective ways of tuning the parameters that increase the model generalizability.

	model	best_score	best_params
0	svm	0.866556	{'C': 20, 'kernel': 'rbf'}
1	random_forest	0.859795	{'n_estimators': 10}
2	logistic_regression	0.826592	{'C': 5}
3	decision_tree	0.878007	{'criterion': 'entropy'}

Table 1: Shows the models and best parameters after performing GridSearchCV technique.

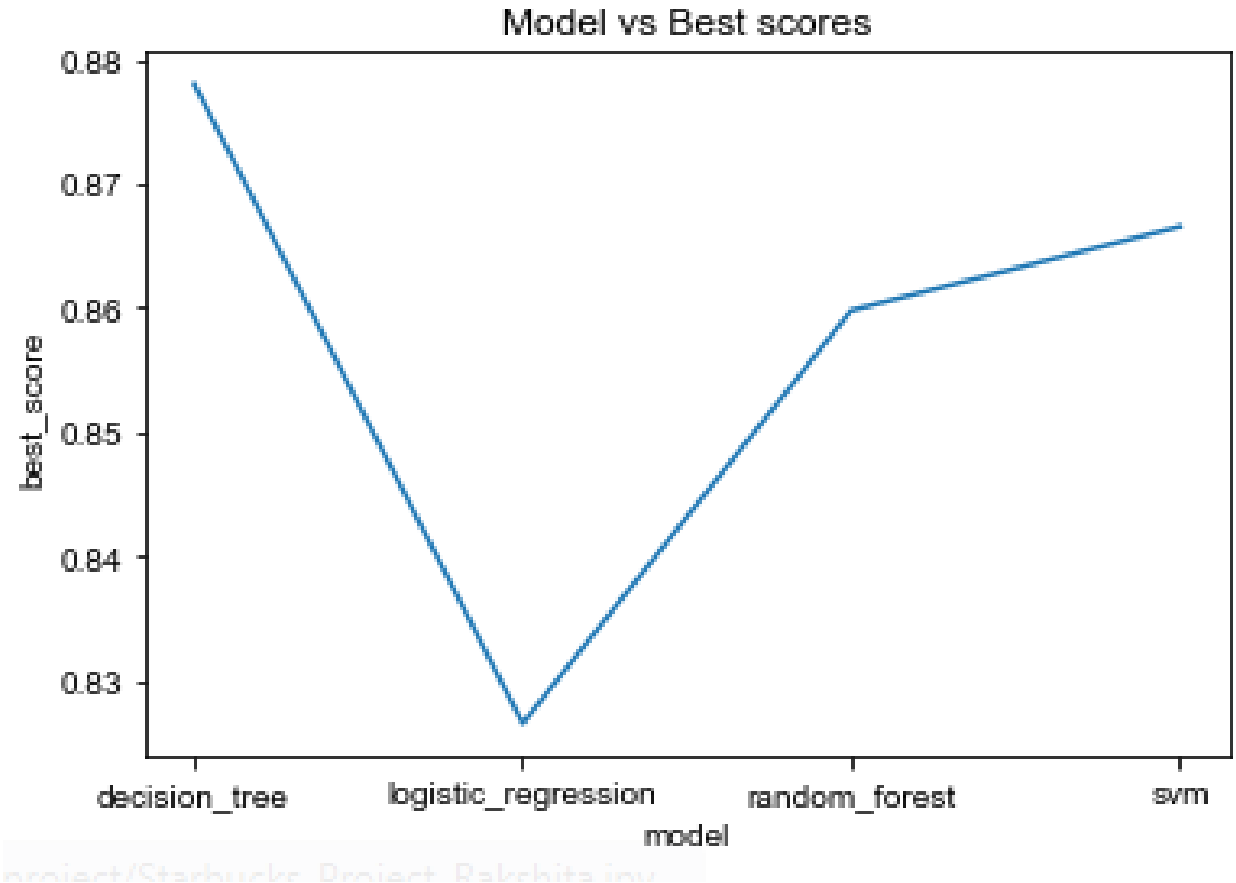

Figure 11: Shows the model performances

Table 1 and Figure 11 shows that the Decision Tree Classifier gives the best scores.

Performance Metrics

Metrics used in this project are the common ones used to calculate any classification model's performance. Performance evaluation of a classification model is based on the number of test records correctly and incorrectly predicted by the model.

Accuracy is the most frequent classification evaluation metric. It works well in balanced datasets. *Accuracy* measures the percentage of cases that the model has classified correctly. Accuracy can be misleading, though. It can make a dysfunctional model look like it's a good one.

The **accuracy** metric does not work well when classes are unbalanced. For problems with unbalanced classes it is much better to use **precision, recall and F1**. These metrics give a better idea of the quality of the model. **Precision** informs about the **quality** of the machine learning model in classification tasks.

I achieved the following scores for the **Decision Tree model**.

Accuracy: 87.80072577

Precision: 83.84215227

Recall: 87.06280351

F1-score: 85.422131794

Result

The problem that I chose to solve was to build a model that predicts whether a customer will respond to an offer.

The following observations were made after the exploratory data analysis:

The age group 45-60 is the most common in customers. Most Customers have income ranging between 50000-70000. Discount offer is more popular because not only the absolute number of 'offer completed' is slightly higher than BOGO offer, its overall completed/received rate is also about 7% higher.

Females are more likely to complete the bogo or discount offer and males have a low offer completion rate It seems to be that people with age greater than 80 don't use the app much or they may not drink many beverages.

Buy one get one free offer is less likely to get wasted and most of the informational offers are most likely to get wasted.

Customers who joined in 2017 are less likely to waste an offer. The ideal time for the offer to be open should be 168 hours ie. 7 days Less number of offers is wasted if offers are sent through social media or web.

The most common offer type among all age groups is the BOGO , followed by the Discount Offers. Whereas, the least common offer to be sent is the informational offers. I believe that BOGO offers are more attractive compared to other offers provided by Starbucks.

I chose the "DecisionTreeClassifier" which is the best performing classifier algorithm among the other classifiers I used. Built the model and an accuracy of 87.8 % was achieved by the model.

Conclusion

Performed an exploratory data analysis on the datasets looking at how the different demographics of Starbucks Rewards users responded to the different offer types

Through the exploratory data analysis of the Starbucks data, I have answered the following business questions:
Q1. In what age group do most of Starbucks customers fall?
Q2. What is the income range of most of the customers?
Q3. Which offer is more popular among the customers?
Q4. Which gender has better offer completion rate in Bogo and discount?
Q5. Which offer is more likely to get wasted?
Q6. Customers who joined in which year are most likely to complete the offers?
Q7. How long should the offer validity last?
Q8. Through which medium (social,mobile,web,email) should the offers be sent to the customers to get the maximum number of offers completed?
Q9. What is the common offer among each age group (teenagers, young-adults, adults and elderly)?

These suggestions could be really beneficial for the company to take valuable business decisions. Personalizing offers on basis of type of customers can really help the company make increased sales.
I preprocessed the data to ensure it was appropriate for the predictive algorithms. Then I used various models to predict whether an offer would be successful based on a variety of features about the offer and the user.

Future work

The project is open to check the performance with other types of classification algorithms such as *XGBoost* for example. In addition, more demographic information related to customers would be helpful to improve the models. Creating an application that applies the best classifier model is an interesting idea to consider. This project can be used to predict whether or not a certain type of offer is suitable for a certain type of customer before investing in a marketing campaign.

References

https://www.kaggle.com/blacktile/starbucks-app-customer-reward-program-data

https://www.kaggle.com/shailjakhurana/starbucks-offers-analysis

https://towardsdatascience.com/why-you-should-do-feature-engineering-first-hyperparameter-tuning-second-as-a-data-scientist-334be5eb276c

https://www.webcontentdevelopment.com/the-starbucks-marketing-model/

https://medium.com/swlh/starbucks-capstone-challenge-offer-analysis-and-success-prediction-78574e915dbf