Outline: [Analysis of Startbucks offers]

Keyword: [Data Scientist Nanodegree Capstone project]
Author: [Indu Girish]
Publish Date: [24-02-2022]

[Analysis of Startbucks offers]

Introduction

Giving offers is a common marketing strategy to attract potential customers and increase sales. Understanding what kind of offers really excites the customers and sending personalised offers to them will skyrocket the business.

Here is an analysis of Starbucks offers to find out what is the best offer that can be given to a customer. The data set contains simulated data that mimics customer behaviour 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.

Problem Statement

Objective of the project is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. Not all users receive the same offer, and that is the challenge to solve with this data set.

Some users might make a purchase through the app without having received an offer or seen an offer. So the approach here is to first find the users who have received an offer and viewed it, and then find users who have completed that offer within the validity period of that offer. This gives the actual dataset of users who have been influenced by Starbucks offers and responded to it. We can analyse this data to find out which type of offers can be given to which group of users. Also creates a machine learning model to predict the offer type that can be given to a demographic group.

Data Exploration

The data is contained in three files:

- *portfolio: There are 10 different offers with details such as offer id, type of offer, difficulty indicating the minimum required spend to complete an offer, reward given for completing an offer, time duration for offer to be open, in days and channels used for sending the offer.*

	channels	difficulty	duration	id	offer_type	reward
0	[email, mobile, social]	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10
1	[web, email, mobile, social]	10	5	4d5c57ea9a6940dd891ad53e9dbe8da0	bogo	10
2	[web, email, mobile]	0	4	3f207df678b143eea3cee63160fa8bed	informational	0
3	[web, email, mobile]	5	7	9b98b8c7a33c4b65b9aebfe6a799e6d9	bogo	5
4	[web, email]	20	10	0b1e1539f2cc45b7b9fa7c272da2e1d7	discount	5
5	[web, email, mobile, social]	7	7	2298d6c36e964ae4a3e7e9706d1fb8c2	discount	3
6	[web, email, mobile, social]	10	10	fafdcd668e3743c1bb461111dcafc2a4	discount	2
7	[email, mobile, social]	0	3	5a8bc65990b245e5a138643cd4eb9837	informational	0
8	[web, email, mobile, social]	5	5	f19421c1d4aa40978ebb69ca19b0e20d	bogo	5
9	[web, email, mobile]	10	7	2906b810c7d4411798c6938adc9daaa5	discount	2

- *profile: It contains the demographic data for 17000 customers with details such as age, gender, customer id, date when customer created an app account, and income.*

	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

- *transcript - It contains records for transactions, offers received, offers viewed, and offers completed, with a value dictionary of offer id or transaction amount depending on the record, customer id, and time in hours since start of test.*

	person	event	value	time
0	78afa995795e4d85b5d9ceeca43f5fef	offer received	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}	0
1	a03223e636434f42ac4c3df47e8bac43	offer received	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}	0
2	e2127556f4f64592b11af22de27a7932	offer received	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}	0
3	8ec6ce2a7e7949b1bf142def7d0e0586	offer received	{'offer id': 'fafdcd668e3743c1bb461111dcafc2a4'}	0
4	68617ca6246f4fbc85e91a2a49552598	offer received	{'offer id': '4d5c57ea9a6940dd891ad53e9dbe8da0'}	0

Data Cleaning

portfolio

There is no missing value in the portfolio table. Following steps are done as part of cleaning the dataset.

- *Renamed column 'id' to 'offer_id'*
- *'channels' column is one-hot encoded.*

	reward	difficulty	duration	offer_type	offer_id	email	social	web	mobile
0	10	10	7	bogo	ae264e3637204a6fb9bb56bc8210ddfd	1	1	0	1
1	10	10	5	bogo	4d5c57ea9a6940dd891ad53e9dbe8da0	1	1	1	1
2	0	0	4	informational	3f207df678b143eea3cee63160fa8bed	1	0	1	1
3	5	5	7	bogo	9b98b8c7a33c4b65b9aebfe6a799e6d9	1	0	1	1
4	5	20	10	discount	0b1e1539f2cc45b7b9fa7c272da2e1d7	1	0	1	0
5	3	7	7	discount	2298d6c36e964ae4a3e7e9706d1fb8c2	1	1	1	1
6	2	10	10	discount	fafdcd668e3743c1bb461111dcafc2a4	1	1	1	1
7	0	0	3	informational	5a8bc65990b245e5a138643cd4eb9837	1	1	0	1
8	5	5	5	bogo	f19421c1d4aa40978ebb69ca19b0e20d	1	1	1	1
9	2	10	7	discount	2906b810c7d4411798c6938adc9daaa5	1	0	1	1

profile

There are some records with nan values for 'income' and 'gender' in the profile table. Also, the 'age' for all these records is 118, and it could be a mistake. So we can delete these records.
Following steps are done as part of cleaning the dataset.

- *Rename column 'id' to 'customer_id'*
- *Drop nan values*
- *convert 'became_member_on' to 'year_of_joining'*
- *Convert 'age' to categorical data*
- *Convert 'income' to categorical data*

	gender	age	customer_id	income	year_of_joining
0	F	50s	0610b486422d4921ae7d2bf64640c50b	above 80k	2017
1	F	70s	78afa995795e4d85b5d9ceeca43f5fef	above 80k	2017
2	M	60s	e2127556f4f64592b11af22de27a7932	70k-80k	2018
3	M	60s	389bc3fa690240e798340f5a15918d5c	50k-60k	2018
4	M	50s	2eeac8d8feae4a8cad5a6af0499a211d	50k-60k	2017

transcript

Following steps are done as part of cleaning the dataset.

- *Split 'value' column and create columns for offer_id, amount and reward*
- *Rename column 'person' to 'customer_id'*
- *Drop records with customer_id not in profile table*

	event	customer_id	time	amount	offer_id	reward_received
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	NaN	9b98b8c7a33c4b65b9aebfe6a799e6d9	NaN
2	offer received	e2127556f4f64592b11af22de27a7932	0	NaN	2906b810c7d4411798c6938adc9daaa5	NaN
5	offer received	389bc3fa690240e798340f5a15918d5c	0	NaN	f19421c1d4aa40978ebb69ca19b0e20d	NaN
7	offer received	2eeac8d8feae4a8cad5a6af0499a211d	0	NaN	3f207df678b143eea3cee63160fa8bed	NaN
8	offer received	aa4862eba776480b8bb9c68455b8c2e1	0	NaN	0b1e1539f2cc45b7b9fa7c272da2e1d7	NaN

Data Analysis

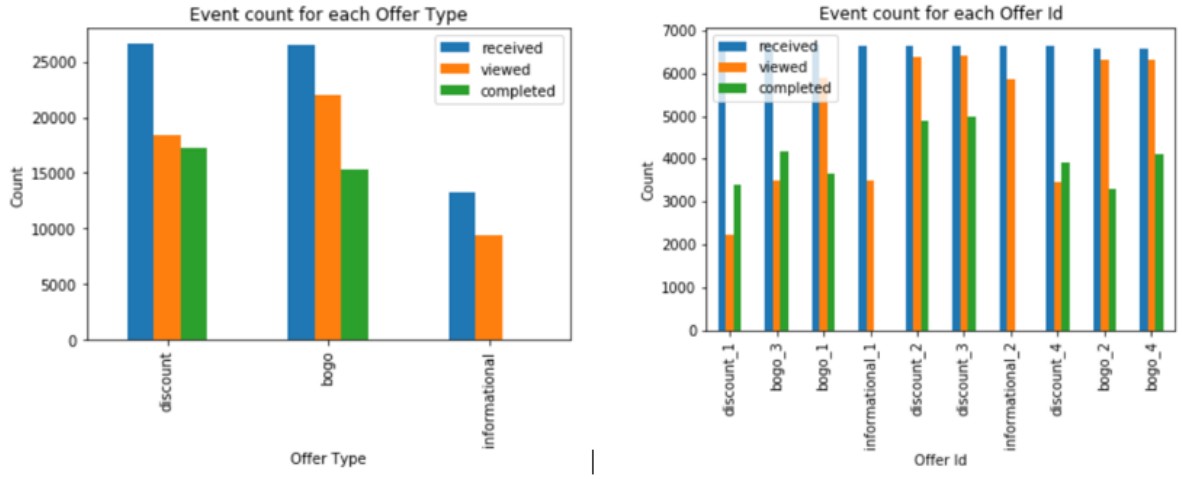
Pre-processing

To find out the customers who are influenced by an offer, we need to merge transcript and portfolio data. Then I renamed offer ids to more readable forms like bogo_1, bogo_2, etc. The merged table looks as shown below.

event	customer_id	time	amount	offer_id	reward_received	difficulty	duration	offer_type	reward	email	web	mobile	social
offer received	78afa995795e4d85b5d9ceeca43f5fef	0	NaN	bogo_3	NaN	5.0	7.0	bogo	5.0	1.0	1.0	1.0	0.0
offer received	e2127556f4f64592b11af22de27a7932	0	NaN	discount_4	NaN	10.0	7.0	discount	2.0	1.0	1.0	1.0	0.0
offer received	389bc3fa690240e798340f5a15918d5c	0	NaN	bogo_4	NaN	5.0	5.0	bogo	5.0	1.0	1.0	1.0	1.0
offer received	2eeac8d8feae4a8cad5a6af0499a211d	0	NaN	informational_1	NaN	0.0	4.0	informational	0.0	1.0	1.0	1.0	0.0
offer received	aa4862eba776480b8bb9c68455b8c2e1	0	NaN	discount_1	NaN	20.0	10.0	discount	5.0	1.0	1.0	0.0	0.0

How many customers have received, viewed and completed offers?

All offers are sent to almost equal numbers of customers. But the response rate is different.



Here we can see that the offer completed count is more than the offer viewed count for offers like bogo_3, discount_1, discount_4 . It means that many customers have completed the offer without knowing that they have received it. So, we need to find customers who actually have seen the offer and then completed it.

Customers who have responded to the offers

A customer is said to be influenced by an offer only if he/she has viewed the offer received. A customer who received an offer, never viewed it but completed the offer, is not influenced by the offer.

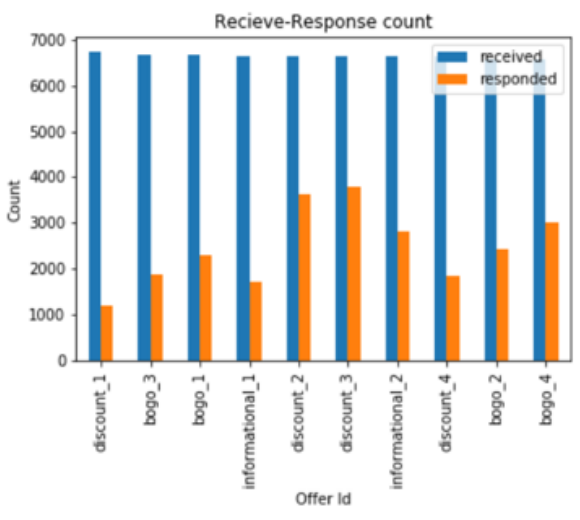
For discount and bogo offers, I have grouped the data by 'customer_id' and 'offer_id'. For each group, if there is a 'offer received' event, calculate the validity period (using 'time' and 'duration') of that offer. Then, if there are 'offer viewed' and 'offer completed' events within this validity period, we can say that the customer is influenced by the offer. We cannot use this method for informational offers, as there is no 'offer completed' event. So it is assumed that if a customer is influenced by an informational offer, if he/she viewed the offer and made a transaction (with amount greater than 'difficulty' level) within the validity period of that offer.

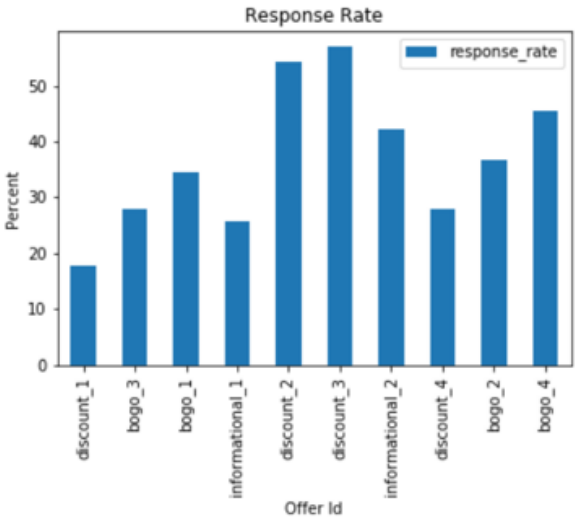
I have created a response table for customer_id with the offer_id and offer_type they have completed.

	customer_id	offer_id	offer_type
0	0011e0d4e6b944f998e987f904e8c1e5	bogo_3	bogo
1	0011e0d4e6b944f998e987f904e8c1e5	discount_1	discount
2	0011e0d4e6b944f998e987f904e8c1e5	discount_2	discount
3	0020c2b971eb4e9188eac86d93036a77	bogo_2	bogo
4	0020c2b971eb4e9188eac86d93036a77	discount_3	discount

Received and responded count and response rate for different offer ids:

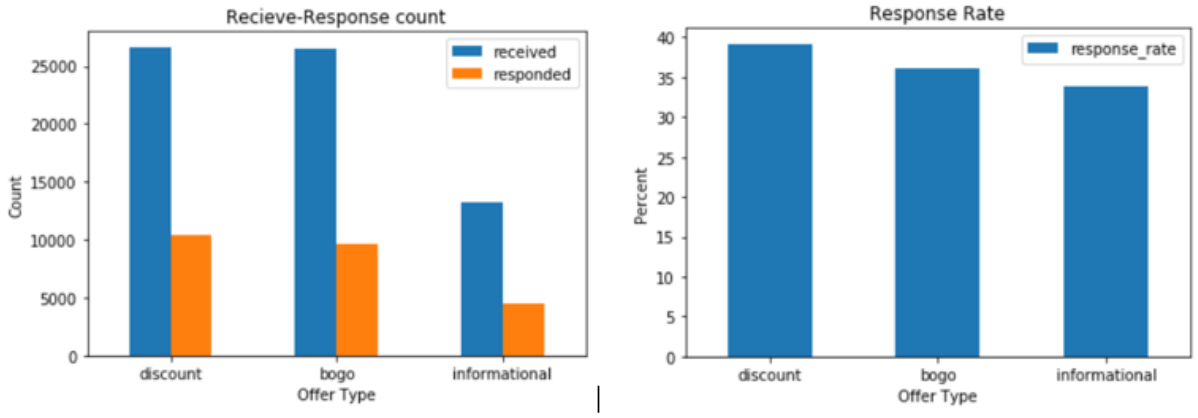
offer_id	received	responded	response_rate
discount_1	6726	1201	17.856081
bogo_3	6685	1861	27.838444
bogo_1	6683	2300	34.415682
informational_1	6657	1708	25.657203
discount_2	6655	3618	54.365139
discount_3	6652	3784	56.885147
informational_2	6643	2798	42.119524
discount_4	6631	1848	27.869100
bogo_2	6593	2422	36.735932
bogo_4	6576	2993	45.513990





Received and responded count and response rate for different offer types:

offer_type	received	responded	response_rate
discount	26664	10451	39.195170
bogo	26537	9576	36.085466
informational	13300	4506	33.879699

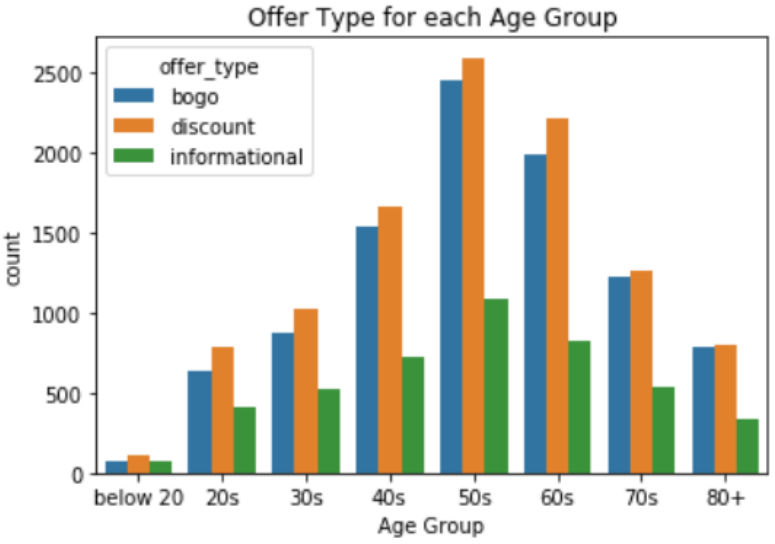


All types of offers have a response rate of above 33%, and discount rate tops the list.

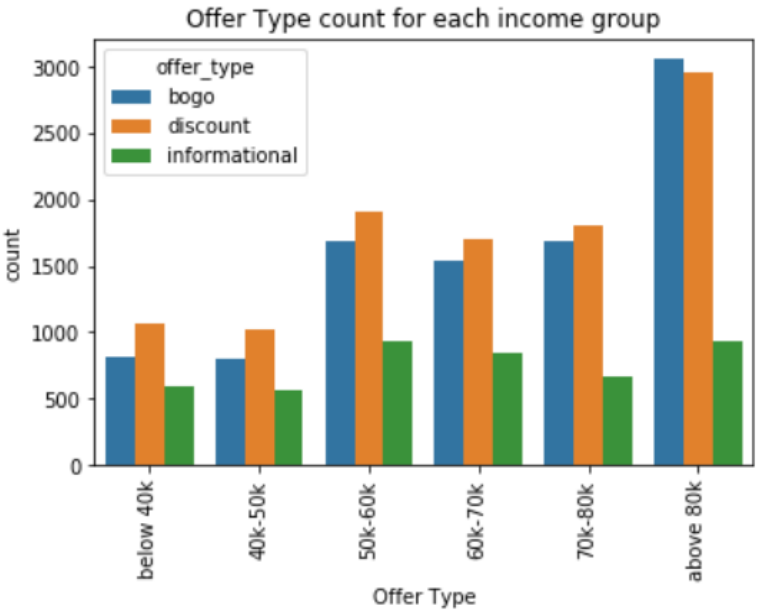
Which type of offers is preferred by different demographic groups?

To find this, I have merged the response table with the customer profile table.

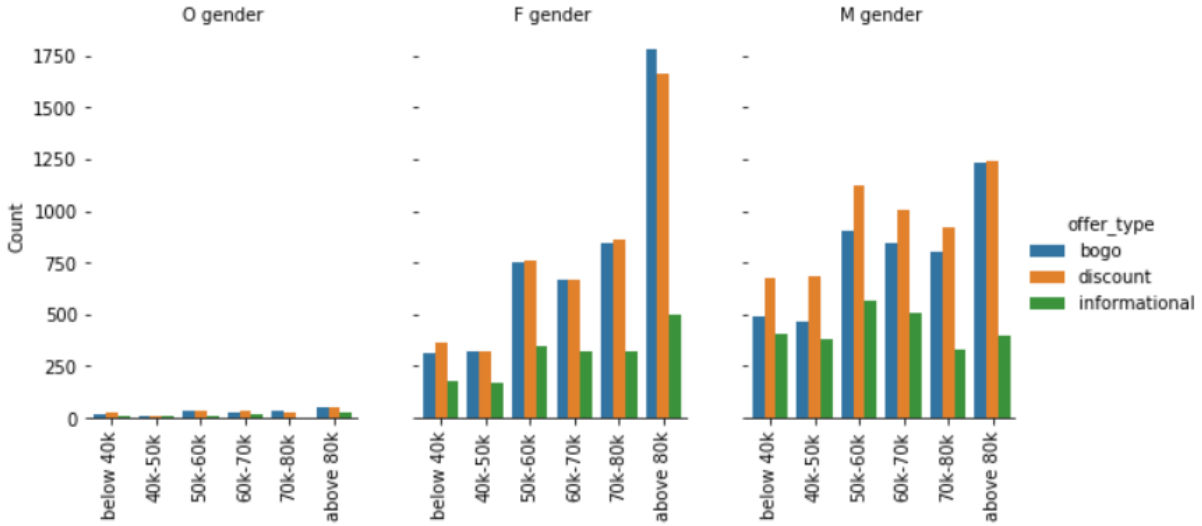
Customers in the age group 50s and 60s are more responsive to offers.



Customers with income above 80k respond more to all types of offer.

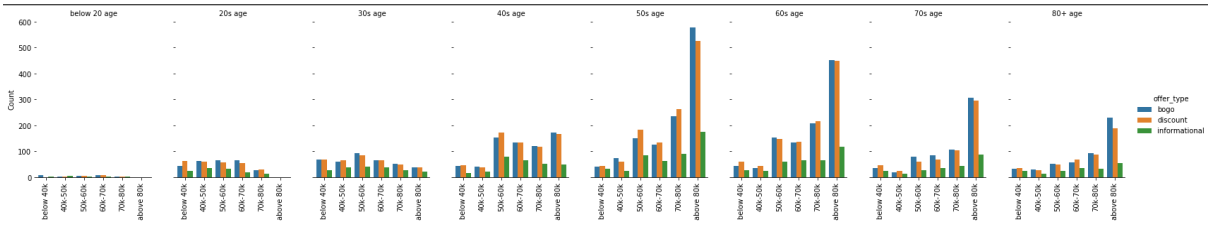


Customers with income greater than 80k are more responsive to offers, irrespective of their gender. For all three types of offers, female users with income greater than 80k respond more. For 'bogo' and 'discount' offers, male users with income greater than 80k respond more, whereas for informational offers, male users in the income category 50k-60k respond more.

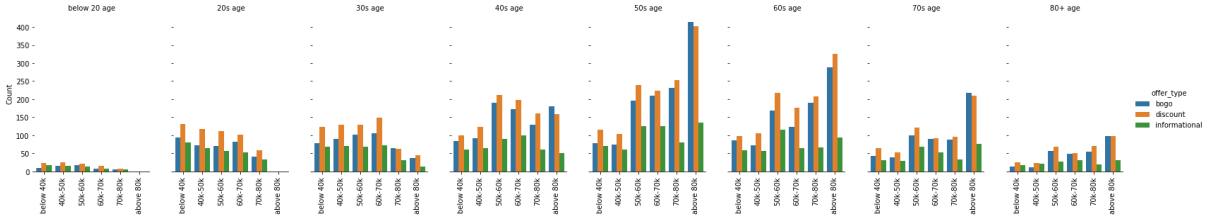


From the age and income-wise distribution of offers for each gender, we can identify the offer type that excites a particular customer based on demographic details.

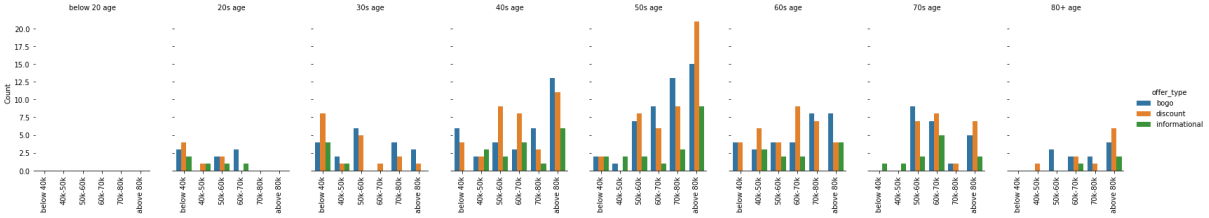
Female customers:



Male customers:



Other customers:



Data Modelling

A machine learning model is created to predict which type of offers should be given to a customer. Data is prepared by converting the categorical columns into numerical. The feature variable consists of 'age', 'year_of_joining', 'gender' and income' and the target variable is 'offer_type'. Since it is a classification problem, accuracy score is taken as the evaluation metric.

The data is splitted into training and test sets and models are built for the following classifiers.

- *Logistic Regression*
- *Decision Tree*
- *Random Forest*
- *Neural Net*
- *AdaBoost*

The accuracy score for training and test sets are as below:

classifier	train_acc	test_acc
Logistic Regression	0.428610	0.433975

Decision Tree	0.437578	0.431366
Random Forest	0.442687	0.427291
Neural Net	0.377140	0.369090
AdaBoost	0.434861	0.430388

Here we can see that the accuracy is low for both training and test set, but comparable. Even with best parameters obtained by GridSearchCV, there isn't any noticeable improvement in accuracy. The main reason for this low accuracy is that our dataset is highly imbalanced. With a more balanced dataset we can get an improved accuracy.

Conclusion

The problem I tried to solve is to find out what we should offer to a customer, based on the demographic details. The more challenging part was to identify the customers who have actually responded to an offer they have received. Once it was clear, then it was a data analysing task which gave the clear information on who responded to what.
I have also built a machine learning model to predict the offer type. But the results were poor, as our dataset was not very balanced. If we have a balanced dataset with a large number of records, then we will be able to get a better result.

For more details on the analysis, check this github repository

---