Starbucks Capstone Project Report

for

Data Scientist Nanodegree Program

Contents

1. Project Background and Description

As a part of learning, Udacity partnered with Starbucks to provide a real-world business problem and simulated data mimicking their customer behavior. Analyzing of data using different AI/ML model is now part of IT world to provide solution and support to business and help them to increase their outcome.

It is very important to understand customer behavior and take actions based on data and this is a key for company success and to earn profit. Right now, companies are only focusing on customer and developing product accordingly

1. What people like?
2. how much they want to pay?
3. what is the gender and age of those people who are interested?

These are very few questions on which companies are working, and the find answer we have to understand historical data for which we have to implement model and to build algorithms according to those Historical data to maximize Companies s' profits.

Starbucks is one of the most well-known companies in the world: a coffeehouse chain with more than 30 thousand stores all over the world. It strives to give his customers always the best service and the best experience. Starbucks has successfully developed a mobile application platform to achieve this. Starbucks use to sends out an offer to users of the mobile app. An offer can be merely an advertisement of a drink or an actual offer such as a discount or BOGO (buy one get one free). This project is focused on tailoring the promotional offers for customers based on their responses to the previous offers and find out which of them are most likely to respond to an offer.

Market campaigns have associated costs. Hence, to be considered a successful campaign, it must generate profit higher than that initial cost. That means, companies expect to have a return on investment (ROI) as high as possible. It is very important for companies to understand to whom to send offer and to whom not as sending offer to customer that are not likely to buy their product. Whereas companies also like to attract new customer with their new marketing campaigns.

Also, it is very important now a days to reward your customer to retain them and to share offer which encourage them to perform business with you. These are few aspects which we want to cover in project and try to find out solution which will cater most of our questions.

2. Problem Statement

Starbucks wants to find a way to give to each customer the right in-app special offer. There are 3 different kinds of offers: Buy One Get One (BOGO), classic Discount or Informational (no real offer, it provides information) on a product. Our goal is to analyze historical data about app usage and offers / orders made by the customer to develop an algorithm that associates each customer to the right offer type. The aim is to create an Analysis model that will predict whether a customer will complete an offer that is sent to him or not. That is, how likely will the customer accept the offer that is sent to them. we will do our statistics analysis and data visualization to understand the role of the features which controlling our model.

3. 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 (ie 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

4. Deliverables

- ✓ data
  - 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
- ✓ clean.csv #File created during processing and analysis
- ✓ combined.csv #File created during processing and analysis
- ✓ final.csv #File created during processing and analysis
- ✓ README.md
- ✓ Starbucks_Capstone_notebook.html #html version of notebook
- ✓ Starbucks_Capstone_notebook.ipynb #notebook file used for this project.

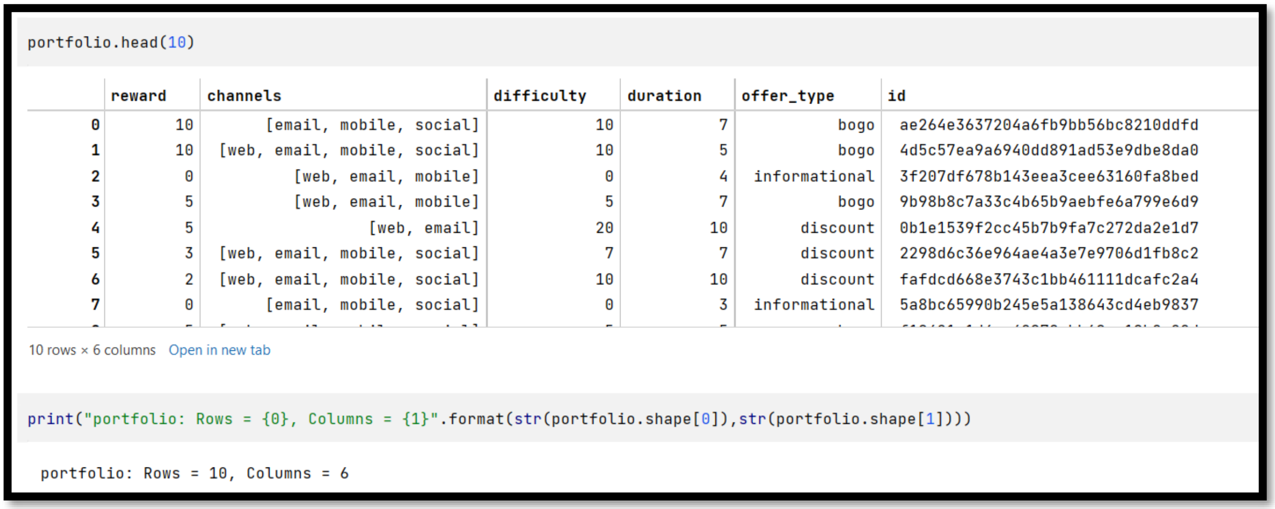
5. Requirements

This project uses Python 3.6 and the following necessary libraries:

- pandas
- matplotlib
- seaborn
- numpy
- progressbar2
- scikit-plot
- sklearn

6. Dataset Analysis

1. portfolio.json



```
portfolio.head(10)
```

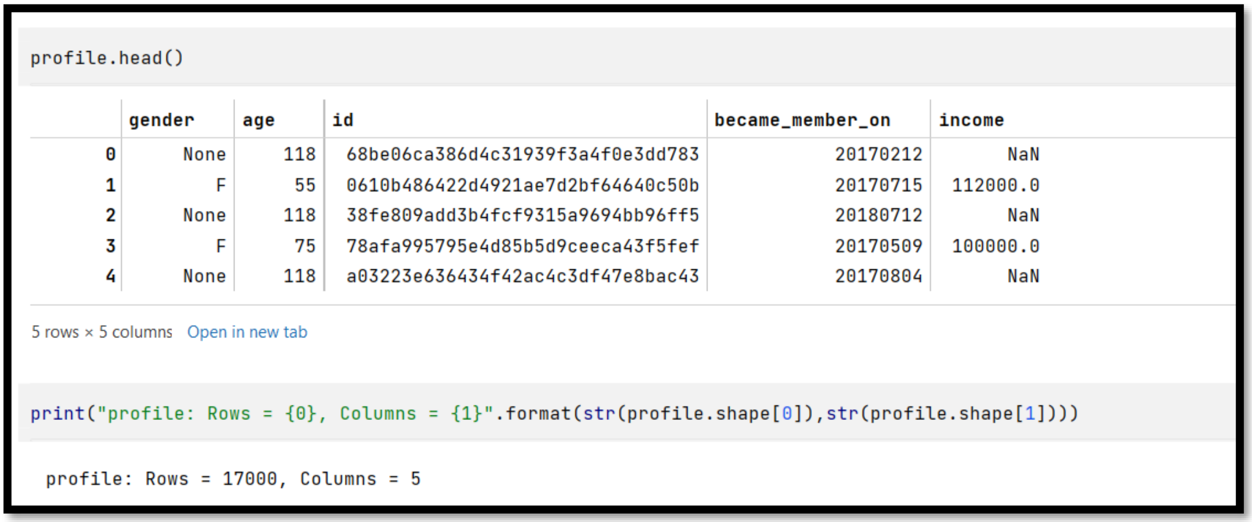
	reward	channels	difficulty	duration	offer_type	id
0	10	[email, mobile, social]	10	7	bogo	ae264e3637204a6fb9bb56bc8210ddfd
1	10	[web, email, mobile, social]	10	5	bogo	4d5c57ea9a6940dd891ad53e9dbe8da0
2	0	[web, email, mobile]	0	4	informational	3f207df678b143eea3cee63160fa8bed
3	5	[web, email, mobile]	5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

10 rows × 6 columns Open in new tab

```
print("portfolio: Rows = {0}, Columns = {1}".format(str(portfolio.shape[0]),str(portfolio.shape[1])))
```

```
portfolio: Rows = 10, Columns = 6
```

2. profile.json



```
profile.head()
```

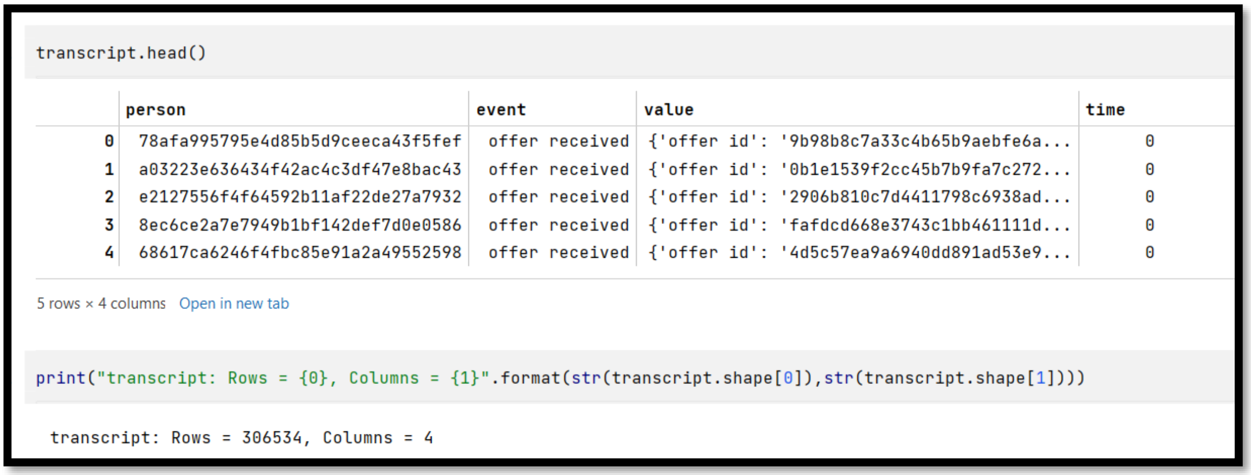
	gender	age	id	became_member_on	income
0	None	118	68be06ca386d4c31939f3a4f0e3dd783	20170212	NaN
1	F	55	0610b486422d4921ae7d2bf64640c50b	20170715	112000.0
2	None	118	38fe809add3b4fcf9315a9694bb96ff5	20180712	NaN
3	F	75	78afa995795e4d85b5d9ceeca43f5fef	20170509	100000.0
4	None	118	a03223e636434f42ac4c3df47e8bac43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

5 rows × 5 columns Open in new tab

```
print("profile: Rows = {0}, Columns = {1}".format(str(profile.shape[0]),str(profile.shape[1])))
```

```
profile: Rows = 17000, Columns = 5
```

3. transcript.json



```
transcript.head()
```

	person	event	value	time
0	78afa995795e4d85b5d9ceeca43f5fef	offer received	{'offer id': '9b98b8c7a33c4b65b9aebfe6a...	0
1	a03223e636434f42ac4c3df47e8bac43	offer received	{'offer id': '0b1e1539f2cc45b7b9fa7c272...	0
2	e2127556f4f64592b11af22de27a7932	offer received	{'offer id': '2906b810c7d4411798c6938ad...	0
3	8ec6ce2a7e7949b1bf142def7d0e0586	offer received	{'offer id': 'fafdcd668e3743c1bb461111d...	0
4	68617ca6246f4fbc85e91a2a49552598	offer received	{'offer id': '4d5c57ea9a6940dd891ad53e9...	0

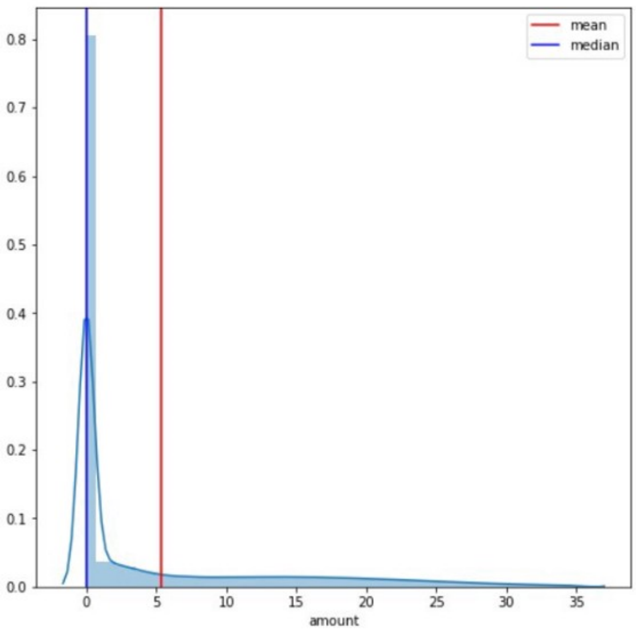
5 rows × 4 columns Open in new tab

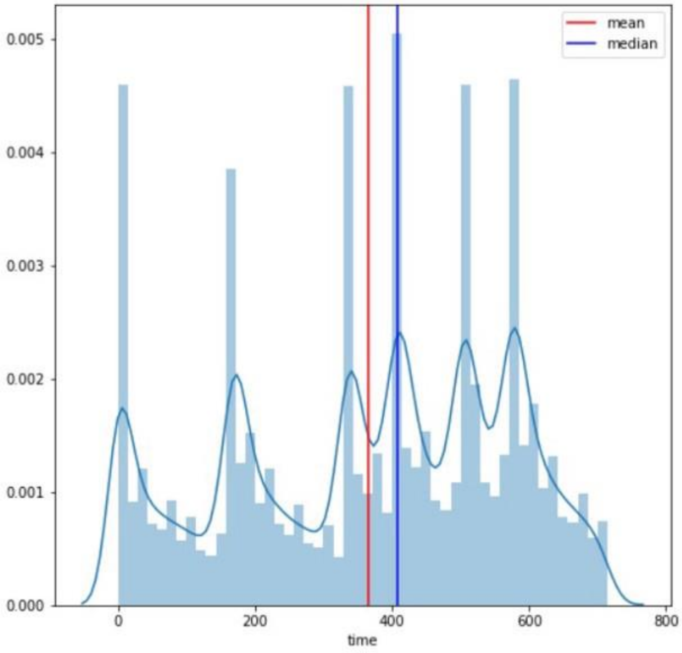
```
print("transcript: Rows = {0}, Columns = {1}".format(str(transcript.shape[0]),str(transcript.shape[1])))
```

```
transcript: Rows = 306534, Columns = 4
```

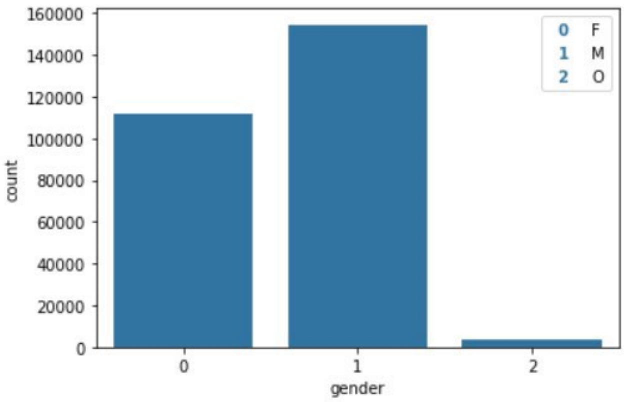
7. Exploratory Data Analysis

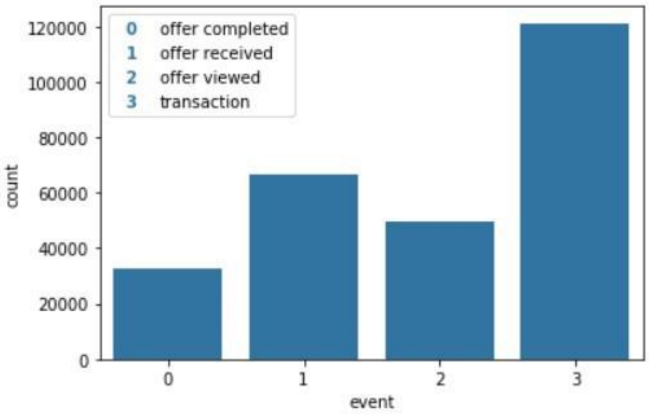
a. Univariate Analysis

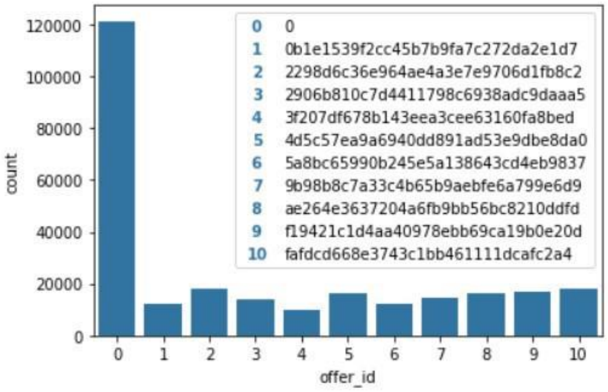


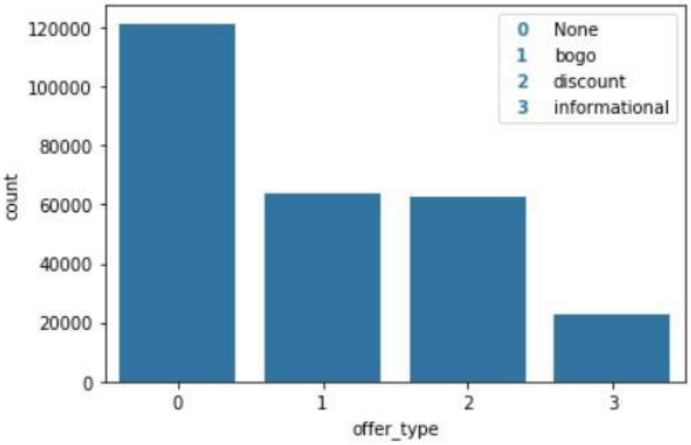


from the graph we can see that after data binning, frequency distribution has some outliner in the amount.



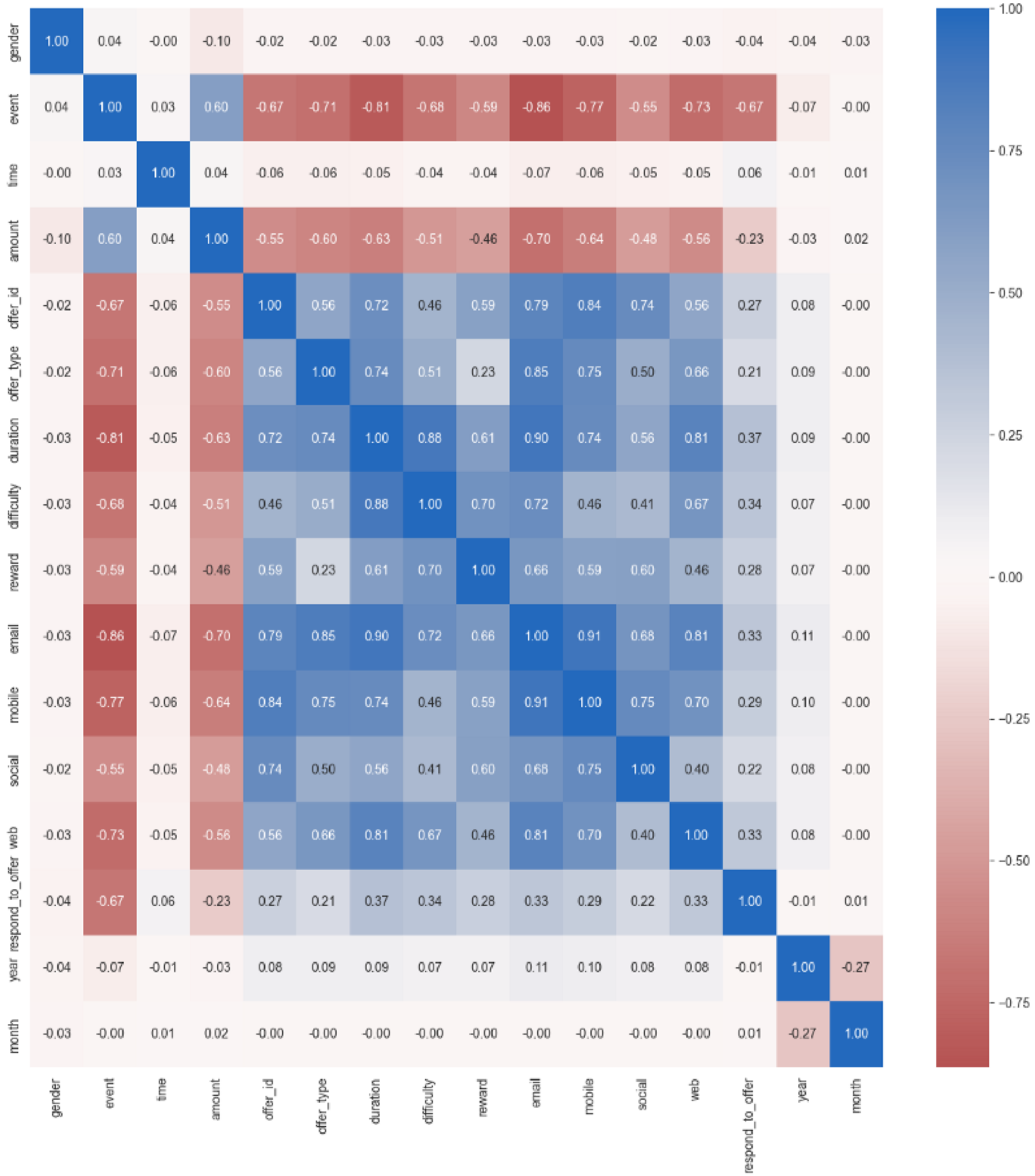






from the above graphs we can conclude that Female are using less app than Male, also very few have opted for offer and most of app user are not getting offers.

b. Bivariate/Multivariate Analysis

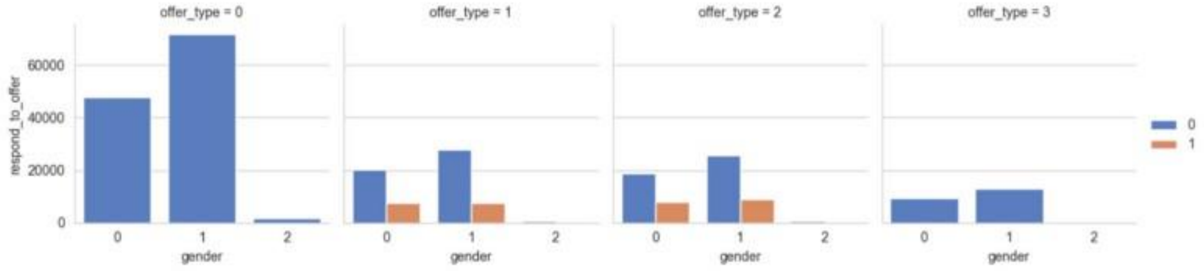


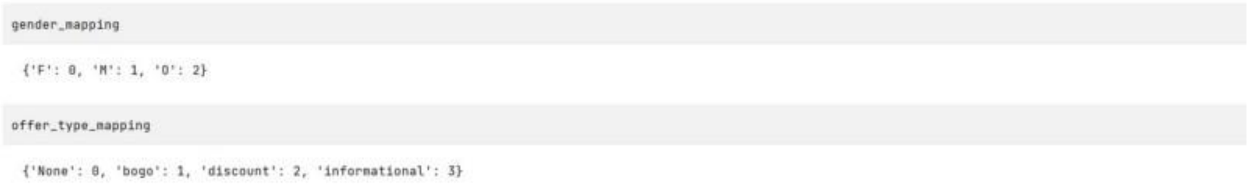
Conclusion for Response to Offer:

1. Event is highly negatively correlated with response_to_offer.

2. duration, difficulty, reward, email, moobile, social and web are slightly positively correlated with response_to_offer.

8. Exploratory Data Analysis





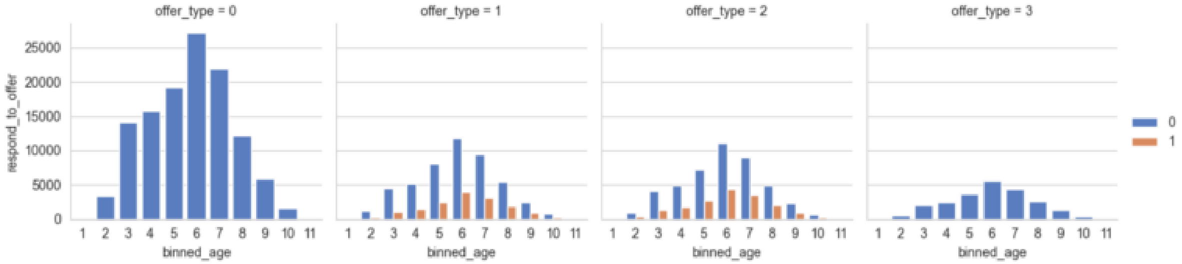
```
gender_mapping
```

```
{'F': 0, 'M': 1, 'O': 2}
```

```
offer_type_mapping
```

```
{'None': 0, 'bogo': 1, 'discount': 2, 'informational': 3}
```

from the graph we can see that both Male and female are completed the offer but mostly offer type ‘BOGO’ and ‘discount’.

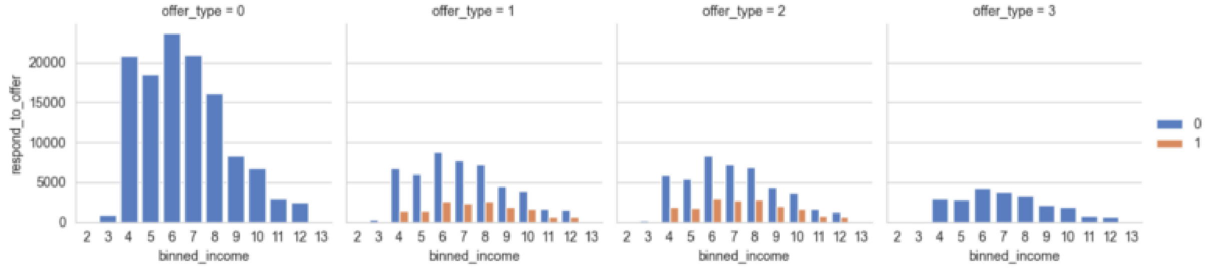


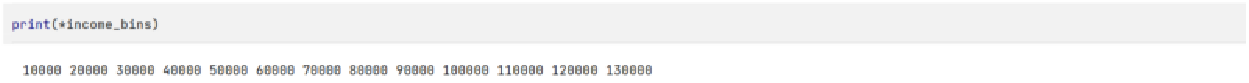
1. People of all ages respond almost equally to ‘BOGO’ and ‘Discount’

2. People of age 50–60 complete the offer — Discount most,

3. Then People of age 60–70 complete the offer — Discount most

4. Offer Type BOGO and Discount has almost similar distribution of response across different age groups.



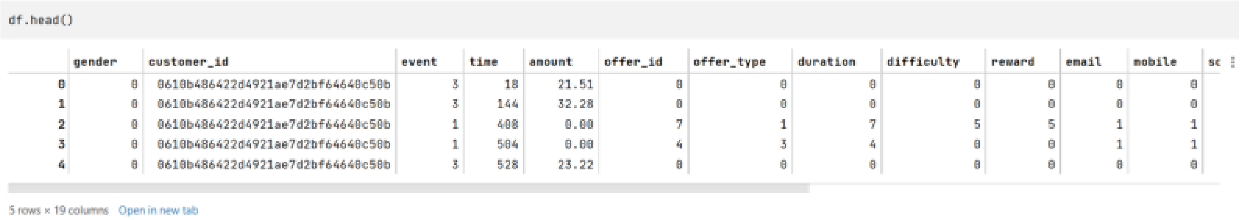


```
print(*income_bins)
```

```
10000 20000 30000 40000 50000 60000 70000 80000 90000 100000 110000 120000 130000
```

From graph we can see that people with income between 60000 and 80000 spend most on offer type ‘BOGO’ and ‘Discount’.

9. Model Building

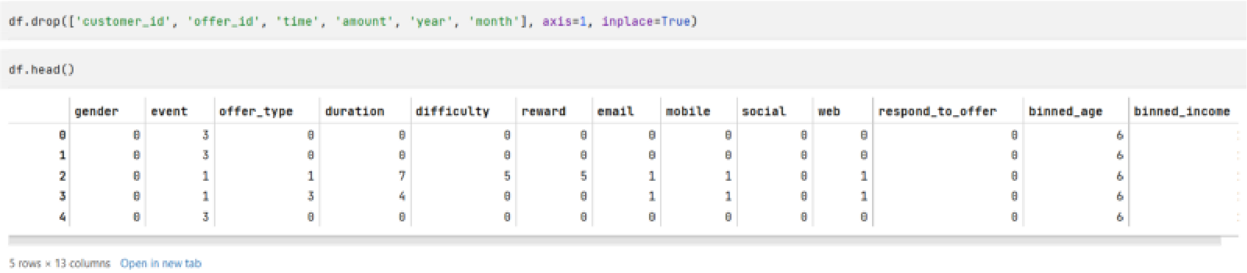


```
df.head()
```

	gender	customer_id	event	time	amount	offer_id	offer_type	duration	difficulty	reward	email	mobile	sc
0	0	0610b486422d4921ae7d2bf64640c50b	3	18	21.51	0	0	0	0	0	0	0	
1	0	0610b486422d4921ae7d2bf64640c50b	3	144	32.28	0	0	0	0	0	0	0	
2	0	0610b486422d4921ae7d2bf64640c50b	1	408	0.00	7	1	7	5	5	1	1	
3	0	0610b486422d4921ae7d2bf64640c50b	1	504	0.00	4	3	4	0	0	1	1	
4	0	0610b486422d4921ae7d2bf64640c50b	3	528	23.22	0	0	0	0	0	0	0	

5 rows × 19 columns Open in new tab

Drop customer_id and offer_id values. Also, as time and amount will be unknown for new customer so we can drop. Drop year and month as they are not correlate much with respond_to_offer



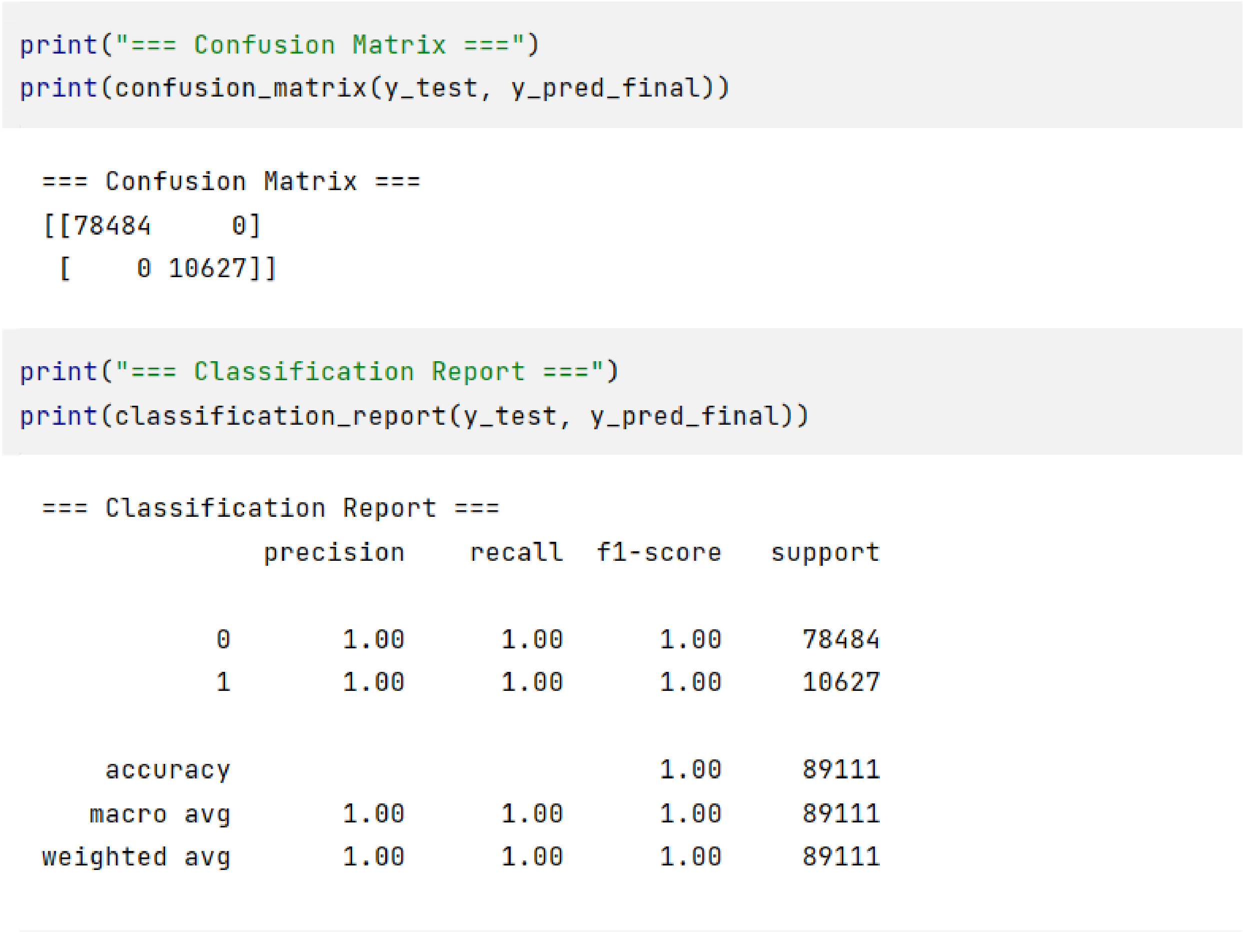
```
df.drop(['customer_id', 'offer_id', 'time', 'amount', 'year', 'month'], axis=1, inplace=True)
```

```
df.head()
```

	gender	event	offer_type	duration	difficulty	reward	email	mobile	social	web	respond_to_offer	binned_age	binned_income
0	0	3	0	0	0	0	0	0	0	0	0	6	
1	0	3	0	0	0	0	0	0	0	0	0	6	
2	0	1	1	7	5	5	1	1	0	1	0	6	
3	0	1	3	4	0	0	1	1	0	1	0	6	
4	0	3	0	0	0	0	0	0	0	0	0	6	

5 rows × 13 columns Open in new tab

10. Hyperparameter Tuning



```
print("=== Confusion Matrix ===")
print(confusion_matrix(y_test, y_pred_final))
```

```
=== Confusion Matrix ===
[[78484     0]
 [    0 10627]]
```

```
print("=== Classification Report ===")
print(classification_report(y_test, y_pred_final))
```

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     78484
           1       1.00      1.00      1.00     10627

    accuracy                           1.00     89111
   macro avg       1.00      1.00      1.00     89111
weighted avg       1.00      1.00      1.00     89111
```

11. Conclusion to Problem statement

While working on dataset provided for this project, we found that
1. Male and Female almost equally complete the offer. So offers should be sent equally among them.

2. The two most completed offer type are ‘BOGO’ and ‘Discount’. So these two should be sent to more people.

3. People of age 50–70 of income between 60000–90000 respond most to offers type ‘BOGO’ and ‘Discount’. So, it will be good to send BOGO and Discount offers to these people.

During analysis we have used Model Random Forest Classifier with hyperparameter tuning to predict whether a customer will complete an offer by making transaction after viewing the offer with the accuracy of 1. I may be getting an accuracy of 1 due to considering only the most important features and dropping all unnecessary features.

12. Conclusion to Problem statement

Thanks to Udacity Data Scientist Nanodegree program