Udacity Data Scientist Capstone Project: Starbucks dataset

Nguyen Thi Thuy

1. Project Definition

This project involves using data analysis skills to enhance the operations of a Starbucksstore. This report presents findings to Starbucks management to provide actionable insights for improving the store's performance.

This will collect data, clean and then predict the purchase offer to which a possible higher level of response or user actions like ‘offer received’, ‘offer viewed’, ‘transaction’ and ‘offer’.

2. Analysis

There are three datasets provided including:

+ Portfolio.json: consists of 10 rows (offers) x 6 columns storing offers sent during 30-day test period.

- channels: (list) web, email, mobile, social

- difficulty: (numeric) money required to be spent to receive reward

- duration: (numeric) time for offer to be open, in days

- id: (string/hash)

- offer_type: (string) bogo, discount, informational

● In a BOGO offer, a user needs to spend a certain amount to get a reward equal to that threshold amount.

● In a discount, a user gains a reward equal to a fraction of the amount spent.

● In an informational offer, there is no reward, but neither is there a requisite amount that the user is expected to spend.

- reward: (numeric) money awarded for the amount spent

	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

+ Profile.json: demographic data for each customer (17000 users x 5 fields)

- age: (numeric) missing value encoded as 118

- became_member_on: (string)

- gender: (categorical) M, F, O, or null

- id: (string/hash)

- income (numeric)

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

+ Transcript.json: records for transactions, offers received, offers viewed, and offers completed (306648 events x 4 fields)

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

- event: (string) offer received, offer viewed, transaction, offer completed

- person: (string/hash)

- time: (numeric) hours after start of test

- value: (dictionary) different values depending on event type

  - offer id: (string/hash) not associated with any "transaction"
  - amount: (numeric) money spent in "transaction"
  - reward: (numeric) money gained from "offer completed"

All dataset are needed to be preprocessed and cleaned for further analysis. The target features for analysis are offer_success, percent_success.

+ For Portfolio.json:

- convert the column 'Channels' into 4 different channels on the basis of different types of channel with binary values.

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

- renaming id column name to offer_id, one-hot encoding of channels and offer_type columns

+ For Profile.json:

- Null values in column gender will be changed to unknown
- renaming id column name to customer_id,
- impute random incomes into income
- replace special values in age column to a random sample
- convert became_member_on to days_as_member

	age	id	income	F	M	O	unknown	days_as_member
0	50	68be06ca386d4c31939f3a4f0e3dd783	105000.0	0	0	0	1	2665
1	55	0610b486422d4921ae7d2bf64640c50b	112000.0	1	0	0	0	2512
2	80	38fe809add3b4fcf9315a9694bb96ff5	56000.0	0	0	0	1	2150
3	75	78afa995795e4d85b5d9ceeca43f5fef	100000.0	1	0	0	0	2579
4	74	a03223e636434f42ac4c3df47e8bac43	67000.0	0	0	0	1	2492

+ For Transcript.json:

- pull out values of offer_id and break into columns
- create a new column showing transaction

	person	offer_id	offer_recieved	offer_viewed	offer_completed	reward	transaction_amount	net_transaction
78afa995795e4d85b5d9ceeca43f5fef	9b98b8c7a33c4b65b9aebfe6a799e6d9	1.0	1.0	1.0	5.0	37.67	32.67	
78afa995795e4d85b5d9ceeca43f5fef	5a8bc65990b245e5a138643cd4eb9837	1.0	1.0	0.0	5.0	49.39	44.39	
78afa995795e4d85b5d9ceeca43f5fef	ae264e3637204a6fb9bb56bc8210ddfd	1.0	1.0	1.0	10.0	48.28	38.28	
78afa995795e4d85b5d9ceeca43f5fef	f19421c1d4aa40978ebb69ca19b0e20d	1.0	1.0	1.0	5.0	48.28	43.28	
a03223e636434f42ac4c3df47e8bac43	0b1e1539f2cc45b7b9fa7c272da2e1d7	1.0	1.0	0.0	5.0	1.09	-3.91	

+ Preprocess the 3 data sets using various methods.

+ Merge 3 data sets

	net_transaction	offer_completed	offer_id	offer_recieved	offer_viewed	person	reward_x
0	32.67	1.0	9b98b8c7a33c4b65b9aebfe6a799e6d9	1.0	1.0	78afa995795e4d85b5d9ceeca43f5fef	5.0
1	13.42	1.0	9b98b8c7a33c4b65b9aebfe6a799e6d9	1.0	1.0	e2127556f4f64592b11af22de27a7932	5.0
2	-1.95	0.0	9b98b8c7a33c4b65b9aebfe6a799e6d9	1.0	1.0	68617ca6246f4fbc85e91a2a49552598	2.0
3	-5.00	0.0	9b98b8c7a33c4b65b9aebfe6a799e6d9	1.0	1.0	389bc3fa690240e798340f5a15918d5c	5.0
4	10.63	1.0	9b98b8c7a33c4b65b9aebfe6a799e6d9	1.0	1.0	389bc3fa690240e798340f5a15918d5c	5.0

transaction_amount	age	income	...	difficulty	duration	reward_y	web	email	mobile	social	bogo	discount	informational
37.67	75	100000.0	...	5	7	5	1	1	1	0	1	0	0
18.42	68	70000.0	...	5	7	5	1	1	1	0	1	0	0
0.05	71	73000.0	...	5	7	5	1	1	1	0	1	0	0
0.00	65	53000.0	...	5	7	5	1	1	1	0	1	0	0
15.63	65	53000.0	...	5	7	5	1	1	1	0	1	0	0

3. Methodology

After preprocessing and cleaning data, the predict model is created.

The data set is split into train data set and test data set to train and test the model, and then print metrics related to the ability of the model.

OLS Regression Results

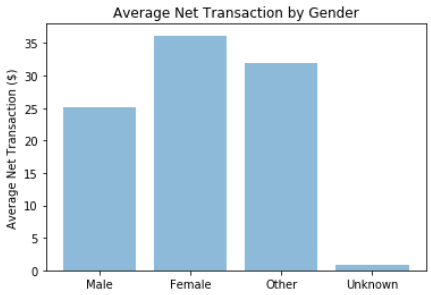
Dep. Variable:	net_transaction	R-squared:	0.107
Model:	OLS	Adj. R-squared:	0.106
Method:	Least Squares	F-statistic:	617.4
Date:	Thu, 18 Jul 2019	Prob (F-statistic):	0.00
Time:	21:47:04	Log-Likelihood:	-3.0833e+05
No. Observations:	56895	AIC:	6.167e+05
Df Residuals:	56883	BIC:	6.168e+05
Df Model:	11		
Covariance Type:	nonrobust		

Finally, many variables are tested using the model and then the data is visualized.

4. Results

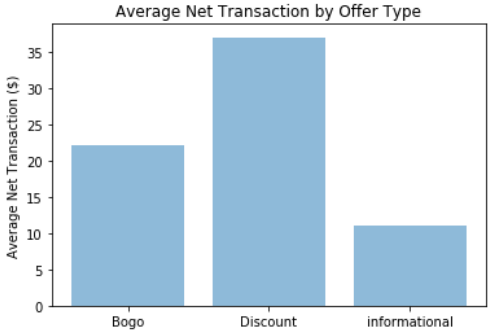
+ For the Net Transactions: The relationships between net transactions and days, gender, income as member, and offer type are visualized.

- Average Net Transaction by Gender:



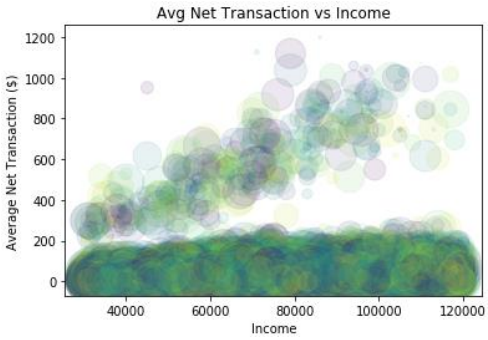
From the graph we can see that the average net transaction of those who's gender is unknown is close to zero. Females on average spend most while males spend least.

- Average Net Transaction by Offer Type:



Net transactions are highest on average for discounts and lowest for informationals.

- Avg Net Transaction vs Income:



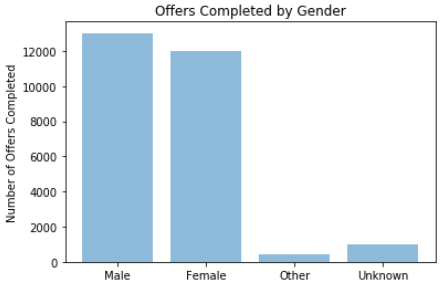
Higher net transactions have a higher probability given a higher income.

+ For Offers Completed: The below visualizations show the relationships between offers completed and income, gender, age, offer type, and days as member.

- Offers Completed by Offer

It can be seen that the numbers of Bogo and dicounts are similar.

- Offers Completed by Gender:



The graph shows that the number of offers done by Males and females are similar.

+ The relationships between net_transaction and oder_completed:

	age	M	F	O	unknown	income	days_as_member	difficulty	duration	reward_y	bogo	discount	informational	tr
age	1	-0.12	0.12	0.0045	0.0058	0.25	0.0082	0.012	0.016	-0.013	-0.012	0.016	-0.0045	
M	-0.12	1	-0.73	-0.12	-0.39	-0.17	0.04	-0.0074	-0.0065	0.0094	0.0056	-0.0068	0.0013	
F	0.12	-0.73	1	-0.088	-0.3	0.19	-0.01	0.0032	0.0025	-0.00091	0.00059	0.0022	-0.0036	
O	0.0045	-0.12	-0.088	1	-0.047	-0.0016	-0.013	-0.0022	0.0018	-0.011	-0.008	0.0014	0.0087	
unknown	0.0058	-0.39	-0.3	-0.047	1	-0.016	-0.04	0.007	0.0054	-0.0086	-0.0063	0.0063	0.00019	
income	0.25	-0.17	0.19	-0.0016	-0.016	1	0.02	0.026	0.025	-0.01	-0.012	0.024	-0.015	
days_as_member	0.0082	0.04	-0.01	-0.013	-0.04	0.02	1	-0.0051	-0.0024	-0.0035	0.0011	-0.0052	0.0052	
difficulty	0.012	-0.0074	0.0032	-0.0022	0.007	0.026	-0.0051	1	0.75	0.5	0.08	0.49	-0.73	
duration	0.016	-0.0065	0.0025	0.0018	0.0054	0.025	-0.0024	0.75	1	0.055	-0.22	0.74	-0.65	
reward_y	-0.013	0.0094	-0.00091	-0.011	-0.0086	-0.01	-0.0035	0.5	0.055	1	0.83	-0.39	-0.58	
bogo	-0.012	0.0056	0.00059	-0.008	-0.0063	-0.012	0.0011	0.08	-0.22	0.83	1	-0.7	-0.41	
discount	0.016	-0.0068	0.0022	0.0014	0.0063	0.024	-0.0052	0.49	0.74	-0.39	-0.7	1	-0.36	
informational	-0.0045	0.0013	-0.0036	0.0087	0.00019	-0.015	0.0052	-0.73	-0.65	-0.58	-0.41	-0.36	1	
transaction_amount	0.051	-0.013	0.14	0.013	-0.18	0.17	0.12	0.13	0.17	0.01	-0.038	0.13	-0.12	
net_transaction	0.05	-0.012	0.13	0.012	-0.17	0.17	0.12	0.13	0.18	-0.011	-0.057	0.15	-0.12	
offer_completed	0.059	0.0078	0.18	0.022	-0.27	0.15	0.17	0.31	0.37	0.12	0.052	0.29	-0.43	

There are a small number of correlations, The maximum is 0.37, and the minimum is 0.17.

Offer Completed:

Positive: Female, income, days_as_member, difficulty, duration, reward, discount.
Negative: Unknown gender.

Noteworthy: There is a positive correlation around age, female, and income with eachother.

Net Transaction:

Positive: Female, income, days_as_member, difficulty, duration, discount.

Negative: Unknown gender, informational.

5. Conclusion

Through this project, I tried to analyze the datasets provided by Starbucks and then built a model that can predict whether a customer would complete the offer or just view it.

I found that younger people, people with lower salaries, relatively new members, or men are less likely to complete offers.

To increase the revenue, Starbucks needs to focus more on offers featured with higher completed rate.

The code can be found in my github: https://github.com/Giga85/Starbucks_Project

Medium: https://medium.com/@giga16785/udacity-data-scientist-capstone-project-starbucks-dataset-f770febaeb1f

Thank You