Udacity final Starbucks Capstone project

Overview:

It is the Starbucks Capstone Challenge of the Data Scientist Nanodegree in Udacity. We get the dataset from the program that creates the data simulates how people make purchasing decisions and how those decisions are influenced by promotional offers. We want to make a recommendation engine that recommends Starbucks which offers should be sent to a particular customer.

Problem Statement:

The problem that I chose to solve was to build a model that predicts whether a customer will respond to an offer. My strategy for solving this problem has four steps. First, I combined the offer portfolio, customer profile, and transaction data. Each row of this combined dataset describes an offer's attributes, customer demographic data, and whether the offer was successful. Second, I assessed the accuracy and F1-score of a naive model that assumes all offers were successful. This provided me a baseline for evaluating the performance of models that I constructed. Accuracy measures how well a model correctly predicts whether an offer is successful. However, if the percentage of successful or unsuccessful offers is very low, accuracy is not a good measure of model performance. For this situation, evaluating a model's precision and recall provides better insight to its performance. I chose the F1-score metric because it is "a weighted average of the precision and recall metrics". Third, I compared the performance of logistic regression, random forest, and gradient boosting models. Fourth, I refined the hyperparameters of the model that had the highest accuracy and F1-score.

Metrics:

The metric I used this project is F1_score. we used it to compare the models performance as F1_score is usually more useful metric because it is "a weighted average of the precision and recall metrics".

After completing this project we will answer the below business questions:

1- who are responsive to coffee offers? (Gender/Age)

2- how can income affects number of responsive customers to the offers?

3- which offer is preferred to customers based on age/gender/income?

4- Knowing all above, we need to predict the future offer which will attract more customers

Data Exploration

We're given 3 data sets to work with

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

First I brought the data in, explored data structure and value statistics

	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

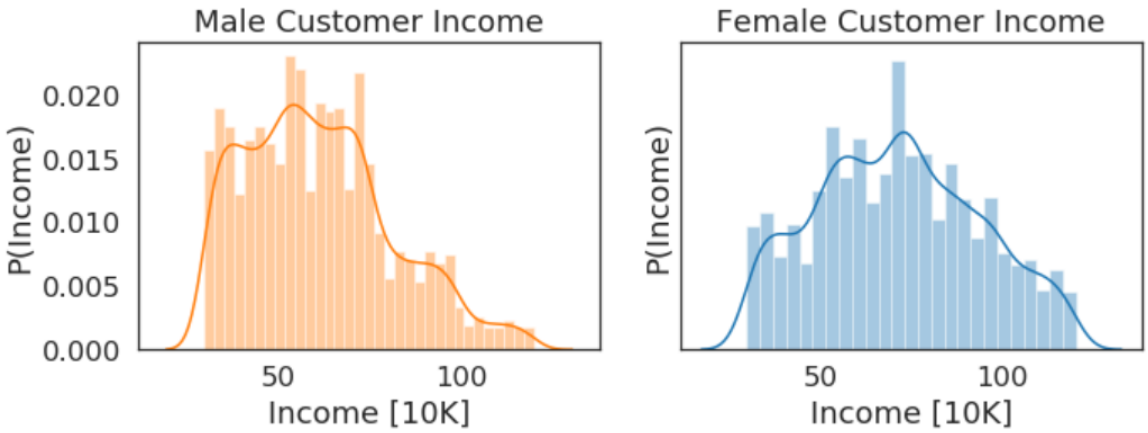
	age	became_member_on	gender	id	income
0	118	20170212	None	68be06ca386d4c31939f3a4f0e3dd783	NaN
1	55	20170715	F	0610b486422d4921ae7d2bf64640c50b	112000.0
2	118	20180712	None	38fe809add3b4fcf9315a9694bb96ff5	NaN
3	75	20170509	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0
4	118	20170804	None	a03223e636434f42ac4c3df47e8bac43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

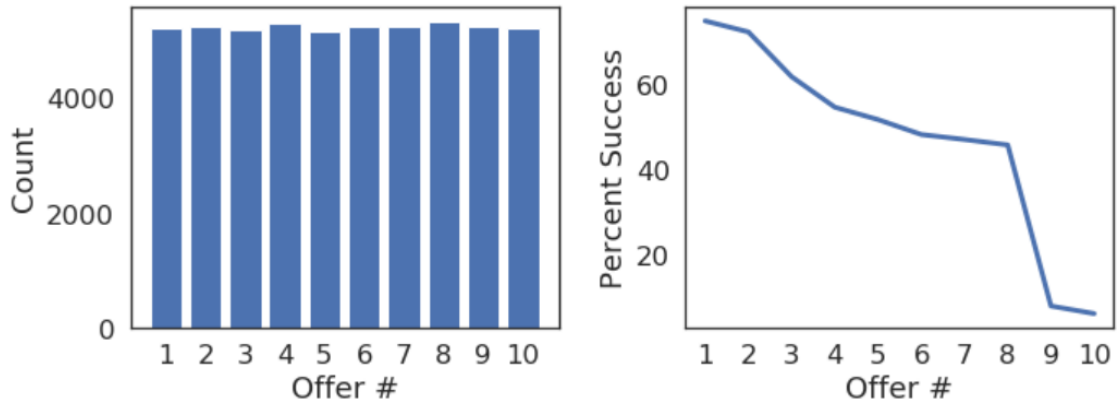
	event	person	time	value
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}
1	offer received	a03223e636434f42ac4c3df47e8bac43	0	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}
2	offer received	e2127556f4f64592b11af22de27a7932	0	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}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

Through initial data exploration, I identified multiple data issues to be taken care of at later preparation stage

- Porfolio data contains categorical values, which needs one-hot encoding process
- Profile data contains null value and extremely high age (118)
- Transcript data has mixture of offer id and amount under value column, etc.

Next we plotted data to see if any additional patterns are spotted





Based on initial learning, I'm ready to move into data preparation

Data Preparation

The goal for data preparation is to get the data ready for modeling exercise. Before you start the preparation process, you'll need to have good level of understanding on how you want to approach the modeling, in order to minimize wasted efforts in the early stage.

Since my goal is to perform classification studies to understand which top features play more important role than remaining ones, I'll need to convert data into numeric feature as much as I can.

For **portfolio** data set, I'll need to separate channel method, transpose categorical offer_type via one-hot encoding, also convert offer duration from days to hours to be consistent with time in transcript data

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

For **profile** data, I'll need to remove outlier with age of 118 years old, as well as null value in gender and income fields, convert categorical gender field, etc.

	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 data is the trickiest to handle, as it has mixture of values within value field, and depending on your needs, you could perform additional processing / calculation towards event type, time, etc.

	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'}

In this study, since my interests are in features that can be used for classifications, I wanted to focus on offer activities behaviors in connection with offer type and user profile. So I separated the offer data from transaction data, and converted them into numerical feature set.

Once I completed initial data cleansing, I turned my attention to integrate data to get it ready for modeling.

One key thing to consider here is to separate effective offer vs. others. The definition of effective offers is as follows:

· Offer needs to be completed within defined offer timeframe

· Offer needs to be viewed prior to completion

Sometimes customers would receive the offer, then never bothered to check it or without even realizing it, still went ahead to complete the offer. For this case we should consider it as ineffective offer, since customer made the qualifying purchases without taking offer into consideration.

This makes the data processing a bit more tricky, we had to look for closest viewing and completion history against the same offer they received, then make sure both events occurred in the right order, to label that as an effective offer.

	offerid	customerid	timedays	completed	received	viewed
0	9b98b8c7a33c4b65b9aebfe6a799e6d9	78afa995795e4d85b5d9ceeca43f5fef	0.0	0	1	0
1	2906b810c7d4411798c6938adc9daaa5	e2127556f4f64592b11af22de27a7932	0.0	0	1	0
2	f19421c1d4aa40978ebb69ca19b0e20d	389bc3fa690240e798340f5a15918d5c	0.0	0	1	0
3	3f207df678b143eea3cee63160fa8bed	2eeac8d8feae4a8cad5a6af0499a211d	0.0	0	1	0
4	0b1e1539f2cc45b7b9fa7c272da2e1d7	aa4862eba776480b8bb9c68455b8c2e1	0.0	0	1	0

The last step before modeling is to combine all features and predictor variable into one single data set as shown below

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

Used Features for the models:

'age'
'gender'
'income'
'year'
'month'
'discount offer'
'bogo_offer'

Prediction

With data in hand, I started out calculating model performance for Naive Predictor Performance, and intend to use the results as baseline to measure future optimized model performance.

```
Naive predictor accuracy: 0.471
Naive predictor f1-score: 0.640
```

Results

1- who are responsive to coffee offers?

we have found that the males are more responsive to coffee offers. and customers on age range 50–65 this answer is proven with charts above

2- how can income affects number of responsive customers to the offers?

obviously, the data showed that male are more responsive to the offers. and when we tried to analyse the data more, we found that male customers have a higher income than females which justify why the number of males consuming offers is higher than females

3- which offer is preferred to customers based on age/gender/income?

Males are responding more to the discount offer, but both males and females have a high response to BOGO offer and almost has the equal number of responses from both males and females. Starbucks should consider this when offering & promoting an offer (Highest response of males)

4- Knowing all above, we need to predict the future offer which will attract more customers

after analysing all the data above and applying ML models on the cleaned data sets, we decided to use Linear regression *model on predicting the preferred offer and which is more attractive to the customer to buy and the result was BOGO*

As a result of training and testing of 4 models shown above, we decided to do the prediction based on linear regression model. the model has the most reasonable F1_Score for both offers and it has an excellent performance. The most important features on this model was age, gender and income.

Improvements

To make the results even better, I think we need to improve the data. not to accept Nan values or fix it on a way or another. We can get more data so we can predict a more accurate best offer to use we can have some extra data like the city / neighbourhood /branch the transaction were done on and have different predictions for each branch. the more data Starbucks can gather from customers the more accurate results ML can provide.

Final Thoughts

I learned a great deal from Starbucks capstone project, including:

- How to think from customer and corporation perspective
- Learn to figure out the best way to approach the problem
- Construct data based on real life scenario, etc.
- Be creative in getting what you want

The data processing part took me a long time to think through and implement, which leaves less time than I desired to work on other parts. At the same time, I found the time well spent as now I have much better understanding on how to approach data processing to better support downstream modeling.

There are many other things I'd like to continue building out and enhancing upon, such as:

- Take transaction amount into consideration
- Further separate offer type to conduct individual study
- Perform more EDA study to look from different angel for demographic groups
- Build predictive model to predict the likelihood of a certain demographic group would respond to specific offer

Overall, I really enjoyed Udacity machine learning nanoprogram, and found it really helpful in my future DS journey.