Starbucks Capstone Project Proposal

Deepthi M | January 17, 2022

Udacity – AWS Machine Learning Engineer Nanodegree Program

Background

Starbucks, a leading specialty coffee brand and coffee store chain based in the USA was founded in 1971 by Gerald Baldwin, Gorden Bowker, and Ziev Siegl. Starbucks is further enhancing its customer experience with the introduction of a new Starbucks Rewards program.

There are three types of offers that can be sent: buy-one-get-one (BOGO), discount, and 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. Offers can be delivered via multiple channels.

Problem Statement

A 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 its 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.

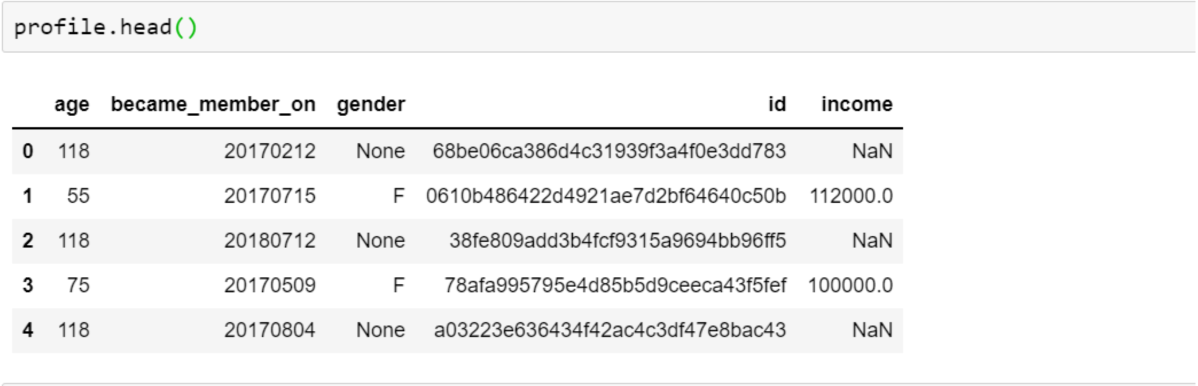
The problem statement is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type.

Dataset

profile.json

Rewards program users (17000 users x 5 fields)

- gender: (categorical) M, F, O, or null
- age: (numeric) missing value encoded as 118
- id: (string/hash)
- became_member_on: (date) format YYYYMMDD
- income: (numeric)



```
profile.head()
```

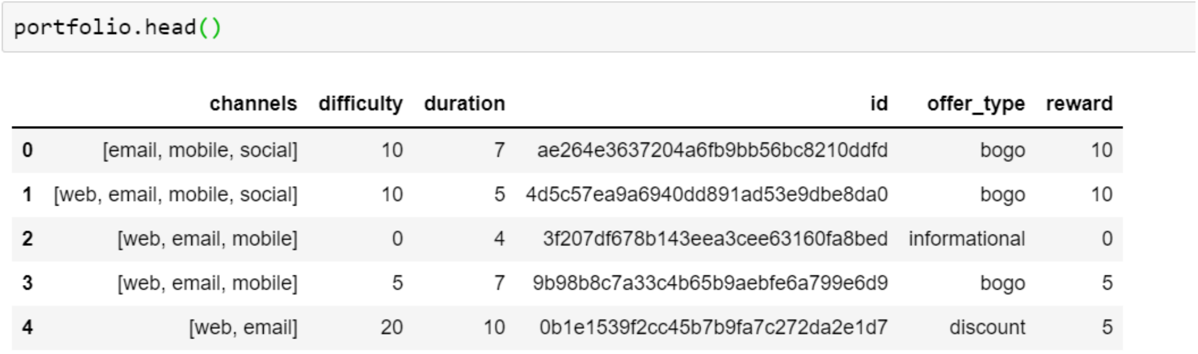
	age	became_member_on	gender	id	income
0	118	20170212	None	68be06ca386d4c31939f3a4f0e3dd783	NaN
1	55	20170715	F	0610b486422d4921ae7d2bf64640c50b	112000.0
2	118	20180712	None	38fe809add3b4fcf9315a9694bb96ff5	NaN
3	75	20170509	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0
4	118	20170804	None	a03223e636434f42ac4c3df47e8bac43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

First five rows of the profile file

portfolio.json

Offers sent during 30-day test period (10 offers x 6 fields)

- reward: (numeric) money awarded for the amount spent
- 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
- offer_type: (string) bogo, discount, informational
- id: (string/hash)



```
portfolio.head()
```

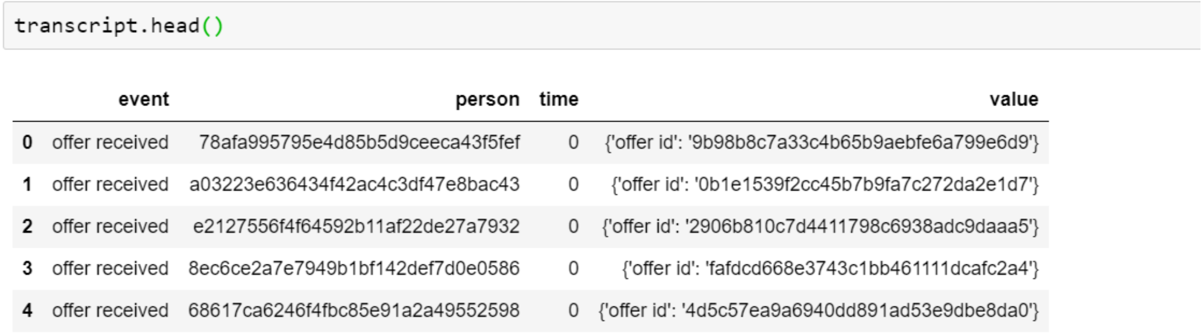
	channels	difficulty	duration	id	offer_type	reward
0	[email, mobile, social]	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10
1	[web, email, mobile, social]	10	5	4d5c57ea9a6940dd891ad53e9dbe8da0	bogo	10
2	[web, email, mobile]	0	4	3f207df678b143eea3cee63160fa8bed	informational	0
3	[web, email, mobile]	5	7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

First five rows of the portfolio file

transcript.json

Event log (306648 events x 4 fields)

- person: (string/hash)
- event: (string) offer received, offer viewed, transaction, offer completed
- 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"
- time: (numeric) hours after start of test



```
transcript.head()
```

	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'}

First five rows of the transcript file

Solution

The first step would be to perform Exploratory Data Analysis (EDA) and describe the dataset. This EDA would expose various trends, patterns and relationships that are not visibly apparent. It will also help to identify obvious errors as well as detect possible outliers.

The second step would be data visualization to understand the dataset better and get a hold of the quantity of each demographic, each type of offer or event.

To identify which groups of people are most responsive to each type of offer, and how best to present each type of offer, the benchmark model will be decided, followed by the other models to determine which model best represents the data in hand.

Benchmark Model

A baseline model provides a point of comparison for the more advanced models that will be evaluated later on. There are few requirements for a good baseline model:

- Baseline model should be simple.
- Simple models are less likely to overfit
- Baseline model should be interpretable.

The baseline model would be a quick and simple model, which is the Support Vector Machine. The model will be evaluated with svm.score ()

Models

To find out a more accurate response of a customer to any offer, I will be using the DecisionTreeClassifier using the score method again to evaluate the metrics.

Evaluation Metrics

The evaluation metric is score (). This returns the mean accuracy on the given test data and labels.

Project Flow

1. The entire Jupyter notebook will be run using an Amazon Sagemaker instance.
2. The input files are uploaded to Amazon S3.
3. Exploratory Data Analysis is performed on the dataset.
4. The dataset is visualized using various graphs and charts.
5. The dataset is then cleaned to clear out any null or empty values.
6. The dataset is also made uniform across the three files. (E.g., columns, index, etc.)
7. Various models are trained including one baseline or benchmark model.
8. Evaluating the models and then choosing the best one.
9. Summarize findings.
10. Publish project work in a detailed blog post.