Starbucks Capstone Project Proposal

Youssef Medhat | 11 May 2023

Udacity – AWS Machine Learning Engineer Nanodegree

○ Domain Background

Starbucks is an enthusiastic retailer of coffee and other beverages with its corporate headquarters in Seattle, Washington. The company is listed as the 121st Fortune 500 company for 2019. Registered users of their mobile application can order coffee for pickup while on the go, pay in-store using the app, and get rewards points. Also, this app provides these users with promos for extra points. The promotional offer could simply be a drink marketing or it could be a real deal, like a discount or a BOGO (buy one, get one free) deal. The goal of this project is to identify the customers who are most likely to respond to an offer by personalizing promotional offers for them based on their answers to prior offers.

○ Problem Statement

My objective is to evaluate which kind of offer to deliver to each user based on their response to the offers that have already been made to them. The objective is to utilize the data set provided by Starbucks, which was collected over 30 days, to address the fact that not all users receive the same offer. Also, I'll create a machine-learning model that predicts how a customer will react to an offer.

○ Datasets and Inputs

The simulated data in this data set closely resembles consumer activity on the Starbucks Rewards mobile app. Starbucks delivers offers to customers who use its mobile app every few days. The data set is provided in form of three JSON files:

- portfolio.json - containing offer ids and metadata offer (duration, type, etc.)

- profile.json - demographic data for each customer

- transcript.json - records for transactions, offers received, offers viewed, and offers completed

Below is the file's schema and an explanation of each variable:

portfolio.json

- reward: (numeric) money awarded for the amount spent
- channels: (list) web, email, mobile, social
- difficulty: (numeric) money required to be spent to receive the reward
- the duration: (numeric) time for the offer to be open, in days
- offer_type: (string) bogo, discount, informational
- id: (string/hash)

```
In [3]: portfolio.head()
```

Out[3]:

	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

profile.json

- 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)

```
In [9]: profile.head()
```

Out[9]:

	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.json

- person: (string/hash)

- event: (string) offer received, offer viewed, transaction, offer completed
- value: (dictionary) different values depending on event type
- time: (numeric) hours after the start of the test

```
In [13]: transcript.head()
Out[13]:
```

	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

The portfolio.json contains an offer_type column, which describes the types of offers that Starbucks is looking to potentially send its customers:

1) BOGO (Buy-One-Get-One): This offer enables a customer to receive an extra and equal product at no additional cost. The customer must spend a certain threshold to make this reward available.
2) Informational: This offer doesn't necessarily include a reward, but rather an opportunity for a customer to purchase a certain object given a requisite amount of money.
3) Discount: With this offer, a customer is given a reward that knocks a certain percentage off the original cost of the product they're choosing to purchase, subject to limitations.

○ Solution Statement

The offers that the customers are most interested in will be identified in order to determine which offers should be sent to them. I'll also think about exploratory data analysis to cover a few topics like:

1) most responded offer
2) response to an offer
3) age & gender groups that are greatly interested in offers

Both the overall population and the level of the individual's customized experience will be covered in this discussion.

I'll be using models like RandomForestClassifier and DecisionTreeClassifier to identify which model best represents the data we currently have to predict the proper response of a customer to an offer. Also, I'll be training these models on the standalone platform especially on a local computer's Jupyter Notebook Environment.

○ Benchmark Model

A rapid and reasonably accurate model can be used as a benchmark. To build the benchmark swiftly and reliably for binary classification machine learning tasks, I'll utilize the KNeighborsClassifier. The model's performance will be assessed using the F1 score as the assessment measure.

○ Evaluation Metrics

To evaluate the effectiveness of the strategy and identify which model produces the best outcomes, I will use the F1 score as the model metric. It can be understood as the weighted average of recall and precision. The traditional or balanced F-score (F1 score), which has a greatest value of 1 and a worst value of 0, is the harmonic mean of precision and recall.

○ Project Design

The overall outline of how I'll approach this project is as follows:

1) Setting up the workspace in the Jupyter environment.
2) The data should be cleaned up as needed for modeling.
3) Examining the data in-depth using exploratory data analysis.
4) Building a variety of models to find the one that best fits the data.
5) Utilizing evaluation metrics and benchmark models to assure rationality.
6) Write a thorough blog post to summarize the research and project work.