Starbucks Capstone Project Proposal

Udacity - Machine Learning Engineer Nanodegree

Muhammad Umar Farrukh

June 2024

1 Domain Background

Starbucks is one of the word renowned coffee, they have huge variety of hot and cold beverages. They have their main head offices in Seattle, USA. They have IOS and Android mobile applications which allow their customers to order from anywhere in the world. These apps allow their customers to find the nearest starbucks coffee shop and pick the beverage at their convenience. Their app have alot of promotional offers going on from time to time. They do alot of advertisements for different seasonal drinks and buy one get one free offers. This capstone project will help tailor the promotional offers to specific customers who would be likely to respond to that offer

Main three varieties of offers are covered in the portfolio section which are

1. BOGO referring to Buy One Get One in which user has to spend certainamount to qualify for the reward
2. Discount offer helps user gain a fractional amount of the total amount hehas spend
3. Informational offer provides information to customer about certain offerbeing activated but doesnot has any reward or amount linked with it Refrence for these three types of offers was checked using valuecount function of data frame and a snippet is provided below to show the offers available in portfolio

Similarly communication of these offers is made through four channels which is through web, email, mobile and social media. These all are digital based channels through which communication is being made but time and value is spend on these communications so an ROI is expected once these promotional offers are being sent to the customers. That is where consumer preference based marketing campaigns implemented through Artificial Intelligence will come in action [1]. Similarly in this project we will try to build a machine learning model that will tailor the personalised offer for a customer and will help to reduce down the costs along with a increased revenue. This project is very important part for my learning as it will add huge benefit for me and my career, i have also read several

other research papers such as Machine learning in marketing [3] defines the purpose and ambition of mahcine learning behind marketing campaigns, similarly another research paper [2] discusses the importance of self learning models in neuro marketing campaigns. These researches and the overall data quest led me to choose this project for my capstone project. Similarly antoher research paper [4] which discusses the campaigns launched by Starbucks on twitter or X and how the sentiments of the customers are studied using the machine learning models.

2 Problem Statement

Main Objective of this project is to tailor offers according to customers based on their responses to previous offers sent to them. All users wont be receiving the same offers, in this case a machine learning model implementation will help determine the best offer for a customer and their response to that offer. Data set consists of 30 days of promotional offers being given to the customers and how they have utilized those offers.

3 Solution Statement

Using Machine Learning techniques such as Exploratory Data Analysis, Hypothesis creation , Model Analysis and predictability will help me to achieve this task of finding most relevant offer according to customers need. Using the EDA technique i will visualize which offers were most accepted by the customers, how well they responded to it and what particular demographics are responding to most offers. Similarly hypothesis testing will help me to check different models and their outputs with the most important feature groups i have selected for that hypothesis. I ll start off with very basic models than advance to complex tree based models such as random forest or xgboost or catboost. For training purposes 80 percent of the data would be separated and trained with the basic models using aws sage maker notebook whereas 20 percent of the data would be sepertaed for testing purpose and that will be done also using standalong sage maker notebooks. In sage-maker notebooks all the cells will be executed and bench mark models will be saved from training practice to s3 bucket using pickle library and then would be reloaded and tested in the sagemaker notebook again.

4 Data Set's and Input

Data set is provided in three main json files which represents that most data is in form of unstructured data. The three main files provided are profile.json, portfolio.json and transcript.json.

Portfolio.json will help us to analyse what was offered to customer for how much duration, and what was the reward given to customer and through what channel communication which will help to analyse what is most relevant channel for communication suited for customer. Some analysis and shape of the portfolio data frame is shown below in the images. Using the value counts of offer type in portfolio helped us to know what three types of offers are available, similarly difficulty variable descriptive analysis helped us to know more about the variation of difficulty levels associated with the offers. portfolio.json 1. id (string) - offer id

2. offertype (string) - type of offer ie BOGO, discount, informational
3. difficulty (int) - minimum required spend to complete an offer
4. reward (int) - reward given for completing an offer
5. duration (int) - time for offer to be open, in days
6. channels (list of strings)

```
portfolio['offer_type'].value_counts().to_frame().style.bar()
```

	offer_type
bogo	4
discount	4
informational	2

```
portfolio.head()
```

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

```
portfolio['difficulty'].describe()
```

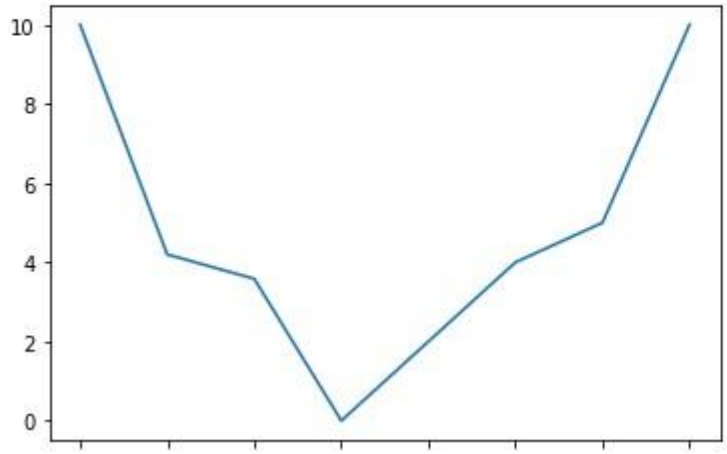
```
count    10.000000
mean      7.700000
std       5.831905
min       0.000000
25%       5.000000
50%       8.500000
75%      10.000000
max      20.000000
Name: difficulty, dtype: float64
```

```
portfolio.shape
```

```
(10, 6)
```

```
portfolio['reward'].describe().plot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7d63fc418128>
```



Profile data consists of demographics features of the customers fore.g what is their age, gender, income, id which will help to tailor specific event based offers for customers, schema of profile data frame is shown below in the bullet points. profile.json

1. age (int) - age of the customer
2. becamememberon (int) - date when customer created an app account
3. gender (str) - gender of the customer (note some entries contain 'O' forother

4. rather than M or F)
5. id (str) - customer id
6. income (float) - customer's income

Some data analysis of the profile data frame is shown below, in which age,gender,income and shape of data frame is checked:

```
]: profile.shape
]: (17000, 5)

]: profile['age'].describe()
]: count    17000.000000
   mean        62.531412
   std         26.738580
   min         18.000000
   25%         45.000000
   50%         58.000000
   75%         73.000000
   max        118.000000
   Name: age, dtype: float64

]: profile['gender'].value_counts().to_frame().style.bar()
]:
```

	gender
M	8484
F	6129
O	212

```
]: profile['income'].describe()
]: count     14825.000000
   mean      65404.991568
   std       21598.299410
   min       30000.000000
   25%       49000.000000
   50%       64000.000000
   75%       80000.000000
   max      120000.000000
   Name: income, dtype: float64
```

```
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

```
profile.shape
```

```
(17000, 5)
```

Third is the transcript.json data which will help us to analze how the customer reacted on that offer and wether he went for that offer or not. Important key id is customer id which will help connect all these three data together and aggregate the values at a customer level. transcript.json

1. event (str) - record description (ie transaction, offer received, offer viewed,etc.)
2. person (str) - customer id
3. time (int) - time in hours since start of test. The data begins at time t=0
4. value - (dict of strings) - either an offer id or transaction amount dependingon the record

Third is the transcript.json data which will help us to analze how the customer reacted on that offer and wether he went for that offer or not. Important key id is customer id which will help connect all these three data together and aggregate the values at a customer level

```
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'}

```
transcript['event'].value_counts()
```

```
transaction        138953
offer received      76277
offer viewed        57725
offer completed     33579
Name: event, dtype: int64
```

```
transcript['time'].describe()
```

```
count    306534.000000
mean        366.382940
std         200.326314
min           0.000000
25%         186.000000
50%         408.000000
75%         528.000000
```

5 Benchmark Model

A fairly accurate model which can help assess implementation of the hypothesis, several different models will be implemented according to different hypothesis. Using the decile ranking i ll check the overall distribution of customers according to the evaluation metrics used and will select the top 3 or 4 deciles, similarly i ll also check for the precision recall f1 score , accuracy for the models, the one which will not overfit and underfit will be selected as the benchmark model

6 Evaluation Metrics

I ll be using precision, recall, accuracy for the evaluation of my models. These metrics will help me to evaluate wether my model is overfitting or underfitting or how i need to re-assess or change the parameters or hyper parameters involved in my modelling approaches.

7 Project Design

These are the general steps i will be taking for my project flow:

1. Set up of environment in which jupyter lab and s3 bucket along with theparticular IAM policies will be set up
2. Implementing of pre-processing steps which will involve removing outliers,null or NA values, similarly changing and checking of the data types of the

variables , then finally aggregating or creating new features in the data sets

3. Implementing Exploratory data analysis and forming hypothesis using theEDA visuals
4. Implementing several different modelling techniques to find the best modelfor the project
5. Analysis of the benchmark model and writing explanation of the modelalong with what evaluation metrics are helping to justify the picking of benchmarked model
6. Writing all the results in form of Blog post on medium

References

[1] B Senthil Arasu, B Jonath Backia Seelan, and N Thamaraiselvan. "A machine learning-based approach to enhancing social media marketing". In: *Computers & Electrical Engineering* 86 (2020), p. 106723.

[2] Adam Hakim et al. "Machines learn neuromarketing: Improving preference prediction from self-reports using multiple EEG measures and machine learning". In: *International Journal of Research in Marketing* 38.3 (2021), pp. 770–791.

[3] Eric WT Ngai and Yuanyuan Wu. "Machine learning in marketing: A literature review, conceptual framework, and research agenda". In: *Journal of Business Research* 145 (2022), pp. 35–48.

[4] Hamid Shirdastian, Michel Laroche, and Marie-Odile Richard. "Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter". In: *International Journal of Information Management* 48 (2019), pp. 291–307.