Machine Learning Engineer Nanodegree

Capstone Proposal

Proposal

Domain Background

- Starbucks Corporation is an American multinational chain of coffeehouses and roastery reserves headquartered in Seattle, Washington. As the world's largest coffeehouse chain, Starbucks is seen to be the main representation of the United States' second wave of coffee culture. Moreover Starbucks comes in fortune 500 companies and is ranked 227th in the year 2020.Starbucks have an mobile application which has various function and is used by the people to order online coffee via the app for pickup, pay for the purchase via the app and collect reward points. This app also provide a membership **"My Starbucks Rewards™ membership"**, after paying through the app the user receives free Stars/Bonus points that can be used to redeem a free drink of the user's choice. This app also offers various promotions to the users which includes Discount in a discount, a user gains a reward equal to a fraction of the amount spent on drinks ,BOGO (Buy One Get One Free) ,in a BOGO offer, a user needs to spend a certain amount to get a reward equal to that threshold amount and Informational offer which basically includes any release of new product and there is no reward, but neither is there a requisite amount that the user is expected to spend. In this project the basic task is to use the data to identify which groups of people are most responsive to each type of offer, and how best to present each type of offer.

Problem Statement

We will be exploring the Starbuck's Dataset in which we will determine how people make purchasing decisions and how those decisions are influenced by promotional offers.The task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. Every offer has a validity period before the offer expires. As an example, a BOGO offer might be valid for only 5 days. You'll see in the data set that informational offers have a validity period even though these ads are merely providing information about a product.

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.

Datasets and Inputs

The data is contained in three files:

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

Here is the schema and explanation of each variable in the files:

portfolio.json

- id (string) - offer id
- offer_type (string) - type of offer ie BOGO, discount, informational
- difficulty (int) - minimum required spend to complete an offer
- reward (int) - reward given for completing an offer
- duration (int) - time for offer to be open, in days
- channels (list of strings)

profile.json

- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
- id (str) - customer id
- income (float) - customer's income

transcript.json

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

Solution Statement

To Find out of the 3 offer which will be suitable for the user using the previous orders and which offer interests the customer more. I will be using Exploratory Data Analysis (EDA) to cover points like What is the proportion of client who have completed the offers based on Gender , Age , Income Level. Which one is the most responded offer, how good is the response to an offer.To find out the best response for a particular user I will be using models like Decision Tree and Random Forest and find which model fits the best and impacted the promotional offer completion in customers.

Benchmark Model

As the data provided has both input and output this type of model comes in supervised learning, the model best suited for benchmark is KNeighbours as it is fast and accurate for this type of problem.

Evaluation Metrics

I will use F1 score as an evaluation metrics in this case to determine which model will suite and performs better.The **F1 Score** is the 2*((precision*recall) / (precision + recall)). It is also called the F Score or the F Measure ,the F1 score conveys the balance between the precision and the recall.

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}$$

TP= number of True Positives

FP= number of False positives

FN= number of false negatives

Project Design

Here’s the basic outline of the approach used in the project:-

1) Data Exploration and Pre-processing :-

    i) cleaning the data

    ii) processing the data and merging data from offer portfolio, customer profile, and transaction for analysing.

2)Perform *Exploratory Data Analysis* on the Data.

3) Building different machine learning model to determine which is the most suitable.

4) Using Evaluation Metric for determination of best model

5) Summarize the prediction