MACHINE LEARNING ENGINEER NANODEGREE.

Capstone Project

Sameerna Joshi
22-04-2021

Proposal

Domain background.

- Starbucks the coffee place, 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' third wave of coffee culture. Starbucks has a mobile application where the user can order coffee online or place an order online and go pick it up from the store.
- “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.

Problem Statement

With the Starbucks dataset I am trying to make an effort to understand the Starbucks customer’s buying behaviour.
What influences their purchasing decisions? Do they respond to promotional offers.
Which demographic groups respond to which offer types.

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

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

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.

I would like to explore K Nearest Neighbor, Random forest and XGBoost. I believe these models are fast and accurate to these type of classification problems.

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 exploring different demographic traits with their relevant purchasing patterns.
3) Building different machine learning models K Nearest Neighbor, Random forest and XGBoost.
4) Using Evaluation Metric for determining the best model
5) Summarize the prediction