Machine Learning Engineer Nanodegree

Udacity

Capstone Project Proposal

Mohamed Amr Farouk

December 30st, 2022

I. Domain Background

This project derives from the marketing team of Starbucks to keep in touch with the customer and to know what leads them to buy

Aiming to increase and reward the customers registered in its platform, Starbucks periodically sends individual messages containing offers related to its products.

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: e-mail, social media, on the web, or via the Starbucks's app.

So we will try to get a pattern from the data to find out how can the marketing team increase the ROI (return on investing) by knowing which feature affects the customer more

The company do not want to spend money on marketing in the wrong direction so we will build the machine learning model to make sure that the propriate offer is delivered to the customer through the propriate so we make sure that the customer will use the offer so that will benifet the marketing campaign by increasing the profit and the customer satisfaction by providing him the best fit offer

Considering the recent advances of artificial intelligence and the massive amount of data gathered over the years; this is a topic that could be widely improved by intelligent systems since they can analyze a large amount of data and understand patterns sometimes hidden for the human perception.

The last section of this document – VIII. Reference – holds a list of the papers I have found helpful to formulate this project proposal

REFERANCE: G. Theocharous, P. S. Thomas, and M. Ghavamzadeh, "Personalized Ad Recommendation Systems for Life-Time Value Optimization with Guarantees," in IJCAI, 2015.

II. Problem Statement

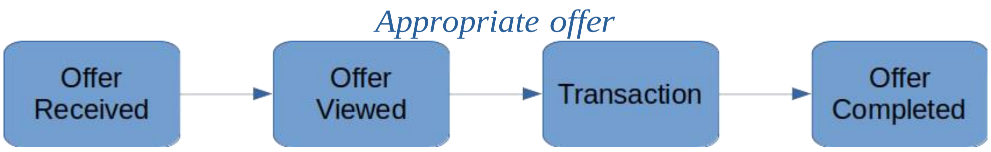
Starbucks, as well as any other company, invests money in marketing campaigns expecting to have a profit higher than the assumed before. So, identifying the most relevant offer to the correct customers is crucial for a successful campaign.

However, some targeted customers do not even see the offer sent to them, which may be a problem with the channel chosen. Other ones do not buy anything, despite seeing the offer, what might be a problem with the offer type sent, or maybe that is not a customer to be considered as a target.

There are other cases where the customers identify themselves with the offer, which leads them to try new products or spend more money than usual. Those are the situations to be identified and pursued.

The problem this project proposes to solve is finding the most appropriate offer for each one of the customers, which means finding the offer that is more likely to lead the customer to buy Starbucks products.

In the context of this project, an appropriate offer is that one where the customer sees the offer received and buys products under its influence, completing the offer lifecycle.



If a customer does not see an offer, it is not an appropriate one. If he or she sees the offer but does not complete it, it is not appropriate as well, since it did not lead the consumer to buy products. Similarly, if the customer buys some products, completes an offer, and receives a reward before visualizing that offer, it is not considered effective because the customer was not under the influence of that offer when decided to make a purchase.

III. Datasets and Inputs

a. Dataset overview

The data set used in this project is provided by Udacity and Starbucks as part of the Machine Learning Engineer Nanodegree program. It contains simulated data that mimics customer behavior on the Starbucks rewards mobile app.

The program used to create the data simulates how people make purchasing decisions and how those decisions are influenced by promotional offers.

Each person in the simulation has some hidden traits that influence their purchasing patterns and are associated with their observable traits. People produce various events, including receiving offers, opening offers, and making purchases.

As a simplification, there are no explicit products to track. Only the amounts of each transaction or offer are recorded.

b. Data Dictionary

The data is contained in three files:

- **profile.json**
  Demographic data for each rewards program
  users Size: 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)

- **portfolio.json**
  Offer ids and meta data about each offer sent during 30-day test period
  Size: 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)

- **transcript.json**

  Event log containing records for transactions, offers received, offers viewed, and offers completed

  Size: 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

IV. Solution Statement

In order to face the problem stated above, this project proposes to apply machine learning techniques to find patterns in customers' behavior by analyzing the transcriptions of their relationship with Starbucks.

More specifically machine learning model will be trained to predict how customers may react when receiving each one of the available offers: if they will complete the offer cycle or not. So, it will be possible to identify which one is more suitable for each customer.

The final result of this project is a direct marketing system that, given a customer, is able to predict the likelihood of each offer be completed.

The steps of the solution :

1. **Get data of each json file**
2. **Merge between the data to get related data in one data frame**
3. **Explore the data**
4. **Clean the data and handle the null values**
5. **Normalize and scale the data to be easier for training the models**
6. **Try different machine learning model to get the model of higher value**
7. **Deploy the model using AWS sagemaker end point**
8. **We can call the model using AWS lambda**

V. Benchmark Model

A more traditional model will be trained in the same dataset used by the intended SGD calssifier,

Basically, an SGD analyzes a static input and makes predictions

A naive bais model might understand an offer as adequate because it produced a good result in the past, and tends to repeat that action every time. However, it may not be able to detect whether the same offer becomes inadequate when sent a second time to the same customer, perhaps owing to the fact that the customer does not want to repeat the same purchase forever.

In another case, the customer is conditioned to buy products, so no offer sending is necessary anymore. However, that SGD Classifier keeps suggesting the same offer over again.

Building and training both models allows us to compare the predictions made considering only the static user state (SGD) and those made based on the customer history (Naïve bias). Then, we will be able to evaluate whether the problem stated is better addressed

I. Evaluation Metrics

The accuracy of each model will be measured to evaluate the performance of the the model. By using the same metric

$$accuracy = \sum_{i=1}^{n} \frac{p_i}{n} \approx 80\%$$

p_i^c = correctly classified instance

n = total number of offers sent

I. Project Design

The theoretical workflow for approaching the solution stated includes several machine learning techniques, following the guideline sections below.

a. Data loading and exploration

Load files and present some data visualization in order to understand the distribution and characteristics of the data, and possibly identify inconsistencies.

b. Data cleaning and pre-processing

Having analyzed the data, handle data to fix possible issues found.

Handling the null values

c. Feature engineering and data transformation

Prepare the data to correspond to the problem stated and feed the neural networks. The transcription records must be structured and labeled as appropriate offer or not.

d. Splitting the data into training, validation, and testing sets

Training: which will be used to train the model

Validation: which will be used to test the model while the training process to make sure its getting better and does not overfit

Testing: which will be used to test the model after the training process

e. Defining and training different model as :

SGD classifier, Naïve bias, etc with different hyperparameters

f. Evaluating and comparing model performances

Comparison between the accuracy of these models so we can choose the best of them