Starbucks Capstone Project

Domain Background

Many aspects from what it was learned will be used. Such as data cleaning, data discovery (with EDA), hyperparamenters tuning, model deployment, creation of endpoint and lambda functions, besides others tools and processes.

Customers' retention is a common subject in many fields of study, since usually is expensive to get a new client than retaining one. Several studies were done in this field to avoid churn.

This one is a really interesting concept that can be applied in several companies in its way, that is one of the many things that motivates me in order to develop this project in applied the knowledge learn so far.

Project Overview

The data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be:

- Merely an ***advertisement*** for a drink or an actual offer such as a ***discount***;
- ***BOGO*** (buy one get one free);
- 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.

Problem

The task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. This data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products. Even gathering outside data if it is pertinent in order to achieve better performance for the model.

Solution and Project Design

First, even before trying to get insights from the data, the solution will begin with a ***cleaning of data*** in order to be easier and faster to work with it. After the structuring of the data, an ***EDA*** will be performed, in interest of getting features understanding, besides featuring engineering.

With the data worked and insights in hand, a pipeline will be developed in pursuance to achieve the best model with the best performance, inside this pipeline will have:

1. Pre-processing of the data;
2. Tuning the model in order to find the best hyperparamenters;
3. Training the model;
4. Saving the model in order to develop an endpoint;
5. Developing a lambda function to invoke the API calling the endpoint.

Few models will be tested in order to keep the best one.

Datasets and Inputs

This data set was provided through a project agreement with Starbucks in the Udacity plataform.

portfolio.json (10 x 6)

- 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 (17000 x 6)

- 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 (306534 x 7)

- 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

Below, on the images, it is possible to understand a bit of the disposition of the data. Understanding a bit of the distribution regarding the profiles and offers.

Figure 1. Kde of age and incode + Bar plot of age and year of membership.

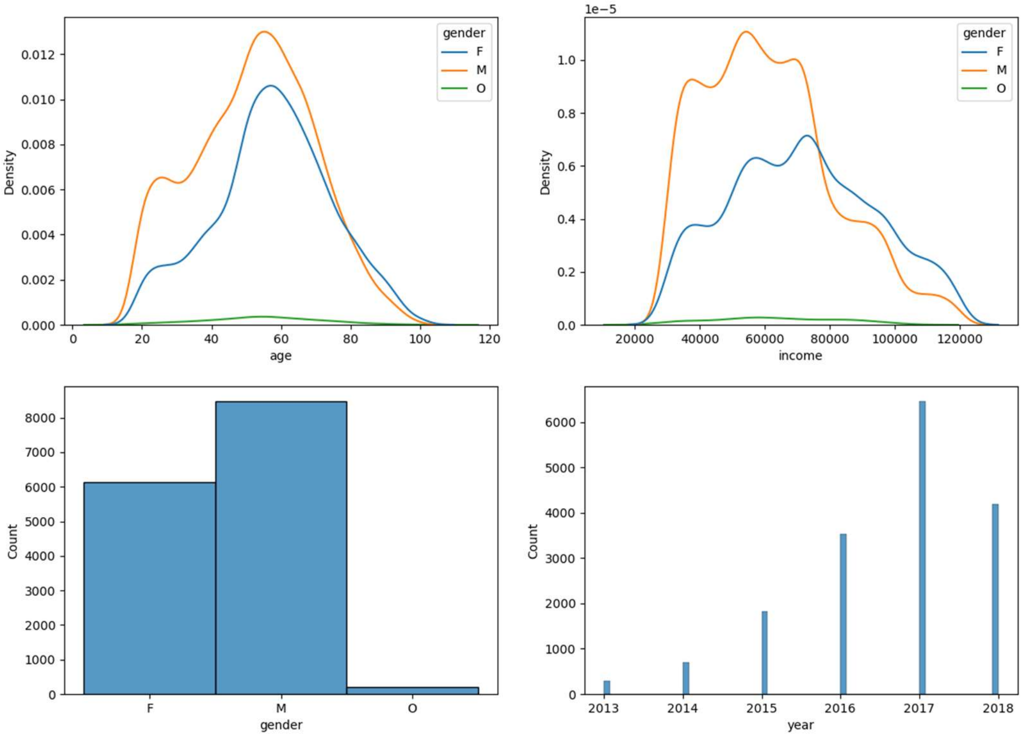
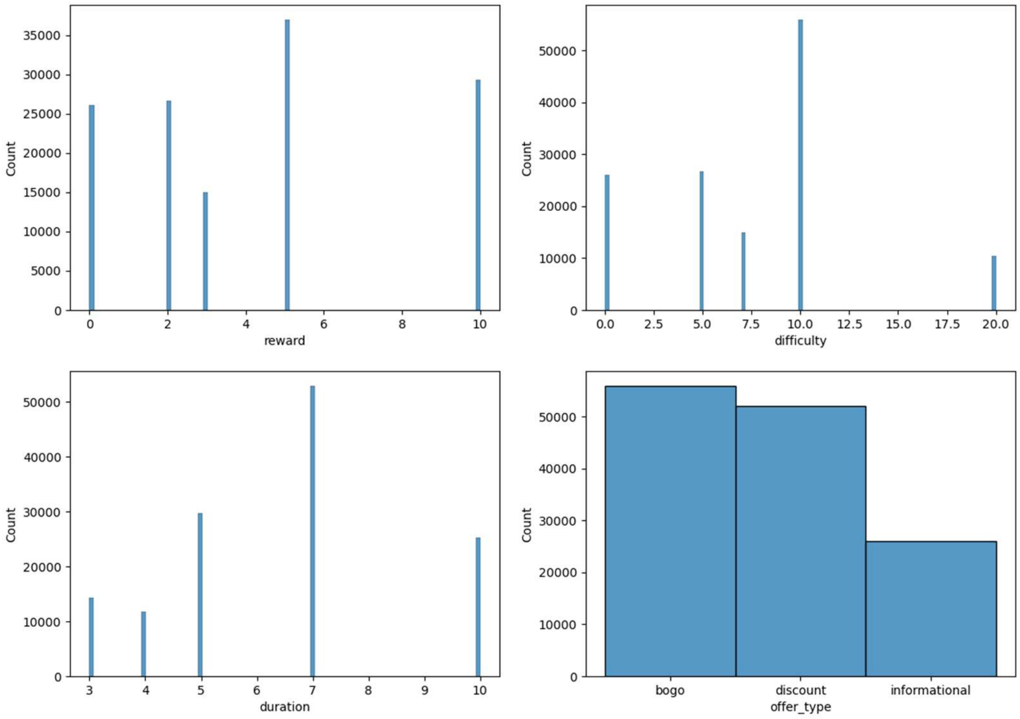


Figure 2. Bar plot of reward + difficulty + duration + type of the offers.



Benchmark Model

The benchmark that will be looked in order to make the comparison are simple business rules, like for example.

- Giving discount for everyone on her/his second beverage after a certain period of time (like a week), comparing the adherence of that rule and the model;
- Offers after a given value is spent, like comparing the adherence to any offer after the client spent 10 dolars, for example;
- Offers to everyone who is client longer than a certain period, for example, a year, six months and so on.

In general simple and naïve filters to see how the adherence is compare to it, in order to see if it makes sense for the stakeholder to use a model instead of the intuition.

Evaluation Metrics

The main evaluation metrics, that will be looked at, will be the ones listed below. It is important to keep track of accuracy, but also looking at those additional metrics in order of not have a bias vision of the process.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall \ = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$F1\ Score = \frac{2 * (Precision + Recall)}{Precision + Recall}$$

TP	True positive
FP	False positive
TN	True negative
FN	False negative