Starbucks Capstone Proposal

Overview

This project is a real-life marketing strategy study based on simulated data set that mimics customer behavior on the Starbucks rewards mobile app. The goal is to discover what are the better offers at an individual level, not for the population as a whole.

At the age of Customer, data analytics provides insights into customers' experience and behavior. I am really interested in this domain as there is a wide range of use cases for this type of study. It doesn't matter whether you buy a cup of coffee or you buy a new car, the data analysis will help companies to create effective marketing strategies, influence targeted customers and drive sales.

Problem Statement

I am a cost conscious consumer. I like to take advantage of promotions and discounts such as udacity's 10% off tuition promotion. It would be interested in seeing how many people are like me in this study.

In this project, Starbucks periodically sends three different types of promotional offers to customers: buy-one-get-one (BOGO), discount, and informational. We would like to combine transaction, demographic and offer data to answer the following two questions.

1. which groups of people are most responsive to each type of offer, and how best to present each type of offer.
2. How much someone will spend based on demographics and offer types?

By answering these two questions, Starbucks can better target their offers towards customers with higher probability to use the offers. At the same time, customers receive personalized and relevant offers on the mobile app and get better user experience. By doing these, Starbucks could potentially maximize revenue and save on marketing and promotional costs.

Datasets and Inputs

The dataset used in this project was created to simulate how people make purchasing decisions and how those decisions are influenced by promotional offers. There are three data dictionaries.

profile.json

- Rewards program users (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: Offers sent during 30-day test period (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 (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

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.

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.

Solution and Design

To solve the problems, my strategies are listed as follows:

1. Clean up the data, such as check missing data, convert column values from string to int, feature scaling, etc.
2. Merge the portfolio, profile and transaction data dictionaries
3. Separate the merged data into three datasets based on three offer types.
4. Build 3 classification models for the 3 datasets to answer the first question in the problem statement section. The model can predict whether the customer will respond to a specific offer type.
5. For each model, I plan to use different algorithms such as Random Forest, XGBoost and will compare them with benchmark simple logistic regression model
6. Assess models using accuracy and F1-score.

7. Build one regression model to predict the amount a customer would spend given that the customer is responding to the offer. It could provide an answer to the second question in the problem statement section. I plan to use different algorithms such as Random Forest, XGBoost comparing with benchmark simple linear regression model
8. Assess the regression model using R-squared.
9. If possible, tune hyperparameters of the best models with highest accuracy and F1-score.
10. Perform feature importance analysis, model may be improved if only using top N features.
11. Performed some exploratory data analysis to extract further insights.

Accuracy is usually used to assess the performance of classification models. In this project, False negatives are important as we don't want to miss the responsive customers. I will use F1 score beside accuracy . If the class distribution is imbalanced, F1-score is a better metric to evaluate our model on.

A benchmark model is desirable so we can compare our model to it. I am not able to find any publications with obvious methodology against which I can benchmark, I plan to compare my classification model to a very simple logistic regression model which can be used as a baseline model. I will also compare my regression model to simple linear regression model as a baseline model.