Starbuck Project Proposal

August 21, 2021

1 Domain Background

The purpose of this project is to analyze from simulated Starbucks' customers data provided by Starbucks and Udacity, in order to gain insight on the relationship of the customers' attributes and their response to promotional offers being given to them.

Once in a while, Starbucks sends promotional offer to its mobile customers and the data gained from it are being used to simulate the dataset this project is based on.

From a business perspective, it is important to understand whether an offer is effective and how to personalize offers based on customers' attributes. This personalization could improve the efficacy of the promotional offer itself and might even increase the revenue, if more people are being attracted to buy based on that personalized offer.

Some research has been conducted using machine learning model to classify things based on marketing data. It is a good practice to learn from them before solving problems in the marketing area and using the marketing data. The followings are some of them:

- https://www.researchgate.net/publication/282657577_Marketing_Research_Data_Classification_by_Means_of_Machine_Learning_Methods
- https://www.researchgate.net/publication/260707025_Using_Neural_Networks_for_Marketing_Research_Data_Classification

Also, this project is a great fit for students of Data Science or Machine Learning to tinker on, since it would widen their experience on a different kind of dataset and also for them to engineer features that matter and algorithm that would perform best.

2 Problem Statement

Would a customer respond to a particular offer?

- The problem of this project would be a classification problem: there needs to be a classification of whether a promotional offer is going to make a customer reponds or not.
- An approach to this problem would be to see if there could be a pattern emerged from customer's attributes and the promotional offer's data (duration, rewards, etc.) to determine whether a customer would respond to a promotional offer: customer's attributes and promotional offer's data to be the inputs and a binary classification of responding or not would be the output.

3 Datasets

There are three types of dataset provided in this project:

1. `portfolio.json`: containing offer ids and meta data about each offer (duration, type, etc.) It has 10 unique offers or data points.

- 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)

2. `profile.json`: demographic data for each customer. It has 17,000 of unique customers or data points.

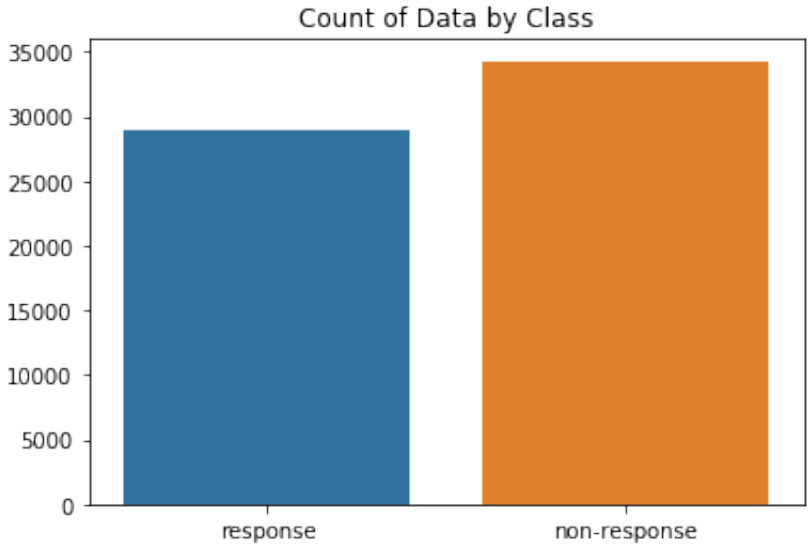
- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
- id (str) - customer id
- income (float) - customer's income

3. `transcript.json`: records for transactions, offers received, offers viewed, and offers completed. It has 306,534 recorded events or data points.

- 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

One way to approach the problem is to join all the tables and group them based on customer ID and offer ID and label each of those grouped data point with a value of `0` if the customer does not respond and `1` otherwise.

A preliminary exploration on the joined dataset grouped by customers ID and offer ID shows the following:



- The figure above shows that there is no significant imbalance in the total data points for each class: 28,996 for the number of customers who respond and 34,292 for the ones who do not.

4 Solution Statement

- In order to determine whether a customer would respond to a particular offer, a model needs to be built to predict response based on that particular customer's attributes.

5 Benchmark Model

In this project, a Logistic Regression algorithm would be used to be the benchmark algorithm in predicting whether a customer would respond to a particular offer. This is due to that Logistic Regression is a basic algorithm in solving a binary classification problem, such as that of the problem statement above.

6 Evaluation Metrics

The metric to be used for the evaluation of this project would be the accuracy level, since it is more important to maximize the true positives and true negatives, rather than to minimize the false positives or false negatives. This choice is also in line with the number of data points in each class, which is balanced.

7 Project Design

1. Data Exploration and Analysis

In this step, the datasets need to be explored and analyzed not only for some missing or inappropriate values, but also the features' datatypes and their distribution visuals.

2. Data Preprocessing

In this step, the datasets would be cleaned or imputed if necessary to compensate the missing or inappropriate values.

3. Features Engineering

Choosing which features to include in the model training or even developing ones, if necessary, depending on the insight from the data exploration step, would be done in this step.

4. Train the Benchmark Model

In order to set a baseline model to compare the performance of the other ones, the benchmark model of Logistic Regression would be trained with the training dataset produced from the previous step.

5. Test Other Model Algorithm

To find the best model which could predict user response from a promotional offer, the followings supervised machine learning algorithms for classification problem would be trained, with possibly grid-search or cross-validation. * Naive Bayes * Decision Tree * SVM * Random Forest

6. Evaluate the Model

After training all the models as outlined above, evaluation of the accuracy of each model would then be conducted.

7. Final Analysis

This step would analyze the overall results based on the performance of the models trained previously and would lay out some analysis on why the accuracy scores obtained are so and how to improve the models in the future.