

Capstone Project

Machine Learning Engineering Nanodegree

Starbucks Offers Classification Model

Runsang Yu
June, 2022

Domain Background

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 or BOGO (buy one get one free). Some users might not receive any offer during certain weeks.

A data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app including transaction, demographic and offer is given 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.

Some basic data analysis are done on simulated data including which offer are most popular and who are mostly likely to complete an offer. A model is built to predict whether or not someone will respond to an offer to help better customer targeting in the future.

Problem Statement

There might be a pattern behind different demographic people responding an offer, if we build a model to predict a customer's likelihood to respond to an offer, this will improve customer targeting strategy.

Dataset and Inputs

The data is contained in three files:

- portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
- profile.json - demographic data for each customer
- transcript.json - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

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

Various machine learning models including logistic regression, K Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier are used to pick out the best performance one to predict whether a customer is responding on an offer or not and feature importance are generated to illustrate which variable has the biggest impact on the model.

Benchmark Model

There are no existing benchmark model for this project, but several machine learning algorithms will be tested to pick out the best performance one.

Evaluation Metrics

A classification model is built to predict whether a customer is responding an offer or not. The dataset will be splitted into training and testing dataset and confusion matrix, accuracy score, recall, precision, f1 score, AUC will be generated on both datasets to ensure there's no overfitting problem happening.

Project Design

There will be lots of efforts to clean the original dataset into training dataset. The final training dataset will be in a offer_id & person_id level, with all the features related including offer reward, offer difficulty, offer duration, offer sent by email/mobile/social/web, customer age, customer

income, customer gender, customer staying length since first registered with Starbucks, offer type (BOGO, discount, informational), a label is generated on offer successful or not. Various machine learning models including logistic regression, K Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier are used to pick out the best performance one to predict whether a customer is responding to an offer or not and feature importance is generated to illustrate which variable has the biggest impact on the model.

Exploratory Data Analysis

Which offer are most popular?

Through analysis, Buy one get one offers are most likely to be received, and completed by customers, whilst discount offers are most viewed by customers.

Who are most likely to complete an offer?

It shows that men will complete more offers compared with women, but it seems that women are more likely to complete 5-6 offers compared to men. High income people complete more offers than other income level people. The older people are, the more likely they finish an offer. People who are quite new to Starbucks are less likely to finish an offer.

Model Result

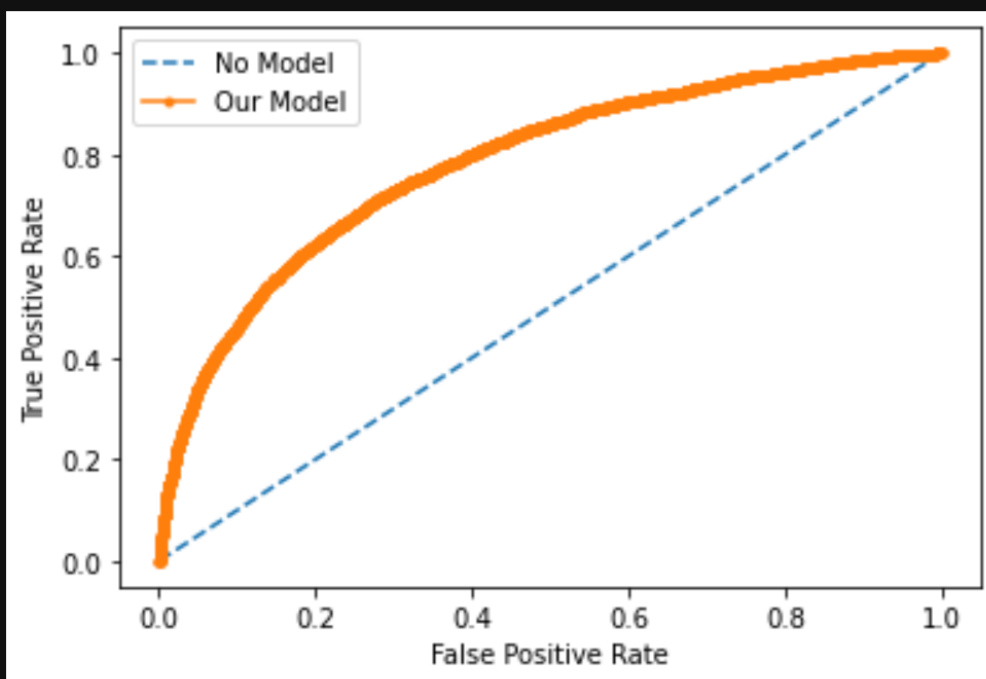
A classification model is built to predict if a customer is going to respond to an offer or not. The features used in the model include offer reward, offer difficulty, offer duration, offer sent by email/mobile/social/web, customer age, customer income, customer gender, customer staying length since first registered with Starbucks, offer type (BOGO, discount, informational), a label is generated on offer successful or not. Various machine learning models including logistic regression, K Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier are used to pick out the best performance one and the gradient boosting model shows to be the best performance model. After hyperparameter tuning, it still shows the default gradient boosting model outperforms all the other models. Accuracy on the training dataset is 72%, and accuracy on the testing dataset is 72% too, so there's no overfitting problem. Details about the result are shown in picture1. Shap value is calculated to infer the feature importance on the model which is shown in picture2. A high level of the "member staying length" has the highest and positive impact on the likelihood that customers' respond to an offer, followed by offer sent by social, high income etc all are more likely to make customers respond to the offers.

Performance on train dataset: 0.7243045112781955
Performance on test dataset: 0.7174648522667468

Confusion Matrix on test dataset:

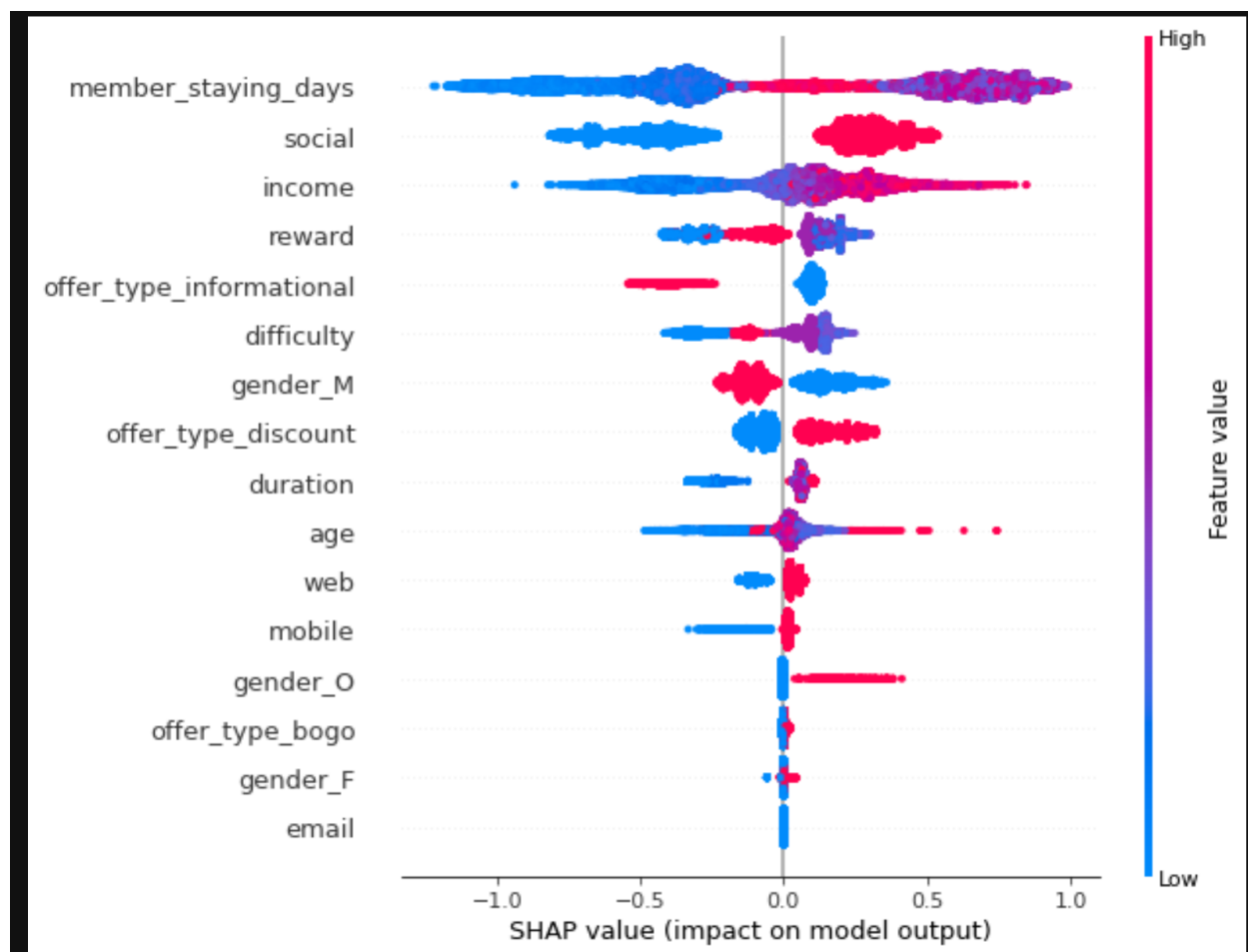
```
[[5499 1650]  
 [2108 4044]]
```

ROC curve on test dataset:



AUC score on test dataset: 0.7852761855623409
Precision on test dataset: 0.7102212855637513
Recall on test dataset: 0.6573472041612484
F1 score on test dataset: 0.6827621137936857
10-fold CV score on whole dataset: 0.7155831278692899

Picture1. Gradient Boosting Model performance



Picture2. Shap Values