

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 and also some initial findings:

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)

There are only 9 offers in total, and offer type consists of BOGO, informational, and discounts. The channels columns needs to be transformed to one-hot-coding to feed into the model later, and there are no missing values in this table.

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

There are no duplicates for customer ID in profile data. 12.79% data are missing in terms of gender and income. Became_member_on should be a date field and a column can be created for customer staying length, the median of customer staying length is 1788 days, max staying length is 3253 days, minimum is 1430 days. In addition, there are 2,175 customers who are aged 118, these records miss income and gender too so are removed from the dataset. In terms of gender, 6,129 are females, 8,484 are males, and 212 are O. Income median is 64k, with a max of 120k and minimum of 30k. The buckets can be created for age, income and staying length for further exploratory analysis.

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

There are various columns consist of dictionaries, key is offer_id and amount, dictionary can be converted into columns. There are only three types of events which are offer received, offer reviewed and offer completed. Time column represents for hours which can be transformed to days so that is align with profile data.

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

Logistic regression is picked as the benchmark model. Several other machine learning algorithms will be tested to see if it's outperformed than the benchmark model.

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 on an offer or not and feature importance are generated to illustrate which variable has the biggest impact on the model.