Machine Learning Engineer Nanodegree

Udacity

Capstone Proposal

Aya Tarek Ali
May 17th, 2020

Table of contents:

1. Domain Background

This project is based on simulated data of customer behavior and offers reactions driven from Starbucks rewards mobile APP.
Starbucks periodically sends offers and advertisement information about the new products to the app users. These offers can be one of three types: Buy One Get One (BOGO), a special discount, or an informational message and some users may not receive any offers for certain period. Each of these offers has a validity period, even the informational ones.

Completed offers don’t mean necessarily that they were actually completed, for example: a customers may receive an offer “Spend 10 dollars and receive free 10 dollars to spend”, but never really views this offer, then the customer spends 12 dollars so there will be an offer completion record, however the customer wasn’t influenced by that offer because he didn’t view the offer.

Understanding the customer behavior will help us in taking decisions of which customers to target with which offers, and which customers will more likely to complete a specific offer.

2. Problem Statement

Merchants like Starbucks spend money on marketing campaigns, however not all customers respond to the offers the same way. Some customers don’t view the offers because they are not interested, others may be not interested in that particular offer they received. The problem is to identify whether the customer will respond and complete these offers or not.

My approach for solving this issue, will be to predict the customer response based on his demographic group features (Age, Gender, Income, App Join year and his response to previous offers.

3. Datasets and Inputs

The data that we will use in this project is contained in three files:

Portfolio.json

Contains the offer ids and Meta data about each offer, there are 10 different offers with 6 features:

- id (string) - offer id
- offer_type (string) - type of offer (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 (validity)
- channels (list of strings) - web, email, mobile, social,…

profile.json

Contains the demographic data for 17,000 customers, each with 5 features:

- 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, Notice that the missing record has age=118, gender=none and income =NaN.

transcript.json

Contains data about the customer transactions and offer status for 306534 events with 4 features:

- 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 (if the event is an offer) or transaction amount (if the event is a transaction)

4. Solution Statement

In this project we will be using machine learning algorithms to study, explore and analyze the customers' behavior data from Starbucks reward app.

What we need is to predict the customer response to the offers being sent, based on the offer type and information. This is a supervised classification problem that we can use more than one model and choose the one with the highest accuracy.

For this problem I will use Logistic Regression, Decision tree and Random forest models and compare the training and testing accuracy to choose the best model.

5. Benchmark Model

As mentioned in the Solution Statement, We will train two additional models Decision tree and Random forest against our baseline model the Logistic Regression using same dataset and compare the accuracy of the three models.

6. Evaluation Metrics

We will evaluate the models by measuring the accuracy of each model using accuracy and confusion matrix by calculating the true positive, true negative, false positive and false negative.
We will choose the model with the highest accuracy, TP and the lowest FN count, as we need to make sure not to get that the customer won't respond to an offer when he actually responded to.

7. Project Design

The project theoretical workflow includes multiple steps and machine learning methodologies that starts with the data exploration till the predicted offer value:

a. **Data loading and understanding**
   We will load the data files and explore what data we have by visualizing and printing the data features.
b. **Data cleaning and preparation**
   We need to remove the null values, remove the outliers and apply one hot encoding if needed.
c. **Data analysis**
   After cleaning and preparing the data, we need to combine the three files into one dataset and start visualizing the each feature to understand the relation between the customer profile, transactions and the offers he receive.
d. **Splitting the data**
   First we split the data into training and testing datasets, then apply the cross validation over the training to find the best folds with the higher accuracy to be used in training and evaluating the models.
e. **Define and train the models**
f. **Save the trained models**
g. **Load and test the models**
h. **Evaluate and compare the accuracy & confusion matrix**