Capstone Project

Machine Learning Nanodegree Program

Dipan Sutradhar

Dated: May 20, 2020

Project Overview

Starbucks Capstone Challenge is a capstone project of the Udacity Machine Learning Nanodegree program. This project addresses a classification challenge that predicts the customers' demographics influence on responding to a Starbucks mobile reward program. The data set includes Starbucks mobile app reward data with simulated customer demographics.

Problem Statement

Like any other campaigning program, the success of the Starbucks reward program depends on customers' participation in the program. Starbucks wants to analyze and identify if there is any influence on customers' demographics. Eventually, develop a model to predict the customers' responsiveness based on their demographic and offer details. This will help Starbucks to back their advertisement decision by data.

Data Exploration

Starbucks datasets separated into three such as portfolio, profile, and transcript. The exploration of these data set is as below:

- Portfolio Dataset
  - 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 Dataset
  - 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 Dataset
  - 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

The solution of this classification problem requires determining and preparing the features that the classification model should be trained on. Process starts with visualizing the data to find out the data distribution. Then cleaning up any NaN or missing data. As the goal to develop classification predictive models, any categorical data needs to be reformatted.

For this challenge, bellow features will be used to train the model:

- Customer Details: Age, Income, Gender, and membership days
  - Gender - male, female, or other
  - Offer Details: Difficulty, Duration, Channel, Offer Type

Finally, dataset is randomized and split into test and training dataset.

Benchmark Model

At training stage, classifier such as DecisionTreeClassifier and SVC from scikit-learn package are leverage for training and benchmarking. Models are trained using this testing dataset along the ground truth as label.

After the training completed, models predict the test data set without label. Accuracy of the prediction is measured with the help of scikit-learn *accuracy* util.

Conclusion

At last, this project allows Starbucks taking data driven decision with the help of trained models. This process starts from exploring data, cleaning, preparing training and test data, train models, and analyzing result. Experiment recommends any possible improvement on data and model.