Starbucks Capstone Project Proposal

By

Shankhadeep Banerjee

Machine Learning Nanodegree Program

Domain Background

Starbucks Corporation is an American global coffee company and coffeehouse chain based in Seattle, Washington. Starbucks is the largest coffeehouse company in the world, with 20,891 stores in 62 countries.

Starbucks sends out personalized offers in such a way that the targeted user ends up paying attention and not ignore them. This is where Machine Learning comes into play.

The project will aim towards maximizing the profit for Starbucks by using Machine Learning to predict whether a particular user will complete the offer or not based on the history of data that is provided in the form of dataset.

The topic of Consumer behaviour research is what has motivated me to select this capstone project as a part of Udacity Machine Learning Nanodegree Program.

Problem Statement

The aim of this project is to create a Machine Learning model that will predict whether or not an user will complete an offer that is sent to him. That is, how likely will the customer accept the offer that is sent to them. This will improve the conversion rate of the offers and in-turn increase the overall profit for Starbucks.

We will use the dataset provided in the problem for this task. This falls under supervised binary classification problem.

Datasets and Inputs

The dataset is in the form of three files:

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

This is the schema of the three files:

portfolio.json (record of all offers available to the customers)

1. id (string) - offer id
2. offer_type (string) - type of offer (BOGO, discount, informational)
3. difficulty (int) - minimum required spend to complete an offer
4. reward (int) - reward given for completing an offer
5. duration (int) - time for offer to be open, in days
6. channels (list of strings)

profile.json (contains information about participant customers in the reward program)

1. age (int) - age of the customer
2. became_member_on (int) - date when customer created an app account
3. gender (str) - gender of the customer (‘M’/’F’/’O’)
4. id (str) - customer id
5. income (float) - customer's income

transcript.json (records of all the events during the test period)

1. event (str) - record description (transaction, offer received, offer viewed, etc.)
2. person (str) - customer id
3. time (int) - time in hours since start of test. The data begins at time t=0
4. value - (dict of strings) - either an offer id or transaction amount depending on the record

Solution Statement

The solution will be divided into 4 sub-parts:

1. **Data Pre-processing –** This will involve removing or processing any missing values or outliers present in the data, also after processing we will be able to analyse the data better.
2. **Data Exploration –** This step will involve a deep study of data which will help us to determine the interdependencies and distribution of data and also understand the bias terms present in the data.
3. **Building the model –** We will build the machine learning model on basis of the above understanding. Since it falls under a supervised binary classification problem, we can use any of the classification algorithms like: Logistic regression, SVM, Decision tree etc. Ensemble methods like XGBoost can also be used to maximize the efficiency of the model.
4. **Evaluation –** This step will involve evaluating the performance of the model upon two different metrics: accuracy and F1 score.

Benchmark Model

An evaluation function will be developed to compare predictions over the customers in the dataset with the real result to provide statistics about accuracy and F1 score of the different prediction models.

Evaluation Metrics

20% of the dataset will be used as test data to test the accuracy of the model trained over the rest 80% of the data. In addition to that, we will use F1 score as evaluation metrics to determine the performance of the models using Precision – Recall.

Project Design & Presentation

First, we will pre-process the three datasets to remove any missing values or outliers present in the data.

Then we will start exploratory analysis of the dataset to search for any co-relation and particularities among the data and find which of the data are useful to us, and will exclude rest of the data which doesn’t affect the outcome.

Then we will prepare the training and testing datasets and design a model to fit the dataset and predict the outcome.

Based on the outcome we will calculate the accuracy from the test data and F1 score, and tune the hyperparameters if required.

Finally, we will try implementing different models along with ensemble methods to see which model provides maximum accuracy on the dataset.