Starbucks Capstone Project Proposal Udacity Machine Learning Engineer Nanodegree

Cole Lineberry

June 28, 2022

Table of Contents

Domain Background

Starbucks is a world wide coffee maker and one of the Fortune 500 companies. They have thousands of customers who buy their coffee daily. These customers have the option to use a mobile application to purchase coffee. The company has gathered lots of customer data and demographics from these mobile apps. The marketing team at Starbucks has been offering some buy one get one free promotions (BOGO). These are hard to predict if they will be successful. Not all users receive an offer every week and not all users receive the same offer.

Problem Statement

Starbucks would like to know which offers to send to which demographic groups based on previous responses to offers. They would also like to be able to predict which offers will most likely be completed/converted by these groups. Starbucks has compiled 30 days worth of data.

If we assume past performance is a measure of future conversion rates we can apply machine learning algorithms to predict these future outcomes.

- Machine Learning task: The problem is defined as a classification task, which demographics are most likely to convert.
- Input: Previous customer responses and associated demographic data.
- Output: Predict which individuals are most likely to convert in the marketing campaign.

Datasets and inputs

The Dataset contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be an ad for a drink or an actual offer for a discount or BOGO. Some users might not receive any offer during certain weeks.

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

I will be building a machine learning model to predict the customer response to an offer. This will be based on the customers previous responses, taking into account customer demographics. This will include doing some data analysis at the beginning of the project to best set up the data for demographics vs customer response. Then I will build a machine learning model based on training data and compare that to a test set of data.

For this project I am dividing the work into three tasks. The first task will be to analyze the data to discover any correlations for demographics to conversion rates. The second task will be to build a benchmark model using KNearestNeighbors classifier so that we can establish a baseline for the last task. The last task will be to build a predictive model using RandomForest and/or DecisionTree classification algorithms.

Benchmark Model

I will use a KNearestNeighbors classifier to create a benchmark for the machine learning model. KNearestNeighbor is a good algorithm for this as it will create a nice binary classification that will be easy to test against. This will provide a way to tell if my predictor is performing at or better than the benchmark algorithm.

Evaluation Metrics

The KNearestNeighbor algorithm will provide a good way to classify the data as either a positive response, no response, or incomplete response. This will provide an F1 score which will be a number between 1 and 0 with anything closer to 1 being the most likely response. F1 provides the balance between precision and recall and is calculated with 2((precision recall)/(precision +recall)

Project Design

This project will be designed and built within a jupyter notebook. The project will be laid out in the following steps,

1. Data preparation: Obtain the data sets and remove any values that are not necessary for later use. Combine the three files into usable dataframes based on id and customer.
2. Data Exploration: Create visualizations on the data to get a better picture of the overall data set. Perform any tuning or refining to better figure out what the best predictive model for this type of data will be. Observe the relationship between gender and marketing campaign conversion.

3. Data transformation: Make the necessary data wrangling operations to be able to feed the training and test data into a machine learning algorithm.
4. Train model: Use the values from the training set to train the model
5. Predict outcomes: Use the model to predict outcomes based on test data.
6. Refinement: Tweek hyperparameters to optimize algorithms. For the randomForest classifier we may need to change to oversampling or undersampling if the data turns out to be imbalanced. I may even switch to a Balanced Random Forest algorithm to solve this problem.
7. Conclusion: Compare the benchmark to the predicted values and then summarize the findings.