Machine Learning Engineer Nanodegree

Sharmin Ahmed

October 12, 2020

Starbucks Capstone Project Proposal

Domain Background

As of early 2020, Starbucks company operates over 30,000 locations worldwide in more than 70 countries, being the largest coffee house chain. This only became possible because of Starbucks' customer-centric culture. In 2011, with the launching of the Starbucks app, the company has granted its customers the ability to order, pay, and collect their beverages without the torture of queuing. Once every few days, Starbucks would send 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 offers during certain weeks.

By effectively introducing these a referral program, and a reward system, Starbucks is attracting millions of customers. For company profit and building a better user experience, Starbucks has to customize the offers for the users. Because, if everyone gets the same offer they might not use them. They offers can be like giving more discounts to heavy users, free offers, limited discounts, etc.

As I, myself is a coffee lover this problem piqued my interest. Addressing this problem will give me insights into marketing schemes and the pattern of users' behavior for an offer.

Problem Statement

The problem is to determine what offers to send to which customers. It can be based on their history such as previous purchases, genders, age groups, locations, etc. The goal of the project is to ensure each customer with relevant offers that will lead the customer to buy Starbucks products.

Datasets and Inputs

The dataset for this project is provided by Udacity and Starbucks. This data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app which is contained in three files:

1) portfolio.json: Containing offer ids and metadata about each offer (duration, type, etc.). Data size is: 17000 users x 5 fields

- **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)** - web, email, mobile, social

2) profile.json: Demographic data for each customer. Data size is: 10 offers x 6 fields

- **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

3) transcript.json: Records for transactions, offers received, offers viewed, and offers complete. Data size is: 306648 events x 4 fields

- **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

To address this problem, my strategy is to apply machine learning models such as logistic regression, support vector machine, random forest regression, etc to predict the likelihood of a customer to complete an offer. In addition, I will compare the models and find the best one to address this problem.

Benchmark Model

In real-world applications, logistic regression is the most commonly used for addressing classification problems[1]. That's why I will build a logistic regression model and use it as a benchmark.

Evaluation Metrics

Since this is a classification problem, I will use the following evaluation matrices:

- **Precision:** It is the number of true positives divided by the number of true positives plus the number of false positives.
- **Recall:** It is the number of true positives divided by the number of true positives plus the number of false negatives.

[1] **Yang, Y. and Loog, M., 2018. A benchmark and comparison of active learning for logistic regression. *Patte#rn Recognition*, *83*, pp.401-415.**

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- **F-1 score:** F-1 score is a way of combining the precision, and it is defined as the harmonic mean of the model's precision and recall.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Project Design

- Environment setup
- Initial data cleansing
- Data distribution and characteristic visualization
- Feature engineering and exploratory analysis of data
- Training on the benchmark model(logistic regression)
- Training on other models
- Hyper-parameter tuning
- Comparing model performances
- Analyzing results