UDACITY MACHINE LEARNING ENGINEER NANODEGREE CAPSTONE PROJECT

@alwzhang | March 7, 2022

OVERVIEW

This project proposal is an essential starting point to the successful completion of the Capstone Project for Udacity's Machine Learning Engineer Nanodegree. The goal of the proposal is to introduce the background of the domain, finalize a business problem statement, explore the data, recommend a Data Science/Machine Learning solution, decide on a baseline model, define a set of evaluation metrics for model performance, and finally, an outline of the end-to-end development.

1. Domain Background

Starbucks was established in Seattle, Washington in 1971, by three students who attended the University of San Francisco. When the first Starbucks was open, it was a small shop in the Seattle's historic Pike Place Market, selling specialty coffee beans and freshly brewed coffee. Since then, Starbucks has quickly expanded into a global enterprise that is well known for its coffeehouses, where customers can enjoy its authentic beverages, food, packaged coffee, and more. As of 2019, Starbucks was ranked 121st in the list of Fortune 500.

One of Starbucks key marketing strategies is sending promotions and advertisements via Starbucks mobile app. Customers can order coffee for pickup, load money and pay from the app, and collect rewards points. Promotional offer and bonus points are incentives that drive more orders/sales. A customer might get offers such as:

- Informational offer (i.e., advertisement on newly launched products)
- Discount offer (i.e., percentage off the original price)
- Buy One Get One Free (BOGO) offer

Discount and BOGO offers require customers to make a minimum purchase before they can redeem the offer. Each offer has an expiration date. This project aims to effectively promote the right offers to the right customers, based on their responses on previous offers and their preferences.

2. Problem Statement

The goal of sending customers promotions and offers is to increase their purchases. However, it is not a desirable strategy to send all offers to all customers at the same time. The goal of the project is to take advantage of the transactions and demographics data to determine the offers that should be targeted to different groups of customers.

Following questions will be considered to better understand the business problem statement:

- What are the main factors that contribute to customers making purchases?
- Are offers a way to increase customer engagement?

- What kinds of offer are the most popular?
- What populations are more interested in offers?
- What offers should we recommend to different groups of customers?

3. Datasets & Inputs

The following datasets contain simulated data that mimic customer behaviors on the Starbucks app, which are:

- portfolio.json contains the details of each offer: duration, reward, type, etc
    - 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 contains demographic information of customer
    - 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 contains all customers activity: transactions, offers received, offers viewed, and offers completed.
    - 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

These datasets were cleaned and merged in a way that each row includes customers activity, customers demographics and offers metadata.

4. Solution Statement

This problem can be solved by classification machine learning models, such as Random Forest, Gradient Boosting Trees etc. The model should be capable of recommending the best offers to the users with reliable model performance.

5. Benchmark Model

Logistic regression is a great machine learning model for binary classification problems. Its simplicity and explicability make it an ideal candidate for a benchmark model.

6. Evaluation Metrics

Following metrics will be considered to evaluate the model:

- Precision: proportion of positive cases that were correctly identified.
- Recall: proportion of actual positive cases which are correctly identified.
- F1 score: a harmonic mean of precision and recall.

7. Project Design

The project will follow a classic machine learning project workflow:

- Frame the problem and look at the big picture
- Environment setup for Jupyter and relevant packages
- Explore and visualize the data to gain insights
- Prepare and clean the data to better expose the underlying data patterns
- Feature engineering
- Build a baseline model
- Explore more model options, train models, and shortlist the best ones
- Tune the model
- Evaluate the model performance
- Present final results, tidy the code, and document the whole process