Machine Learning Engineering Nano Degree

Capstone Proposal

Udacity, March 2020

Ali Rafieh

Domain Background

Starbuck is most famous chain store in the world in coffee industry. One big part of company development, in gaining new customers and keeping existing customers, comes from targeting advertisement. This ads should be subjective and attractive.
For this matter, Starbucks using its mobile app, for customers whose using that. Company every few days, send out an offer to its customers by 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).
In this project, we are going to help Starbucks in this targeting offer technics by using and exploring the datasets company provided.

Problem Statement

The data set that is going to be used for this project are simulated data that mimics customer behavior on the Starbucks rewards mobile app. As said, 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 offer during certain weeks.
Not all users receive the same offer, or any offer.
The goal is determine which kind of offer, if any, to send to each customer based on their purchases and interaction with the previously sent offers. So, building a machine learning model that predicts how Starbucks customers will respond to an offer based on demographics and offer type, is subject that will followed.
For demographic data for each customer, three type of classification supervised machine learning models, feeding in the data from three combine data (portfolio, profile, transactional) will be used: GaussianNB, Decision Tree and Support Vector Machine (SVM). Finally a Logistic Regression apply on data.

Datasets and Inputs

The data consists of 3 files containing simulated data that mimics customer behavior on the Starbucks Rewards mobile app.
Portfolio.json contains info about the offers, profile.json contains info about the customers, and transcript.json contains info about customer purchases and interaction with the offers.
The data contain information about 10 offers: 4 BOGO, 4 discount, and 2 informational. It consist of 17,000 customers and a transcript containing 306,534 purchases and offer interactions.
A customer can interact with an offer by receiving it, viewing it, or completing it. It is possible for a customer to complete some offers without viewing them.
To split the customer data into training/validation/testing sets , a 60/20/20 split percentage, respectively for the customers will be used. So, 10.2k customers will be for training, 3.4k will be for validation and 3.4k for testing.
The dataset seems balanced. To determine this looked at following value counts for all events listed in transcript.json :

transaction	138953
offer received	76277
offer viewed	57725
offer completed	33579

Percentage customer who received an offer and complete it are 55.97% ((76,277 – 33,579) / 76,277). That means that 55.79% of the people completed their offers, while 44.03% received offers but did not complete. These percentages are close enough to consider this a balanced dataset.

Following describe the different datasets 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

portfolio.json

Range Index:(10, 6)

- 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

Range Index: (17000, 5)

- 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

Range Index: (306534, 4)

- 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

As this data set is a simplified version of the real Starbucks app, so the underlying simulator only has one product whereas Starbucks actually sells dozens of products. By using this simple-simulated-dataset, we can draw demographic chart for each customer and then by using classification supervised learning ML models, feed in data from three combine data (portfolio, profile, transactional).
I will use supervised learning models to determine the propensity for a customer to complete an offer. GaussianNB, Decision Tree and Support Vector Machine (SVM). Finally a Logistic Regression apply on data.
Logistic Regression is a common technique used for binary classification problems. Propensity models are a form of binary classification, since they are concerned whether a customer is likely to respond to an offer or not.

Support Vector Machines (SVMs) attempt to find the best dividing hyperplanes in the data to determine whether to send an offer or not.

Benchmark Model

A logistic regression model will serve as the benchmark model in this project. Logistic regression is possibly the most popular algorithm for binary classification problems in industry.

Evaluation metrics

The performance of models will be measured using two metrics, accuracy and F1 score.

Project Design

At very high level project will be done in three phase:

- Wrangling
- Exploratory Analysis
- Data Modeling

Each of these steps break out into several more steps, which following steps highlight their bold items:

1. Create a conda environment on my local computer and install necessary packages.
2. Download the data from the Udacity Workspace to my computer.
3. Perform high-level exploration of the data.
    1. Graph distribution of variables to better understand the data.
        1. Examples: Gender distribution, Age distribution, Income distribution, event distribution, etc.
    2. Graph relationships between data.
        1. Examples: total transactions by age, total offers completed by age, total transactions by income, total offers completed by income, etc.
4. Clean the data, like:
    1. Replace age 118 with NaN, since that is the default value.
5. Perform feature engineering to prepare for modeling. like:

    1. profile.json
        1. Convert “became_member_on” column to number of days they have been a member or extract year.
    2. portfolio.json
        1. Convert channel list into separate columns, with 1 indicating in the list and 0 otherwise.
    3. transcript.json
        1. Convert “value” into separate columns for “offer_id” and “amt_spent” depending on if value is an offer id or part of a transaction.
6. Build our benchmark logistic regression model and use the evaluation metric on it.
    1. This will be the benchmark used when comparing with the other models.
7. Build other models using different supervised-learning algorithms (Support Vector Machines, GaussianNB and Decision Tree).
8. Perform hyperparameter optimization
9. Run the model against the test set.
10. Use the evaluation metrics on each model and compare with our benchmark.
11. Compile results into a report and blog post.