Machine Learning Engineer Nanodegree

Capstone Proposal

Mengjia Lyu June 1, 2020

Domain Background

Starbucks is the largest coffeehouse chain in the world. Business insights are invaluable when it comes to targeted advertising. In order to maximize the effectiveness of app advertising, we need to be able to predict with reasonable accuracy what kinds of offers are most likely to lead into purchases for different customers. Machine learning algorithms can be employed to learn from transactional data showing user purchases made on the app and make predictions.

Problem Statement

The task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. More specifically, the goal is to build a model that predicts whether or not someone will respond to an offer based on the metadata available about the customer. As whether or not someone will make a purchase is inherently binary, binary classification models can be used to tackle the problem.

Datasets and Inputs

The dataset is provided by Starbucks and the data is contained in three *json* 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

You'll also want to take into account that some demographic groups will make purchases even if they don't receive an offer. From a business perspective, if a customer is going to make a 10 dollar purchase without an offer anyway, you wouldn't want to send a buy 10 dollars get 2 dollars off offer. You'll want to try to assess what a certain demographic group will buy when not receiving any offers. First I will preprocess and cleanse the data, then I'll train several possible models and pick out the one that has the most predictive power.

Evaluation Metrics

In the absence of related literature in the specific domain, the final model can be benchmarked against relatively *naive* models such logistic regression or naive Bayes classifier.

Project Design

First and foremost, data cleaning is the first step. As mentioned, I'll want to take into account that some demographic groups will make purchases even if they don't receive an offer. In other words, I need to assess what a certain demographic group will buy when not receiving any offers.

As it is a classification problem, available options include logistic regression, decision tree, SVM, naive Bayes, neural network and random forest. I can then pick out the best model by cross-validation and eventually optimizes the hyper-parameters of the selected model.

References

- Starbucks