Machine Learning Engineer Nanodegree

Capstone Proposal

Pedro Gabriel Salazar do Nascimiento
March 3rd, 2020

Proposal

Domain Background

Starbucks Corporation is an American coffee company and coffeehouse chain. Starbucks was founded in Seattle, Washington, in 1971. As of early 2019, the company operates over 30,000 locations worldwide.

Starbucks joined the Capstone Project as a way to see how to profit from Machine Learning to create value from the data about clients and their consuming behavior recollected during one month.

The topic of consumer behavior research together with Machine Learning is what has motivated me to select this project for the Capstone Project part of the Udacity Machine Learning Engineer Nanodegree Program.

Problem Statement

Starbucks wants to personalize the offers sent to a client to avoid falling into "Spam" (sending every single offer to everyone... what it usually ends up in people not paying attention to what they receive and even getting a bad impression of the company). Moreover, sending the appropriate offer to a client will make that client have a better image of the company and willing to profit from the offers and thus consuming more (and more gladly).

Therefore, my project will analyze the data provided by Starbucks and will create a prediction model that will let us know if one client will take (complete) one offer. The data will be used to train the prediction model as well as to test the accuracy of the predictions.

Datasets and Inputs

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

My solution to the problem is to construct a Machine Learning model to predict if a client will take a certain type of offer (discount of BOGO – Buy One Get One-). This way, Starbucks can decide whether to send a discount offer, BOGO offer, both or none of them to a client.

The characteristics of the client that my model will use to make the prediction will be:

- age
- gender
- income
- time as client (represented by the year of membership)

Benchmark and Evaluation Metrics

I will use 80% of my model data to train my prediction model and the other 20% to evaluate the accuracy of the prediction model.

I will create my own evaluation function to provide statistics about the % of successful predictions, false positives and false negatives.

This function will allow me to benchmark different prediction models.

I will select and evaluate two different prediction models (from scikit-learn library).

Nevertheless, the prepared Train and Test Datasets and the Evaluation Function could be use to evaluate and benchmark a lot of different prediction models (even self-developed ones).

Project Design

(approx. 1 page)

First, I will explore the data in the three datasets, see the size of the datasets, the type of content and the kind of values.

During that exploration, I will start performing some transformation of the data, changing format or disposition, to make the data easier to work with and to analyze.

Then, I will analyze the data. The distribution of the values inside the different fields. My aim will be to discover regularities and particularities in order to decide which data will serve to my purpose and which not. In other words, which data I consider could be used to predict if a client will take (complete) an offer.

During this analyzing and decision process I will filter the data, dispose some part of it and probably perform some more transformation of the datasets.

After deciding which data will serve to my purpose, I will prepare the Train and Test Datasets to train the prediction model and to evaluate its suitability.

As commented in the previous chapter, I will develop a method to evaluate the accuracy (and hence the suitability) of a prediction model.

I will then select a couple of prediction models, train them and evaluate them.