Machine Learning Engineer Capstone Proposal

Background

Starbucks is a worlwide brand with coffe stores in multiple countries. To incentivize consumption there are two main types of offers that they promote `BOGO` (buy one get another) and `discount` offers. Usually they have a limited duration, for example, two weeks and a difficulty (the amount of money to spend). However, they also involve a reward, for example in BOGO offers your reward is the same as the amount that you bought.

Problem Statement

Starbucks has collected some data that simulates the real behavior of their customers. In this way, they aim to develop a machine learning solution that allows them to take the decision of going or not forward with a certain offer.

The data

Characteristics of their customers, offer's description and historic transactions from the 30 days test period are provided. Next, we give a brief description of the data.

Profile Data. Customer description

- `age`: Customer age
- `gender`: Customer gender
- `id`: Unique customer identifier
- `became_member_on`: Date as Year/Month/Day when the customer became a Starbucks member.
- `income`: Customer income.

Portfolio Data. Offer description

- `reward`: Awarded money for the amount spent
- `channels`: Offer's distributed channels.
- `difficulty`: Money to be spent to receive a reward.
- `duration`: Offer's time validity.

- `offer_type`: Type of offer
- `id`: Unique offer identifier.

  Transcript Data. The transactional data

- `person`. Unique customer identifier.
- `event`. Event registered. Offer Received, Offer Viewed, Offer Completed or Transaction.
- `value`. Offer id or transaction amount
- `time`. Hours since test started.

How to solve the problem?

There are different ways to approach the problem, but, we're going to focus on calculating the completition rate, that is,

$$CR = \frac{\text{\# Offers Completed}}{\text{\# Offers received}}$$

or the attractiviness rate

$$AR = \frac{\text{\# Offers Viewed}}{\text{\# Offers received}}$$

depending on what our final goal is after exploring the data we're going to decide to predict this value for each offer.

How to compare our solution? (Benchmark)

We will initially go with a simple model like a Linear Regression Model, and then improve on that with a non-linear one like a Support Vector Machine or a Decision Tree. However, one way to compare will be also to test the model on previous offers or new offers.

How to evaluate the model (metrics)

To measure the accuracy we decide to go with the $MAPE$ or mean absolute percentage error, that is,

$$MAPE = \frac{1}{n} \sum_{n} \left| \frac{y_n - \hat{y_n}}{y_n} \right|,$$

where y_n is the real value and $\hat{y_n}$ the prediction. Additionally, we could measure how much variability is capable of explaining the model through the R^2 (r-squared).

Project Design

To understand the relationships between the variables and possible interaction effects, we're first going to perform a visual exploratory analysis, in case we would like to formally prove an insight, we will use statistical test's.

Gaining familiarity with the data will allow us to propose new ways to build variables based on the given ones, that is, feature engineer.

Finally, a mathematical solution will be formulated to solve the regression problem. This step will highly depend on the size of the data avaible. For example, we don't have enough samples we will not implement more robust solutions like Decision Trees or Neural Networks.