Machine Learning Engineer Nanodegree
Capstone Proposal

Matteo Felici

December 29th, 2019

Domain Background

"There's an app for that" was the new motto from Apple in 2009. 10 years later, this statement is proven to be true, and every company has to create a proprietary app to sell his business. That's why analysis on app usage is more crucial than ever to leverage the business: understand the customers' behavior browsing the app, predict his needs and give the correct response.
Starbucks is one of the most well-known companies in the world: a coffeehouse chain with more than 30 thousand stores all over the world. It strives to give his customers always the best service and the best experience; as a side feature, Starbucks offers his free app to make orders online, predict the waiting time and receive special offers.
I work in a web company, so my goal is to leverage my experience in analyzing this type of data, to replicate the same ideas in my everyday job.

Problem Statement

Starbucks wants to find a way **to give to each customer the right in-app special offer**. Our goal is to analyze historical data about app usage and offers / orders made by the customer to develop an algorithm that associates each customer to the right offer type. We can assess the performance of the project by measuring the correct association by applying the model to past data.

Datasets and Inputs

There are 3 available data sources, given along the Capstone instructions. The first one is **portfolio**: it contains the list of all available offers to propose to the customer. Each offer can be a *discount*, a *BOGO (Buy One Get One)* or *Informational* (no real offer), and we've got the details about discount, reward and period of the offer.
The next data source is **profile**, the list of all customers that interacted with the app. For each profile, the dataset contains some personal information like sex and income.

Finally, there is the **transcript** dataset: it has the list of all actions on the app relative to special offers, plus all the customers' transactions. For each record, we've got a dictionary of metadata, like *offer_id* and *amount spent*.

Solution Statement

My strategy is to develop a Machine Learning model to predict which is the best type of offer for each customer (with "best" I mean the type of offer that makes the customer **more propense to convert**).
I will develop a model for each offer type and then combine the results to have a "best action" for each app user.

Benchmark Model

As the benchmark result, we can extrapolate the current Conversion Rate of the offer received. Leaving out the informational offers, which have no real "conversion", the CR on the viewed offers is **43% for BOGO, 56% for Discount** (37% and 42% on all the received offers).

Evaluation Metrics

I'm going to evaluate the model comparing the "best action" proposed to the fact that the customer has actually completed that kind of offer.
I will use different statistic measures:

- *Precision / Recall*, the percentage of true positives (conversions predicted correctly) on all positives (predicted as conversions) / on all conversions. These two measures contrast each other: tuning the model to grow the precision results in a smaller recall and vice-versa.
- *F1-score*, that combines the two previous measures.
- *Area Under the ROC Curve (AUC)*, a measure that calculates the area under the Receiving Operating Characteristic Curve. This particular curve accounts that higher probabilities are associated with true positives and vice-versa.

Project Design

First, there is the **data preparation** step: we look at the data sources, understand their content and cleanse the data. For example, we aim at recreating the customer journey (from the *received offer* to the relative *transaction*) through the *transcript* dataset. Moreover, we have to join all the different pieces of information coming from the 3 data sources. Finally, we create the target variable, which is the base of all our analyses.

The next step is **data exploration**. We analyze the newly formed datasets to understand the distributions of the features and their relationship, especially with the target. We have to investigate possible missing values, data skewness and categorical features with too many categories.
Then, we need to tackle the **data preprocessing** part. After analyzing the data, we transform the starting dataset through different stages: missing imputation, categories encoding, data standardization.
After that, we **develop the model**. We create 2 different Machine Learning models, one to predict the *BOGO* propensity, the other for the *Discount* counterpart. For each model, we try different algorithms, such as Gradient Boosting, Support Vector Machine and Neural Network. Moreover, for each algorithm, we **tune the hyper-parameters** to find the set that gives the best performances.
Then, we **compare the models** and choose the best one for each type of offer.
With the 2 best models we **combine the results** in order to obtain a single type of offer to give to each customer.
Finally, we **measure the performances** of the built process and compare them with the current benchmark, to understand if the proposed solution is viable to implement the current offer attribution process.