Machine Learning Engineer Nanodegree Capstone Proposal

Abhimanyu Rajput

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, I tried to analyze and make model to predict the best offer to give a Starbucks customer. First I explored the data and see what I have to change before start the analysis. Then I did some exploratory analysis on the data after cleaning. After that I trained the data, then choose one model and improved it to get better results. In conclusion, I think that Starbucks needs to focus more on adults and Males. Also, offer more *BOGO* and *discounts* to their customers.

Evaluation Metrics

Since we have a simple classification problem, I will use accuracy to evaluate my models. We want to see how well our model by seeing the number of correct predictions vs total number of predicitons. Why choose accuracy? First let's define accuracy, the ratio of the correctly labeled subjects to the whole pool of subjects. Also,

accuracy answers questions like: How many students did we correctly label out of all the students?

It's similar to our situation right? because we want to see how many customers use Starbucks offers. Furthermore, Accuracy = (TP+TN)/(TP+FP+FN+TN). Not to forget, that this is a simple classification problem, so this is my opinion and reasoning on why to use the easiest (accuracy).

Reference :

First : https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124

Second : https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Cost function :-

Absolute Error :-

Absolute Error is the amount of error in your measurements. It is the difference between the measured value and "true" value. For example, if a scale states 90 pounds but you know your true weight is 89 pounds, then the scale has an absolute error of 90 lbs – 89 lbs = 1 lbs.

This can be caused by your scale not measuring the exact amount you are trying to measure. For example, your scale may be accurate to the nearest pound. If you weigh 89.6 lbs, the scale may "round up" and give you 90 lbs. In this case the absolute error is 90 lbs – 89.6 lbs = .4 lbs.

Reference : https://www.statisticshowto.com/absolute-error/

Mean absolute percentage error :-

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a loss function for regression problems in machine learning.

Reference : https://en.wikipedia.org/wiki/Mean_absolute_percentage_error

Project Design

The process of our analysis will be by using the CRISP-DM Process (Cross Industry Process for Data Mining) : Define Business understanding, Data understanding,Analyze the data,Modeling the data,Compare model performance and finally selecting one model and improving it.

The first step is Business understanding : The objective here is to find patterns and show when and where to give specific offer to a specific customer.

The second step is Data understanding : First, *portfolio* dataframe ,Then, *profile* dataframe , Finally, *transcript* dataframe:

The third step is Data preparation and wrangling : Based on what we have seen in the previous step, there needs to be some work to prepare the data for analysis and modeling.

The fourth step is Analyzing the data : For this part, it will be divided into Univariate Exploration and Multivariate Exploration.

The fifth step is Modeling the data : I tried to make a model that can identify which kind of offers we should give a customer. Because my model will guess the *offer_type*, I will only get those transcripts with offer id's. So I will ignore all transactions without offer id's.

The sixth step is compare model performance : Now that we have trained the data, it's time to evaluate their performance based on accuracy.

The seventh step is model improvements : : After using *Grid Search* with *Logistic Regression* we managed to get better results.

Finally,conclusion and Improvements.