Machine Learning Engineer Nanodegree

Capstone Proposal

Omar Hazem

omarhazem6@gmail.com

1-9-2020

Domain Background

Starbucks Corporation is an American multinational chain of coffeehouses and roastery reserves headquartered in Seattle, Washington. As the world's largest coffeehouse chain, Starbucks is seen to be the main representation of the United States' second wave of coffee culture, As of early 2020, the company operates over 30,000 locations worldwide in more than 70 countries. Starbucks locations serve hot and cold drinks.**Starbucks** offer free application to provide best in class services to its customers which in return leverage its business , from this services Starbucks offer its clients offers to attract them to spend specific money or do specific action to redeem their price.

Due to that Data Analysis is needed to know customers preferences and provide them personalized experience and provide them with offers which they are interested in .

Problem Statement

Starbucks want to provide its customers personalized offers that they are most probably to take action to redeem them , we can do that through analysis of data recorded earlier to predict every customer own preferences

Datasets and Inputs

3 Dataset are provided , Details in table Below

Portfolio dataset :	● reward: (numeric) money awarded for the amount spent ● channels: (list) web, email, mobile, social ● difficulty: (numeric) money required to be spent to receive reward ● duration: (numeric) time for offer to be open, in days ● offer_type: (string) bogo, discount, informational ○ There are three types of offers that can be sent: buy-one-get-one (BOGO), discount, and informational. In a BOGO offer, a user needs to spend a certain amount to get a reward equal to that threshold amount. In a discount, a user gains a reward equal to a fraction of the amount ● id: (string/hash)
Profile dataset	Information about rewards program users ● gender: (categorical) M, F, O, or null ● age: (numeric) missing value encoded as 118 ● id: (string/hash) ● became_member_on: (date) format YYYYMMDD

	● income: (numeric)
Transcript dataset	Customers’ transactions info. ● person: (string/hash) ● event: (string) offer received, offer viewed, transaction, offer completed ● value: (dictionary) different values depending on event type ● offer id: (string/hash) not associated with any "transaction" ● amount: (numeric) money spent in "transaction" ● reward: (numeric) money gained from "offer completed" ● time: (numeric) hours after start of test

Solution Statement

Creation and deployment of Machine learning algorithm to predict which offer is the best to offer for each client

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

To Evaluate such machine learning algorithm I am going to depend on some measures (statistical measures):

***Accuracy** : ratio of correctly predicted observation to the total observations.*

Accuracy = TP+TN/TP+FP+FN+TN

***Precision** - ratio of correctly predicted positive observations to the total predicted positive observations.*

Precision = TP/TP+FP

***Recall** (Sensitivity) - ratio of correctly predicted positive observations to the all observations in actual class*

Recall = TP/TP+FN

***F1 score** - F1 Score is the weighted average of Precision and Recall.This score takes both false positives and false negatives into account.*

F1 Score = 2(Recall * Precision) / (Recall + Precision)*

Project Design

First Step : **Data exploration and analysis** : Take a look at the three datasets and understand what they represent , investigate possible missing values, and see how datasets can be joined and merged together to get one general and complete combined datasets, This step also contain data visualization and plotting to further understand the data

Second step : **Data cleaning** : Make any data cleaning operations to get cleaned datasets with no missing or misleading values

Third step : **Prepare Data for Modeling** : See which features present the most value and which columns to abandon , make any needed categorical to numerical conversion or any hot encoding needed to prepare the data to be feed to different models

Fourth Step : **Develop Models** : Develop different machine learning models(classification models) to get the predictions from different models, This step contain the model creatin , training and predicting process

Fifth Step : **Models Evaluation** : In this step models are evaluated , best one is chosen