CAPSTONE PROPOSAL

Vanessa Ezeoke

April 14, 2021

Proposal (The Starbucks Project)

DOMAIN BACKGROUND:

Driving customers sales behavior in the retail space via digital channels (such as email, mobile and social media, etc.) has been shown to be effective and has become an important tool in creating longevity of a brand, increasing profits and engagements, as well as building long lasting relationships with customers. This challenge was a test done by the Starbucks team, to gather data over a period with which customers were either provided with information or offered discounts or Buy One Get One (BOGO) free deals on Starbucks products.

PROBLEM STATEMENT:

The ultimate problem is to decide the customers who are likely to take up an offer (discount or BOGO), view informational offers shared by Starbucks and the customer who are very unlikely to respond to any engagement by Starbucks.

After these can be provided, the solutions would be solved:

- Understanding the customers and their propensity to taking up offers to purchase more products and the specific type of offers they prefer.
- Understanding the customers and their propensity to viewing information from Starbucks.

DATASET AND INPUTS:

The data received are on Udacity workspace. It 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 (17,000 customers x 5 details including their income, age, date of membership and gender)
- transcript.json - records for transactions, offers received, offers viewed, and offers completed (306534 transactions X 4 rows with details on when, how much was spent and which customer did the spending)

SOLUTION STATEMENT:

The solution to this problem is a supervised learning approach. After data clean up and feature engineering, two models will be built one to classify customers who are likely to read informational offers and the other to classify customers who are likely to take up promotional offers on purchases. This will enable the Starbucks team to do more targeted marketing to optimize sale and personalization of their offerings to customers.

BENCHMARK MODEL:

A way to bench mark the model will be to use a Logistic Model Classifier model for both models. For the first model with 4 classes, An accuracy of 25% will show that the model is randomly guessing the classes. For the second Model other than the logistic baseline model, an assumption will also be made that all customers are interested in receiving promotional offers. There are two classes for the second model so the bench mark will be 50% accuracy for the assumption.

A SET OF EVALUATION METRICS:

The evaluation of the models will be based on the Accuracy, precision and recall of the multiclass classifications performed by the model. The higher the values of the above, the better the model is at predicting whether a customer should be sent a promotional offer, informational offer or not contacted. The AUC-ROC of the model will also be provided this will be relative to the response of customers to promotional offers. A Receiver Operating Characteristic (ROC) curve is used to plot the True Positive Rate against the False Positive Rate. AN ROC of 50% or below for any of the predictions is an indicator that the solution cannot accurately distinguish between positive and negative class points ie the classifier is predicting random class or constant class for all the data points.

OUTLINE OF PROJECT DESIGN:

Project is broken down into 3 sections:

- Data Clean up and feature engineering
- Exploratory Data analysis (EDA)
- Modelling

Data Clean Up:

Several methods will be used here:

- Label Encoding and one hot encoding for categorical variables
- Replacing or dropping Nan values depending on their value to the total dataset. If shown to be providing using insights, Nan values will be replaced by Mean or most frequently occurring values or a new category “Unknown – U”
- Replacing outliers with the mean.
- Create new fields such as time taken to view and offer, time taken to respond to an offer, customer’s offer conversion rates and which the offers customers have been shown to take up. The last is what will be used to classify the customers in the model.

Exploratory Data Analysis:

This will entail an entry point into the world of solving this problem. An in-depth analysis of the demographic (age, income, gender, etc.) and customer behavior (interested offers, transactions, response to offers, view rates, etc.). AN example is the create a pair plot of customers who are interested in all types of offers, checking their incomes, age groups, gender distributions, view rates, etc. to see if a pattern emerges.

Modelling:

This entails breaking the model into a train and test and scaling. Smote oversampling will be also done on the minority data set if the data set is imbalanced. A series of models will be used to predict against the train set. All models will be evaluated and the model with the best performance will be tuned further for optimal performance. Feature importance of the models will be checked, and final comments provided.