Capstone Proposal

Domain Background

Founded in 1971 in the city of Seattle, USA, Starbucks's coffeehouses have become a beacon for coffee lovers everywhere[1]. As a marketing strategy, Starbucks introduced a loyalty program called Starbucks Rewards since 2009, and till the end of 2018 the program owns over 16 million active members.

Starbucks sends out different offers to users of this program. The offers include a discount, BOGO (buy one get one free), or merely an informational advertisement for its product. Instead of sending out offers randomly, it's better to send the offer that will most likely to increase customers' buying tendency. With proper unsupervised and supervised algorithms, we can create a powerful framework that allows us to best understand customers' needs and behaviors[2].

Problem Statement

Some customers viewed the offers, while others didn't. For those who viewed the offers, some redeemed them while others didn't. Furthermore, some customers made the purchase without viewing the offers. Regarding different customer behaviors, I want to build a machine learning model that predicts whether or not a customer will respond to an offer.

Datasets and Inputs

Udacity provides 3 files in json format.

“portfolio.json” contains information of the offers, including offer id, types, period of validity, the amount a customer must spend to redeem it, the amount a customer will receive after redeeming it, and channels.

“profile.json” contains information of the members, including member id, gender, age, the date of registration, and income.

“transcript.json” contains logs of transactions, including member id, offer id, event description, and time.

Solution Statement

My approach to this problem is to predict the probability of a customer to respond to each offer, and three classification algorithms will be used: logistic regression, support vector machines, and gradient boosting decision tree.

Benchmark Model

I will use logistic regression as benchmark model and tune the hyper parameters of the other models for higher scores.

Evaluation Metrics

I will use accuracy and recall score to evaluate the models’ performance, because it would not hurt much to send offers to customers who will buy the coffee at original price, but it would be devastating not to send offers to

customers who will only buy the coffee with discounts. As a result, the focus is on increasing true negative counts as well as decreasing false negative counts.

Project Design

The steps of the project are:

- Preprocess the data: in this step I will evaluate the quality the data in each file by cleaning, visualizing, and initial interpreting. The cleaning part includes filling out missing values, unifying data types, scaling values, etc.. The visualizing part includes drawing the distribution of each single feature and try to find possible trends, and drawing multiple features together and see if any correlation exists. The interpreting part includes reasoning out the trends and correlation of the visualizations and coming up with more ideas. Finally I will concatenate 3 files into one big data frame, and create new features necessary for the model.
- Train the model: in this step I will use the three algorithms mentioned before to train a classification model.
- Evaluate the model: in this step I will choose the model with highest recall score.

Reference

[1] www.starbucks.com

[2] https://dl.acm.org/doi/10.1145/3419604.3419794