Starbucks_Capstone_Proposal

January 24, 2020

Table of Contents

1 Starbucks Capstone Proposal

Udacity ML Engineering Nanodegree

1.1 Domain Background

This project revolves around the Starbucks dataset provided through Udacity. I chose this problem because I have some experience in predicting customer outcomes from self-reported data as it relates to product use and engagement. I was also interested in the extent that demographics will actually have in offer completion. Does income or age matter at all in terms of who completes offers or who simply walks in a buys something.

From the Udacity project outline:

Your task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type.

- As such this project will try to predict each of the four event types: transact, complete offer, view offer, and receive offer. With a model that can predict transaction, completion, view, and receive then we can see what types of features have predictive power for a given outcome.
- It machine learning terms, this is a multilabel classification task. Well known algorithms such as XGB, GBM, and to a lesser extent SVM can be used successfully for multilabel classification. Given the size of the dataset and relatively large number of features, a tree boosted algorithm will probably a good chioce.

1.2 Problem Statement

Simply put, I'm going to build a tree boosted algorithm to predict who will complete and offer, who will transact, and who will view/recieve but not complete.

- These are major events that can be tracked directly with the data. If a customer chooses to not opt into an offer and completes is anyway, that is only a win for business as they have spent more money and recieved $0 in reward. It also tells us something about their purchase habits. If they only transact, then again they have spent money with the need for reward. Lastly, if they complete an offer after viewing, then they are responding well to advertising and are likely spending more, or at least the same, amount of money to receive the incentive.
- Additionally, by knowing who is likely to complete, transact, view, or recieve only then we can analyze the business practices around duration, marketing, and difficulty to determine changes in engagment for those who do not complete or transact.

Once a model is build and evaluated, we can view the feature importance to understand what features are important to outcome.

1.3 Datasets and Inputs

The datasets are provided by Udacity and Starbucks. They include: - profile.csv - portfolio.csv - transcript.csv

These will provide the basis for the multilabel classifier. No further inputs will be needed.

1.4 Solution Statement

I will label the outcomes of transact, complete offer, view, and recieve only as target labels for a multilabel classifer using a tree boosted algorithm (XGB). The output for any observation in the training or validation set will a prediction if the customer will transact, complete, view, or receive only. I will them use the model weights to determine the predictive power each feature in the combined dataset.

1.5 Benchmark Model

Given the lack of pre-determined answer to this problem, I'll simply benchmark this algorithm against a random distribution of target label predictions. The goal here is to such that solution model is significantly better than random guessing.

1.6 Evaluation Metrics

I like to compute all the basic evaluation metrics for an ML model. Often, individual models will have trade-offs in preformance. Those trade-offs need to carefully weighed to understand which model will best fit the needs of the problem. The following metrics will be calculated during cross vaildation and displayed in dictionary form as follows:

- 'f1_score': {'mean': ###, 'sd': ###},
  'recall': {'mean': ###, 'sd': ###},
  'precision': {'mean': ###, 'sd': ###},
  'specificity': {'mean': ###, 'sd': ###},
  'balanced_accuracy': {'mean': ###, 'sd': ###},
  'accuracy': {'mean': ###, 'sd': ###}}

1.7 Project Design

This project will follow the basic data science workflow. - Data collation, preprocessing, cleaning. Given the categorical features in these datasets, encoding will be required. - Modeling. This will include building out a model testing framework that uses cross validation so as to be able to thorough compare various models. - Analyze results. Using the best model, determine which features best predict the defined outcomes. - Ideas for future work