Capstone Proposal

Domain Background

Nowadays Understanding user behavior and taking actions based on data is a key for customer success and profitability. Starbuck coffee, a coffee chain in America has successfully developed a mobile application platform to achieve this. Once every few days, Starbuck sends out an offer to users of the mobile app. An offer can be merely an advertisement fo a drink or an actual offer such as a discount or BOGO (buy one get one free). Some users might not receive any offers during certain weeks.

Problem Statement

I will build a binary classifier which predicts wether or not a user will complete the offer or not by using machine learning predictors.

Datasets and Inputs

There is three files.

- portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
- profile.json - demographic data for each customer
- transcript.json - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

portfolio.json

- id (string) - offer id
- offer_type (string) - type of offer ie BOGO, discount, informational
- difficulty (int) - minimum required spend to complete an offer
- reward (int) - reward given for completing an offer
- duration (int) - time for offer to be open, in days
- channels (list of strings)
- profile.json*
- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer (note some entries contain 'O' for other rather than

M or F)
- id (str) - customer id
- income (float) - customer's income
- transcript.json*
- event (str) - record description (ie transaction, offer received, offer viewed, etc.)
- person (str) - customer id
- time (int) - time in hours since start of test. The data begins at time t=0
- value - (dict of strings) - either an offer id or transaction amount depending on the record

Solution Statement

Given this problem is supervised learning, potential model options can be support vector machine(SVM), decision tree, random forest, gradient boosting, logistic regression. I will first use a XGBoost model to see how much improvements I can make. Second, I will apply deep learning model. Additionally, another potential approach is time series approach with DeepAR.

Benchmark Model

I will use a LinearLearner as a benchmark model to classifier if a customer will use an offer or not.

Evaluation Metrics

I will use ROC-AUC to evaluate models assuming the data set is balanced according to this video,

Project Design

The potential process to achieve this goal is as follows. First, I will analyze and preprocess the data set. During this phase I will combine data to find out what offer was completed by which demographics. Second, I will use logistics regression to build a benchmark model. After that, I will explore different potential solutions mentioned in the solution statement section. Finally, I will compare them using ROC-AUC to find out which model performed the best.