Submission Information

Author: Tim Gorman

Project: Starbucks Capstone Challenge

Domain Background

As part of the Udacity Machine Learning Engineer Course, Starbucks has provided a data science experiment for us to attempt. This experiment is about understanding what the best offer is for each customer demographic that can be found in the Starbucks app at an invidualized, personalized level. The way that the challenge is presented leaves the door open for different approaches to this challenge. For example, I could build a machine learning model that predicts how much someone will spend based on demographics and offer type, I could build a model that predicts whether or not someone will respond to an offer, or I could decide not to build a machine learning model at all and instead define something like a rules engine.

Generally speaking, this project falls under marketing analytics which is the field of optimizing marketing campaigns for increased return on marketing investment. This interests me because marketing analytics is part of my everyday job at Huntington National Bank (HNB). In my role at HNB, I support model building and model deployment for marketing campaigns. Tackling this project will provide me with relevant experience to the problems I'm faced with every day at work.

Problem Statement

As described in the previous section, there are multiple ways to analyze the Starbucks dataset. For this project, I have chosen to build a model that predicts whether or not individuals will accept the different offers presented through the Starbucks app. We will define the solution, benchmark model and evaluation metrics in a later section.

Datasets and Inputs

The data is provided in the AWS Starbucks Capstone Challenge work space. The data 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 * 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

To use this data I will need to download it from the provided workspace and upload it into my AWS Account for this section of th class.

Solution Statement

To model this problem I will use a package called LightGBM to predict whether or not individuals will accept the different offers presented through the Starbucks app. LightGBM is a gradient boosting framework that uses tree based learning algorithms. Tree based algorithms are well-suited for tabulaur data like in this Starbucks dataset and LightGBM in particular is advantageous because it is designed to be fast, have low memory usage, allow for parallel, distributed, and GPU learning, and handle large-scale data.

Benchmark Model

I will use a logistic regression as a simple model to benchmark my lightgbm model. It is one of the simplistic classification models available and is capable of operating on tabular data like in this problem. I will compare the AUC (area under curve) score between the two models.

Evaluation Metrics

I will measure success based on the AUC score (area under the ROC curve). This is an appropriate metric for classification problems that will give me a sense of the false positive and the true positive rate (extractable from the ROC Curve). Beyond that, AUC is a desirable metric for the following reasons: (1) AUC is scale-invariant (doesn't depend on absolute value of scores) and (2) AUC is classification-threshold-invariant (measures quality of prediction independent of selected classification threshold). I will additionally consider other metrics such as F1 Score as I work through the problem so that I can better inform my assessment.

Project Design

My solution will be developed using Sagemaker Studio in the AWS account associated with this. After appropriately analyzing and cleaning the data in a notebook, I will split the data into a training sets, validation sets, testing sets. I will then save those data sets in S3 to be used be a Sagemaker Training job. I will develop two scripts that will allow for training inference with the Logistic Regression and LightGBM models, respectively. Then I will train and test the models using sagemaker processing. As mentioned before, I will be using AUC score to compare each models ability to successfully predict whether or not a customer will act on an offer. Each model will be trained on the same training set and tested on the same testing set. Once training and testing are finished I will deploy the light gbm model using a sagemaker endpoint and then I will invoke that endpoint using example input data.