Udacity Capstone: Google Analytics Customer Revenue Prediction

Background:

This project's aim is to analyze Google Merchandise Store data, specifically the customer dataset, in order to run predictions for revenue per customer. This is a form of a customer lifetime value model. Essentially, this project's aim is to use machine learning, specifically regression, in order to predict how much a customer may spend in the store given customer attributes and segments.

Problem Statement:

Can we accurately predict how much a customer will spend, thus allowing marketers to target high spending customers with more advertisements?

Data Sets and Inputs:

The data that is being used is transaction data for a subset of dates. The full dataset covers data from December 1st 2018 to January 31st 2019; however, this data is quite large, thus I will use a subset of the data. I will be predicting visitors' spend, thus our primary key will be fullVisitorId.

The schema of the data is as follows:

- fullVisitorId- A unique identifier for each user of the Google Merchandise Store.
- channelGrouping - The channel via which the user came to the Store.
- date - The date on which the user visited the Store.
- device - The specifications for the device used to access the Store.
- geoNetwork - This section contains information about the geography of the user.
- socialEngagementType - Engagement type, either "Socially Engaged" or "Not Socially Engaged".
- totals - This section contains aggregate values across the session.
- trafficSource - This section contains information about the Traffic Source from which the session originated.
- visitId - An identifier for this session. This is part of the value usually stored as the _utmb cookie. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
- visitNumber - The session number for this user. If this is the first session, then this is set to 1.
- visitStartTime - The timestamp (expressed as POSIX time).
- hits - This row and nested fields are populated for any and all types of hits. Provides a record of all page visits.

- customDimensions - This section contains any user-level or session-level custom dimensions that are set for a session. This is a repeated field and has an entry for each dimension that is set.
- totals - This set of columns mostly includes high-level aggregate data.

Solution Statement

An XGBoost Regressor will be used in order to minimize the RMSE and accurately predict a customer's revenue.

Benchmark Model

There are many benchmark models on Kaggle for this competition--many of which use LightGBM; however, I will actually use a Linear Regression as a baseline test. This will be a non-regularized linear model.

Evaluation Metrics

To evaluate our accuracy, I will be using Root Mean Squared Error (RMSE) (similar to many regression-type problems). Our outcome however, will be the natural log of revenue plus one and our prediction will just be the natural log of the predicted revenue (no plus one).

Project Design

This will be fairly straightforward and standard in terms of project design. The majority of training can be done in one notebook; however, I may decide to use a RandomForest Regressor rather than an XGBRegressor, and thus I will have a sci-kit train model.py file. EDA and such will be done in the main notebook.