Project Proposal

Problem Statement:

In many business, only small percentage of customers produce most of the revenue. The marketing teams are challenged to make appropriate investments in promotional strategies to attract customers. We are challenged to analyze a Google Merchandise Store (G-Store) customer dataset to predict the revenue per customer. As a result, the outcome will hopefully bring better decision making in the market team and better use of marketing budgets for other companies.

Performance Measure:

In this competition, the performance measure will be how well we can predict the natural log of sum of all transaction per user. We will be likely to use root mean square error and mean absolute error for our metric.

Outline:

1. Data Wrangling
    a. Explore the dataset
    b. Clean up missing or single value column
2. Exploratory data analysis
    a. Discover and visualize the data to gain insights
    b. How many features are meaningful?
    c. How are the values in each feature grouped?
    d. Find out anything different from customers who spent money vs those who didn't
3. Machine Learning
    a. Prep numerical and categorical features
    b. Determine which algorithm to use
    c. Better evaluation using cross-validation
    d. Use gridsearchCV to tune the hyperparameters
4. Launch, Monitor and Maintain model

Deliverables:

Git-hub Repository:

- IPython Jupyter Notebook
- Google Slides
- Comprehensive Report

Data Set:

- 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.