MSAI 349 - Project Proposal

Google Analytics Customer Revenue Prediction

Mayank Malik and Amit Adate

1 Introduction

This write-up is a proposal for MSAI 349, Machine Learning. The project is our contribution to a featured prediction competition by Google hosted on Kaggle. The links to the competition homepage and the data provided by Google are provided at the end of the document.

2 The Task

In this project we plan to participate in a live kaggle competition, Google Analytics Customer Revenue Prediction. In this competition, we're challenged to analyze a Google Merchandise Store customer dataset to predict revenue per customer. The task is to predict total revenue per user. This project will be done under an open source license.

It is a live Kaggle competition (launched September 23) with more than 3000 Teams already registered. As it is a live ongoing competition, no prior solutions are present and it is one of the genuine issues that Google wants to analyze. One of the key reasons that we chose this task was the flexibility that Kaggle provides as a platform to elevate and enhance models with consistent proceedings. In addition, Kaggle provides a percentile standing with a live leader board, this will allow us to get a working knowledge of our standing throughout the challenge.

3 The Data

The data is provided by Google, it is posted on kaggle. However it is not an off the shelf dataset, this dataset requires a lot of rework. We are going to invest a considerable amount of time in data preparation. For Instance, many of the fields are in json format, hence they require conversion from Json to flattened CSV format. There are multiple columns which contain JSON blobs of varying depth.

Kaggle platform facilitates teams to work on their projects in their own kernels, where we can discuss data, discover public code and techniques, and work on our own models. The data provided is about 1.5 GB csv files, for testing and training. The data fields in the given files are

- **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.
- **sessionId** - A unique identifier for this visit to the store.
- **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).

4 Features and Attributes

Although we are planning to do feature importance using Gini Index, Entropy etc, we believe there are a few variables that will be more likely to have more importance to total revenue than other. Some of the important features are Total pageviews, Number of hits, Visit number, Country and City.

We need to run more tests to compute feature importance of all attributes. We will be able to report more about which attributes are important after we build a baseline model.

5 Initial Approach

Since it a regression task, we will start using with basic regression techniques such as Linear regression, decision tree, KNN etc . However, then we will try ensembling techniques : Bagging - Random Forest or Boosting - Adpative boosting etc.

- **Data pre-processing** :

  Json to flatten CSV conversion. - Many of the fields are in json format, hence they require conversion from Json to flattened CSV format.

  Conversion of categorical variable to numerical variable. We are planning to do one hot encoding . However, we might use other encoding techniques if the features are very large.

Implementing standardization and normalization techniques over the dataset. Normalization is undertaken in order to scale the features to a certain range of values. Standardization is undertaken in order to scale the features in such a way that their mean becomes 0 and standard deviation becomes 1.

• **Evaluation Metric** :

Submissions are scored on the root mean squared error. RMSE is defined as:

RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$

The evaluation metric is decided by Kaggle contest creators in order to create a generalized metric to gauge each submission.

6 Links

Link - View the Competition, Kaggle

Link - View the Data, Google Analytics