Machine Learning Engineer Nanodegree

Capstone Proposal

Andreas Erga
November 21st, 2018

Domain Background

Today it's become a rule of thumb that a typical marketing department uses 30-50% of its' marketing budget on activities that bear little to no results. By analyzing and accurately segment customers according to propensity to spend a business a marketing department can optimize and allocated its marketing budget according to activities that generate income while at the same time free users of unwanted ads and promotions.

Problem Statement

The 80/20 rule has proven true for many businesses–only a small percentage of customers produce most of the revenue. As such, marketing teams are challenged to make appropriate investments in promotional strategies.

The challenge is to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer. If successful we can provide data that is actionable results and a better use of marketing budgets.

This task is from the Google Analytics Customer Revenue Prediction competition hostet by Kaggle.
Competition page

Datasets and Inputs

The dataset consists of two files, a training and test set. Each row in the dataset is one visit to the store.

- 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

The task is to predict the natural log of the sum of all transactions per fullVisitorIds during the periode December 1st, 2018 to January 31st, 2019.

$$y_{user} = \sum_{i=1}^{n} transaction_{user_i}$$

$$target_{user} = \ln(y_{user} + 1)$$

Benchmark Model

XGboost is currently one of the most widely used models in competitions. To investigate why i plan to compare XGboost to a linear regression model. I'll compare the accuracy (root mean squared error) for both models to see which is more effective of predicting revenue per customer. A well as the speed of the two models.

Evaluation Metrics

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

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where y hat is the natural log of the predicted revenue for a customer and y is the natural log of the actual summed revenue value plus one.

For each fullVisitorId in the test set, the task is to predict the natural log of their total revenue in PredictedLogRevenue.

Project Design

The first step will be to attain data. Easy step since the data is provided by the Kaggle competition Google Analytics Customer Revenue Prediction. Load in the training and test data.

Given the size of the files it is prudent to look data that safely can be removed from the set. Columns with constant values bring no information to the model and is therefore safe to delete from the set.

To get an overview of the task and prediction problem some initial exploration and analysis a pluss before preparing the data for the models for the final ABT table.

Within the training and testing, a pipeline is set up to assemble several steps feature selection, cross-validation and parameter tuning.

The results for the two models will finally be compared in the final step.