Google Analytics Customer Revenue Prediction[i]

Domain

Google Analytics Customer Revenue Prediction is a Kaggle competition to predict how much GStore customers will spend on buying Google branded merchandise. Lot of marketing dollars are spent on retaining customers, and better understanding of customers spending habits will help the marketing teams with allocation of marketing budgets. Revenue prediction also helps with retention of customers by providing sales and promotions and with financial planning. This competition is sponsored by R Studio and Google Cloud and developers are challenged to demonstrate the impact by using machine learning.

In my research I have noticed that many people have worked on similar projects, one interesting project was Revenue Forecasting for Enterprise Products (Bansal, n.d.); another one is on Municipal Revenue Prediction by Ensembles of Neural Network and Support Vector Machine (PETR HÁJEK, n.d.) and one on predicting stock prices, Equity forecast: Predicting long term stock price movement using machine learning (Milosevic, n.d.). I also came across a paper which used CNN for Sales forecast, Sales Forecast in E-commerce using Convolutional Neural Network (Wang, n.d.)

Problem Statement

The criteria for this competition is to predict the natural log of the sum of all transactions per user. For every user in the test data, the target is to calculate

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

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

Datasets and Inputs

The data for the competition is provided by Google Merchandise store (GStore) on Kaggle. The data is split in two csv files train.csv and test.csv. The train.csv file has over 900,000 rows and test.csv file has over 800,000 rows. For the capstone project I will use partial data of last 3 months starting from May 1, 2017 it will be over 200,000 rows from train.csv file. I will randomly split the data into 80 – 20, 20% will be used for testing and rest of the data will be used for training the model. My initial research into the data did not reveal that it there was time element involved, so I have chosen random split.

Each row in the dataset corresponds to one visit to the store and each user is uniquely identified by fullVisitorId field. Some of the columns in this file contain JSON blob, one of those column totals has a blob that has a field called transactionRevenue. This field has the revenue information that I will try to predict.

Solution Statement

For the solution I will first process the data by extracting the columns from JSON blobs. Then I will check for null values and remove outliers if any. Next, I will perform feature selection, and try to understand data using visualization. Since this is a regression problem, for training purposes I will compare Linear regression, ADABoost, Random Forest and SVM. Since this problem is part of competition, time permitting I will train Light GBM and Neural Network to predict natural log of transactional revenue.

Finally, I will select the best model for this problem.

Benchmark Model

I will use random forest model a benchmark for this problem and I will try to beat its performance with other models mentioned above.

Evaluation Metrics

I will be using the same evaluation criteria as the competition has set for this project, Root Mean Squared Error.

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

y is the natural log of actual summed up revenue per customer plus one

y hat is natural log of predicted revenue for a customer

I will use sklearn mean_squared_error function and apply a square root function to it.

Project Design

I will start with exploring the data and pre-processing it to remove outliers, then perform data visualization to see if there are any skewed features that would require normalization.

To train the models I will choose the models listed above and compare their results. I will evaluate the models by performing cross validation and grid search techniques. Based on the results I will pick the best model.

---

[i] Kaggle Competition
https://www.kaggle.com/c/ga-customer-revenue-prediction
https://en.wikipedia.org/wiki/Marketing_mix
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html
https://arxiv.org/ftp/arxiv/papers/1701/1701.06624.pdf
https://www.researchgate.net/profile/Petr_Hajek8/publication/228945890_Municipal_revenue_prediction_by_ensembles_of_neural_networks_and_support_vector_machines/links/55daf31c08aeb38e8a8a2f02/Municipal-revenue-prediction-by-ensembles-of-neural-networks-and-support-vector-machines.pdf
https://arxiv.org/ftp/arxiv/papers/1603/1603.00751.pdf
https://arxiv.org/pdf/1708.07946.pdf