Capstone Proposal
Google Analytics Customer Revenue Prediction

Domain Background

The project is based on E-commerce customer behavior analysis. As Internet usage increases worldwide, one can expect to see a concomitant rise in the use of this interactive tool in consumers' shopping decisions [1]. Customer analysis plays an integral part in making marketing decisions and strategic planning of any E-commerce business. Internet marketing is the process of building and maintaining customer relationships through online activities to facilitate the exchange of ideas, products, and services that satisfy the goals of both parties [2]. Customer analysis is conventionally done from demography, behavioral, attitude, physiological and social data of the customer.

Problem Statement

The problem statement is "Google Analytics Customer Revenue Prediction" for Google Merchandise Store a Kaggle competition [3]. The goal of the project is to predict the revenue generated by the customer based on the data obtained. The revenue generated is based on 80/20 rule for most business that is most of the revenue generated by a small percentage of customers. Customer analysis can help make strategic marketing decisions by figuring out the pattern of high revenue customers.

Databases and Input

Google Merchandise Store dataset has a total of 12 attributes (some are JSON collection set). The data include attributes of 2 Dates, 4 numeric identifiers and 7 categorical. The 4 categorical attributes (**'device'**, **'geoNetwork'**, **'totals'**, **'trafficSource**) are in **JSON set** and can be divided into multiple attributes for analysis. Totals JSON attribute have **TransactionRevenue** values which will be the Response/Predictive attribute.

Each of the twelve attributes which are given as part of the Google Merchandise Store dataset is explained in the table below and the dataset has user data from the year 2016 to 2017.

Attribute	**Description**
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).

A solution statement

The problem statement is to revenue values, therefore, the regression model is used to solve it. LightGBM is a gradient boosting framework which uses tree-based algorithms. It produces much more complex trees by following leaf wise split approach rather than a level-wise approach which is the main factor in achieving higher accuracy [4]. Comparing to other boosting frameworks like XGBOOST, LightGBM is faster and has high performance, therefore, its suitable to use for "Google Analytics Customer Revenue Prediction" a supervised learning problem.

Benchmark Model

The benchmark model chosen is random forest tree regressor as we are predicting continuous values. Random forest is one of the Ensemble methods as they build many individual trees used for both classification and regression applications that are built using some fraction of the data. The output is predicted by combining the outputs of these individual trees, by voting or average or anything suitable. Random forest is taken as the benchmark model as they can predict continuous values and can perform better than simple regression models and non-ensemble models like decision trees.

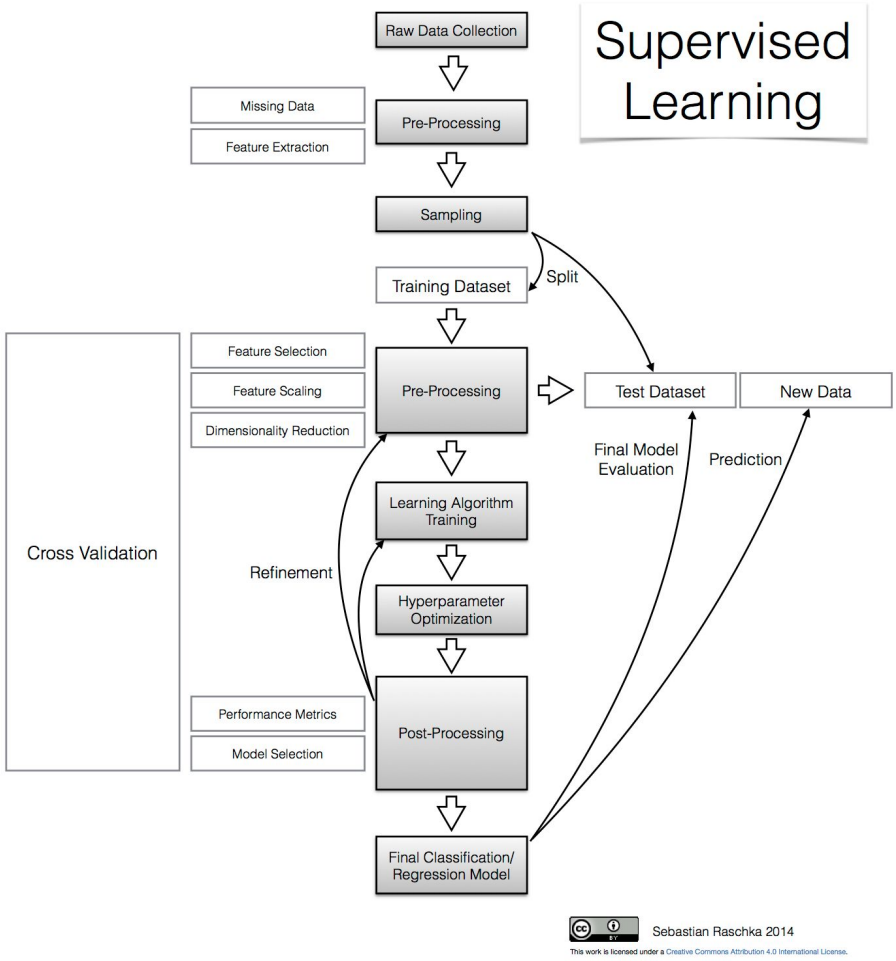
Evaluation Metrics

RMSE is chosen as the evaluation metric to measure the performance of the model for the predicted. RMSE value can range from 0 to ∞ and are indifferent to the direction of errors and it is defined as negatively-oriented scores, therefore lower RMSE values are desirable. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors and penalizing large errors through the defined least-square terms proves to be very effective in improving model performance. [5]

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Project Design

The supervised learning problem “Google Analytics Customer Revenue Prediction” will follow the steps explained in this section and the below flowchart outlines the process followed.



Data Collection

This Project is taken from Kaggle, therefore, the dataset is downloaded from the “Google Analytics Customer Revenue Prediction” competition. The dataset downloaded is in CSV format and is used as the raw data input.

Data Exploration

Data Exploration is the pre-processing step in which the dataset is loaded and exploration steps are done which are

- The attributes are converted to the required data type
- The attributes with JSON format are split up into different columns
- The attributes with most missing data are explored
- The correlation within the attributes and the required revenue is checked
- The attributes are plotted for visualization

Feature Creation

After the data exploration, the insights can be used to select the needed features in the dataset. The feature selection and creation are done and the dataset is prepared to be fed as input to create the model. In this Project the natural log of the sum of all transactions per user is to be predicted there it is taken as the target feature from the “Totals” attribute.

Model Selection

The data is cleaned and features are selected from the pre-processing step and the train and test data are separated. The model selection is the most critical part as different algorithms have to be experimented to find the best one for the dataset. In this project, the different algorithms which are to be tested are Random Forest XGBOOST and LightGBM. The algorithms will be tuned using hyperparameter optimization, the model is finalized by comparing the different models using the evaluation metrics.

Model Evaluation

Model evaluation is done using the evaluation metrics. In this project, RMSE is chosen as the evaluation metric and is used to tune the parameters of the chosen model to find the best model. Model evaluation can also help to choose between different algorithms which are used for the prediction.

References

1. G. Brashear, Thomas & Kashyap, Vishal & D. Musante, Michael & Donthu, Naveen. (2009). “A Profile of the Internet Shopper: Evidence from Six Countries”. *The Journal of Marketing Theory and Practice*. 17. 267-282. 10.2753/MTP1069-6679170305.
2. R.A. Mohammed, R.J. Fisher, B.J. Jaworshi, A.M. Cahill, Internet Marketing-Building Advantage in a Network Economy, International edition, McGraw-Hill/Irwin MarketspaceU, 2002, pp. 203 – 313.
3. Google Analytics Customer Revenue Prediction - https://www.kaggle.com/c/ga-customer-revenue-prediction
4. Which algorithm takes the crown: Light GBM vs XGBOOST? - https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/
5. Chai, T. and Draxler, R. R.: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7, 1247-1250, https://doi.org/10.5194/gmd-7-1247-2014, 2014.