ML COURSE PROJECT

Google Analytics Customer Revenue Prediction

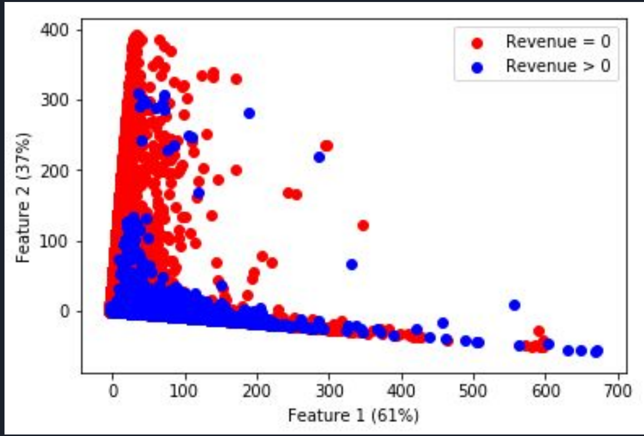
Team Members
Nihesh Anderson
Shravika Mittal
Pragya Prakash

Problem Statement

- Kaggle challenge - Google Analytics Customer Revenue Prediction
- Given the details of customers hitting GStore website and their transaction details, the objective is to predict the log revenue per user.
- Evaluation Metric: RMSE Error of all users, grouped by unique FullVisitorIDs

Dataset

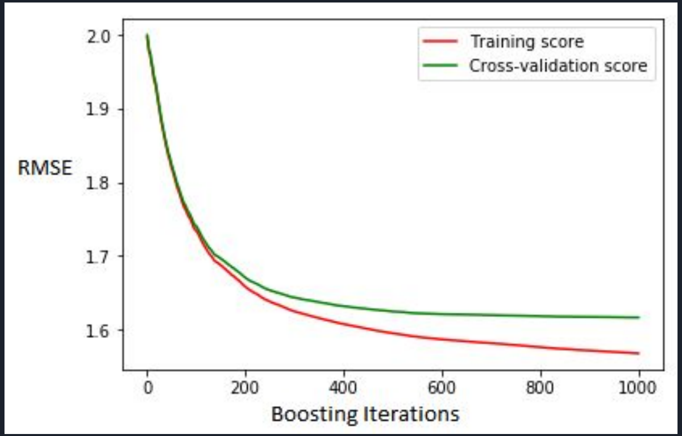
1. Almost 99% of the transactions do not generate revenue
    a. Transactions generating revenue - 11515
    b. Transactions not generating revenue - 892138
2. One user can have multiple transaction entries
    a. Unique users - 714167 (9996 users generate revenue)
3. Data visualization using PCA
    a. Number of features reduced to 2 using PCA
    b. Some overlap can be seen between the data points generating revenue and those which do not generate revenue



** More insights in backup slides

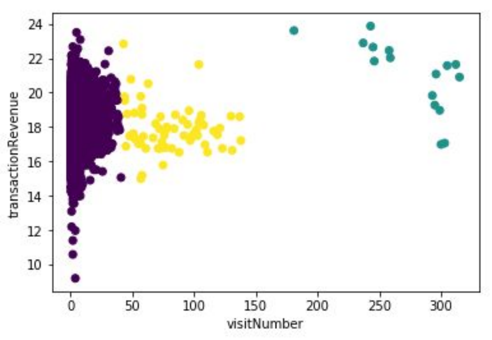
Approaches Used

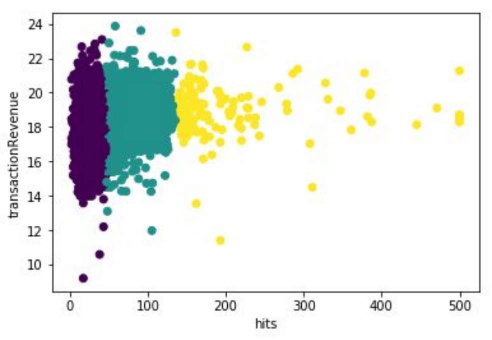
- Baseline Model - LASSO
- Robust SVM
- Dense Neural Network
- Random Forest Regressor
- Light Gradient Boosting Machines (LGBM)

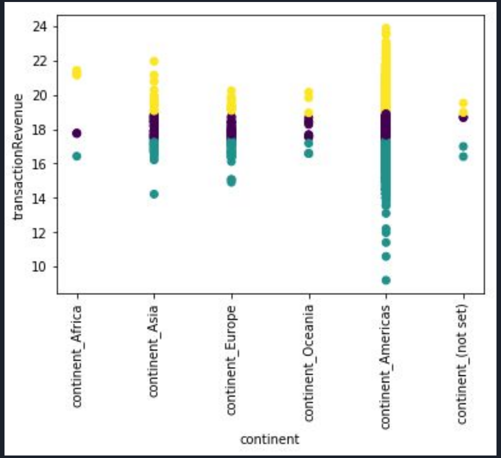


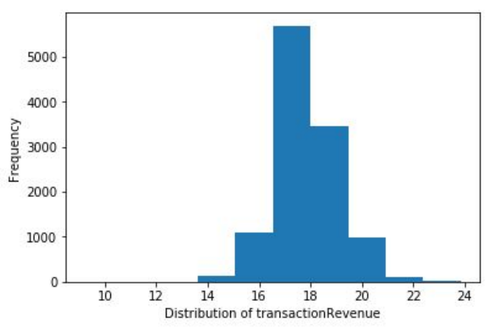
Plot of training and validation rmse scores at intervals of 50 boosting rounds of our best model: Light GBM

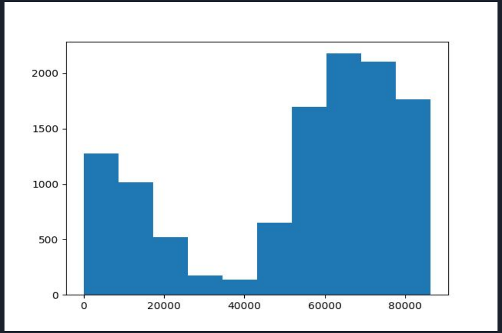
Exploratory Data Analysis









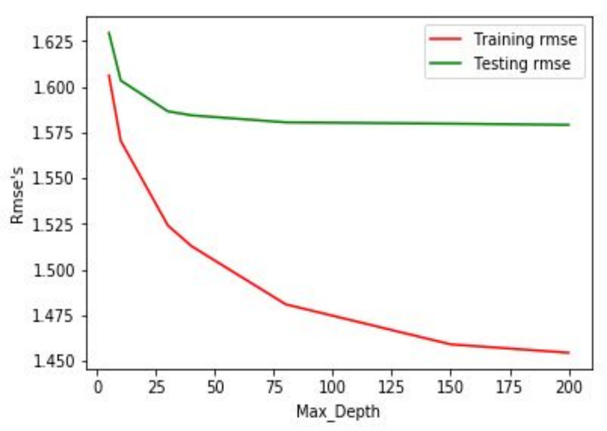


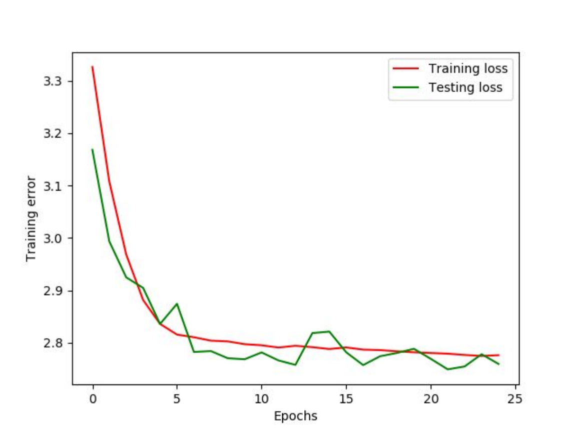
Results

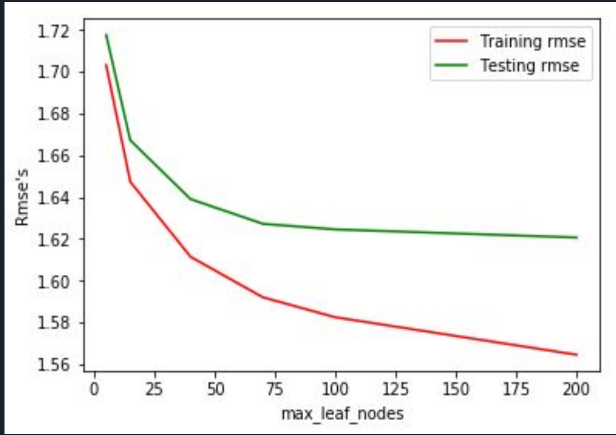
- Interim evaluation result - 1.7000 (RMSE on test data)
- Kaggle benchmark - 1.4500

Model Used	RMSE on Train data	RMSE on Test data
Baseline - LASSO	1.8159	1.8273
SVM	1.6912	1.7400
Random Forest Regressor	1.6114	1.6390
Neural Network (MLP)	1.6029	1.6282
Light GBM	**1.5265**	**1.5812**

Model Analysis and Parameter Tuning







LightGBM Training

- Max_leaf_nodes = 40, max_depth = 15
- More depth leads to more accuracy on train set, but poor generalisation, i.e., overfitting
- Max-Depth ensures good fit.

Neural Network Training

- Trained with high batch size
- Training was stopped at 12 epochs
- Early stopping ensures that the model doesn't overfit.

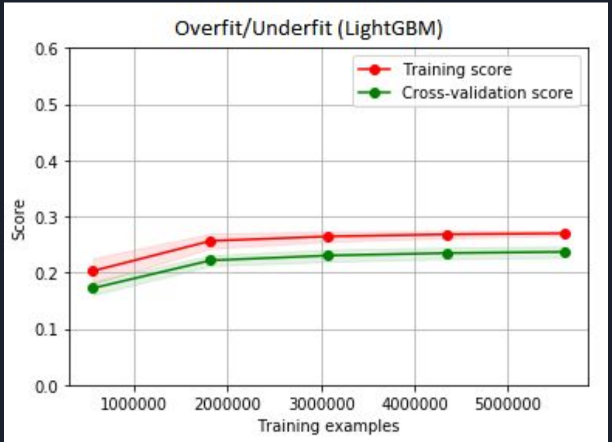
Random Forest Training

- Max_leaf_nodes = 40 (GridsearchCV)
- More leaf nodes leads to more complexity causing overfitting.
- Above trend can be seen in the graph.

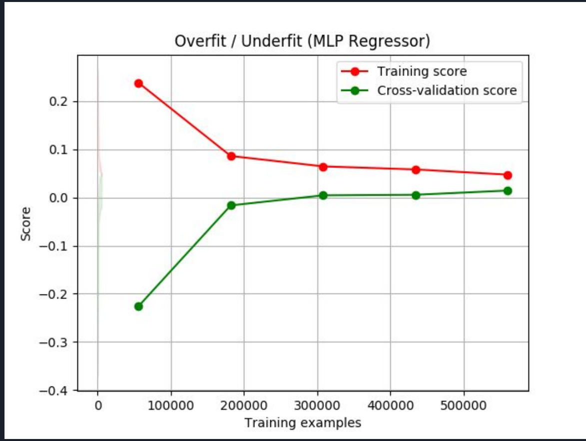
Ablative Analysis - Test Error

Component	LightGBM	Neural Network	Random Forest
Overall System	1.5812	1.6283	1.6390
isMobile	1.5855	1.6545	1.6381
hits	1.6133	1.6602	1.6420
continent	1.6142	1.7000	1.6425
pageviews	1.9097	1.9476	1.9181

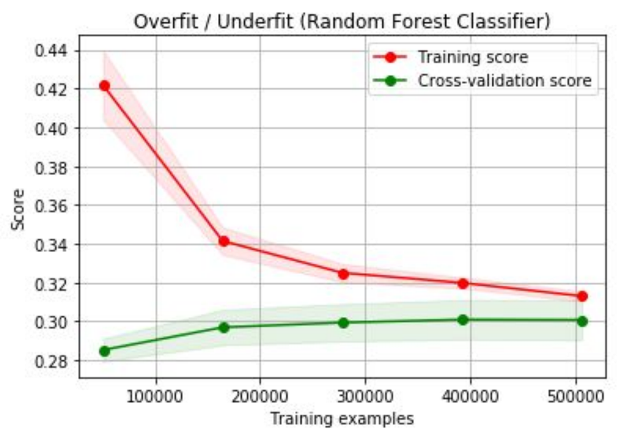
Learning Curves



Light GBM learning curve



Neural Network learning curve



Random Forest learning curve

Individual Contribution

Nihesh Anderson

- Data processing (Parsing unstructured data, encoding categorical data)
- Exploratory Data Analysis
- Trained Lasso Regression Baseline
- Tuning Neural Network Architecture using Keras

Shravika Mittal

- Exploratory Data Analysis
- Trained Decision Trees and Random Forests using scikit learn
- Feature importance from Random Forest used for feature engineering

Pragya Prakash

- Exploratory Data Analysis
- Trained SVM using scikit learn
- Trained LightGBM using lightgbm library