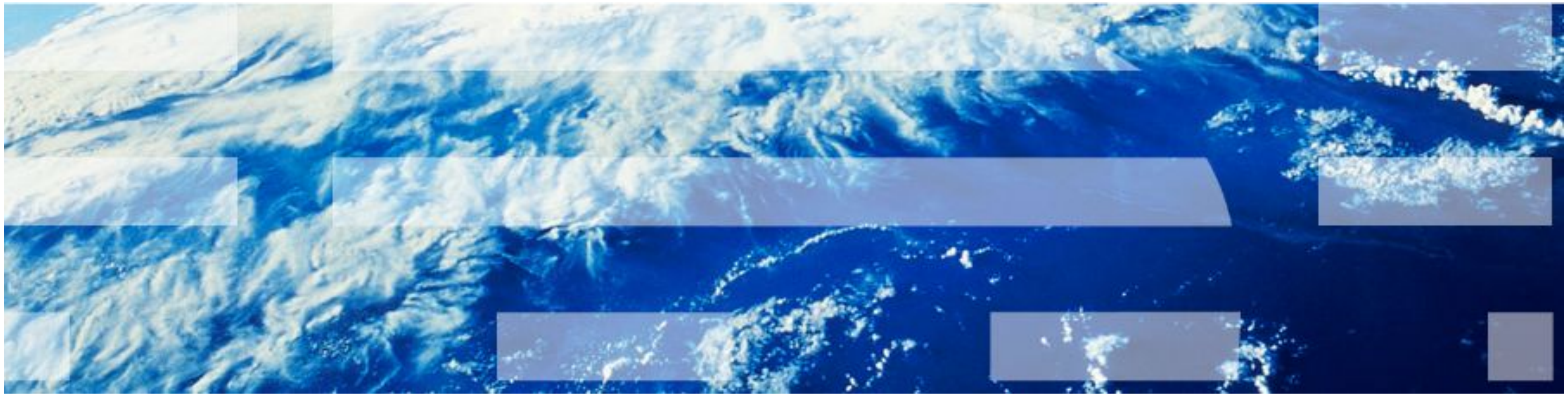
E6893 Big Data Analytics:
Google Analytics Customer Revenue Prediction

Project ID : 2018-29
Team Members (with UNI): Zhejing Hu(zh2290), Yuchong Wang(yw3081), Chi Ma(cm3700)



Motivation

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.

In our project, we are going to analyze the Google Merchandise Store customer dataset to predict revenue per customer. We hope the outcome will bring actionable operational changes and a better use of marketing budgets for those companies who choose to use data analysis on top of Google Analytics data.

Dataset, Algorithm, and Tools

Dataset

- **Around 2.1 million rows and 12 fields** 21G for training and 7G for testing.
- **Traffic source data**: Information about where website users originate. This includes information about organic traffic, paid search traffic, and display traffic.
- **Content data**: Information about the behavior of users on the site. This includes the URLs of pages that users look at, and how they interact with page content.
- **Transaction data**: Information about the transactions that occur on the Google Merchandise Store website

Algorithm

- Linear regression, regression tree, Gradient Boosting, LGBM, Xgboost, Catboost, Ensemble

Tools

- Tableau, Jupyter notebook, Spark, GCP

Exploratory Data Analysis

Target Variable: log(sum(yi) + 1)

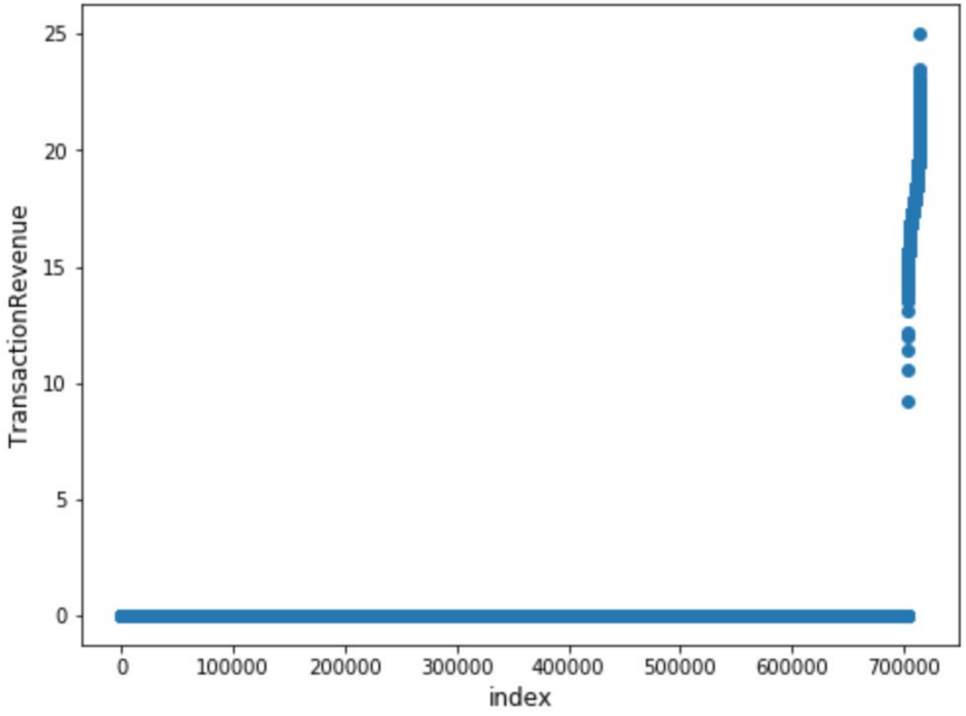


Figure 1

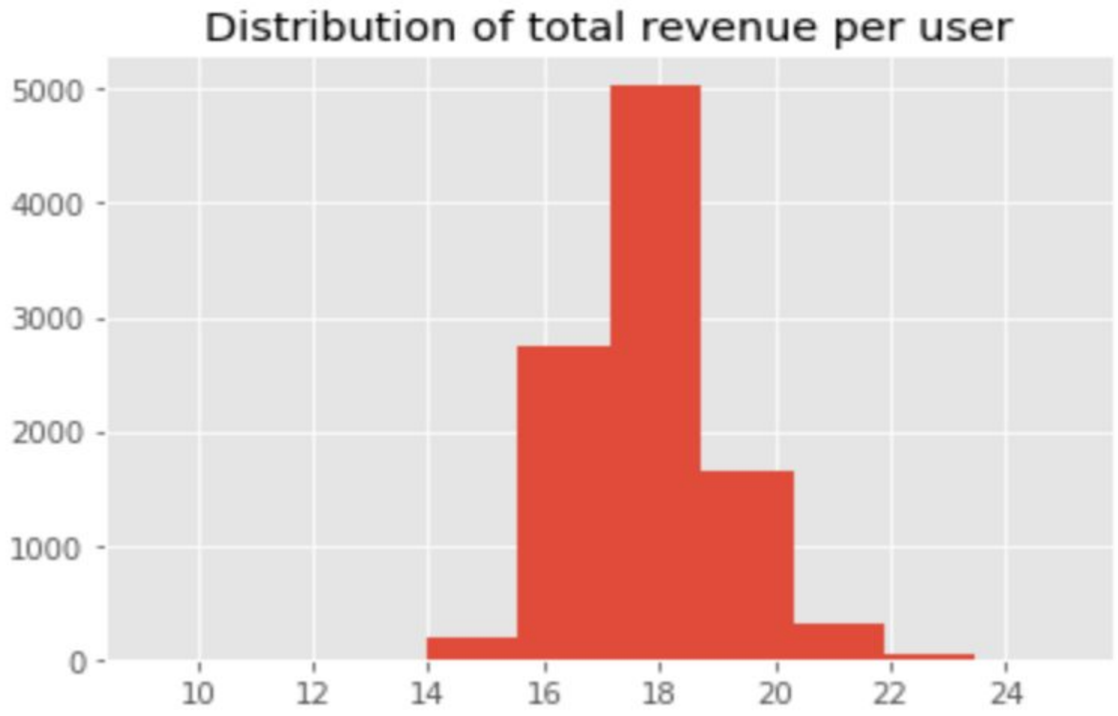


Figure 2

Exploratory Data Analysis - HTML Link

Data Preprocessing

- Convert all the json fields to a flattened csv format which generates more features
- Remove features with unique value
- Remove features with more than 80% missing values
- Impute missing data with median, mean or certain values
- Turn text features to lowercase and remove all the punctuations

Feature Engineering

- Generate some new features like weekday, month, transaction status...
- Create buckets for some of the categorical features with too much categories based on domain knowledge
- Convert other categorical features into dummy variables
- Turn text features like geo_networkDomain into TF-IDF scores
- Create aggregated features(sum, min, max, mean, median) per user

Evaluation

Root Mean Squared Error (RMSE):

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

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

Model Building

- **Linear Regression (Elastic Net)**
  - Baseline Model
  - Only a few parameters are non-zero
  - Validation RMSE: 1.786

- **Regression Tree**
  - Most Important Features: totals_pageviews, trafficSource_adContent,totals_hits
  - Validation RMSE: 1.658

- **Gradient Boosting**
  - Validation RMSE: 1.527

Model Building

- **XGBoost:**
  - Validation RMSE: 1.492
- **CatBoost**
  - Validation RMSE: 1.488
- **LightGBM**
  - Validation RMSE: 1.462
- **LightGBM(Parameter tuning)**
  - Number of leaves, Number of round, Max depth,etc.
  - Validation RMSE (After tuning): 1.40221

Results

```
Training until validation scores don't improve for 100 rounds.
[500]   training's rmse: 1.42971        valid_1's rmse: 1.43317
[1000]  training's rmse: 1.38804        valid_1's rmse: 1.4128
[1500]  training's rmse: 1.36425        valid_1's rmse: 1.40775
[2000]  training's rmse: 1.34426        valid_1's rmse: 1.40636
[2500]  training's rmse: 1.32702        valid_1's rmse: 1.40525
[3000]  training's rmse: 1.30918        valid_1's rmse: 1.404
[3500]  training's rmse: 1.29293        valid_1's rmse: 1.40308
[4000]  training's rmse: 1.27813        valid_1's rmse: 1.40239
Early stopping, best iteration is:
[4251]  training's rmse: 1.27123        valid_1's rmse: 1.40221
LGBM: RMSE val: 1.40221  - RMSE train: 1.27123
```

- LightGBM performs better than other models and achieved the lowest RMSE after parameter tuning
- Ensemble performs better than single models, but requires longer training time

Next Step:

- Parameter Tuning on other models
- Feature Engineering
  - Create more insightful features
  - Drop some unrelated features

Q & A

Thank you

Youtube link: public

https://youtu.be/u4T2H4f2Y9g

https://www.youtube.com/watch?v=3kdq
t0dhLk4