ICS 691 Final Report:
Google Merchandise Store Customer Revenue Prediction

Predict how much GStore customers will spend

Xue Gong

The structure of the presentation

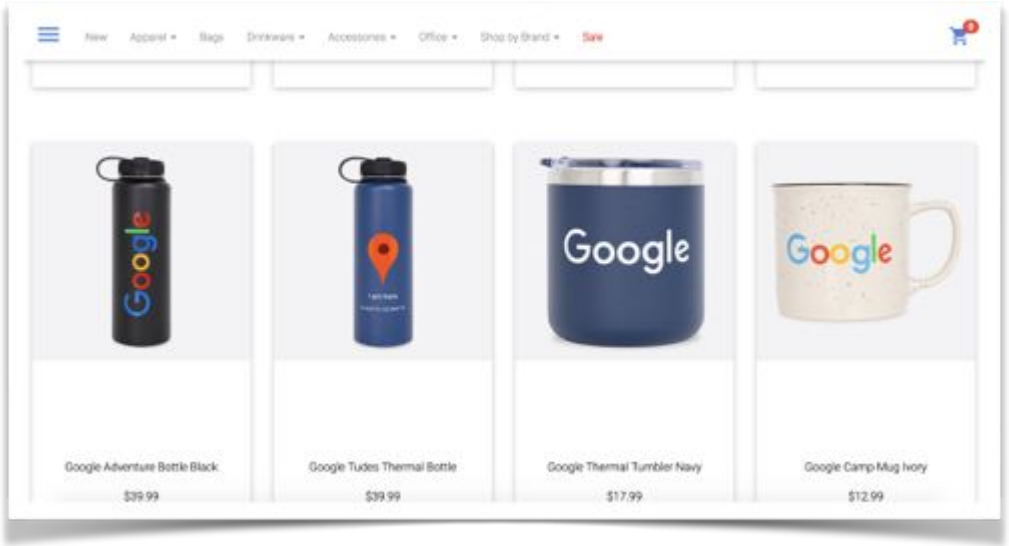
- Problem Statement
- Data Description
- Methodology
- Empirical Results
- Conclusion

Problem Statement (1)

The 80/20 rule always exists in the businesses, that is 20 percentage of customers contribute 80 percentage of the revenue.

Understand the customers' preferences and characteristics and furthermore to make appropriate investments in promotional strategies for the target customers is important for any business.

Problem Statement (2)



In this study, we take the Google Merchandise Store (Gstore) data as an example to analyse the large customer data and predict for the Revenue in the future.

Data Description -- Time Period

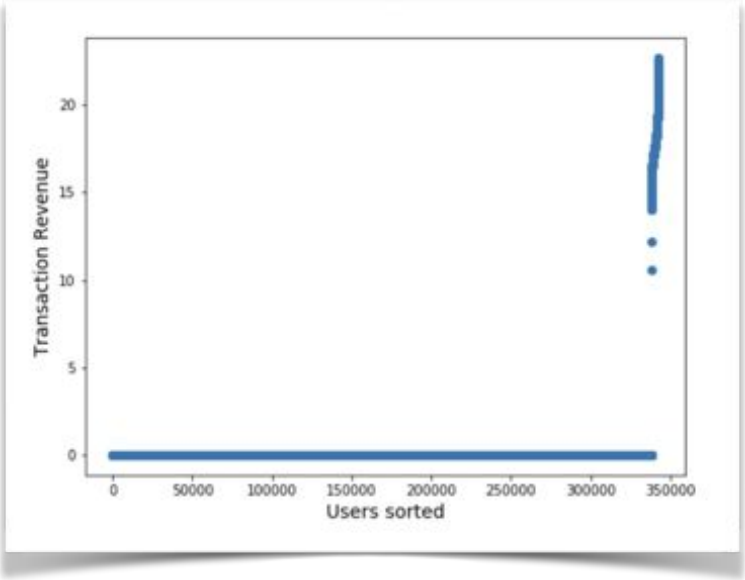
I have used the first 500,000 rows of the data set from the Kaggle competition, which is from August 3rd, 2016 to April 29th, 2018. And I separate them into the training and test set.

(**source:**https://www.kaggle.com/c/ga-customer-revenue-prediction/data)

Data Description -- Important Field

- **fullVisitorId**- A unique identifier for each user of the Google Merchandise 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.
- **totals** - This section contains aggregate values across the session. (Revenue)
- **trafficSource** - This section contains information about the Traffic Source from which the session originated.
- **visitStartTime** - The timestamp (expressed as POSIX time).

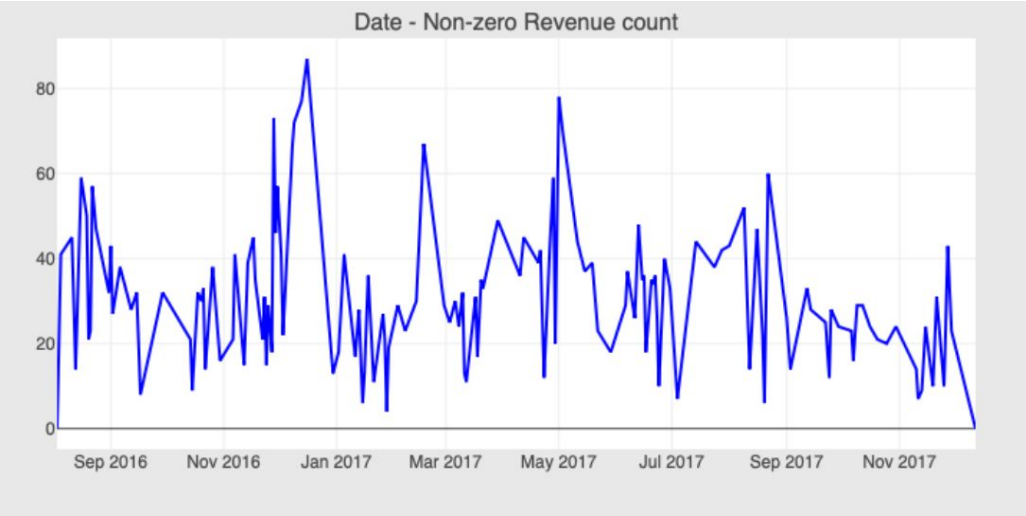
Data Description -- Does 80/20 rule exist in our data?



The statistics show that the ratio of revenue generating customers to all customers is in the ratio as 1.24%. Around 1% of the users contribute to the total revenue.

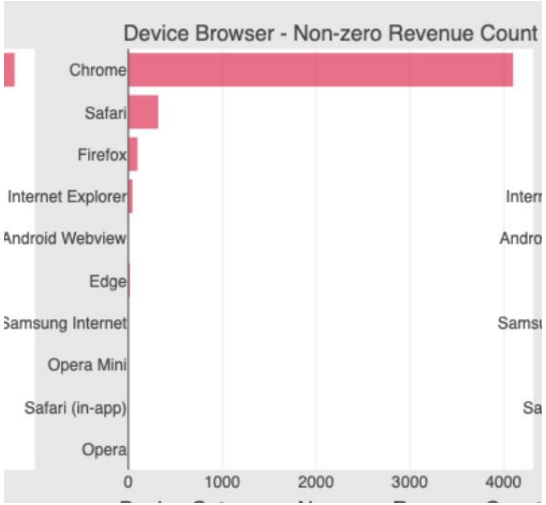
This proves that the 80/20 rule exist in our case.

Data Description--customer counts in training data



it seems there are some seasonal pattern, in the end of each month, there is a peak for the non-zero revenue.

Data Description--Device Analysis

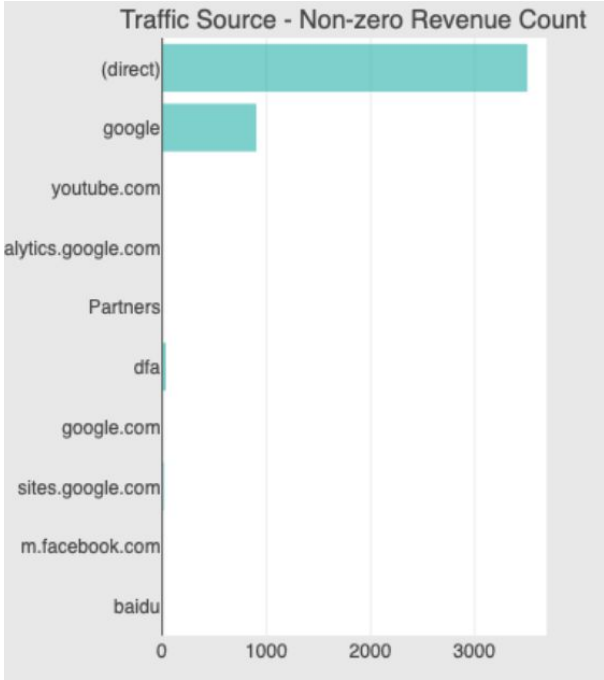


The chrome users spent the most!

Data Description--Geographic Information

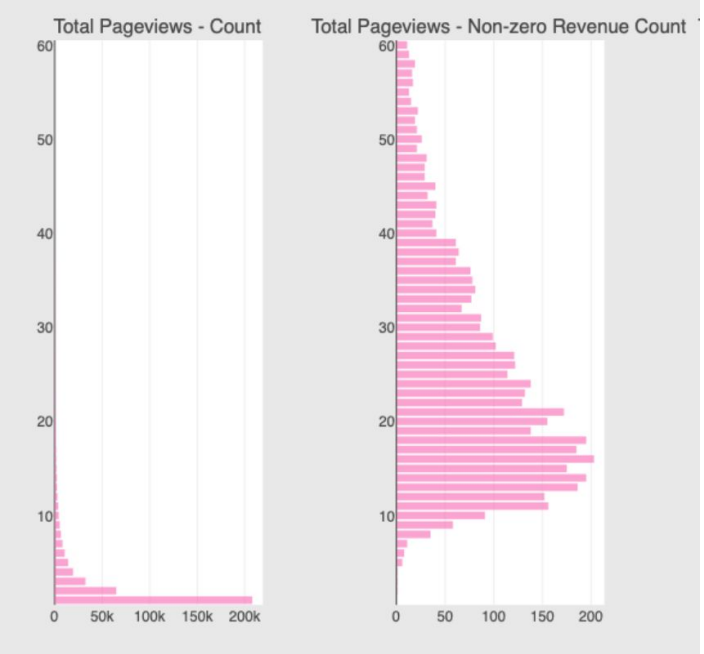
The people in North America spend the most!

Data Description -- Traffic Source



The most traffic source are from the direct visit to the website, the second one is from google.

Data Description -- Page Views



Most of the users do not click or review the page more than 10 times, however, within the data that the customer who bought something, the average number of click or the page review is around 20.

Methodology -- Prediction Accuracy Measurement

I will compare the predicted value and real data by the Root Mean Square Error (**RMSE**), which is the standard deviation of the residuals (prediction errors) , the equations are:

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

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

Methodology --- Model Description

I will use baseline model, LightGBM model and Random Forest Regression to predict the revenue in the test data and compare the performances of the models. We use the last two month in the training data as the validation data.

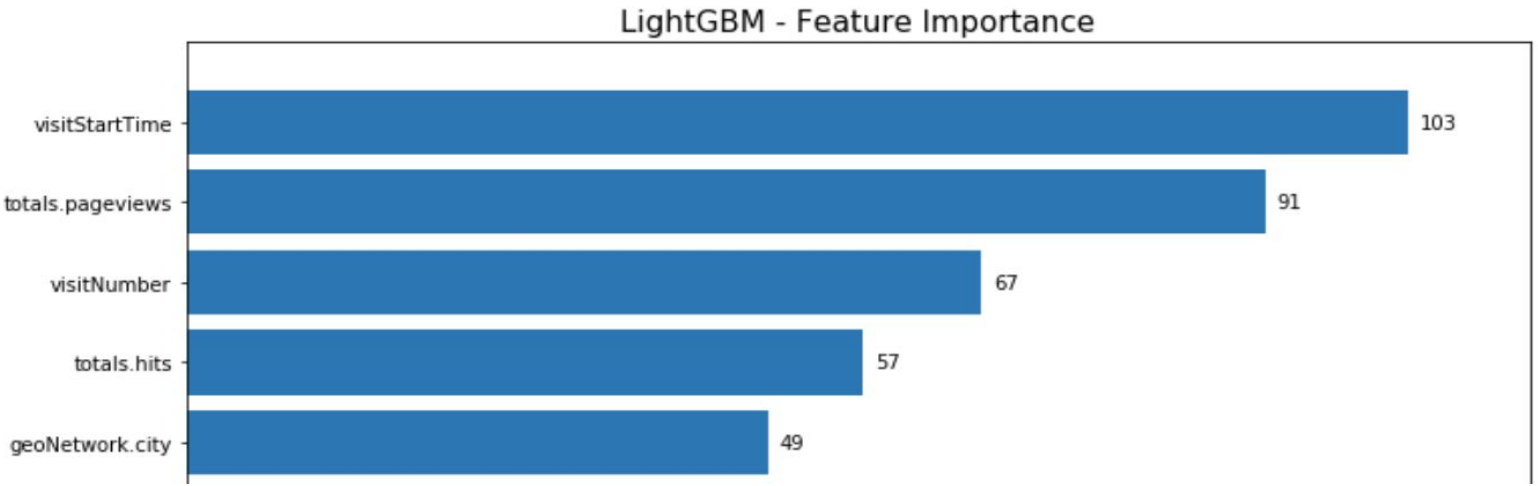
Baseline model, LightGBM model and Random Forest Regression

Baseline model: based on the naive assumption that the customers who bought products before will buy in the future.

Light GBM: Light GBM is a relatively new method for the big data analysis, it would grow the tree vertically whereas other algorithms grow trees horizontally, which makes LightGBM an effective method in processing large-scale data and features.

Random Forest model: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. It is a flexible, easy to use machine learning algorithm.

Empirical Results -- the most important features for lightGBM



Empirical Results -- the performance comparison

	RMSE	Time
Baseline model	1.664	-
LightGBM	0.019	1min 37s
RandomForest	0.0005	3min 20s

Conclusion

In this study, we predict the Gstore revenue per customer by using different features in the period of August 3rd, 2016 to April 29th, 2018, since the data set is so large, we chose the lightGBM, and random forest model to predict the revenue. our results show that Random Forest has more accuracy while the lightGBM model is faster.