Google Analytics Customer Revenue Prediction

Hao Shen

2020/11/6

Personal statement

My career goal after graduation is to be a data related consultant for consulting company such as PwC or KPMG. So, basically the project suitable for my career development should be related to commercial problems. And the Google Analytics Customer Revenue Prediction is exactly such kind of project.

The reason Google proposed this project is that they hope to better design marketing strategies by predicting the revenue from each customer, so that they can achieve greater revenue growth with smaller marketing expenses and thereby create more profits for shareholders. However, in this project, participants are only required to complete the step of predicting the revenue from each user.

In addition to being related to career goals, this project is attractive to me since its starting point is the 80/20 rule, which is, in this scenario, 80% of the revenue of GStore comes from 20% of customers. The 80/20 rule can be applied to most business scenarios, and its core idea is to grasp and solve the main aspects of problems. Interestingly, when I read others' Notebooks on Kaggle, I found that although there are a lot of content about data processing and modeling, almost no one mentions this important rule. So, I am wondering if it is possible to use this as a prior information to improve the prediction effect of the model

Question

What I am supposed to answer is by analyzing a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer. And from my personal interest, I would add more commercial related background as prior information try to its performance.

The data source

The data sets is available from Kaggle. The data set contains detailed transactions information such as date, geography, device, time, etc. from GStore and is consisted by two sub-sets:

- one is the train set from August 1st 2016 to April 30th 2018, with 13 columns, 1.71 million observations and a size of 23.67GB;
- the other is test set from May 1st 2018 to October 15th 2018, with also 13 columns but only 0.402 million observations and a size of 7.04GB.

Proposed Timeline of work

- EDA: November 14th, 2020, Saturday
- Data Processing: November 21th, 2020, Saturday
- Modeling and Validation: November 28th, 2020, Saturday
- Write up: December 5th, 2020, Saturday