Google GStore Customer Transaction Analysis

Yu Tian, Meixing Dong, Zishen Li, Duo Zhang

I. DATASET INTRODUCTION

GStore dataset covers around 900k customer shopping records of Google Merchandise Store which is provided by Google and posted on Kaggle community. The dataset includes different type of variable such as numeric, string as well as integrated formated data such as json. The dataset is not tidy and contain a large number of missing value, which will be discussed in preliminary result part.

II. PROJECT SUMMARY

Our goal is trying to provide advice on increasing the Gstores revenue based on analysis of Google Merchandise Store customer dataset. We are going to focus on following questions.

A. Statistical description about the GStrore dataset

We would like to know more about the dataset and explore the basic but interesting statistical facts such as but not limited to:

- Geographical distribution of GStores sales and accesses
- The trend of user accesses and corresponding transaction on GStore
- The influence of different customer feature such as browser, operating system, device on total transaction of each user during a certain period of time.

For this part of the project, tidying methods on data will be employed such as handling missing data as well as expanding integrated variable such as json type variable to regular multiple variable. Exploratory data analysis will be used to find out distribution and summary of each variable. Data visualization will also be conducted to find potential relationship cross variables as well as between different customer feature and total transaction.

B. Prediction of total transaction for each user

In this part of project, we are going to predict the natural log of total transaction for each user based on their behavior on Gstore website. Since our target is continuous variable, we are going to use regression machine learning methods and try to make prediction. Here are several steps we are going to implement for our prediction problem:

- Pre-process dataset such as handling missing value and outliner and deal with categorical data using one hot encoding method.
- Divide training dataset into training and validation dataset for validation purpose.
- Apply linear regression as baseline model and try to make feature selection using stepwise regression to find out most important variables.
- Add regularization for linear regression and implement cross validation for avoiding overfitting.
- Implement other regression model such as kernel smoothing, polynomial regression,decision tree regression,
- Conduct feature engineering to find out new features which may improve accuracy of final prediction.
- Use test dataset for comparing different model and technique and find out best model for transaction prediction.

III. PRELIMINARY RESULTS

We successfully loaded whole dataset and found that the train dataset contains 12 columns and 903653 rows. The test dataset has 12 columns and 804684 rows. After parsing the JSON format columns, we get 55 variables in total with 49 character variables, 5 numerical variables and 1 boolean variable. We also notice that there exist some variables that appear in the train dataset but lost in the test dataset. We may remove those variables when conduct the data processing. Our target have just 1.3% of non-null values; 6 columns with 97%+ of missing values; 4 columns with 50%+ of missing values; 1 column with 22.22%; 1 column with 0.004%.

REFERENCES

[1] https://www.kaggle.com/c/ga-customer-revenue-prediction

[2] https://en.wikipedia.org/wiki/Regression_analysis