FLIP oo PROJECT REPORT

Bing Liu[1]

[1] Jilin University, China
[2] Deakin University, Australia

Introduction

This is a demand forecasting problem. Participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bike share program in Washington, D.C. Data is provided as follow.

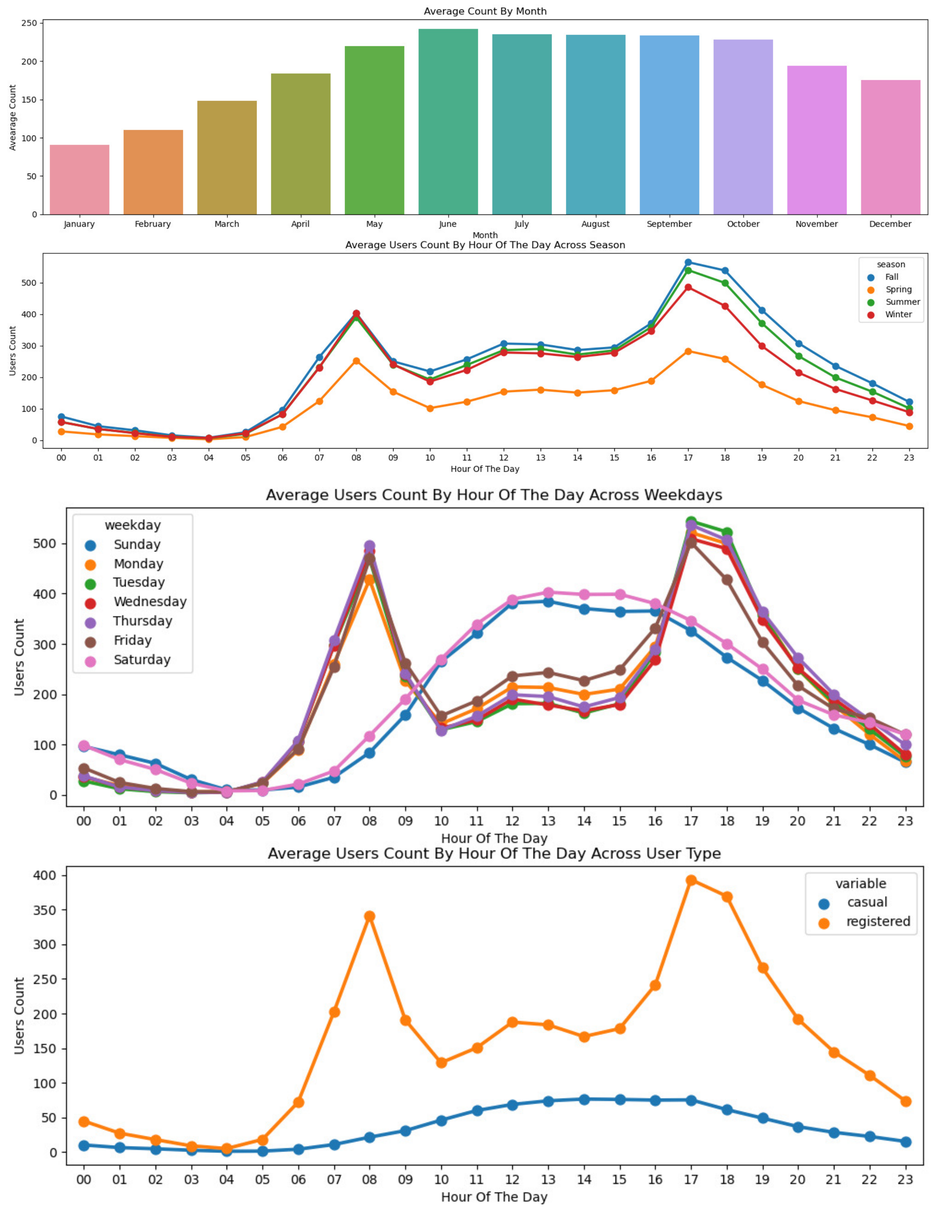
- **train.csv** – it contains 10886 rows and 12 columns. Each row represents bike rental data for a certain hour. Each column indicates the current conditions
- **test.csv** – it contains 6493 rows and 9 columns. Compared with the train data, there are fewer "casual","registered" and "count" columns.
- **sampleSubmission.csv** – it clarifies the data submission format. It just contains 2 columns that is "datetime" and "count".

Data processing

- **Missing values analysis** Use "missingno" to visualize missing value in the dataset.
- **Outliers analysis** Analyze the relationship between demand and date, season, hour of the day, and working day or not.
- **Correlation analysis** Analyze the correlation between count and temp, atemp, humidity and windspeed.

Data Visualization

Data visualization of the relationship between demand and date, season, hour of the day, and working day or not is follow, which show some information about data feature.

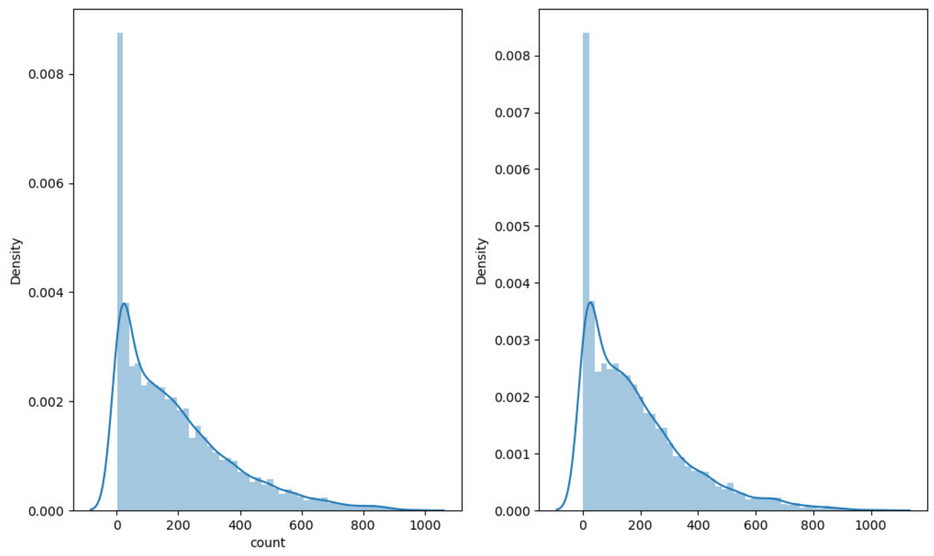


Feature Engineering

- **Time feature processing** Split the given date into "date, hour, year, weekday, month".
- **Feature selection** According to the data analysis results, select the more relevant features as the input data of the training model. The significant features contain temp, humidity, windspeed, atemp.
- **Training set and test set division** Divide train set and test set according to whether there is count attribute.

Model and Experiment

I have choose the Ensemble Model - Gradient Boost. Compare the distribution of train and test results.It confirms visually that the model has not predicted really bad and not suffering from major overfitting problem.



Conclusion

Loss function Using RMSLE to calculate the error, it penalizes under-prediction even more.

RMSLE Value RMSLE Value For Gradient Boost: 0.189973542608

Rank The score of my submission in kaggle is 0.41867. Ranked 428 among 3242 teams.

Acknowledgement

- Flip00 project report, 26/04/2021, changchun. China