

## **Literature Review**

Bike sharing system has appeared more and more on the street of the cities in order to meet the demand of public transportation in the last short distance to the destination. Also, Bike sharing system is so popular around all the world that most of major modern cities and campuses have been operated. Among researches for the bike system, predicting the demanding number of future bike shares is one of the most important and necessary tasks. During our project, the goal is aimed to forecast the amount of usage of bikes by a certain model.

Many recent studies have focused on the problem of prediction demand of shared bikes in the system. In order to complete prediction, the bike sharing system is supposed to satisfy the real-time and high accuracy requirements. Some researchers utilized Spark Machine Learning framework to predict the number of rental bikes to optimize the accuracy of model [1]. Firstly, the author collected three types of data including Citi Bike, Weather and Holiday, which the data have been SQL and outlier processed. Then, three predicative models including multiple linear regression, decision tree and random forests are constructed to analyze and train the processed data. During the experiment, applying to machine learning Spark ecosystem, the author used 70 % data as training data while 30% as the test data. In the result, after testing, random forests model has the lowest root mean error (RMSE). At last, the researcher optimized the result further by applying logarithmic optimization to the model.

Throughout the literatures, there is consistent evidence that forecast number of shared bikes are influenced not by holiday, weather, and temperature [2] in general, but by user information such as gender, birth year and user type [1]. The strengths of the literature are that the author consider more factors like user information and process useful and reasonable row data. Moreover, the optimization improves the accuracy the result. While there has been much research on algorithm optimization, few researchers have taken the influence of different factors and the real-time data into consideration [2]. The weakness of the literature is that the author ignored the timeliness of data and station selection. Furthermore, the mythology is limited to predict the number in the new area where lack of the existing shared bike number. Usually, station clustering and demand prediction in every station should be evaluated and analyzed as an integration [3].

In conclusion, it is challenging to collect the external factors in the future study because there are too many factors affecting users whether to use shared bike. However, we should consider multiple factors as much as possible to improve the accuracy. Comparing with our project's database, more factors are taken into consideration including time, number of new bike shares grouped by hour, temperature, humidity, wind speed, types of the weather, holiday, weekend, season [4]. In our project, we will just employ historical usage of bikes with some necessary factors to predict the number, which are very similar with the literature because it is easier to collect data and the predicted numbers are closer to the real.

## Reference

- [1] Z. Kang, Y. Zuo, Z. Huang, F. Zhou and P. Chen, "Research on the Forecast of Shared Bicycle Rental Demand Based on Spark Machine Learning Framework," 2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES), Anyang, 2017, pp. 219-222, doi: 10.1109/DCABES.2017.55.
  
- [2] F. Lin, S. Wang, J. Jiang, W. Fan and Y. Sun, "Predicting public bicycle rental number using multi-source data," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 1502-1509, doi: 10.1109/IJCNN.2017.7966030.
  
- [3] J. Huang, X. Wang and H. Sun, "Central Station Based Demand Prediction in a Bike Sharing System," 2019 20th IEEE International Conference on Mobile Data Management (MDM), Hong Kong, Hong Kong, 2019, pp. 346-348, doi: 10.1109/MDM.2019.00-38.
  
- [4] <https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset>