Seoul Bike Sharing Demand Prediction

Ziyu Liu, Kexin Guo, Miaojin Hu

1 Introduction

Bike-sharing is widely regarded as an environmental-friendly transportation option and is considered to be one of the remedies for both air pollution and traffic congestion(Zhang et al. 2021). However, increasing the utilization rate of shared bikes remains a challenge. Although many studies have focused on the impact of subjective factors on bike leasing(Kaplan et al. 2015), there has been limited research on objective factors. This study aims to explore these objective factors. Seoul was chosen as the research focus due to its data being highly representative as an international metropolis. Our analysis will encompass the rental patterns and factors influencing bike usage over a one-year period. This study will provide important guidance for the bike-sharing industry, aiding in the effective allocation of bicycles and the attraction of customers.

2 Data

2.1 Data Description

This data set comes from Kaggle: https://www.kaggle.com/datasets/joebeachcapital/seoul-bike-sharing/ data, which contains the number of bikes rented per hour every day in Seoul from 12/01/2017 to 11/30/2018, as well as weather information (Temperature, Humidity, Wind speed, Visibility, Dew point temperature, Solar radiation, Snowfall, Rainfall).

Table 1 provides a comprehensive overview of the descriptive information of the pre-cleaning data of 8760 hours. For continuous variables, we used mean (standard deviation) to describe; For categorical variables, we used frequency (percent) to describe.

2.2 Data Pre-processing

Several critical aspects of our data processing procedure needed to be emphasized. First is that we choose to use the number of rented bikes in multiple hours as the respond, and the rest variables will be considered as the predictors. And then we handle the NA values and outliers to avoid side effects on following model fitting. We also noticed the Functioning day column is a binary variable indicating the availability of shared bikes at specific time of a day, so we do filtering on this variable to exclude unfunctional bikes from our analysis. Additionally, correlation analysis between the predictors was done to avoid the possible effect of multicolliearity. We exclude Solar Radiation and Rainfall, Snowfall variables due to too many zeros. For the rainfall and snowfall, they also closely related with the humidity and temperature. And Dew point temperature and Visibility are removed because of their high-correlation with other covariates, for example temperature. In summary, our approach involves a systematic and rigorous treatment of the data set, ensuring that the selected sample is representative of the whole accessible rental bike population. The careful preprocessing steps and considerations for potential confounders underscore the commitment to methodological robustness and the pursuit of reliable inferential outcomes.

Characteristics	Overall (n=8760)
Rented bike count	729.16 (642.35)
Hour	11.51 (6.92)
Temperature	12.77 (12.10)
Humidity	58.15 (20.48)
Wind speed (ms)	1.73 (1.03)
Visibility (10m)	1433.87 (609.05)
Dew point temperature	3.94 (13.24)
Solar radiation ($MJ.m^2$)	0.57 (0.87)
Rainfall (mm)	0.15 (1.13)
Snow fall (cm)	0.08 (0.44)
Seasons	
Spring	2208 (25.21%)
Summer	2208 (25.21%)
Autumn	2184 (24.93%)
Winter	2160 (24.65%)
Holiday	
Yes	432 (4.93%)
No	8328 (95.07%)
Functioning day	
Yes	8465 (96.63%)
No	295 (3.37%)

Table 1: Baseline characteristics overall

2.3 Methods

2.3.1 Model used

Firstly, assuming a linear relationship between predictors and the expected count, we used linear regression to fit the model. Since the expected count is a non-negative integer variable, the linear regression may not be an ideal model for our situation. So, a Poisson generalized linear regression model was considered. When using a Poisson generalized linear regression model, overdispersion is a possible issue, which means that we need to check the mean and variance among the observations. For example, we used 9 am observations, whose mean is around 668 and variance is around 149,626. The variance is much larger than the mean, so a Poisson distribution may be inappropriate here. In this case, a negative binomial distribution might be a better choice.

Besides the parametric regression, we also consider some non-parametric methods. In the following part, we will use Random Forest(RF), Kernel Regression (KR) with multiple variables, and the K-Nearest Neighbors (KNN). All these models' complexities, underlying assumptions, and interpretabilities should be further considered.

2.3.2 Tune parameters

For the non-parametric models we mentioned above, tuning parameters is an essential part of the work that makes the model fit well. We separated the data into the training and testing data to find the optimal hyperparameter. We used the 10-fold cross-validation among the training data to find the parameter. KNN needs a hyperparameter K to determine how many neighbors are used to predict. The hyperparameter used in the RF is the number of trees that are used to make the final decision. KR's bandwidth to fit in the local will be determined using the least square cross-validation. We need to know how many times to restart finding the extrema of the least square cross-validation function from different initial points to optimize the

bandwidth. We will all use cross-validation to find the optimal value for the K, the number of trees, and the initial points. The metric we use is the mean square error (MSE), which will be used to tune the parameters.

2.3.3 Computation enhancement

In our project, we will predict the hour-specific rental counts, which means that we need to fit 24 times for each model. Besides that, cross-validation is also a time-consuming process. So, to conquer the time challenge we will meet, we used the multi-core parallel to complete the project. Firstly, we wrote a demo code with the five hyperparameters for RF and KNN from 0 a.m. to 5 a.m. We run the code both with and without using the parallel. The code without using parallel spent 17.18 seconds, and the code used the parallel spent 3.73 seconds. The parallel enhances the speed significantly. So, for the project, we used the parallel to conquer the computational challenge we met. The device we used is 12-core and 18G memory.

3 Results

3.1 Parameter selection

We used 15 hyperparameters for each nonparametric model. The candidate K used in KNN are 1, 6, 11, 16. . . , 66, 71. The candidate numbers of trees used in random forests are 100, 200, .., 1400, 1500. The candidate initial points used to find the kernel regression bandwidth are 1, 2,. . . , 14, 15. Figure 1 shows how models performed with different hyperparameters. We can see that KNN has the best performance when k = 11, the RF has the best performance when the number of trees is 210, and the kernel regression has the best performance when the number of initial points is 3. So, in the following model comparison, we will use those optimal values to fit in those models.

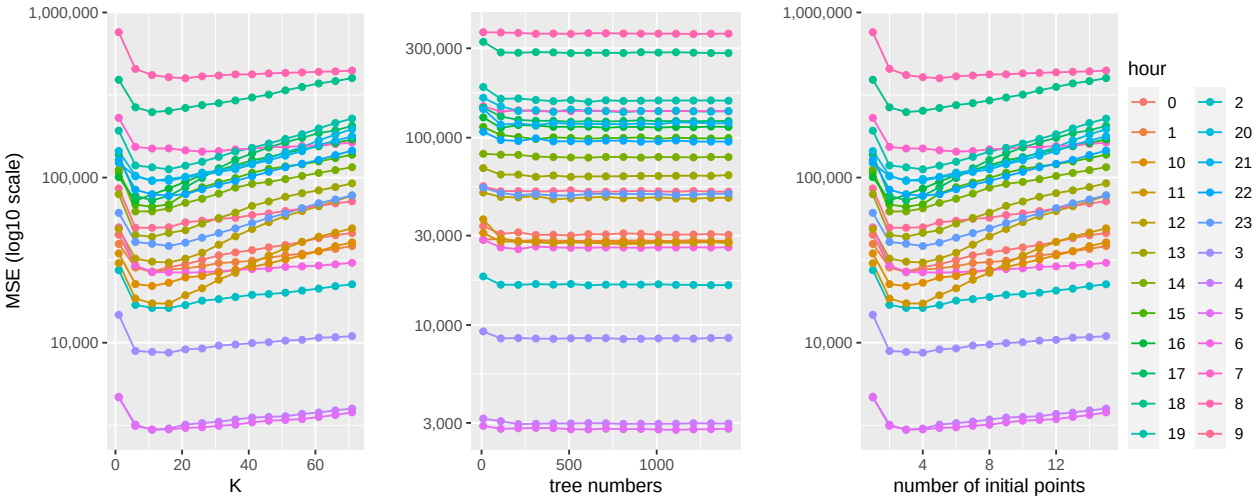

Figure 1: Tune the parameter for different model. The left one is the cross validation results used to tune the K neighbors used to fit in the KNN model, the middle one is the cross-validation used to tune number of trees used to fit the Random Forest, and the right one is the cross validation results used to tune the number of initial values to fit the kernel regression model

3.2 Model comparison

Figure 2 displays Mean Squared Error (MSE) on a log scale for different predictive models across different hours. Model performance varies throughout the day, reflecting changing data patterns such as variability

with the time of day. No single model consistently outperforms the others across all hours, but there are some general trends: The "Kern Regression" (Kernel Regression) and "RF" (Random Forest) models appear to perform similarly across the day, with MSE values that fluctuate but remain in the middle range compared to the other models. Lowest MSE occurs around hour 5, rising until hour 10, with a general increase towards day's end. No clear "best" model emerges, rather, the choice of model might depend on the specific hour of the day. For instance, at hour 5, "KNN" or "LM" might be preferred for their lower MSE. It's also important to note that while MSE is a useful metric for comparison, it doesn't capture all aspects of model performance, such as bias-variance tradeoff, model complexity, or interpretability.

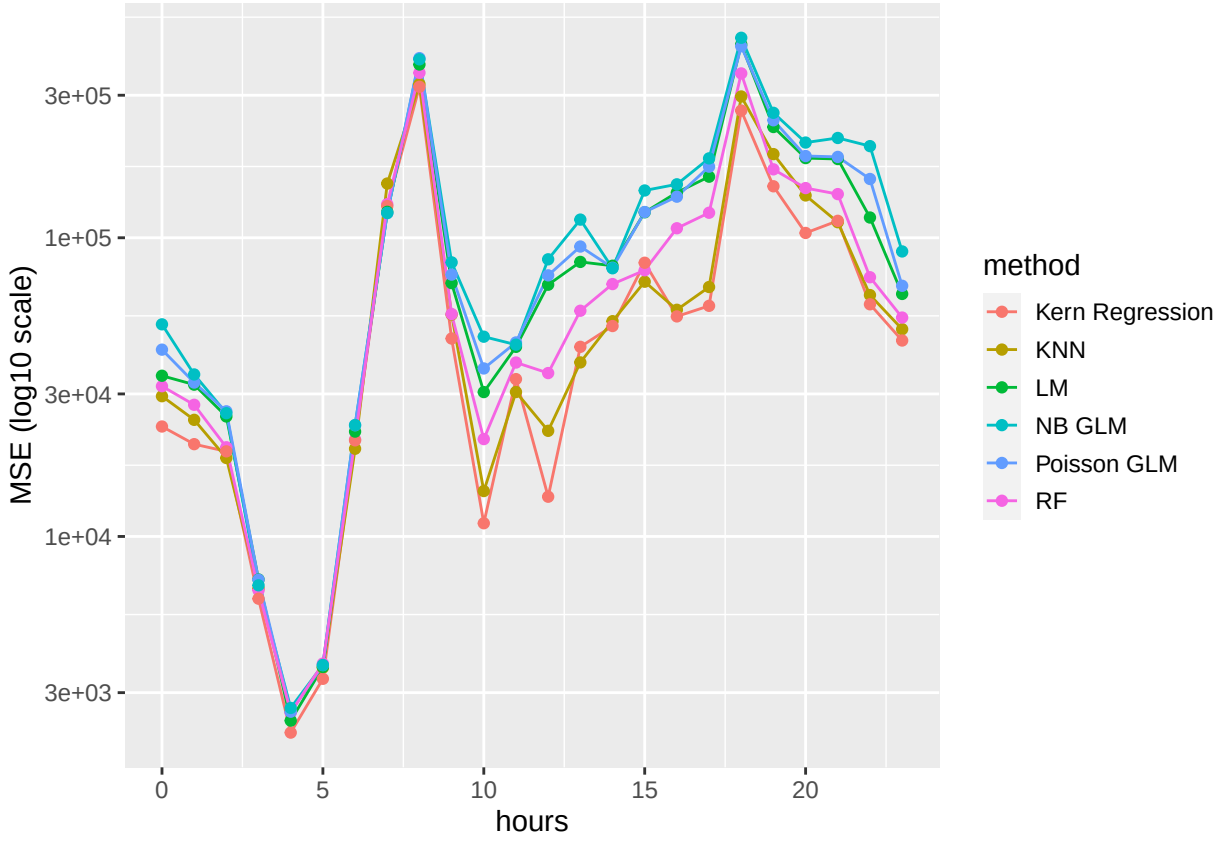

Figure 2: Different models' performance in different hours

3.3 Interpretation

Based on the results, we found that the KNN model and Kernel Regression model performed better. In this case, we may good to predict the results using those two models. However, as a common issue for the nonparametric model compared with the parametric models, the non-parametric model is hard to interpret. So, we decided to chose the linear regression model, which also performs well, to interpret. In this paper, we will use 5 a.m. as an example to see how the linear regression fits the data in Table 2; different hours may have different coefficients. The linear regression model showed that temperature, humidity, wind speed, seasons, and Holidays are the five main factors affecting bicycle rentals($p<.0.5$, respectively). Specifically, temperature has a significant positive effect on the number of bicycles rented, meaning that the number of bicycles rented typically increases as the temperature rises. Conversely, humidity generally has a negative effect on the number of bicycles rented, meaning that high humidity may reduce people's willingness to rent bicycles. Wind speed also has an effect on bike rentals at certain times of the year. In addition, seasons and holidays can also have a significant impact on bicycle rentals.

	Estimate	Pr(>|t|)
(Intercept)	185.05	9.255908e-19
Temperature	3.62	1.732835e-10
Humidity	-1.08	1.480361e-09
Wind speed	-11.71	2.907788e-03
Spring	-43.36	3.734985e-07
Summer	41.91	1.211326e-04
Winter	-64.24	1.069984e-07
No holiday	30.11	2.729474e-02

Table 2: This table shows 5am linear regression coefficients and p values

4 Discussion

In conclusion, these models reveal how multiple factors influence sharing-bike rentals. For example, days with warmer temperatures, lower humidity, and moderate wind speeds may have more bike rental activity, especially on non-holiday days and during certain specific seasons. This information can be very useful for city planners and bike share service providers to better understand and forecast bike rental demand.

Our program has many significant advantages: first, unlike many studies that focus only on peak hours, our study covers 24-hour bike rentals, providing a more comprehensive guidance to the bike-sharing service industry; Second, we used parallel programming techniques, which drastically reduced the running time of the program, improving the overall efficiency.

Nevertheless, our study has some limitations. In particular, the mean square errors of all six models were relatively large, suggesting that these all six models did not fit very well. Therefore, in future studies, we plan to explore more different models to improve the predictive power and accuracy.

Our current code running time using the local 12-core 18G memory device is 2.15 mins. In our quest to enhance the performance of our model, we have always been attempted to leverage parallel computation by utilizing more cores, with a notable example being the integration of up to 36 cores in a server named Greatlakes. However, we faced with a setback as the server consistently operated crowded at high capacity, causing delays in the queue. Despite this, we acknowledge the potential benefits of employing more cores for parallel computation, and remain committed to exploring the Greatlakes server in the future, once it becomes more accessible and less congested.

Another challenge we encountered is the interpretation of results from non-parametric models. Given the lack of interpretable parameters, non-parametric models reach results through patterns and relationships in the data, typically relying on distance metrics or similarities between data points, making it difficult to understand the influence of specific features on the model's predictions. For future improvement strategies, feature importance analysis can be employed. By assessing the contribution of each feature to the model's predictions, insights into which variables are more influential can thus be gained.

5 References

Kaplan, Sigal, Francesco Manca, Thomas Alexander Sick Nielsen, and Carlo Giacomo Prato. 2015. "Intentions to Use Bike-Sharing for Holiday Cycling: An Application of the Theory of Planned Behavior." *Tourism Management* 47: 34–46. https://doi.org/10.1016/j.tourman.2014.08.017.

Zhang, X, J Wang, X Long, and W Li. 2021. "Understanding the Intention to Use Bike-Sharing System: A Case Study in Xi'an, China." *PLoS One* 16 (12): e0258790. https://doi.org/10.1371/journal.pone.0258790.

6 Contribution

Ziyu Liu: code of all non-parametric model, model comparison and parallel, data visualization, report writing, github; Kexin Guo: code of lm model, report writing, create table; Miaojin Hu: code of glm model, create table report writing

7 Github

https://github.com/ziyuliu1999/biostat625_final/tree/main