Prediction Analysis for Seoul Bike Sharing Demand

Kai Wang

Ryerson University

Abstract

Bike renting service is one of the important components in the modern city for establishing a living and comfortable environment. In this report, the Seoul Bike Sharing Demand Data set is used to be analyzed to determine the most influential factors on rented bike counts. This dataset demonstrated the rented bike counts under different weather conditions and time intervals. The prediction models for the rented bike counts can help to improve the provided service of rented bikes for the rented bike company. The regression models, multiple linear regression, decision tree regression, and KNN regression, were built using R language. Through the multiple linear regression model, the stepwise regression was applied to find the best model depending on their AIC scores. All three regression models were applied 10-fold cross-validation to enhance the performance. Before modeling, the data inspection processes such as correlation among the independent variables, variances, missing data, et cetera were checked through data visualization. In the end, the RMSE, MAE and R-Squared were the metrics evaluation for all the model to compare their performance.

Introduction

Bike sharing plays an important role all over the world, especially in urban cities. A bicycle, as a transportation tool, is essential to be a solution to global climate issues comparing to other transportations. It is used to reduce carbon emissions and overcome traffic jams because citizens can ride on a bike instead of vehicles. As a consequence, more and more bike sharing services have been springing up in modern cities. Bike sharing systems are the systems that citizens can rent bicycles for daily use inside cities. The rental companies installed bike stations all over the cities to satisfy the demand. The prediction analysis for the bike sharing demand becomes a crucial step in order to improve the provided service of rented bikes.

In this report, the prediction analysis was demonstrated as regression models using the Seoul Bike Sharing Demand Data set. The dataset recorded the rental counts of the sharing bike under various weather conditions and time intervals. The multiple linear regression, decision tree regression, and KNN regression models were established to study the most influential factors on rented bike counts and predict the patterns how the factors affecting the rental counts.

Literature Review

Along with the growth of bike rental services all over the world, multiple studies have been completed to provide excellent prediction models in different urban cities. There are some studies involving the same data as mine and similar methods. Both Seoul bike rental and Capital Bikeshare program datasets were analyzed to establish five different statistical models, Cubist model, regularized random forest, regression trees, K-nearest neighbors, and conditional inference tree (Sathishkumar and Yongyun, 2020). They found out that the hour, temperature, and humidity are the most effective variables on the rental counts. Another approach to deal with bike sharing dataset was a low-dimensional model (Guido et al, 2020). This study applied a hierarchical procedure to spatial clustering and temporal clustering on the data. Then, they aggregated and decomposed two clusters to build the models and gain over 75% accuracy. Furthermore, the learning methods such as long short-term memory neural networks were applied on the bike sharing dataset instead of the traditional algorithm. It built dynamic and visualized models on the map of the New York cities (Youru et al, 2019). One of the studies created a web application with a multi-agent system based on bike share system in Salamanca data (Álvaro et al, 2017). This system make integration on different types of bike-sharing data in order to provide the forecast and visualization. They achieved the best model by Random Forest Regressor. To dig into the bike-sharing problem, bike rental counts are not the only dependent variable needed to be considered. The condition of the returning bike should be also considered included in the mode so that the operational efficiencies can be improved (Weiguo et al, 2020). They used Spatial-Temporal Dynamic Interval Network to analyze the time and location of the renting and returning.

Approach

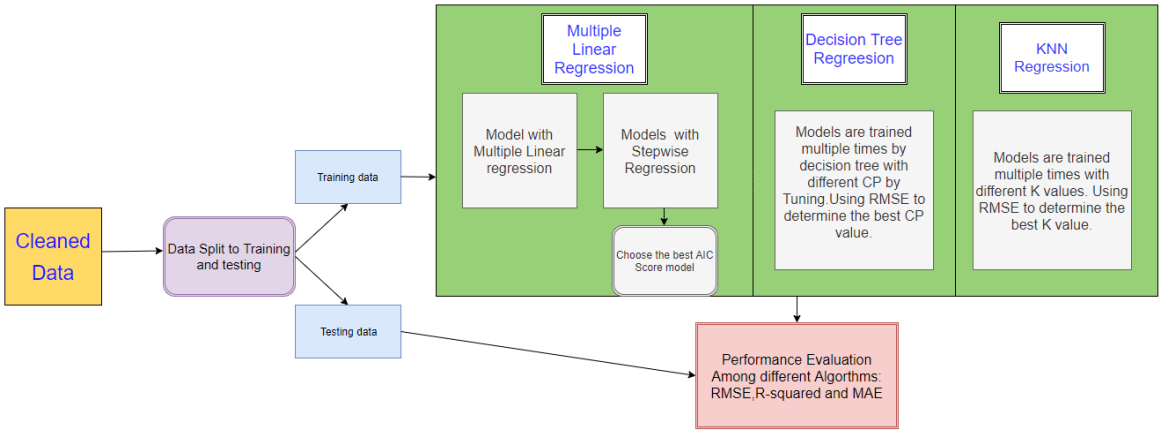

Figure 1 An overview of modeling procedure.

Train-test split dataset

Since the Seoul Bike Sharing Demand Data has been cleaned on the last part, the first step was to split the cleaned data to training data and testing data. The training data was used to create models with different algorithms. The testing data was used to examine the performances of the models. The separation was performed randomly. Testing data was 25% and training data was 75% of the original data.

Multiple Linear Regression

Linear regression is a statistical technique that is used to indicate the linear relationship between an independent variable and a dependent variable (James et al, 2013 , p. 71). However, the separate linear regression models for different independent variables towards one dependent variable were not satisfactory. Therefore, multiple linear regression can indicate the linear relationship of several independent variables towards one dependent variable at one time.

The form of multiple linear regression for n independent variables is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Y is the dependent variable.

X_n are independent variables.

β_0 is y-intercept.

β_n are slope coefficients for X_n.

ϵ are residuals.

k-fold cross-validation was used to get the best performances (James et al, 2013, p. 176). It resampled the training data for k times. The k was taken as 10. Root mean squared error was the root squared of means of the square of residuals between the observed values and the predicted values (Neill et al, 2018). It was one of the common measurements of the models. When the RMSEs are smaller, the models are better.

The formula of root mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

N is the number of the data.

$Predicted_i$ is the predicted value from the model.

$Actual_i$ is the actual value from testing data.

Then, the algorithm applied to improve the performance of the multiple linear regression was stepwise regression. Stepwise regression included backward elimination and forward selection (James et al, 2013, p. 207). Forward selection started from a model without any independent variables. Then, added independent variables one at a time. Backward elimination was reversed. It began from the model with all the independent variables and removes independent variables one at a time. Stepwise regression combined these two methods to have the best performance. During the stepwise regression, the Akaike information criterion (AIC) was used to evaluate their performance in every step. AIC is a mathematical method to determine if the model was improved comparing to the last model (James et al, 2013, p. 212). The lower AIC values were, the better the model was. When the model was added or dropped an independent variable, the AIC of the model would be either increased or decreased. If the AIC was increased, the execution was remained. If the AIC was decreased, the execution was deleted and moved on to next independent variable. Therefore, the model with the least AIC value was selected for the next steps. 10-fold cross-validation was also included.

Decision Tree Regression

A decision tree was a tree-like model to predict and demonstrate the factors that affected the dependent variable. A decision tree was a non-parametric algorithm classification (James et al, 2013, p. 303). Decision tree regression was a regression with a decision tree strategy to predict the dependent variable which was continuous numerical. The major factor that influenced the performance of the decision trees was their cost complexity (James et al, 2013, p. 309). Complexity parameters in the decision tree were used to determine the size of the tree in order to gain an ideal tree model. The tuning grid was specified manually to find the best complexity parameters for the model. In every tuning, 10-fold cross-validation was used to achieve the best performances.

The best model was chosen with the best CP for the next steps. The best CP was selected according to the RMSE when applying the model in testing data.

KNN Regression

K-nearest neighbors (KNN) was a non-parametric algorithm (James et al, 2013, p. 39). This algorithm classified the data depending on how its neighbor was classified. Also, KNN can be used both in regression and classification (James et al, 2013, p. 105). The dependent values in KNN regression were taken by computing the average of nearest k-neighbor groups from the training data. K in KNN was used to determine the numbers of the neighbors. The tuning length was set to 10. In every tuning, 10-fold cross-validation was used to achieve the best performances.

The best model was chosen with the best K value for the next steps. The best value was measured according to the RMSE, which was similar to the decision tree process.

Performance Evaluation Comparison for Three Prediction Models Using Testing Data

Multiple linear regression, decision tree regression, and KNN regression have been applied to establish the best three models on their own. Their corresponding RMSEs, MAEs and Coefficients of determination(R-squared) were calculated and compared to decide the best performance among them. The mean absolute error (MAE) was similar to RMSE but instead of square the difference, it took the absolute and there was no square root (Masters, 2018). When the MAEs are smaller, the models are better.

The formula of root mean squared error:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

N is the number of the data.

y_i is the predicted value from the model.

x_i is the actual value from testing data.

Coefficients of determination(R-squared) also was a commonly used metrics evaluation for regression model (Masters, 2018). It indicated better models if their values were close to 1.

The formula of R-squared:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

y_i is the actual value from testing data.

$\hat{y}_i$ is the predicted value from the model.

$\bar{y}$ is the mean of the actual value from testing data.

The Figure 1 showed an overview of the process visually.

Dataset

The data are downloaded from UCI Machine Learning Repository. The dataset is named Seoul Bike Sharing Demand Data Set[1]. It was used to make prediction models for the dependent variable 'rental bike count'. The data recorded the rented bike from December 1st, 2017 to November 30th, 2018. There are 8760 observations and 14 variables with no missing values.

Rental bike counts indicated the count of the bikes being rented within one-hour intervals. Table 1 gave an overview of the attributes. At first, when I went through all the attributes, I have noticed the attribute called functioning days. Functioning day means whether the bike rental services are available. I compared the number of observations in non-functioning days and the number of observations with 0 rental bike

[1] UCI Machine Learning Repository: Seoul Bike Sharing Demand Data Set. https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand.

counts. Both of them are 295 observations. Then, I filtered all the observations with non-functioning days and counted the number of observations with rental bike counts. As a result, they are still the same. Therefore, I concluded that all the observations with 0 rental bike count are only at the hour when there is no bike rental service in these. As a consequence, the non-functioning days' observations are dropped because the 0 bike rental count only happened when the services were not functioning. Furthermore, the variable 'function day' was not in the model since all of them were functioning after dropping 0 bike rental count observations.

Seoul Bike Sharing Demand Data Set

Variable Names	types	Example
Date	Character	"01/12/2017" "01/12/2017" "01/12/2017" "01/12/2017" …
rental_count	Interger	254 204 173 107 78 100 181 460 930 490 …
Hour	Interger	0 1 2 3 4 5 6 7 8 9 ...
Temperature	Numerical	-5.2 -5.5 -6 -6.2 -6 -6.4 -6.6 -7.4 -7.6 -6.5 ...
Humidity	Interger	37 38 39 40 36 37 35 38 37 27 …
Wind_speed	Numerical	2.2 0.8 1 0.9 2.3 1.5 1.3 0.9 1.1 0.5 …
Visibility_in_10m	Interger	2000 2000 2000 2000 2000 2000 2000 2000 2000 1928 …
Dew_point_temperature	Numerical	-17.6 -17.6 -17.7 -17.6 -18.6 -18.7 -19.5 -19.3 -19.8- 22.4 ...
Solar_Radiation	Numerical	0 0 0 0 0 0 0 0 0.01 0.23 …
Rainfall_in_mm	Numerical	0 0 0 0 0 0 0 0 0 0 …
Snowfall_in_cm	Numerical	0 0 0 0 0 0 0 0 0 0 ...
Seasons	Character	"Winter" "Winter" "Winter" "Winter" …
Holiday	Character	"No Holiday" "No Holiday" "No Holiday" "No Holiday" …
Functioning_Day	Character	"Yes" "Yes" "Yes" "Yes" …

Table 1 Seoul Bike data attributes and their data types and examples

The dependent variable 'rental bike count' has a minimum of 2 counts, a maximum of 3556 counts, a median of 542 counts, the first quartile of 214 and the third quartile of 1084. From the Figure 2, it had right-skewed distribution, and a few outliers but did not need to be dropped.

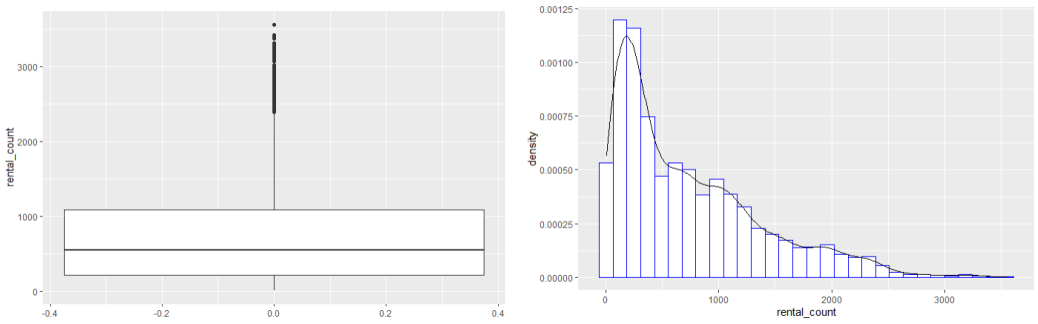

Figure 2 histogram with density distribution and Boxplot for rental bike count in Seoul Bike Sharing Demand Data

The attribute ‘hour’ is the integer from 0 to 23. From the Figure 3, the busiest hours are around 18 o’clock and 8 o’clock, which matched the rush hours. The attribute ‘seasons’ has winter, autumn, spring, and summer. According to the Figure 4, the rental counts in summer are highest and in winter are lowest. The attribute ‘Holiday’ means whether it is a holiday. According to the Figure 5, the rental counts in non-holidays are higher than holidays. The rest of the attributes are weather conditions. The attribute ‘temperature’ is from – 17.8 Celsius degree to 39.4 Celsius degree. The attribute ‘humidity’ is from 0% to 98%. The attribute ‘wind speed’ is from 0m/s to 7.4m/s. The attribute ‘visibility’ is 270m to 20000m. The attribute ‘dew temperature’ is from -30.6 Celsius degree to 27.2 Celsius degree. The attribute ‘solar radiation’ is from 0 MJ/m2 to 3.42 MJ/m2. The attribute ‘rainfall’ is from 0mm to 35mm. The attribute ‘snowfall’ is from 0cm to 8.8cm. All the outliers for the variables have been checked and it was not necessary to drop any observations.

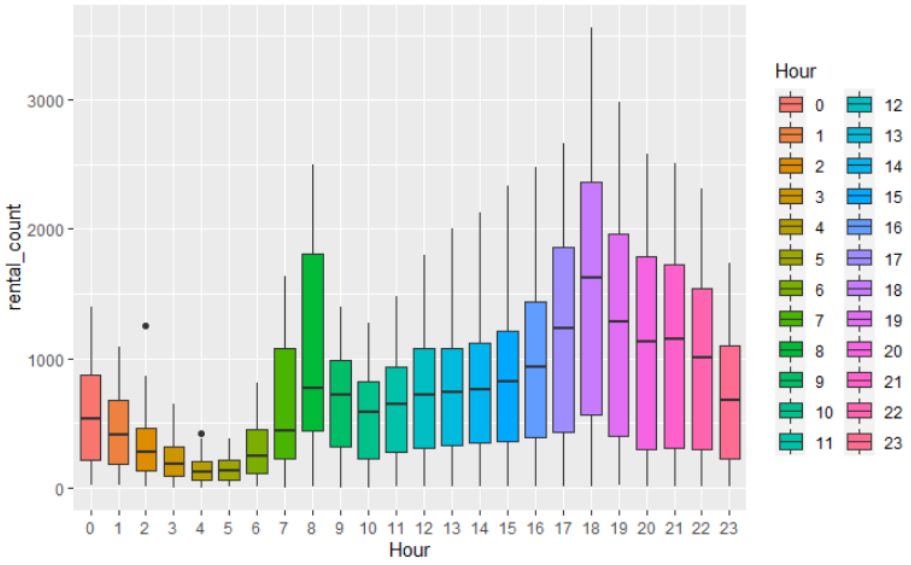

Figure 3 Boxplot for different hour intervals toward bike rental counts in Seoul Bike Sharing Demand Data

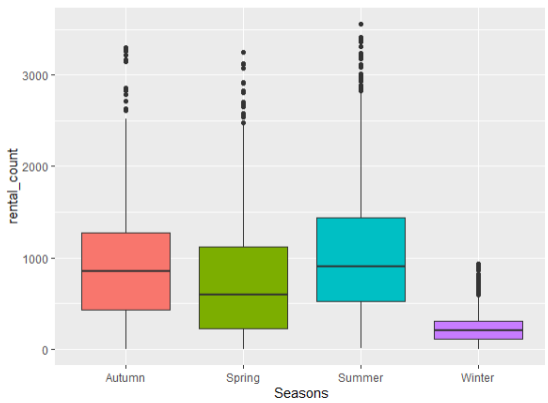

Figure 4 Boxplot for different seasons verse bike rental counts in Seoul Bike Sharing Demand Data

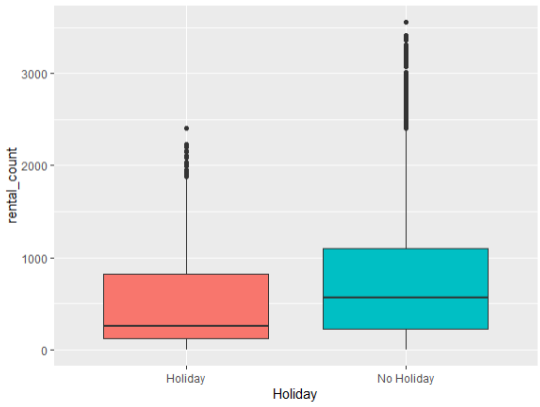

Figure 5 Boxplot for holiday or non-holiday verse bike rental counts in Seoul Bike Sharing Demand Data

After examining a correlation test between the independent variables, the dew temperature and temperature were highly correlated according to Figure 6. Their correlation coefficient was 0.91. Therefore, the dew temperature variable was dropped when building the multiple linear regression model because the temperature variable contained and included the information of the dew temperature variable. Meanwhile, the preliminary analysis indicated the temperature with correlation coefficient of 0.56 and hours with correlation coefficient of 0.43 had the most significant impact on bike rental counts according to Figure 6.

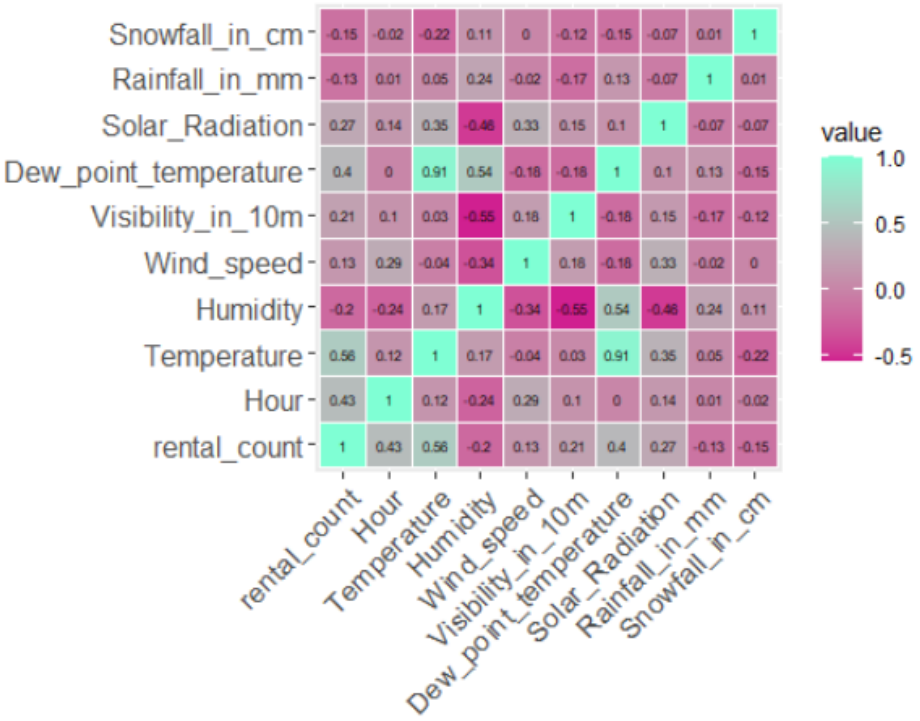

Figure 6 Correlation test heatmap among variables in Seoul Bike Sharing Demand Data

Analysis and Results

Multiple linear regression and Stepwise Regression

The multiple linear regression model was demonstrated in Figure 7. 10-fold cross-validation had applied. Most of the attributes had high significance except snowfall and visibility. The p-value of snowfall was 0.093. It was not significant under a significance level of 0.05. The visibility had the same conclusion. It had only a 0.45 p-value. This result concluded that visibility and snowfall had no significant or only slightly affecting on rental counts. The formula of this regression model was round to the second decimal point and was following:

$$\begin{aligned} Rental\ count = {} & 57.35 + 28.71Hour + 26.61Temperature - 8.43Humidity \\ & + 19.22Wind\ speed - 0.0076Visibility - 79.69Solar\ radiation \\ & - 63.43\ Rainfall + 18.82Snowfall + 110.79Seasons \\ & + 128.08Holiday \end{aligned}$$

Applied the model above with testing data, it achieved 438.215 as its RMSE, 0.535 as its R-squared, and 326.996 as its MAE, displayed in table 3.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        57.345482  55.244293   1.038 0.29928
Hour               28.706993   0.746413  38.460  < 2e-1
Temperature        26.609709   0.579858  45.890  < 2e-1
Humidity           -8.431415   0.372911 -22.610  < 2e-1
Wind_speed         19.218473   5.236120   3.670 0.00024
Visibility_in_10m  -0.007595   0.009984  -0.761 0.44685
Solar_Radiation   -79.693998   7.577910 -10.517  < 2e-1
Rainfall_in_mm    -63.431446   4.375804 -14.496  < 2e-1
Snowfall_in_cm     18.817067  11.200192   1.680 0.09298
Seasons           110.785998   5.623107  19.702  < 2e-1
Holiday           128.077284  22.298547   5.744 9.58e-0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 437.8 on 8454 degrees of freed
Multiple R-squared:  0.536,     Adjusted R-squared:  0.5
F-statistic: 976.5 on 10 and 8454 DF,  p-value: < 2.2e-1
```

Figure 7 multiple linear regression model result

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        39.0364    49.7244   0.785 0.432443
Hour               28.7498     0.7443  38.628  < 2e-16 ***
Temperature        26.5632     0.5766  46.068  < 2e-16 ***
Humidity           -8.2726     0.3090 -26.773  < 2e-16 ***
Wind_speed         18.8501     5.2135   3.616 0.000301 ***
Solar_Radiation   -78.3414     7.3662 -10.635  < 2e-16 ***
Rainfall_in_mm    -63.2961     4.3721 -14.477  < 2e-16 ***
Snowfall_in_cm     18.9046    11.1993   1.688 0.091446 .
Seasons           109.9105     5.5039  19.969  < 2e-16 ***
Holiday           128.2228    22.2972   5.751  9.2e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 437.8 on 8455 degrees of freedom
Multiple R-squared:  0.536,     Adjusted R-squared:  0.5355
F-statistic:  1085 on 9 and 8455 DF,  p-value: < 2.2e-16
```

Figure 8 multiple linear regression model result applied stepwise regression

A stepwise regression model was expected to enhance the performances of the multiple linear regression model. 10-fold cross-validation had applied. From Figure 8, the model was similar to the multiple linear regression model. Only the visibility was dropped during the stepwise process. The formula of this regression model was round to the second decimal point and was following:

$$\begin{aligned} Rental\ count = {} & 39.04 + 28.75Hour + 26.56Temperature - 8.27Humidity \\ & + 18.85Wind\ speed - 78.34Solar\ radiation - 63.30\ Rainfall \\ & + 18.90Snowfall + 109.91Seasons + 128.22Holiday \end{aligned}$$

Applied the model above with testing data, it achieved 438.212 as its RMSE, 0.535 as its R-squared, and 326.862 as its MAE, displayed in table 3.

The smaller RMSEs were, the better models were. According to the RMSE metric, the stepwise model had increased 0.000685% performance compared to the original multiple linear regression. The tiny difference was explained by only dropping one non-significant attribute 'visibility'. In short, the stepwise process did not improve much for the multiple linear regression model because most attributes were essential in the model.

	1	2	3	4	5	6	7	8
Hour		*	*	*	*	*	*	*
Temperature	*	*	*	*	*	*	*	*
Humidity			*	*	*	*	*	*
Wind_speed								*
Visibility_in_10m								
Solar_Radiation						*	*	*
Rainfall_in_mm					*	*	*	*
Snowfall_in_cm								
Seasons				*	*	*	*	*
Holiday							*	*

Table 2 best combination subsets and the importance of the attribute.

In Table 2, * denotes the variables included in the subset. For example, the best combination of 2 attributes was Hour and Temperature. The best combination of the 3 attributes was Hour, Temperature, and Humidity. The more * the attributes had, the more important the attributes were. The independent attributes temperature had the greatest influence on the bike rental count. Then, the independent attributes hours and humidity were next. This result matched the correlation test in Figure 6. The independent attributes temperature and hours that had the highest correlations were the effective variables on the rental counts.

Decision Tree Regression

Figure 7 was used to find the best complexity parameter (CP). The tuning parameters were set as 10. There are 10 CPs across the figure. As we can see from Figure 7, the RMSEs were increasing as the CPs increased. Therefore, the best CP was the smallest CP. In this case, the CP was set in 0.00954.

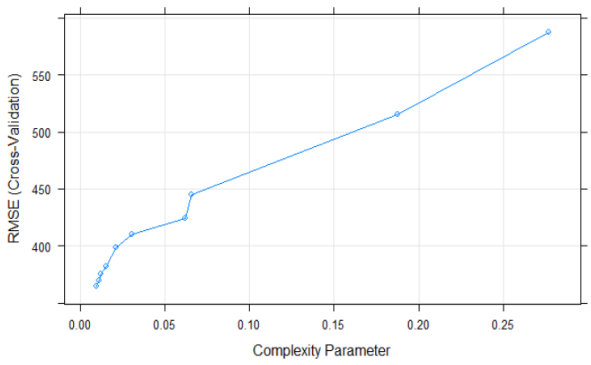

Figure 7 complexity parameter results depending on RMSE for decision tree regression.

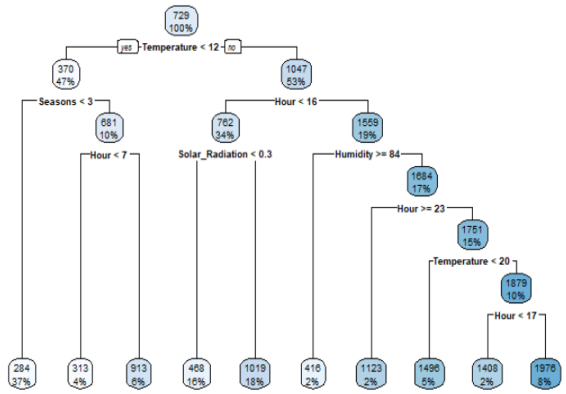

Figure 8 The tree model of the decision tree regression

Then, Figure 8 displayed the tree model with 0.00954 CP. 10-fold cross-validation had applied. As we can see, the attributes Hour and Temperature were shown the most frequently in this tree model, which indicated those two attributes were the most significant ones. Applied test data on the decision tree regression model, achieved 364.577 as its RMSE, 0.679 as its R-squared, and 259.350 as its MAE, displayed in table 3.

KNN Regression

Figure 9 was used to find the best number of the neighbors. The tune length was set at 10. There are 10 neighbors across the figure. As we can see from the Figure 9, the RMSEs were decreasing first then increasing. Therefore, the lowest point was the best number of the neighbors. It was 7 with lowest RMSE. 10-fold cross-validation had applied.

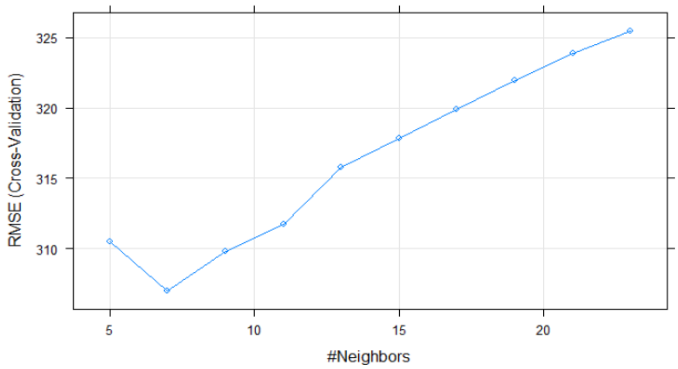

Figure 9 The number of the neighbors results depending on RMSE for KNN regression.

After applying the KNN model to test data, it achieved 297.135 as its RMSE,0.786 as its R-squared, and 194.320 as its MAE, displayed in table 3

Analysis Among Models

	RMSE	Rsquared	MAE
Multiple linear regression	438.215	0.535	326.966
Stepwise Regression	438.212	0.535	326.862
Decision Tree Regression	364.577	0.679	259.350
KNN Regression	297.135	0.786	194.320

Table 3 The comparison among models along the evaluation metrics, RMSE, R-squared and MAE.

Table 3 represented the best model that had been built was the KNN regression model. It has the lowest RMSE value, which was 297.135, the lowest MAE value, which was 194.320, and the highest R-squared value, which was 0.786. After KNN regression, decision tree regression was the second-best model. The multiple linear regression model and stepwise regression model had high similarity since their only difference was the attributes 'Visibility'. According to their RMSE, the performance of the KNN regression model was 22.70% better than decision tree regression and 47.48% better than the multiple linear regression model. According to their MAE, the performance of the KNN regression model was 33.47% better than decision tree regression and 68.21% better than the multiple linear regression model. In short, the KNN regression model had the least error and significantly better performance compared to other models. The R-squared of KNN regression was larger than 0.7, which indicated a strong effect regression model. The multiple linear regression model with 0.535 R-squared and

Decision Tree Regression with 0.679 R-squared indicated moderate effect regression models.

Conclusion

This study focused on the prediction model for Seoul bike sharing demand. The results showed KNN regression algorithm had the best performance compared to decision tree regression and multiple linear regression. Analysis of the important attributes concludes that temperature was the top-ranked independent variable as the most influential variable on the rental counts. After the rank of temperature, the independent variable hour and humidity were the next ranks.

The best model, which was KNN regression, had 0.786 R-squared. Though this R-squared was generally satisfied, it can still have improvement of its performance with other deep learning algorithms. Therefore, other deep learning methods were suggested to further model establishment. Another direction of studying can be the location analysis for the bike distribution in cities. Figure 3 indicted the high demand during the rush hours. The location analysis can determine the best place during rush hours. Also, figure 4 expressed clearly that there was much less demand during the wintertime. The bike rental company may consider taking some bikes back to storage during the wintertime due to weather conditions and less demand.

Reference

Cantelmo, G., Kucharski, R., & Antoniou, C. (2020). Low-Dimensional Model for Bike-Sharing Demand Forecasting that Explicitly Accounts for Weather Data. *Transportation Research Record: Journal of the Transportation Research Board, 2674*(8), 132-144. doi:10.1177/0361198120932160

E, S. V., & Cho, Y. (2020). A rule-based model for Seoul Bike sharing demand prediction using weather data. *European Journal of Remote Sensing, 53*(Sup1), 166-183. doi:10.1080/22797254.2020.1725789

Li, Y., Zhu, Z., Kong, D., Xu, M., & Zhao, Y. (2019). Learning Heterogeneous Spatial-Temporal Representation for Bike-Sharing Demand Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 1004-1011. doi:10.1609/aaai.v33i01.33011004

Lozano, Á, Paz, J. D., González, G. V., Iglesia, D., & Bajo, J. (2018). Multi-Agent System for Demand Prediction and Trip Visualization in Bike Sharing Systems. *Applied Sciences, 8*(1), 67. doi:10.3390/app8010067

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: with applications in R.* Springer.

Masters, T. (2018). *Assessing and Improving Prediction and Classification: Theory and Algorithms in C++*. Apress.

Neill, S. P., & Hashemi, M. R. (2018). *Fundamentals of ocean renewable energy: generating electricity from the sea.* Academic Press, an imprint of Elsevier.

Pian, W., Wu, Y., & Kou, Z. (2021). STDI-Net: Spatial-Temporal Network with Dynamic Interval Mapping for Bike Sharing Demand Prediction. *From Data to Models and Back Lecture Notes in Computer Science,* 38-53. doi:10.1007/978-3-030-70650-0_3