****

Principles of Statistical Modeling Project

Spring 2022

**Data analysis for Seoul bike-sharing service**

**Author:** *Ali Afsharian* **Instructor:** *Prof. Dr. Stefan Kettemann*

**Abstract**

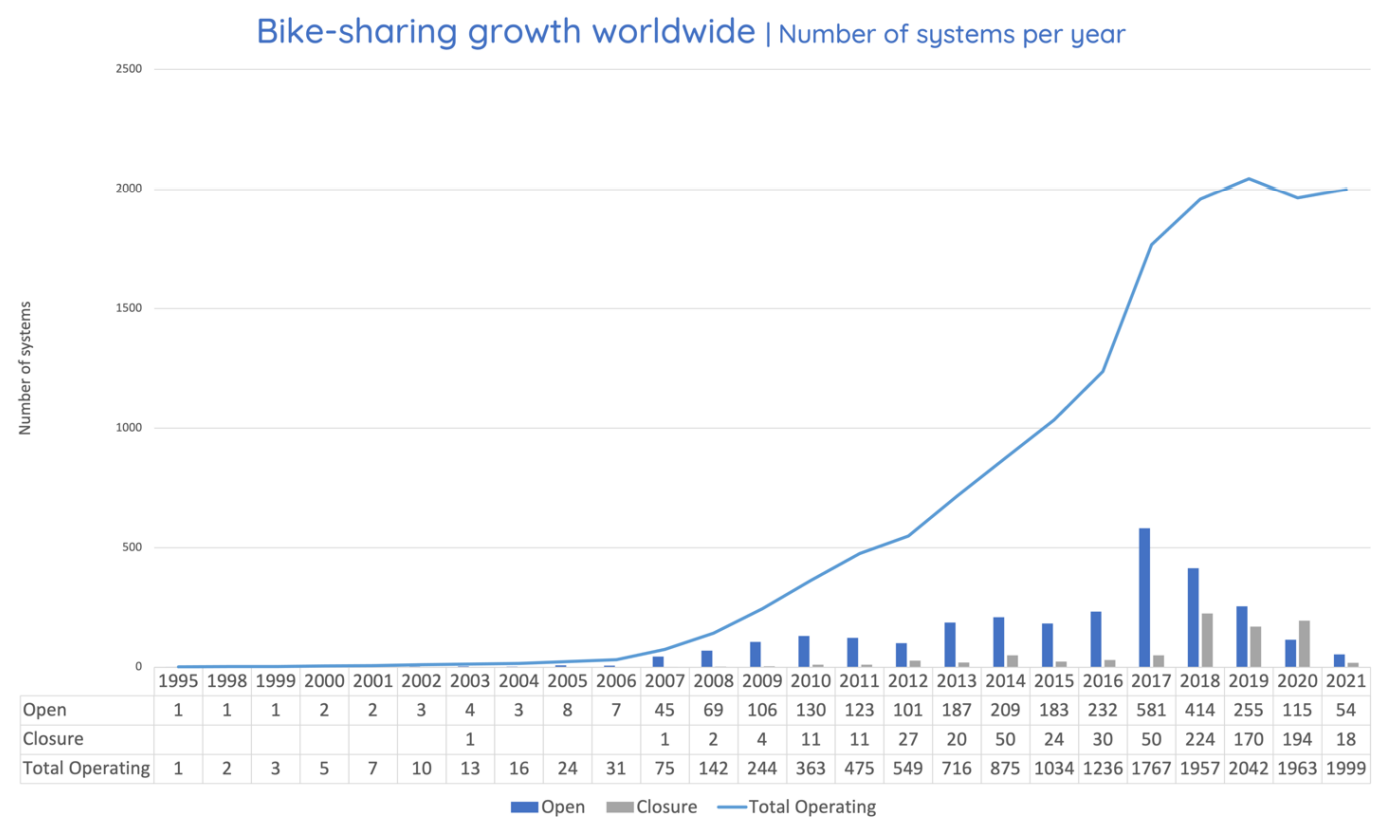
Today, rental bikes are offered in a wide range of urban cities to enhance comfortableness and environmental sustainability. Availability and accessibility of rental bikes to the public are two important aspects of a successful bike-sharing scheme. In order to have a stable supply of rental bikes and reduce waiting time, it is essential to make a prediction for future rental bikes demand. This project first discusses an exploratory data analysis for hourly rental bike demand in Seoul based on different predictors such as weather information and time, and explores the distribution characteristic of temperature and wind speed data. Moreover, two statistical regression model, linear regression and gradient boosting machine, were developed and their performances were evaluated with three evaluation indices on training and test sets. Features importance is explained for both methods and cross validation method is used to find the best hyperparameters for gradient boosting algorithm. The gradient boosting machine with optimized parameter can give the best results with value of 0.95 on the training set and 0.87 on the test set. All codes are given at the end of the report.

**Introduction**

Bike-sharing is a concept originally from revolutionary 1960s. It was initially grow slowly, but the development in technology and bike tracking systems accelerate this progression [1]. According to the Meddin Bike-Sharing World Map, there were more than 10 million bikes shared in diverse kinds of schemes by August 2021 [2]. Although, our data belongs to the year between 2017 and 2018, it is still interesting to look at the recent report. Fig. 1 shows the growth of bike sharing systems from 1995 until 2021. It can be seen that the number of systems increase dramatically in recent years. Although because of the COVID-19 pandemic, the number of system closure exceeded the number of launches in 2020, the number of opening have again grown in 2021. We know that this report is limited to the open bike-sharing systems in the world, however, ranking systems offer an interesting overview.

A bike-sharing system provides bike rentals which are returned automatically at a network of kiosks located throughout a city. In these systems, people are able to rent a bike from one place and at the end of their trip return the bike to the different spot. Considering global attention to the climate issues, introducing bike-sharing services receives increasing attention in the recent decades. These services provide numerous benefits, such as promoting cycling, reducing greenhouse gas emissions, improving public health and reducing traffic congestion. [?]

With the growing number of users, it is important to have an accurate prediction of available bikes in the stations in order to make the system function consistently. Different studies have been done to predict the bike usage using Spatio-temporal data and weather data. Sathishkumar et al. [3] presents five statistical models for bike sharing demand prediction based on weather data on two different data set. In another study, Sathishkumar et al. [4] compares the performances of different algorithms in demand forecasting of Seoul bike sharing. A short-term prediction for a case in Suzhou, China using recurrent neural networks (RNNs) and Random Forest methods is proposed by Wang et al. [5]. This report provides a data analysis on a dataset contains the rented bicycles in the Seoul bike sharing system and weather information, and investigates two predictive models for demand forecasting.



**Fig. 1.** Number of bike-sharing systems worldwide per year [2]

**Data description:**

Originally, rented bike data for one year (2017 December to November 2018) comes from the Seoul Public Data Park website of South Korea, where the hourly public rental history of Seoul bikes is available [6]. The dataset used in this report downloaded from the UCI Machine Learning Repository [7]. It contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information. Dataset contains 8760 rows and 14 columns in total. Table 1 lists all features in the dataset with corresponding data type and data value space.

Moreover, the dataset has been already cleaned without any missing values which means no further steps for data cleaning required. Table 2 describes the dataset and missing values for each feature.

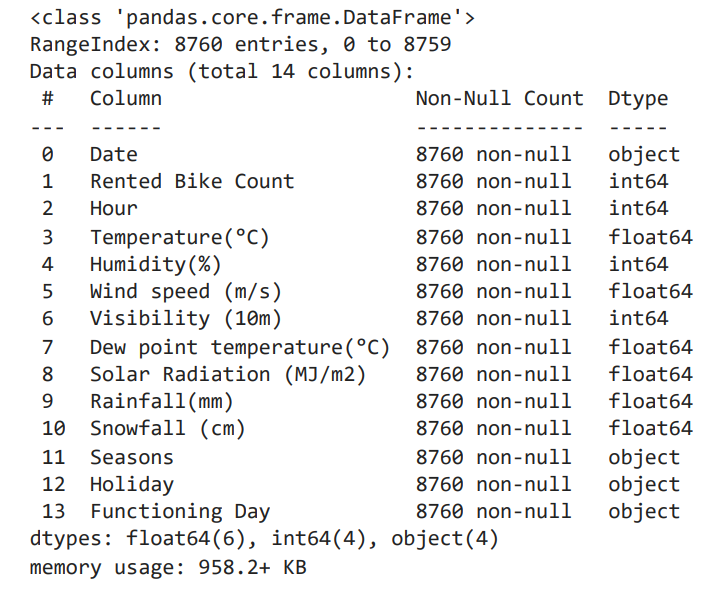
Mathematical formalism of the dataset is described in the following.

* **Universe Ω:** Universe consists the Seoul city, sharing bikes, bike sharing system provider, and collection of all atmospheric developments that start which is compatible with the available physical measurements
* **Elementary Events *ω*:** Each atmospheric development and any situation where someone use a sharing bike is an observation opportunity.
* **Measurable Function (RV-function):** Procedure of measuring climate data and sharing bike usage data as well as corresponding date.
* **Action of actually measuring Data:** Action of actually carrying out this procedure and collecting the number of bikes and weather data.
* **Data Value Space S:** The data value space would be the collection of weather information and the number of used bikes as well as corresponding date. Data value space achieved from the Eq. (1).

(1)

|  |  |  |
| --- | --- | --- |
| Parameters | Type | Data Value Space |
| Date | day-month-year |  |
| Rented Bike | Integer |  |
| Hour | Integer |  |
| Temperature | Continuous |  |
| Humidity | Continuous |  |
| Wind Speed | Continuous |  |
| Visibility | Continuous |  |
| Dew point temperature | Continuous |  |
| Solar radiation | Continuous |  |
| Rainfall | Continuous |  |
| Snowfall | Continuous |  |
| Season | Categorical |  |
| Holiday | Categorical |  |
| Functional Day | Categorical |  |

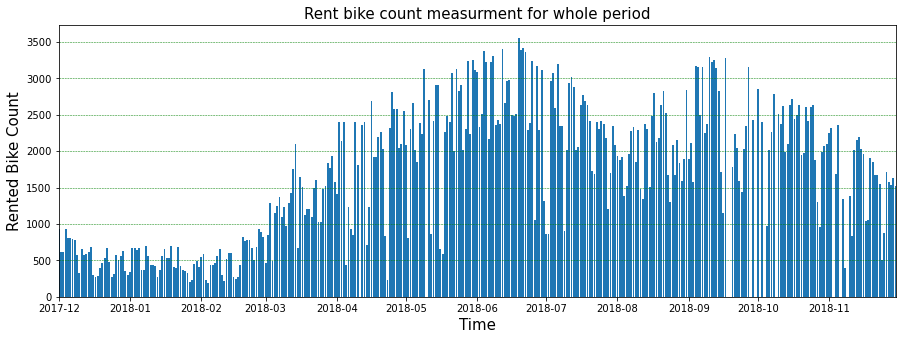
**Table 1.** Data variables and description



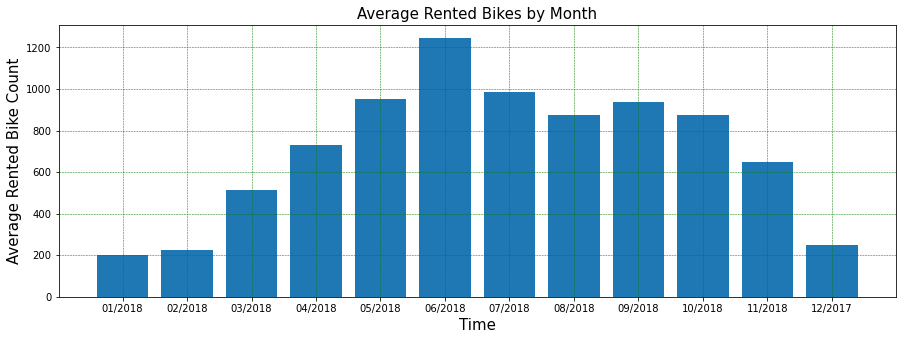
**Table 2.** A summary of the dataset and missing values

**Data Exploration:**

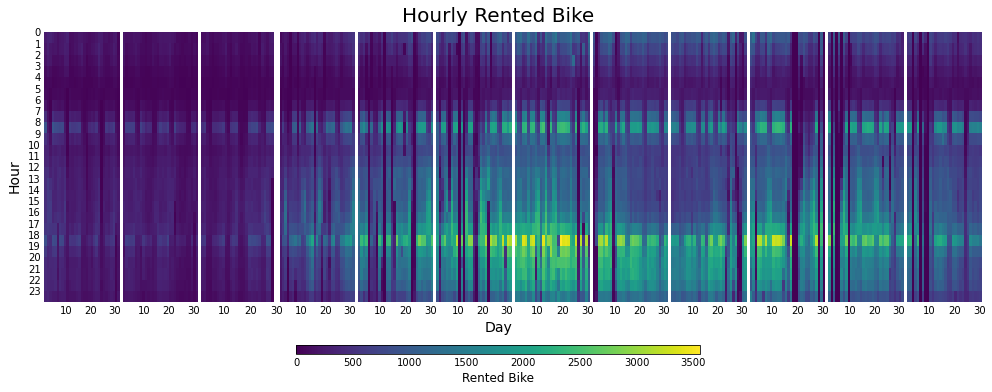
Fig. 2 displays the number of rented bikes in the whole period. According to the figure, rental bike counts vary dramatically from hour to hour. To have a better understanding, I plotted the average count for each month in Fig. 3, and the heat map for hourly rented bike during the whole period from December 2017 until November 2018 in Fig. 4. From the results, it can be seen that the average demand for shared bikes is high during summer season and less during the winter season. By looking at heat map, we can also see a pattern related to hours of the day and the rental bikes use. This indicates that the usage increases at around 8AM and 6PM which can be related to the peak hours in Seoul. The histogram of rental bikes is shown in Fig. 5, and we can that there is a large tail at the right of the diagram which indicates that the data has right skewed distribution. Fig. 6 shows the box plot for the target value with lower whisker 0, upper whisker 2400, and the median value of 505. It is not necessary to remove outliers from the data because there are few of them. Totally there are 147 rows that contains rented bike count larger than 2400.



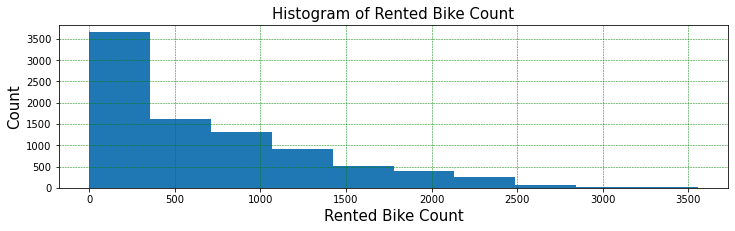
**Fig. 2.** Hourly rented bike count in the whole period



**Fig. 3.** Average rented bikes by month

****

**Fig. 4.** Heat Map for hourly rented bikes

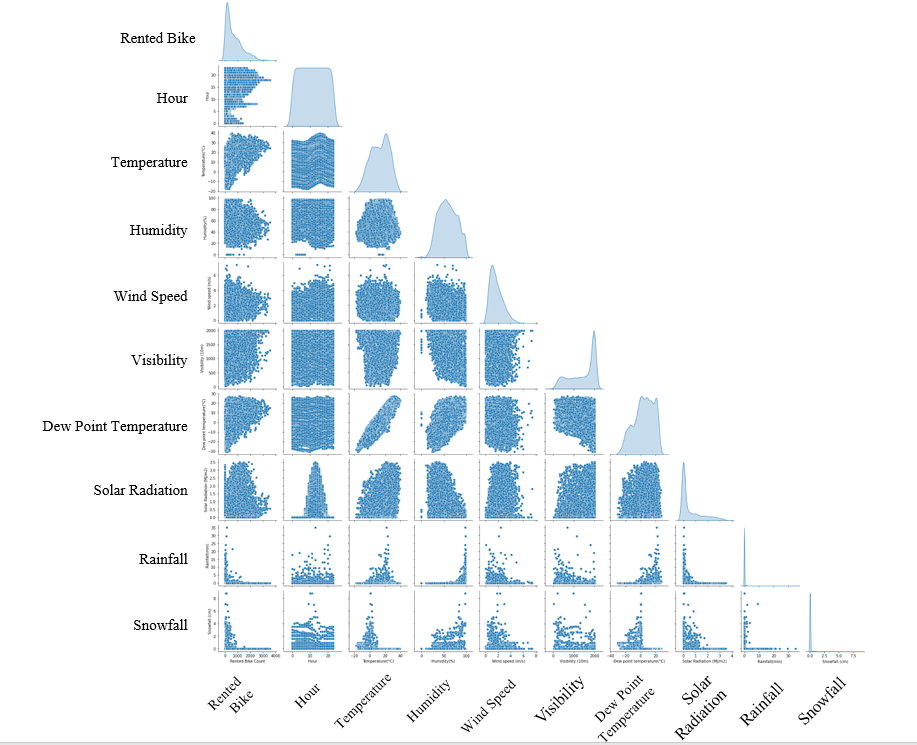


**Fig. 5.** Histogram of rented bikes



**Fig. 6.** Boxplot of the rented bikes

Fig. 7 displays the scatter plot for each pair of numerical features in the dataset and the distribution function in the diagonal. Considering scatter plots along with Fig. 8 which presents the Pearson correlation coefficient on a heat map, we can find that temperature feature has the highest correlation with the rented bike count (0.54). It simply means that the by increasing the temperature, the demand for rented bike will be raised. In the second position, the number of rented bike has a positive correlation with the hour (0.41). Moreover, humidity, rainfall and snowfall show negative correlation with bike count for -0.2, -0.12 and -0.14 respectively. We can conclude that these features influence the number of rented bike negatively but not significantly. Temperature and dew point temperature are features with highest correlation coefficient with 0.91. Their linear relation can also be seen in the scatter plot. There is a negligible negative correlation -0.036 between temperature and wind speed which means they are almost uncorrelated. I choose these two variable as an example to fit distribution functions and to find the joint pdf in the next part.

**Fig. 7.** Scatter plots of all numerical features

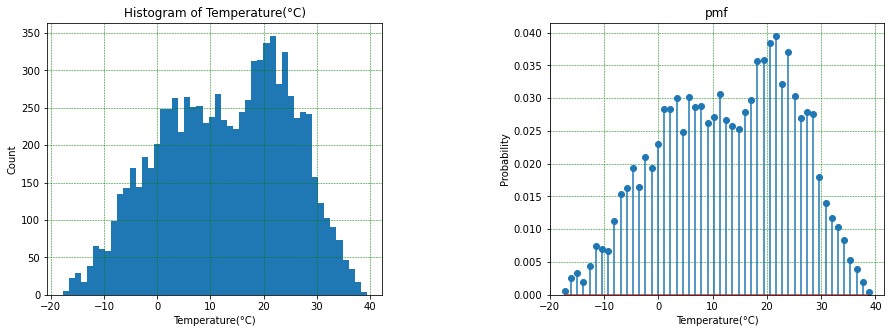


**Fig. 8.** Correlation heatmap

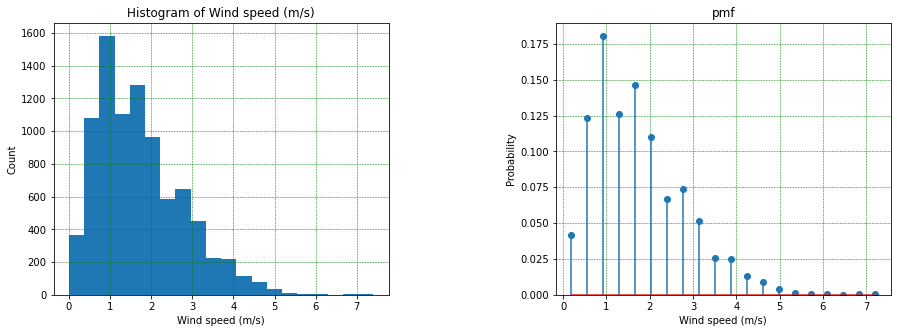
**Analyzing “Temperature” and “Wind Speed” features:**

In this part I choose temperature and wind speed data to study their probability mass functions, probability distribution functions as well as their joint probability. Fig 9 and 10 display the histogram and probability mass function for temperature and wind speed respectively. It can be seen from the histogram that the temperature data has two peaks, and it would be more accurate if we fit a multimodal normal distribution to the data. However, I chose the normal distribution with one mode to find and compare the properties such as mean and standard deviation. Moreover, the lower number of wind speed has higher probability mass function and the data is right skewed.

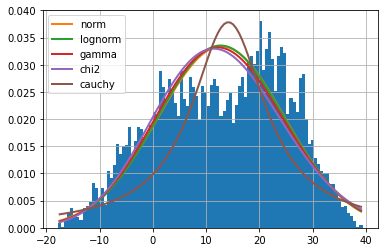
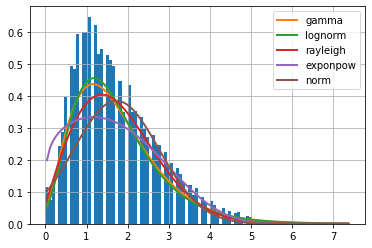
In the next step, I fit different kind of popular density function to these features. As it can be seen in Fig 11 and 12, the best estimations for temperature and wind speed density functions are normal function gamma function respectively. The comparison between the real data and fitted distribution is given in the table ?? and ??.



**Fig. 9.** (a) Histogram of Temperature (b) Probability mass function of Temperature data



**Fig. 10.** (a) Histogram of Wind Speed (b) Probability mass function of Wind Speed data

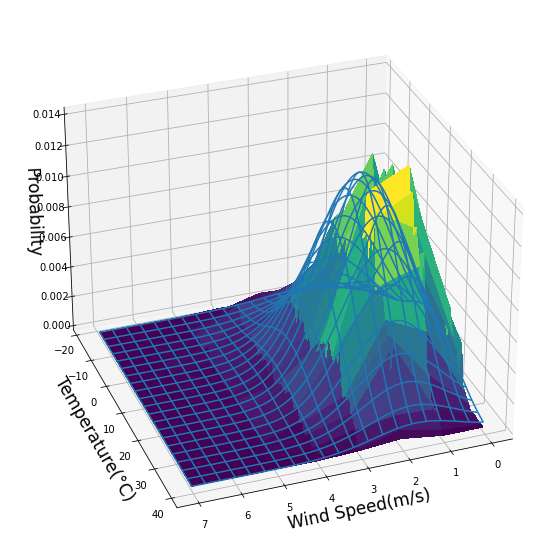
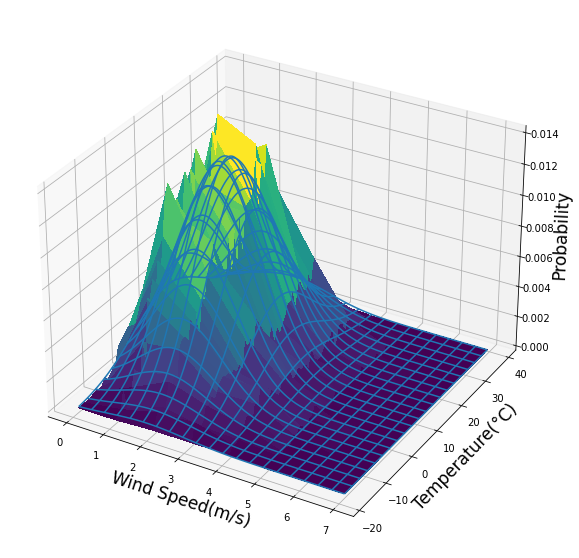


**Fig. 11.** Fitted distribution functions to the (a) Temperature data (b) Wind Speed data

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Temperature** | μ | σ | Skewness | Kurtosis |  | **Wind Speed** | μ | σ | Skewness | Kurtosis |
| Normal | 12.883 | 11.944 | 0 | 0 |  | Gamma | 1.725 | 1.05 | 1.085 | 1.766 |
| Real data | 12.883 | 11.944 | -0.883 | -0.198 |  | Real data | 1.725 | 1.036 | 0.726 | 0.891 |

**Table. ??.** Properties of distribution functions

For the next move, I round the measurements for temperature and wind speed and make discrete sets with bin size equal to 1 in order to find the joint probability. Calculated joint probability from Eq. ?? is shown in the Fig. 13 in two different views. Moreover, covariance matrix and means are calculated to find the multivariate Gaussian distribution (Eq. ? and ?). The derived multivariate Gaussian distribution is plotted in the Fig. 14 in the blue wire frame. We can conclude that the multivariate Gaussian could be a reasonable model for this joint probability qualitatively.

****

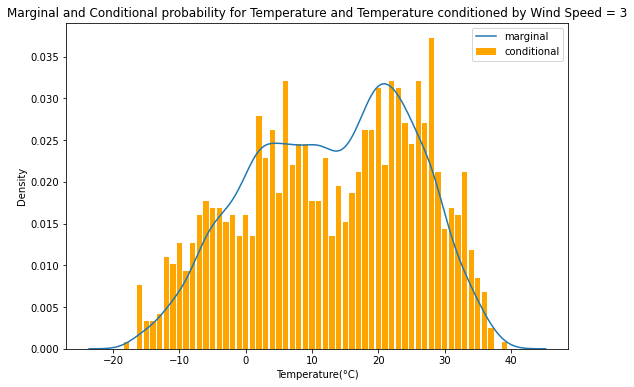
**Fig. 13.** The joint probability of Temperature -Wind Speed and Multivariate Gaussian distribution

(?)

**(?)**

Furthermore, I calculated the production of marginal distributions and compare with joint probability. The average percent error

By looking at Fig. ??, we can see a similar pattern and close values between the marginal probability for Temperature and conditional probability for Temperature conditioned by Wind Speed equals to 3. Since their joint probability is almost equal to the production of marginal probabilities, it can be concluded that Temperature and Wind Speed are independent variables.

****

**Evaluation Metrics:**

The evaluation metrics used in this report are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Rsquard (. RMSE is a standard deviation of prediction errors or residuals. It shows how far predictions fall from observed values using Euclidean distance. MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. R-squared values range from 0 to 1 and determines proportion of variance in the dependent variable that can be explained by the independent variables. The equations for calculating RMSE, Rsquard, and MAE are given in the following.

(?)

(?)

(?)

Here, is the actual measurement value, is the predicted value, is the average of the sample and is the sample size.

**Method Description:**

**1. Linear Regression:**

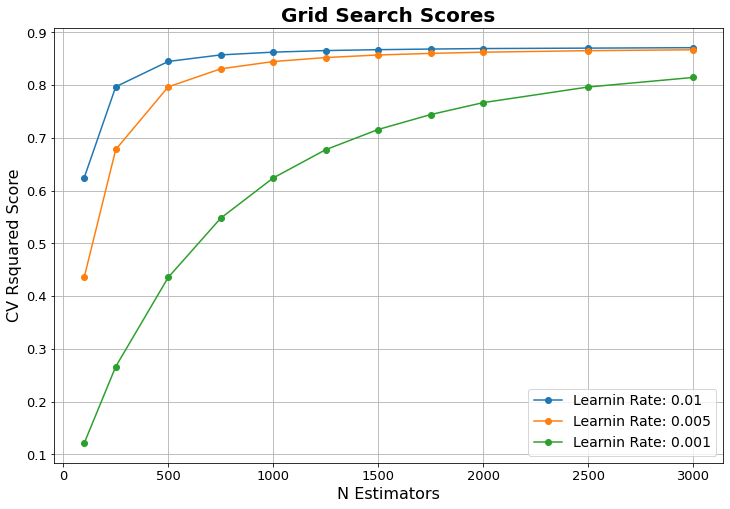
Linear regression is a conventional method for prediction problems. Linear regression models the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables). It is called multiple linear regression when more than one independent variable are considered. The basic mathematical model of linear regression is assumed as in Eq.??.

Here, to are unknown parameters in the model, known as the regression coefficient, and is the term of error.

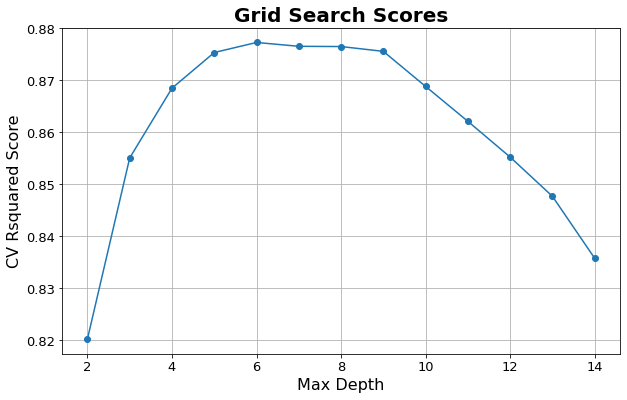
**2. Gradient Boosting Machine:**

**Model Development:**

In order to increase the accuracy of the prediction it is necessary to tune the hyperparameters of each regression model. The Scikit-learn library in Python provide a grid search to try all combination of possible parameter value and find the optimal parameters for a model. Hyperparameters for Gradient Boosting Machine includes the learning rate, number of estimators and maximum depth of the tree. The grid search runs for learning rate ranges from 0.01 to 0.001, number of estimators ranges from 100 to 3000, and maximum depth ranges from 2 to 14. As shown in Fig. 14 and Fig. 15, it turns out that the optimal values are 0.01, 1750, and 6 respectively.

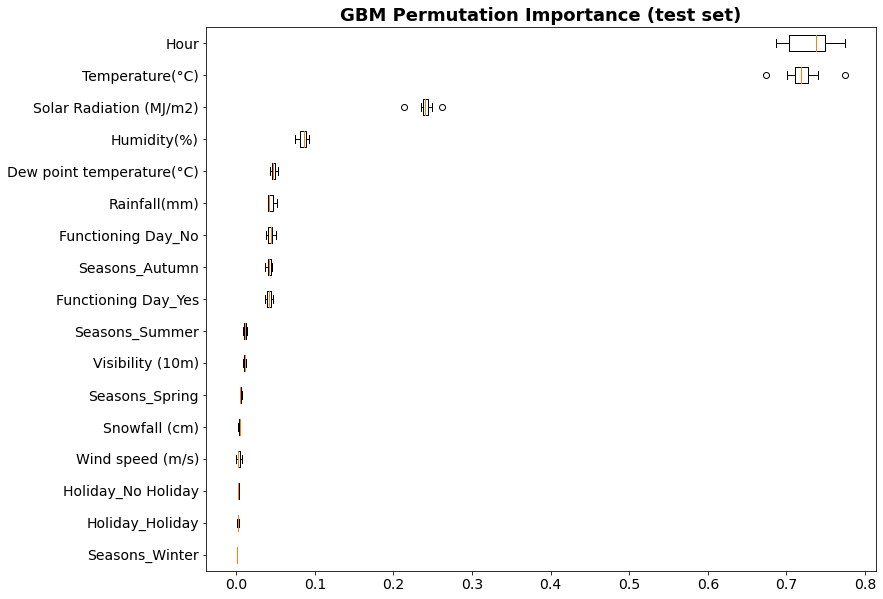
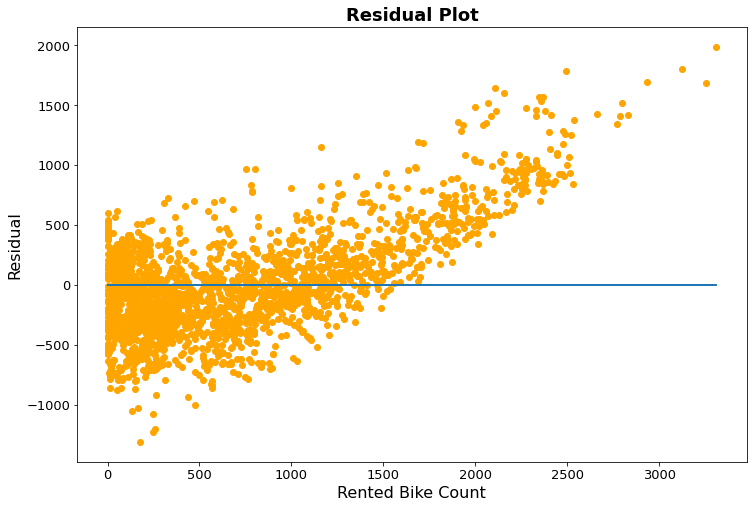
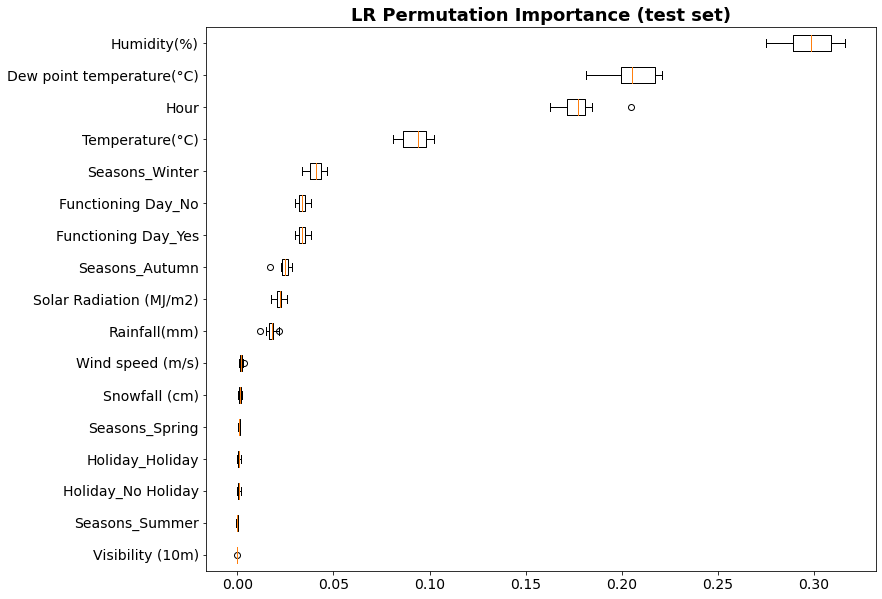


**Fig. 14.** Grid search results for GBM based on learning rate and number of estimators

****

**Fig. 15.** Grid search results for GBM based on maximum depth

**Results and Discussion:**

****

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training | | |  | Testing | | |
|  | RMSE | MAE |  |  | RMSE | MAE |
| LM | 0.55 | 434.04 | 322.40 |  | 0.56 | 428.15 | 323.97 |
| GBM | 0.95 | 137.97 | 88.92 |  | 0.87 | 227.48 | 139.81 |

**Table 4.** Models Performance

**Feature selection / dimensionality reduction techniques:**

**Conclusion and outlook:**

**References:**

[1] DeMaio, P. Bike-sharing: History, Impacts, Models of Provision, and Future. *Journal of Public Transportation* **12,** 41–56 (2009).

[2] "The Meddin Bike-sharing World Map." Russell Meddin, Paul DeMaio, [Oliver O’Brien](https://orcid.org/0000-0002-3413-0853), Renata Rabello, Chumin Yu, Jess Seamon, Thiago Benicchio. Accessed [05-05-2022]. <http://bikesharingworldmap.com/>

[3] V E, S. & Cho, Y. A rule-based model for Seoul Bike sharing demand prediction using weather data. European Journal of Remote Sensing 53, 166–183 (2020)

[4] E, S. V., Park, J. & Cho, Y. Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications 153, 353–366 (2020)

[5] Wang, B. & Kim, I. Short-term prediction for bike-sharing service using machine learning. in Transportation Research Procedia 34, 171–178 (Elsevier B.V., 2018)

[6] <http://data.seoul.go.kr/>

[7] https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

(Introduction), (main topic), (method description), (conclusions and outlook)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bike-Sharing Country Rank  (by number of active systems, mid-2021) | | | | | | |
| Rank | **Country** | **Systems** |  | **Rank** | **Country** | **Systems** |
| 1 | China | 673 |  | **26** | Mexico | 13 |
| 2 | United States | 174 |  | **27** | Australia | 13 |
| 3 | Germany | 107 |  | **28** | South Korea | 13 |
| 4 | Italy | 104 |  | **29** | Denmark | 13 |
| 5 | France | 70 |  | **30** | Norway | 11 |
| 6 | Spain | 67 |  | **31** | Slovenia | 9 |
| 7 | Poland | 63 |  | **32** | Ireland | 9 |
| 8 | United Kingdom | 40 |  | **33** | Argentina | 8 |
| 9 | India | 33 |  | **34** | Belarus | 8 |
| 10 | Japan | 33 |  | **35** | Slovakia | 8 |
| 11 | Switzerland | 29 |  | **36** | Hungary | 7 |
| 12 | Greece | 27 |  | **37** | Romania | 6 |
| 13 | Brazil | 27 |  | **38** | Belgium | 6 |
| 14 | Czech Republic | 27 |  | **39** | Ukraine | 6 |
| 15 | Taiwan | 24 |  | **40** | Liechtenstein | 6 |
| 16 | Austria | 20 |  | **41** | Israel | 5 |
| 17 | Netherlands | 19 |  | **42** | Indonesia | 5 |
| 18 | Finland | 17 |  | **43** | Kazakhstan | 5 |
| 19 | Sweden | 17 |  | **44** | Malaysia | 4 |
| 20 | Russia | 16 |  | **45** | Chile | 4 |
| 21 | Turkey | 16 |  | **46** | Bosnia and Herzegovina | 3 |
| 22 | Canada | 16 |  | **47** | New Zealand | 3 |
| 23 | Croatia | 15 |  | **48** | Estonia | 3 |
| 24 | Colombia | 15 |  | **49** | Cyprus | 3 |
| 25 | Portugal | 14 |  | **50** | United Arab Emirates | 3 |

**Table 1.** Rank countries with more active systems by mid-2021