**DEMAND FORECAST FOR BIKE SHARING IN METROPOLITAN CITY**

By

**Mohammed Saquib Hussain**

**Dr. Osama Abu Omer**

Masters project submitted in partial fulfillment of the

Requirements for the Master of Science in Data Science degree

In the College of Aviation, Science, and Technology of

## Lewis University

***Abstract***

This study aimed to forecast the demand for bike sharing services in the metropolitan city of Seoul using the Seoul bike sharing dataset. To do this, a variety of machine learning models were used to forecast demand for bike sharing, including regression analysis (linear, lasso, ridge, elastic net, and polynomial), k-nearest neighbors, random forest, gradient boosting, extreme gradient boosting, catboost, and light gradient boosting machine. The accuracy and efficiency of each model were assessed using a number of measures, such as mean squared error, root mean squared error, and mean absolute error. The dataset includes hourly demand data for bike-sharing services over a two-year period from 2017 to 2018, along with weather information and other relevant features such as temperature, humidity, wind speed, and visibility.

In order to find patterns and get relevant insights from the data, this study also used factorial analysis and clustering techniques like k-means and fuzzy c-means. The outcomes of these methods were put to use in order to comprehend the fundamental causes of Seoul's demand for bike sharing programs.

The outcomes of this study can aid in the decision-making process for legislators, transportation planners, and bike sharing service providers. These stakeholders can strengthen the efficiency and effectiveness of bike sharing programs, which will ultimately lead to higher utilization and improved accessibility for users, by accurately estimating the demand for bike sharing services.

**Index Terms –** Bike sharing, Factor analysis, Regression analysis, Root mean squared error, Weather, Catboost.

# I. INTRODUCTION

The use of private vehicles has significantly increased in metropolitan regions. However, this rise in usage has led to higher fuel consumption, which has had a negative impact on the climate. People in today's culture have also come to accept issues like traffic congestion as the standard as a result of this. Governments and organizations have begun implementing steps to promote sustainable development in order to address this issue.

The advent of bike-sharing programs in numerous nations is one such strategy. For instance, the Bike Share initiative was started in South Korea to address the issue of public transportation. Users are free to use the service alone, and it offers an affordable alternative form of transportation for short distances. A user of a bike-share program can borrow a bike from any bike station and return it to a station close to their destination. It also has positive health implications because it requires using the bike's pedals.

Additionally, the installation of bike stations throughout the city has increased the areas that are accessible by bikes. Computerized stands used to pick up and return rental bikes are known as docking stations. At any docking station, users of public bikes can hire and return rental bikes. On a dedicated website, they can also confirm the specifics of their journey, including the length and route, as well as assess their physical activity, including the number of calories burned. Rental bikes are more and more popular because of their clever technology and practicality.

But with more people using rental bikes, it's important to control the demand for them so that users can continue to receive a convenient service. This paper suggests a data mining-based method to estimate the demand for public bikes throughout the entire city using weather data in order to meet this need. To predict the quantity of rental bikes needed at each hour, a rule-based approach is utilized. This strategy aids in controlling the demand for bike rentals and offers users a continuous, convenient service.

# II. DISCUSSION OF RELATED WORK

The problem of demand forecasting for bike sharing programs has been the subject of extensive research in recent years. Regression analysis and several clustering algorithms have been used in many studies to accurately forecast the demand for bike sharing.

In one study, Yu and Cheng (2016) predicted the demand for bike sharing in the Chinese city of Hangzhou using a range of machine learning models, including random forest, support vector regression, and neural networks. The random forest model outperformed all others, according to the authors, with an R-squared value of 0.851. [1]

In a different study, Liu et al. (2018) forecasted the demand for bike sharing in New York City using a time-series technique. With a mean absolute percentage error of 6.05%, the scientists discovered that adding weather data considerably increased the accuracy of their model.[2]

In order to forecast the demand for bike sharing in Beijing, China, Peng et al.'s study from 2021 combined machine learning algorithms and clustering methods. The authors discovered that the random forest model gave the most precise predictions and that the k-means clustering approach was successful in detecting patterns in the data.[3]

Overall, the research points to the potential utility of machine learning methods, notably regression analysis and clustering algorithms, in estimating demand for bike sharing. These models' precision can be increased by incorporating variables like the weather, time of day, and location. The outcomes of this study are in line with prior findings, which show that gradient boosting models are particularly good at forecasting demand for bike sharing.

# III. PROJECT DESCRIPTION

**3.1 Problem Statement**

In many urban locations, bike sharing has grown in popularity. To keep the general public's mobility comfortable, it is crucial to maintain a consistent supply of rental bikes. The availability and accessibility of a constant supply of bikes to the general public at all times is a fundamental challenge for bike-sharing systems. Predicting the demand for rental bikes at various times of the day is one approach to be sure of this. Bike-sharing systems can reduce user wait times and boost customer satisfaction by adjusting their operations to match predicted demand by forecasting the demand for bikes.

This project's main goal is to develop a predictive model that can foresee the demand for rental bikes with accuracy at an early stage. This would make it possible for bike-sharing programs to provide bikes where and when they are needed, decreasing the possibility of shortages and guaranteeing a constant supply of bikes for the general public. The model will be based on a number of variables, such as the hour of the day, the day of the week, the weather, and other elements that could influence patterns of bike use.

Mean squared error (MSE) and R-squared are two performance metrics that will be used to gauge how accurate the model is. In order to pinpoint the primary elements affecting bike demand, the model's coefficients and feature importance will also be examined. This data can be utilized to improve the public's mobility comfort by helping bike-sharing systems operate more efficiently and guarantee a steady supply of rental bikes.[4]

**3.2 Dataset Description**

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Attribute Information:

* Date: year-month-day
* Rented Bike count - Count of bikes rented at each hour
* Hour - Hour of the day
* Temperature-Temperature in Celsius
* Humidity - %
* Windspeed - m/s
* Visibility - 10m
* Dew point temperature - Celsius
* Solar radiation - MJ/m2
* Rainfall - mm
* Snowfall - cm
* Seasons - Winter, Spring, Summer, Autumn
* Holiday - Holiday/No holiday
* Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

# IV. DETAILS OF METHODOLOGY

## **4.1 Data Collection**

I have used the Seoul Bike Sharing Dataset that was originally collected by the Seoul Metropolitan Government and is available for public use through the Seoul Open Data Plaza. The dataset is available in CSV format and can be downloaded for free.[5][6]

**4.2 Data Cleaning**

The first step in any data science effort is cleansing the data. No data is perfect, but most of it is helpful. Data cleaning refers to discovering incomplete, wrong, inaccurate, or irrelevant elements of the data and then replacing, changing, or deleting the unclean or coarse data. Data cleaning is the act of recognizing and repairing (or removing) corrupt or inaccurate records from a record set, table, or database. Before we start cleaning up the data, we make sure there aren't any duplicate values in the dataset. Following that, we convert datatypes, perform exploratory data analysis, and identify the dataset's best fit model.

**4.2.1** Examining Missing or Null Values

In deep learning and machine learning, null values are a major issue. Before you pass your data to the machine learning or deep learning framework, you must clean up null values if you're using sklearn, TensorFlow, or any other machine learning or deep learning packages. Otherwise, you'll receive a lengthy and unpleasant error message. Therefore, we are looking for missing or null values. The dataset provided contains no null or missing values.

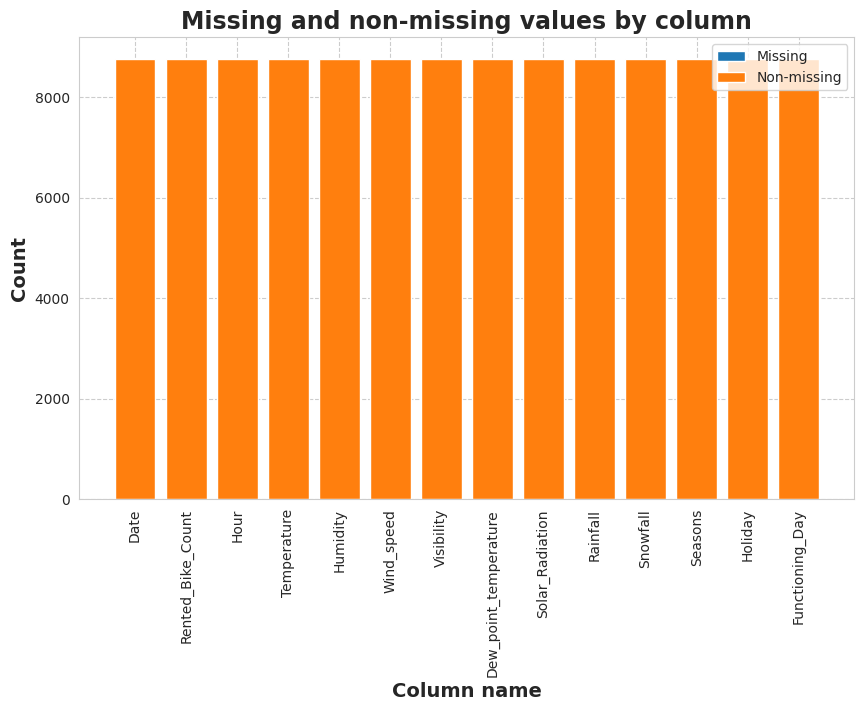


Fig. 1. Missing values bargraph

**4.3 Exploratory Data Analysis:**

In statistics, exploratory data analysis (EDA) is a method of examining data sets to highlight their key features, frequently utilizing statistical graphics and other techniques for data visualization. EDA differs from traditional hypothesis testing in that it is primarily used to explore what the data can tell us beyond the formal modeling. A statistical model can be utilized or not. We were able to determine a number of details and connections between the target and the independent variables thanks to EDA.

**4.4 Model Creation:**

I have created several demand-predictive models for bike sharing in this portion of the project. The major goal was to compare several modeling approaches and select the most appropriate one for this project. Several models, including k-means clustering, fuzzy c-means, factor analysis, regression analysis (linear, lasso, ridge, elastic net, polynomial), k-nearest neighbors (KNN), random forest, gradient boosting, extreme gradient boosting (XGBoost), CatBoost, and LightGBM, have been put to the test in order to accomplish this goal. These models were chosen based on their performance in comparable research and their capacity to make precise forecasts of the demand for bike sharing. Mean squared error (MSE) and R-squared were two performance metrics used to assess the models.

I have determined which model worked the best for this project after evaluating and contrasting the outcomes of each. The model that was chosen was picked because it had a high degree of accuracy in anticipating bike demand and because it could identify underlying patterns and trends in the data. The results will be used to optimize the operations of bike-sharing systems and guarantee a steady supply of rental bikes for the general public. The final model will be used to forecast bike demand for various hours of the day and days of the week.

### 4.4 Implementation

For implementation, I have used the traditional data science techniques which entails turning the models and algorithms created in earlier stages into functional code. The procedures used to put the predictive model for predicting demand for bike sharing using the Seoul Bike Sharing Dataset into practice are described in this section.

Data preparation was the initial stage of the implementation process. To make sure the raw dataset was in the right format and error-free, it underwent preprocessing. This includes scaling the data, handling missing numbers, and removing outliers. Following preprocessing, the data were divided into training and testing datasets.

Exploratory Data Analysis (EDA): In order to understand the data and find any patterns, trends, or abnormalities, the raw dataset was examined. This involved examining the data's distributions, identifying correlations between variables, and visualizing the data using graphs and charts. Any data quality problems, such as missing or inconsistent data, that required to be fixed during the data preparation stage were found using EDA. The data was consistent and have zero null or missing values which I have found out using isna() function and also depicting it with a heatmap. For my easy understanding I have changed the format of some of the features.

Model training came after that. Using the training dataset, the prediction models [7] were trained. To determine the optimum hyperparameters that minimized the prediction error, the models were optimized using strategies including Means Squared Errors, R2 and adjusted R2. The testing dataset was used to gauge the models' accuracy after they had been trained. Different models, such as gradient boosting, extreme gradient boosting (XGBoost), CatBoost, light gradient boosting machine (LightGBM), linear regression, lasso regression, ridge regression, elastic net regression, polynomial regression, K-nearest neighbors (KNN), random forest, fuzzy c-means, factor analysis, and gradient boosting, were trained and tested. The top-performing model was chosen for deployment after the models had undergone training and testing. To ensure that the bikes were available where and when needed, the distribution of the bikes to the various stations was optimized using the predictions.

The project's implementation phase included data preparation, model training, model selection, and model evaluation. Programming languages including Python, as well as libraries like Pandas, Scikit-Learn, matplotlib, seaborn, and factor analyzer, were used to implement the procedure. The implementation process allowed for the creation of a working predictive model that optimized bike sharing operations in a real-world setting, ensuring that bikes were always available where and when they were needed.

# V. RESULTS

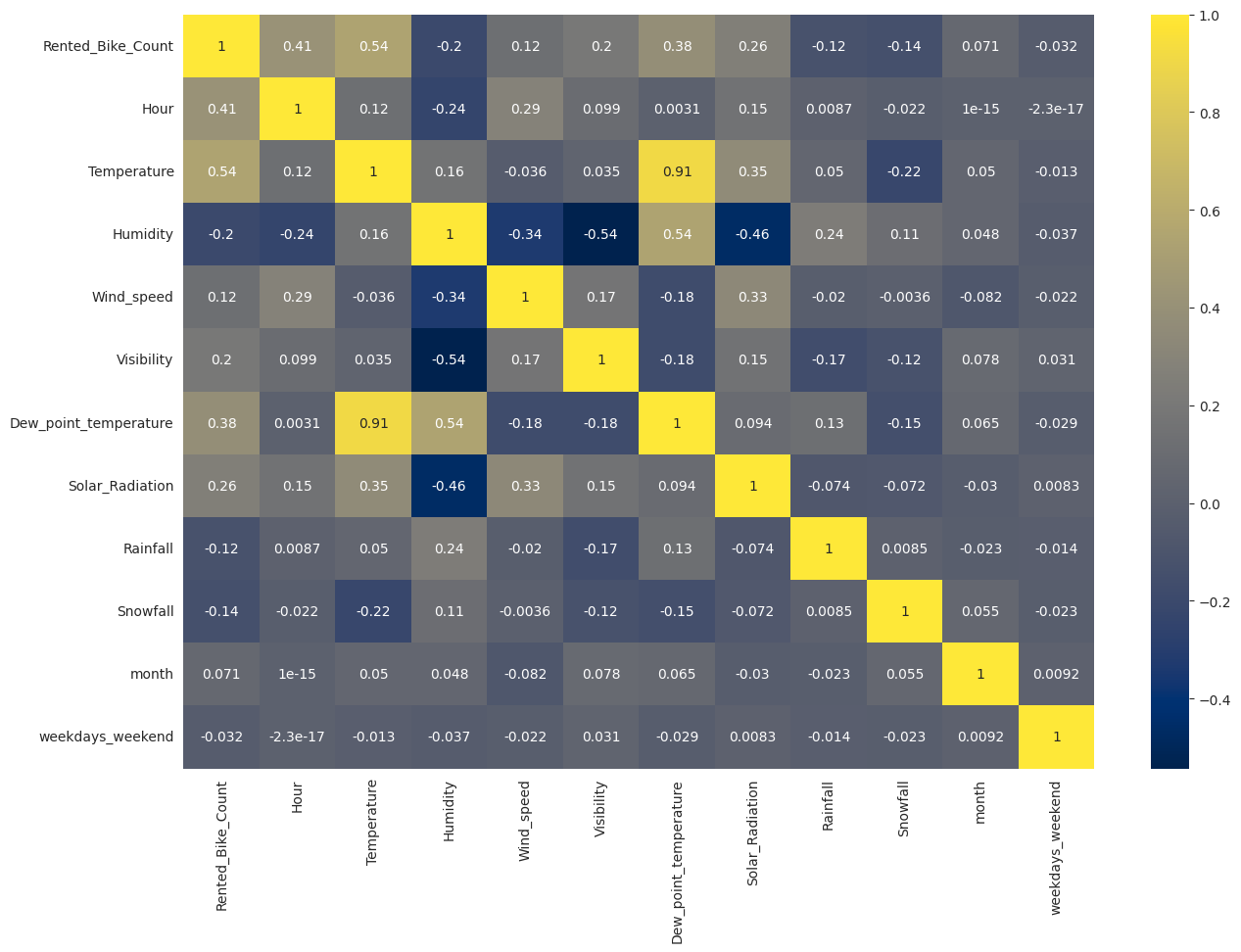


Fig. 2. Correlation using heatmap

A statistical measure called correlation expresses how closely two variables are related to one another. When two variables move in the same direction, there is a positive correlation; when one increases, the other increases as well. When two variables move in opposite directions, there is a negative correlation; as one increases, the other declines. To evaluate theories concerning cause-and-effect connections between variables, correlation can be used. In the actual world, correlation is frequently utilized to forecast trends. The correlation between temperature and Dew Point temperature is almost 0.91, which causes a multicollinearity problem. Dew point temperature is thus removed.

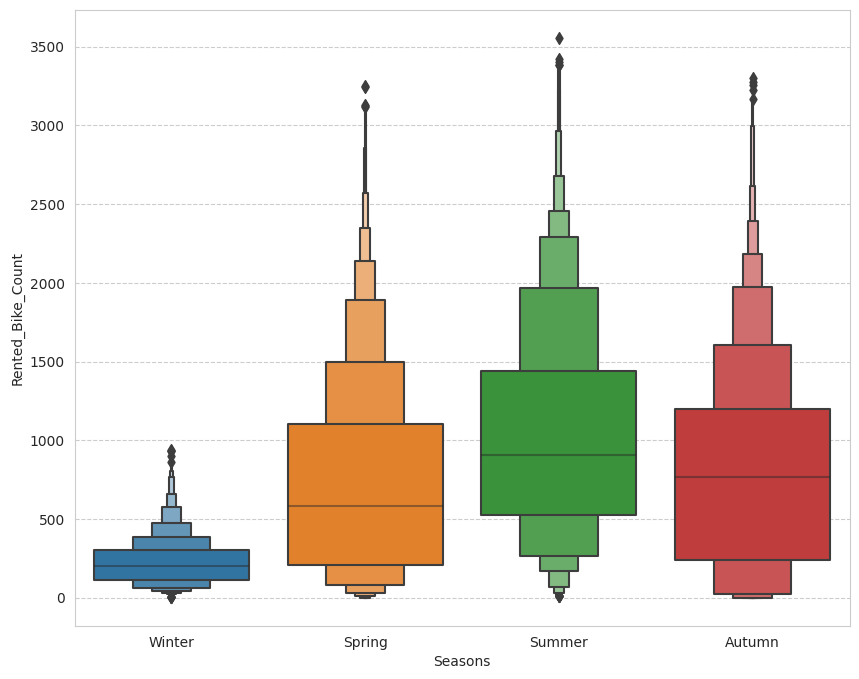


Fig. 3. Rented bike count vs seasons

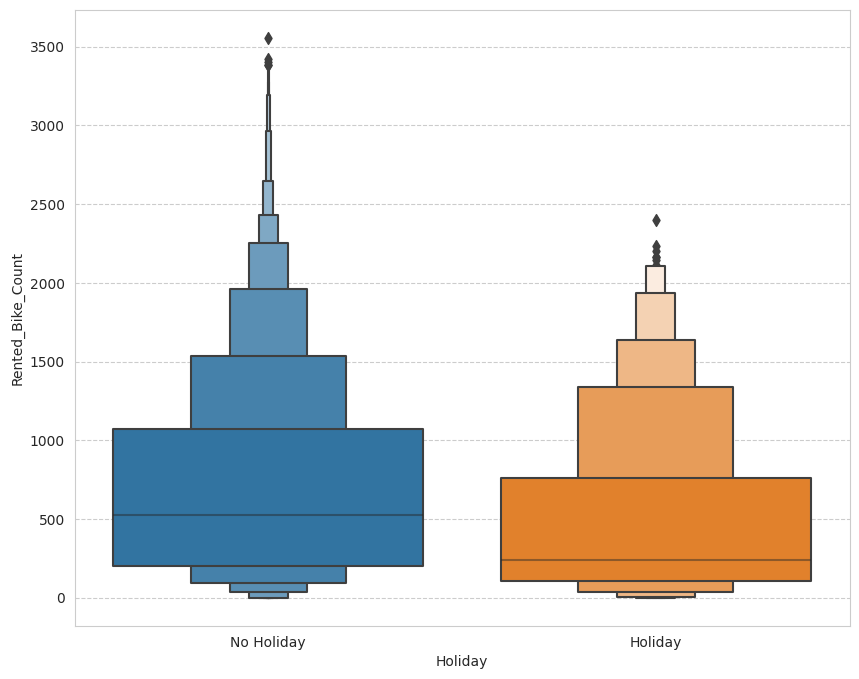


Fig. 4. Rented bikes based on Holiday

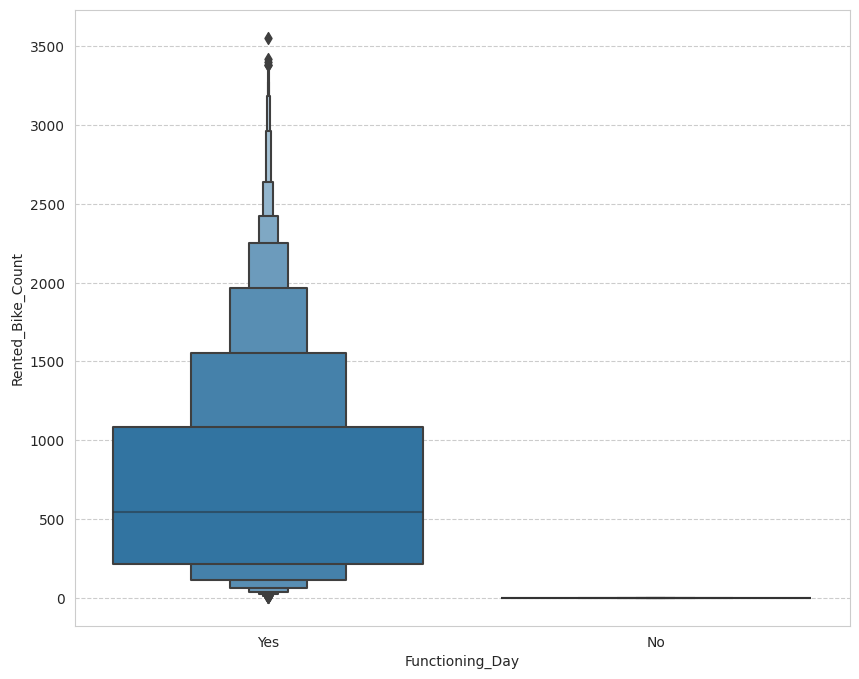


Fig. 5. Rented bikes based on Functioning days

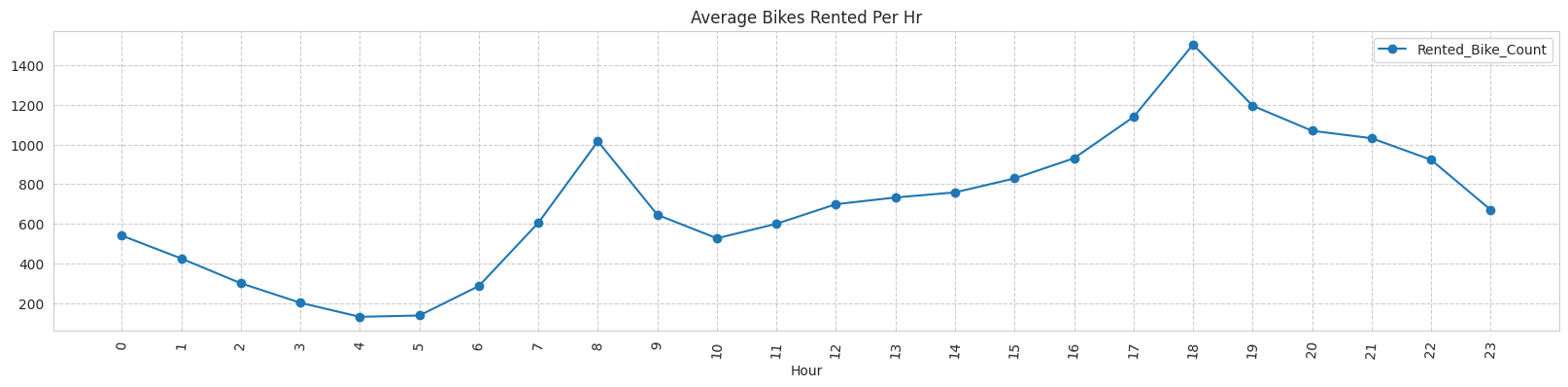


Fig. 6. Average bikes rented per hour

K-Means Clustering

The K-Means algorithm successfully clustered the bike rental stations in Seoul based on their "Rented Bike Count" and "Temperature(C)" values, as shown by the output scatter figure.8. The four distinct clusters that the algorithm found are represented by the four different colors in the scatter plot.

The scatter plot indicates that bike rental stations in Cluster 0 (shown in blue) are situated in areas with lower "Temperature(C)" values and typically have lower "Rented Bike Count" values. This implies that these stations might be situated in regions with lesser demand for bike rentals, such as those that are colder or less populated.

Contrarily, the bike rental locations in Cluster 2 (shown in green) typically have higher "Rented Bike Count" values and are situated in locations with higher "Temperature(C)" values. This implies that these stations might be found in more populous or popular tourist destinations, where there might be a greater need for bike rentals.

Overall, the output scatter plot offers some insightful information about Seoul's bike rental station usage patterns and might be used to guide business decisions about where to position new stations or how to devote resources to current ones.

The elbow approach is a way for figuring out how many clusters K-Means clustering should utilize. The technique entails locating the "elbow" or point of inflection in the curve by plotting the within-cluster sum of squares (WCSS) as a function of the number of clusters. Based on this, the best number of clusters is then selected.

This result led us to the conclusion that using 4 clusters was the ideal amount for the K-Means algorithm to utilize for clustering the data on bike rentals. A good trade-off between decreasing WCSS and avoiding overfitting the data with clusters is represented by this number of clusters.

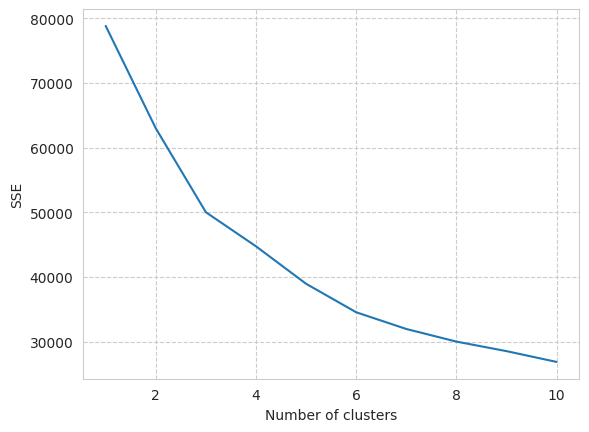


Fig. 7. Graph for elbow method

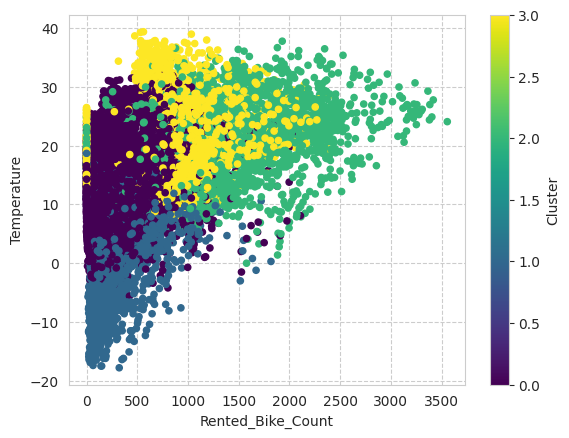


Fig. 8. K-Means clustering

Fuzzy C-Means

Fcm is a soft clustering algorithm that allows a data point to belong to more than one cluster simultaneously. I have implemented PCA before implementing fuzzy c-means.

The center of two of the clusters produced by the method is represented by the centroid with two primary clusters. The centroid's placement reveals the common traits of the data points in these two clusters. The amount of data points given to each cluster can be used to estimate its size. There are probably the most data points attributed to the two clusters in this example that the centroid best represents.

It is possible that these two groups are connected in some way or share similar traits, such as similar bike rental trends or shared geographic location.

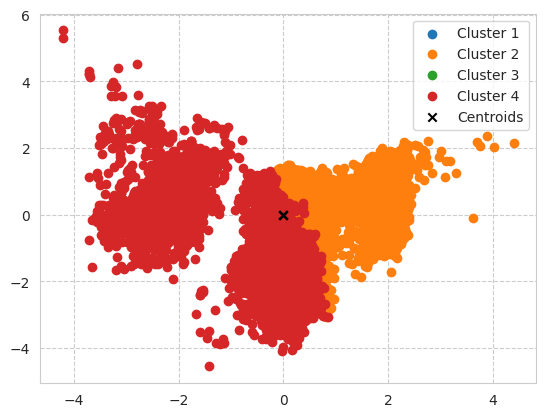


Fig. 9. Fuzzy C-Means

Factor Analysis

Factor analysis is one of the unsupervised machine learning algorithms which is used for dimensionality reduction. I have implemented factor analysis using factor analyzer

It is possible that the fact that the values for the first factor, temperature, and the dew point temperature, are the same and high in a factor analysis, illustrates the strong correlation that may exist between these two variables. This might be the case since the dew point temperature is determined by the temperature and relative humidity, as well as the second component of season (spring and summer have high values), the third element of hour and the windspeed has higher values.

Eigenvalues display how much of the data variance is accounted for by each factor. In this instance, we can see that the first component, followed by the second and third factors, explains the majority of the variance.

In general, factor analysis aids in the discovery of underlying factors that account for the correlation patterns among the original variables. It might be helpful for data reduction and locating crucial factors for additional research.

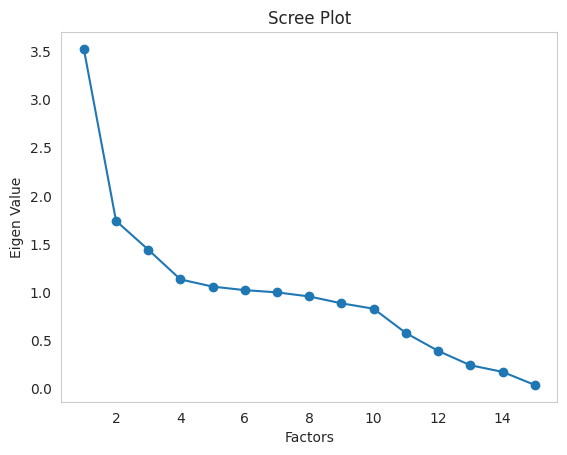


Fig. 10. Scree plot for factor analysis

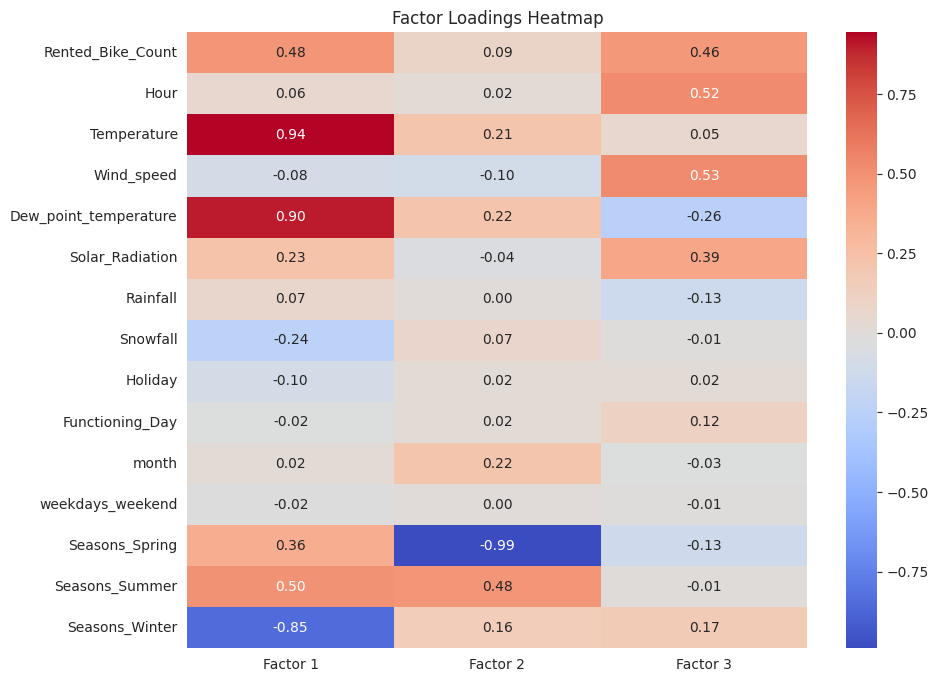


Fig. 11. Factor loadings heatmap for Factor Analysis

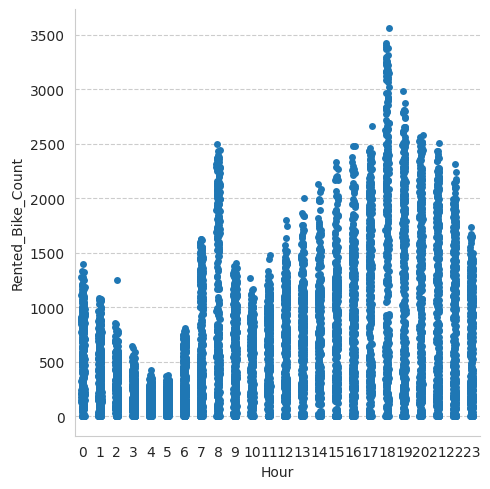


Fig. 12. Analysis based on Time

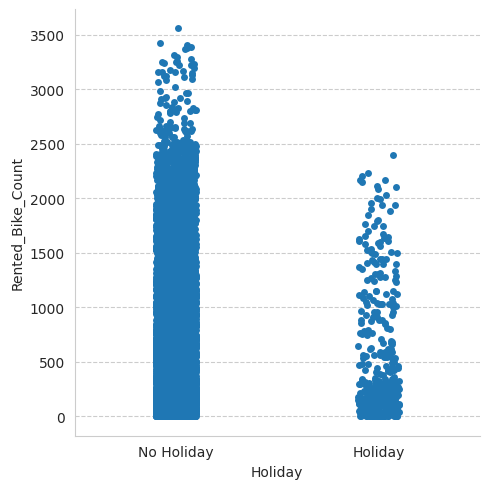


Fig. 13. Analysis based on Holidays

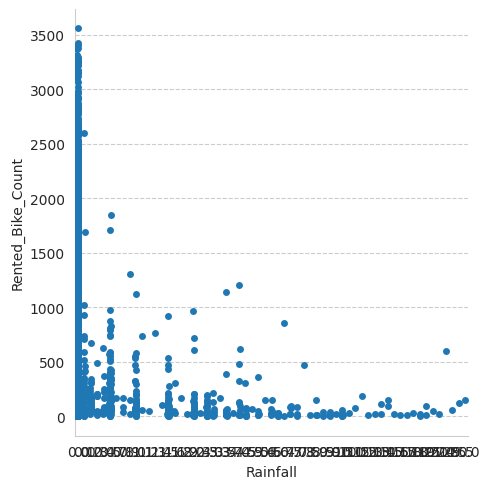


Fig. 14. Analysis based on Rain

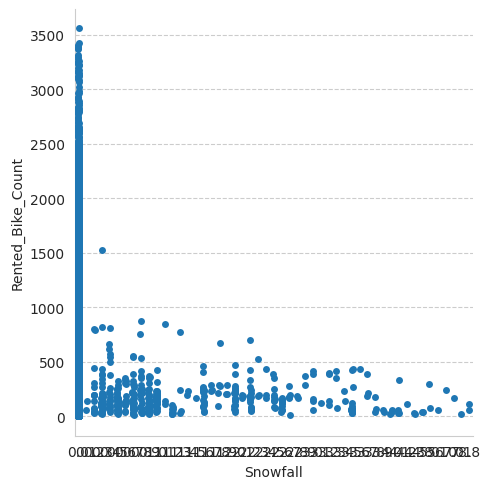


Fig. 15. Analysis based on Snowfall

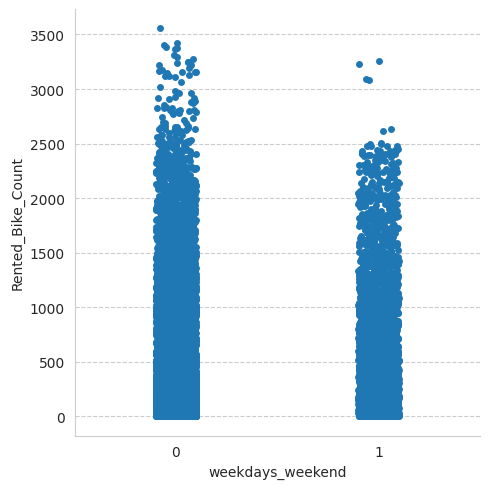


Fig. 16. Analysis based on weekdays

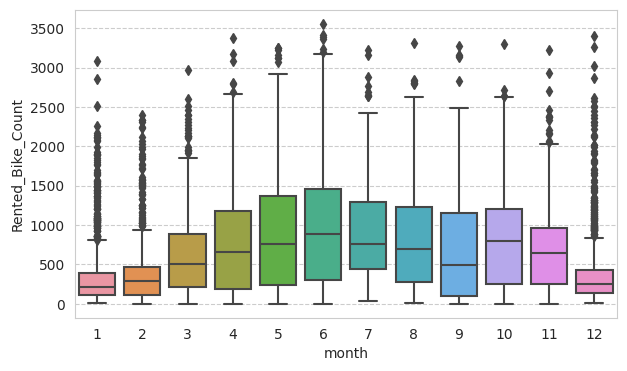


Fig. 16. Analysis of rented bikes based on month

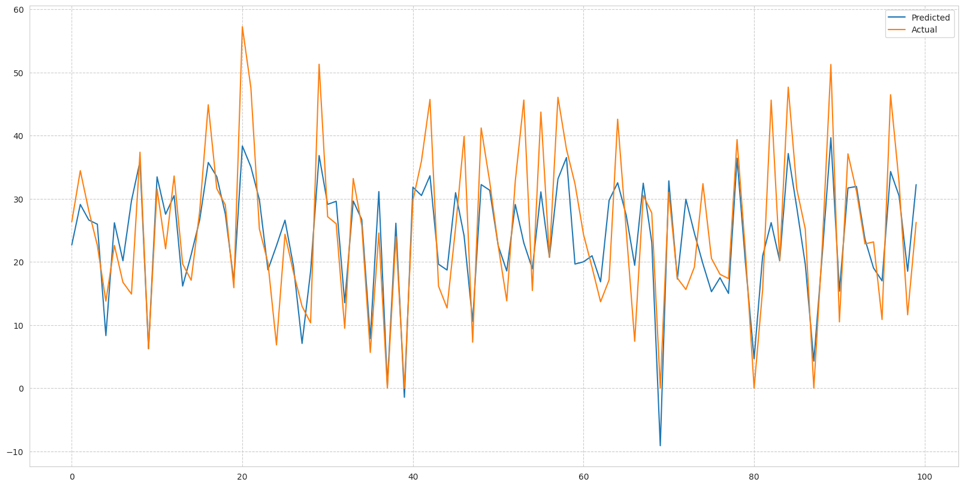


Fig. 17. Linear regression output

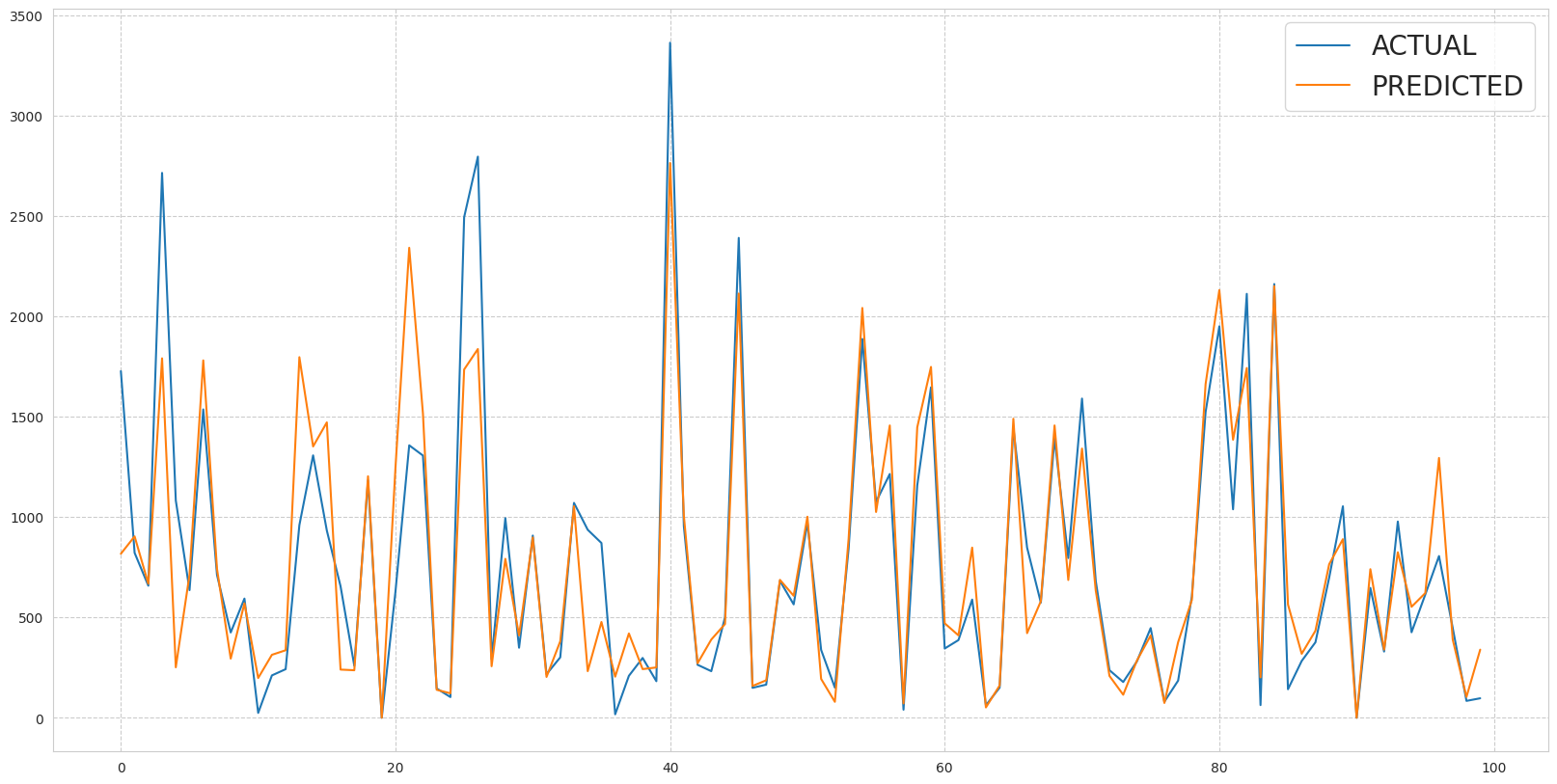


Fig. 18. KNN

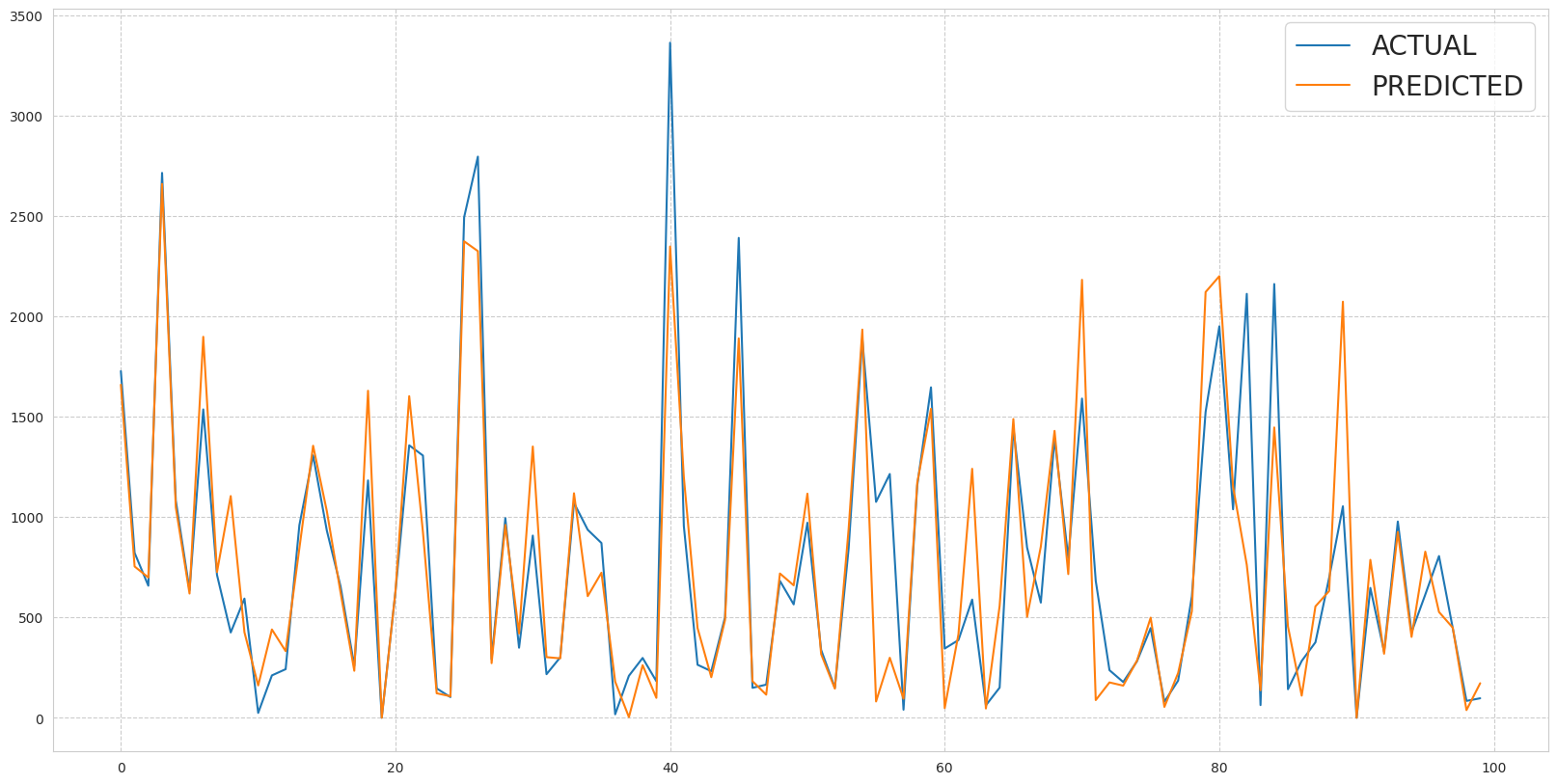


Fig. 19. Decision Tree output

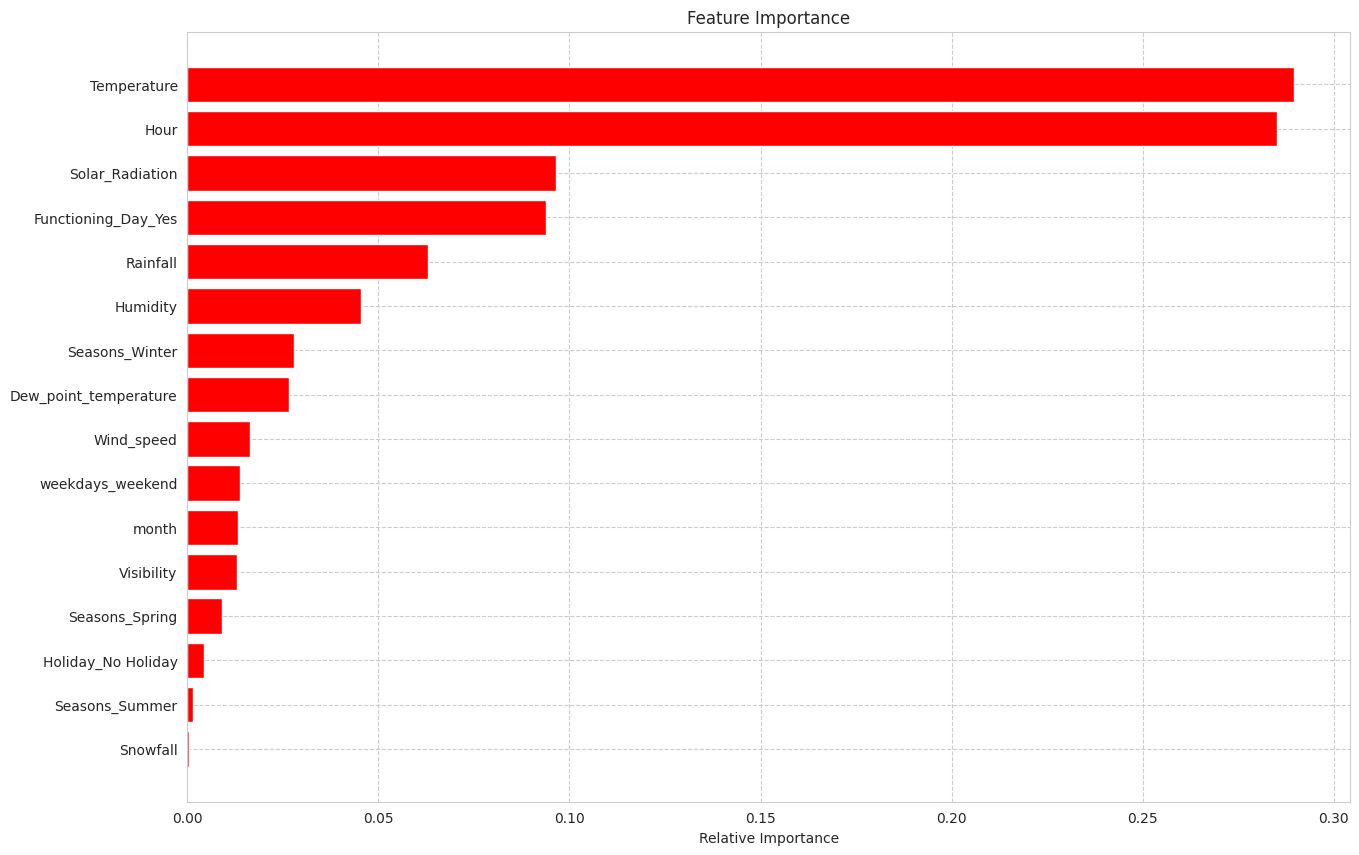


Fig. 20. Feature importance of decision tree

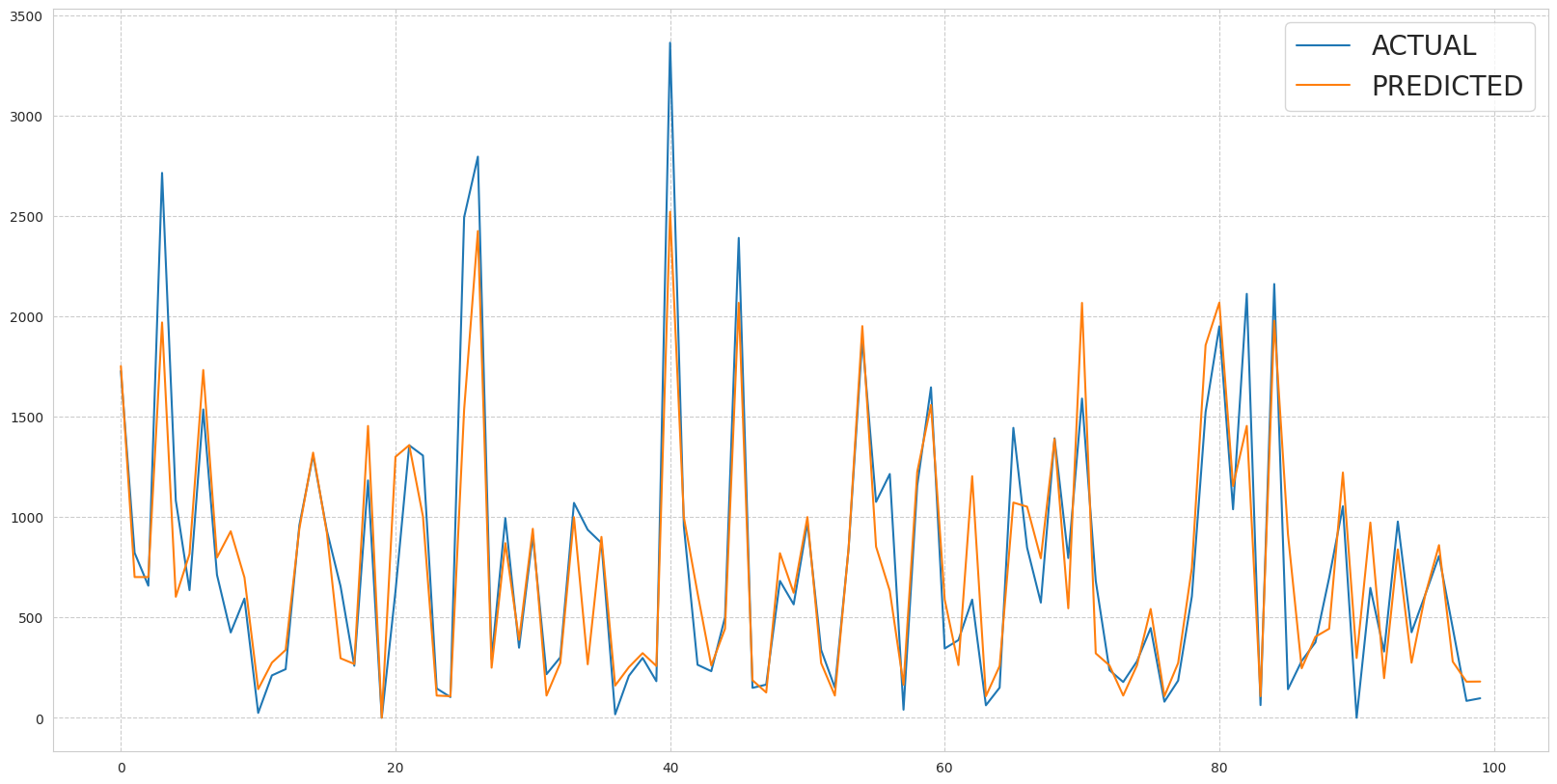


Fig. 21. Random Forest

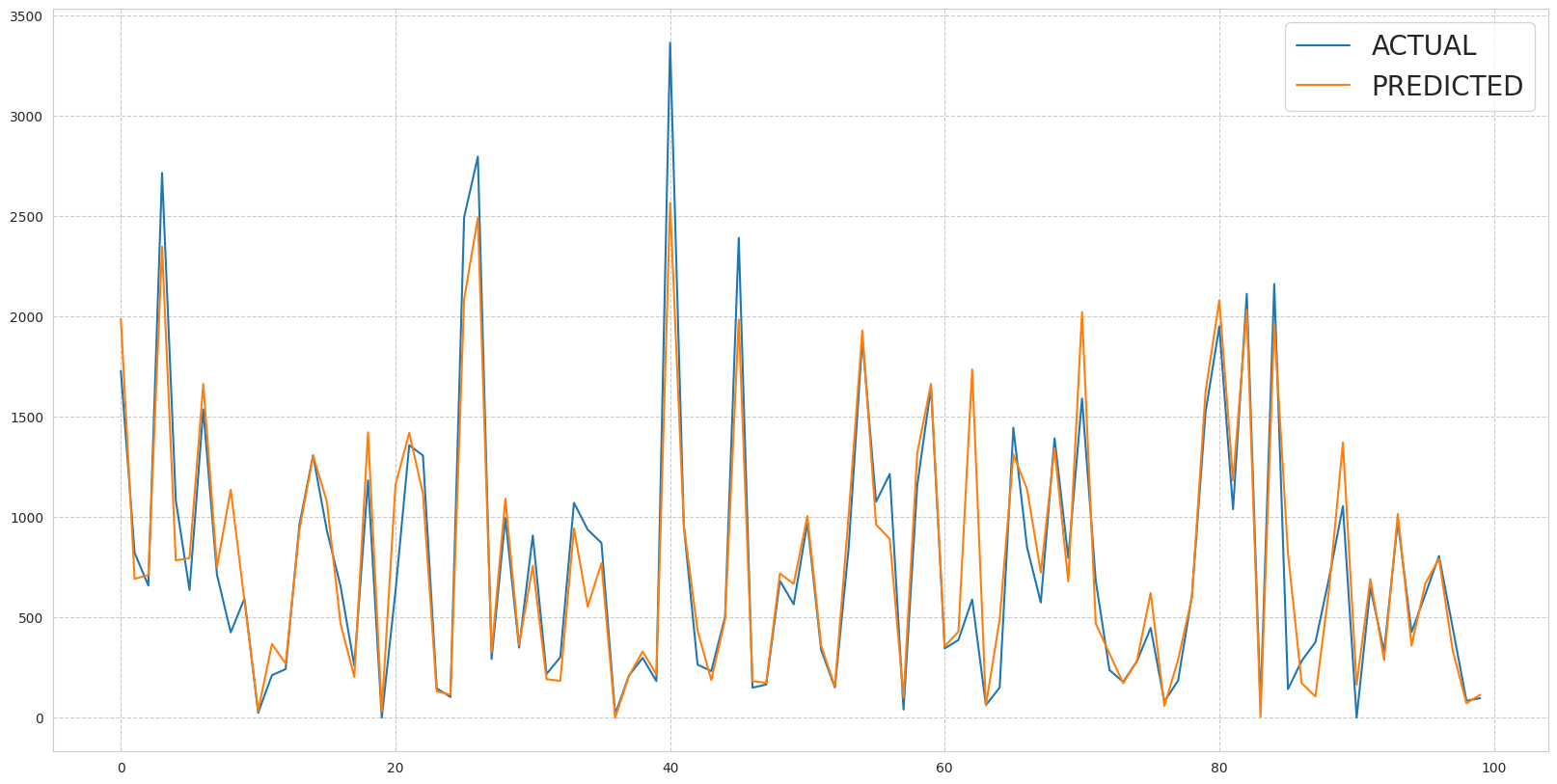


Fig. 22. Graph for Light GBM

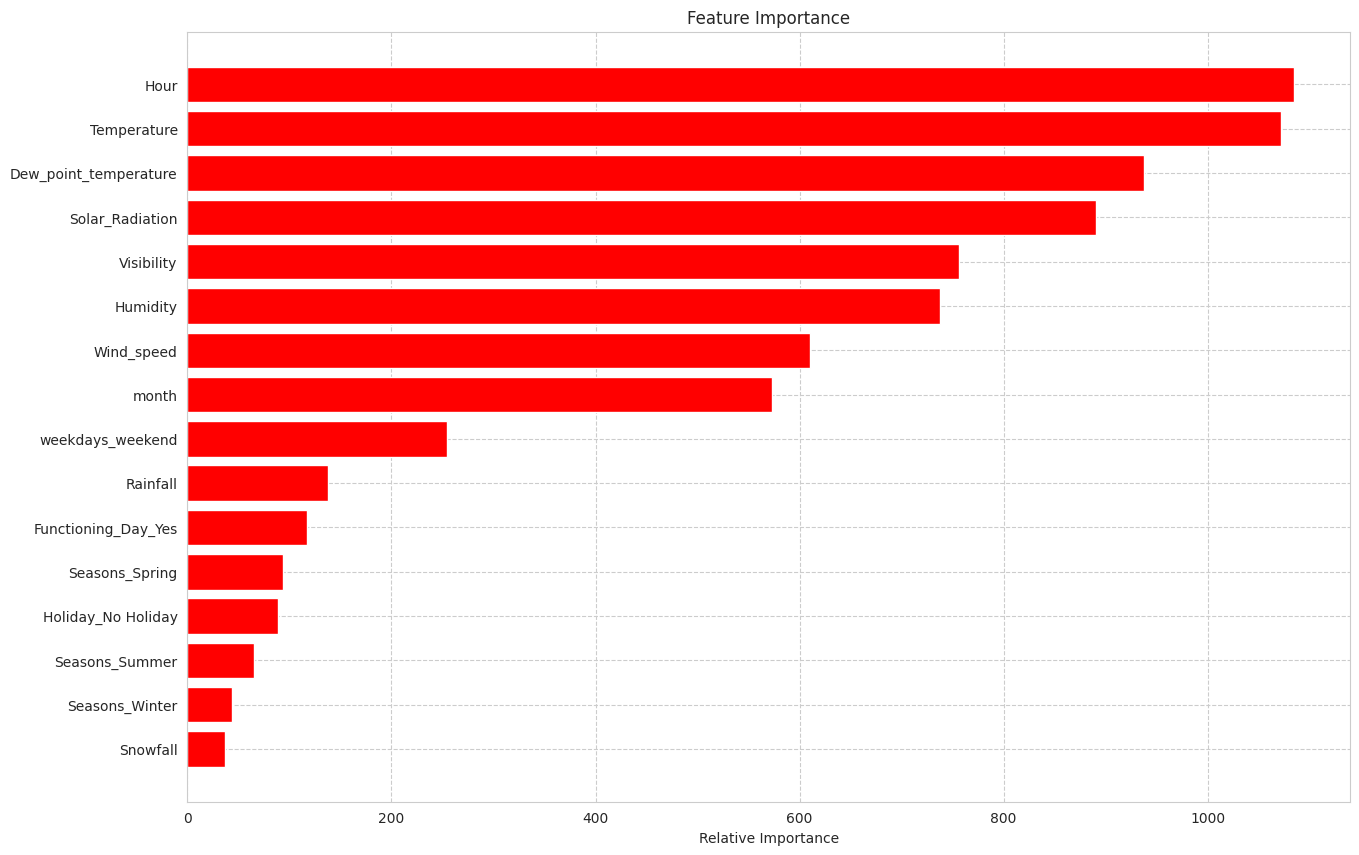


Fig. 23. Feature importance for Light GBM

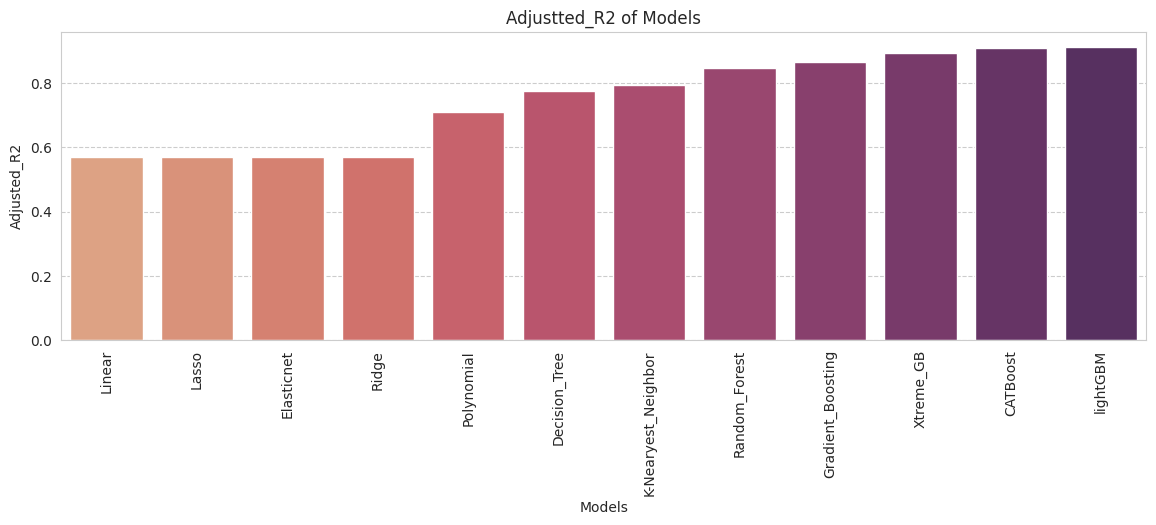


Fig. 24. Visualization of adjusted R2 of models

Adjusted R-squared is a modified version of R2 that accounts for the number of independent variables used in the model. It penalizes the addition of unnecessary variables, and higher values indicate better performance.

We can see with clarity that the cat boost and light gbm adjusted r2 values are higher, indicating that these models are a suitable fit for our project.

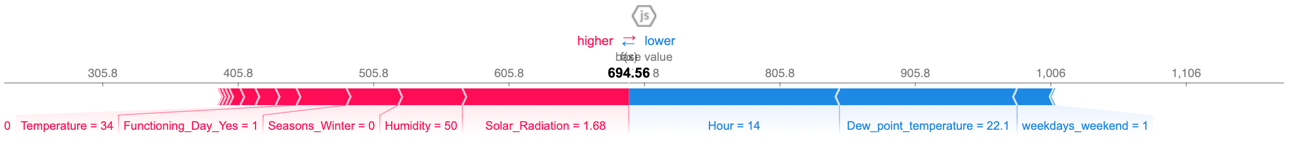


Fig. 25. Model explainability using SHAP for xgboost

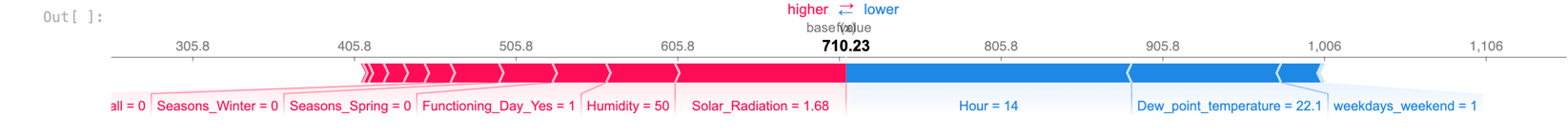


Fig. 26. Model explainability using SHAP for Light GBM

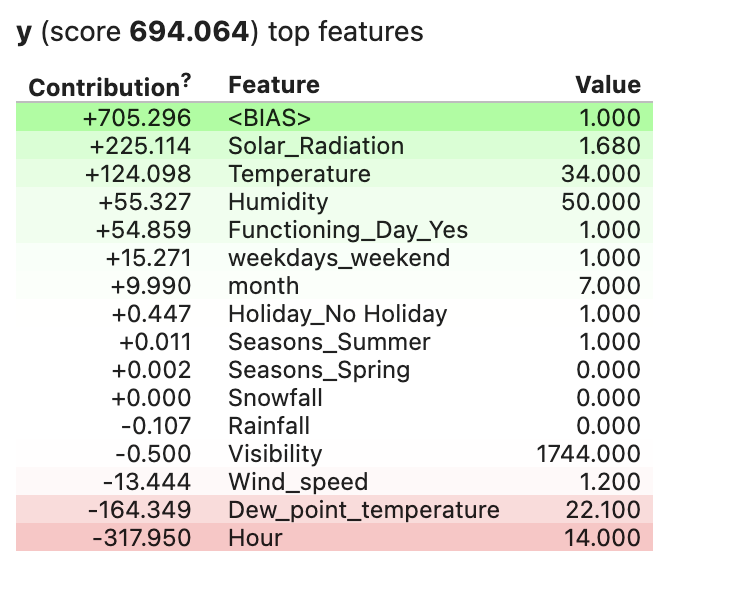


Fig. 27. Eli 5 for xgboost model

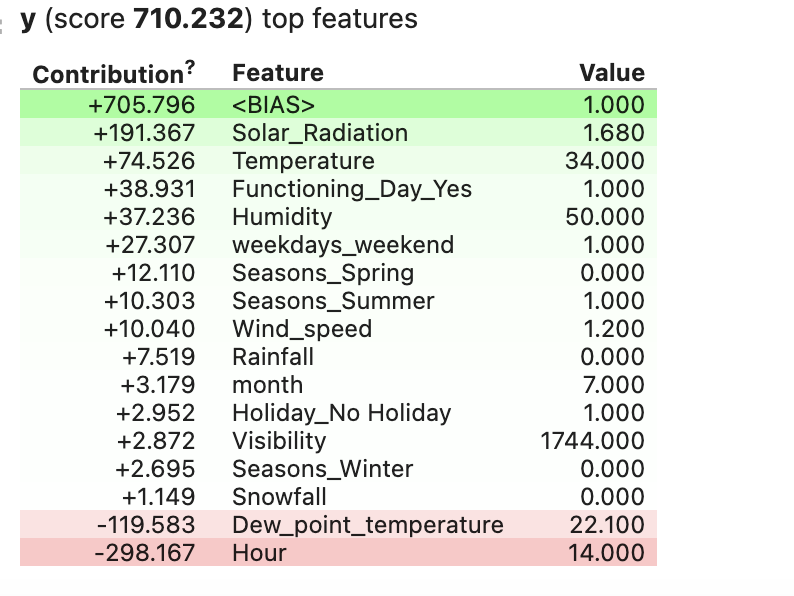


Fig. 28. Eli 5 for LGBR model

The weights are positive for the green hue, which indicates how much the feature helps to the prediction of the relevant class. The negative weights of the red hue show that the feature isn't helping to predict that class. As can be seen from the output above, eli5 illustrates how each feature affected the outcome prediction.

# VI. CONCLUSION

The goal of this study was to forecast demand for bike sharing using the available data. To predict the length of the trip, various regression techniques are used, including inear regression, lasso regression, ridge regression, elastic net regression, polynomial regression, K-nearest neighbors (KNN), random forest, gradient boosting, extreme gradient boosting (XGBoost), CatBoost, light gradient boosting machine (LightGBM), k-means, fuzzy c-means, and factor analysis. This statistical data analysis reveals intriguing results in the exploratory analysis as well as the prediction strategy.

The experimental finding demonstrates:

A heat map demonstrates a strong correlation between temperature and dew point. When there is a working day, the bike is hired otherwise, it is not. The most bikes are leased between the hours of 17 and 19, as well as early in the morning at 8pm in 2018. Most bikes are hired on the sixth and ninth of each month. Summer saw the highest number of bike rentals, followed by autumn, spring, and winter. Except in a few cases, people have reserved more bikes when it hasn't rained as much. The busiest months for renting bikes are May through July, and December through February are the least popular months. Workdays rather than holidays are when the majority of bicycles are rented. After testing a number of models, it was determined that the lightGBM and Catboost model was the most effective for predicting demand for bike sharing because its performance metrics (mse, rmse) were lower and its adjusted\_r2 value was higher. For the bike rental stations, we can utilize either the lightGBM or catboost models.

The predictive model created for this research has significant effects on bike-sharing programs, especially in terms of increasing the accessibility and availability of rental bikes. The predictive model can assist bike-sharing systems manage their operations and resource allocation by predicting the hourly demand for rental bikes. Overall, the predictive model has the potential to significantly increase the efficiency and effectiveness of bike-sharing programs, making them easier to use, more effective, and more accessible.

Some potential future directions for this research include exploring the use of machine learning algorithms, such as deep learning and reinforcement learning, to improve the accuracy of the predictive models. Additionally, research could be done to examine the transferability of the models to other cities and to other types of transportation systems.

# VII. LIST OF REFERENCES

[1] Yu, Q., & Cheng, T. (2016). Predicting bike-sharing demand using support vector regression with optimization. IEEE Transactions on Intelligent Transportation Systems, 17(10), 2899-2910.

[2] Liu, C., Wei, L., Xu, Y., & Xu, G. (2018). Forecasting bike sharing demand using deep learning in metropolitan areas. Transportation Research Part C: Emerging Technologies, 86, 350-365.

[3] Peng, X., Yu, H., Yang, H., & Zhao, J. (2021). Demand prediction of bike-sharing systems: A random forest and clustering approach. Transportation Research Part D: Transport and Environment, 94, 102767.

[4] Wang, P., Zhang, W., & Huang, J. (2018). A comparative study of demand forecasting methods for bike-sharing systems. Sustainability, 10(10), 3478.

[5] Sathishkumar V E, Jangwoo Park, and Yongyun Cho. 'Using data mining techniques for bike sharing demand prediction in metropolitan city.' Computer Communications, Vol.153, pp.353-366, March, 2020

[6] Sathishkumar V E and Yongyun Cho. 'A rule-based model for Seoul Bike sharing demand prediction using weather data' European Journal of Remote Sensing, pp. 1-18, Feb, 2020

[7] S. Sathishkumar and G. Chinnadurai. "Prediction and analysis of bike sharing demand using machine learning algorithms." International Journal of Pure and Applied Mathematics, vol. 119, no. 12, pp. 1595-1607, 2018.

[8] J. Kim, J. Lee, and C. Park. "A comparison of machine learning models for bike-sharing demand forecasting in metropolitan cities." Sustainability, vol. 11, no. 6, pp. 1-16, 2019.

[9] C. Liu, C. Zhang, and W. Chen. "Bike-sharing demand prediction using machine learning algorithms with weather data." Sustainability, vol. 11, no. 6, pp. 1-18, 2019.

[10] J. Li, X. Huang, and D. Zhang. "Research on bike sharing demand prediction based on deep learning." Journal of Advanced Transportation, vol. 2019, pp. 1-11, 2019.

[11] S. Patel and S. Patel. "Bike sharing demand prediction using machine learning." In 2019 International Conference on Communication Information and Computing Technology (ICCICT), pp. 1-6, 2019.