

Seoul Bike Sharing Demand Prediction

Pavithra K

Data science trainee

AlmaBetter, Bangalore

Abstract:

Bike sharing systems are innovative ways of renting bicycles for use without the onus of ownership. A pay per use system, the bike sharing model is either works in two modes: users can get a membership for cheaper rates or they can pay for the bicycles on an ad-hoc basis. The users of bike sharing systems can pick up bicycles from a kiosk in one location and return them to a kiosk in possibly any location of the city

Exploratory Analysis process makes us to understand how temperature, humidity, seasons, snowfall, raining, seasons, effect the demand for bike sharing.

The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes using machine learning regression algorithms considering previous trends to forecast the demand for bike sharing.

Introduction:

Bike-sharing systems are shared transport services that allow individuals to rent or borrow bikes on a short or long-term basis. These services can be free, for a fee or

require the individual to leave a deposit when checking out the bike. Most bike-sharing systems have a network of docks where the user can check out and check in bikes but the exact nature of operations depend on the kind of bike-sharing system. It is cost effective, Less Congestion on City Streets, Shorter Commute Times, Healthy Alternative and enjoying riding. The data set consists of complete one years' worth of hourly rental data.

Our goal is to analyze and predict bike count required at each hour for the stable supply of rental bikes based on historical data to discover key factors affecting the rental bike count. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Problem Statement:

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make

the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

The goal of business is

Maximize: Availability of bikes

Minimize: Waiting time of customers.

Data description:

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility,

Dewpoint, temp, solar radiation, Snowfall, Rainfall) the number of bikes rented per hour and date information.

Attribute Information: The dataset contains 8760 observation and 14 features

- Date: year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Wind Speed - 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 - (Non-Functional Day), Fun (Functional Day).

Methodology followed:

1. **Importing dataset:** Here we use pandas libraries to import data.
 2. **Exploratory Data analysis:** Exploratory Data Analysis (EDA) is a critical component involved while working with data. EDA is used to comprehensively understand the data and discover all its characteristics, summarizing data, identifying patterns and relationships, and detecting outliers, typically by employing visual techniques. This is an important step in data analysis that can be used on both qualitative and quantitative data. This makes it possible for you to understand your data more thoroughly and find interesting patterns in it.
- **Understanding about data source:** To utilize the data in well manner

the knowledge of data source and the data collection process is important. to what purpose data has been collected for research purpose or for randomly collected.

- **Understanding about attributes of data:** This includes understanding the dataset's data type, what each column represents, and any other relevant information. This understanding is critical for properly performing an EDA because it will help you know what to look for and how to analyze the data.

- **Data cleaning:** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. It includes following steps:

- ✓ **Dealing with missing values:** Sometimes we may find some data are missing in the dataset. We know that we can use the `isnull().sum()` and `notnull().sum()` functions from the pandas library to determine count of null values present. Based on amount of data present either we can drop the missing rows, columns or replace missing values with calculated mean, mode or median of respective.

- ✓ **Dealing with duplicate rows:** Sometimes we may find some data are missing in the dataset. We know that we can use the `isnull().sum()` and `notnull().sum()` functions from the pandas library to determine count of null values present. Based on amount of data present either we can drop the missing rows, columns or replace missing values with calculated mean, mode or median of respective.

- ✓ **Dealing with incorrect format:** It is important to check our data is in correct formats (int, float, text, or other). We can use following pandas function, `df.dtypes()` to check data types and `.astype()` to change the data type `df.dtypes`-will return data types in our data frame.

- ✓ **Dealing with Anomalies/Outliers:** Outliers are unusual values in your dataset, handling them is important because they can distort statistical analyses and violate their assumptions. Outlier are very informative it is important to know reason of cause of outliers it may cause due to Data entry errors/ Measurement errors/ instrument errors/Sampling errors/Data processing error/Natural novelties in data

- **Data Analysis:**

- ✓ **Univariate analysis:** The univariate analysis of data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. It is done using Histogram, Pie chart, Bar chart, Boxplot.
- ✓ **Bivariant analysis:** As its name indicates this type of analysis includes two variables for analysis. Through this we can find the how two variables are related to each other
- ✓ **Multivariate analysis:** Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set. Using pairplot, heatmap on correlation matrix

5. Encoding of categorical column: Feature encoding of categorical variables binary because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

6. Binning: one of the popular techniques of feature engineering, "binning", can be used to normalize the noisy data. This process involves segmenting different features into bins.

7. Transformation: Transform helps in handling the skewed data, and it makes the distribution more approximate to normal after transformation. It also reduces the effects of outliers on the data, as because of the normalization of magnitude differences, a model becomes much robust.

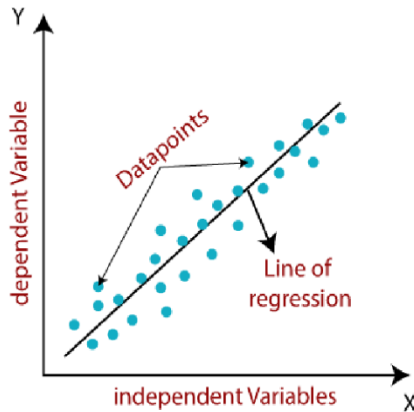
8. Splitting the dataset into training and testing dataset.

7. Fitting different Regression models:

For modelling we tried various regression algorithms like:

1. Linear Regression:

Linear regression is supervised machine learning algorithm. It is a simple statistical regression technique to find a linear relation of independent variable with target variables or we can say that, Regression models a target prediction value based on independent variables.



It is mathematically represented by

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

Where \hat{Y} target or dependent variable, X_1, X_2, \dots are independent variables

β_0 is intercept of line, β_1, β_2 are model parameters, ε is error.

The goal of linear regression is to get best fit line (regression line) with best parameters ($\beta_0, \beta_1, \beta_2$) with minimal residual error (i.e., difference between actual and predicted value) or we can say that minimum cost function(J) the RMSE value between actual y and predicted value (pred).

$$J = \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2$$

The optimal value of cost function is obtained by Gradient descent which is a first order iterative algorithm used to update the model parameter using hyperparameter learning rate which will decide how large the steps to be taken to

update the parameter in order to get global minimum.

The assumptions of linear regression:

- Linear relation between independent and dependent variable.
- No or minimal multicollinearity:
- Homoscedasticity of error term:
- Normal distribution of error term

The assumptions:

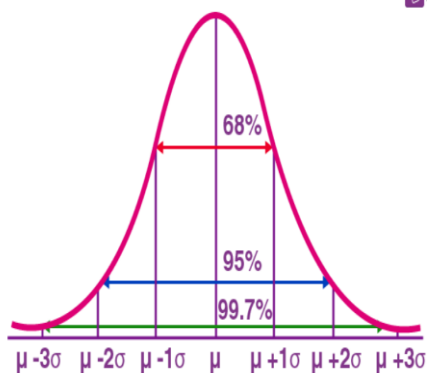
1. Multicollinearity between the independent variables should be zero or minimal. It can be checked using correlation matrix or vif.

VIF (variance influence factor): It is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results.

$$VIF_i = \frac{1}{1 - R_i^2}$$

- VIF equal to 1 = variables are not correlated
- VIF between 1 and 5 = variables are moderately correlated
- VIF greater than 5 = variables are highly correlated.

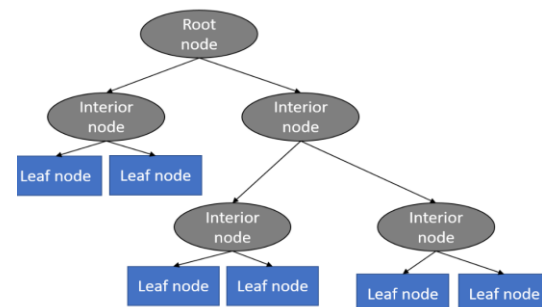
2. **Homoscedasticity** describes how similar or how far the data deviates from the mean. This is an important assumption to make because parametric statistical tests are sensitive to differences. Opposite of homoscedasticity is heteroscedasticity that caused due to extreme outliers, which will adversely affect the regression problem.
3. **Normal distribution:** A probability function that specifies how the values of a variable are distributed is called the normal distribution. It is symmetric since most of the observations assemble around the central peak of the curve. The probabilities for values of the distribution are distant from the mean narrow off evenly in both directions.



2. Decision Tree:

Decision tree supervised machine learning non parametric algorithm. This algorithm used for regression as well as for classification algorithm.” Learning simple decision rules in the form of tree inferred from the data

features”. It has tree like structure consisting of nodes and branches. Where internal node represents the features of dataset, branches represent decision rules and leaf node represents outcome. The leaf node may have continuous (Regression) as well as categorical values(classification).



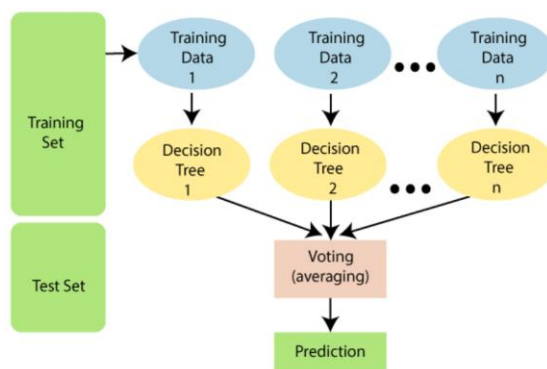
- ✓ **Root Node:** Root node is from where the decision tree starts.
- ✓ **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- ✓ **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes
- ✓ **Branch/Sub Tree:** A tree formed by splitting the tree.
- ✓ **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- ✓ **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

Decision trees have an advantage that it is easy to understand, lesser data cleaning is

required, non-linearity does not affect the model's performance and the number of hyper-parameters to be tuned is almost null. However, it may have an over-fitting problem, which can be resolved using the **Random Forest** algorithm which will be explained in the next article.

4. Random Forest :

Supervised machine learning algorithm used for regression as well as for classification technique. It is based on the concept of **ensemble learning**. Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly (row replacement method) selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the averaging the predicted output of all the output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



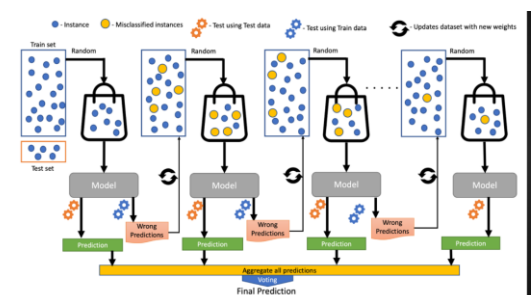
- It works well even if it contains missing or null values in it.

- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

5. XGBoost Regressor.

The XGBoost algorithm performs well in machine learning competitions because of its robust handling of a variety of data types, relationships, distributions, and the variety of hyperparameters that you can fine-tune. You can use XGBoost for regression, classification (binary and multiclass), and ranking problems. Before moving ahead we should know about boosting.

Boosting is an ensemble learning technique to build a strong regressor from several weak regressor in series. Boosting algorithms play a crucial role in dealing with bias-variance trade-off. Unlike bagging algorithms, which only controls for high variance in a model, boosting controls both the aspects (bias & variance) and is more effective.



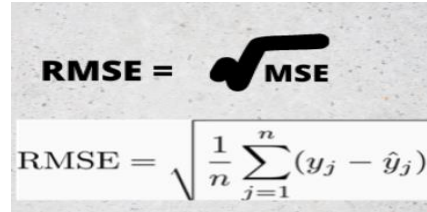
- A loss function should be improved, which implies bringing down the loss function better than the result.
- To make expectations, weak learners are used in the model
- Decision trees are utilized in this, and they are utilized in a jealous way, which alludes to picking the best-divided focuses considering Gini Impurity and so forth or to limit the loss function
- The additive model is utilized to gather every one of the frail models, limiting the loss function.
- Trees are added each, ensuring existing trees are not changed in the decision tree. Regularly angle plummet process is utilized to find the best hyper boundaries, post which loads are refreshed further.

8.Evaluation metrics:

1)MSE(Mean of Squared Error): Mean squared error states that finding the squared difference between actual(y) and predicted value(y hat).

$$MSE = \frac{1}{n} \sum \underbrace{(y - \hat{y})^2}_{\substack{\text{The square of the difference} \\ \text{between actual and} \\ \text{predicted}}}$$

2)RMSE(Root Mean Squared Error): As name indicates rmse is square root of mean squared difference between actual(y) and predicted value(y hat).



$$RMSE = \sqrt{MSE}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

3)R2 or coefficient of determination: R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.

$$R^2 \text{ Squared} = 1 - \frac{SSr}{SSm}$$

SSr = Squared sum error of regression line

SSm = Squared sum error of mean line

4)Adjusted R2: The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1} \right) \times (1 - R^2) \right]$$

where:

n = number of observations

k = number of independent variables

R_a^2 = adjusted R^2

Hyperparameters tuning for better accuracy:

Hyperparameters tuning for decision tree, random forest and XGBoost algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree based models.

Seoul Bike Sharing Demand Prediction:

Here we are reporting our work on "**Seoul Bike Sharing Demand Prediction**". It contains inference of analysis done on data to find major factors affecting demand for bike sharing and build the prediction models.

- **Importing the dataset:** Seoul Bike Sharing Demand dataset in .csv (comma-separated values file) is imported using pandas' library.
- **Identifying the Data Source and Data Collection:** Seoul Bike Sharing Demand is dataset collected by researcher of service provider to maximize Maximize the Availability of bikes, Minimize the Waiting time of customers and also understand the factors affecting the demand for bike sharing.
- **Dataset Information (Identifying the Variables):** This understanding Seoul Bike Sharing Demand is dataset is critically important for properly performing an EDA and to build suitable prediction model because it will help you know what to look for and how to analyze the data and to understand which type of algorithm need to be built.
- **Understanding Seoul Bike Sharing Demand dataset:**

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

Attribute Information: The dataset contains 8760 observation and 14 features

- ❖ **Date:** is object type dataset and contains 365 unique dates. i.e, one year of information 365 from 01/12/2017 to 30/11/2018
- ❖ **Rented Bike Count:** It is target variable and datatype is int64. it is about count of number of bikes that are demanded to share per hour.
- ❖ **Hour:** Hours of the day that is in 24 format and its datatype is int64.
- ❖ **Temperature(°C):** Temperature of area where the bike sharing demand. Datatype is float. The temperature varies from -17°C to 39.4°C. It is represented by measuring unit "°C".
- ❖ **Humidity (%):** Humidity of area while booking a bike represented by unit %. Datatype is int64
- ❖ **Wind speed (m/s):** speed of wind while booking bike measured using unit m/s.
- ❖ **Visibility (10m):** Visibility of road in the area.
- ❖ **Dew point temperature(°C):** dew point is frequently cited as a more

accurate way of measuring the humidity and comfort of the measured in °C

- ❖ **Solar Radiation (MJ/m2):** Sunlight during the booking a bike. It is measured in MJ/m2
- ❖ **Rainfall(mm):** rainfall in the area of booking. Measured in mm
- ❖ **Snowfall (cm):** snow fall in the area of booking. It is measured in cm
- ❖ **Seasons:** booking day belongs to which season. This data set included all four-season winter, summer, autumn, and spring seasons. It is categorical datatype.
- ❖ **Holiday:** The day which was booked is a holiday or working day. Datatype is categorical.
- ❖ **Functioning Day:** the working day or not datatype is categorical.

- **Data Cleaning Seoul Bike Sharing Demand dataset:**

- ✓ **Dealing with missing values:**
Our **Bike Sharing** dataset does not contain any null values (i.e, there is zero number of missing value). So, there is no need to drop or imputation of null value with calculated values.

- ✓ **Dealing with duplicate rows:**

Our **Bike Sharing dataset** does not contain any duplicate rows. So, no need to deal with duplicate values.

- ✓ **Dealing with incorrect format:** Date feature in bike sharing dataset is converted to timestamps. Feature name are renamed because the feature name contains special characters

- ✓ **Dealing with Anomalies/Outliers:**

On conducting univariant analysis on continuous data we come to know that few feature suffer from skewness, left skewed features are Dew_point_temp, Visibility, Temperature, and right skewed features are Bike_Count, Wind_speed, Solar_Radiation, Rainfall, Snowfall. So, we used power transformation to reduce skewness of data.

- **Data Analysis (Univariate and bivariate analysis on Bike Sharing dataset).**

- ❖ We encountered many features that suffer from skewness of data even though these are extreme points we cannot lose them because they play an important role for bike count.
- ❖ **Hour:** Analysis on hour basis, we come to know that, Peak hours for bike sharing demand is between 4pm- 10pm. Very high around 5.30pm -6.30pm i.e, evening time. This feature is converted into categorical feature.
- ❖ **Temperature** (temperature, dew_point_temp, humidity): On analysis average bike sharing demand increases as temperature increases. But there is average demand of bike sharing as temperature goes above 32 degrees. In terms of as dew_point_temp increases demand for bike sharing also increases. The humidity is also related with temperature when humidity is high demand is less
- ❖ Demand for bike sharing increases with increase of visibility.
- ❖ When **snow fall** is less the demand for bike sharing is high as the snow fall increases the demand decreases. On heavy snow fall roads may be closed
- ❖ As **rainfall** increases the demand for bike sharing decreases. Solar radiation is not much effective in demand of bike sharing. As **wind speed** increases demand for rental bike decreases.
- ❖ **Seasons:** Average demand for rented bike is high during summer and autumn when compared with spring but it is very less during winter seasons.
- ❖ **Holiday:** Average demand for rented bike is high during normal day than during holiday.
- ❖ **Functioning day:** we have observed that bike sharing demand is high functioning day(week_day). Also, observed

that there is no demand for bike sharing during non-functioning day.

- **Dealing with skewness:** While conducting univariate analysis we encountered with skewness in distribution of few features. To overcome this we will use power transformation (will make the probability distribution of a variable more Gaussian).
- **Binning:** Converting hour feature into categorical. The feature includes four classes (bins) morning, day_time, evening and night
- ❖ **Correlation:** Dew_point_temp, Snowfall and temperature is highly correlated with each other. But we can drop dew_point_temp because it is less correlated with bike_count on comparison with temperature. As we know there is correlation exist between temperature, humidity and solar_radiation as temperature increase humidity increases. Due high humidity, or high rain fall or high snow fall the demand for bike sharing will increase we can analyze

this by observing above heat map that is these three are negatively correlated with bike_count. As we know there is correlation exist between temperature, humidity and solar_radiation as temperature increase humidity increases. Due high humidity, or high rain fall or high snow fall the demand for bike sharing will increase we can analyze this by observing above heat map that is these three are negatively correlated with bike_count. Using correlation metrics, we come to know that there few independent features that are highly correlated with each other like temperature and dew_point_temp. we will drop highly correlated independent feature using vif.

- **Checking linear relation of independent variable with target variable:** From 'Temperature', 'Wind_speed', 'Visibility', 'Dew_point_temp', 'Solar_Radiation', are linearly related in positive direction with demand for bike i.e bike count. 'Humidity' 'Rainfall' and 'Snowfall' is

linearly related but in negative direction with demand for bike i.e bike count.

- **Converting categorical columns into numerical columns.** categorical variables like seasons, holiday, function day and hour we change it with numerical database.
- **Splitting the dataset into training and testing dataset:** In train test split we take x as dependent variables and y take as independent variable then train the model.

Model building.

1.Linear Regression model:

Checking for assumption is satisfied our not:

- ❖ **Linear relation between independent and target variable:** linearly related with in positive direction with demand for bike are temperature, wind_speed, visibility, dew_point_temp, wind speed and solar radiation. 'Humidity' 'Rainfall' and 'Snowfall' are linearly related but in negative direction with bike count.

❖ **Multi-collinearity:**

Dew_point_temp, Snowfall and temperature is highly correlated with each other. But we can drop dew_point_temp because it is less correlated with bike_count on comparison with temperature. Using VIF we dropped highly correlated feature.

- ❖ The residuals follow almost normal distribution but mean of residuals close to zero.
- ❖ Assumption of homoscedasticity (equal distribution over the average line) is true here.

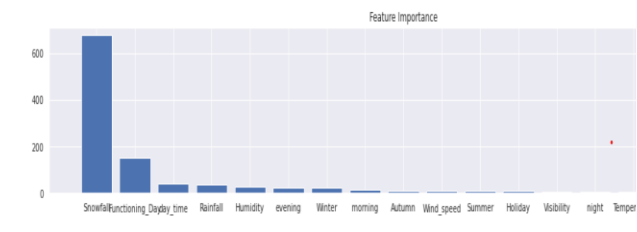
Model performance:

We train model by **linear regression** and we get results as follows:

Model R-Square of train set: 0.9177249
Model MSE of train set: 34060.9312
model rmse of train set:184.5560
Adjusted R2: 0.9175

Model R-Square of test set: 0.918720
Model MSE of test set: 34447.02666
model rmse of test set :185.5991
Adjusted R2: 0.9179

Feature importance obtain from linear regression model:



Decision Tree model:

Since this is non parametric algorithm. We used hyperparameter tuning to get best accuracy. We train model by **Decision Tree** and we get results as follows:

Model R-Square of train set: 0.9985
Model MSE of train set: 599.3153
model rmse of train set:24.4809
Adjusted R2: 0.9985

Model R-Square of test set: 0.9985
Model MSE of test set: 592.8859
model rmse of test set:24.3492
Adjusted R2: 0.9985

Feature importance obtain from Decision tree model:



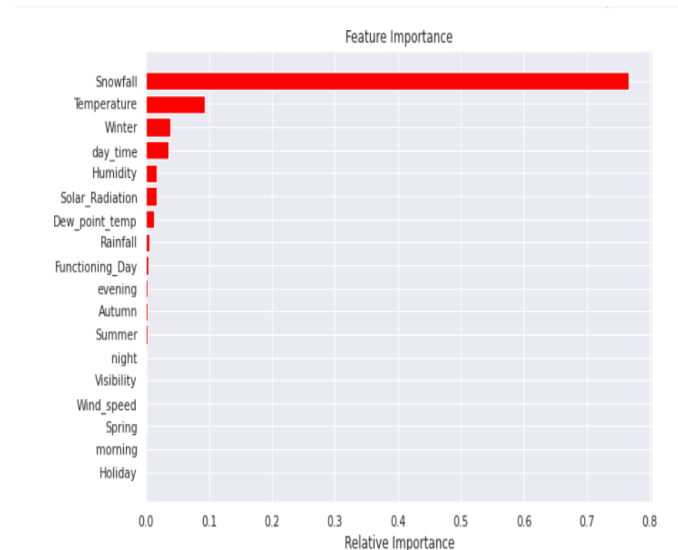
Random Forest model:

Since this is non parametric algorithm. We used hyperparameter tuning to get best accuracy. We train model by **Random regression** and we get results as follows:

Model R-Square of train set: 0.9738
Model MSE of train set: 10907.7438
model rmse of train set:104.44014
Adjusted R2: 0.97382

Model R-Square of test set: 0.97198
Model MSE of test set: 11442.37415
model rmse of test set:106.9690
Adjusted R2: 0.97169

Feature importance obtain from Random Forest model:



XGBoost:

We used hyperparameter tuning to get best accuracy. We train model by **XGBoost** and we get results as follows:

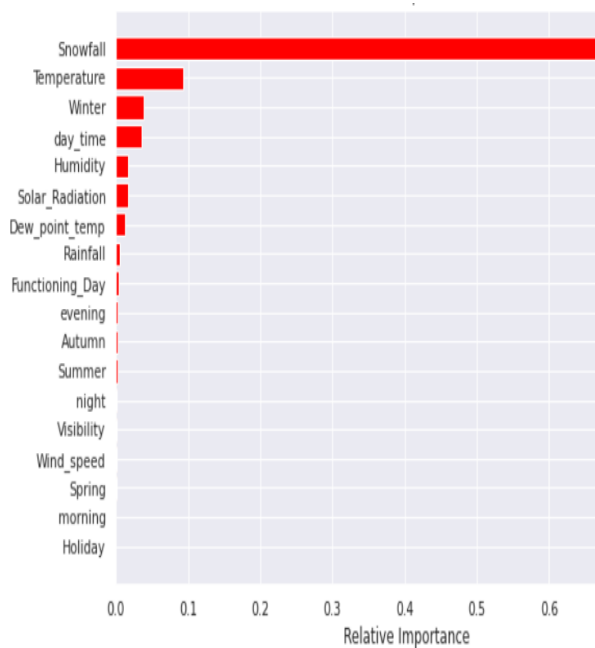
Model R-Square of train set: 0.999995
Model MSE of train set: 1.910359
model rmse of train set:1.382157
Adjusted R2: 0.9999954

Model R-Square of test set: 0.99998
Model MSE of test set: 5.07880
model rmse of test set:2.25362
Adjusted R 2: 0.999987

bike sharing at the evening, functioning is also highly another dominating factor.

The crucial part is the prediction of bike count we developed many models and got best accuracy with XGBoost algorithm. Having model R-Square of test set: 0.99998, Adjusted R2: 0.999987

Feature importance obtained from XGBoost:



Conclusion:

On analysis of data there are many factors affect the demand for bike sharing like weather (it includes temperature, snowfall, rain, visibility) is the major dominating feature, and time there is high demand for