**Bike Sharing Prediction**

**IRSHAD**

**Data science trainees,**

**Alma Better, Bangalore**

**Abstract:**

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand prediction. In recent days, Pubic rental bike sharing is becoming popular because of is increased comfortableness and environmental sustainability. Data used include Seoul Bike and Capital Bike share program data. Both data have weather data associated with it for each hour. For both the dataset, five statistical models were trained with optimized hyper parameters using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b) Regularized Random Forest (c) Classification and Regression Trees (d) K Nearest Neighboure Conditional Inference Tree. Multiple evaluation indices such as R2 , Root Mean Squared Error, Mean Absolute Error and Coefficient of Variation were used to measure the prediction performance of the regression models. The results show that the rule-based model CUBIST was able to explain about 95 and 89% of the Variance (R2 ) in the testing set of Seoul Bike data and Capital Bike share program data respectively. An analysis with variable importance was carried to analyse the most significant variables for all the models developed with the two datasets considered. The variable importance results have shown that Temperature and Hour of the day are the most influential variables in the hourly rental bike demand prediction.

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

This research paper presents a rule-based regression predictive model for bike sharing demand

prediction. In recent days, Pubic rental bike sharing is becoming popular because of is

increased comfortableness and environmental sustainability. Data used include Seoul Bike

and Capital Bikeshare program data. Both data have weather data associated with it for each

hour. For both the dataset, ﬁve statistical models were trained with optimized hyperparameters

using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b)

Regularized Random Forest (c) Classiﬁcation and Regression Trees (d) K Nearest Neighbour (e)

Conditional Inference Tree. Multiple evaluation indices such as R

2

, Root Mean Squared Error,

Mean Absolute Error and Coeﬃcient of Variation were used to measure the prediction perfor-

mance of the regression models. The results show that the rule-based model CUBIST was able

to explain about 95 and 89% of the Variance (R

2

) in the testing set of Seoul Bike data and

Capital Bikeshare program data respectively. An analysis with variable importance was carried

to analyse the most signiﬁcant variables for all the models developed with the two datasets

considered. The variable importance results have shown that Temperature and Hour of the day

are the most inﬂuential variables in the hourly rental bike demand prediction

**1.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.

**2. Introduction**

### The dataset contains weather information (Temperature, Humidity, Wind Speed, Visibility, Dew point , Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Many countries have bike sharing system, such as Dareungi is a bike sharing system in South Korea, which started in the year 2015, known as Seoul bike in English. It was started to overcome issues like greater oil prices, congestion in traffic and pollution in the environment and to develop a healthy environment for citizen of Seoul to live.

### Seoul city is now equipped with 1500 bike renting station which are operational round the clock. With the help of internet-enabled device or mobile phone, people can know the number of bikes available for the people to rent

## **3. Features depend on the bike sharing demand**

* **Temperature(°C)**
* **Humidity (%)**
* **Rainfall**
* **Seasons**
* **Holiday**
* **Functioning day**

## **4. Benefits for bike sharing**

The benefits of bike sharing schemes include transport flexibility, reductions to vehicle emissions, health benefits, reduced congestion and fuel consumption, and financial savings for individuals

* Improved Air Quality. Bike share systems are great for the environment.
* Convenience. Bike share systems are most definitely convenient. ...
* Healthier Population
* No Helmets carry problem

# **5. How bike sharing demand works**

## Using these Bike Sharing systems, people rent a bike from one location and return it to a different or same place on need basis. People can rent a bike through membership (mostly regular users) or on demand basis (mostly casual users). This process is controlled by a network of automated kiosk across the city.

Bike sharing relies on a system of self-service bike stations. Users typically check out a bike using a membership or credit/debit card. They can then ride to their destination and park the bike in a nearby docking station.

**6. Steps involved:**

* **Exploratory Data Analysis**

After loading the data set we performed this method by comparing our target variable that is **Rented Bike Count** with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

* **Null values Treatment**

Our dataset may be contains a large number of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

In our data set there is no null values present in it.

* **Encoding of categorical columns**

We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

* **Feature Selection**

In these steps we used algorithms like Linear regression , lasso regression, ridge regression, decision tree, random forest etc. to check the results of each feature i.e which feature is more important compared to our model and which is of less importance.

* **Standardization of features**

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

* **Fitting different models**

For modelling we tried various algorithms like:

1. **Logistic Regression**
2. **Lasso Regression**
3. **Ridge Regression**
4. **Elastic Net**
5. **Random Forest Classifier**
6. **Decision Tree**
7. **Gradient Boosting**

* **Tuning the hyper parameters for better accuracy**

Tuning the hyper parameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree based models

like Random Forest Classifier..

**7.1. Algorithms:**

1. **Logistic Regression:**

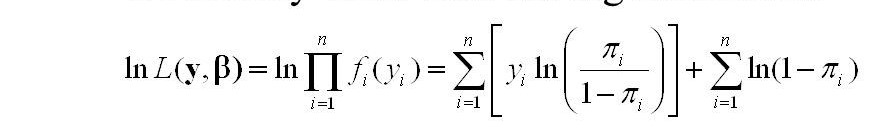
Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.

The function used in Logistic Regression is sigmoid function or the logistic function given by:

f(x)= 1/1+e ^(-x)



The optimization algorithm used is: Maximum Log Likelihood. We mostly take log likelihood in Logistic:

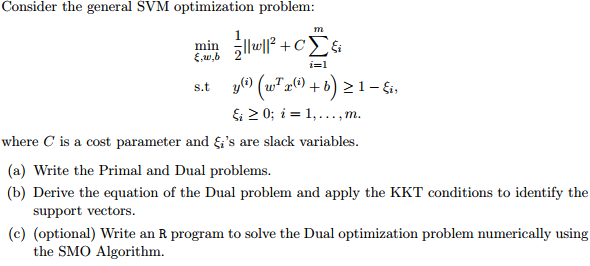


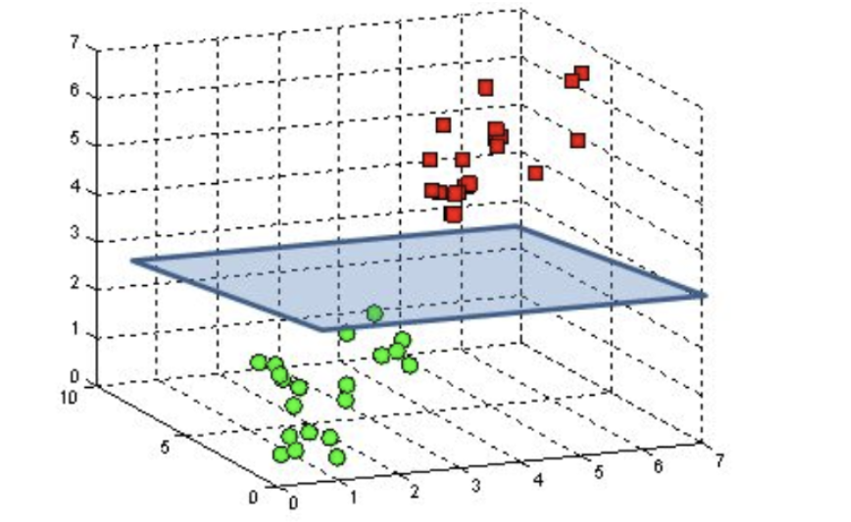
.

1. **Support Vector Machine Classifier:**

SVM is used mostly when the data cannot be linearly separated by logistic regression and the data has noise. This can be done by separating the data with a hyperplane at a higher order dimension.

In SVM we use the optimization algorithm as:





We use hinge loss to deal with the noise when the data isn’t linearly separable.

Kernel functions can be used to map data to higher dimensions when there is inherent non linearity.

1. **Random Forest Regressor:**

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single mode



1. **XG Boost-**

To understand XG Boost we have to know gradient boosting beforehand.

* **Gradient Boosting-**

Gradient boosted trees consider the special case where the simple model is a decision tree

****

In this case, there are going to be 2 kinds of parameters P: the weights at each leaf, w, and the number of leaves T in each tree (so that in the above example, T=3 and w=[2, 0.1, -1]).

When building a decision tree, a challenge is to decide how to split a current leaf. For instance, in the above image, how could I add another layer to the (age > 15) leaf? A ‘greedy’ way to do this is to consider every possible split on the remaining features (so, gender and occupation), and calculate the new loss for each split; you could then pick the tree which most reduces your loss.

**XG Boost** is one of the fastest implementations of gradient boosting. trees. It does this by tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits). XG Boost tackles this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits.

**7.2. Model performance:**

Model can be evaluated by various metrics such as:

1. **Confusion Matrix**-

The confusion matrix is a table that summarizes how successful the classification model is at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

1. **Precision/Recall**-

Precision is the ratio of correct positive predictions to the overall number of positive predictions : TP/TP+FP

Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP

1. **Accuracy**-

Accuracy is given by the number of correctly classified examples divided by the total number

of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

1. **Area under ROC Curve(AUC)**-

ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance.

**7.3. Hyper parameter tuning:**

Hyper parameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyper parameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyper parameters grid that can be adjusted according to the business problem. Hyper parameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV, Randomized Search CV and Bayesian Optimization for hyper parameter tuning. This also results in cross validation and in our case we divided the dataset into different folds. The best performance improvement among the three was by Bayesian Optimization.

1. **Grid Search CV-**Grid Search combines a selection of hyper parameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.
2. **Randomized Search CV-** In Random Search, the hyper parameters are chosen at random within a range of values that it can assume. The advantage of this method is that there is a greater chance of finding regions of the cost minimization space with more suitable hyper parameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyper parameters is beyond the scientist’s control

**8. Conclusion:**

During the time of our analysis, we initially did EDA on all the features of our datset. We first analysed our dependent variable, 'Rented Bike Count' and also transformed it. Next we analysed categorical variable and dropped the variable who had majority of one class, we also analysed numerical variable, found out the correlation, distribution and their relationship with the dependent variable. We also removed some numerical features who had mostly 0 values and hot encoded the categorical variables.

Next we implemented 7 machine learning algorithms Linear Regression, lasso, ridge, elastic net, decision tree, Random Forest and XG Boost. We did hyper parameter tuning to improve our model performance. The results of our evaluation are: