# Bike Sharing Demand Prediction

Shrinidhi Choragi
Data Science Trainee, AlmaBetter
Bangalore.

## Abstract

Bike rental predictions forecast the demand for bike rentals in dependency on weather conditions like the temperature and calendric information like holidays and functional days. To make predictions machine learning is used. Generally, in bike rental systems, the administrators must know how many bikes are to be placed in each station, knowing this count enables them in arranging the required number of bikes at the stations and decide whether a particular station needs to have the extra number of bikes or not. Therefore the following study on prediction associates to enhance their administration given clients' input.

Regression analysis is explored to achieve the same. It is a statistical technique for determining the relationship between a single dependent (criterion) variable and one or more independent (predictor) variables. The analysis yields a predicted value for the criterion resulting from a linear combination of the predictors.
The process of regression analysis for prediction helps to understand this type of study more easily.

*Keywords*: *machine learning*, *prediction*, *regression.*

## Problem Statement

The objective of this project is to predict bike rental count/ forecast bike rental demand required at each hour based on bike usage patterns with the environmental and seasonal data history.
It is a regression problem.

Some of the questions to be explored through this study:

- What is the relation between the features and the bike rental count?
- Which regressive model gives the most optimum predictions?
- What features influence the most in predicting the bike rental count?

## Introduction

Several bike/scooter ride-sharing facilities (e.g., Rapido, Bird, Capital Bikeshare, CitiBike) have started up lately, especially in metropolitan cities like San Francisco, New York, Chicago, and Los Angeles.

One of the most important problems from a business point of view is to predict the bike demand on any particular day. While having excess bikes results in wastage of resources (both concerning bike maintenance and the land/bike stand required for parking and security), having fewer bikes leads to revenue loss (ranging from a short-term loss due to missing out on immediate customers to potential long-term loss due to loss in future customer base). Thus, having an estimate of the demands would enable the efficient functioning of these companies.

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 and provides many alternatives to commuters in metropolises. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The

majority of bicycle rental businesses are clustered around heavily trafficked tourist spots. However, with increased rails-to-trails projects and traffic congestion, there are many more bicycle paths away from resort areas, office space, and residential areas that are creating excellent new rental opportunities.

There are a lot of advantages to bike rental, it is convenient because it permits people not to keep the bike all day long, whether it is at work or school. Furthermore, it is the healthiest way to travel and it has many environmental benefits. The bike rental service has great potential as a business opportunity.

Therefore this study on the prediction model helps in enhancing the administration of rental services.

## Dataset

The dataset contains **8760** observations, **13** predictors, and a target variable 'Rented Bike Count'.

The predictors/features describe various environmental factors and weather information factors like Temperature, Humidity, Wind speed, Visibility, Dew point temperature, Solar radiation, Snowfall, and Rainfall, and information regarding holidays, functioning days, and dates information. The target variable is bikes that are rented per hour as a function of weather conditions. The dataset presents the company's data between December the 1st of 2017 and finishes one year later. The goal of the company Seoul Bike is to provide the city with a stable supply of rental bikes. It becomes a major concern to keep users satisfied.

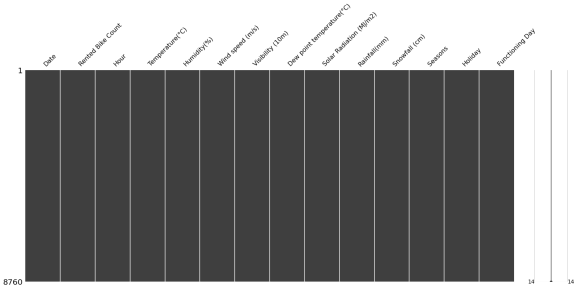
## Methodology

Exploratory Data Analysis

Data pre-processing gives the feel of the data. If the data is messy, it is improved by sorting and deleting extra rows and columns. This stage generally involves data cleaning, merging, sorting, looking for outliers, looking for missing values in the data, and imputing missing values.

1. Missing Value Analysis

Missing value analysis was performed on the dataset. No missing values were found.




2. Outlier Analysis

There were outliers present in columns namely:

- Rented Bike Count (target variable)

The outliers of the target variable were treated using ***Square Root Transformation***. It also converts the *rightly skewed* data to *normally distributed* data.

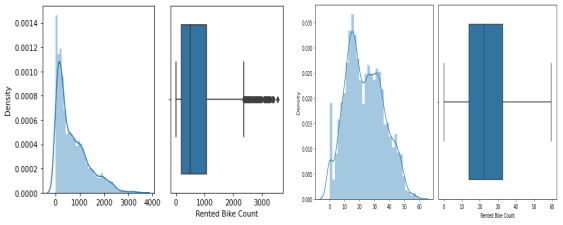



**Figure: The density distribution and boxplot for target variable before and after square root transformation(Outlier Treatment).**

- Features- WindSpeed, Rainfall, Snowfall, Solar radiation.

The outliers found in independent numerical variables were treated using a ***Robust scaler***, during data modeling.

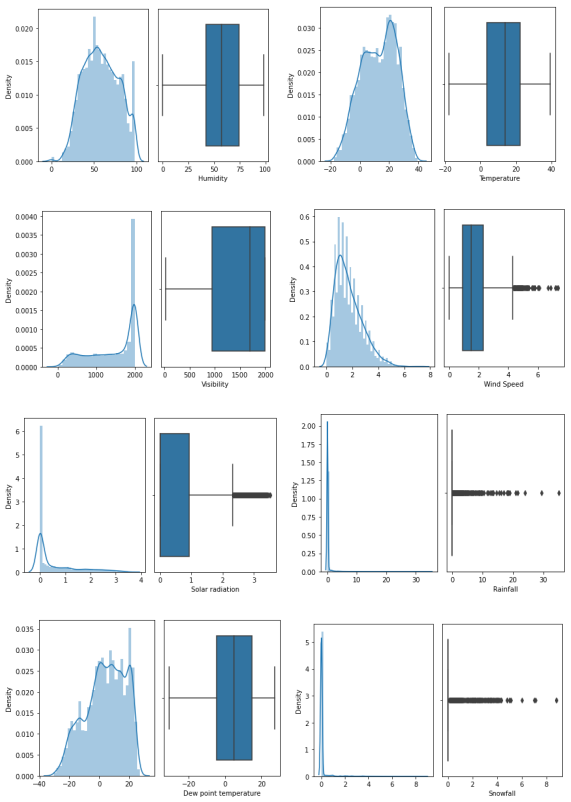



**Figure: The density distribution and boxplot for numerical features- Humidity, Temperature, Visibility, Wind speed, Solar radiation, Rainfall, Dew point temperature, Snowfall.**

3. <u>Correlation Analysis</u>

This requires only numerical variables. There should be no correlation between independent variables but a high correlation between independent and dependent variables. So, the correlation matrix shows the variables and their correlation.

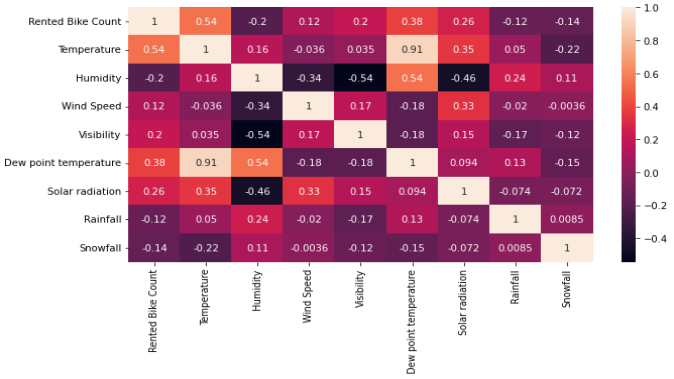



**Figure: Correlation matrix**

The correlation can be visualized in the regression plots.

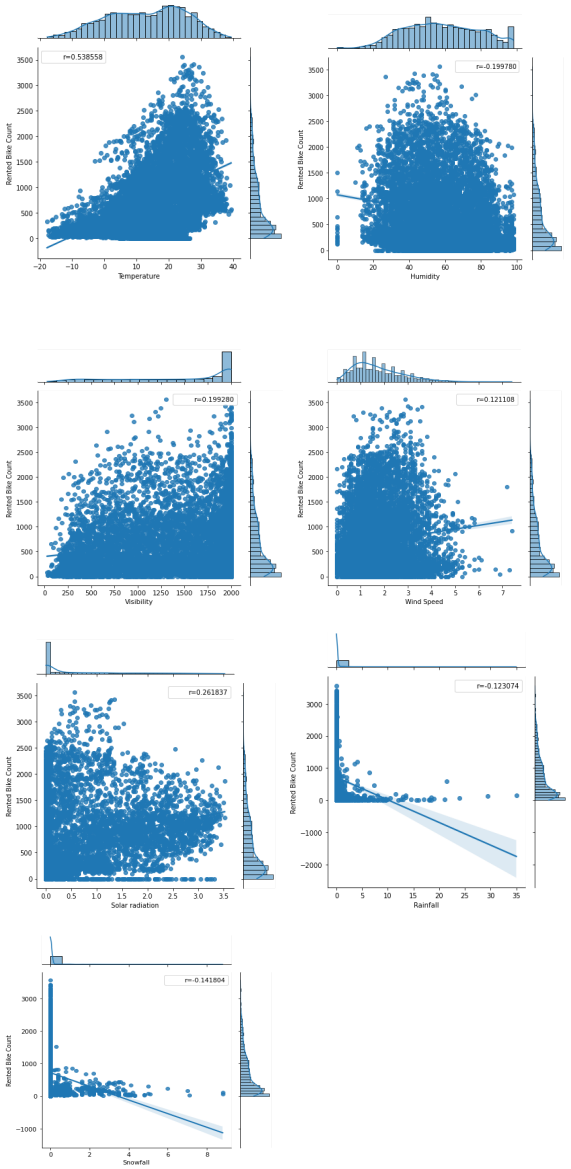


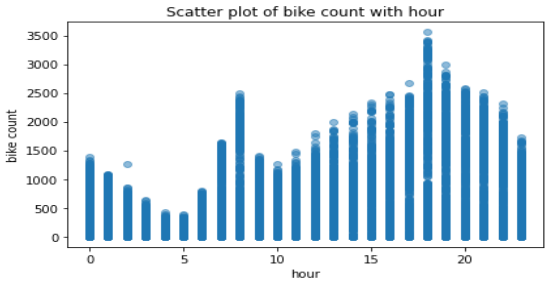

**Figure: Regression plot with correlation coefficient for numerical features against target variable.**

- The temperature correlates (0.54) with the count of bike rents.
- Temperature and dew-point temperature are highly correlated. One of the features could be dropped later.
- Humidity is positively correlated(0.54) with dew-point temperature as much as it is negatively correlated(-0.54) with Visibility.
- Solar Radiation has a slight negative correlation with Humidity.

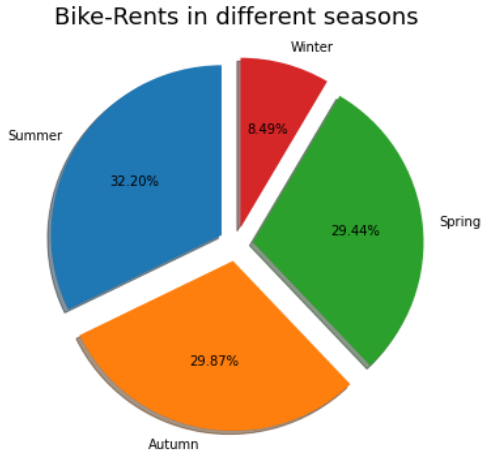
4. Variable Analysis

*Bike Rental Count Analysis: Count v/s Hour of the day*

- Bike rents tend to be the highest during the peak hours like 8:00 am and 6:00 pm.

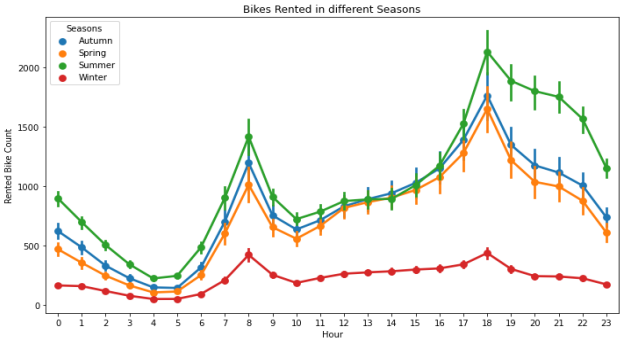



*Bike Rental Count Analysis: In different seasons*



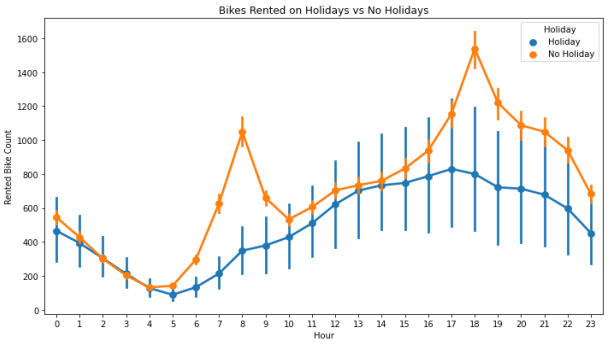

- Bike rental count is highest during summers indicating the ideal temperature for use of bikes.

*Bike Rental Count Analysis: Throughout the day for all seasons*



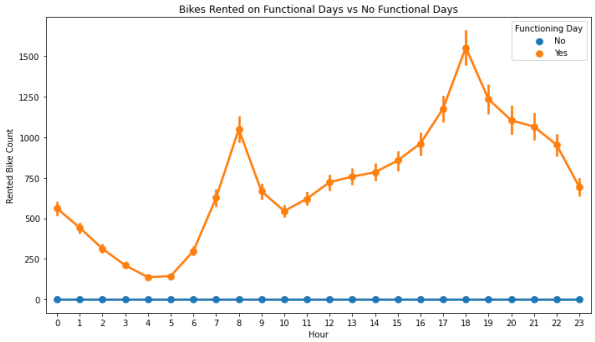

- Winter has comparatively lesser bike-rents
- 8:00 am and 6:00 pm are the peak hours of bike rent in all seasons.
- Bike rents in the morning are lesser than that in the evening
- Summers have the highest bike rents.

*Bike Rental Count Analysis: Holiday and No Holiday*




- On No Holidays, bike rents start around 5:00 am and are highest during peak hours.
- On Holidays the bike rents are casually increasing from 8:00 am to 5:00 pm and decrease afterward.

*Bike Rental Count Analysis: Functional Day and No Functional Day*




- No bike rents were observed on No Functional days.
- Bike rents follow a regular trend on functional days.

What is the regression analysis?

To establish the possible relationship among different variables(features), various modes of statistical approaches are implemented, known as *regression analysis*. To understand how the variation in an independent variable can impact the dependent variable, regression analysis is specially designed.

*Regression analysis sets up an equation to explain the significant relationship between one or more predictors and response variables and also to estimate current observations where the analytical significance of the relationship between predictor and the dependent variable is derived.*

Assumptions of a Regression Model.

1. There should be a linear and additive relationship between the dependent variable and the independent (predictor) variable(s). An additive relationship suggests that the effect of $X^1$ on Y is independent of other variables.

2. No Autocorrelation: There should be no correlation between the residual (error) terms. Autocorrelation is the presence of correlation in error terms that drastically reduces the model's accuracy. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.

3. No Multicollinearity: There shouldn't be a correlation between independent variables. When the independent variables are highly correlated to each other, then variables are said to possess multicollinearity.

It is a common assumption that is tested before selecting the variables for the regression model.

Why Multi-Collinearity is a problem?

- When independent variables are highly correlated, change in one variable would cause change to another and so the model results fluctuate significantly.
- The unstable nature of the model may cause overfitting.

Collinearity issue is considered severe if correlation > 0.8 between any two variables or Variance inflation factor(VIF) > 5.

Variance Inflation Factor(VIF) is used to check multicollinearity for each independent variable. It is a measure of multicollinearity in the set of multiple regression variables. The higher the value of VIF the higher correlation between these variables.

4. Homoscedasticity: The error terms must have constant variance. When the variation in the dependent variables is not even crosswise the values of independent variables result in heteroscedasticity.

If heteroscedasticity exists, the residual vs fitted values plot would exhibit a funnel shape pattern.

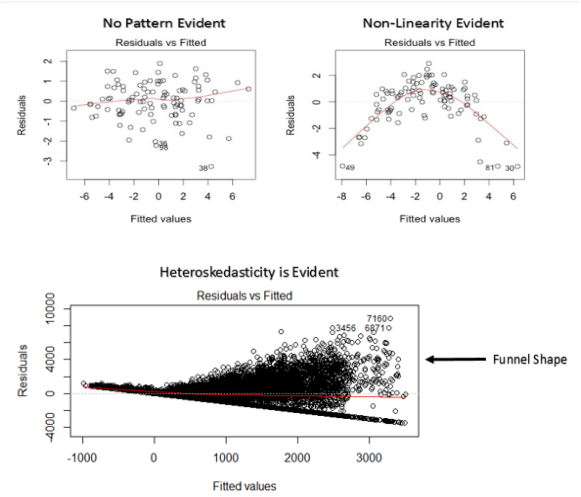



**Figure: The scatter-plot shows the distribution of residuals (errors) vs fitted values (predicted values).**

It reveals various useful insights including outliers. The outliers in this plot are labeled by their observation number which makes them easy to detect. Two major things are to be learned:

- If there exists any pattern (maybe, a parabolic shape) in this plot, consider it as a sign of non-linearity in the data. It means that the model doesn't capture non-linear effects.
- If a funnel shape is evident in the plot, consider it as the sign of non-constant variance i.e. heteroscedasticity.

To overcome the issue of non-linearity, a non-linear transformation of predictors such as log (X), $\sqrt{X}$, or $X^2$ to transform the dependent variable is done. To overcome heteroscedasticity, possible ways are to transform the response variable such as log(Y) or $\sqrt{Y}$ or apply the weighted least square method.

5. The error terms must be normally distributed. If the error terms are non-normally distributed, confidence intervals may become too wide or narrow. Once the confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on the minimization of least squares. The presence of non–normal distribution suggests that there are a few unusual data points that must be studied closely to make a better model.

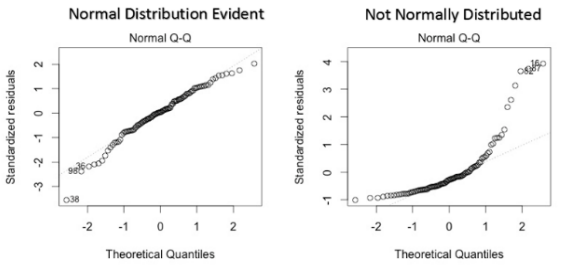


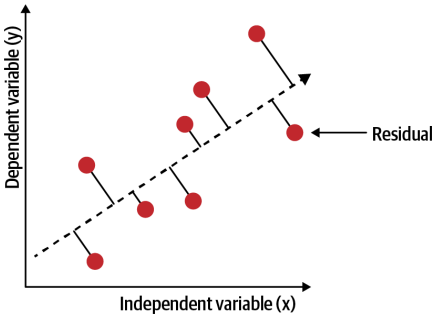

**Figure: The q-q or quantile-quantile is a scatter plot that helps us validate the assumption of normal distribution in a dataset.**

Using this plot we can infer if the data comes from a normal distribution. If yes, the plot would show a fairly straight line. The absence of normality in the errors can be seen with deviation in the straight line.

## Regression Algorithms

Types of regression analysis can be selected on the attributes, target variables, or the shape and nature of the regression curve that exhibit the relationship between dependent and independent variables.

### 1. ***Linear Regression***




It is the most straightforward regression technique used for predictive analysis, a linear approach for featuring the relationship between the response and predictors or descriptive variables.

Although, linear regression faces the issue of overfitting, and it shows a best-fitted straight

line (technically a hyperplane depending on the number of features) that possesses an equation:

**Y = *b*X+C,**

where Y is a dependent variable;
X is the independent variable;
b as the slope of the line;
C as intercept.

Since our goal is to build a model that can make accurate predictions, simple linear regression suffers from two major flaws:

- It's prone to overfit with many input features.
- It cannot easily express non-linear relationships.

Suppose the number of features is more for the observations. In that case, the coefficients memorize the observations and the model would have perfect accuracy on the training data but perform poorly on unseen data. It hasn't learned the true underlying patterns having memorized the noise in the training data.

## ***2. Regularized Regression***

Regularization is a technique used to prevent overfitting by artificially penalizing model coefficients. It can discourage large coefficients (by dampening them). It can also remove features entirely (by setting their coefficients to 0). The "strength" of the penalty is tunable.

### ***a. Ridge regression***

Ridge Regression penalizes the squared size of coefficients. Practically, this leads to smaller coefficients, but it doesn't force them to 0. In other words, Ridge offers ***feature shrinkage***. Again, the "strength" of the penalty should be tuned. A stronger penalty leads to coefficients pushed closer to zero.
In simple words, sometimes the regression model becomes too complex and approaches overfitting, so it is worthwhile to minimize the variance in the model and save it from overfitting by correcting the size of the coefficients.

### ***b. Lasso regression***

LASSO, stands for Least Absolute Shrinkage and Selection Operator.
Lasso regression penalizes the absolute size of coefficients. Practically, this leads to coefficients that can be exactly 0. Thus, Lasso offers automatic **feature selection** because it can completely remove some features. The "strength" of the penalty should be tuned. A stronger penalty leads to more coefficients pushed to zero. It is a widely used regression analysis to perform both variable selection and regularization.

### ***c. ElasticNet Regression***

Elastic-Net is a compromise between Lasso and Ridge.
It is the mixture of ridge and lasso regression that brings out a grouping effect when highly correlated predictors approach to be in or out in the model combinedly. It is recommended to be used when the number of predictors is greater than the number of observations.
Elastic-Net penalizes a mix of both absolute and squared size. The ratio of the two penalty types should be tuned.

## ***3. Decision Trees***

Decision trees model data as a "tree" of hierarchical branches. They make branches until they reach "leaves" that represent predictions.

Due to their branching structure, decision trees can easily model nonlinear relationships. Unfortunately, decision trees suffer from a major flaw as well. If you allow them to grow limitlessly, they can completely "memorize" the training data, just by creating more and more branches. As a result, individual unconstrained decision trees are very prone to overfitting.

**Ensemble Models**

The ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

- Bagging

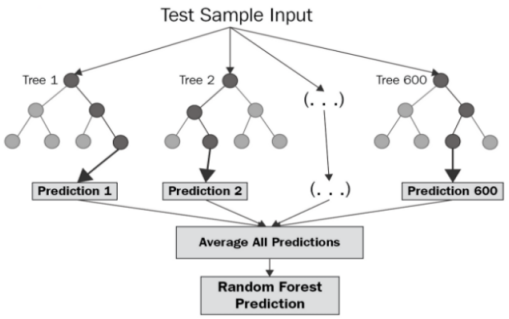
Also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once. After several data samples are generated, these weak models are then trained independently to form a strong model.

- Boosting

It is an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors. In boosting, a random sample of data is selected, fitted with a model, and then trained sequentially—that is, each model tries to compensate for the weaknesses of its predecessor. With each iteration, the weak rules from each model are combined to form one, strict prediction rule.

***Bagging:***

***Random Forest Regressor***




Random Forest Regression is a supervised learning algorithm that uses the ensemble learning method for regression. A Random Forest operates by constructing several decision trees during training time and outputting the mean of them as the prediction of all the trees.

The process of Random Forest is summarized as follows:

- Pick at random $k$ data points from the training set.
- Build a decision tree associated with these $k$ data points.
- Choose the number N of trees you want to build and repeat steps 1 and 2.
- For a new data point, make each one of your N-tree trees predict the value of $y$ for the data point in question and assign the new data point to the average across all of the predicted $y$ values.

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include no interpretability, overfitting may easily occur, and the number of trees to include in the model is to be chosen.

***Boosting:***

***a: Gradient Boosting Regressor***

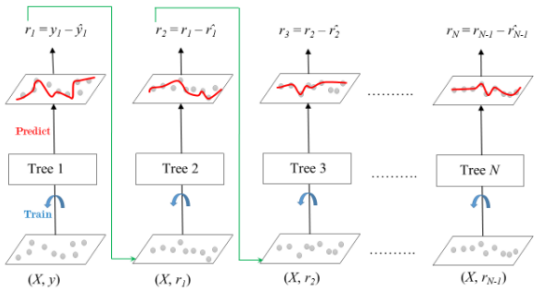



**Figure: Gradied Boosting Procedure**

In gradient boosting, each predictor corrects its predecessor's error.

The ensemble consists of N trees. Tree1 is trained using the feature matrix X and the labels y. The predictions labeled y1' are used to determine the training set residual errors r1. Tree2 is then trained using the feature matrix X and the residual errors r1 of Tree1 as labels. The predicted results' r1' are then used to determine the residual r2. The process is repeated until all the N trees forming the ensemble are trained. There is an important parameter used in this technique known as **Shrinkage**.

It refers to the fact that the prediction of each tree in the ensemble is shrunk after it is multiplied by the learning rate (eta) which ranges between 0 to 1. There is a trade-off between eta and number of estimators, decreasing learning rate needs to be compensated with increasing estimators in order to reach certain model performance. Since all trees are trained now, predictions can be made.

Each tree predicts a label and final prediction is given by the formula:

***y(pred) = y1 + (eta * r1) + (eta * r2) + ..... + (eta * rN)***

Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, "*gradient boosting*," as the loss gradient is minimized as the model is fit, much like a neural network.

***b: Light Gradient Boosting***

LightGBM, short for Light Gradient Boosted Machine, is a library developed at Microsoft that provides an efficient implementation of the gradient boosting algorithm.

The primary benefit of the LightGBM is the changes it does to the training algorithm that make the process dramatically faster, and in many cases, result in a more effective model.

***c: CatBoost Regressor***

CatBoost is a third-party library developed at Yandex that provides an efficient implementation of the gradient boosting algorithm.

The primary benefit of the CatBoost (in addition to computational speed improvements) is support for categorical input variables. This gives the library its name CatBoost for "*Category Gradient Boosting*."

**Data Preparation and Modeling**

**Feature Selection**: Multicollinearity Test

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

The higher the value of VIF the higher correlation between this variable and the rest. Following the points below the features are selected based on correlation and VIF values.

- VIF is always greater or equal to 1.
- If VIF = 1 ➪ Not correlated to any of the variables.
- If 1 < VIF < 5 ➪ Moderately correlated.
- VIF > 5 ➪ Highly correlated.
- If there are multiple variables with VIF greater than 5, then remove one of them and repeat the process.

**Encoding categorical columns**

One Hot Encoding is used to produce binary integers of 0 and 1 to encode our categorical features.
The categorical features namely season, hour, month, holiday, and functioning day are encoded.

**Data Splitting**

The dataset is split into train and test data in the ratio of 75:25 resp. The model is fit on the training data and tested on the newest data. Sklearn's *train_test_split* is used. The *random state* hyperparameter in the train_test_split() function controls the shuffling process, and checks and validates the data when running the code multiple times. Setting random_state to a fixed value will guarantee that the same sequence of random numbers is generated each time you run the code.
The resulting dataset is subjected to scaling.
Dataset now consists:
Train: *6560 observations and 51 features.*
Test: *2190 observations and 51 features.*

**Feature scaling**

Feature scaling is a method used to normalize the range of independent variables or features of data. It is also known as data normalization and is generally performed during the data preprocessing step.

The concept of standardization or standard coefficients come into the picture when independent variables or predictor for a particular model are expressed in different units. Here, we have independent features namely Temperature(°C), Humidity(%), Wind speed (m/s), Visibility (10m), Dew point temperature(°C), Solar Radiation (MJ/m2), Rainfall(mm), and Snowfall (cm) with respective units.

It would not be a fair comparison to rank these predictors based on unstandardized coefficients (that are obtained through the regression model) since the units for all the predictors are different.
Also, because regularization techniques manipulate the value of the coefficients, this makes the model performance sensitive to the scale of features. Therefore, features should be transformed to the same scale.
The data is initially scaled with ***Robust Scaler*** to handle the outliers in the features and then subjected to ***Minmax Scaler*** to normalize the data.
It is important to note that the scaler is *fit using the training set only and then applying the transform to both the training and testing set*. So, the dataset should be split first.

**Model fitting**

The following models have been studied and implemented on the given dataset:

- Linear Regression
- Regularized Regression
    - Lasso Regression
    - Ridge Regression
    - Elastic Net Regression
- Decision Tree regression
- Random Forest Regression
- Gradient Boosting Regression
- Light Gradient Boosting Regression
- CatBoost Regression

**Hyperparameter Tuning**

Hyperparameters are parameters that are explicitly specified and control the training process. Model optimization necessitates the use of hyperparameters.

Cross Validation is explored in the project. The k-fold cross-validation method implies performing cross-validation, where the input data is split into k subsets of data. The ML model is trained on all but one (k-1) of the subsets, and then evaluates the model on the subset that was not used for training. This process is repeated k times, with a different subset reserved for evaluation (and excluded from training) each time.

To avoid overfitting, there is a need for distinct data for training and assessing the model.

GridSearch is a method for performing hyperparameter optimization.

## Evaluation Metrics and Visualizations

**R-Square**

$R^2$ score tells us how well our model is fitted to the data by comparing it to the average line of the dependent variable. $R^2$ measures, "How much the change in output variable (y) is explained by the change in input variable(x).

$$\text{R-Square} = 1 - \frac{\sum(Y\_actual - Y\_predicted)^2}{\sum(Y\_actual - Y\_mean)^2}$$

It is always between 0 and 1. In general, the higher the $R^2$, the more robust will be the model. One disadvantage of R-squared is that it can only increase as predictors are added to the regression model. To cure this, we use "Adjusted R-squared".

**Adjusted R-squared**

Adjusted R square calculates the proportion of the variation in the dependent variable accounted by the explanatory variables. It incorporates the model's degrees of freedom. It decreases as predictors are added if the increase in model fit does not make up for the loss of degrees of freedom. Likewise, it will increase as predictors are added if the increase in model fit is worthwhile. Adjusted R-squared should always be used with models with more than one predictor variable. It is interpreted as the proportion of total variance that is explained by the model.

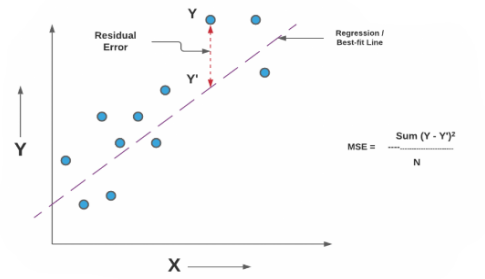
$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

where:
n = number of observations
k = number of independent variables
$R_a^2$ = adjusted $R^2$

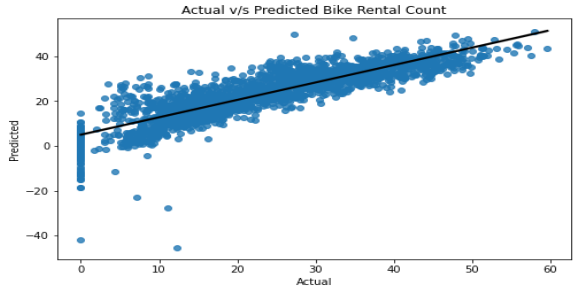
**MSE-Mean Square Error**

The Mean Squared Error measures how close a line is to a set of data points. Mean square error is calculated by taking the average, specifically the mean, of errors squared.

The lesser the MSE ➪ Smaller is the error ➪ Better the estimator.



## Scatter Plot




Scatter plots of *Actual vs Predicted* are one of the richest forms of data visualization. Ideally, all the points should be close to a regressed diagonal line.
The graph displays the actual value of the target on the x-axis and the value of the target predicted by the model on the y-axis.

If the model was perfect, all points would be on the diagonal line. This would mean that predicted values are exactly equal to actual values. The error on a data point is the vertical distance between the point and the diagonal.

## Residual Plot

The presence of non-constant variance in the error terms results in heteroscedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Looks like, these values get too much weight, thereby disproportionately influencing the model's performance. When this phenomenon occurs, the confidence interval for out-of-sample prediction tends to be unrealistically wide or narrow.

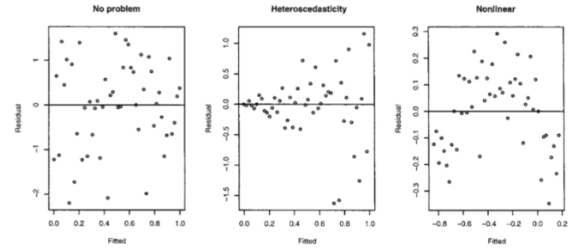
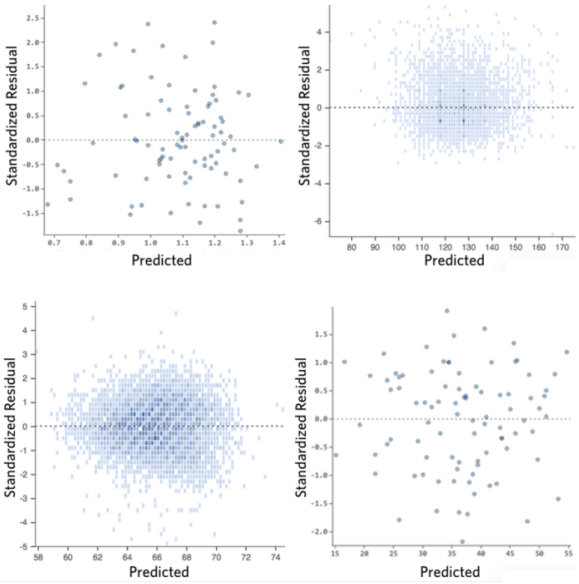



**Figure: Types of Residual vs Fitted value plots**

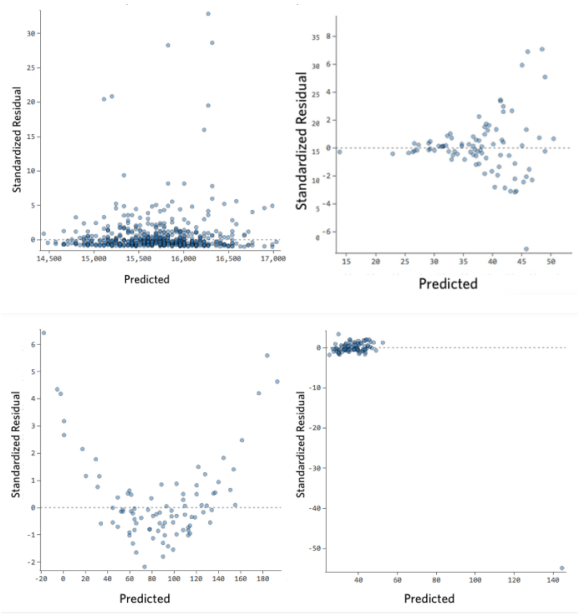
The first plot suggests no change to the current model required since it seems to indicate that the residuals and the fitted values are uncorrelated, as they should be in a homoscedastic linear model with normally distributed errors. The second shows non-constant variance and the third indicates some non-linearity, which should prompt some change in the structural form of the model.
Therefore, the second and third plots, which seem to indicate dependency between the residuals and the fitted values, suggest a different model.

Ideally, the plot of the residuals looks like one of these:




- They are pretty symmetrically distributed, tending to cluster towards the middle of the plot.
- They're clustered around the lower single digits of the y-axis.
- In general, there aren't any clear patterns.

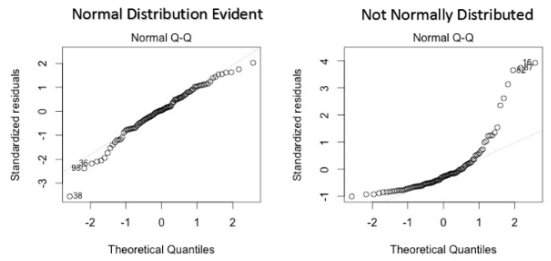
Here are some residual plots that don't meet those requirements:




These plots aren't evenly distributed vertically, or they have an outlier, or they have a clear shape to them.

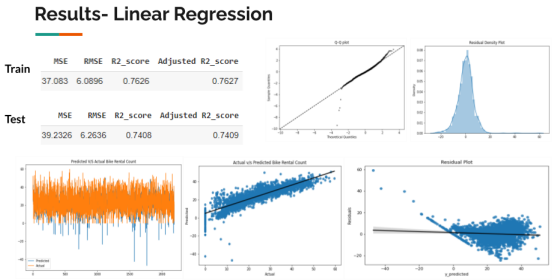
If a clear pattern or trend is detected in the residuals, then the model has room for improvement.

## Q-Q Residual plot




This q-q or quantile-quantile is a scatter plot which helps us validate the assumption of normal distribution in a data set. Using this plot we can infer if the data comes from a normal distribution. If yes, the plot would show fairly straight line that aligns closely to the dotted 45° line. Absence of normality in the errors can be seen with deviation in the straight line that skew drastically from the line.

## Results




* No overfitting is observed.
* Performance improvement can be expected.
* Heteroscedasticity is observed from residual plot.
* The density distribution of residuals show that residuals are normally distributed, that is validated by the straight line in q-q residual plot.

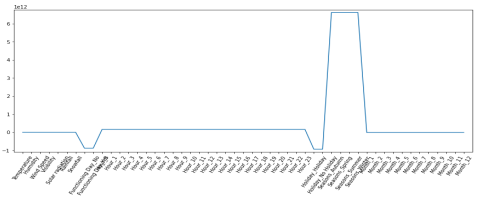
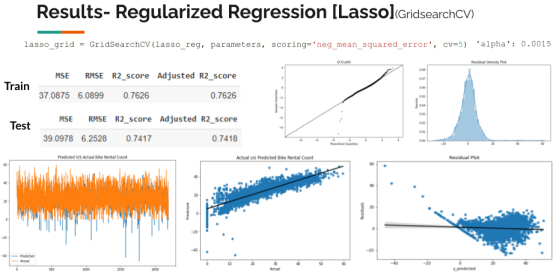


**Figure: Linear Regression Coefficients plot.**

* Few features have high estimated values of coefficients value compared to that of others.
* The larger the absolute value of the coefficient, the more power it has to change the predicted response, resulting in a higher variance.




* No prominent improvement in performance.

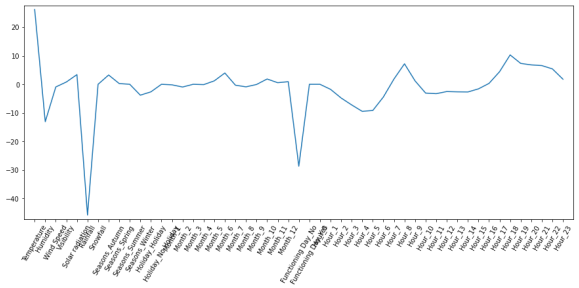
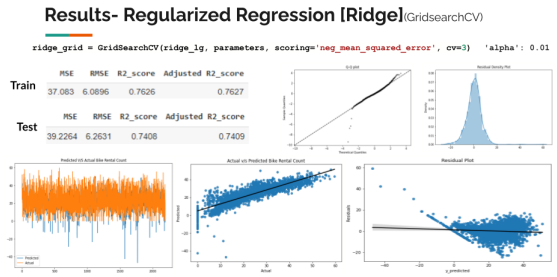


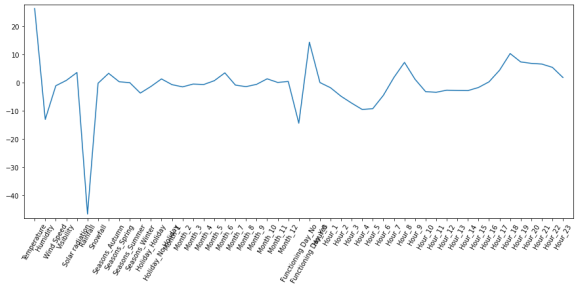
**Figure: Lasso Regression Coefficients plot.**

* According to the coefficients the features namely, *humidity*, *rainfall*, *and functioning day_no* have a negative relation with the dependent variable, indicating that high rainfall, high humidity, and non-functioning hours are mostly not favorable for bike rents.

* Whereas, the features namely *hour_8 and hour_18* have a high positive relation to the dependent variable, indicating that the peak hours demand a high number of bike rents.

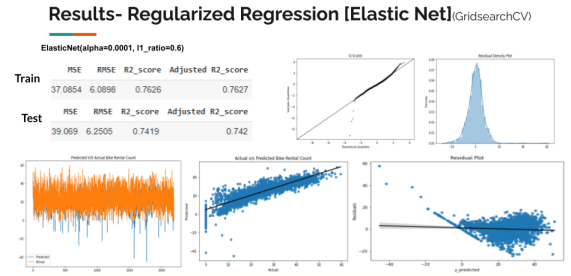



* No prominent improvement in performance.

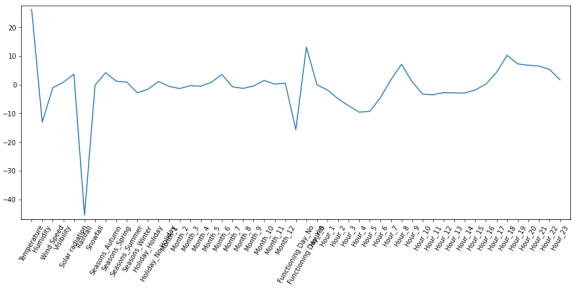


* According to the coefficients of the Ridge Regression model, the features namely, *humidity*, *rainfall*, *and functioning day_no* have a negative relation with the dependent variable, indicating that high rainfall, high humidity, and non-functioning hours are mostly not favorable for bike rents.

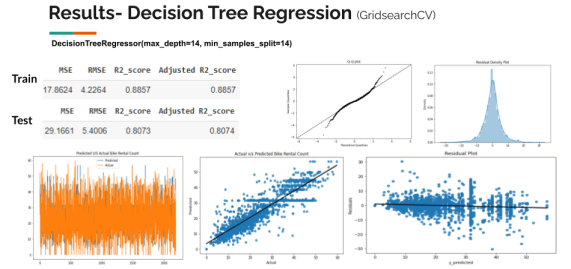
* Whereas, the features namely *hour_8, hour_18, and functional day_yes* have a high positive relation with the dependent variable, indicating that the peak hours and functioning hours demand a high number of bike rents.




* No prominent improvement in performance.



* The coefficients are almost similar to that of the ridge regression.




* There is a significant improvement in the performance.
* Model is overfitted.
* The residuals are normally distributed and no heteroscedasticity is observed.

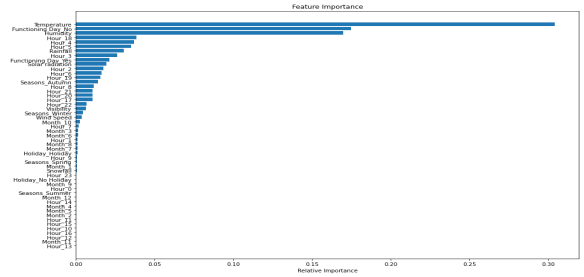
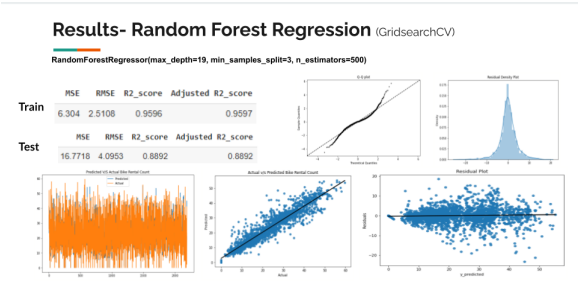



**Figure: Feature importance as predicted by decision tree regressor**

* Most important features in the prediction include: *Temperature*, *Functioning Day_yes*, *Humidity*, *Hour_18*, *Hour_4*, *Hour_5*, *Rainfall*, and so on.




* There is further improvement in the performance.
* Residuals are normally distributed.
* There exists homoscedasticity. The residual plot exhibits the ideal performance.
* Points are pretty symmetrically distributed, tending to cluster towards the middle of the plot without any clear patterns.

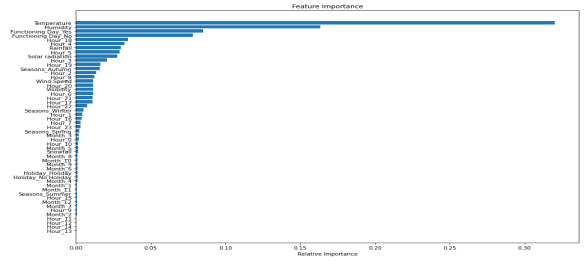
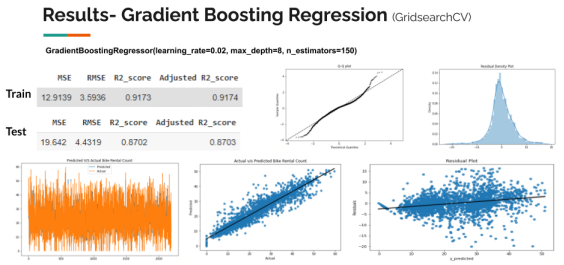



**Figure: Feature importance as predicted by Random Forest regressor**

* Most important features in the prediction include *Temperature*, *Humidity*, *Functioning Day_No*, *Functioning Day_Yes*, *Hour_18*, *Hour_4*, *Rainfall*, and so on.




* Better generalization on test data, indicating optimally fit model.
* Normalized residual distribution and homoscedasticity exist.

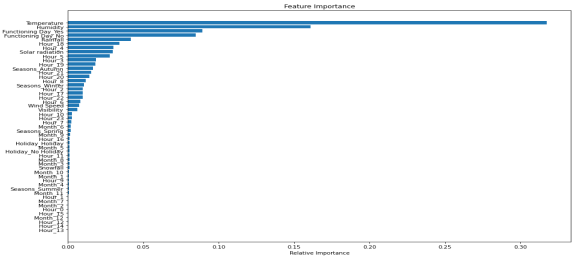
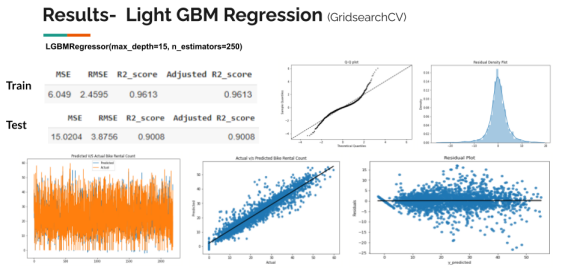



**Figure: Feature importance as predicted by Gradient Boosting regressor**

* Most important features in the prediction include *Temperature*, *Humidity*, *Functioning Day_Yes*, *Functioning Day_No*, *Rainfall*, *Hour_18*, *Hour_4*, and so on.




* Better performance compared to that of gradient boosting regressor.
* Homoscedastic and normal distribution of residuals.

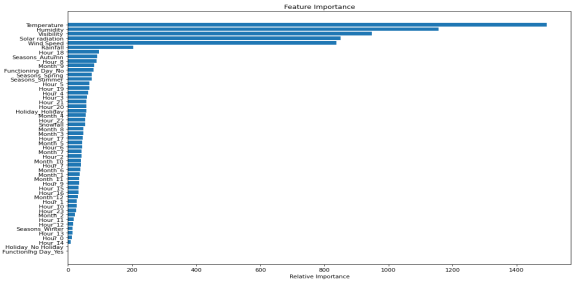
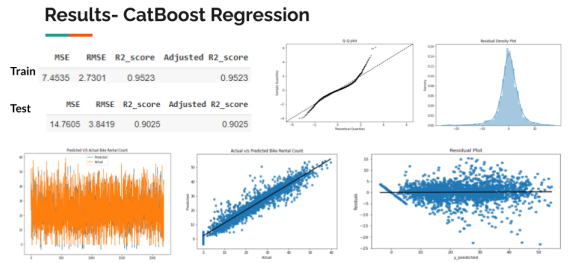



**Figure: Feature importance as predicted by LGB regressor**

* Most important features in the prediction include *Temperature*, *Humidity*, *Visibility*, *Solar radiation*, *Windspeed*, *Rainfall*, *Hour_18*, and so on.

* It can be observed that the numerical features top the predictive analysis.




* By far the best model in terms of performance and generalization.
* Homoscedastic and normal distribution of residuals.

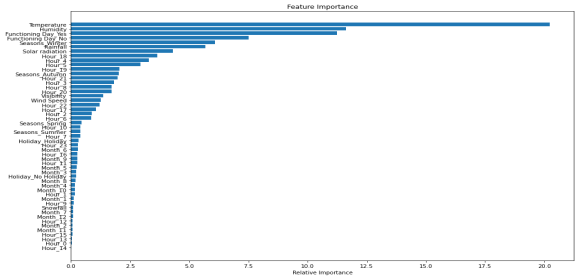



**Figure: Feature importance as predicted by CatBoost regressor**

* Most important features in the prediction include *Temperature*, *Humidity*, *Functioning Day_Yes*, *Functioning Day_No*, *Seasons Winter*, *Rainfall*, *Solar radiation*, and so on.

## Summary

| Model | MSE-train | MSE-test | Adjusted R2_score-train | Adjusted R2_score-test |
|---|---|---|---|---|
| Linear Regression | 37.0831 | 39.2343 | 0.7627 | 0.7409 |
| Lasso Regression GridSearchCV | 37.0875 | 39.0978 | 0.7626 | 0.7418 |
| Ridge Regression GridSearchCV | 37.0830 | 39.2264 | 0.7627 | 0.7409 |
| Elastic-Net GridSearchCV | 37.0854 | 39.0690 | 0.7627 | 0.7420 |
| Decision Tree Regression | 30.7392 | 36.9465 | 0.8033 | 0.7560 |
| Decision Tree GridSearchCV | 17.8624 | 29.0001 | 0.8857 | 0.8085 |
| Random Forest Regression | 2.3587 | 16.2013 | 0.9849 | 0.8930 |
| Random Forest GridSearchCV | 6.3040 | 16.7718 | 0.9597 | 0.8892 |
| Gradient Boosting Regression | 21.1626 | 22.2409 | 0.8646 | 0.8531 |
| Gradient Boosting GridSearchCV | 12.9139 | 19.6146 | 0.9174 | 0.8705 |
| Light Gradient Boosting GridSearchCV | 6.0490 | 15.0204 | 0.9613 | 0.9008 |
| CatBoost Regression | 7.4535 | 14.7605 | 0.9523 | 0.9025 |

* The dataset with 10 numerical features and 4 categorical features very well explains the bike rental count in the city and is further explored and analyzed to draw some meaningful predictions/insights.
* The outliers in the features have been handled by robust scaling.
* The exploratory data analysis gives insights into the dependency between features, their relationship with the target variable, and various trends.

    * Bike rental count is seen to follow a certain trend to grow high during peak hours, seasons, and functioning hours.

    * Bike rental count tends to increase with the increase in temperature, indicating that people tend to avoid bikes during low temperatures.

* Various models have been fitted onto the prepared dataset and the following comparisons can be made:

    * Linear Regression model fails to capture the details in the data and the extreme high coefficients of certain features can bias the model towards the same. Hence looking for regularization. It also suffers from heteroscedasticity.

    * Regularization helped in regularizing the model coefficients and smoothening the coefficients plot. Though no prominent

improvements in the performance are seen after regularization.

***** Non-linear model- Decision Tree succeeded in capturing the non-linearities in the data that the linear model couldn't capture, hence leading to the improved performance, which was further improved by hyperparameter tuning but led to an overfitting model.

***** Random forest model got the further improvement in the performance but was still overfit, which was further reduced by hyperparameter tuning. The residual plot exhibits homoscedasticity.

***** Started off boosting ensemble models with '*Gradient Boosting*' algorithm that performed comparatively weaker than previous non-linear models.

The performance improved highly on hyperparameter tuning and obtained an optimally fit model with high performance.

Further carried on with the 'LGB' model and 'CatBoost' model resulting in the most optimal fit models.

***** It is safe to conclude that models namely '***Random Forest GridSearchCV***', '***Gradient Boosting GridsearchCV***', '***Light Gradient Boosting***', ***and*** '***CatBoost***' are fit to be considered for application/implementation purposes.

***** **Temperature** and **Humidity** remain the most important features in ensemble predictive models [Random Forest, Gradient Boosting, Light GBM, CatBoost]. This implies that temperature and humidity are the most crucial features in deciding the bike rental count.

***** Functioning Day, Solar radiation, Rainfall, and Hour_18 are the others among the most important features in most of the optimally performing models.

***** This implies the fact that the prediction of bike count is high during functioning hours.

***** The higher the rainfall, the lesser the bike rents.

***** Higher the solar radiation, the more the bike rents.

***** And 6:00 pm being one of the peak hours is one among the most important factors in prediction.

## Conclusion

The study shows that the rent of bikes is influenced by a lot of features. It is understood that many Koreans usually and mainly rent bikes during peak hours, so it is supposed that the main use is to go to school or work. Many conditions contribute to the variation of the number of rents like the day of the week, the hour of the day, season, and weather conditions. And as expected more people are set to rent bikes when the weather is favorable. Bike rents tend to be higher during the peak hours of the favorable day.

The prediction could be enhanced if provided with the location of the bike stands in the city and traffic details in real-time.

References


1. Wikipedia
2. GeeksforGeeks
3. Analytics Vidhya
4. StackOverflow