**Seoul Rented Bike Count Prediction**

**Saransh Srivastava, Jai Harish S,**

**Pranil Thorat, Harish Patil**

**Data science trainees,**

**Alma Better, Bangalore**

**Abstract:**

The goal of this project is to combine the historical bike usage patterns with the weather data to forecast bike rental demand. The data set consists of hourly rental data spanning two years.

Exploratory Data Analysis is done on the dataset and compare the target variable with the other variables to find the distribution of graph. We look for null values which were not found and outliers. We also perform correlation analysis to extract out the important and relevant feature from dataset and later perform train test split to train the model.

The main objective is to build a predictive model, which could help to train a model to predict the number of bike rentals per hour given the weather conditions. This would in turn help to predicting quickly and efficiently.

**1.Problem Statement**

The main objective is to build a predictive model, which could help to train a model to predict the number of bike rentals of the year given the weather conditions. This would in turn help to predicting quickly and efficiently.

The dataset contains following columns:

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

**2. Introduction**

Bike sharing systems are a means of renting bicycles where the process of renting bike on the hourly basis is given in this dataset. Certain features will play a major role on Bike Count so by taking care of that we have to make a model that can predict Rental Bike Count by using the independent features present in the dataset.

## **3. Factors Affecting**

## Following are the factors affecting to the number of bike rentals:

1. **Weather**: We observe higher bike rentals when the weather (ie humidity, windspeed solar radiations) is more clear and sunny. We also notice that there is a single instance where there were rentals under heavy rain/snow condition this maybe happen because of outliers in the dataset.
2. **Seasons**: Bike rental counts across the 4 seasons ie. Fall spring summer and winter. Bike reservations are highest during the Summer season and least during the Spring season.
3. **Working Day**: Bike rental counts on working and non-working days and we observed that the outliers are present in working day.
4. **Holiday**: Bike rental counts on holidays and non-holidays. Holidays correspond to non-working days. Also, outliers are present in non-holidays.
5. **Temperature**: We observed that there is increase in the bikes rented counts with temperature with a small decrease at the highest temperature. Temperature between 32 and 36 degrees Celsius seems to be the ideal temperature.
6. **Hours**: We observed that there is a peak in the bike rentals counts at around 8am morning and at around 5pm evening.

.

**4. Steps involved:**

The following steps are involved in the project

1. **Exploratory Data Analysis**:

After loading and reading the dataset in notebook, we performed EDA. Comparing target variable which is bike rentals counts with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables and also, we observed the distribution of variables. It gave us a better idea that how feature behaves with the target variable.

1. **Null values Treatment and Outliers:**

Dataset contains a no null values to disturb the accuracy but outliers are present which can disturb the accuracy. So Again, we use z-score to remove outliers.

1. **Numerical and categorical Features:** With the help of exploratory data analysis, we analyzed the categorical as well as numerical features in the dataset.
2. **Label encoding:**

In this dataset some categorical variables like seasons, holiday and function day, we change it with numerical database.

1. **Correlation Analysis:** We plot the heatmap to find the correlation between both dependent variable and independent variables.
2. **Train test Split:**

In train test split we take x as dependent variables and y take as independent variable then train the model.

1. **Fitting different models**

For modelling we tried various classification algorithms like:

1. **Linear Regression**
2. **Polynomial Regression**
3. **Random Forest Regressor**
4. **Tuning the hyperparameters for better accuracy**

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting.

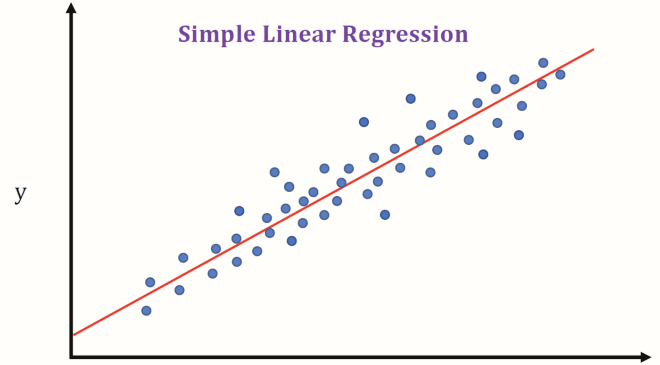
**5. Algorithms:**

1. **Linear Regression:**

Linear Regression is an attractive model because the representation is so simple. The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric. The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g., moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient. For example, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1\*x

So, we establish a relationship b/w independent & dependent variable by fitting a best fit line.



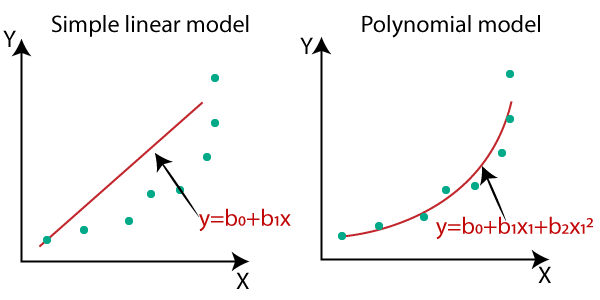
1. **Polynomial Linear Regression:**

In simple linear regression algorithm only work when the relationship between the data is linear but suppose if we have non-linear data then Linear regression will not capable to draw a best-fit line and It fails in such conditions. consider the below diagram which has a non-linear relationship and you can see the Linear regression results on it, which does not perform well means which do not comes close to reality. Hence, we introduce polynomial regression to overcome this problem, which helps identify the curvilinear relationship between independent and dependent variables. Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression.

Suppose we have X as Independent data and Y as dependent data. Before feeding data to a mode in pre-processing stage we convert the input variables into polynomial terms using some degree.

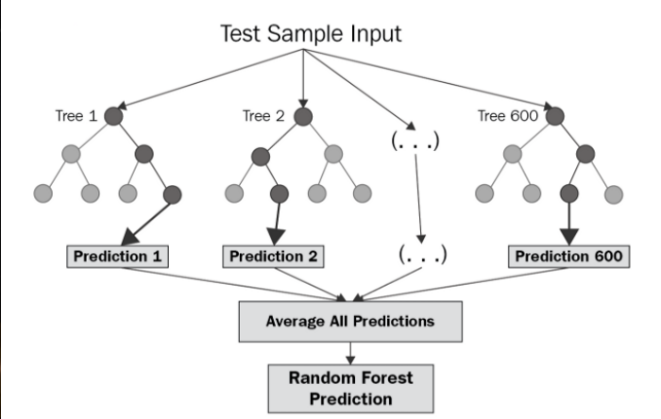
Consider an example my input value is 35 and the degree of a polynomial is 2 so I will find 35 power 0, 35 power 1, and 35 power 2 And this helps to interpret the non-linear relationship in data.  
The equation of polynomial becomes something like this.

**y = a0 + a1x1 + a2x12 + … + anx1n**



1. **Random Forest Regressor:**

Random Forest are an ensemble combination of Decision Tree. In this input data is passed through multiple decision trees. Executes by constructing a different number of Decision Trees at a training time & outputting mean prediction of individual trees.



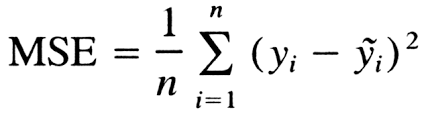
A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include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

**6. Model performance:**

Model can be evaluated by various metrics such as:

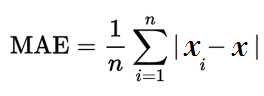
1. **Mean Square Error**-

The [**Mean Squared Error (MSE)**](https://en.wikipedia.org/wiki/Mean_squared_error) or **Mean Squared Deviation (MSD)** of an estimator measures the average of error squares i.e., the average squared difference between the estimated values and true value. It is a risk function, corresponding to the expected value of the squared error loss. It is always non – negative and values close to zero are better. The MSE is the second moment of the error (about the origin) and thus incorporates both the variance of the estimator and its bias.



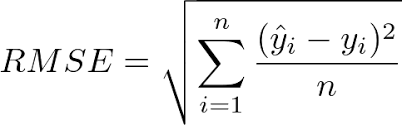
1. **Mean Absolute Error**-

Absolute Error is the amount of error in your measurements.

The **Mean Absolute Error**(MAE) is average of all absolute errors. The formula is:  
[](https://www.statisticshowto.com/wp-content/uploads/2016/10/MAE.png)

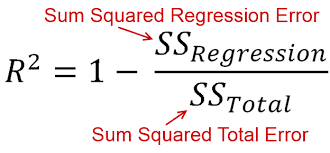
1. **Root Mean Square Error**-

The **root-mean-square deviation** (**RMSD**) or **root-mean-square error** (**RMSE**) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSD represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic function of these differences. RMSD is the square root of the average of squared errors.



1. **R2 Square**-

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.



**7. Hyper parameter tuning:**

Hyperparameters 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 hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV for hyperparameter tuning.

**Grid Search CV-**Grid Search combines a selection of hyperparameters 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.

**8. Conclusion:**

We have reached at the end of our exercise. Starting with loading dataset, treating null values, EDA, Transformation of data, handling skewness in data, Removing Outliers, Removing Multicollinearity, Label Encoding, Feature Engineering and then selecting the features we have implemented three models and our r2 score varies from 56.6% to 87% among all the models used. We have to use Ensemble Techniques as our model was underfitted even after applying hyperparameter tuning. As bagging technique like Random Forest Regression is prone to overfitting, we optimized it by again using Hyperparameter tuning on it.