**Bike Sharing Prediction**

**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 and appropriately modify them by z-score. 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 of the year given the weather conditions. This would in turn help to predicting quickly and efficiently.

***Keywords : Linear regression, Bike rental counts, correlation, Null values, regression model.***

.

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 contain 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 - NoFunc(Non Functional Hours), Fun(Functional hours)

1. **Introduction :**

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis.

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.

1. **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.
7. **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. **One hot 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.

1. **Train test Split :**

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

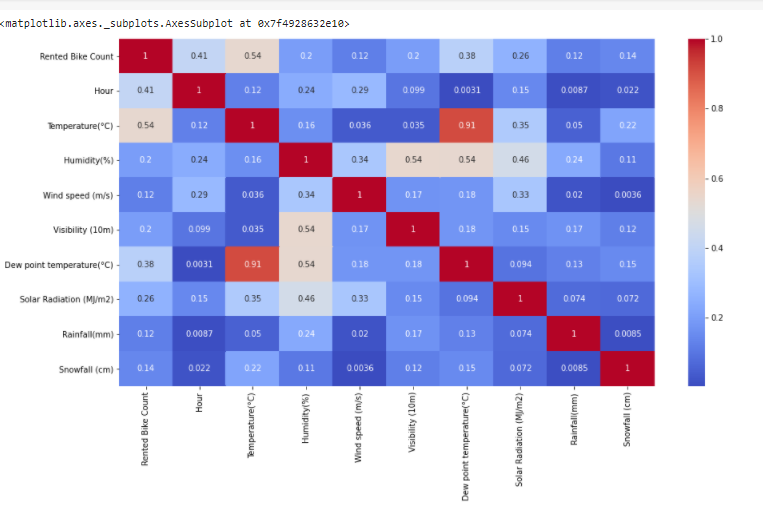
1. **Models :**

We uses 11 modeling to train the data and for predicting the accuracy, RMS and R2.

1. Linear regression
2. Lasso regression
3. Ridge regression
4. Elastic net
5. Polynomial regression
6. KNN and Tree based model
7. Random forest
8. Gradient boosting
9. Extreme gradient boosting
10. Catboost
11. LightBGM
12. **Correlation Analysis :**

We plot the heatmap to find the correlation between all the columns and observed that:

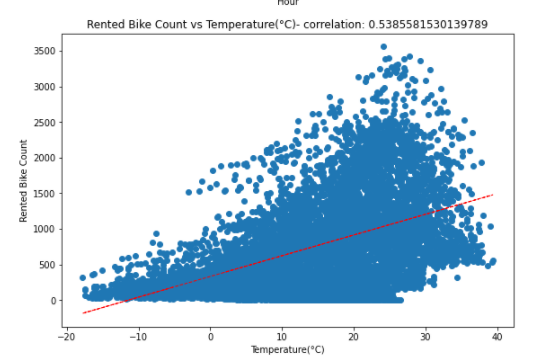
1. Temperatures are highly correlated.
2. There is a positive correlation between bike rentals counts and temperature.
3. We observed a correlation between bike rentals counts and humidity. The more the humidity, the less people prefer to rental bikes.
4. Bike rentals counts has a weak dependence on wind speed.

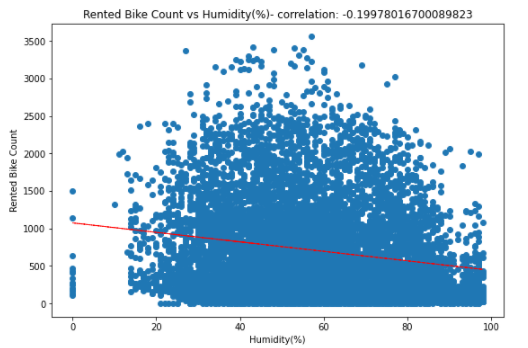
****

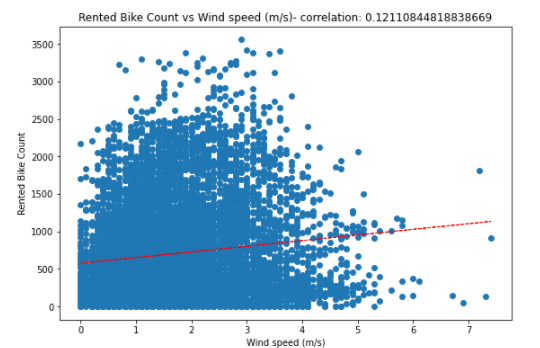
1. **Regression Plot :**

Below are regression plots of the bike rental count with Temperature, Humidity and Wind speed, respectively.

1. There is a positive correlation between bike rentals counts and temperature.
2. We observed a correlation between bike rental counts and humidity. The more the humidity, the less people prefer to rental bikes.
3. Bike rentals counts has a weak dependence on wind speed. As we see in heatmap.



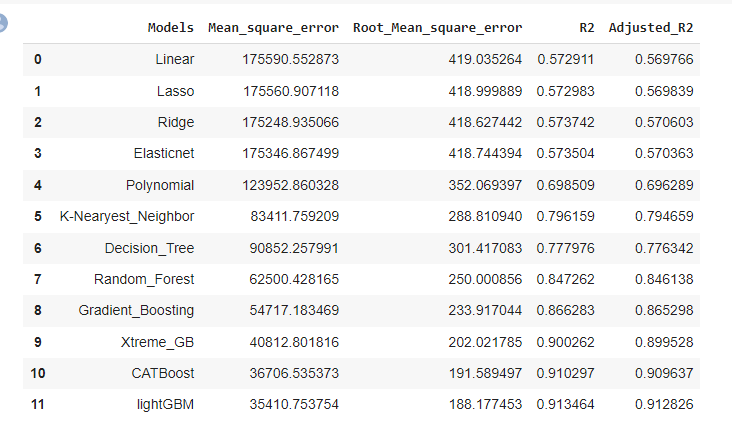




1. **Train test split :**

In the train test split we take two variables ie X and Y where X contain all the independent variables and Y contain dependent variable. Here the independent variable is bike rentals counts and dependent variables is affecting the bike rentals counts like temperature, weather, seasons etc.

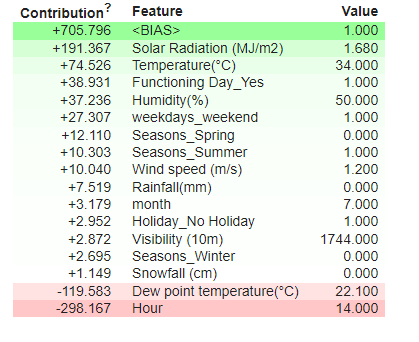
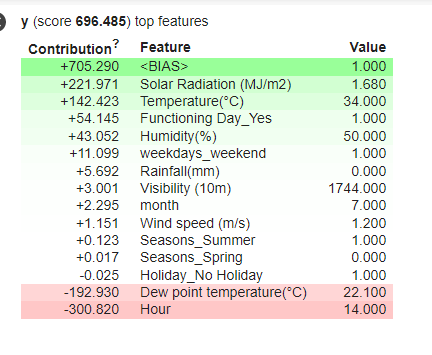
1. **Modeling :**



After performing the various models the lightGBM and Catboost found to be the best model that can be used for the Bike Sharing Demand Prediction since the performance metrics (mse,rmse) shows lower and (r2,adjusted\_r2) shows a higher value for the lightGBM and Catboost models !

# **Model Explainibility**

Model explainability refers to the concept of being able to understand the machine learning model. For example – If a healthcare model is predicting whether a patient is suffering from a particular disease or not. The medical practitioners need to know what parameters the model is taking into account or if the model contains any bias. So, it is necessary that once the model is deployed in the real world. Then, the model developers can explain the model.

****

The green color shows how much the feature contributes to the prediction of the respective class and the weights are positive for the green color.

The red color has negative weights that indicate the feature isn't contributing to the prediction of that class.

It can be observed from the above output, eli5 shows us the contribution of each feature in predicting the output.

1. **Conclusions :**

1. In holiday or non-working days there is demands in rented bikes.

2. There is a surge of high demand in the morning 8AM and in evening 6PM as the people might be going to their work at morning 8AM and returing from their work at the evening 6PM.

3. People prefered more rented bikes in the morning than the evening.

4. When the rainfall was less, people have booked more bikes except some few cases.

5. The Temperature, Hour & Humidity are the most important features that positively drive the total rented bikes count.

6.After performing the various models the lightGBM and Catboost found to be the best model that can be used for the Bike Sharing Demand Prediction since the performance metrics (mse,rmse) shows lower and (r2,adjusted\_r2) shows a higher value for the lightGBM and Catboost models !

7. We can use either lightGBM or catboost model for the bike rental stations.