**Bike Sharing Prediction**

**Shubhangi Devidas Dharmik**

**Data science trainees,**

**AlmaBetter, 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 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.

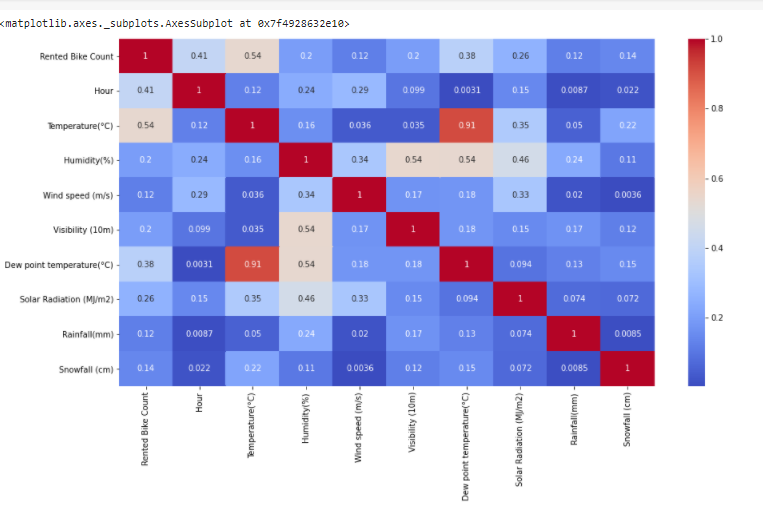
1. **Models :**

We uses 4 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. **Correlation Analysis :**

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

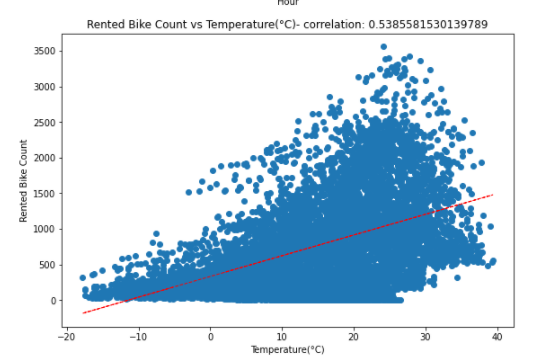
1. Temperatures are highly correlated.
2. There is a positive correlation between bike rentals counts and temperature.
3. We observed a correlation between bike rentals counts and humidity. The more the humidity, the less people prefer to rental bikes.
4. Bike rentals counts has a weak dependence on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

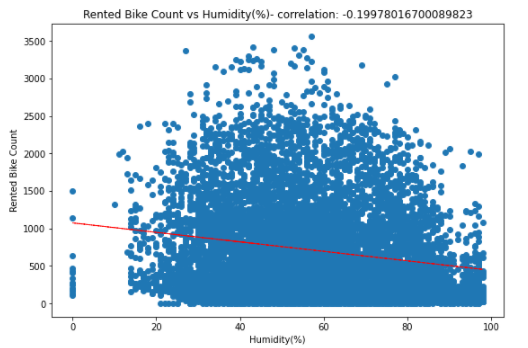
****

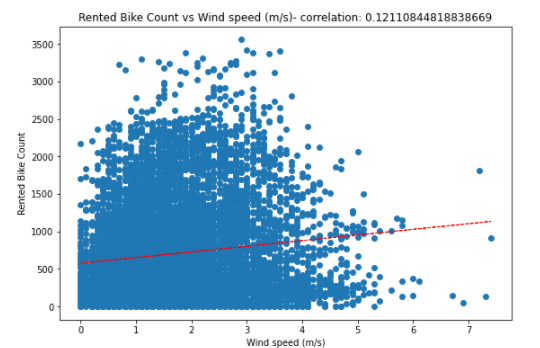
1. **Regression Plot :**

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.





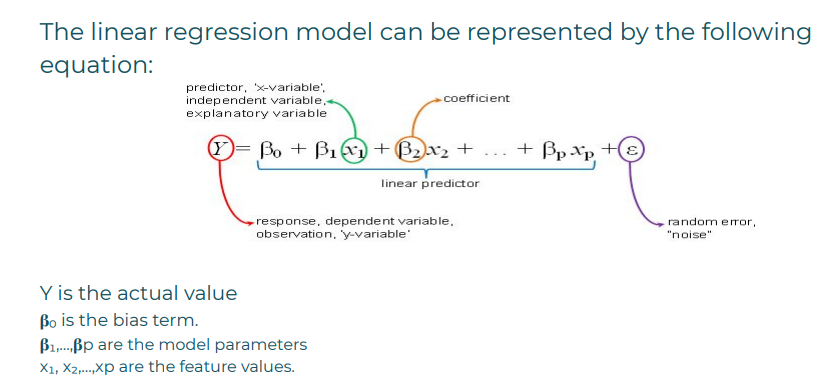


1. **Train test split :**

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 :** 
   1. **Linear Regression :**

The hypothesis of linear regression model represented below:

****

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

1. R Squared for Training Data: 0.55
   1. R Squared for Testing Data: 0.54
   2. RMS for Training Data: 431.48
   3. RMS for Testing Data: 436.59
   4. MAE for Training Data: 322.21
   5. MAE for Testing Data: 326.49
2. **Lasso regression:**

By performing lasso regression we get the results are as follows :

1. R Squared for Training Data: 0.55
2. R Squared for Testing Data: 0.54
3. RMS for Training Data: 431.48
4. RMS for Testing Data: 436.6
5. **Ridge regression:**

By performing ridge regression we get the results are as follows :

* + 1. R Squared for Training Data: 0.55
    2. R Squared for Testing Data: 0.54
    3. RMS for Training Data: 431.49
    4. RMS for Testing Data: 436.63

1. **Elastic Net :**

By performing ridge regression we get the results are as follows :

* + 1. R Squared for Training Data: 0.55
    2. R Squared for Testing Data: 0.54
    3. RMS for Training Data: 431.52
    4. RMS for Testing Data: 436.74

1. **Conclusions :**
2. Bike rental count is mostly correlated with the time of the day as it is peak at 10 am morning and 8 pm at evening.
3. We observed that bike rental count is high during working days than non working day.
4. We see that people generally prefer to bike at moderate to high temperatures. We observed highest rental counts between 32 to 36 degrees Celsius.
5. It is observed that highest number bike rentals counts in Autumn/fall Summer Seasons and the lowest in Spring season.
6. We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day.
7. We observed that with increasing humidity, the number of bike rental counts decreases.
8. Hour of the day holds most importance among all the features for prediction of dataset.
9. We get quite less accuracy because of outliers are present in dataset.