**Bike Sharing Demand Prediction**

Mayur Prakash Parande

Shubham Shrivastava

Praful Mahendra Gedam

**Data science trainees,**

**Alma Better, Bangalore**

**Abstract:**

In today’s upgradation period in business, most of the MNC companies, New start-ups apply the sharing economy concept to give the affordable and better service to customer demand. Bike sharing receives significant attention in worldwide market. Today’s competitive market companies need to develop their prediction model to predict the demand of the customers more accurately. This is the study of to explores a new feature of the bike sharing demand prediction using Machine Learning concept of the available data of the bike sharing company. In this case study apply Linear Regression model to predict the demand of the bike sharing in this challenging market and improve accuracy of the ML algorithms by checking of MAE, RMSE & R2 score and try to get best performance of the algorithm to reducing errors.

***Keywords: Machine Learning, Linear Regression, Analysis, Evaluation of ML model, MSE, MAE, RMSE, R2 score.***

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**

In today’s challenging market Prediction of the accurate demand of the customers is the key factor of the business success. Most of the business is needs prediction of the customer demand for give the better service to the customers. Most of the MNC companies, New Start-ups, or any medium or small business use Machine Learning for get accurate demand of the customers. By the help of the predicted demand companies have more growth in short term. Various business decision depends on the predicted demand for the company.

The bike has attracted worldwide attention in few years ago. In resent business news Rental Bike sharing business is receives significant attention in this challenging worldwide market. So, companies use historical data for prediction demand of the customers using machine learning give the best service. Companies achieve success because they have sufficiently predicted demand based on internal and external data. Data is getting by using the data mining techniques.

Sharing bikes are help transportation that can comprehensively improve environmental problems, traffic jams, sound pollution, air pollution, etc. Now the average distance of a bike trip is around 5-8 km, but its is possible to travel longer distances in connection with public transportation,

In this case study first part collecting of the data and data exploration. In second part Data cleaning, third part Data analysis, forth part Feature Engineering, fifth part Model Building, sixth part Model Evaluation.

**3. Data Description**

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

**Attribute Information:**

* 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)

## **4. Steps:**

* **Data Cleaning:**

In this step we try to clean dataset, check is there any missing value present in dataset or not. Check duplicate values of dataset and remove it, convert date feature into datetime format. we convert the "date" column into 3 different columns i.e. "year", "month", "day". The "year" column in our data set is basically contain the 2 unique number contains the details of from 2017 December to 2018 November so if I consider this is a one year then we don't need the "year" column so we drop it. The other column "day", it contains the details about each day of the month, for our relevance we don't need each day of each month data but we need the data about, if a day is a weekday or a weekend so we convert it into this format and drop the "day" column.

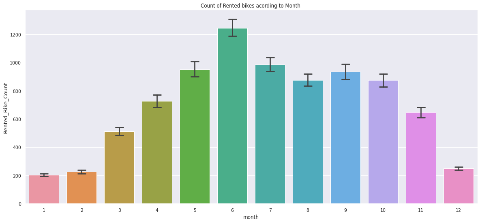
* **Encoding of categorical columns**

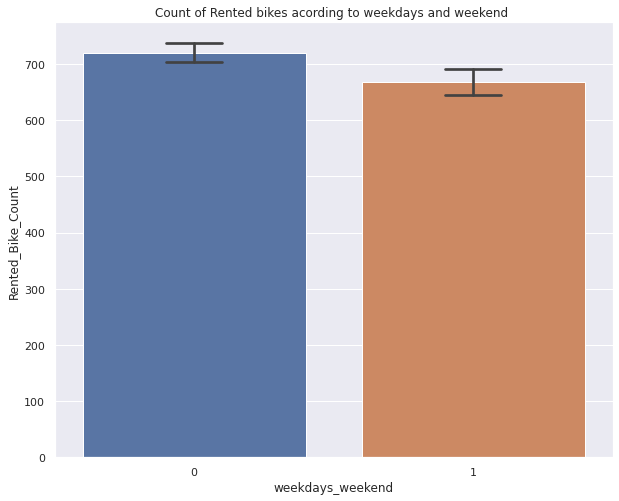
As "Hour", "month", "weekdays\_weekend" column are show as a integer data type but actually it is a category data type. so we need to change this data types if we not then, while doing the further analysis and correlated with this then the values are not actually true so we can mislead by this. Needs to be converted to numerical format.

* **Exploratory Data Analysis**

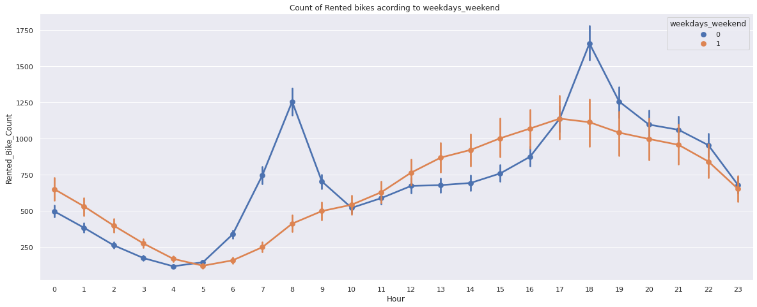
In this step we will analyze dataset by the use of Data visualization by using different graphs. Get idea about features available in dataset.

* **Demand of Bikes Time, Month, Year basis:**





Above plot show the demand of the rental bike in monthly basis. the month 5 to 10 the demand of the rented bike is high as compare to other months. These months are coming inside the summer season.

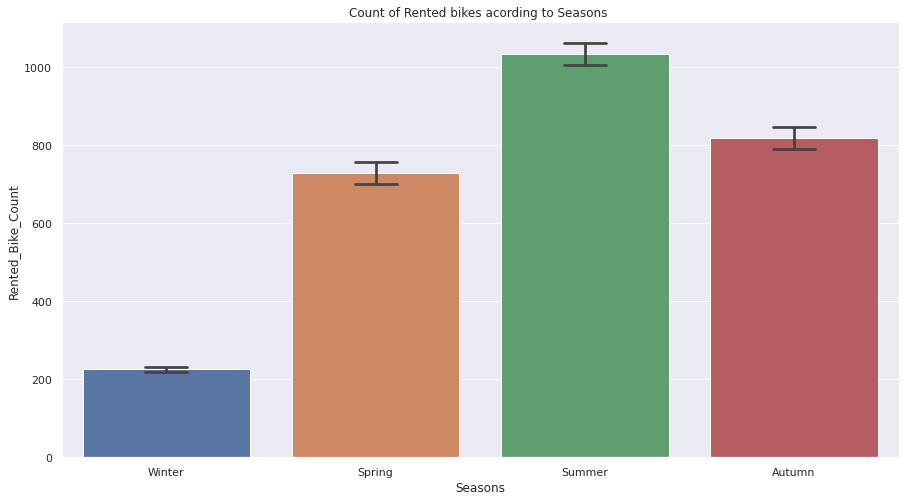


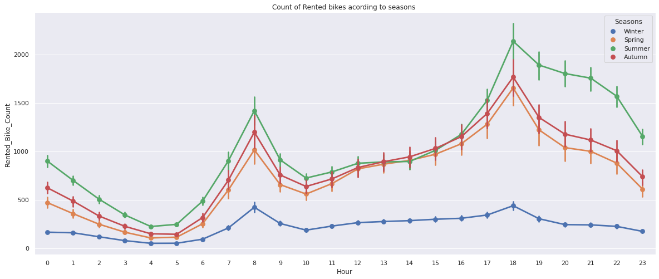
From the above point plot and bar plot we can say that in the week days which represent in blue color show that the demand of the bike higher because of the office.

Peak Time are 7 am to 9 am and 5 pm to 7 pm\*

The orange color represents the weekend days, and it show that the demand of rented bikes is very low specially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases.

* **Demand of the bike season wise:**

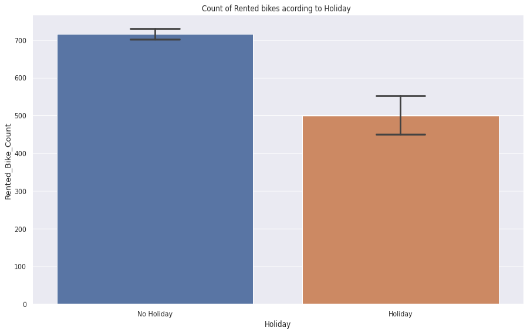




In the above bar plot and point plot which shows the use of rented bike in in four different seasons, and it clearly shows that,

In summer season the use of rented bike is high and peak time is 7am-9am and 7pm-5pm.

In winter season the use of rented bike is very low because of snowfall.

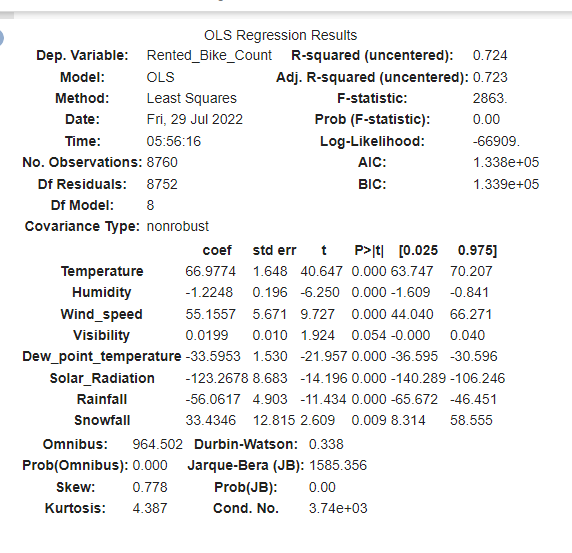


plot shows that in holiday people uses the rented bike from 2pm-8pm.

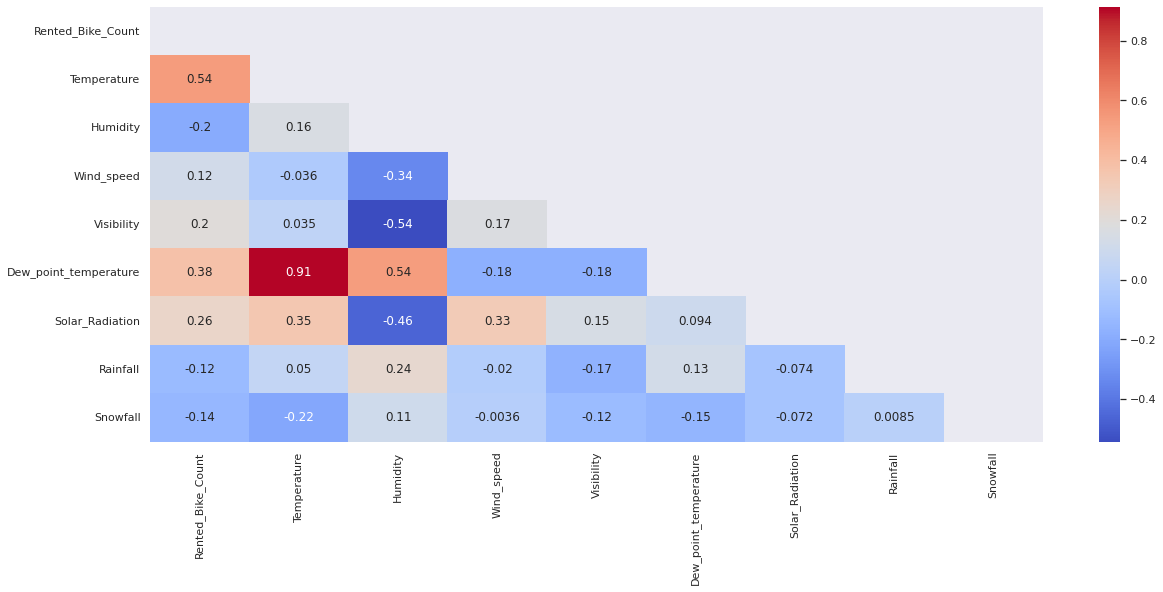
From the analysis numerical features 'Temperature', 'Wind\_speed', 'Visibility', 'Dew\_point\_temperature', 'Solar\_Radiation' are positively relation to the target variable.

which means the rented bike count increases with increase of these features. Rainfall, 'Snowfall', 'Humidity' these features are negatively related with the target variable which means the rented bike count decreases when these features increase.

* **Checking correlation of features:**



* R square and Adj Square are near to each other. 40% of variance in the Rented Bike count is explained by the model.
* For F statistic, P value is less than 0.05 for 5% level of significance.
* P value of dew point temp and visibility are very high and they are not significant.
* Omnibus tests the skewness and kurtosis of the residuals. Here the value of Omnibus is high., it shows we have skewness in our data.
* The condition number is large, 3.11e+04. This might indicate that there are strong multicollinearity or other numerical problems
* Durbin-Watson tests for autocorrelation of the residuals. Here value is less than 0.5. We can say that there exists a positive auto correlation among the variables.



the target variable line the most positively correlated variables to the rent are: temperature, dew point temperature, solar radiation

And most negatively correlated variables are: Humidity, Rainfall

* **Convert Categorical features into Numerical Features:**

In these steps we used algorithms like **A one hot encoding allows the representation of categorical data to be more expressive. Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers. This is required for both input and output variables that are categorical.**

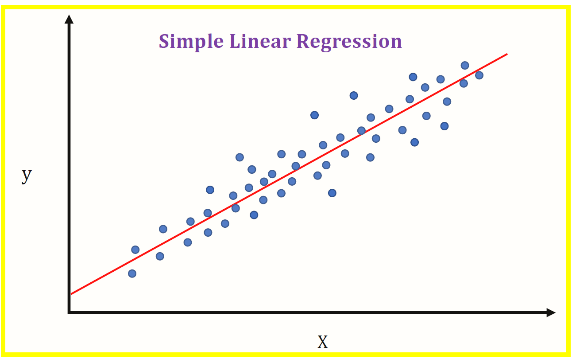
1. **Fitting model**

For modelling we tried Linear Regression algorithm.

* **Train Test Split:** We split data by the use of sklearn model selection import train test split. We split data in 75% Training set and 25% testing set.
* **Algorithms:**

1. **Linear Regression:**

Linear Regression is used to predict the variable values of dependent feature based on the values of independent features. It estimates the coefficient of the linear equation involving one or more than one independent variable that best predict the dependent value. Linear Regression fits a best fit straight line that minimizes the discrepancies between predicted and actual output values. It uses “Least Squares” method to get the best fit line for set of paired data then we get estimated value of X (dependent variable) from Y (Independent variable)



The formula for simple linear regression is

**Y = mX + b**,

where Y is the response (dependent) variable, X is the predictor (independent) variable, m is the estimated slope, and b is the estimated intercept.

1. **Model performance:**

Model can be evaluated by various metrics such as:

**1. Mean Squared Error**-

* The mean squared error (MSE) tells you how close a regression line is to a set of points

MSE= (1/n) \* Σ (actual – forecast)2  
Where:

n = number of items,

Σ = Summation

Actual = original or observed y-value,

Forecast = y-value from regression.

* Model evaluation is calculated

**MSE: 33.27533089591926**

1. **Root Mean Squared Error** -

RMSE (Root Mean Squared Error) is the error rate by the square root of MSE. It is used because MSE is calculated by square so value is too big. By using sqrt root to get it easier for interpretation.

RMSE = √MSE

* Model evaluation is calculated

**RMSE: 5.76847734639907**

1. **Mean Absolute Error**-

Mean absolute error (MAE) is a loss function used for regression. Use MAE when you are doing regression and don't want outliers to play a big role. The loss is the mean over the absolute differences between true and predicted values, deviations in either direction from the true value are treated the same way.

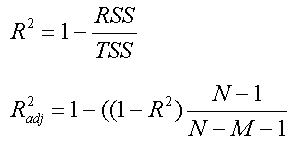
MAE= 1/n) \* Σ |actual – forecast |

* Model evaluation is calculated

**MAE: 4.410178475318181**

1. **R2 and Adj R2-**

R2 shows how well terms (data points) fit a curve or line. Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease



* Model evaluation is calculated

**R2: 0.7893518482962683**

**Adjusted R2: 0.7847297833429184**

1. **Conclusion:**

That's it! We reached the end of our exercise.

Starting with loading the data so far, we have done EDA, null values treatment, encoding of categorical columns, and then model building.

Our model gives lowest MSE on applying Linear Regression on dataset. So, the performance of the model is good in this dataset.

**References-**

1. https://scikit-learn.org/
2. GeeksforGeeks
3. https://www.statisticshowto.com/