Machine Learning

Capstone Project

By

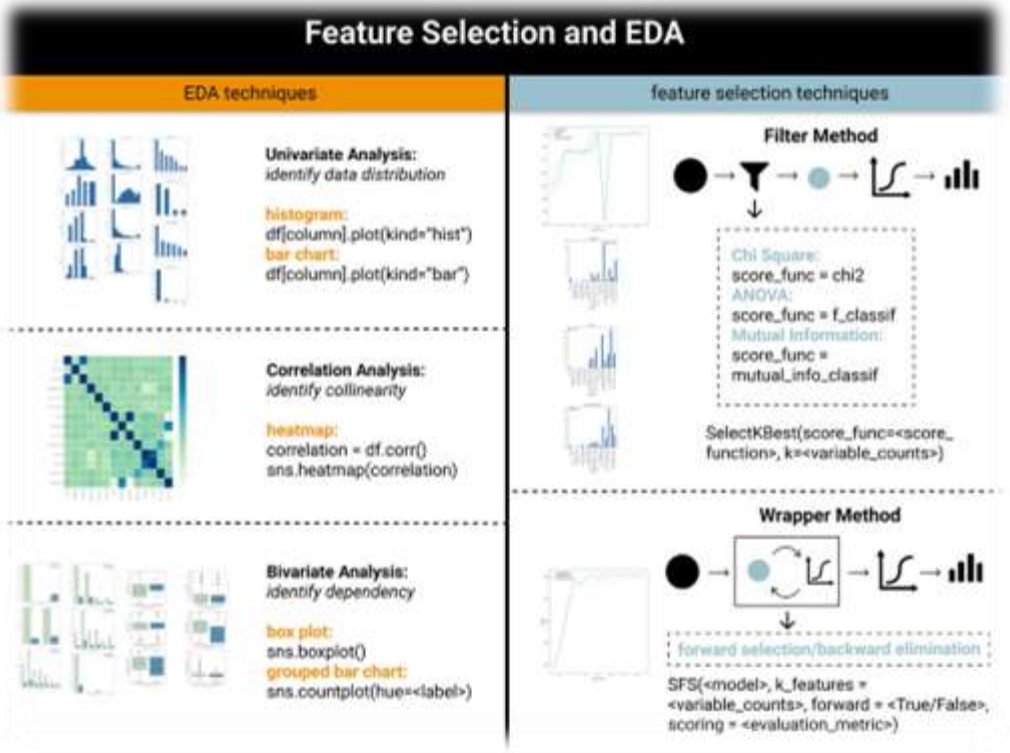
Sushant Jagtap

Team Member: Sushant Jagtap

Akash Bhor

❑ CONTENT

- Defining problem statement
- Data summary
- Insights from our Dataset
- EDA
- Feature engineering
- Applying ML algorithms
- Comparing different ML models
- Challenges
- Conclusion



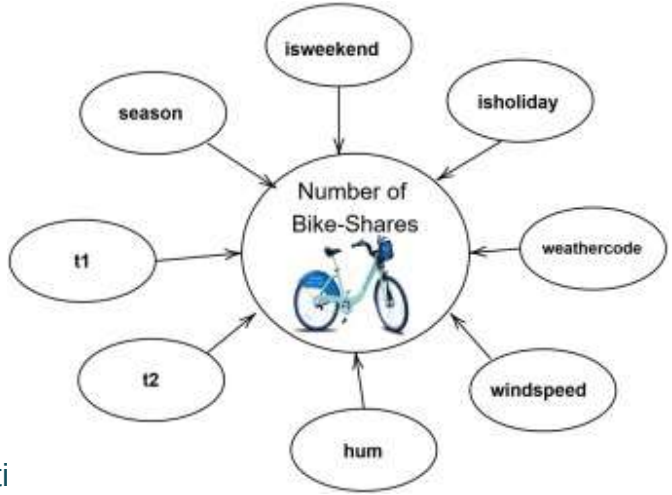
❑ Defining Problem Statement

Rental bikes service is very crucial 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. Therefore we have to predict the number of rental bikes required in each hour for smooth functioning of service.

Data Summary

We are given dataset containing count of rental bikes from December 2017 to November 2018 for each day and each hour of day. Along with count of rental bikes there are following variables also present.

(1) Date : year-month-day
(2) Rented Bike count - Count of bikes rented at each hour
(3) Hour - Hour of the day
(4) Temperature-Temperature in Celsius
(5) Humidity - %
(6) Wind speed - m/s
(7) Visibility - 10m
(8) Dew point temperature - Celsius
(9) Solar radiation - MJ/m2
(10) Rainfall - mm
(11) Snowfall - cm
(12) Seasons - Winter, Spring, Summer, Autumn
(13) Holiday - Holiday/No holiday
(14) Functional Day - No (Non Functional Hours), Yes(Functi

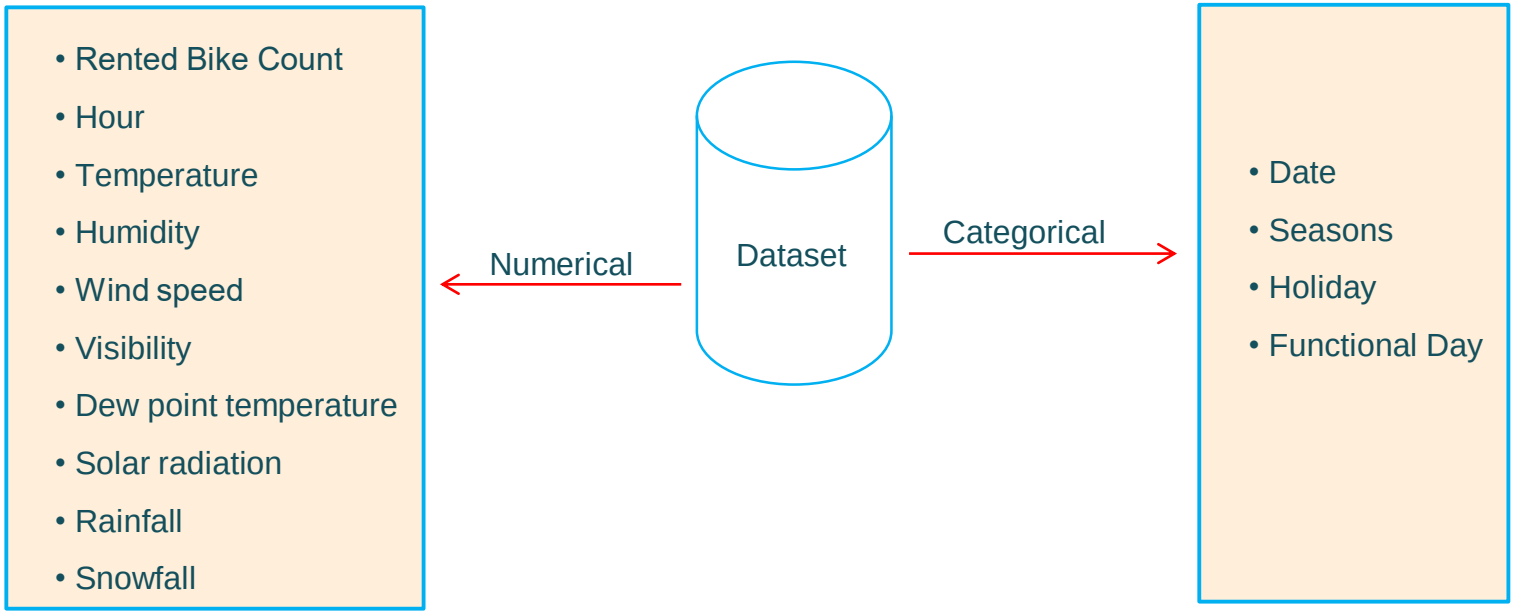


'Rented Bike count' is dependent variable.

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
8755	30/11/2018	1003	19	4.2	34	2.6	1894	-10.3	0.0	0.0	0.0	Autumn	No Holiday	Yes
8756	30/11/2018	764	20	3.4	37	2.3	2000	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8757	30/11/2018	694	21	2.6	39	0.3	1968	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8758	30/11/2018	712	22	2.1	41	1.0	1859	-9.8	0.0	0.0	0.0	Autumn	No Holiday	Yes
8759	30/11/2018	584	23	1.9	43	1.3	1909	-9.3	0.0	0.0	0.0	Autumn	No Holiday	Yes

- ➢ This Dataset contains 8760 lines and 14 columns.
- ➢ Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.
- ➢ One Datetime features ‘Date’.
- ➢ We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall which tells the environment conditions at that particular hour of the day.

Data Summary



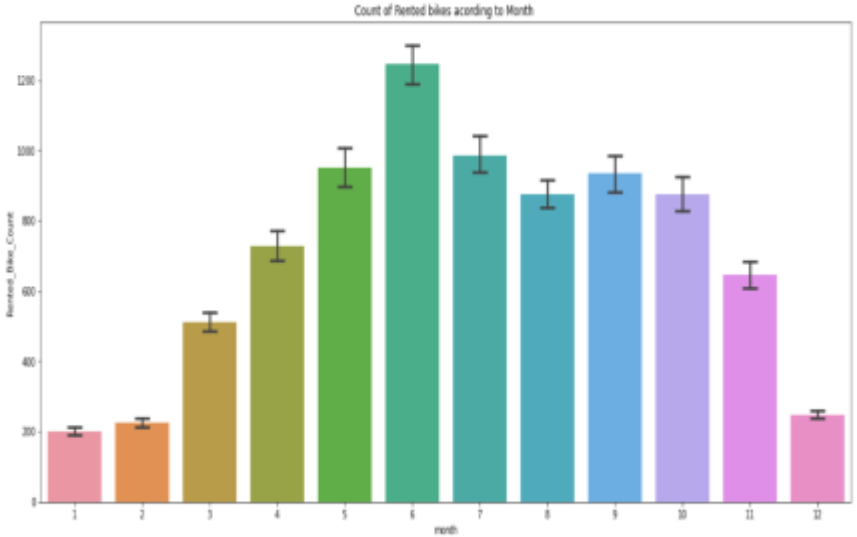
Dependent variable is ‘Rented Bike Count’ that describes the count of rented bikes for each hour.

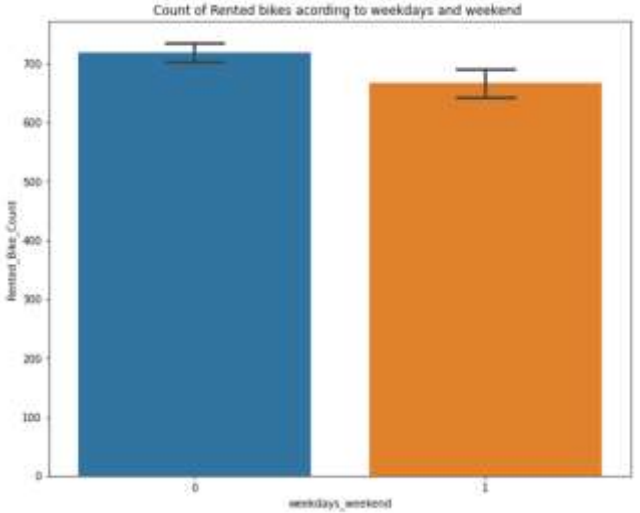
INSIGHTS FROM OUR DATASET

- ➢ There are No Missing Values present
- ➢ There are No Duplicate values present
- ➢ There are No null values.
- ➢ And finally we have 'rented bike count' variable which we need to predict for new observations
- ➢ The dataset shows hourly rental data for one year (1 December 2017 to 31 November(2018)(365 days).
- we consider this as a single year data
- ➢ So we convert the "date" column into 3 different column i.e "year","month","day".
- ➢ We change the name of some features for our convenience , they are as below 'Rented_Bike_Count', 'Hour',
- 'Temperature', 'Humidity', 'Wind_speed', 'Visibility', 'Dew_point_temperature', 'Solar_Radiation', 'Rainfall',
- 'Snowfall', 'Seasons', 'Holiday', 'Functioning_Day', 'month','weekdays_weekend'

EDA

Let us see how the values of ‘Rented Bike Count’ are distributed in given dataset. Distribution of values is highly positively skewed.

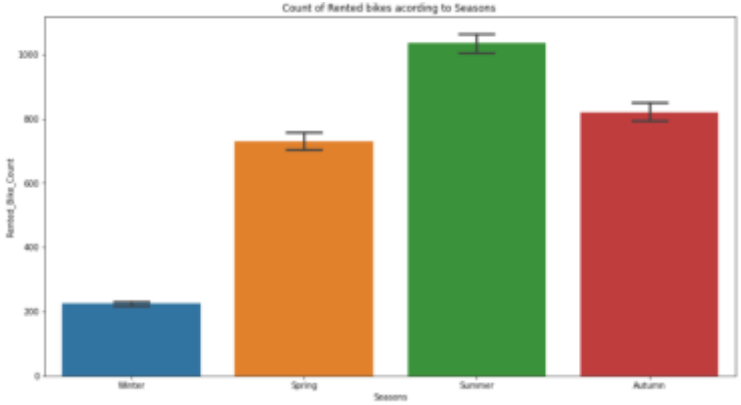


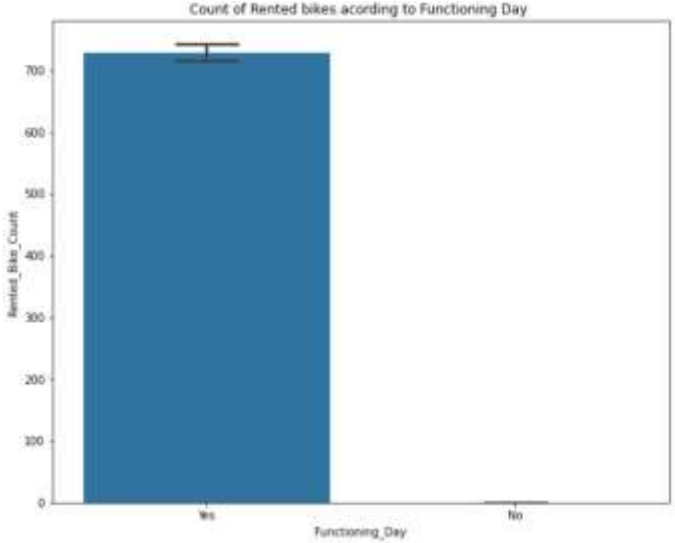


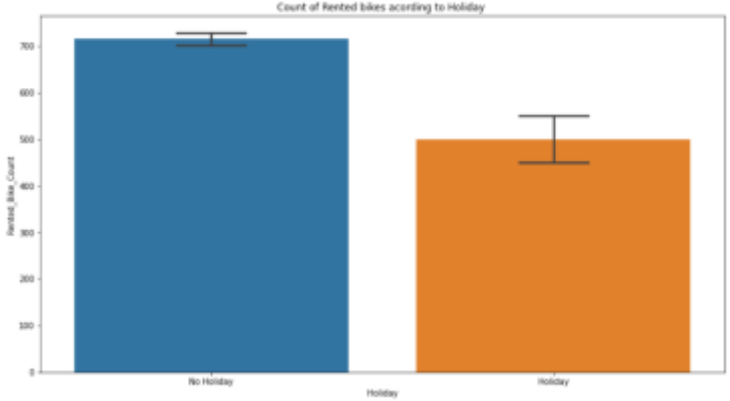
● From the above point plot and bar plot we can say that in the weekdays which represent in blue colour show that the demand of the bike higher because of the office. ● Peak Time are 7 am to 9 am and 5 pm to 7 pm ● The orange color represent the weekend days, and it show that the demand of rented bikes are very low especially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases.

● from the month 5 to 10 the demand of the rented bike is high as compare to other months. these months are comes inside the summer season.

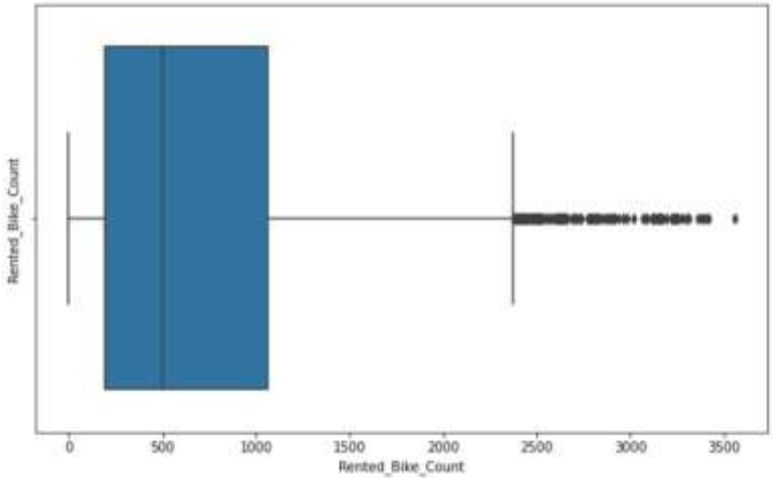
Count of values of categorical features.
Functioning Day and Holiday have highly imbalanced count of values.





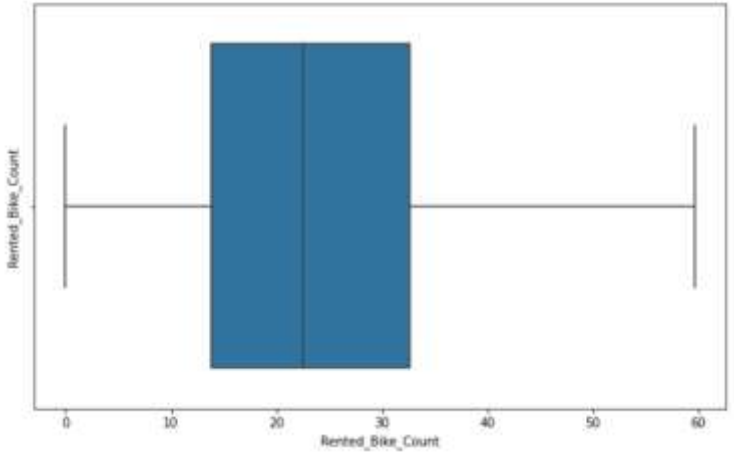


detect outliers in Rented Bike Count column

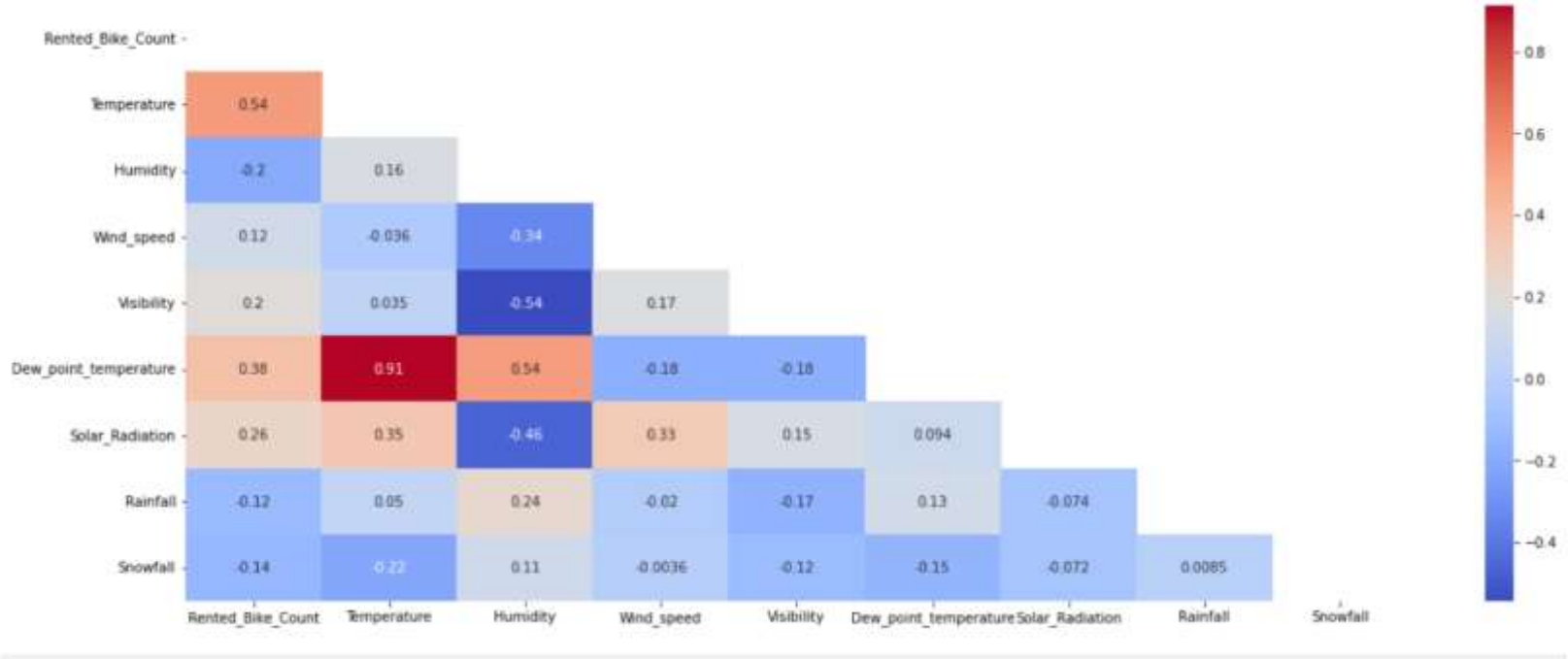



```
#After applying sqrt on Rented Bike Count check wheater we still have outliers
plt.figure(figsize=(10,6))

plt.ylabel('Rented_Bike_Count')
sns.boxplot(x=np.sqrt(df['Rented_Bike_Count']))
plt.show()
```



- The above graph shows that Rented Bike Count has moderate right skewness.
- The above boxplot shows that we have detect outliers in Rented Bike Count column
- Since the assumption of linear regression is that 'the distribution of dependent variable has to be normal', so we should perform Square root operation to make it normal
- After applying Square root to the skewed Rented Bike Count, here we get almost normal distribution.
- After applying Square root to the Rented Bike Count column, we find that there is no outliers present



the most positively correlated variables to the rent are :
the temperature
the dew point temperature
the solar radiation

And most negatively correlated variables are:
Humidity
Rainfall

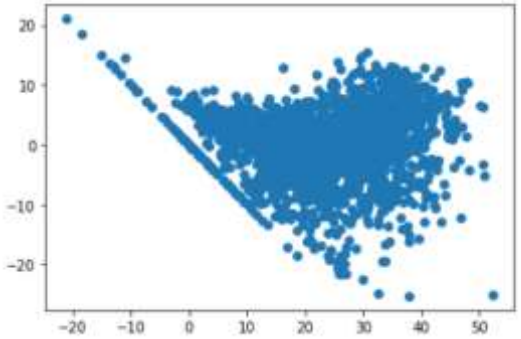
Linear Regression

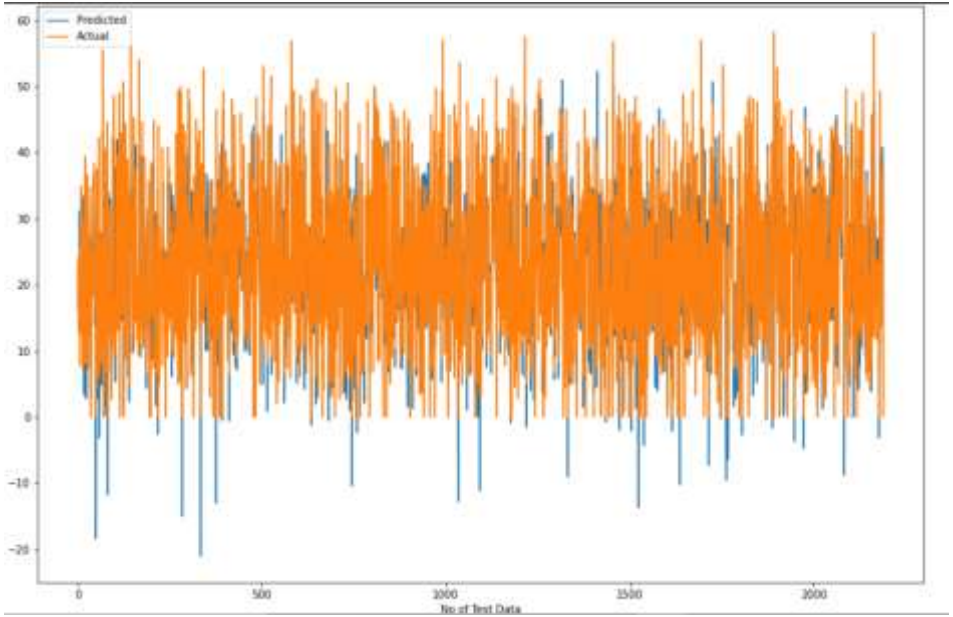
$$y\_pred = \beta_0 + \beta_1 x$$

β_0 and β_1

```
### Heteroscadocity
plt.scatter((y_pred_test),(y_test)-(y_pred_test))
```

```
<matplotlib.collections.PathCollection at 0x7f9b293b9150>
```





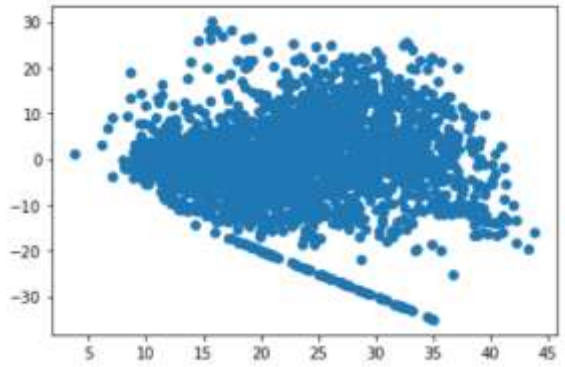
Looks like our r2 score value is 0.77 that means our model is able to capture most of the data variance.

The r2_score for the test set is 0.78. This means our linear model is performing well on the data.

LASSO REGRESSION

```
### Heteroscadacity
plt.scatter((y_pred_test_lasso),(y_test-y_pred_test_lasso))
```

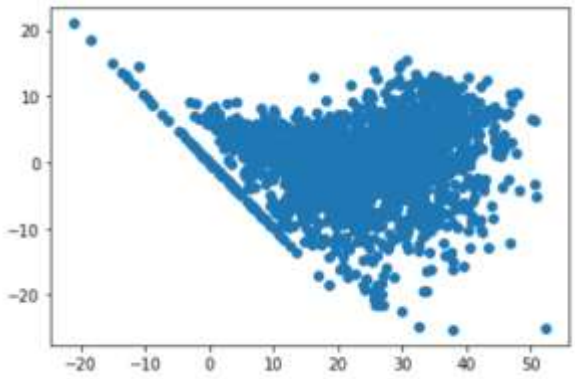
```
<matplotlib.collections.PathCollection at 0x7f9b39169390>
```



RIDGE REGRESSION

```
### Heteroscadacity
plt.scatter((y_pred_test_ridge),(y_test)-(y_pred_test_ridge))
```

```
<matplotlib.collections.PathCollection at 0x7f9b3bf0a310>
```



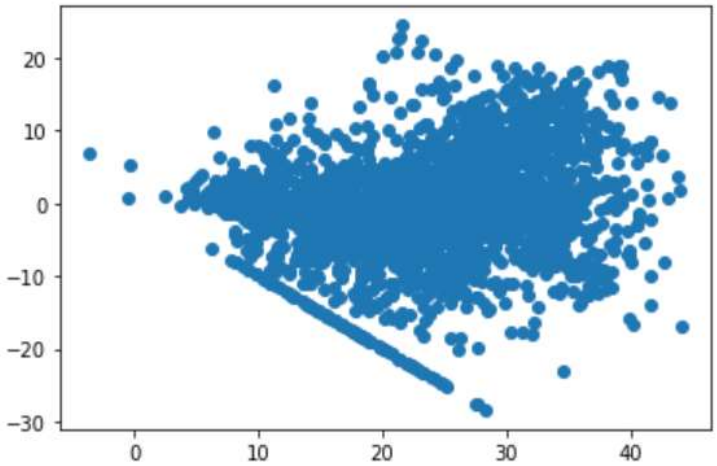
Looks like our r2 score value is 0.40 that means our model is not able to capture most of the data variance.
The r2_score for the test set is 0.38. This means our linear model is not performing well on the data.

Looks like our r2 score value is 0.77 that means our model is able to capture most of the data variance.
The r2_score for the test set is 0.78. This means our linear model is performing well on the data.

ELASTIC NET REGRESSION

```
### Heteroscadacity
plt.scatter((y_pred_test_en),(y_test)-(y_pred_test_en))
```

```
<matplotlib.collections.PathCollection at 0x7f9b3bcc9090>
```



Looks like our r2 score value is 0.62 that means our model is able to capture most of the data variance.
The r2_score for the test set is 0.86. This means our linear model is performing well on the data.

Features transformation

Due to presence of categorical features we cant feed data directly in ML algorithm. We need to transform categorical features that have strings datatype to numerical datatype. For which we have used One-hot encoding and label encoding for categorical features.

Seasons
Summer
Winter
Autumn
Spring

One Hot Encoding →

Summer	Winter	Autumn	Spring
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

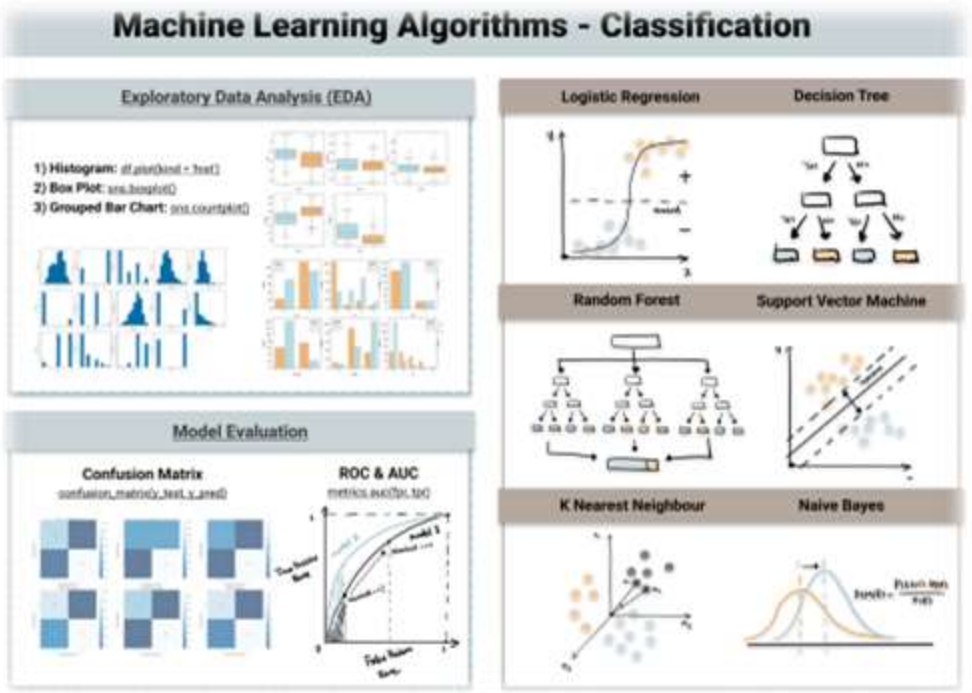
Similarly for ‘Holiday’ and ‘Functional Day’ feature.

❑ Applying ML Algorithms

Since we have to predict the count of rented bikes required per hour. Hence, we have to use regression algorithms.

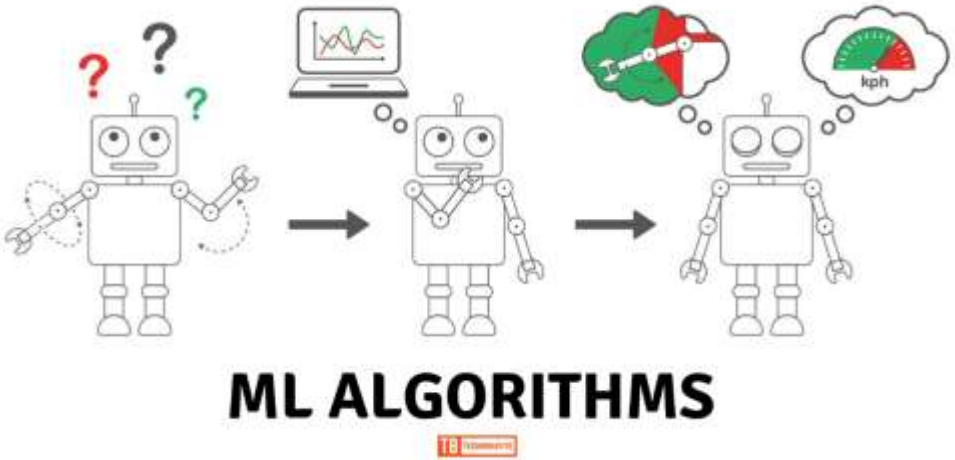
Algorithms that we will use are:

- Decision Tree
- Random Forest
- Linear Regression
- Lasso Regression
- Ridge Regression
- Elastic Net Regression



❑ Applying ML Algorithms

Applying supervised ML algorithms have following steps



Preparing data for model

Training and Hyper parameter tuning

Evaluating model on test data

CHALLENGES

- ➢ Large Dataset to handle.
- ➢ Needs to plot lot of Graphs to analyse.
- ➢ Feature engineering
- ➢ Feature selection
- ➢ Optimising the model
- ➢ Carefully tuned Hyperparameters as it affects the R2 score.

❑ Conclusion

In the given dataset there was no strong linear relation between dependent variable ‘Rented Bike Count’ and independent features. That’s why Linear regression model and its other regularization variant models didn’t performed well.

Out of all models we apply Decision tree and Random forest model are most accurate. Reason for this are no specific relation between features and large data.

Random Forest performed best as it is an ensemble model and result from multiple decision trees is average out to give the prediction.

1. Functioning day is the most influencing feature and temperature is at the second place for LinearRegressor.

2. Temperature is the most important feature for DecisionTree, RandomForest and GradientBoosting Regressor.

3. Functioning day is the most important feature and Winter is the second most for XGBoostRegressor.

4. RMSE Comparisons:
   (a) LinearRegressor RMSE : 370.46
   (b) DecissionTreeRegressor RMSE : 302.53
   (c) RandomForestRegressor RMSE : 290.02
   (d) XGBoostRegressor RMSE : 242.72
   (e) GradientBoostingRegressor RMSE : 248.18

5. The feature temperature is on the top list for all the regressors except XGBoost.

6. XGBoost is acting different from all the regressors as it is considering whether it is winter or not. And is it a working day or not. Though winter is also a function of temperature only but it seems this trick of XGBoost is giving better results.

7. XGBoostRegressor has the Least Root Mean Squared Error. So It can be considered as the best model for given problem.

THANK YOU !...