SEOUL BIKE SHARING DEMAND PREDICTION

Vaitul Sidhdhapara & Drashti Shah
Data Science Trainees,
AlmaBetter, Bangalore

Abstract

This paper presents a rule-based regression predictive model for bike sharing demand prediction. In recent days, Pubic rental bike sharing is becoming popular because of is increased comfortableness and environmental sustainability. Here given data is used included Seoul Bike and data have weather data associated with it for each hour. For the dataset, many statistical models were trained with optimized hyperparameters using a repeated cross validation approach and testing set is used for evaluation:

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Elastic Net Regression
5. Decision Tree Regression
6. Random Forest Regression
7. Gradient Boosting Regression

Multiple evaluation indices such as R^2, Adjusted R^2, Mean Squared Error, Root Mean Squared Error and Mean Absolute Error were used to measure the prediction performance of the regression models. An analysis with variable importance was carried to analyse the most significant variables for all the models developed with the two datasets considered. The variable importance results have shown that Temperature and Hour of the day are the most influential variables in the hourly rental bike demand prediction.

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. Currently, there are over 500 bike-sharing programs around the world.

Several bike/scooter rides sharing facilities (c.g., Bird, Capital Bikeshare, Citi Bike) have started up lately especially in metropolitan cities like San Francisco, New York, Chicago and Los Angeles, and one of the most important problems from a business point of view is to predict the bike demand on any particular day. While having excess bikes results in wastage of resource (both with respect to bike maintenance and the land/bike stand required for parking and security), having fewer bikes leads to revenue loss (ranging from a short-term loss due to missing out on immediate customers to potential longer-term loss due to loss in future customer base). Thus, having an estimate on the demands would enable efficient functioning of these companies.

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.

Problem Statement

The goal of this project is to combine the historical bike usage patterns with the weather data in order to forecast bike rental demand.

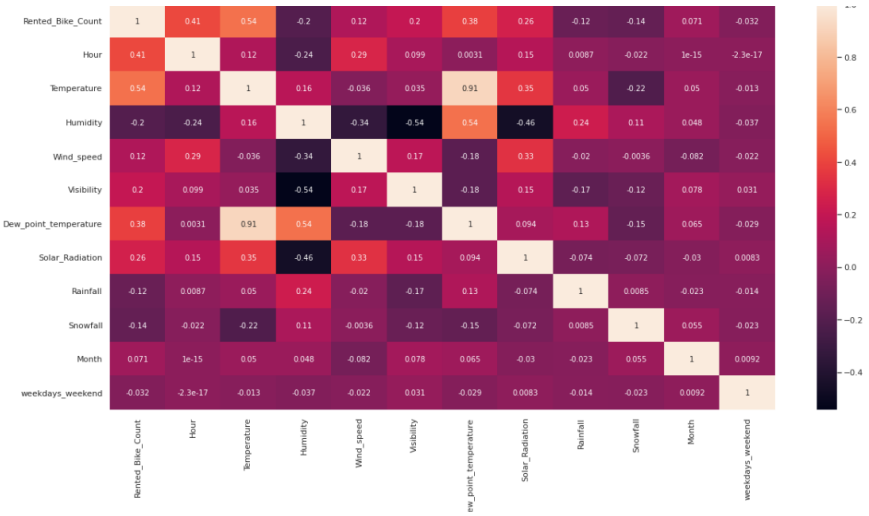
- Target Column to be predicted: 'count'
- Input Columns used as variables (8 columns): ['date', 'season', 'holiday', 'working day', 'weather', 'temp', 'solar radiations', 'humidity', 'windspeed']
- The other two columns (casual and registered) comprise of the split-up of the target column 'count'.

Data Description

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	No (Non-Functional Hours), Yes (Functional hours)

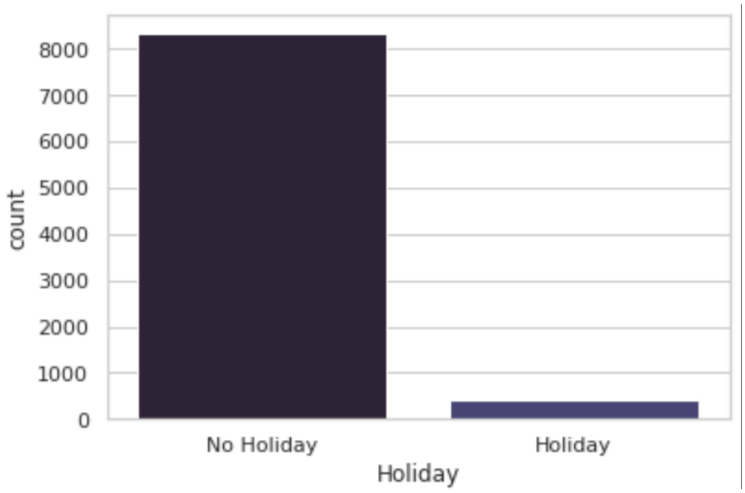
1. Exploratory Data Analysis

Co-relation Map

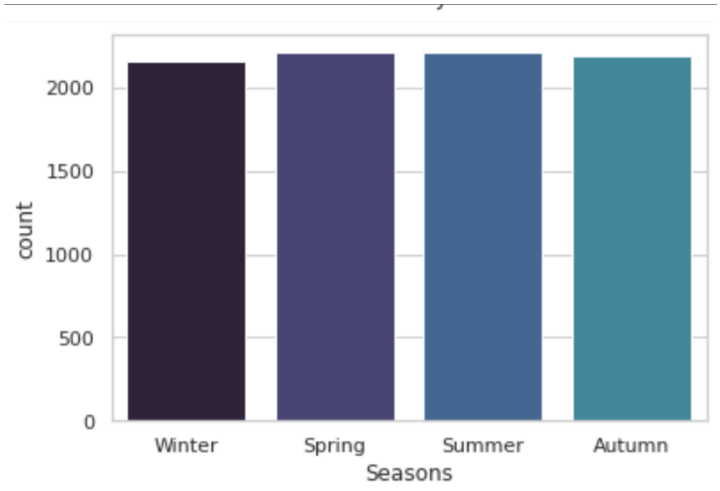


- Humidity has low correlation with visibility due to solar radiation.
- Dew point temperature and temperature are highly related.

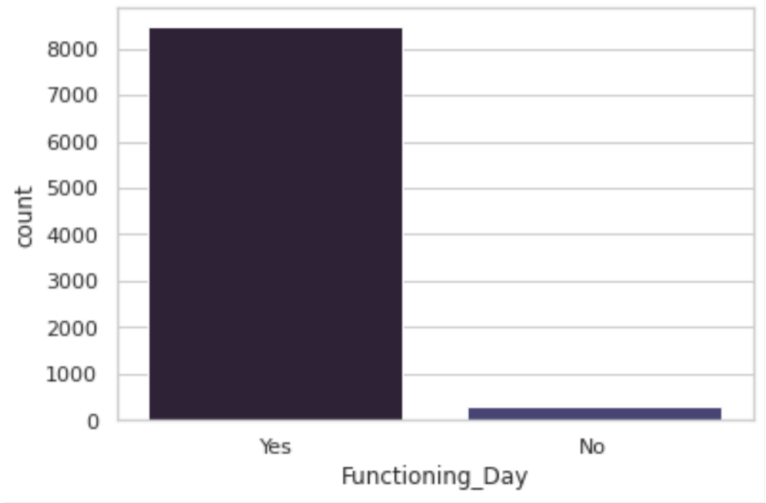
Holiday



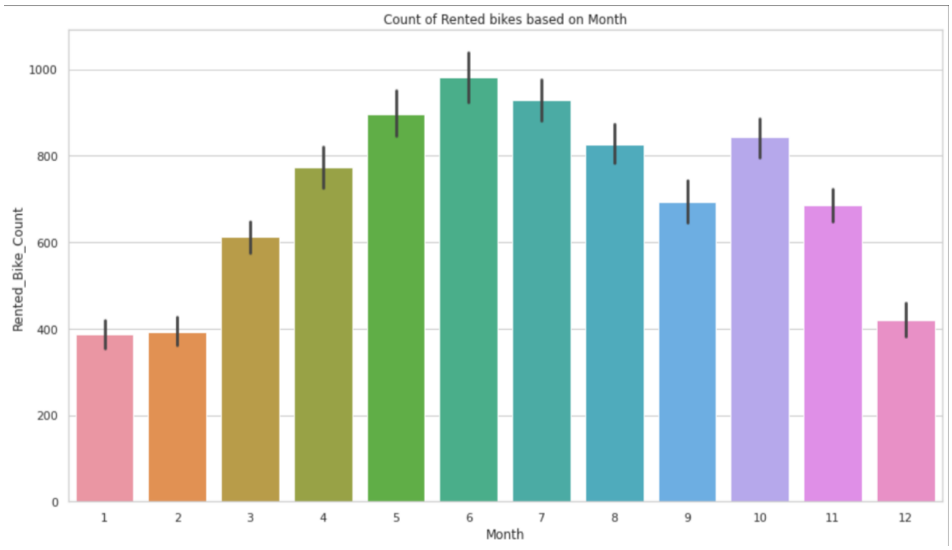
Season



Functioning Day

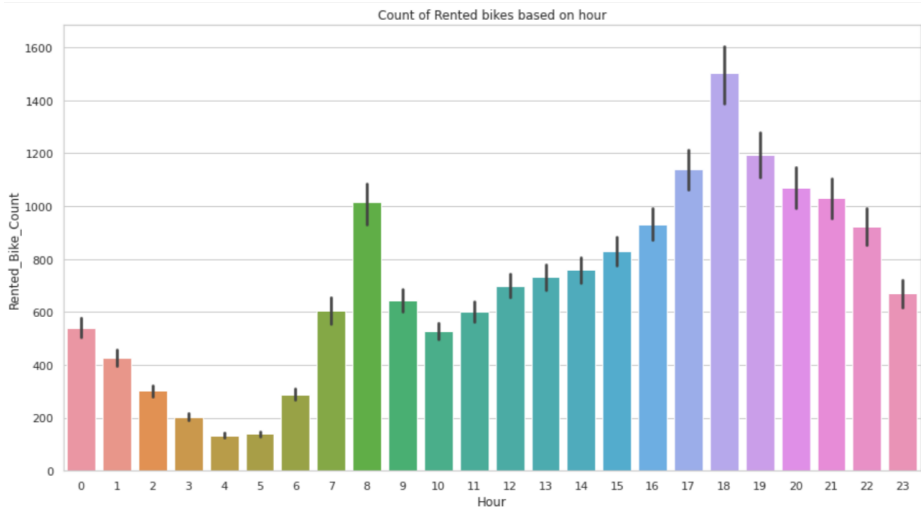


Month & Rented_Bike_Count



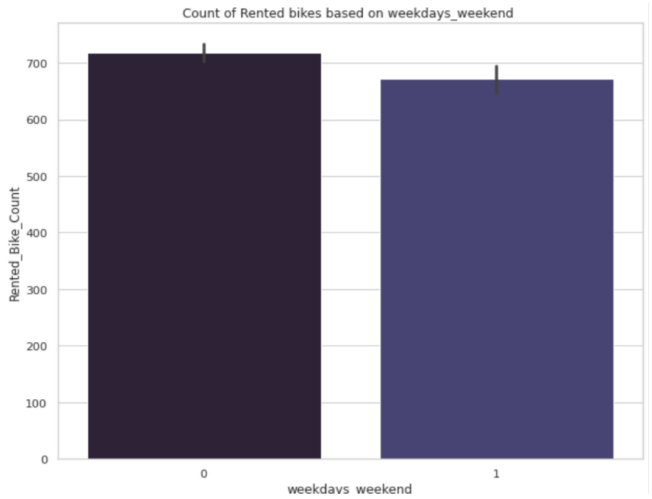
- The demand of the rented bike is high between 5 to 10 months as compare to other months.

Hour & Rented_Bike_Count



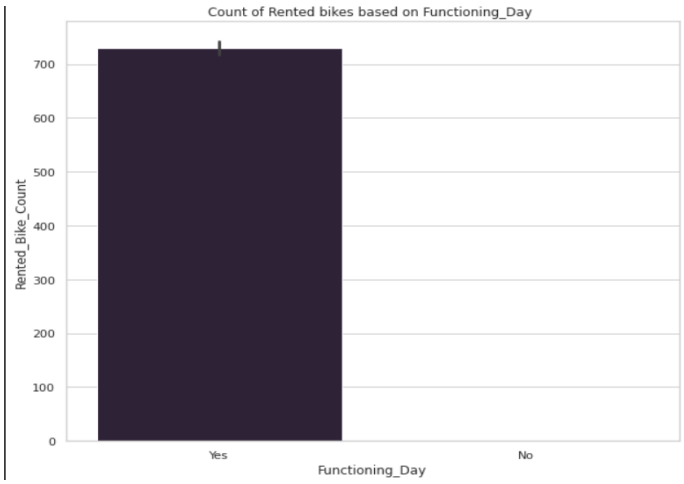
- People generally use rented bikes during their working hour from 7 AM to 9 AM and 5 AM to 8 PM.

Weekend Days & Rented_Bike_Count



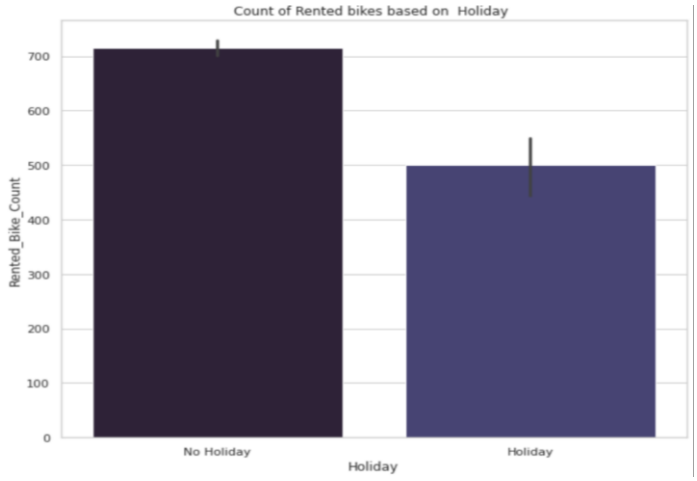
- In the week days the demand of the bike is higher because of the office as compare to the weekend.

Functioning_Day & Rented_Bike_Count

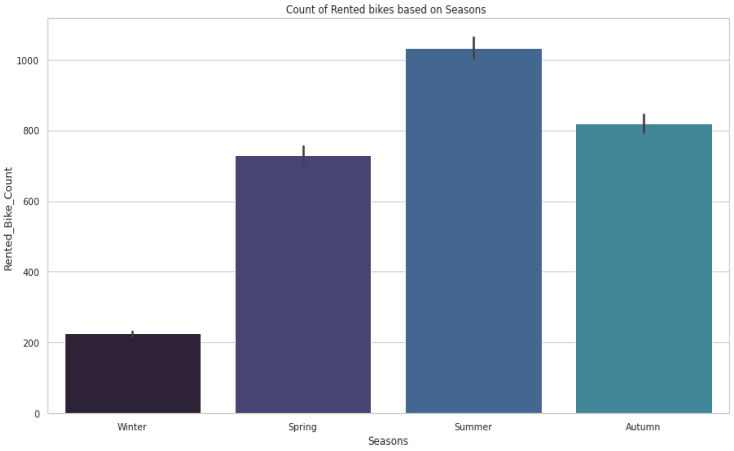


- People use rented bike only in functioning day

Holiday & Rented_Bike_Count

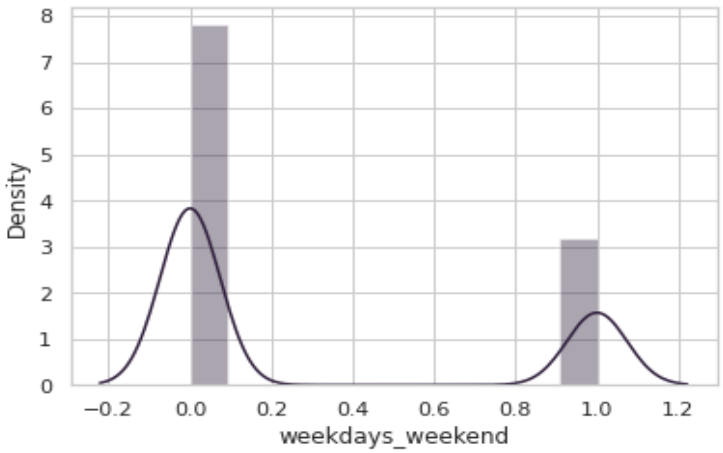


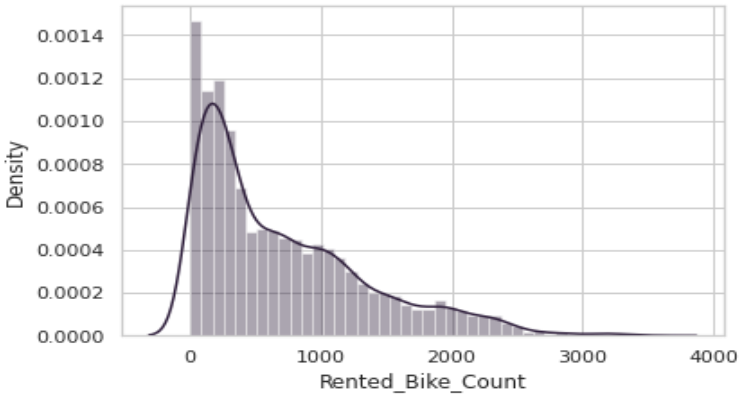
- Use of rented bike is more on 'No holiday' means on working days as compare to 'Holiday'.

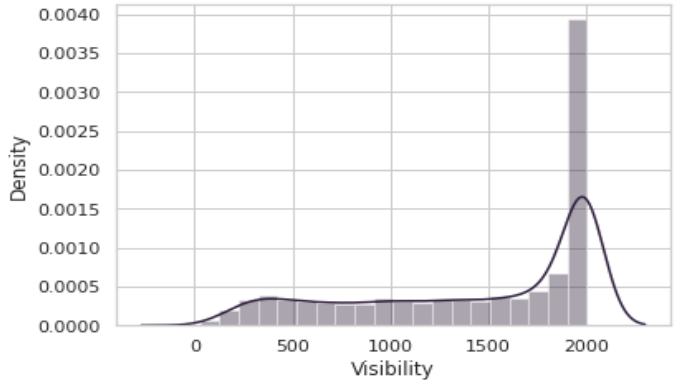


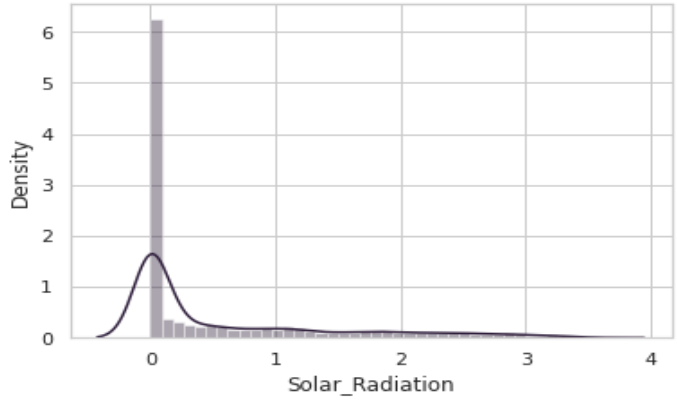
- Maximum number of bike rentals in summer season, while in winter season the bike rented bike count is low.

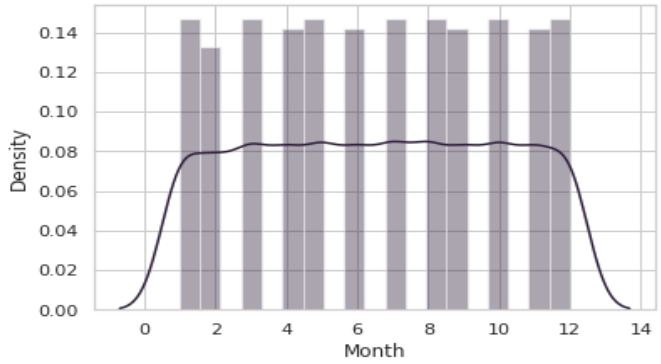
2. Visualizing The Data Distributions

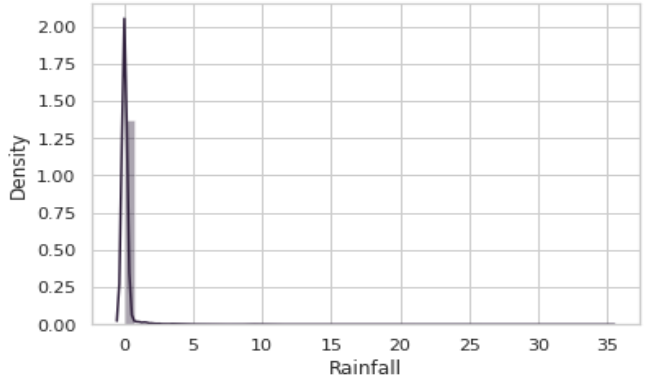


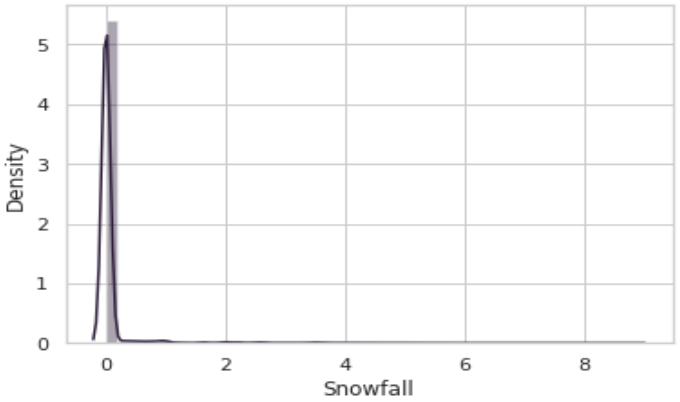


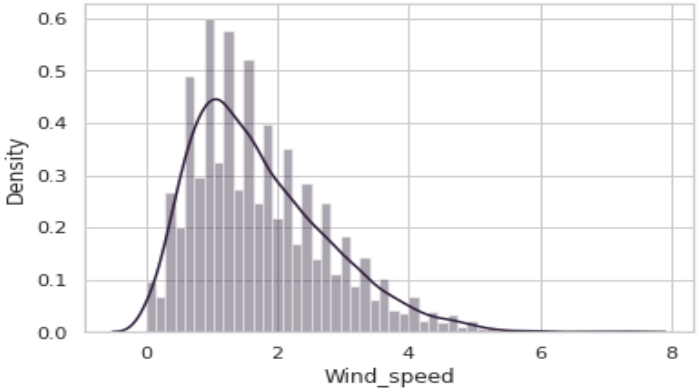

0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
Density
0
500
1000
1500
2000
Visibility

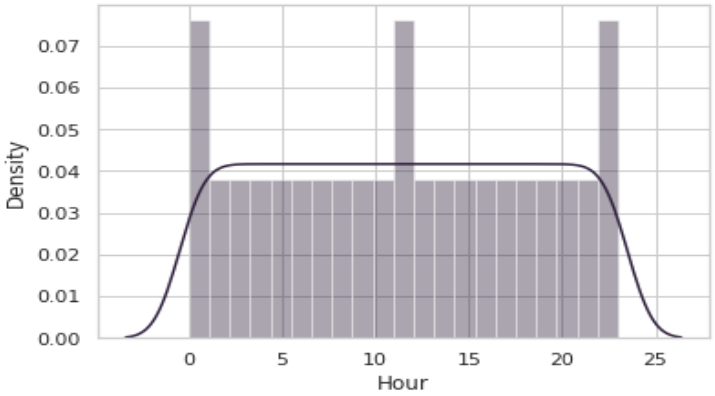

6
5
4
3
2
1
0
Density
0
1
2
3
4
Solar_Radiation

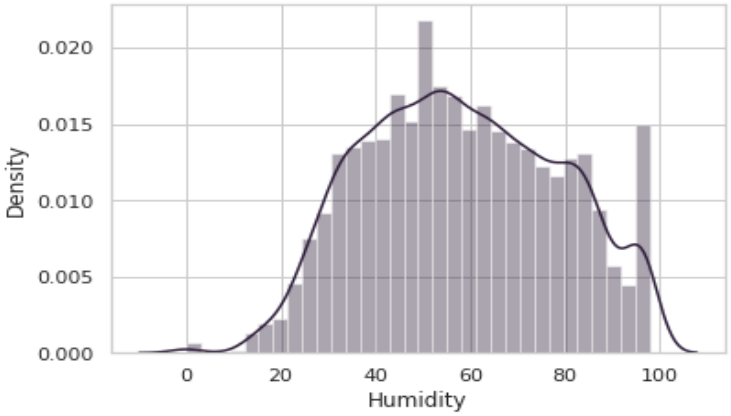

0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
0
2
4
6
8
10
12
14
Month

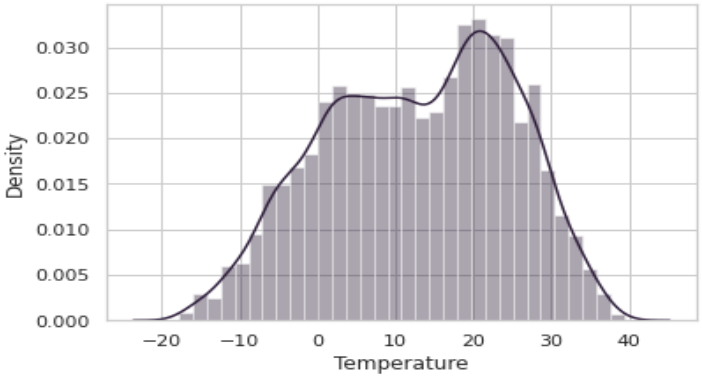

2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Density
0
5
10
15
20
25
30
35
Rainfall


Density
Snowfall


Density
Wind_speed

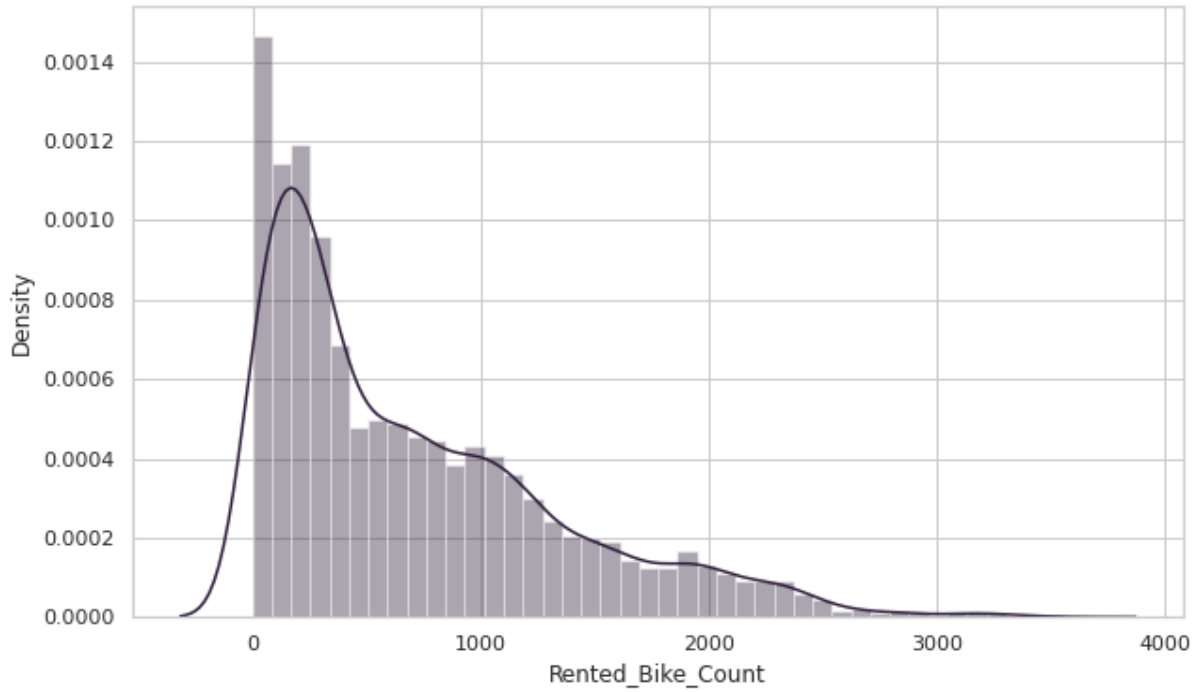

Density
Hour

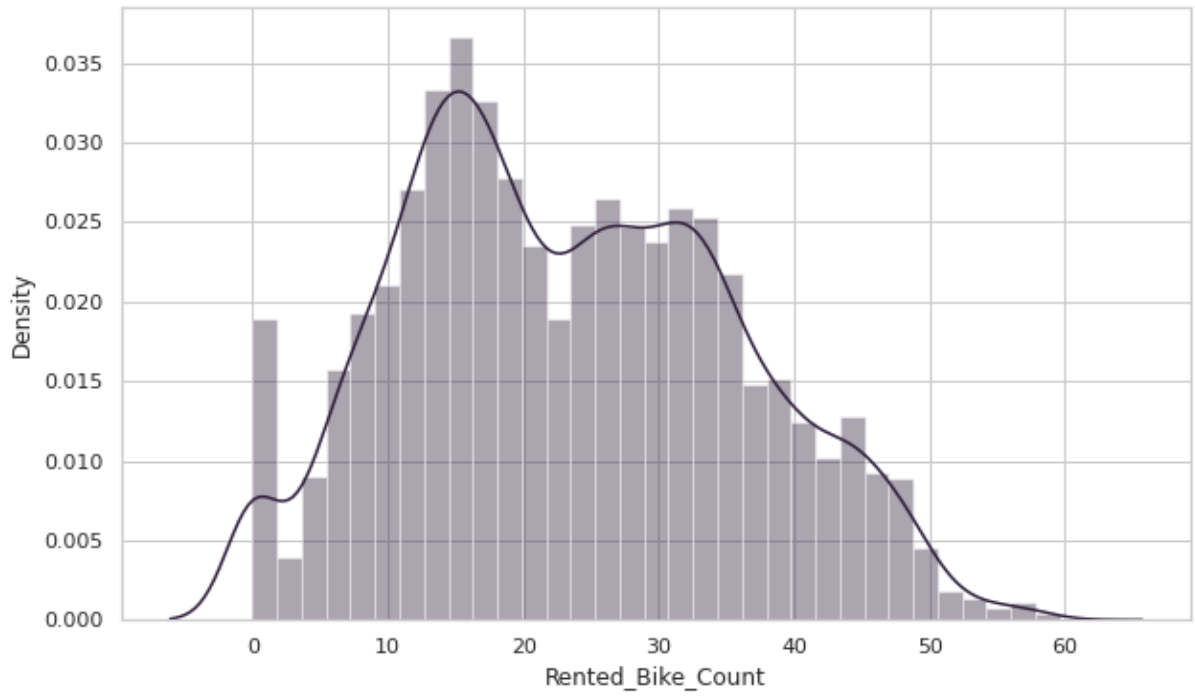

Density
Humidity



- “Temperature” and “Humidity” are normally distributed.
- “Hour” and “Month” columns follow a uniform distribution.
- “Wind Speed”, “Solar Radiation”, “Rainfall” and “Snowfall” are having positively skewed distribution.
- “Visibility” column is negatively skewed.

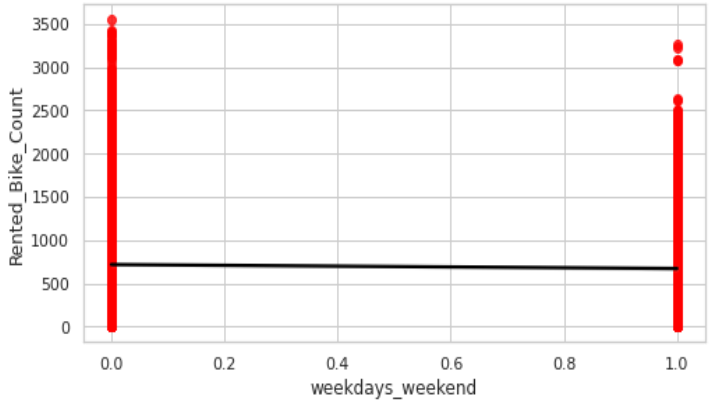
3. Normalising "Rented_Bike_Count" Column Data (Target Variable)

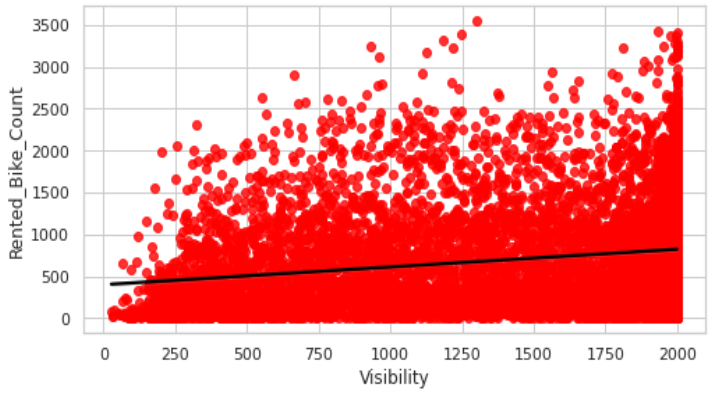


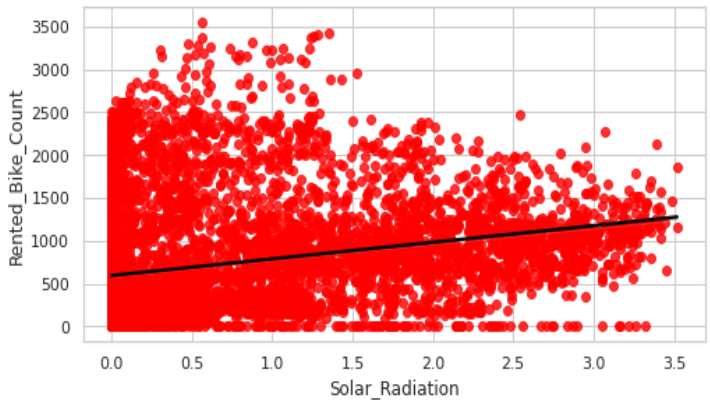


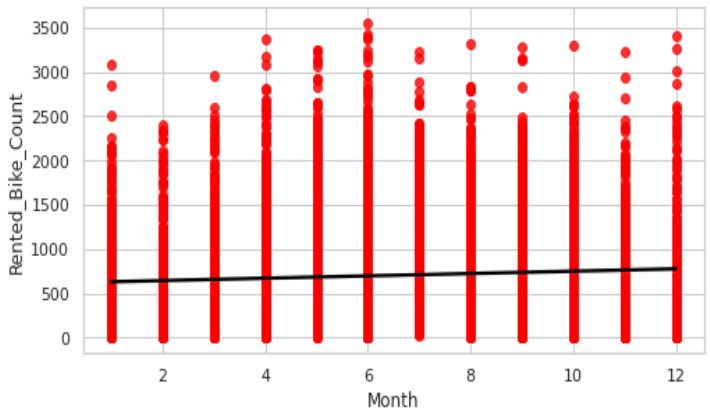
Rented Bike Count has moderate skewness toward right. We already know that the assumption of linear regression tells us that the distribution of dependent variable has to be normal, hence we should perform some operation to make it normal.

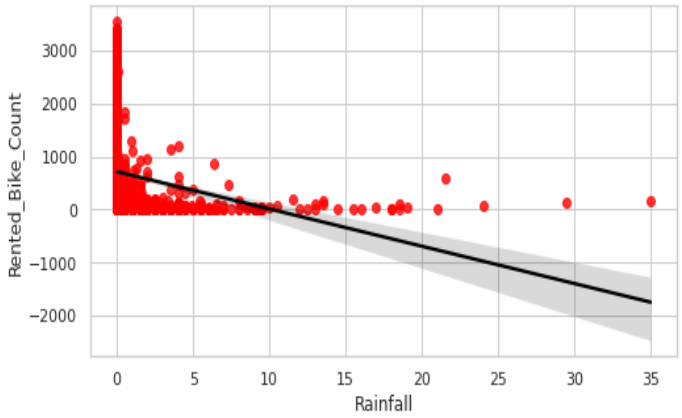
4. Regression Plotting of Features

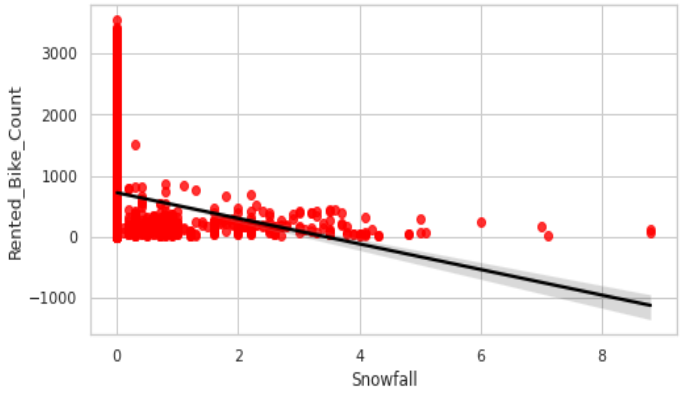


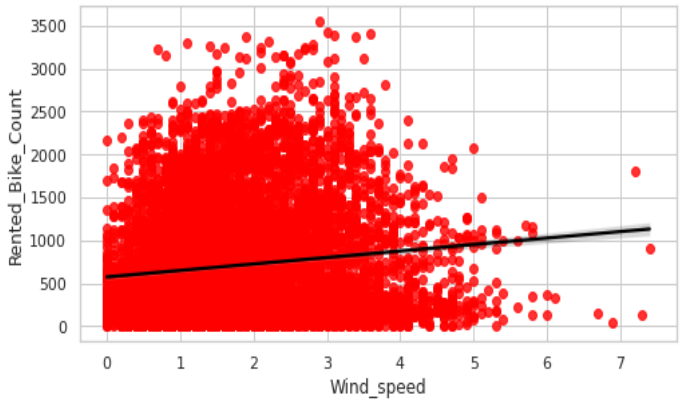


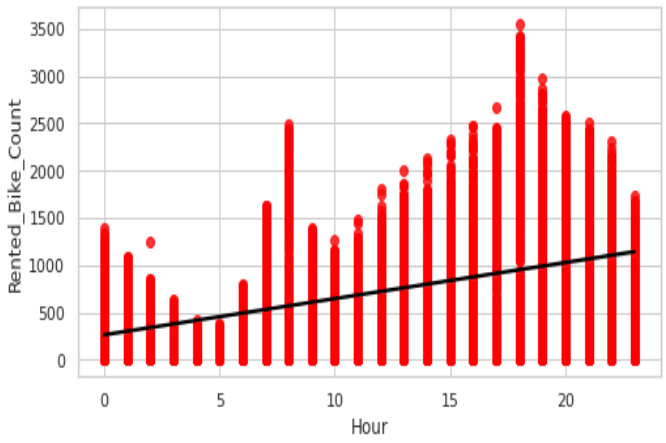


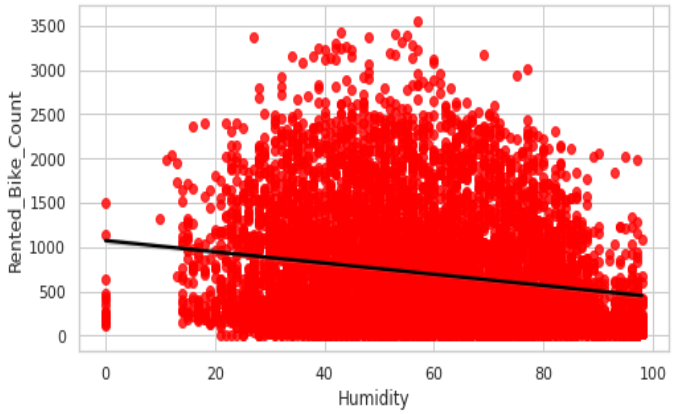


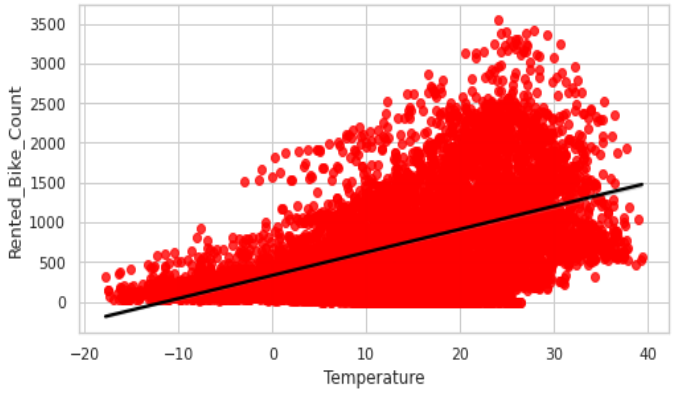

Rented_Bike_Count
3000
2000
1000
0
−1000
−2000
0
5
10
15
20
25
30
35
Rainfall


Rented_Bike_Count
3000
2000
1000
0
−1000
0
2
4
6
8
Snowfall


Rented_Bike_Count
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
Wind_speed


Rented_Bike_Count
3500
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
Hour





- The columns 'Hour', 'Temperature', 'Wind speed', 'Visibility', and 'Solar_Radiation' are positively related to the dependent variable. Which means that the rented bike count increases with increase of these features.
- Whereas, the columns 'Rainfall', 'Snowfall', 'Humidity' are those features which are negatively related with the dependent variable, which implies that the rented bike count decreases when these features increases.

5. Creating Dummy Variable and Performing the One Hot Encoding to the Dataset

For the model fitting of our data, we create some dummy variables and perform the one hot encoding to the dataset.

6. Model Training

Splitting the dataset into training and testing dataset, for the purpose of train the model.

7. Model Building

To get the best model, the training data needs to be passed into these models. We also use the cross validation technique (Hyperparameter). After passing data into these models, we get different evaluation matrices.

Here the list of models that we perform.

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

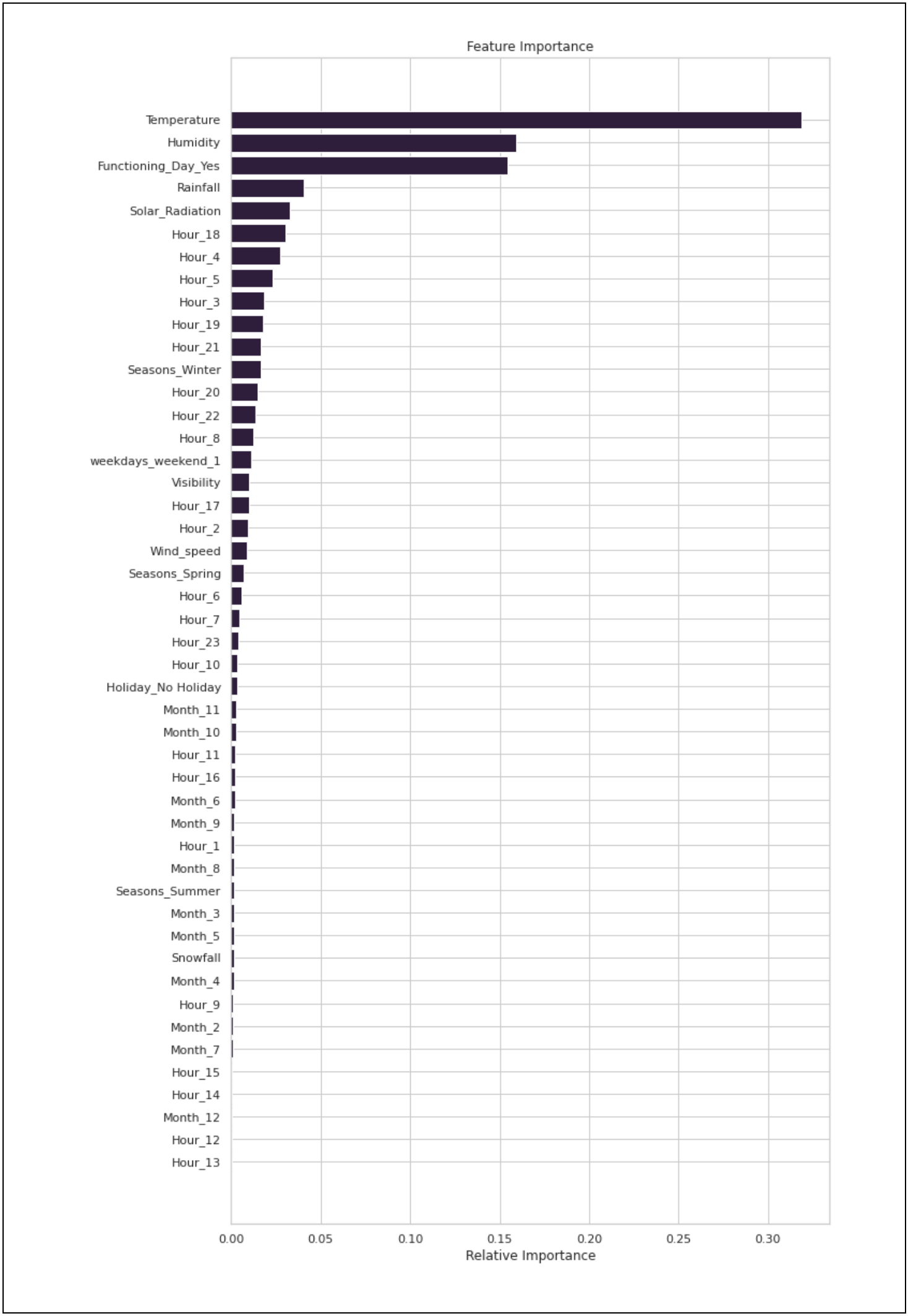
8. Different Evaluation Matrices for Different Models.

		Model	MAE	MSE	RMSE	R2	Adj_R2
Training set	0	Linear regression	4.658	37.606	6.132	0.756	0.75
	1	Ridge regression	4.659	37.608	6.133	0.756	0.75
	2	Lasso regression	4.658	37.608	6.133	0.756	0.75
	3	Elasticnet regression	4.660	37.611	6.133	0.756	0.75
	4	Decision tree regression	4.043	31.921	5.650	0.793	0.79
	5	Random forest regression	3.552	23.851	4.884	0.845	0.84
	6	Gradient Boosting Regression	1.331	3.688	1.921	0.976	0.98
Test set	0	Linear regression	4.658	36.645	6.053	0.768	0.76
	1	Ridge regression	4.661	36.664	6.055	0.768	0.76
	2	Lasso regression	4.659	36.647	6.054	0.768	0.76
	3	Elasticnet regression	4.662	36.674	6.056	0.768	0.76
	4	Decision tree regression	4.731	44.565	6.676	0.718	0.71
	5	Random forest regression	3.877	28.994	5.385	0.816	0.81
	6	Gradient Boosting Regression	2.546	14.252	3.775	0.910	0.91

1. Out of all above models Gradient Boosting Regressor gives the highest R2 score of 98% for Train Set and 91% for Test set.
2. No overfitting is seen.
3. We can deploy Gradient Boosting Regressor model.

9. Feature Importance

- Here we can say that temperature has a highest weightage.
- Best model is Gradient Boosting Regression. For this model very important feature is ‘temperature’ which we can see in the ‘Feature Importance Graph’.


Feature Importance
Temperature
Humidity
Functioning_Day_Yes
Rainfall
Solar_Radiation
Hour_18
Hour_4
Hour_5
Hour_3
Hour_19
Hour_21
Seasons_Winter
Hour_20
Hour_22
Hour_8
weekdays_weekend_1
Visibility
Hour_17
Hour_2
Wind_speed
Seasons_Spring
Hour_6
Hour_7
Hour_23
Hour_10
Holiday_No Holiday
Month_11
Month_10
Hour_11
Hour_16
Month_6
Month_9
Hour_1
Month_8
Seasons_Summer
Month_3
Month_5
Snowfall
Month_4
Hour_9
Month_2
Month_7
Hour_15
Hour_14
Month_12
Hour_12
Hour_13
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Relative Importance

Conclusion

In all of these models, our accuracy ranges from 75% to 98% which can be said to be good for such a large dataset.

This performance could be due to various reasons like the proper pattern of data, large data, or because of the relevant features.

After performing variable importance analysis to find the most significant variables for all the models developed with the given data sets.

We are getting the best results from Gradient Boosting Regression.

For most of the models, temperature and Functioning days were ranked as the most influential variable to predict the rental bike demand at each hour.

Out of all above models Gradient Boosting Regression gives the highest R2 score of 98% for Train Set and 91% for Test set.

No overfitting is seen.

We used 7 Regression Models to predict the bike rental count at any hour of the day - Linear Regression, Ridge, Lasso, Random Forest, Elastic Net, Gradient Boost and Decision Tree Regression