Capstone Project

Bike Sharing Demand Prediction

TEAM DETAILS:
SHRUNGA M
SNEHA H V

Steps Performed

1. Defining the problem statement
2. EDA and Feature Engineering
3. Preparing dataset for modeling
4. Applying the Model
5. Model Evaluation and Selection

Problem Statement

Rental Bikes are being introduced in many urban cities in recent times for the enhancement of mobility comfort. The key factor that decides the profitability of this industry is predicting the number of bikes that would be needed in the next few hour/hours so that the incoming customer demand can be fulfilled. Let's see how this can be accomplished in the coming sections.

Data Pipeline (Contd.)

- **Data Exploration**

Here we have familiarized ourselves with the given dataset. We have taken a look at all the columns, their datatypes and their statistics such as mean/median etc.

- **Exploratory Data Analysis**

Data Preprocessing – Here we have created new columns that seemed useful for our analysis going forward.

Analysis of Dependent and Independent Variable – After Preprocessing we have analyzed every column present in the dataset to identify the trend and relation of them with the dependent variable.

Data Pipeline

- **Model Creation and Evaluation**

This is the last but the most important step as this is where we will create ML models and evaluate their performance. We have created some of the ML Regression Models and performed model training and evaluation. After evaluation, we have tuned them to achieve best predictions.

Conclusion

- **In this final step we have compared all the models and concluded which model best predicts our data. Also, the variables that plays major role is prediction are identified.**

Data Summary

Our Dataset has 8760 rows and 14 columns to begin with. We have the data of every hour for one year from Dec 2017 to Nov 2018. In the preprocessing we have added 4 more columns. Totally we have a dataset of shape (8760 x 17).

The dataset is pretty much clean with no missing values. The columns in the dataset are as follows:

Columns Present in the dataset

•**Date** - Date is in the format of day/month/year. We have data from 2017 Dec to 2018 Nov i.e, 1 year of data.
•**Rented Bike count** - This is our dependent variable. And it gives the information about number of bikes rented per hour.
•**Hour** - We have values from 0-23 (24 hour format) i.e, we have data for each and every hour.
•**Temperature** - Temperature is in Celsius and it gives the temperature reading for every hour.

Data Summary (Contd.)

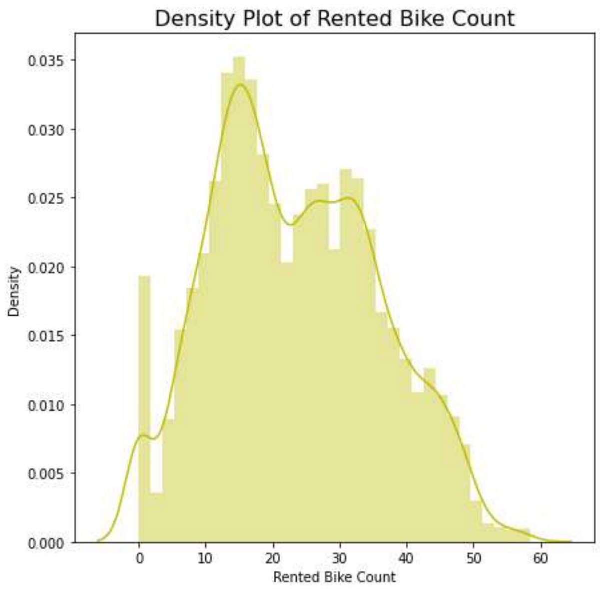
- **Humidity** - Humidity is the amount of water vapor in the air and it is measured in % here.
- **Windspeed** - The speed of wind is measured in m/s
- **Visibility** - Visibility is a measure of the horizontal opacity of the atmosphere at the point of observation and is expressed in terms of the horizontal distance at which a person should be able to see and identify.
- **Dew point temperature** - It is the temperature to which air must be cooled to become saturated with water vapor, assuming constant air pressure and water content. And it is measured in Celsius.
- **Solar radiation** - It is an electromagnetic radiation emitted by the sun. And it is measured by MJ/m2.
- **Rainfall** - It is measured in mm. And it gives the rainfall reading for every hour.
- **Snowfall** - It is measured in cm. And it gives the snowfall reading for every hour.
- **Seasons** - We have 4 different seasons in dataset. They are - Winter, Spring, Summer, Autumn.
- **Holiday** - Gives information about that day whether that day is holiday or not.
- **Functioning Day** - Gives information about that day whether that day is functional day or non-functional day.

Calculated Columns

- **month** – month value extracted from the date column
- **year** – Year value extracted from date column.
- **day** – day value extracted from day column.
- **weekday** – binary column indicating if the day is weekday or weekend.

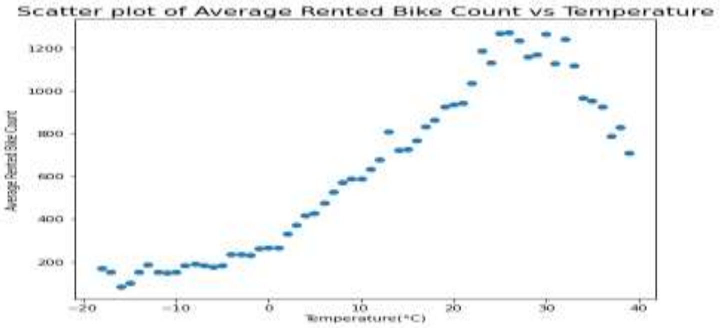
Dependent Variable Analysis

Rented Bike Count is our dependent variable. This is the number of bikes rented on the given day on the given environmental conditions. It is a continuous variable.

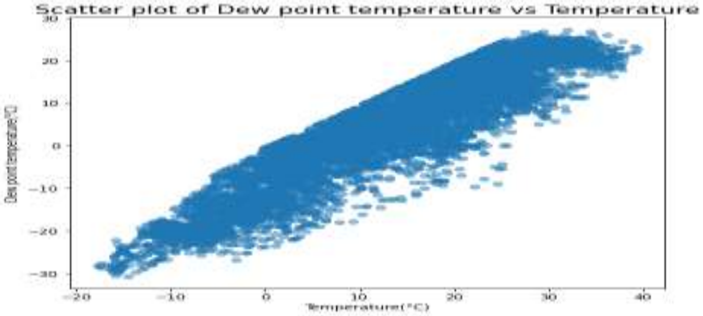


After applying square root to the values, there is no prominent skewness and outliers observed.

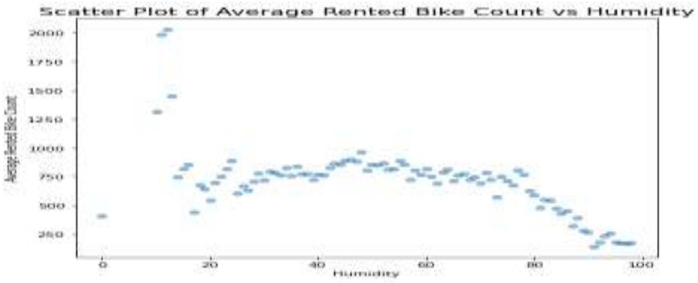
Exploratory Data Analysis



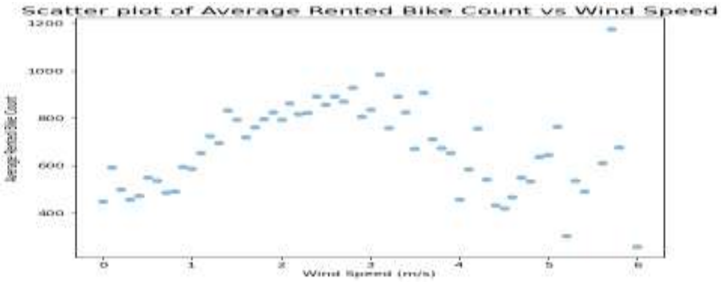
Temperature has positive correlation with Rented bike count



Dew point temp. and Temperature are linear

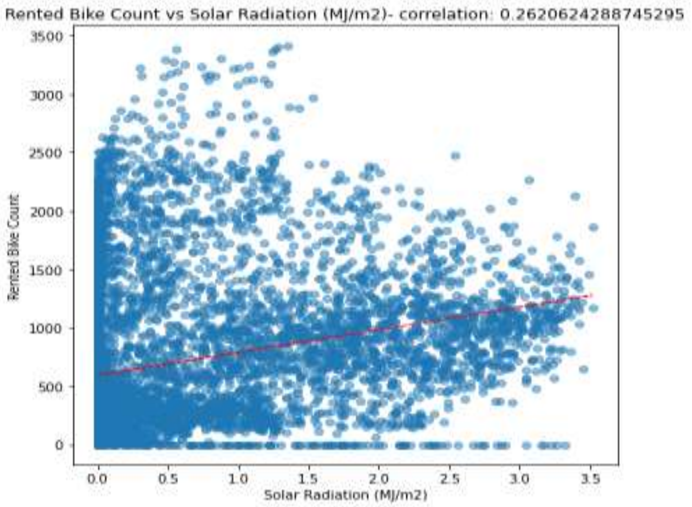


As the humidity rises above 70% bike count drops sharply



Windspeed has sinusoidal relation with Rented bike count

Exploratory Data Analysis (Contd.)



There is a positive correlation between solar radiation and rented bike count

```
[ ] (dataset['Snowfall (cm)'].value_counts().head())/len(dataset)*100
```

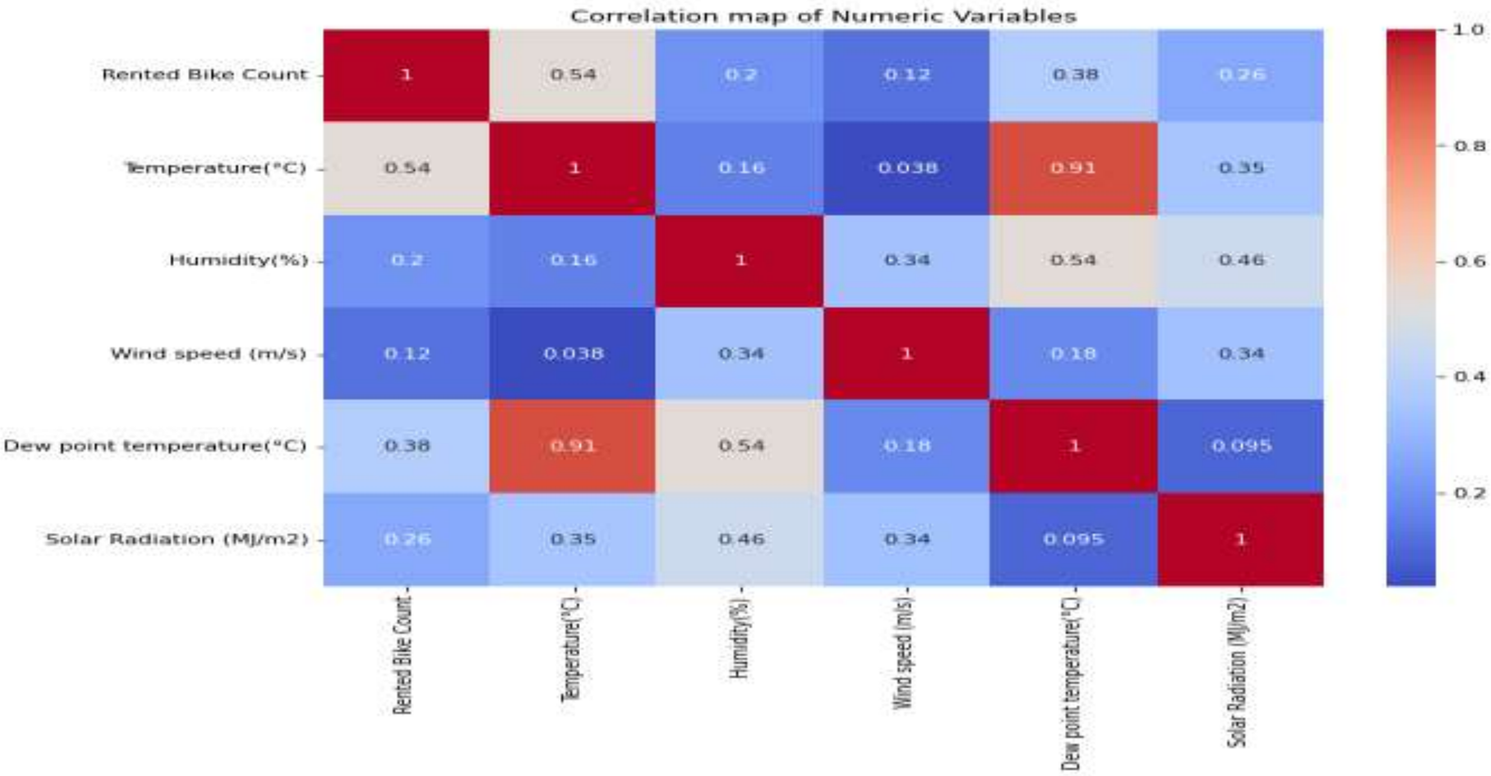
```
0.0    94.950303
0.3     0.479835
1.0     0.434137
0.9     0.388438
0.5     0.388438
Name: Snowfall (cm), dtype: float64
```

```
(dataset['Rainfall(mm)'].value_counts().head())/len(dataset)*100
```

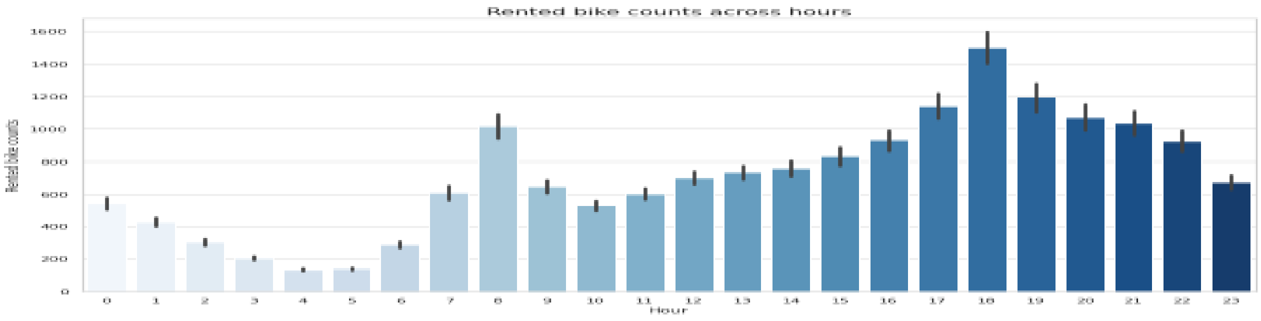
```
0.0    94.002056
0.5     1.313835
1.0     0.754027
1.5     0.639781
0.1     0.525534
Name: Rainfall(mm), dtype: float64
```

Most of the entries in Rainfall and Snowfall are zeros.

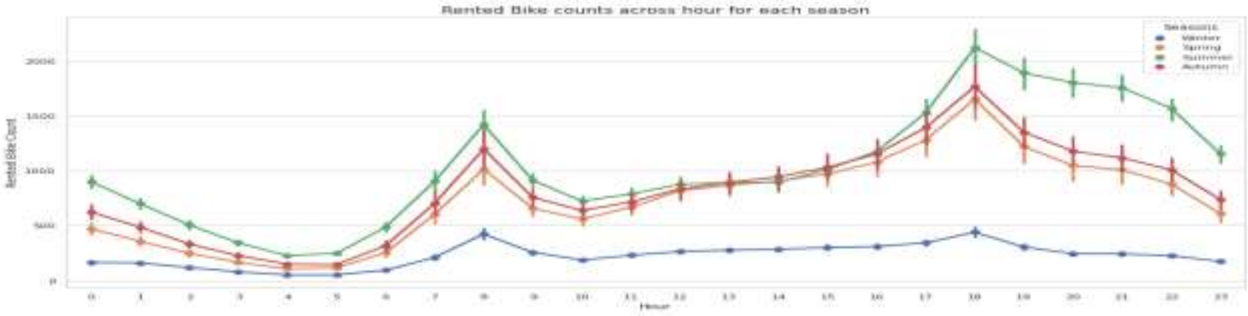
Exploratory Data Analysis (Contd.)



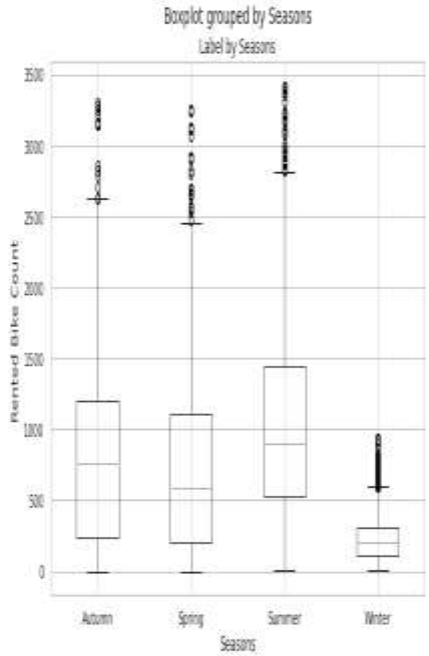
Exploratory Data Analysis (Contd.)



Demand for Rented bikes increases in peak traffic hours

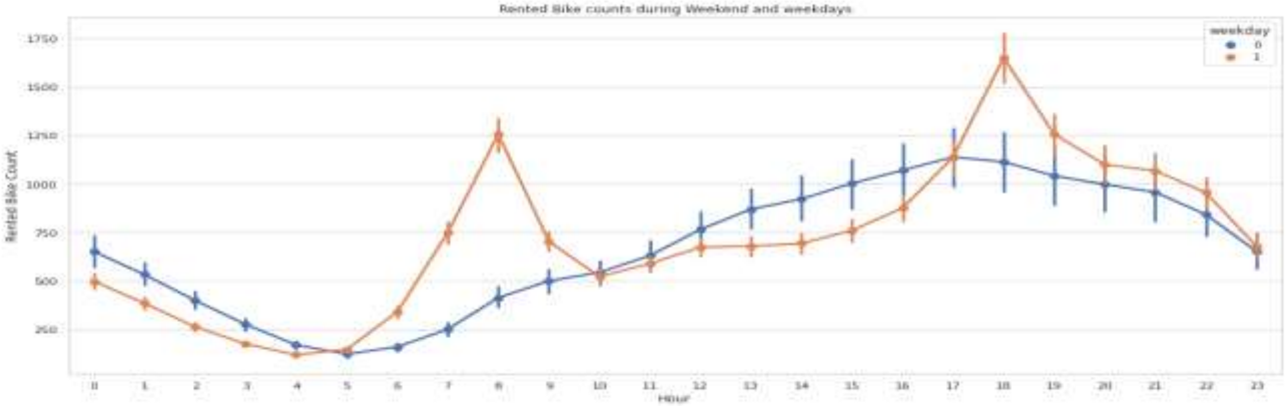


Demand for bikes is very low in Winter and highest in Summer

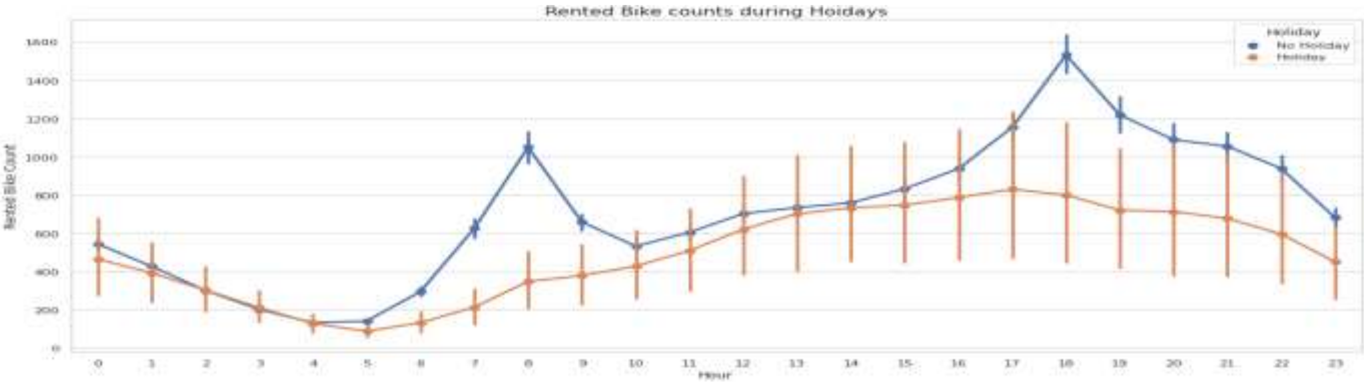


Rented bike count in each season

Exploratory Data Analysis (Contd.)

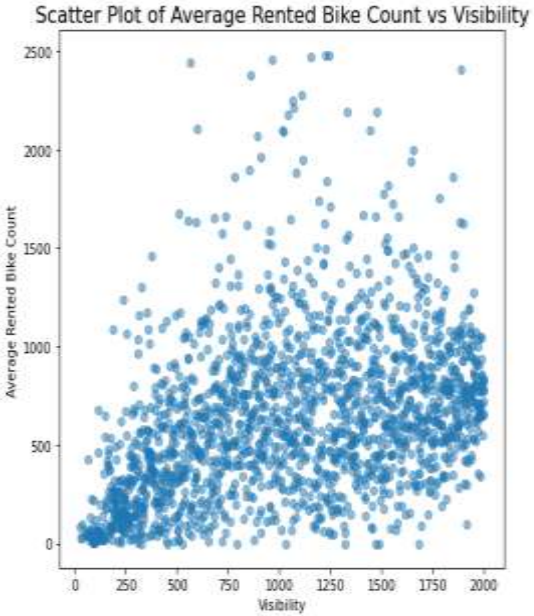


Rental bike counts will be more during Peak hours in Weekdays, and in weekends, rental bike counts will keep on raising during evening times.



Rental bike counts will be more during Peak hours in Holidays, and in non holidays, rental bike counts will keep on raising during evening times.

Exploratory Data Analysis (Contd.)



Visibility has no relation with demand in Rental Bikes

Rented bikes are used only during functioning days.

Preparing Dataset for Modeling

```
Data columns (total 38 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Rented Bike Count           8753 non-null   int64
 1   Temperature(°C)             8753 non-null   float64
 2   Humidity(%)                 8753 non-null   int64
 3   Wind speed (m/s)            8753 non-null   float64
 4   Dew point temperature(°C)   8753 non-null   float64
 5   Solar Radiation (MJ/m2)     8753 non-null   float64
 6   Holiday                     8753 non-null   int64
 7   Functioning Day             8753 non-null   int64
 8   month                       8753 non-null   int64
 9   weekday                     8753 non-null   object
 10  season_Autumn               8753 non-null   uint8
 11  season_Spring               8753 non-null   uint8
 12  season_Summer               8753 non-null   uint8
 13  season_Winter               8753 non-null   uint8
 14  hour_0                      8753 non-null   uint8
 15  hour_1                      8753 non-null   uint8
 16  hour_2                      8753 non-null   uint8
 17  hour_3                      8753 non-null   uint8
 18  hour_4                      8753 non-null   uint8
 19  hour_5                      8753 non-null   uint8
 20  hour_6                      8753 non-null   uint8
 21  hour_7                      8753 non-null   uint8
 22  hour_8                      8753 non-null   uint8
 23  hour_9                      8753 non-null   uint8
 24  hour_10                     8753 non-null   uint8
 25  hour_11                     8753 non-null   uint8
 26  hour_12                     8753 non-null   uint8
 27  hour_13                     8753 non-null   uint8
 28  hour_14                     8753 non-null   uint8
 29  hour_15                     8753 non-null   uint8
 30  hour_16                     8753 non-null   uint8
 31  hour_17                     8753 non-null   uint8
 32  hour_18                     8753 non-null   uint8
 33  hour_19                     8753 non-null   uint8
 34  hour_20                     8753 non-null   uint8
 35  hour_21                     8753 non-null   uint8
 36  hour_22                     8753 non-null   uint8
 37  hour_23                     8753 non-null   uint8
dtypes: float64(4), int64(5), object(1), uint8(28)
```

Task : Regression

Train set : (7002, 36)

Test set : (1751, 36)

Response : Continuous variable(predictions of Rented Bike Count)

Linear Regression Baseline Model

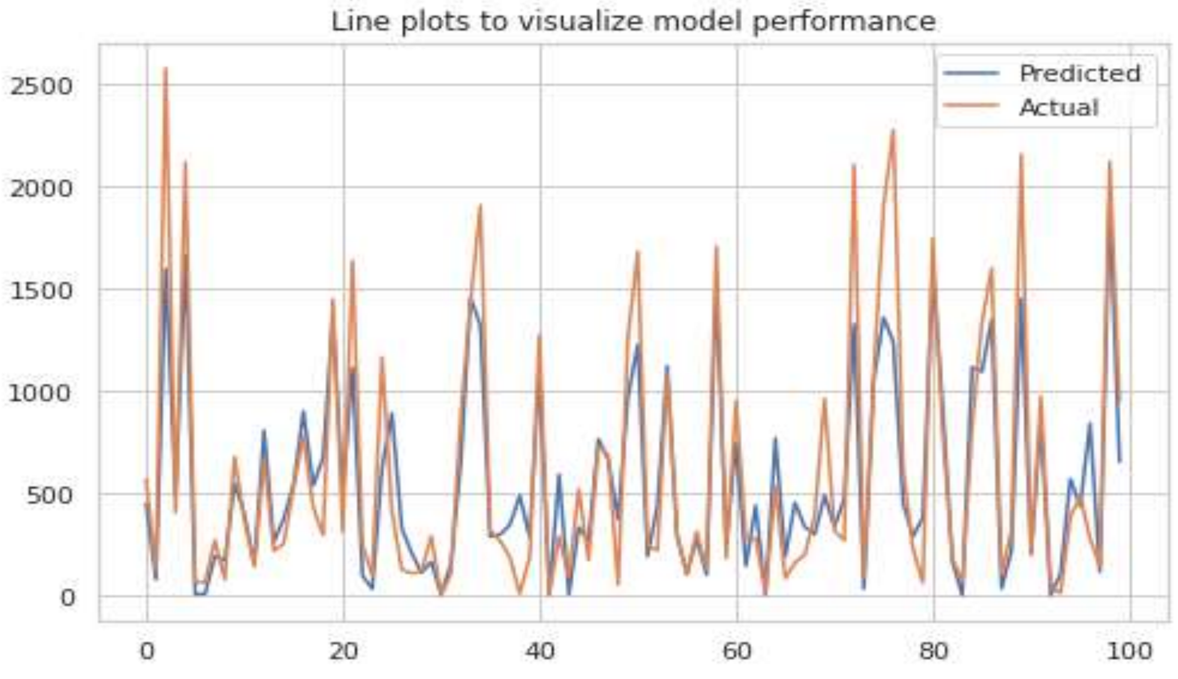
Train :-
MSE : 122281.54226712306
RMSE : 349.6877782638722

Test :-
MSE : 129299.90773112362
RMSE : 359.5829636274828

****

Train :-
R2 : 0.7023660593473617
Adjusted R2 : 0.7008276786060128

Test :-
R2 : 0.7007630130368323
Adjusted R2 : 0.6944779888065674



Regularized Linear Regression

Lasso Regression Scores

Train :-
MSE : 122281.54226712306
RMSE : 349.6877782638722

Test :-
MSE : 129353.08007191615
RMSE : 359.6568921512782

****

Train :-
R2 : 0.7023660593473617
Adjusted R2 : 0.7008276786060128

Test :-
R2 : 0.7006399570244366
Adjusted R2 : 0.6943523481871435

Ridge Regression Scores

Train :-
MSE : 122445.55419891152
RMSE : 349.9222116398322

Test :-
MSE : 129502.28150346558
RMSE : 359.8642542730044

****

Train :-
R2 : 0.7019668534110675
Adjusted R2 : 0.7004264092937378

Test :-
R2 : 0.700294662216336
Adjusted R2 : 0.6939998009793396

Elastic Net Regression Scores

Train :-
MSE : 122331.0845807845
RMSE : 349.7586090159677

Test :-
MSE : 129365.7943897578
RMSE : 359.674567337972

****

Train :-
R2 : 0.7022454730857661
Adjusted R2 : 0.700706469070129

Test :-
R2 : 0.7006105324546206
Adjusted R2 : 0.694322305598358

As seen in the previous slides, neither of linear regression variants could perform well on the data, we will apply tree-based regression algorithms.

Model Evaluation and Selection

	Model	MSE	RMSE	R2_Score	Adjusted_R2
0	Linear Regression	122281.542	349.688	0.702	0.701
1	Lasso Regression	122281.542	349.688	0.702	0.701
2	Ridge Regression	122445.554	349.922	0.702	0.700
3	Elastic Net Regression	122331.085	349.759	0.702	0.701
4	Random Forest Regression	23314.696	152.692	0.943	0.943
5	Gradient Boosting Regression	17849.574	133.602	0.957	0.956
6	XG Boost Regression	5272.463	72.612	0.987	0.987

← **Evaluation metrices for training data**

Evaluation metrices for test data →

	Model	MSE	RMSE	R2_Score	Adjusted_R2
0	Linear Regression	129299.908	359.583	0.701	0.694
1	Lasso Regression	129353.080	359.657	0.701	0.694
2	Ridge Regression	129502.282	359.864	0.700	0.694
3	Elastic Net Regression	129365.794	359.675	0.701	0.694
4	Random Forest Regression	46041.649	214.573	0.893	0.891
5	Gradient Boosting Regression	38227.795	195.519	0.912	0.910
6	XG Boost Regression	40597.381	201.488	0.906	0.904

Model Evaluation and Selection (Contd.)

Observations:

1. **As seen previously Linear Regression is not able to give good/dependable accuracy in predictions.**

2. **Gradient Boost and XG Boost Regressors are able to track most of the data variance.**

3. **From the above we can conclude that Gradient Boosting algorithm tracks most of the data variance without overfitting.**

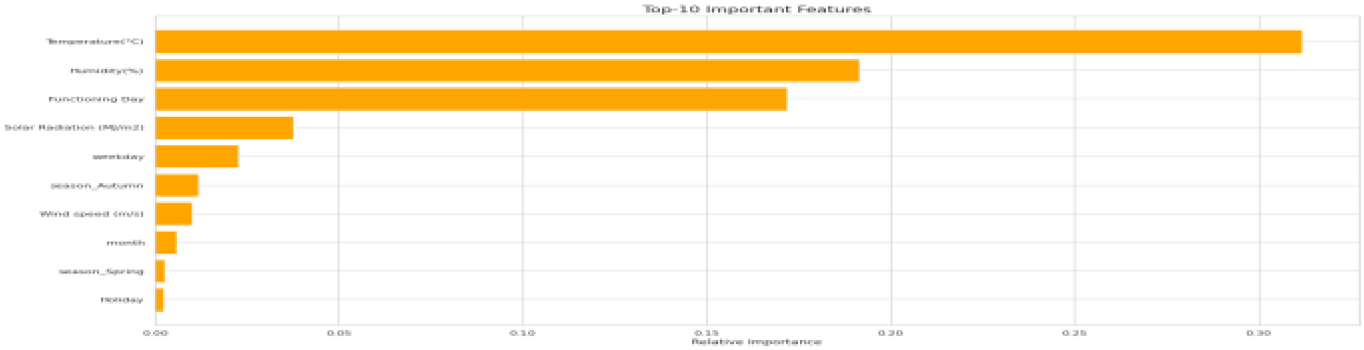
Model Evaluation and Selection (Contd.)

We have chosen GB Regressor as our ML model to predict the Rental Bike demands.

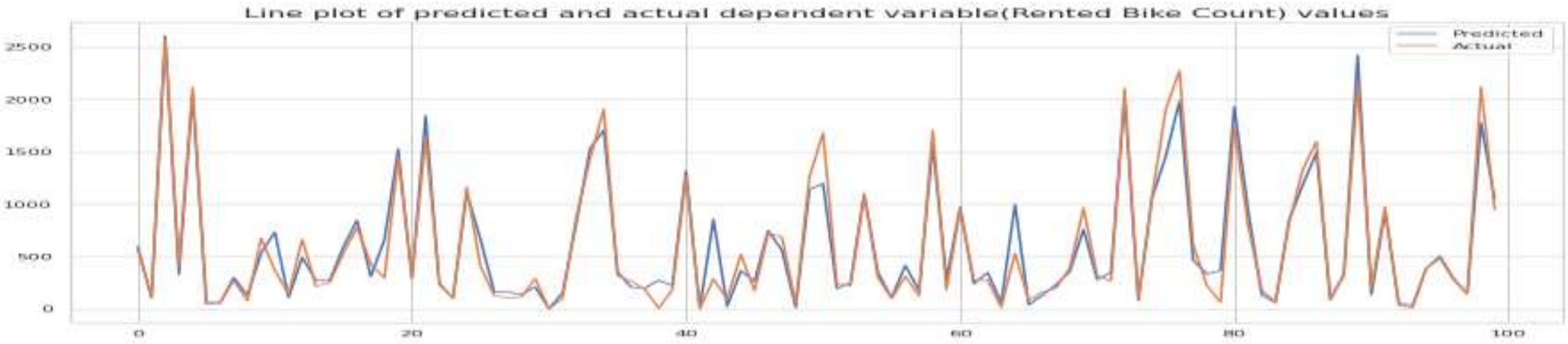
The Best-Fit Hyperparameters are:

max_depth: 9,
min_samples_leaf: 30,
min_samples_split: 30,
n_estimators: 150

Model Evaluation and Selection (Contd.)



Temperature is the dominant feature that explains the demand in Rental Bikes



Conclusion

- **Temperature(°C), Humidity(%) and Functioning Day were found to be the predominant features in predicting the Rental Bike Count. Our current model tries to predict the number of rental bike needed on a given day and hour to satisfy the customer needs.**

- **Linear Regression and its variants were not able to track the data variance with the expected accuracy.**

- **Tree-Based Regression models such as Gradient Boosting and XG Boost models are predicting the Rented bike counts with 91% R2 Score.**

- **Knowing the number of bikes needed to meet the customer demand before hand, helps the companies stock the appropriate number of bikes and offer seamless supply to the customers which will increase the trust on the company and maximize profit.**

Challenges

- **Feature Selection**

  Most of the features doesn′t have considerable correlation with the dependent variable.

- **Computational Time**

  Multiple iterations are run on a single model to tune the hyperparameters.

Q & A

Thank You