Capstone Project - 2

Bike Sharing Demand Prediction

By-Dany Chitturi

Let’s Catch The Defaulters

1. Defining problem statement
2. Exploratory Data Analysis
3. Feature engineering
4. Preparing dataset for modelling
5. Model Building
6. Hyperparameter tuning
7. Model validation and Selection

Problem Statement

Nowadays Rental bike sharing is becoming popular because of the increased comfortableness and availability. It is necessary to make rental bikes available and accessible to the public at the right time as it lessens the waiting time. So providing a stable supply of rental bikes to the public will be a very big challenge. The idea of this project is to create a predictive model that predicts the Rental Bike Count. To accomplish this, We organized the whole series into five parts as follows :

Data Pipeline

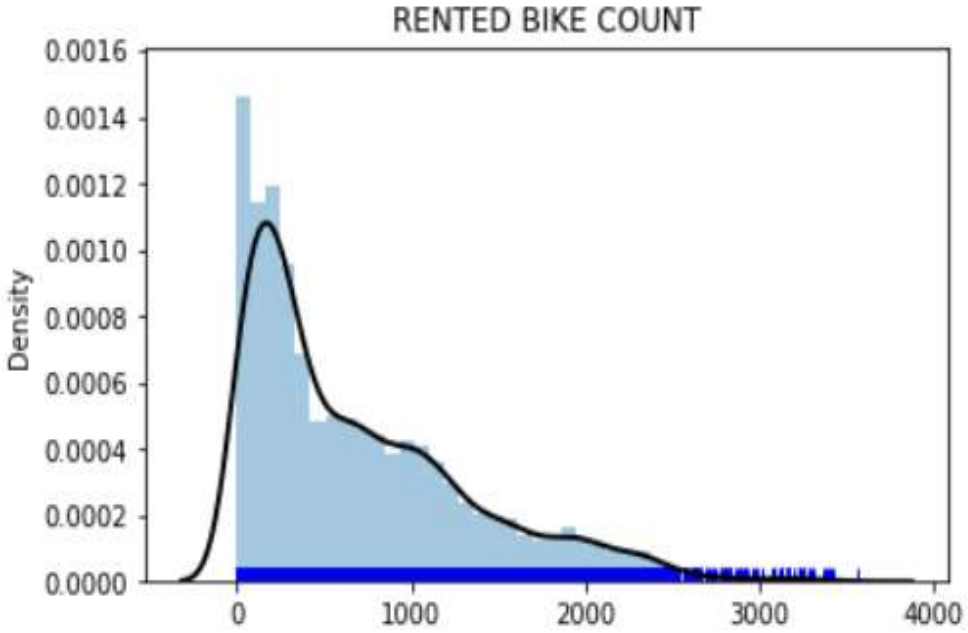
- **Data processing-1:** In this part we've checked and handled the null values and also checked for duplicated records
- **Data processing-2:** In this part, we performed outlier detection, encoded the categorical features, and changed the column containing date values
- **EDA:** In this part, we do some exploratory data analysis(EDA) on all the features in our dataset to uncover the relationship between dependent and independent variable(s).
- **Feature Engineering:** In this part, we've created day, month and year columns from date column, And also applied square root transformation on dependent variable to make it as close to Normal distribution
- **Model building :** Finally, we create five different models. We start with a simple model , then use hyperparameter tuning to get the best optimal parameters.

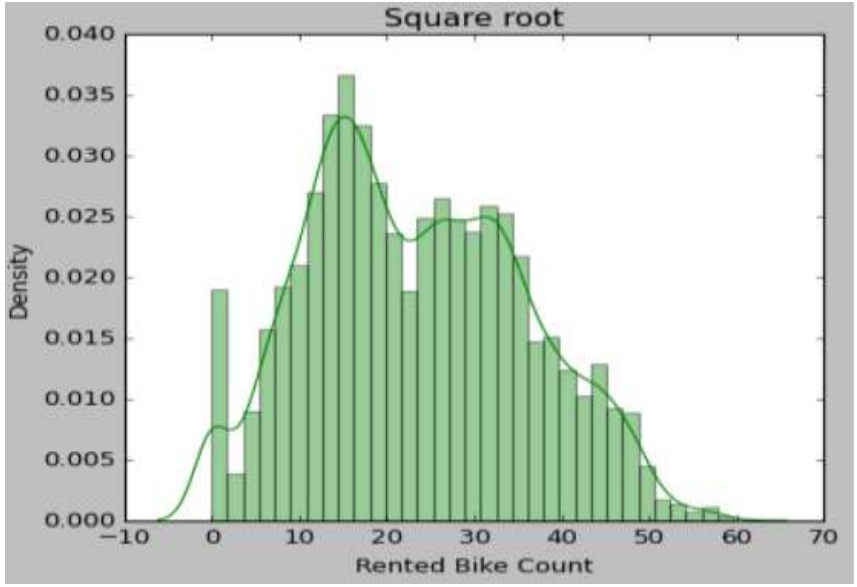
Data Summary

Date: day-month-year
Hour: the hour of the day
Temperature: temperature in Celsius at particular hour
Humidity(%): percentage of humidity in air at particular hour
Wind speed(m/s): wind speed at particular hour
Visibility(m): the distance one can see as determined by weather conditions
Dew point temperature: dew point temperature in Celsius
Solar radiation: the amount of solar radiation (MJ/m2)
Rainfall: total rainfall depth in mm
Snowfall: depth of snowfall in cm
Seasons: winter, spring, summer, autumn
Holiday: holiday/no holiday
Functional Day: yes/no

Define Dependent Variable

Rental Bike Count : count of bikes rented at each hour

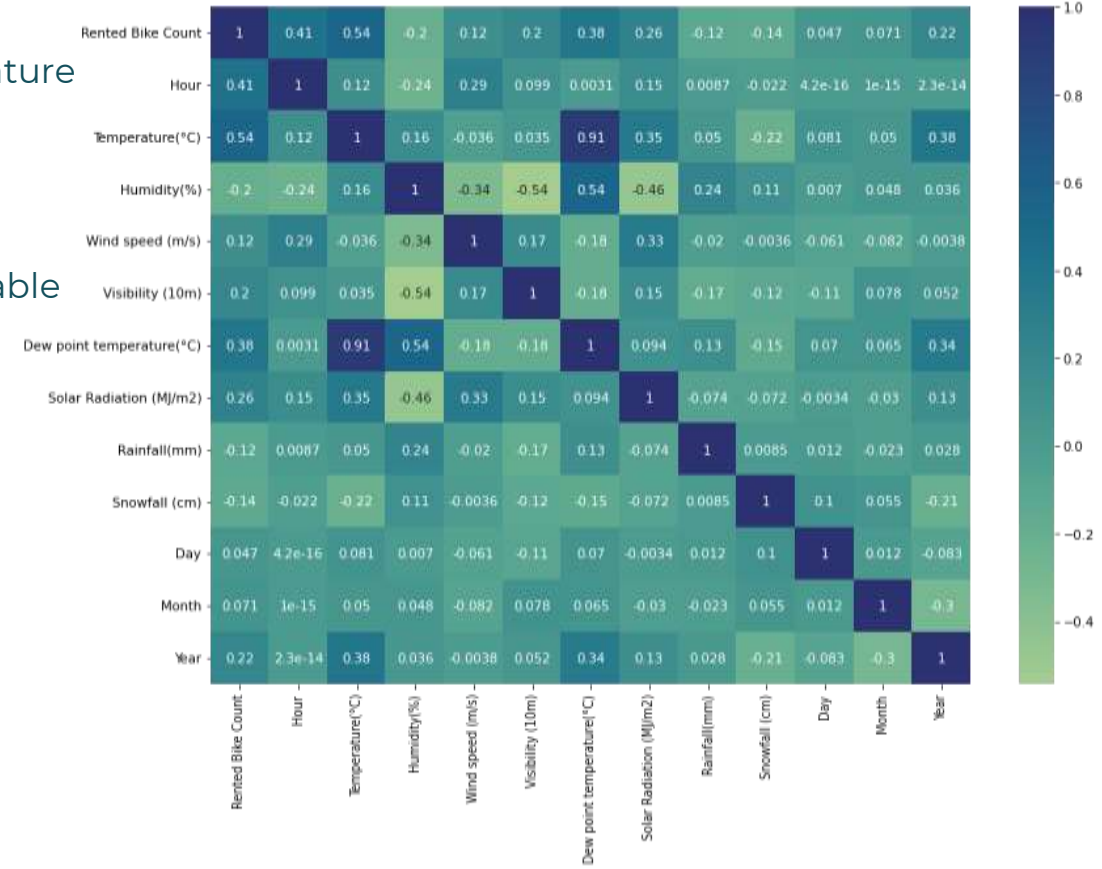




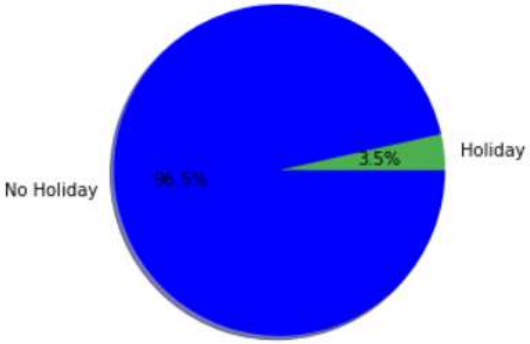
EDA

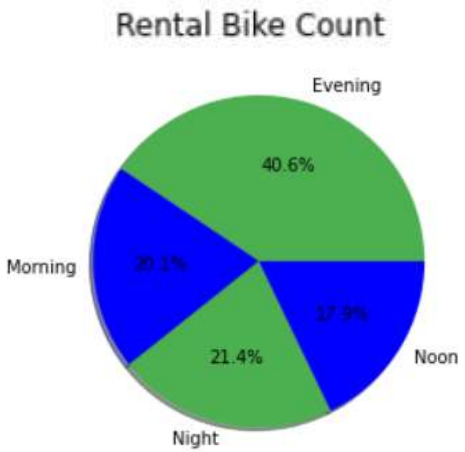
- From the correlation heatmap we can see that Temperature and Dew point temperature are highly correlated to each other

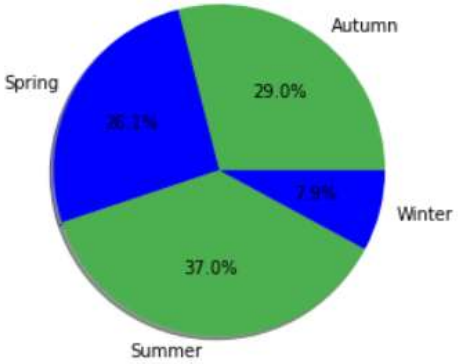
- To avoid multicollinearity, remove Dew point temperature, because it has weak correlation with the dependent variable when compared with Temperature.



EDA

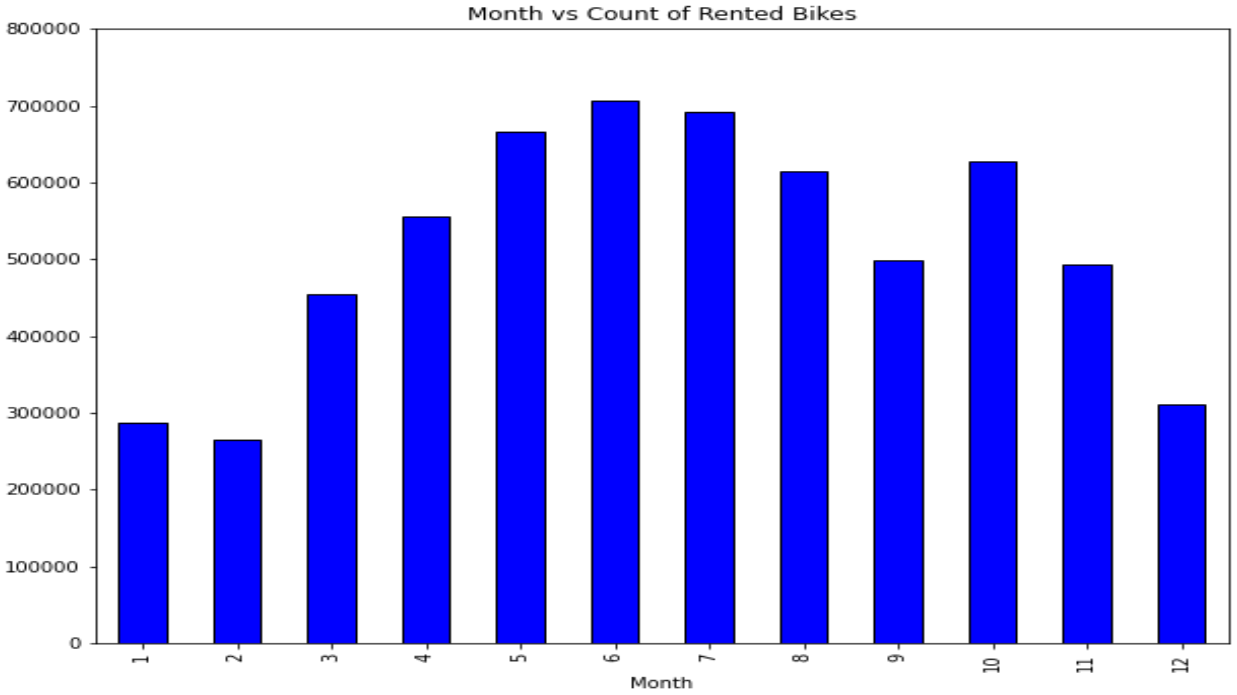






- Demand for rental bikes is high on working days and very low on holidays
- The highest number of bike rents occur in summer while the least bike rents occur in winter
- Demand for rental bikes is high in the evening

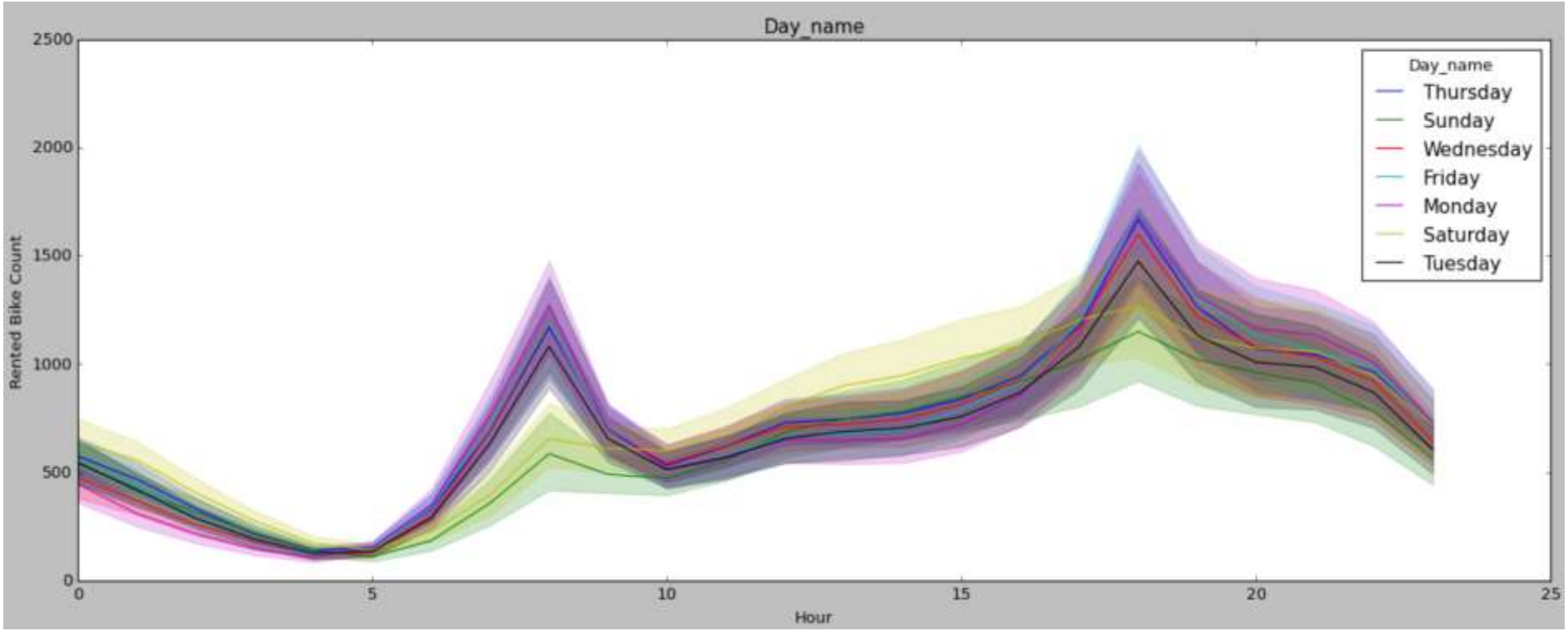
EDA (continued)



- We can see that most numbers of rental bike bookings are during the month's May, June and July

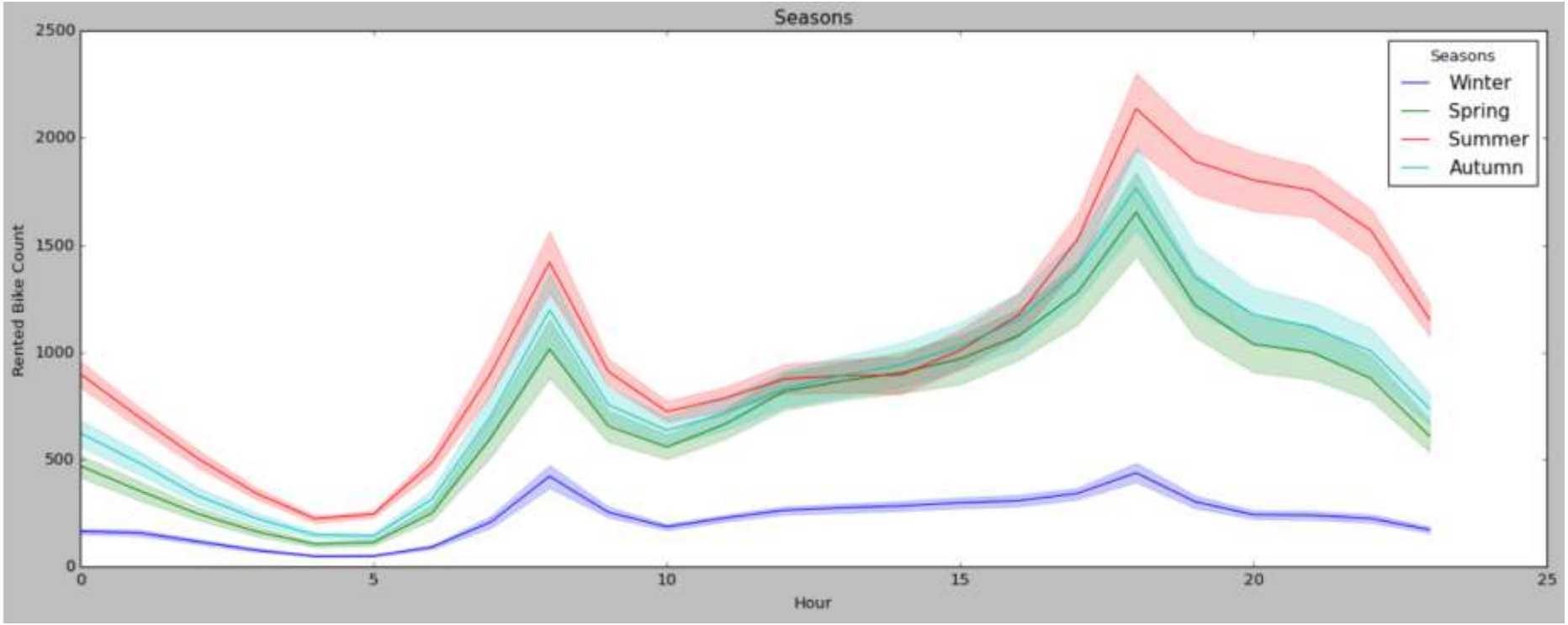
- Months November, December, January and February have the least number of bookings

EDA (continued)



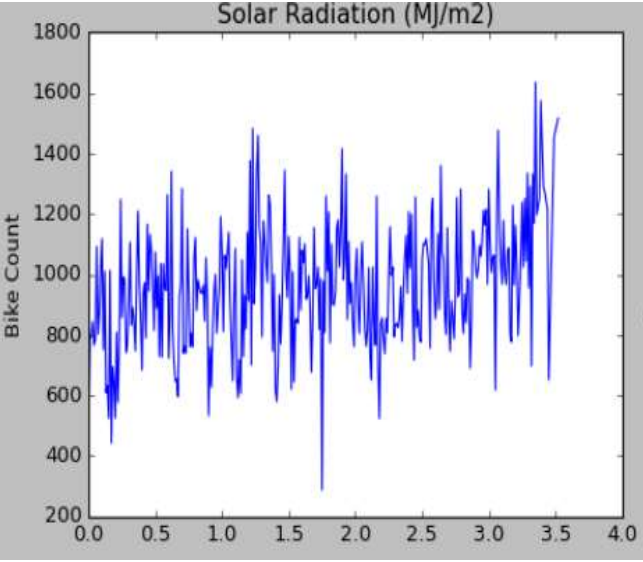
- Demand is high during weekdays, But during weekends the number of bike rents is more between 12.00 to 16.00 hrs than weekdays. On day to day basis, the trend of bike rents is almost similar with slight peaking demands on Thursday while drops on Sunday.

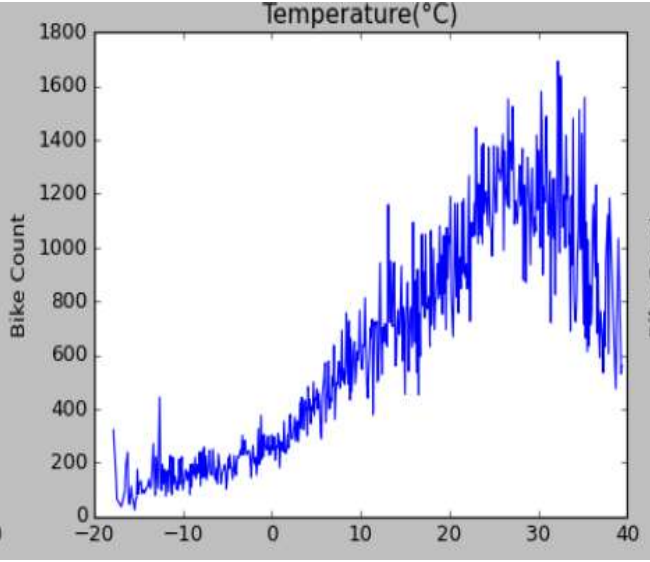
EDA (continued)

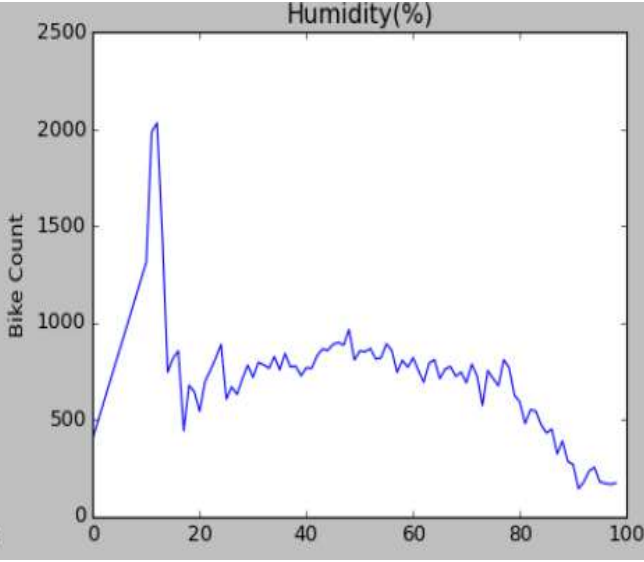


- The number of bike rents peaked at 8’o clock in the morning and 6’o clock in the evening, this may be because of employees who go to work in the morning and return from work in the evening

EDA (continued)

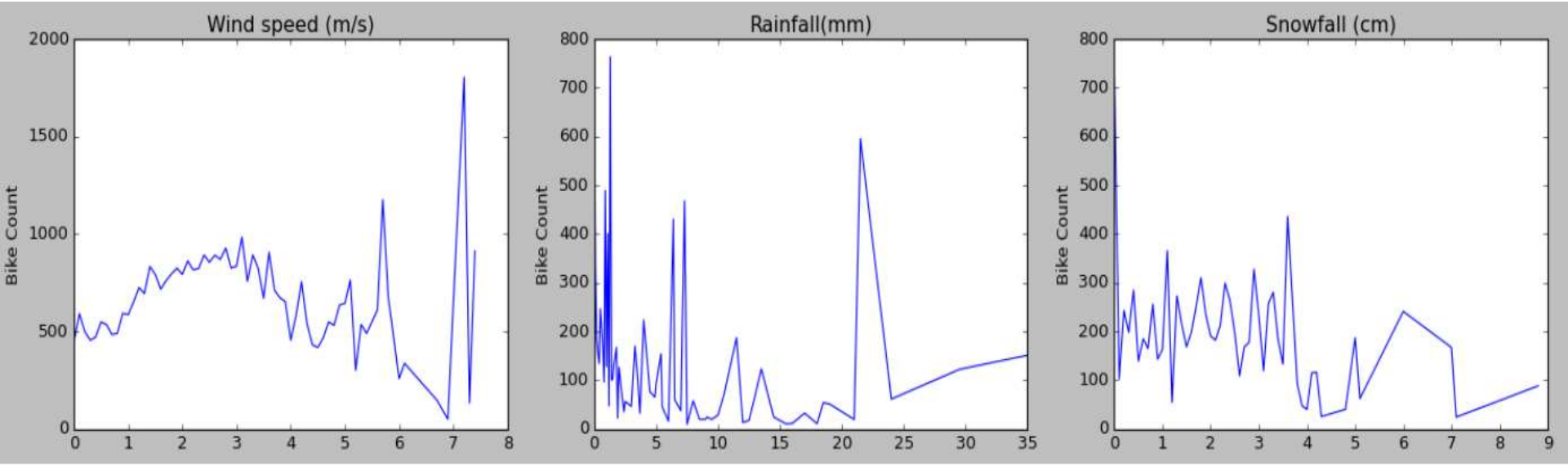






- The 1st plot shows that Rented Bike Bookings and Solar Radiation are positively correlated. From the 2nd plot we can see that in general, more people tend to prefer biking at moderate to high temperatures. however, if the temperature is too hot there is a slight decline in rental bike bookings. The bike counts peak in the afternoon when temperature is the highest, with the most visibility and least humidity

EDA (continued)



- The movement of snowfall and rainfall negatively correlates with rental bike bookings. The bike counts peak when there is minimal rainfall, minimal snowfall, and good windspeed.

Preparing dataset for modelling

Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Holiday	Functioning Day	...	Seasons_Spring	Seasons_Summer	Seasons_Winter
0	-5.2	37	2.2	2000	0.00	0.0	0.0	1	1	...	0	0	1
1	-5.5	38	0.8	2000	0.00	0.0	J.0	1	1	...	0	0	1
2	-6.0	39	1.0	2000	0.00	0.0	0.0	1	1	...	0	0	1
3	-6.2	40	0.9	2000	0.00	0.0	0.0	1	1	...	0	0	1
4	-6.0	36	2.3	2000	0.00	0.0	0.0	1	1	...	0	0	1
5	-6.4	37	1.5	2000	0.00	0.0	0.0	1	1	...	0	0	1
6	-6.6	35	1.3	2000	0.00	0.0	0.0	1	1	...	0	0	1
7	-7.4	38	0.9	2000	0.00	0.0	0.0	1	1	...	0	0	1
8	-7.6	37	1.1	2000	0.01	0.0	0.0	1	1	...	0	0	1
9	-6.5	27	0.5	1928	0.23	0.0	0.0	1	1	...	0	0	1

Selected_features= Hour, Temperature(°C), Humidity(%), Wind speed (m/s), Visibility (10m), Solar Radiation (MJ/m2)‘, 'Rainfall(mm)‘, 'Snowfall (cm)‘, 'Holiday', 'Functioning Day', 'Month‘, 'Year‘ ‘Seasons_Autumn‘, 'Season_Spring‘, 'Season_Summer', 'Season_Winter‘, 'Day_name_Friday‘, 'Day_name_Monday','Day_name_Saturday','Day_name_Sunday','Day_name_Thursday‘, ’Day_name_Tuesday‘ , 'Day_name_Wednesday‘.

Train Set:- (7008, 23) , Test Set:- (1752, 23) , Response:- Count of Rental bikes

Model Building (Baseline Model)

```
Linear Regression
**************************************************
Training_r2_score =  0.6564367793934631
Root_Mean_Squared_Error =  7.286036566385434
Mean_Squared_Error =  53.08632884670564
R-Squared =  0.6524835592012341
Adjusted_R2 =  0.6478580510193062
```

Model Validation and Selection

	Training_score	Mean_square_error	Root_Mean_square_error	R_Squared	Adjusted_R_Squared
Linear	0.656437	53.086329	7.286037	0.652484	0.647858
Lasso	0.609191	58.833721	7.670314	0.614860	0.609733
Ridge	0.656438	53.088690	7.286199	0.652468	0.647842
Decision_Tree	1.000000	26.484386	5.146298	0.826627	0.824319
Random_Forest	0.988867	12.754013	3.571276	0.916509	0.915398

Models Used: Linear Regression, Lasso Regression , Ridge Regression ,Decision Tree and Random Forest

Model Validation and Selection (continued)

Observation 1: As seen in the table above linear regression is not giving great results, Lasso had low R2 score than linear regression this is because of the shrinking of coefficients of some features to zero.

Observation 2: Linear regression and Ridge regression performed equally in terms of R2 score

Observation 3: The Training R2 score of Decision tree is 1 , whereas the test R-squared is 0.82. So we can say that the Decision Tree is overfitting the data.

Model Validation and Selection (continued)

Observation 4: Finally, The ensemble model Random Forest is performing very well with Train R2 Of 0.97 and Test R2 of 0.91

Observation 5: From the above observations, we concluded that we would choose our model from Random Forest Regressor.

Model Validation and Selection (Hyperparameter tuned)

	Mean_square_error	Root_Mean_square_error	Training_score	R_Squared	Adjusted_R_Squared
Rand_Lasso	53.075243	7.285276	0.656283	0.652556	0.647932
Rand_Ridge	53.085469	7.285978	0.656416	0.652489	0.647864
Rand_DecisionTree	20.304380	4.506038	0.953651	0.867082	0.865313
Rand_RandomForest	12.977061	3.602369	0.969158	0.915049	0.913918
Bayes_Lasso	53.075243	7.285276	0.656283	0.652556	0.647932
Bayes_Ridge	53.085877	7.286006	0.656428	0.652487	0.647861
Bayes_DecisionTree	20.135110	4.487216	0.911369	0.868191	0.866436
Bayes_RandomForest	13.003754	3.606072	0.971763	0.914874	0.913741
Grid_Lasso	53.075243	7.285276	0.656283	0.652556	0.647932
Grid_Ridge	53.085877	7.286006	0.656428	0.652487	0.647861
Grid_DecisionTree	23.277733	4.824700	0.905314	0.847618	0.845590
Grid_RandomForest	13.037360	3.610729	0.971862	0.914654	0.913518

Model Validation and Selection (continued)

We had chosen Random Forest Regressor for our prediction, and the best hyperparameters obtained are below.

n_estimators= 250
criterion = ”squared_error”
max_depth= 25
max_features= 18
min_samples_split= 10
max_leaf_nodes = None
Max_samples = None
min_impurity_decrease = 0.0
min_samples_leaf = 1
min_samples_split = 2
min_weight_fraction_leaf = 0.0

Model Evaluation

Random Forest

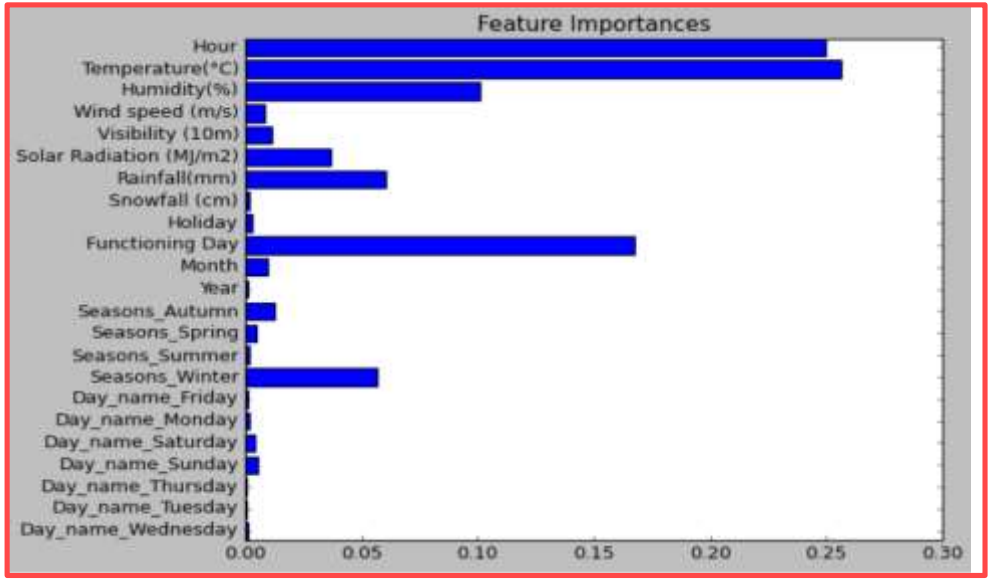
Training_r2_score = 0.972003532260945

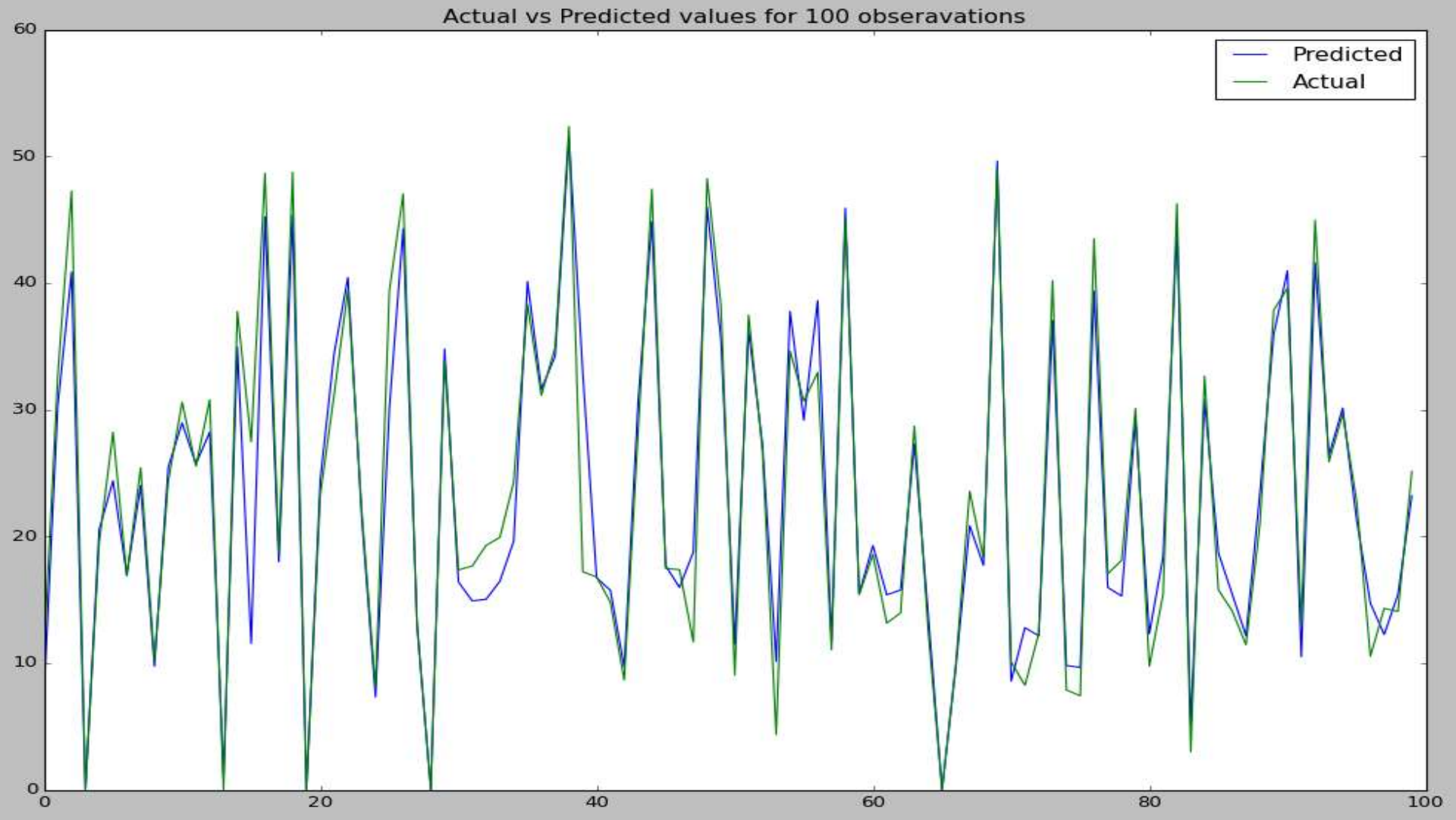
Root_Mean_Squared_Error = 3.5888658226502095

Mean_Squared_Error = 12.879957892986766

R-Squared = 0.915684560943481

Adjusted_R2 = 0.9145623068356686




Actual vs Predicted values for 100 obseravations
Predicted
Actual
60
50
40
30
20
10
0
0
20
40
60
80
100

Conclusion

- On an hourly basis, the bike counts peak in the afternoon (from 15.00 to 20.00). There are two peak occurrences, at 8.00 and 18.00, which are most likely to be caused by workers going to the office in the morning and going back home in the evening

- The demand for bikes will be lower on a rainy day compared to a sunny day. Similarly, higher Snowfall will cause lower demand and vice versa. The bike counts peak in the afternoon when the temperature is at it's highest, with the most visibility, wind speed, and least humidity

- Based on the analysis, we built five models to predict the count of Rented bikes. All the models performed decently, but the tree and ensemble models outperformed all the linear models (Linear Regression, Lasso, Ridge)

- Previously, the Decision Tree is overfitting the data. But after hyperparameter tuning, we avoided the problem of overfitting

- The Hour, Temperature, and Humidity are the most relevant features for predicting the count of Rental bikes.

- Only Decision Tree and Random Forest models are used. But there are many good ones out there even neural network models can be improved further by tuning on hyperparameters.

Thank You