Capstone Project - 2

Seoul Bike Sharing Prediction

Team

Arunav Goswami
Nayanjyoti Sharma
Mohammed Saad Pasha

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.

Content

- Data Pipeline
- Data Description
- Exploratory Data Analysis
- Feature Selection
- Machine Learning Algorithms
- Model Validation and Selection
- Evaluation Matrix of all the Models
- Model Explainability – SHAP
- Challenges
- Conclusion

Data Pipeline

- **Data Processing :** Checking for Missing values and Duplicate values.
- **EDA & Feature Engineering: -** Analyzing each feature individually, creation of new features according to our need, dropping of features by checking correlation and VIF, handling of outliers, standardization and normalization of features.
- **Model Creation and Validation :** Fitting of Machine Learning models into training and testing dataset, evaluation of performance metrics and Hyperparameter Tuning.
- **Model Explainability – SHAP**

Data Description

Dependent variable :

- Rented Bike Count :- Count of bikes rented at each hour.

Independent variables :

- Date - day/month/year
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10 m
- 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)

Data Description

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Date                       8760 non-null   object 
 1   Rented Bike Count          8760 non-null   int64  
 2   Hour                       8760 non-null   int64  
 3   Temperature(°C)            8760 non-null   float64
 4   Humidity(%)                8760 non-null   int64  
 5   Wind speed (m/s)           8760 non-null   float64
 6   Visibility (10m)           8760 non-null   int64  
 7   Dew point temperature(°C)  8760 non-null   float64
 8   Solar Radiation (MJ/m2)    8760 non-null   float64
 9   Rainfall(mm)               8760 non-null   float64
 10  Snowfall (cm)              8760 non-null   float64
 11  Seasons                    8760 non-null   object 
 12  Holiday                    8760 non-null   object 
 13  Functioning Day            8760 non-null   object 
dtypes: float64(6), int64(4), object(4)
memory usage: 958.2+ KB
```

Data Description

	count	mean	std	min	25%	50%	75%	max
Rented Bike Count	8760.0	704.602055	644.997468	0.0	191.00	504.50	1065.25	3556.00
Hour	8760.0	11.500000	6.922582	0.0	5.75	11.50	17.25	23.00
Temperature(°C)	8760.0	12.882922	11.944825	-17.8	3.50	13.70	22.50	39.40
Humidity(%)	8760.0	58.226256	20.362413	0.0	42.00	57.00	74.00	98.00
Wind speed (m/s)	8760.0	1.724909	1.036300	0.0	0.90	1.50	2.30	7.40
Visibility (10m)	8760.0	1436.825799	608.298712	27.0	940.00	1698.00	2000.00	2000.00
Dew point temperature(°C)	8760.0	4.073813	13.060369	-30.6	-4.70	5.10	14.80	27.20
Solar Radiation (MJ/m2)	8760.0	0.569111	0.868746	0.0	0.00	0.01	0.93	3.52
Rainfall(mm)	8760.0	0.148687	1.128193	0.0	0.00	0.00	0.00	35.00
Snowfall (cm)	8760.0	0.075068	0.436746	0.0	0.00	0.00	0.00	8.80

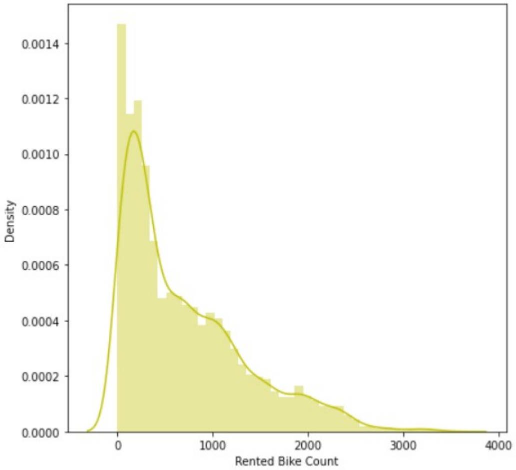
Numerical Data

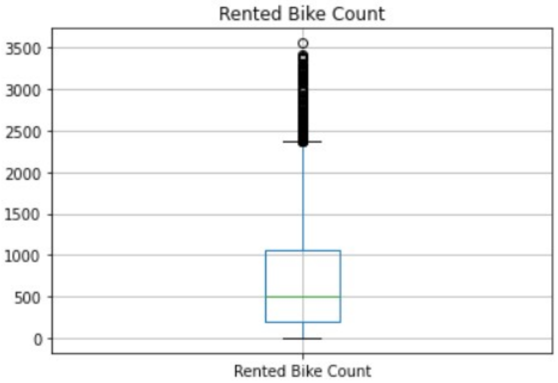
	count	unique	top	freq
Date	8760	365	01/12/2017	24
Seasons	8760	4	Spring	2208
Holiday	8760	2	No Holiday	8328
Functioning Day	8760	2	Yes	8465

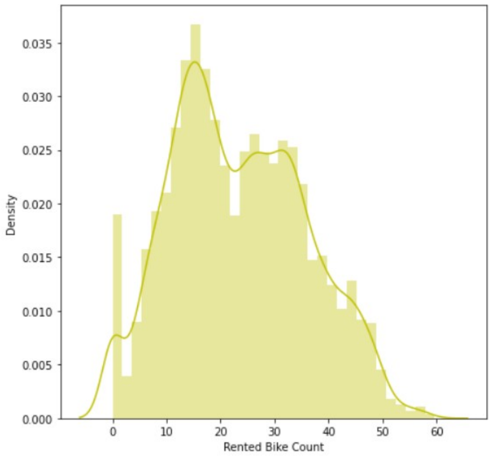
Categorical Data

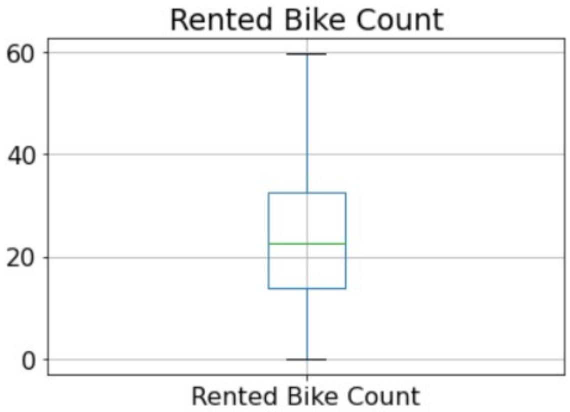
EDA

Square Root Transformation

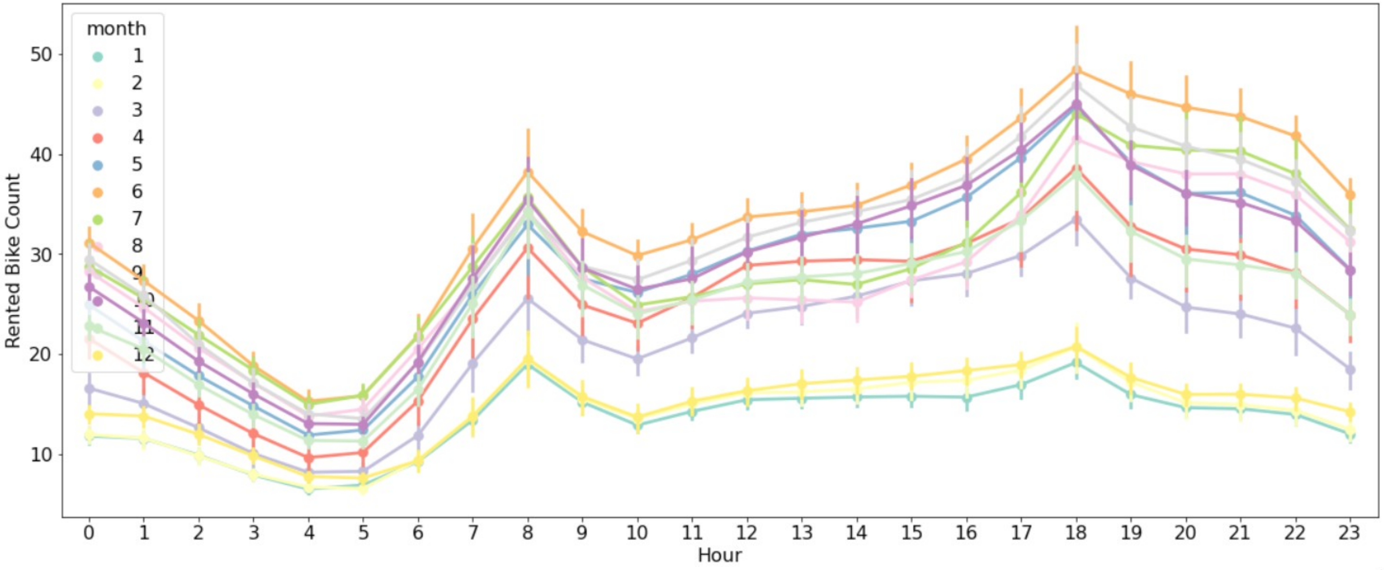


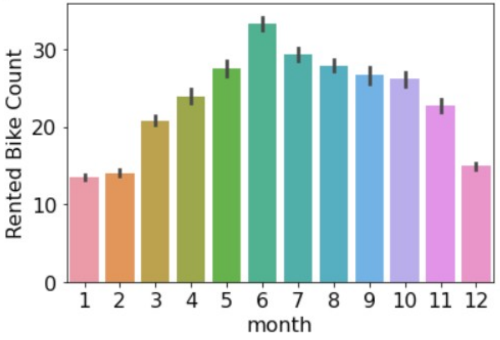






EDA continued.

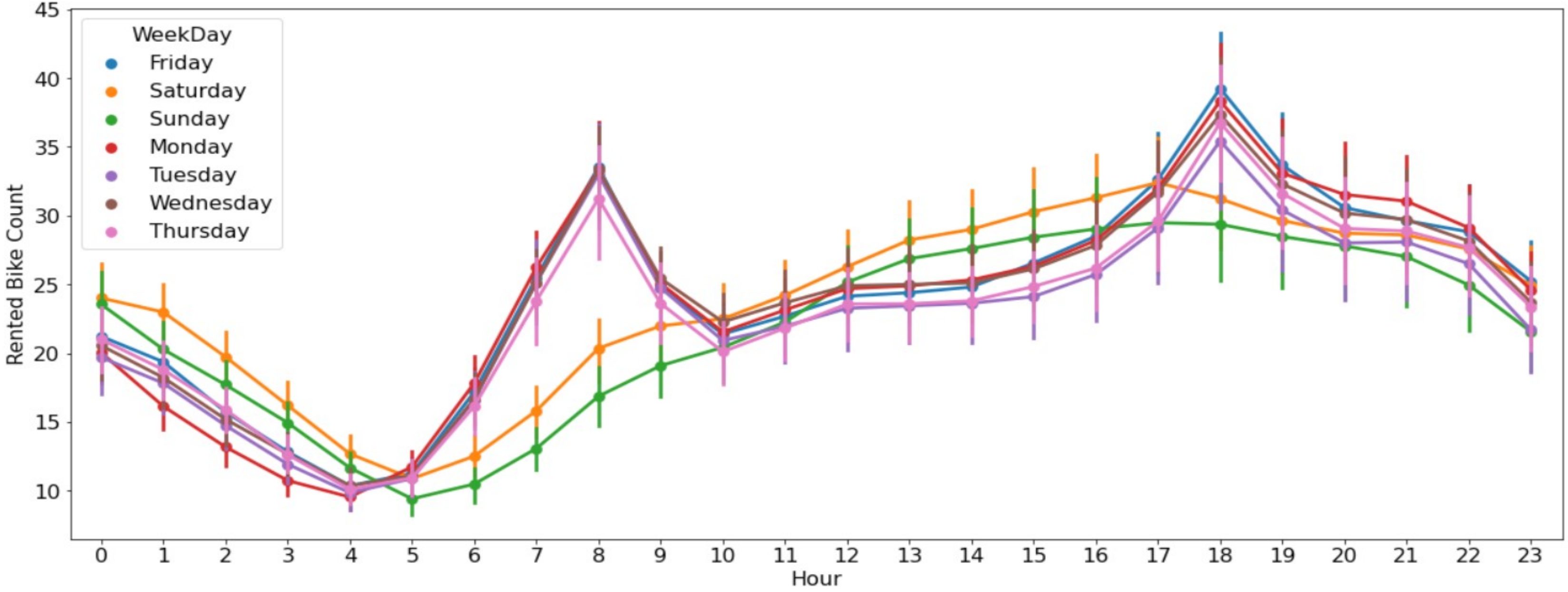




New feature Creation from 'Date' – 'month'

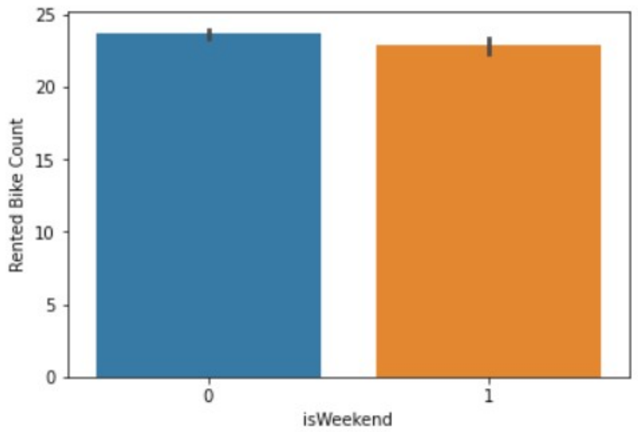
Rental bike count is highest in the month of June.

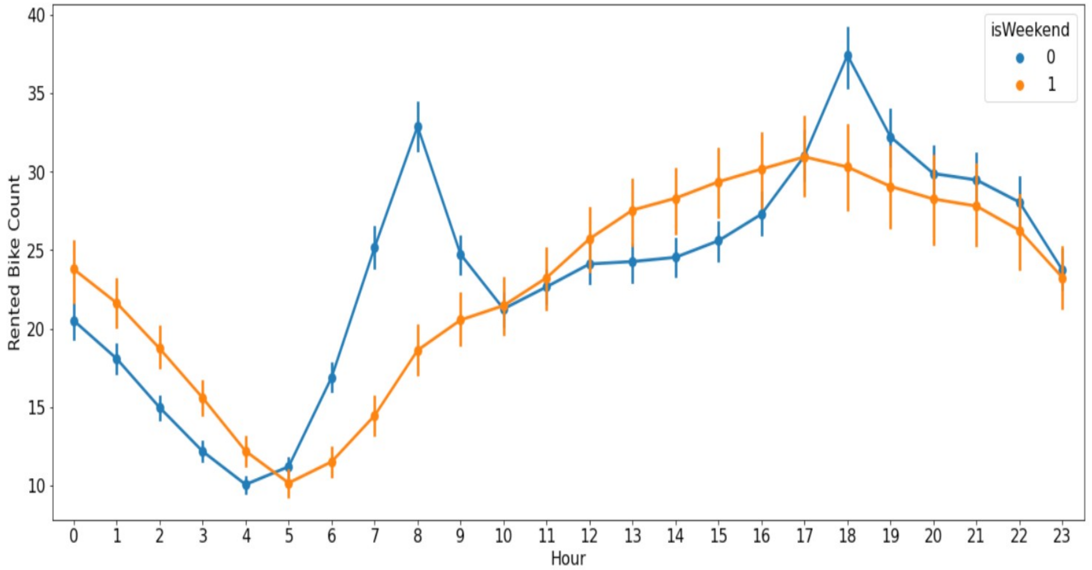
EDA continued.



- From the above plot, we can see that there is a different trend for Saturday & Sunday compared to others.
- So we will create a new categorical feature where we consider Saturday & Sunday as a weekend.

EDA continued.

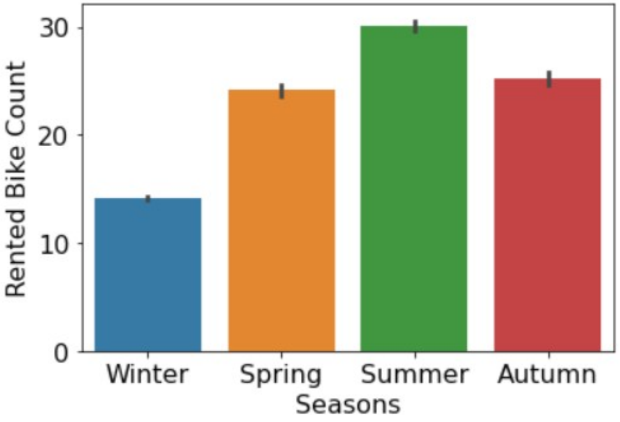


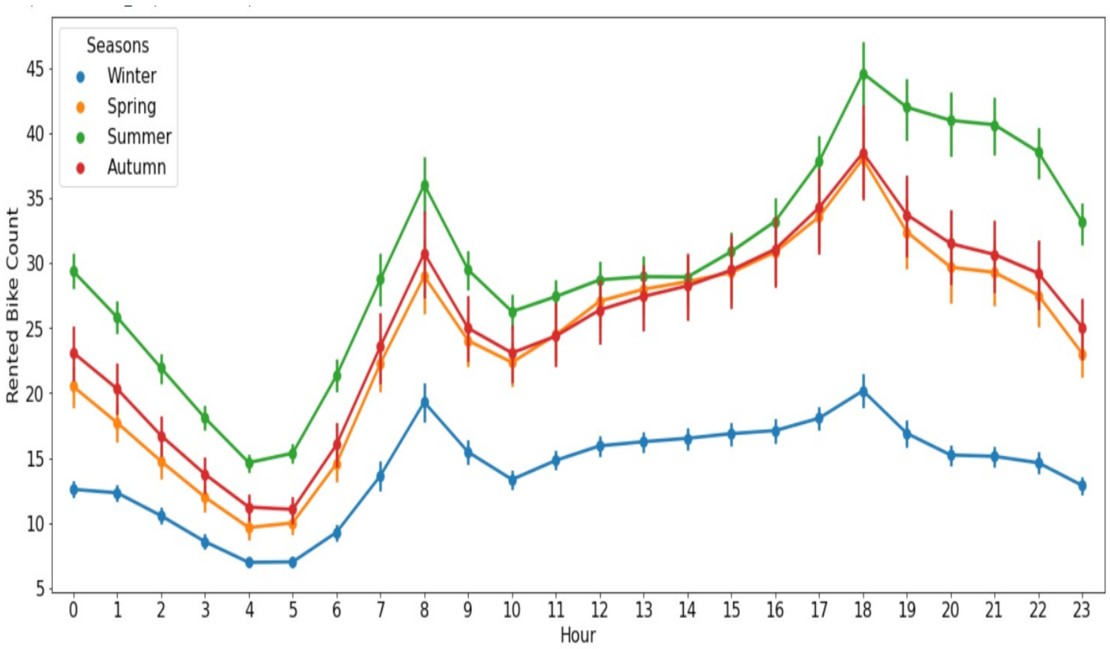


New feature Creation from ‘WeekDay’ – ‘isWeekend’

Since the trend between ‘WeekDay’ and ‘isWeekend’ is same, we will drop the variable ‘WeekDay’.

EDA continued.

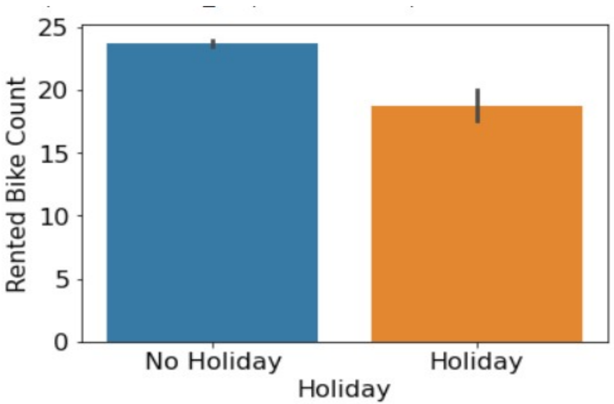


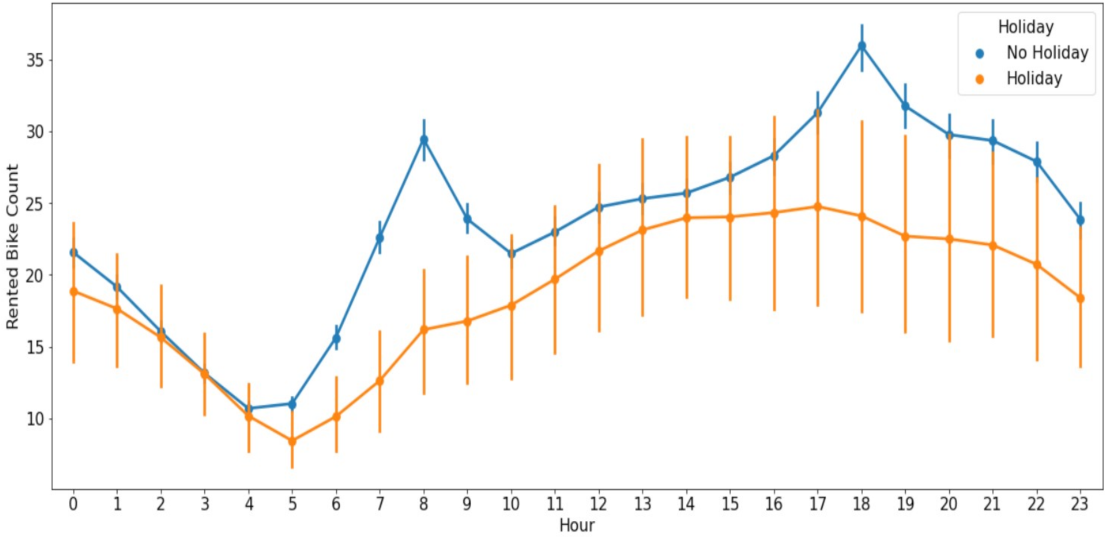


From the above plot, we can conclude that.

- Bike count is lowest during winter season.
- Bike count is highest during summer season.
- above plot, we can conclude that

EDA continued.



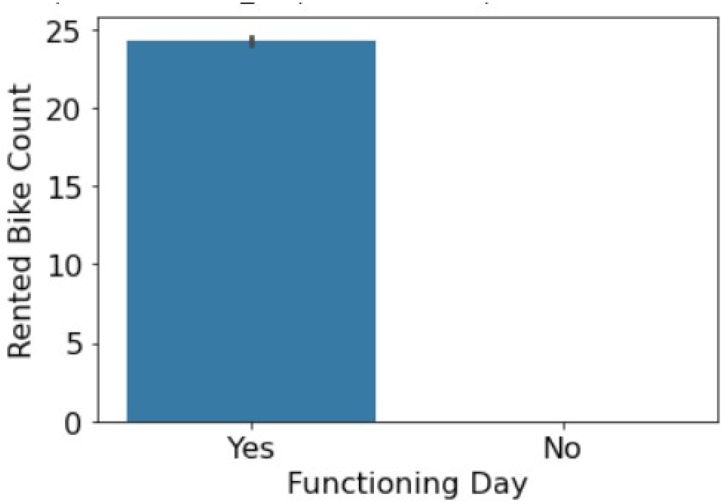


Rented Bike Count vs Holiday

From the above plot, we can see that the rented bike count is lower on holidays compared to the working day.

On working days from 7-9 AM and 5-7 PM, there is a sudden spike.

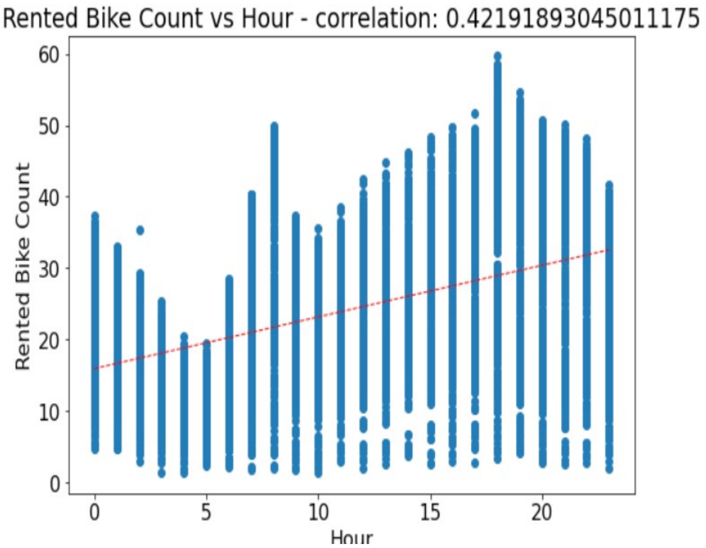
EDA continued.

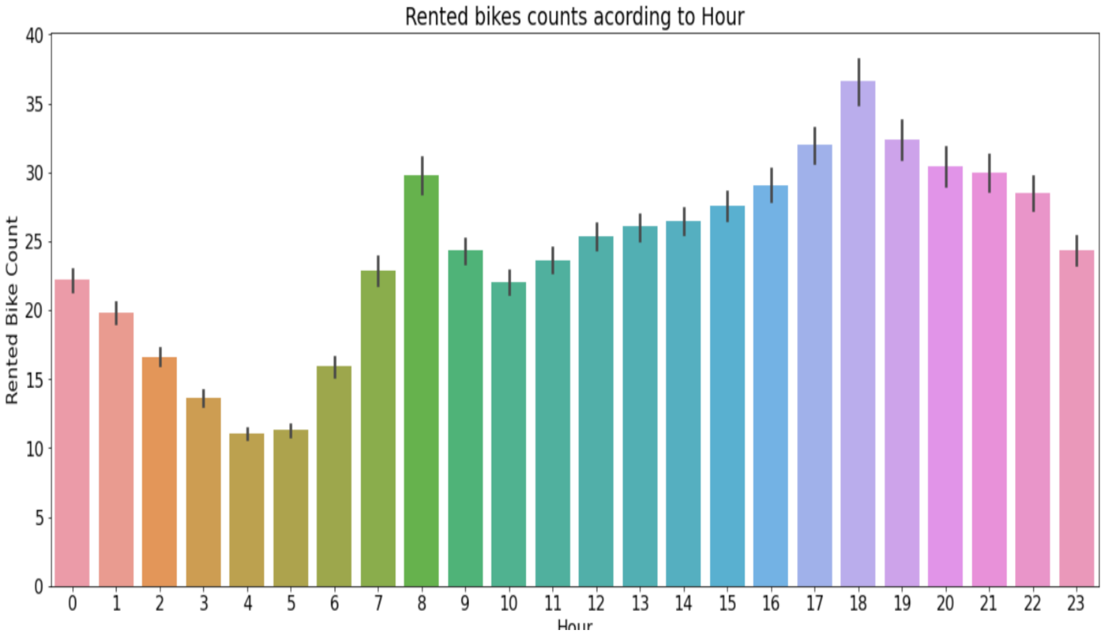


Rented Bike Count vs Functioning Day

- The rented bike count is 0 for a non functioning day.
- We choose to remove the rows with ‘No’ values in the ‘Functioning Day’ feature.
- We drop the "Functioning Day" feature.

EDA continued.

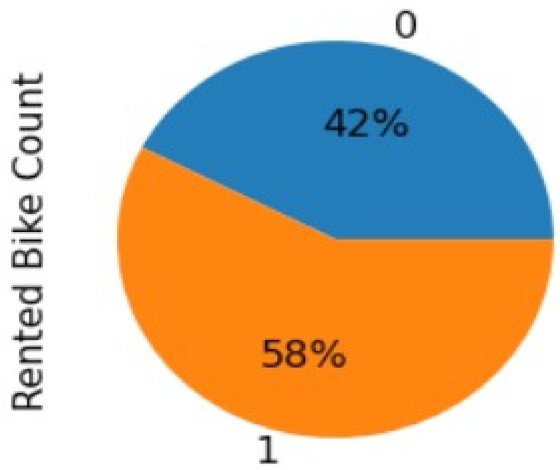




Rented Bike Count vs Hour

There is a sudden spike in bike count between 7-9 AM and 5-7 PM. So we can create a new categorical feature where we take 7 AM to 7 PM as a working hour.

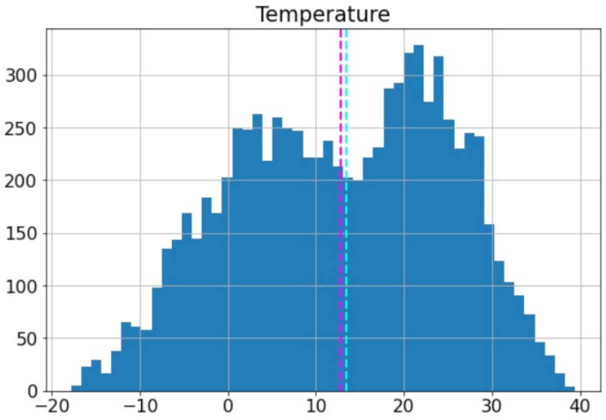
EDA continued.

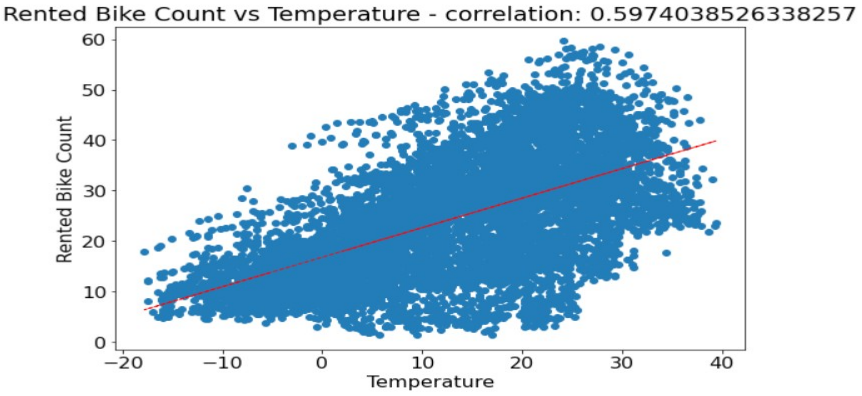


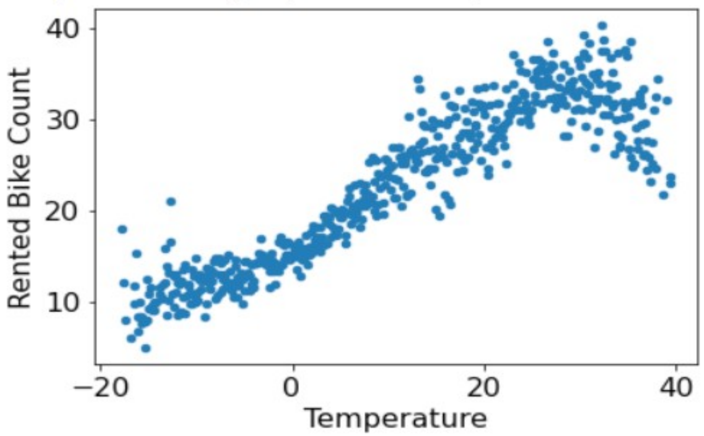
New Feature - Working Hour

During working hours (i.e. 7 AM to 7 PM) rented bike count is high as compared to non working hours.

EDA continued.



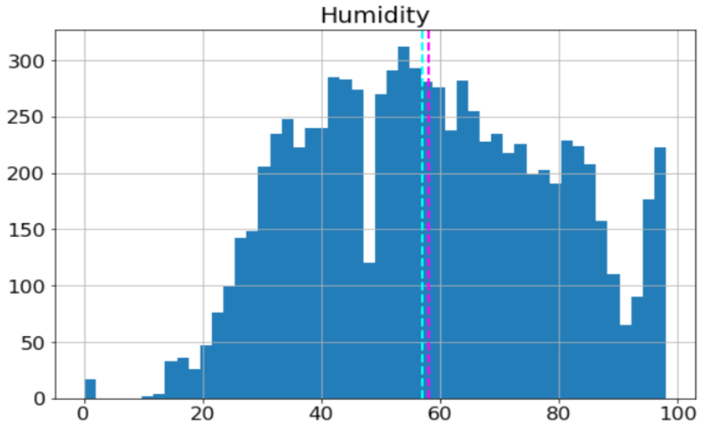


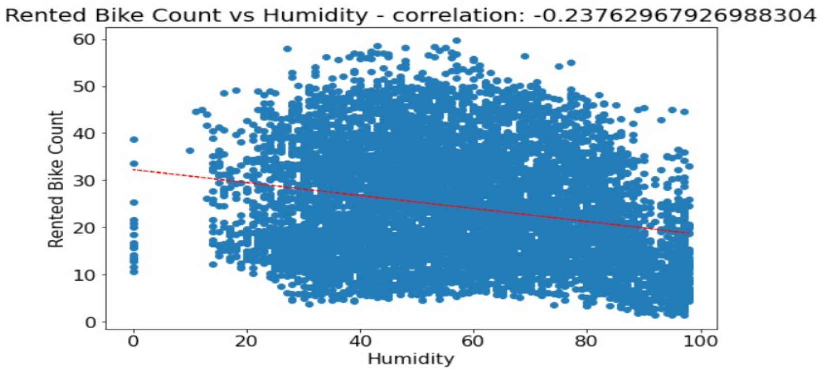


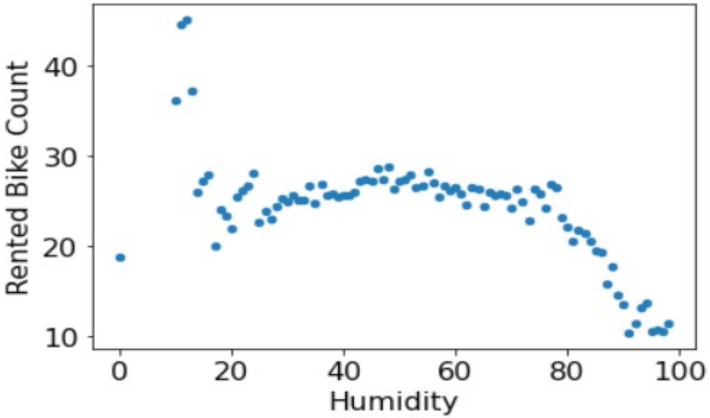
Rented Bike Count vs Temperature

Rented bike count is high between 20-30 °C

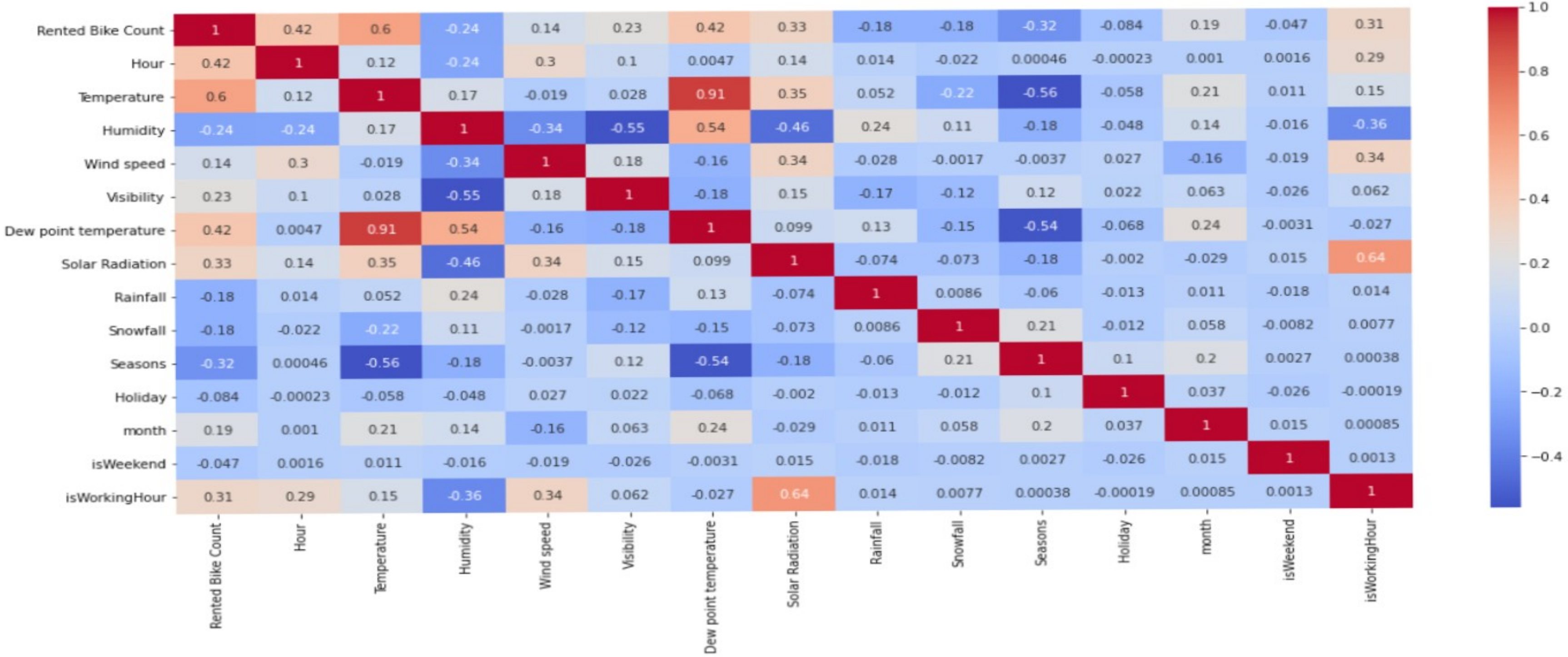
EDA continued.







EDA - Feature Selection



EDA - Feature Selection

	variables	VIF
0	Hour	4.483397
1	Temperature	4.872151
2	Humidity	8.114673
3	Wind speed	11.735593
4	Visibility	6.794825
5	Solar Radiation	3.482968
6	Rainfall	1.088932
7	Snowfall	1.144078
8	Seasons	4.751114
9	Holiday	1.066142
10	month	5.849840
11	isWeekend	1.392026
12	isWorkingHour	3.860307

Dropping ‘Wind Speed’ feature (VIF>10)

	variables	VIF
0	Hour	3.960817
1	Temperature	4.770525
2	Humidity	6.043629
3	Visibility	5.432355
4	Solar Radiation	3.198719
5	Rainfall	1.088726
6	Snowfall	1.144052
7	Seasons	4.735202
8	Holiday	1.064599
9	month	5.784683
10	isWeekend	1.390723
11	isWorkingHour	3.784978

EDA Feature Selection

OLS Regression Results

Dep. Variable:	Rented Bike Count	**R-squared (uncentered):**	0.923
Model:	OLS	**Adj. R-squared (uncentered):**	0.923
Method:	Least Squares	**F-statistic:**	8403.
Date:	Mon, 09 May 2022	**Prob (F-statistic):**	0.00
Time:	10:15:00	**Log-Likelihood:**	-29079.
No. Observations:	8465	**AIC:**	5.818e+04
Df Residuals:	8453	**BIC:**	5.827e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	11.9835	0.284	42.207	0.000	11.427	12.540
x2	32.0371	0.578	55.421	0.000	30.904	33.170
x3	-7.0177	0.409	-17.159	0.000	-7.819	-6.216
x4	4.8912	0.261	18.713	0.000	4.379	5.404
x5	-5.5412	0.515	-10.766	0.000	-6.550	-4.532
x6	-65.4882	2.628	-24.921	0.000	-70.639	-60.337
x7	-4.1369	1.710	-2.419	0.016	-7.489	-0.784
x8	-2.0441	0.269	-7.588	0.000	-2.572	-1.516
x9	-3.1983	0.384	-8.332	0.000	-3.951	-2.446
x10	4.5881	0.293	15.668	0.000	4.014	5.162
x11	-1.2374	0.179	-6.894	0.000	-1.589	-0.886
x12	4.1782	0.224	18.616	0.000	3.738	4.618

Omnibus:	86.187	**Durbin-Watson:**	0.503
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	145.288
Skew:	-0.009	**Prob(JB):**	2.83e-32
Kurtosis:	3.642	**Cond. No.**	49.5

From OLS, the p-value for all features is less than 0.05. So, we will consider all features.

Machine Learning Algorithms

- Linear Regression
- Lasso Regression
- Ridge Regression
- Decision tree
- Random Forest
- Gradient Boost

Model’s Evaluation Matrices

		MSE	RMSE	MAE	R2	Adjusted R2
Train	Linear Regression	54.107034	7.355748	5.740547	0.616818	0.614081
	Lasso Regression	54.107087	7.355752	5.740440	0.616818	0.614081
	Ridge Regression	54.107034	7.355748	5.740547	0.616818	0.614081
	Decision Tree	0.000207	0.014390	0.000333	0.999999	0.999999
	Random Forest	1.326741	1.151842	0.731217	0.990604	0.990537
	Gradient Boosting	12.584937	3.547525	2.513794	0.910875	0.910238
Test	Linear Regression	55.388354	7.442335	5.734993	0.598447	0.595579
	Lasso Regression	55.391033	7.442515	5.735024	0.598427	0.595559
	Ridge Regression	55.388356	7.442335	5.734993	0.598447	0.595579
	Decision Tree	22.253700	4.717383	2.862540	0.838666	0.837513
	Random Forest	11.613644	3.407880	2.102932	0.915804	0.915202
	Gradient Boosting	16.659540	4.081610	2.843235	0.879222	0.878359

Random Forest has the highest R2 & Adjusted R2. So, we will select this model and find the best hyper parameters for it.

Hyperparameters

n_estimators :- number of trees in the random forest
max_features :- number of features in consideration at every split
max_depth :- maximum number of levels allowed in each decision tree
min_samples_split :- minimum sample number to split a node
min_samples_leaf :- minimum sample number that can be stored in a leaf node
bootstrap :- method used to sample data points

Hyperparameter Tuning

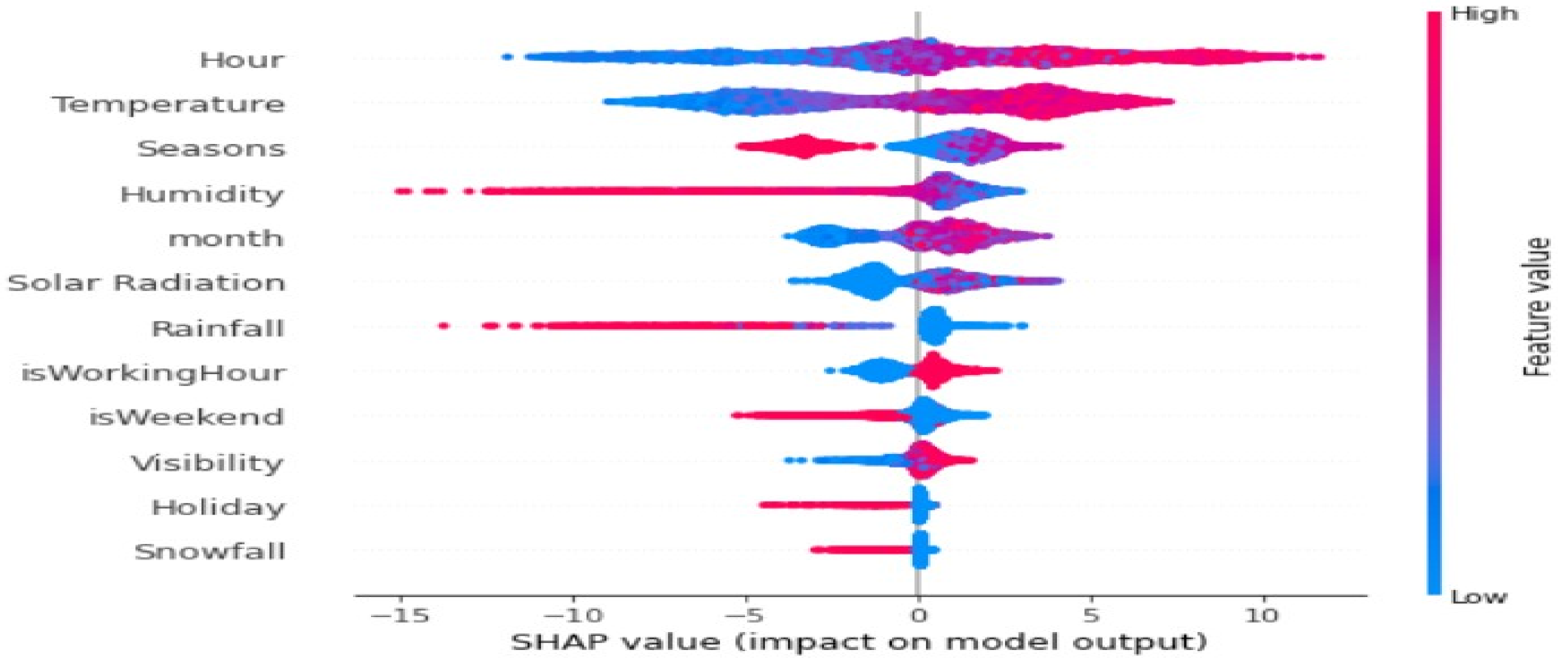
Random Forest → Randomized Search CV and Grid Search CV → Random Forest (Tuned)

```
For Train Data:
MSE  :  1.3267407683053913
RMSE  :  1.1518423365658128
MAE  :  0.7312174470916964
R2  :  0.9906041297437512
Adjusted R2  :  0.990537016384778
----------------------------------------
For Test Data:
MSE  :  11.613644144960146
RMSE  :  3.407879713980549
MAE  :  2.1029315025294806
R2  :  0.9158036975941133
Adjusted R2  :  0.9152022954340713
```

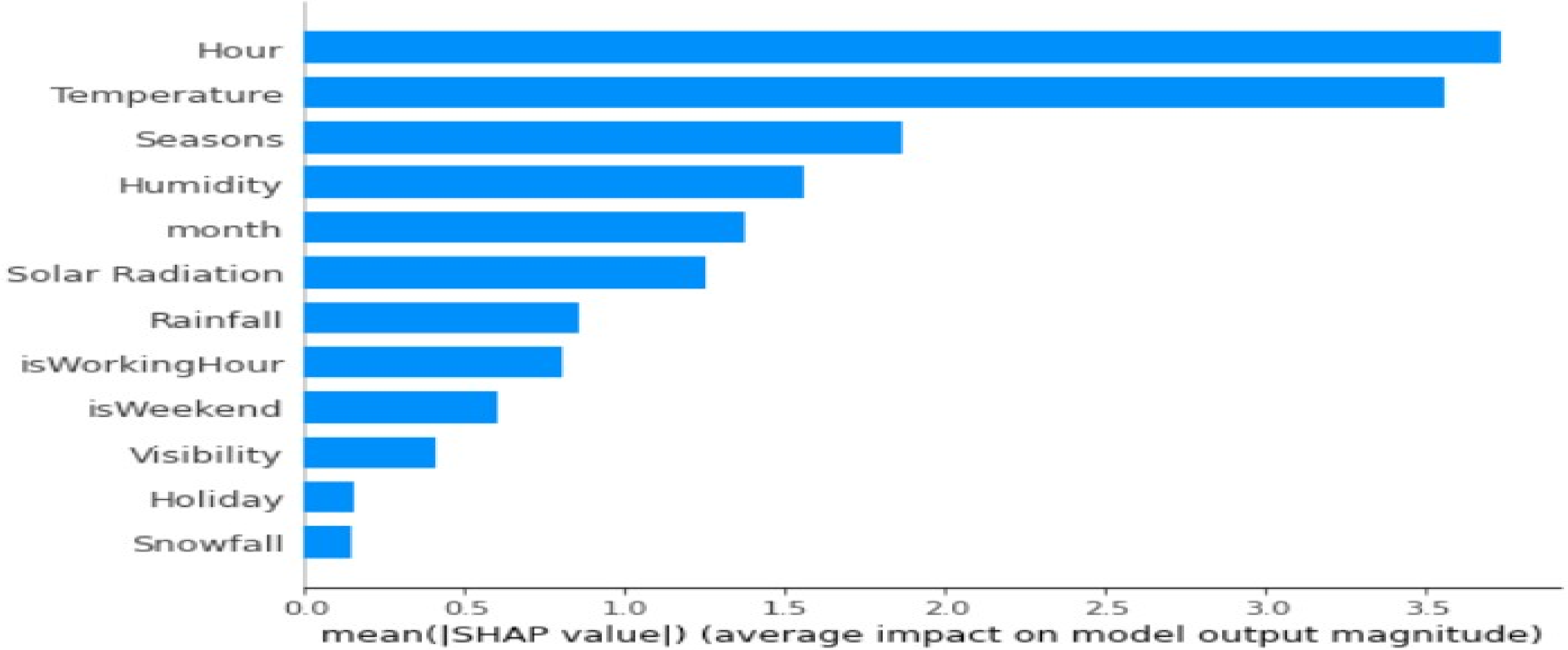
```
{'bootstrap': False,
 'max_depth': 320,
 'max_features': 'sqrt',
 'min_samples_leaf': 1,
 'min_samples_split': 4,
 'n_estimators': 410}
```

```
For Train Data:
MSE  :  0.3937936006293233
RMSE  :  0.6275297607518892
MAE  :  0.41532530159763326
R2  :  0.9972111857360197
Adjusted R2  :  0.9971912656341342
----------------------------------------
For Test Data:
MSE  :  11.238047990312861
RMSE  :  3.3523197923695855
MAE  :  2.1627324212664307
R2  :  0.9185266850582069
Adjusted R2  :  0.9179447328086227
```

Model Explainability - SHAP



Feature Importance



The above plot shows the most important features in decreasing order.

Conclusion

i) We observed that the bike rental count is high on non holiday than on holiday.
ii) During weekdays at 7-9 AM and 5-7 PM, there are sudden spikes in bike count.
iii) The bike count is high at high temperatures.
iv) Rental bike count is highest in the month of June.
iv) In summer the bike count is the highest and in winter it is the lowest.
v) When we compare the RMSE and Adjusted R2 of all the models for test data, Random Forest gives the highest Score where the Adjusted R2 score is 0.91 and RMSE is 3.4. So this model is the best for predicting the bike rental count on hourly basis.

Challenges

- The biggest challenge we had to overcome was the computation time. GridSearch CV and SHAP took almost few hours to execute.

- Some of the features like ‘Rainfall’, ‘Snowfall’, ‘Solar Radiation’ are extremely skewed for which it was difficult to normalize them.

Thank You