Capstone Project
Seoul Bike Sharing Demand Prediction

Team
Arman Alam,
Fathima.K,
Syed Sharin,
Abdul Rahman

AlmaBetter

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 Summary
- Exploratory Data Analysis
  - Visualizing Rented Bike Count, Hour with Respect to different categorical Feature
  - Visualizing Value count of Categorical Features
  - Visualizing Distribution of Data
  - Normalize Dependent Variable
  - Visualizing Regression Plot of Data
- Correlation Analysis
- Evaluation Matrix of All the models
- Model Explainability – SHAP, Lime and Eli5
- Model Validation & Selection
- Challenges
- Conclusion

Data Summary

- Bike sharing has been gaining importance over the last few decades. More and more people are turning to healthier and more liveable cities where activities like bike sharing are easily available. there are many benefits from bike sharing, such as environmental benefits. It was a green way to travel

- The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

- This dataset contains the hourly and daily count of rental bikes between years 2017 and 2018 in Capital bike share system with the corresponding weather and seasonal information. The dataset contains 8760 rows (every hour of each day for 2017 and 2018) and 14 columns (the features which are under consideration).

Exploratory Data Analysis

Rented Bike Count, Hour with Respect to different categorical Feature

Observation

From all these point plot we have observed a lot from every column like :

Season

In the season column, we are able to understand that the demand is low in the winter season.

Holiday

In the Holiday column, The demand is low during holidays, but in no holidays the demand is high, it may be because people use bikes to go to their work.

Functioning Day

In the Functioning Day column, If there is no Functioning Day then there is no demand

Days of week

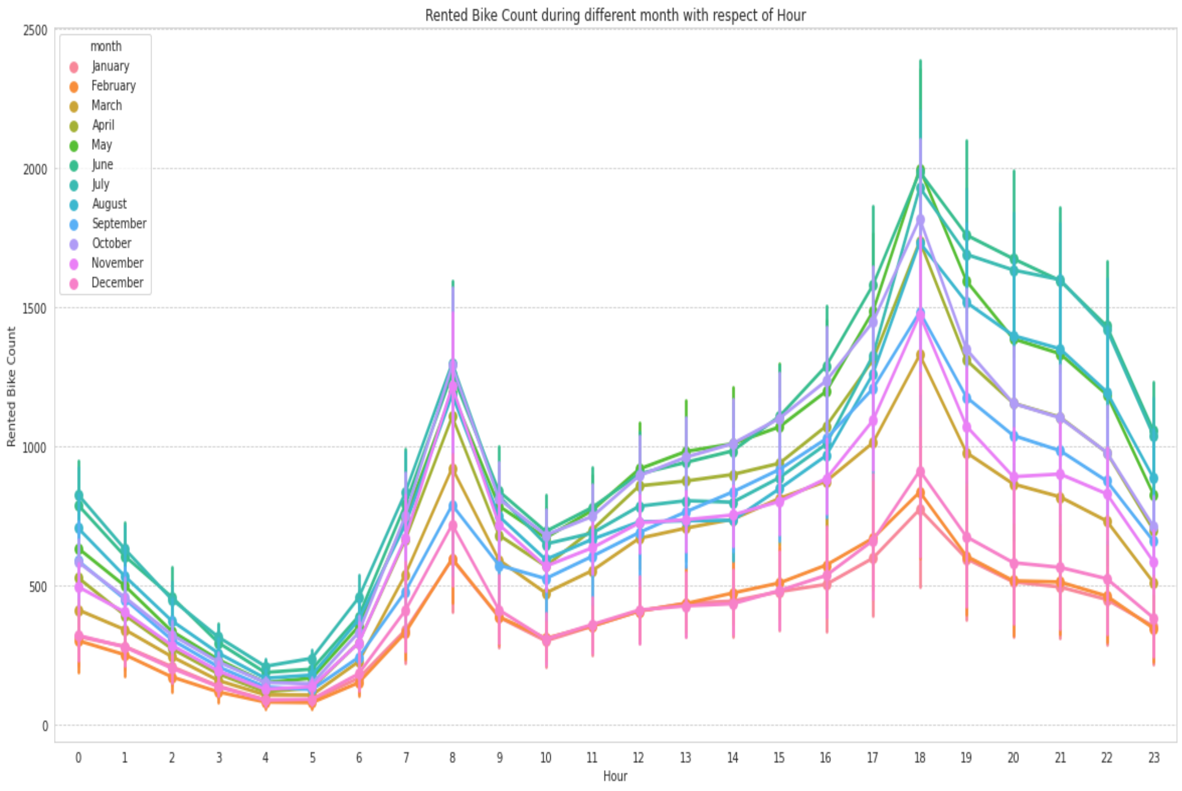
In the Days of week column, We can observe from this column that the pattern of weekdays and weekends is different, in the weekend the demand becomes high in the afternoon. While the demand for office timings is high during weekdays, we can further change this column to weekdays and weekends.

month

In the month column, We can clearly see that the demand is low in December January & February, It is cold in these months and we have already seen in season column that demand is less in winters.

year

The demand was less in 2017 and higher in 2018, it may be because it was new in 2017 and people did not know much about it.



Visualizing Value count Percentage of Each Categorical Features

This Pie plot Shows us how each feature value is distributed

Hour:

Hour is distributed equally

Season:

Season is also equally Distributed

Holiday:

No Holiday comes 95% and Holiday 5%

Functioning Day:

Yes comes 97% and No Comes 3%

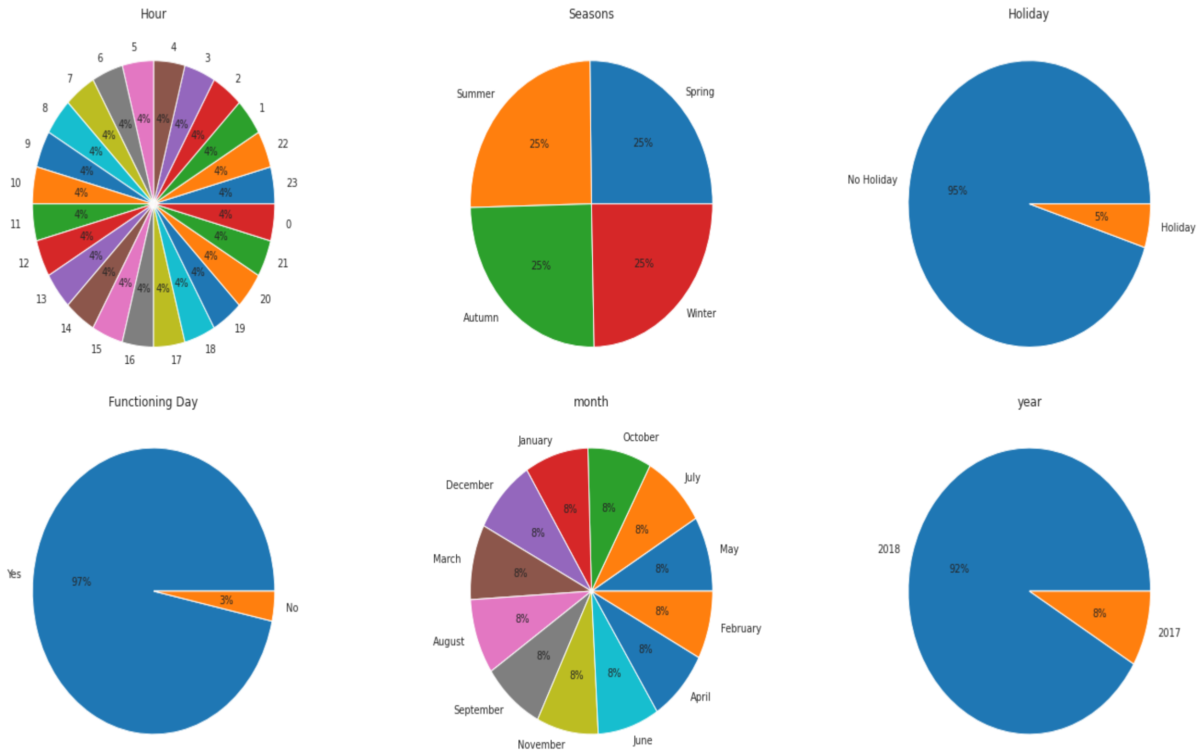
Month:

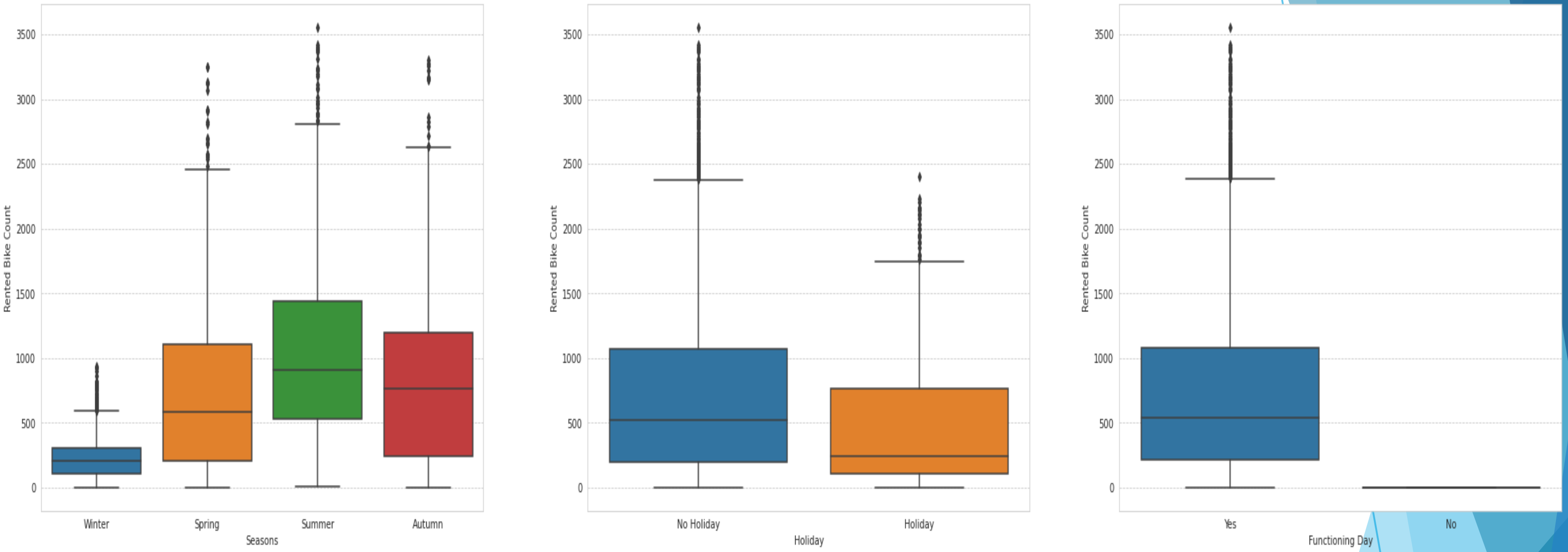
Month is also equally Distributed

Year:

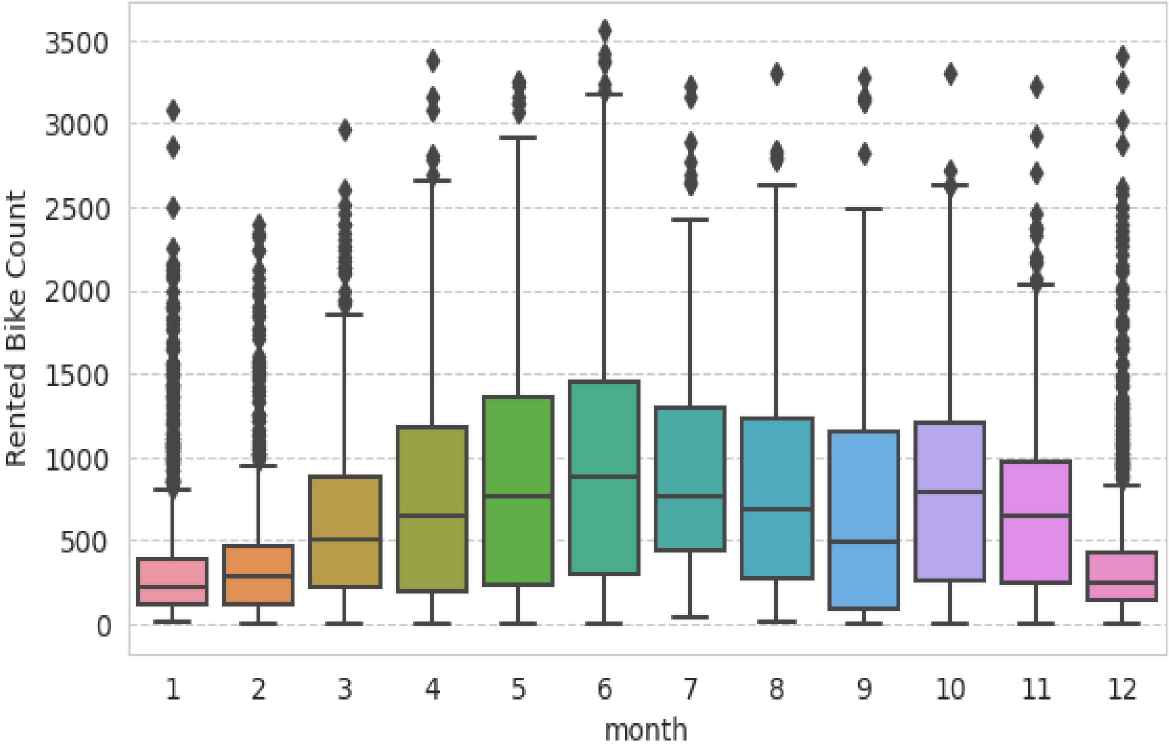
2018 comes 92% and 2017 8%

We think may be in 2017 they are new





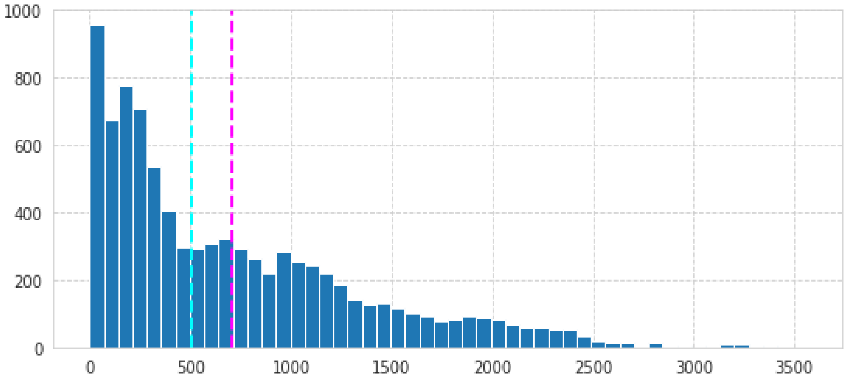
- Less demand on winter seasons
- Slightly Higher demand during Non holidays
- Almost no demand on non functioning day

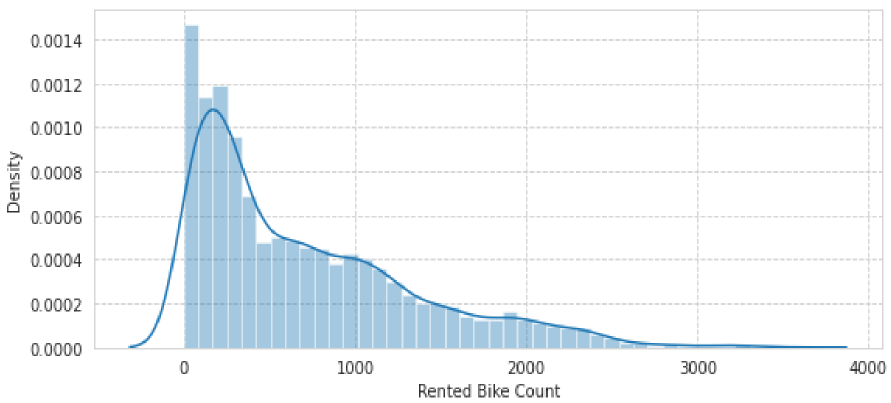


- We can see that there less demand of Rented bike in the month of December, January, February i.e. during winter seasons

- Also demand of bike is maximum during May, June, July i.e Summer seasons

Distribution of Numerical Features





Right skewed columns are

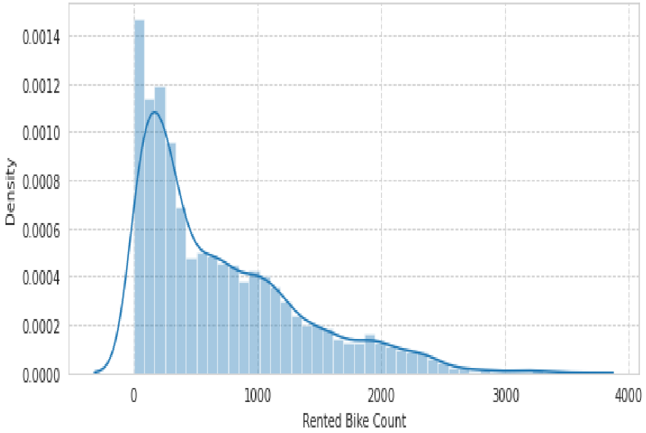
Rented Bike Count (Its also our Dependent variable), Wind speed (m/s), Solar Radiation (MJ/m2), Rainfall(mm), Snowfall (cm),

Left skewed columns are

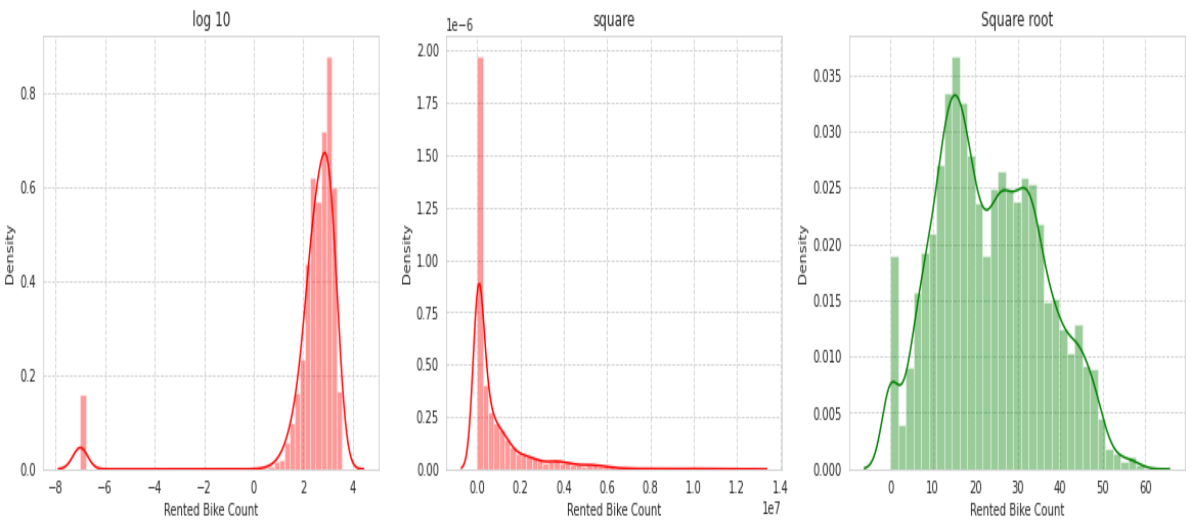
Visibility (10m), Dew point temperature(°C)

Normalize Dependent Variable for Linear Models

Before Transform



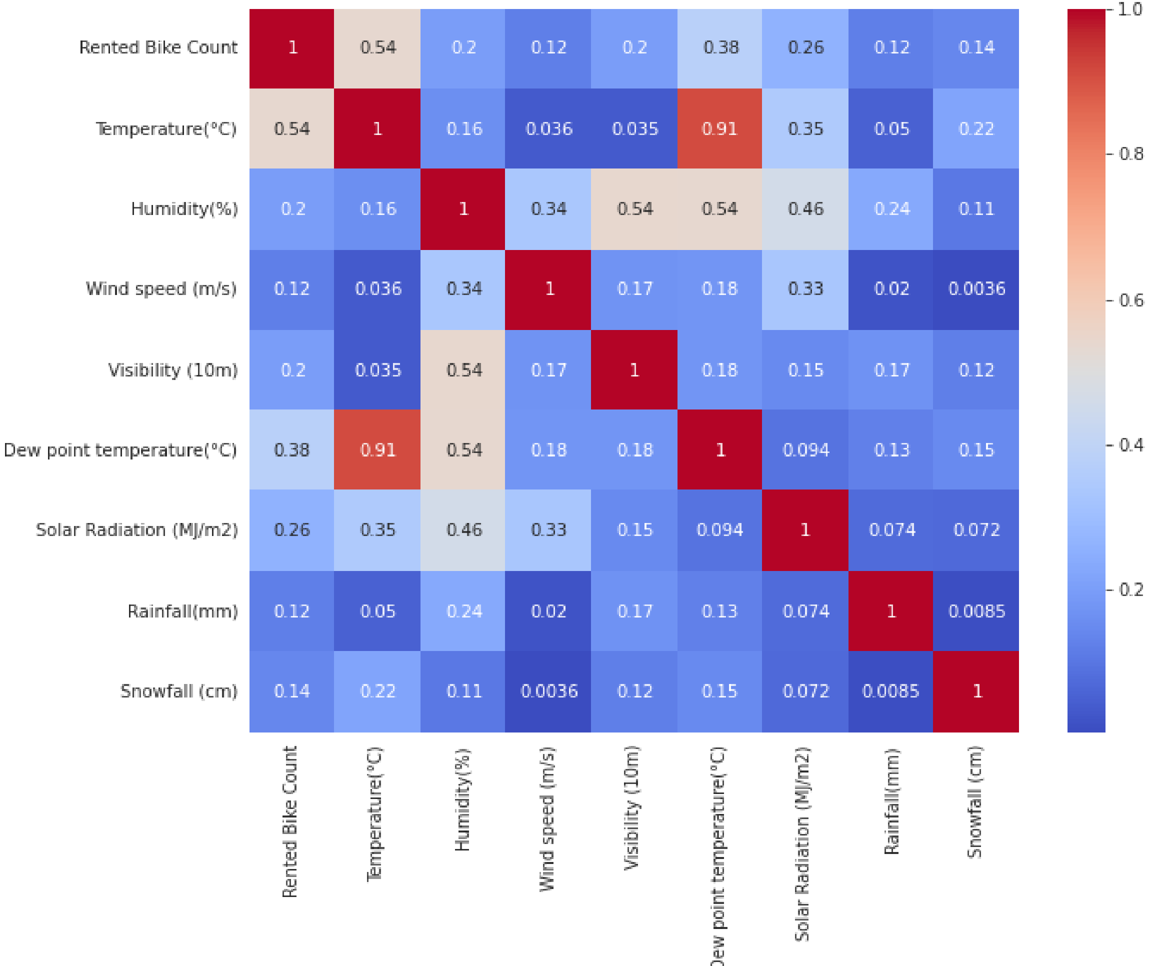
After Transform



Before transformation : **Our dependent variable is right Skewed.**
After transformation : **Our dependent variable in green plot is normalized to some extent: so we will go with square root on our dependent variable**

Correlation Analysis

From the correlation graph with Heat map we saw that dew point temp and temperature is highly correlated. Then we checked VIF and concluded that these two features are affecting VIF score also. so we decided to drop one of these feature and to do this we checked which feature is least correlated with Dependent variable and we identified it to be Dew point temperature and therefore we dropped the Dew point temperature.



Models Performed

List of Models

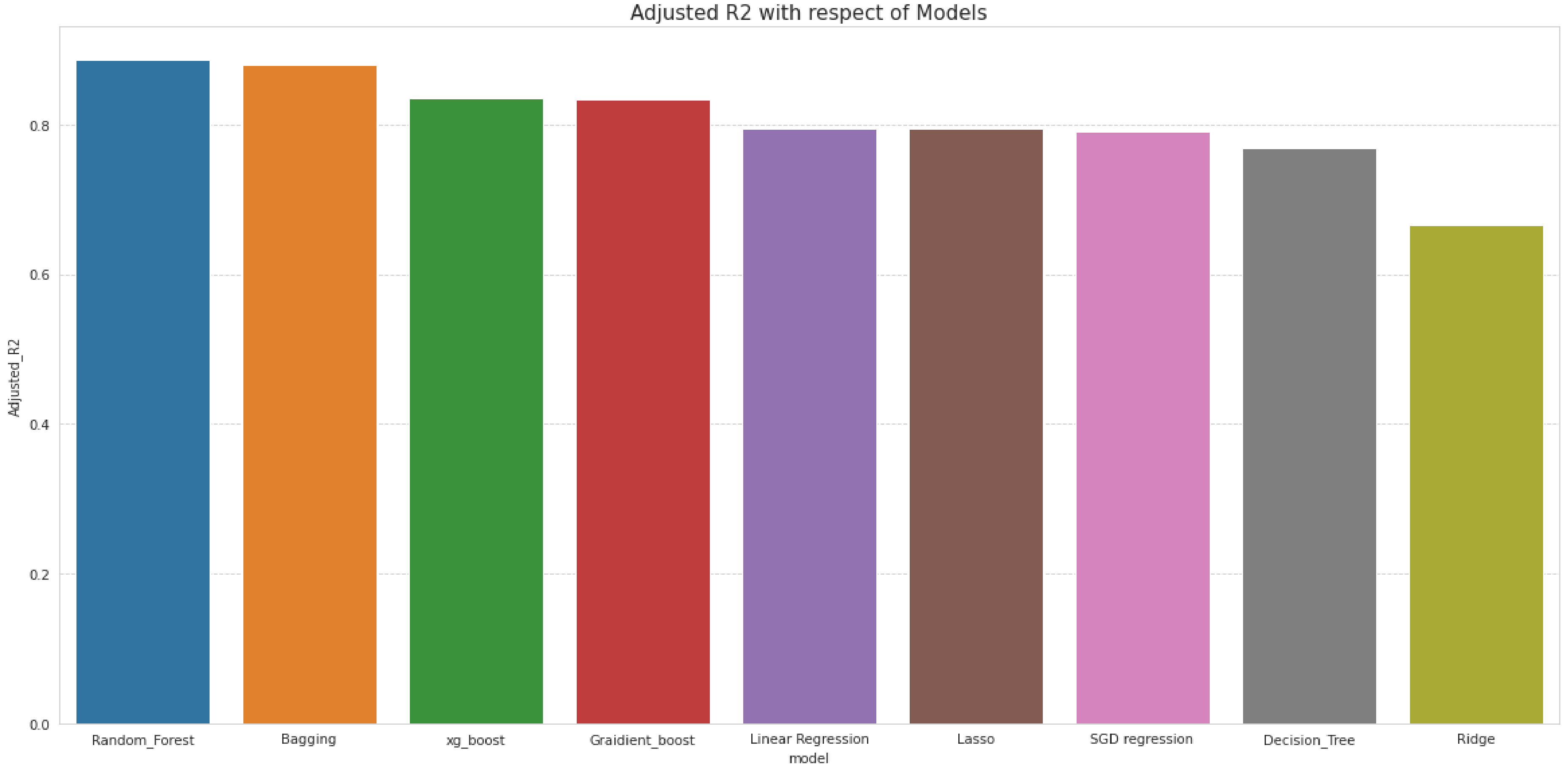
- Linear Regression with regularizations (Lasso & Ridge)
- Polynomial Regression
- Stochastic Gradient Descent Regressor
- Decision tree
- Random Forest
- Bagging Regressor
- Gradient Boosting Regressor
- Extreme Gradient Boosting
- Stacking Regressor

All Models Evaluation without hyperparameter tuning

	model	Mean_Absolute_error	Mean_square_error	Root_Mean_square_error	Training_score	R2	Adjusted_R2
0	Random_Forest	126.154142	45822.448664	214.061787	0.985495	0.888020	0.886621
1	Bagging	133.118813	48561.701772	220.367198	0.979672	0.881326	0.879843
2	xg_boost	169.428017	66493.455641	257.863250	0.864647	0.837504	0.835475
3	Graidient_boost	170.837848	67399.729690	259.614579	0.865264	0.835289	0.833232
4	Linear Regression	4.234345	30.525994	5.525033	0.794813	0.798303	0.793781
5	Lasso	4.234372	30.526526	5.525082	0.794813	0.798299	0.793777
6	SGD regression	4.273215	31.005091	5.568222	0.793467	0.795137	0.790544
7	Decision_Tree	165.736530	93337.902740	305.512525	1.000000	0.771902	0.769053
8	Ridge	5.482710	49.511948	7.036473	0.663220	0.672855	0.665521

The Best model is Random Forest but it is over fitted, that's why We are using Hyperparameter tuning so that we can reduce the overfitting and increase the accuracy.

Adjusted R2 score with respect to models without applying hyper parameter tuning

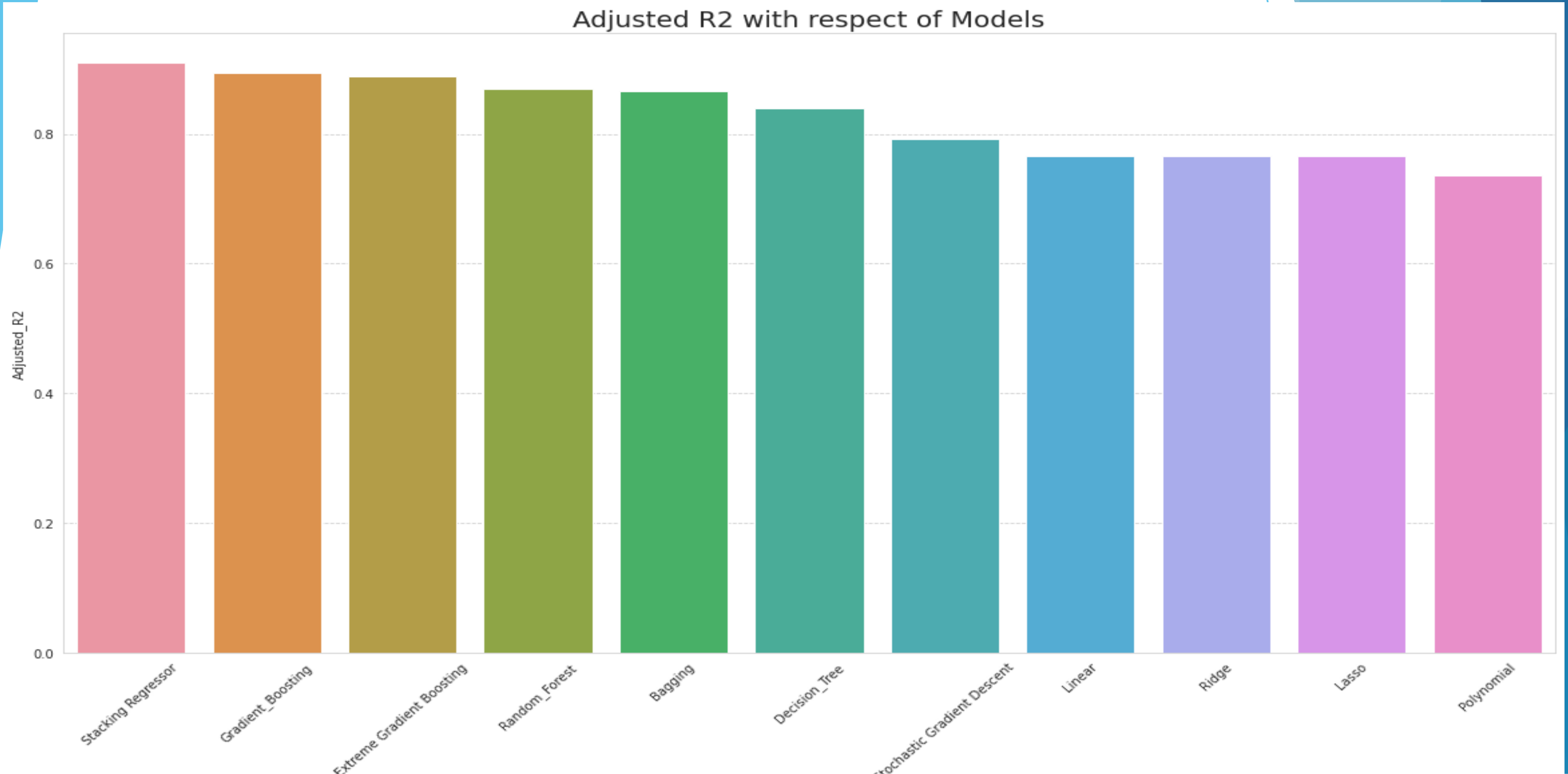


All Models Evaluation with hyperparameter tuning

	Models	Mean_Absolute_error	Mean_square_error	Root_Mean_square_error	Training_score	R2	Adjusted_R2
0	Stacking Regressor	114.475727	36597.556611	191.304879	0.951983	0.910563	0.909446
1	Gradient_Boosting	124.696362	42859.142667	207.024498	0.922070	0.895261	0.893953
2	Extreme Gradient Boosting	135.715404	44777.019534	211.605812	0.928859	0.890575	0.889208
3	Random_Forest	139.582526	53012.618940	230.244694	0.923972	0.870448	0.868831
4	Bagging	141.082279	54003.318533	232.386141	0.934478	0.868027	0.866379
5	Decision_Tree	151.572656	64658.424018	254.280208	0.919458	0.841989	0.840015
6	Stochastic Gradient Descent	4.267835	30.865402	5.555664	0.793195	0.796060	0.791488
7	Linear	207.416668	93868.920316	306.380352	0.794813	0.770604	0.765461
8	Ridge	207.438415	93898.286665	306.428273	0.794813	0.770533	0.765388
9	Lasso	207.483631	93964.528373	306.536341	0.794811	0.770371	0.765223
10	Polynomial	132.812851	47669.588013	218.333662	0.926009	0.883506	0.735471

Top 3 best performing models are :1. Staking Regressor
2. Gradient Boosting
3. Extreme Gradient Boosting

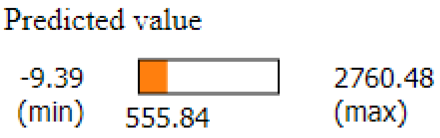
Adjusted R2 score with respect to models after applying hyper parameter tuning

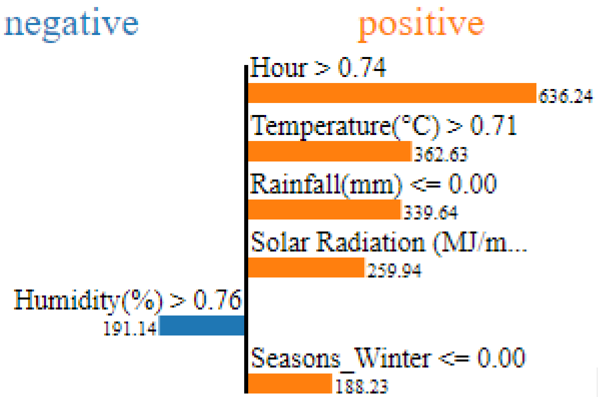


Model Explainability

Explaining Stacking with Lime

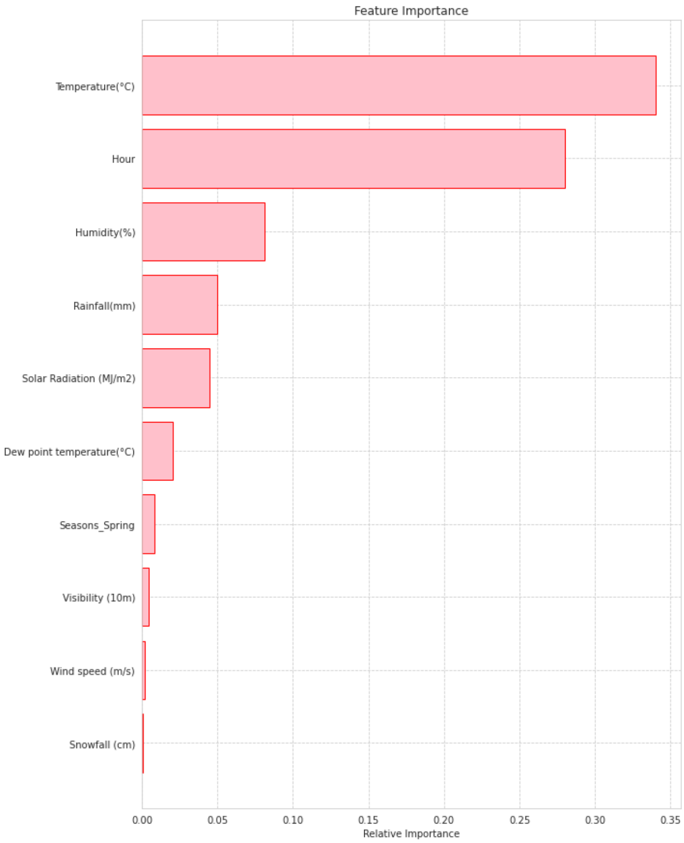
```
Intercept 10.634989571981578
Prediction_local [1606.18373005]
Right: 555.8428679296986
```





Feature	Value
Hour	14.00
Temperature(°C)	34.00
Rainfall(mm)	0.00
Solar Radiation (MJ/m2)	1.68
Humidity(%)	50.00
Seasons_Winter	0.00

Explaining Gradient Boosting



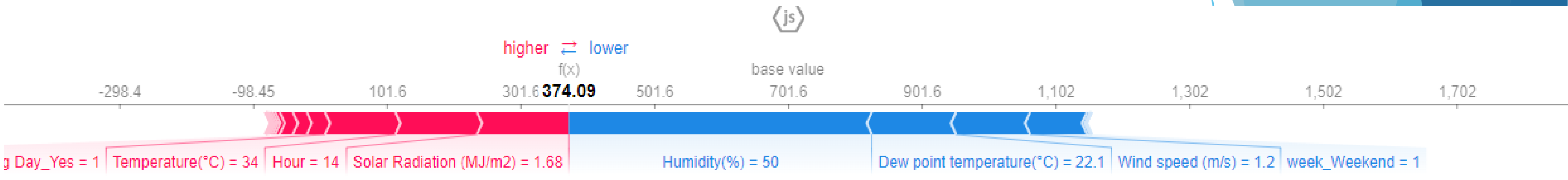
Feature Importance

y (score **293.219**) top features

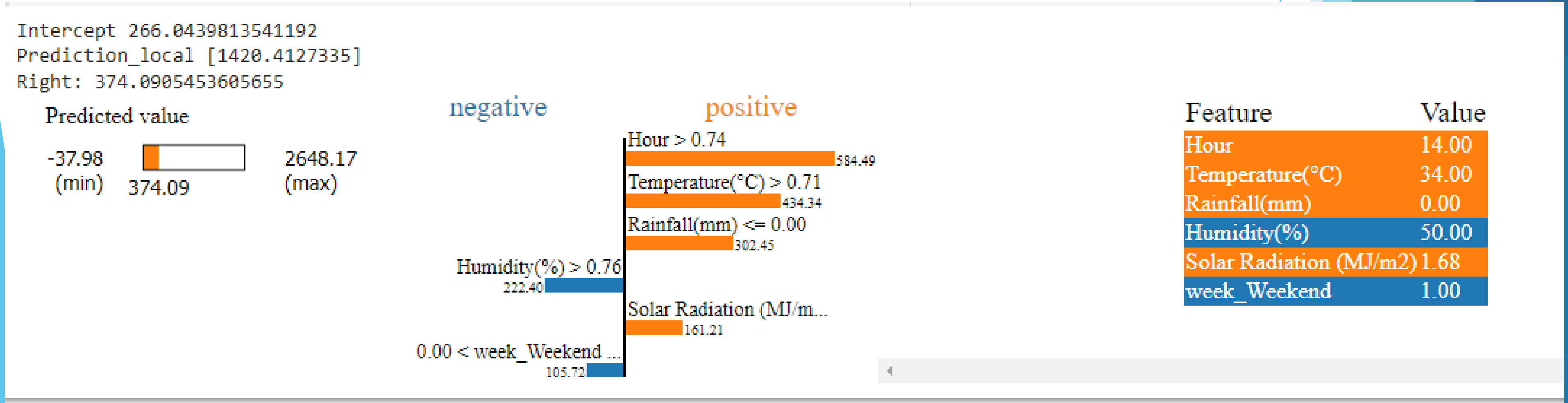
Contribution?	Feature	Value
+635.658	<BIAS>	1.000
+232.275	Hour	1.000
+174.976	Temperature(°C)	28.000
+33.761	Rainfall(mm)	0.000
+32.505	Functioning Day_Yes	1.000
+11.840	month_July	1.000
+5.866	Seasons_Spring	0.000
+2.849	Seasons_Winter	0.000
+1.647	Seasons_Summer	1.000
+1.472	Visibility (10m)	1799.000
+0.855	Holiday_No Holiday	1.000
+0.716	Snowfall (cm)	0.000
+0.503	month_January	0.000
+0.349	month_May	0.000
+0.167	month_September	0.000
+0.102	month_February	0.000
-0.174	month_October	0.000
-0.207	month_December	0.000
-0.260	year_2018	1.000
-0.330	month_March	0.000
-0.388	month_August	0.000
-5.866	month_June	0.000
-13.903	week_Weekend	1.000
-24.599	Wind speed (m/s)	0.600
-35.214	Solar Radiation (MJ/m2)	0.000
-114.390	Dew point temperature(°C)	22.200
-646.990	Humidity(%)	71.000

ELI5

Explaining Gradient Boosting

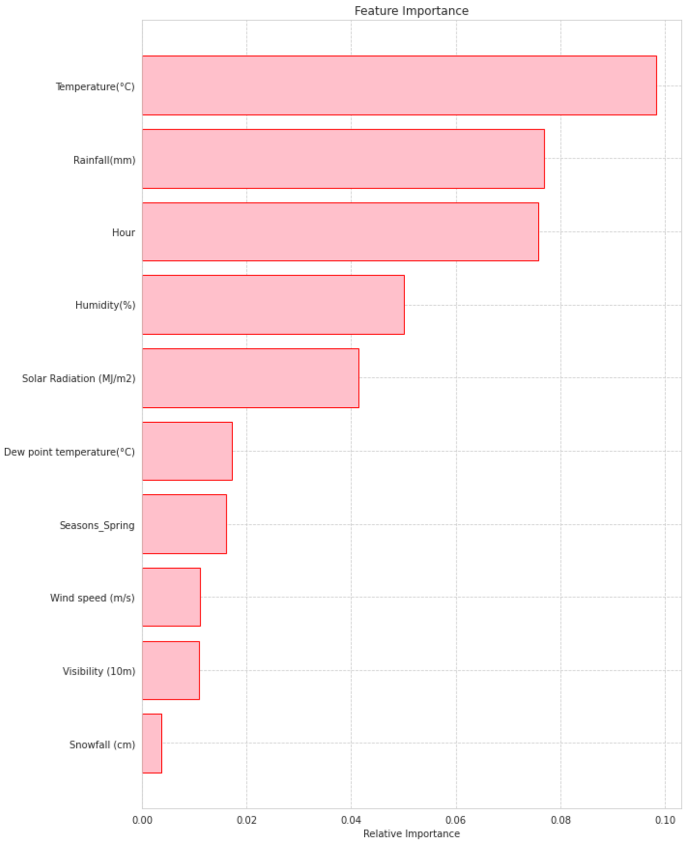


SHAP



LIME

Explaining Extreme Gradient Boosting



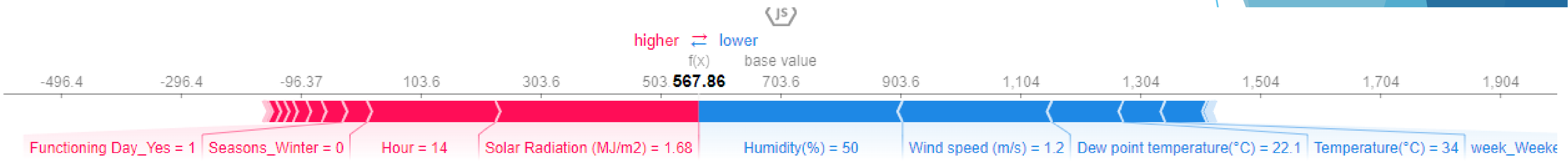
Feature Importance

y (score **567.364**) top features

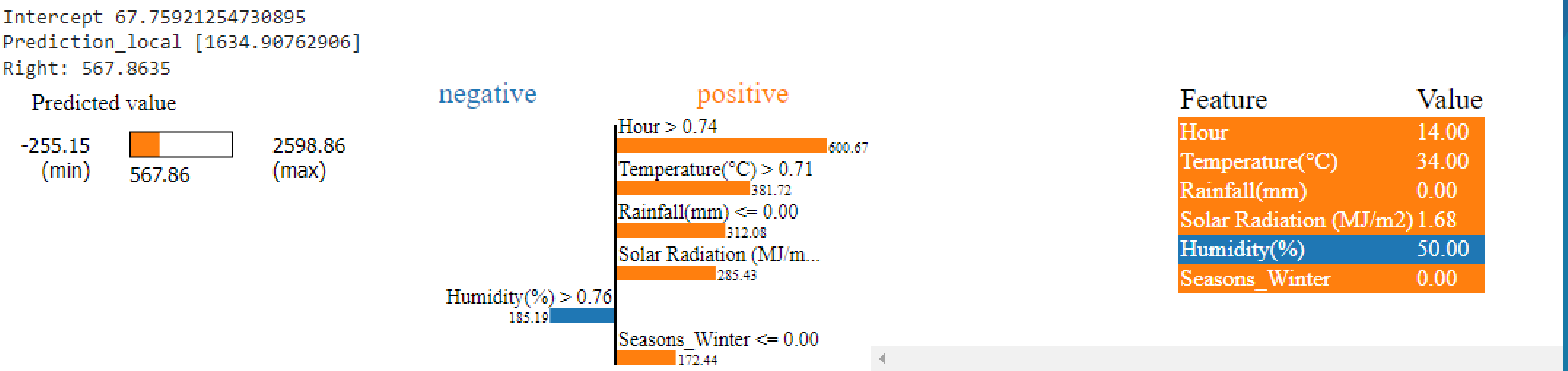
Contribution?	Feature	Value
+703.132	<BIAS>	1.000
+342.393	Solar Radiation (MJ/m2)	1.680
+265.845	Hour	14.000
+66.364	Temperature(°C)	34.000
+37.944	Functioning Day_Yes	1.000
+37.522	month_July	1.000
+36.865	Seasons_Winter	0.000
+34.178	Rainfall(mm)	0.000
+24.854	Seasons_Spring	0.000
+23.034	Seasons_Summer	1.000
+22.177	Visibility (10m)	1744.000
+2.736	Holiday_No Holiday	1.000
+1.875	month_September	0.000
+1.670	month_November	0.000
+0.738	month_February	0.000
+0.652	month_January	0.000
+0.545	Snowfall (cm)	0.000
+0.329	year_2018	1.000
-0.825	month_May	0.000
-2.417	month_March	0.000
-2.587	month_December	0.000
-4.388	month_October	0.000
-5.311	month_August	0.000
-8.658	month_June	0.000
-112.748	Dew point temperature(°C)	22.100
-123.622	week_Weekend	1.000
-244.414	Wind speed (m/s)	1.200
-530.518	Humidity(%)	50.000

ELI5

Explaining Extreme Gradient Boosting



SHAP



LIME

Model Validation & Selection

- **Observation 1:** As seen in the Model Evaluation Matrices table, Linear Regression is not giving great results.

- **Observation 2:** Random forest & Bagging have performed equally good in terms of adjusted r2.

- **Observation 3:** We are getting the best results from Stacking and XGBoost.

Challenges

- A huge amount of data needed to be deal while doing the project which is quite an important task and also even small inferences need to be kept in mind.
- As dataset was quite big enough which led more computation time.

Conclusion

- We observed that bike rental count is high during week days then weekend days.
- The rental bike counts is at its peek at 8 AM in the morning and 6pm in the evening, We can see an increasing trend from 5am to 8 am, the graph touches the peak at 8am and then there is dip in the graph. Later we can see a gradual increase in the demand until 6pm, the demand is highest at 6 pm, and reduces there after until midnight,
- We observed that people prefer to rent bikes at moderate to high temperature, and even when it is little windy,
- it is observed that highest bike rental count is in Autumn and summer seasons and the lowest is in winter season.
- We observed that the bike rentals is highest during the clear days and lowest on snowy and rainy days.
- when we compare the RMSE and Adjusted R2 of all the models, Stacking Regressor gives the highest Score where R2 score is 0.90 and Training score is 0.95 so this model is the best for predicting the bike rental count on daily basis.

THANK YOU