Supervised ML - Regression

Bike Sharing Demand Prediction

By- Ranajay Biswas

Renting a Bike?

Bike renting is now-a-days very common and desirable method of transportation for many. Whether you're visiting for a short trip or you are doing your daily chores, bike sharing has quickly become one of the most popular forms of short distance travelling.

The benefits of bike sharing includes transport flexibility, reductions to vehicle emissions, health benefits, reduced congestion and fuel consumption, and financial savings for individuals. But **the most special quality of public bicycles is the idea of sharing**.

Bike share programs **increase the visibility of cyclists, making riding safer for everyone**. Studies also show that more people riding bikes in urban areas leads to improved bicycling and walking infrastructure.

So, what's the problem?

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.

So, it seems we need to know in advance how many bikes we might need at a given time or day. This makes life so much easier for everyone.

How do we solve this?

We have Seoul Bike Sharing data, which we can use to predict the number of bikes that might be in demand for rentals.

We need to follow some steps in order to get there.

1. Setting the ultimate Goal.
2. Understanding the dataset.
3. EDA & Feature Engineering.
4. Preparing data for modelling.
5. Applying models.
6. Model Validation & Selection.

Ultimate Goal : We need to build a model that will take into account various influencing factors and finally predict the number of bikes that will be in demand for rental. This would seem to solve our problem and help the business to stay prepared.

Data Summary :

The dataset has 8760 observations and 14 variables.

One of which is the dependent variable, which our model will be predicting the value for. The dependent variable is -

Rented Bike Count : The values for this column is continuous and numeric in nature.

Independent Variables :

Now, that we have our dependent variable, it's time to focus on the predictors or Independent variables.

Starting with --->

Date : Data-type is object initially. Contains records from December 1st 2017 to 30th November 2018.

Hour : The values are in 24 hour format. Starting from 0-23. Data-type is integer. This variable helps us understand hourly demand.

Temperature(°C) : Data-type is float. This variable tells us how the customers prefer to go for rental bikes depending on the temperature.

Humidity(%), Dew point temperature, Solar Radiation, Visibility, Wind Speed, Rainfall & Snowfall are also weather related parameters influencing the dependent variables.

Seasons : We will be able to determine the seasonal effect on bike rentals with the help of this variable. There are 4 unique seasons listed in our data i.e Summer, Autumn, Spring and Winter.

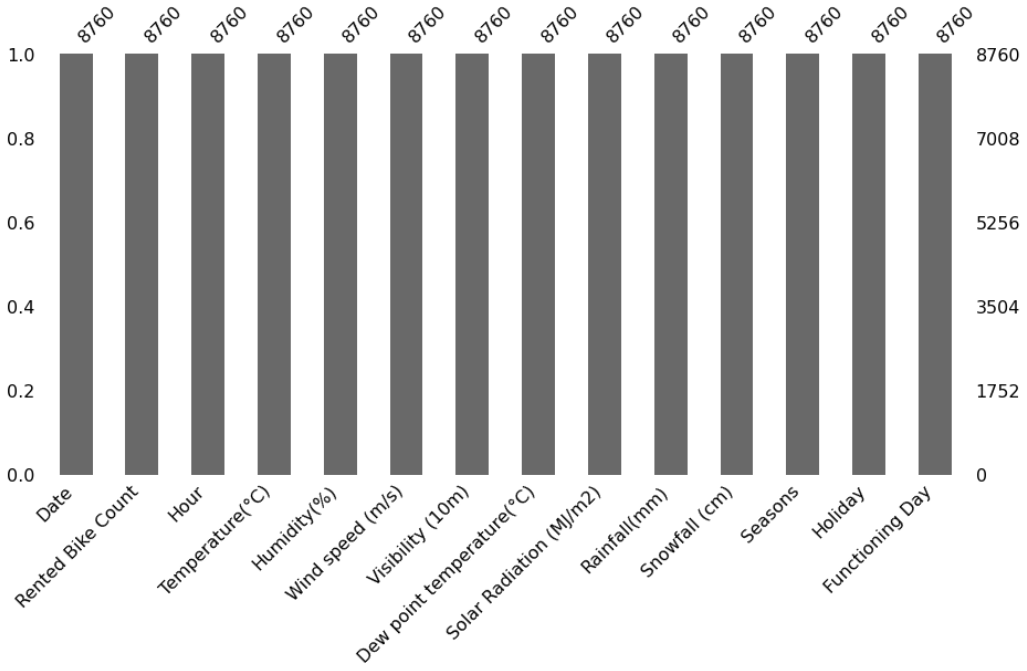
Holiday & Functioning Days columns helps us understand the difference in bike demand depending on whether it's holiday or not.

EDA & Feature Engineering :

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. This is where we have spent most of the time.

EDA Continued ...

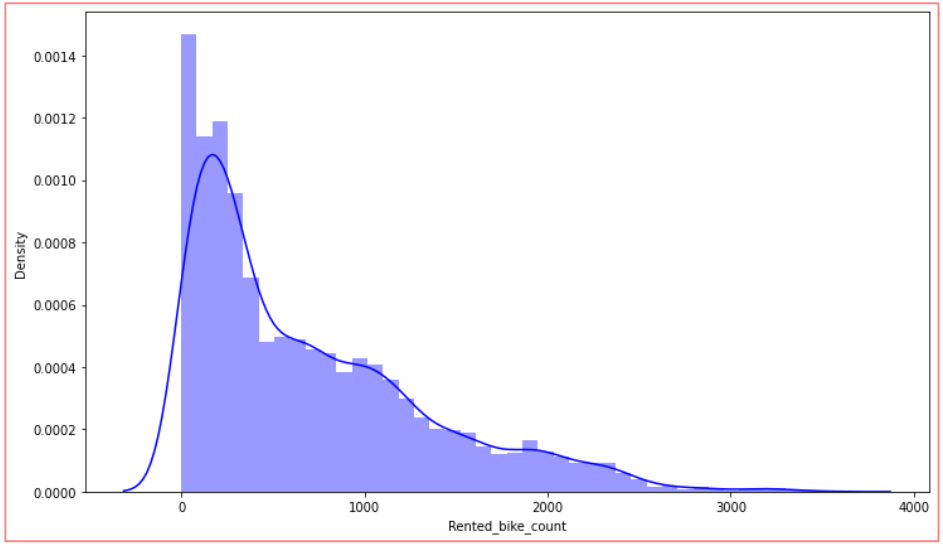
Null and Duplicate Values : First thing we did was to check for missing values in the data.

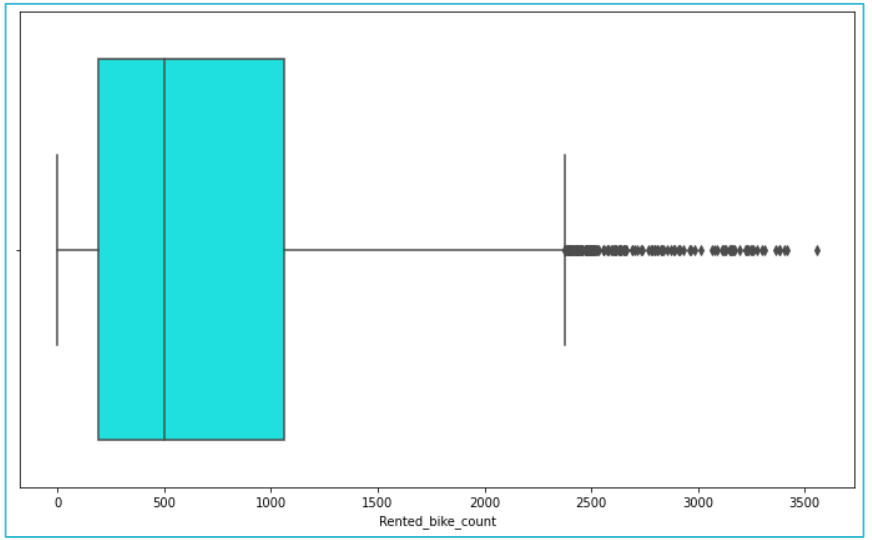


As it is evident from our plot, we fortunately did bot have any missing values in the data. That's why no dropping or imputations were needed.

Next we searched for duplicate values. There were no occurrences of duplicate values as well.

Distribution of Rented Bike Count :

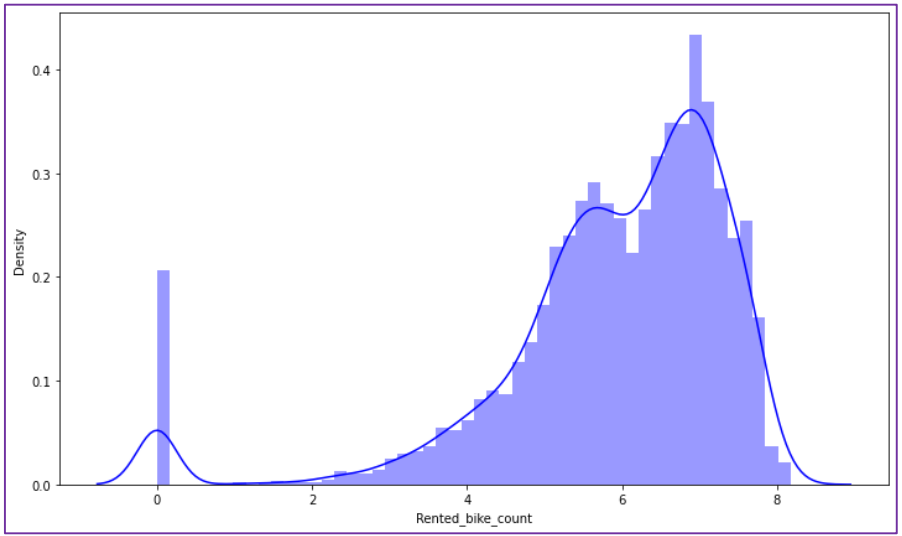


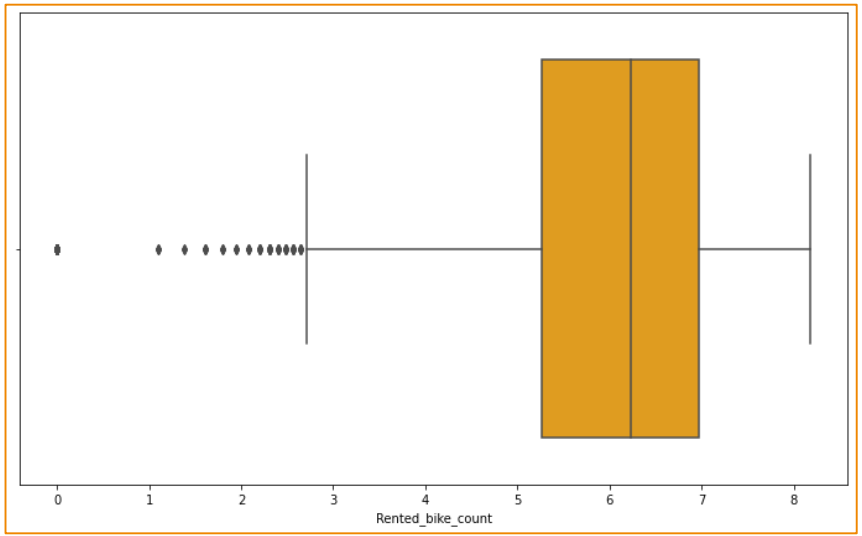


- Distribution for the dependent variable Bike Count is right skewed.
- From the boxplot, it is evident that many outlier values are present.

Since, these values behaving as outliers, could very well be real values, we cannot just get rid of them. Hence, let's try log transformation.

Log Transformed Rented bike Count :





We get these plots after performing the log transformation. Still, skewness and outliers remained. But these values are much more acceptable.

We also tried square root and log10 transformations for the data as well and checked the distributions.

Continued ...

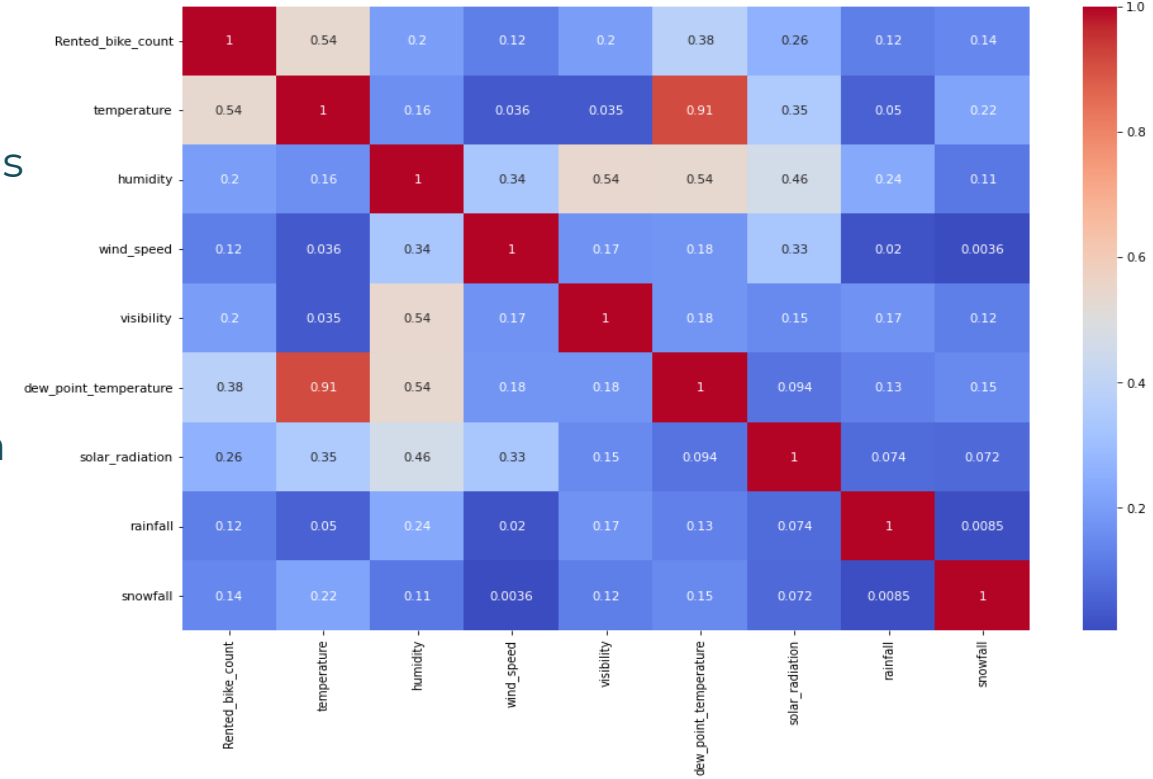
- Original ‘Date’ column is object type in nature. We changed it date-time format and then extracted year, month, name of the day and the dates from this and stored the values in respective columns.
- We dropped the observations that were having the value 0 for rented bike count, since those values only occurred when it was non-functional days and these values were only going to make our regression models worse as we would never be predicting the value 0 for any observations.
- From the ‘day’ column, we made another variable called ‘weekend days’ and stored the value as 1 for Sundays & Saturdays and value 0 for regular week days.
- ‘Holiday’ column values were updated with value of 1 and 0 were assigned for holidays and non-holidays respectively.

- Similar approach was taken for ‘Functioning days’ column as well.
- The original ‘Date’ column was dropped along with ‘year’ & ‘day’ column.
- Data-types for hour, seasons, holiday, functioning day, month, date & weekend days were changed into categorical.

Continued ...

Correlation Heat-map :

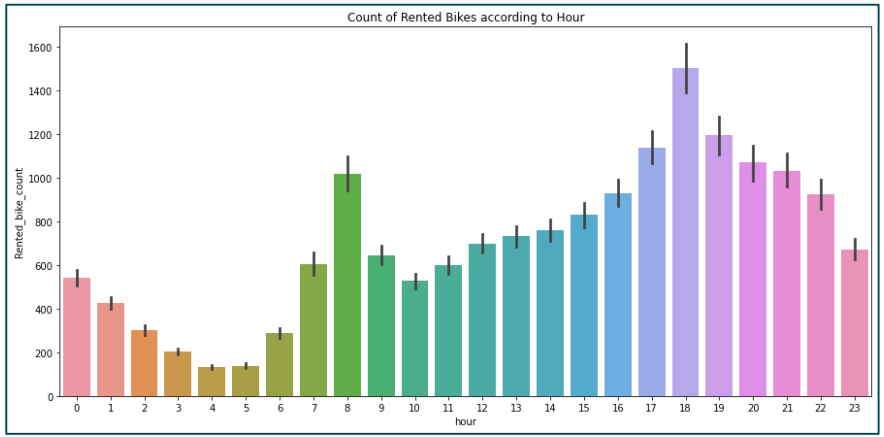
- From the heat-map, it is evident that Temperature has the highest correlation with number of Bike rents.
- Temperature and Dew point temperature are strongly correlated to each other with a correlation value of 0.91
- In order to avoid multi-collinearity, we dropped the feature Dew point temperature.



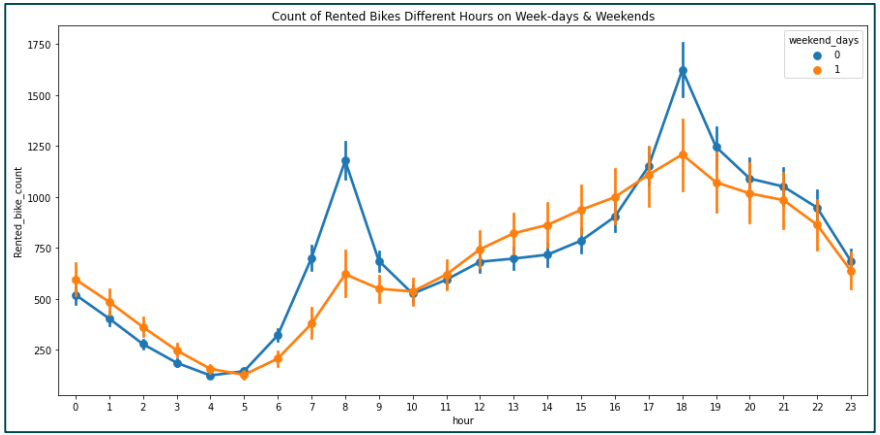
Visualization :

Visualization will help us understand influences, patterns & trends in the data in a very effective way. Let's see what insights we can draw by checking the relations between Rented Bike count and various different features in the data.

Rented Bike Count vs Hour

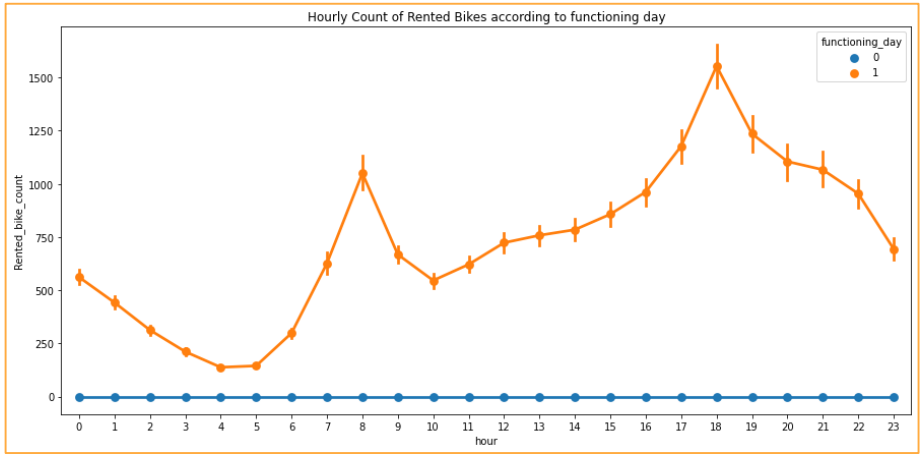


Rented Bike Count vs Hour on Weekends



- *The conclusions we can derive from above plots are that on weekends, bikes rentals have been relatively* lesser.
- *We see a peak around* 8 am *in the morning and around* 5-7 pm *in the evening.* 6pm *is the busiest hour for business. This may be because many of the working professionals prefer to rent bikes after work.*

Visualization :

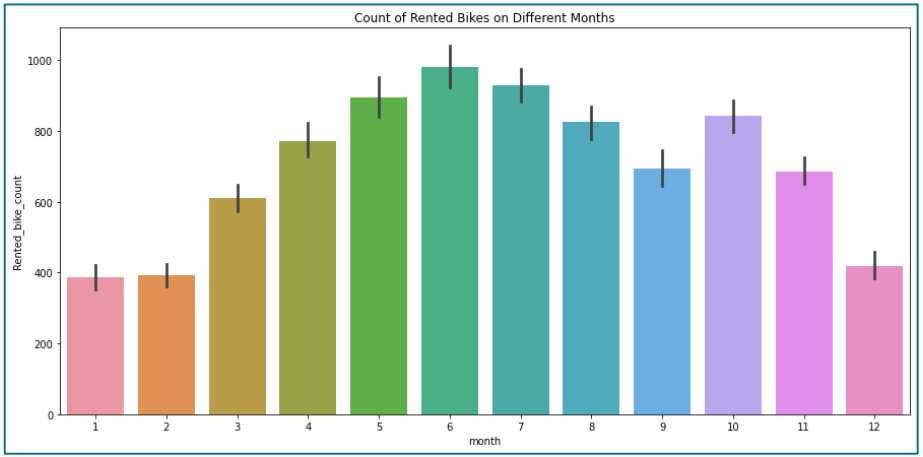


Hourly Count of Rented Bikes based on Functioning Days -

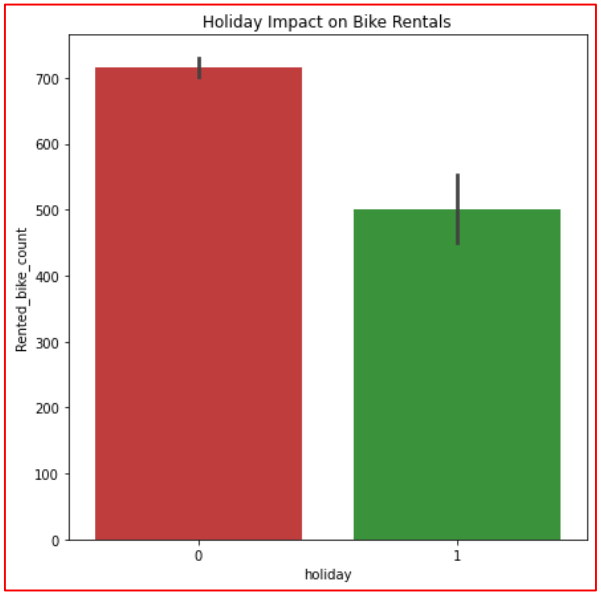
- *For non-functioning days, bike renting services were not available.*

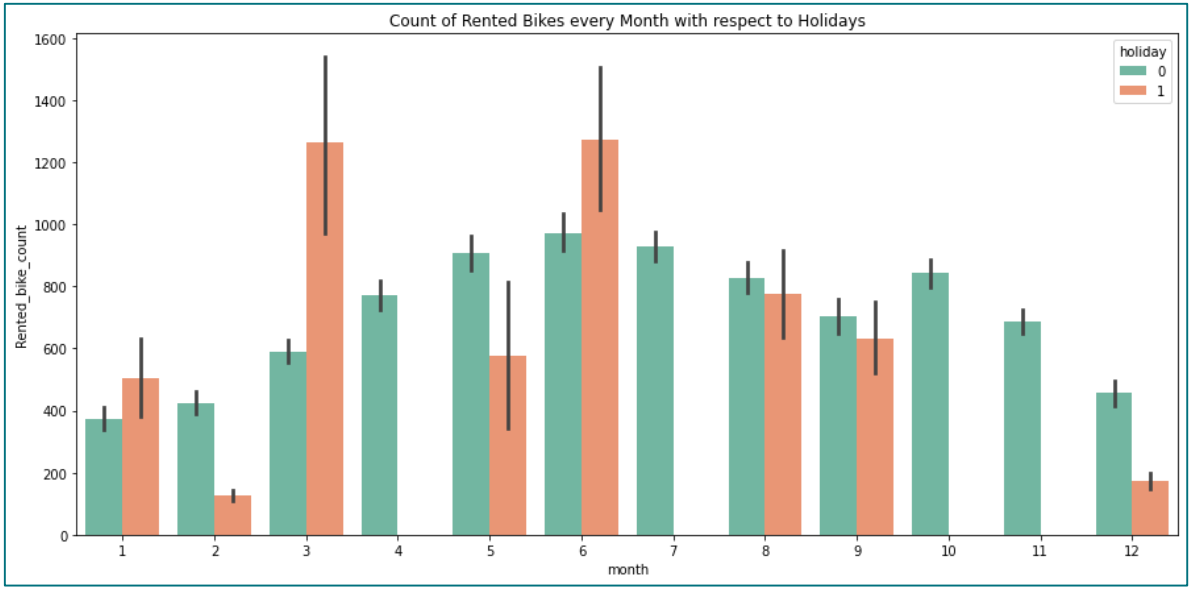
Count vs Months :

- *Months of* January, February *and* December *recorded the least number of bikes rents.*
- *During* May – July, *the numbers go up the most.*
- *Month of* June *recorded the best numbers.*



Visualization :

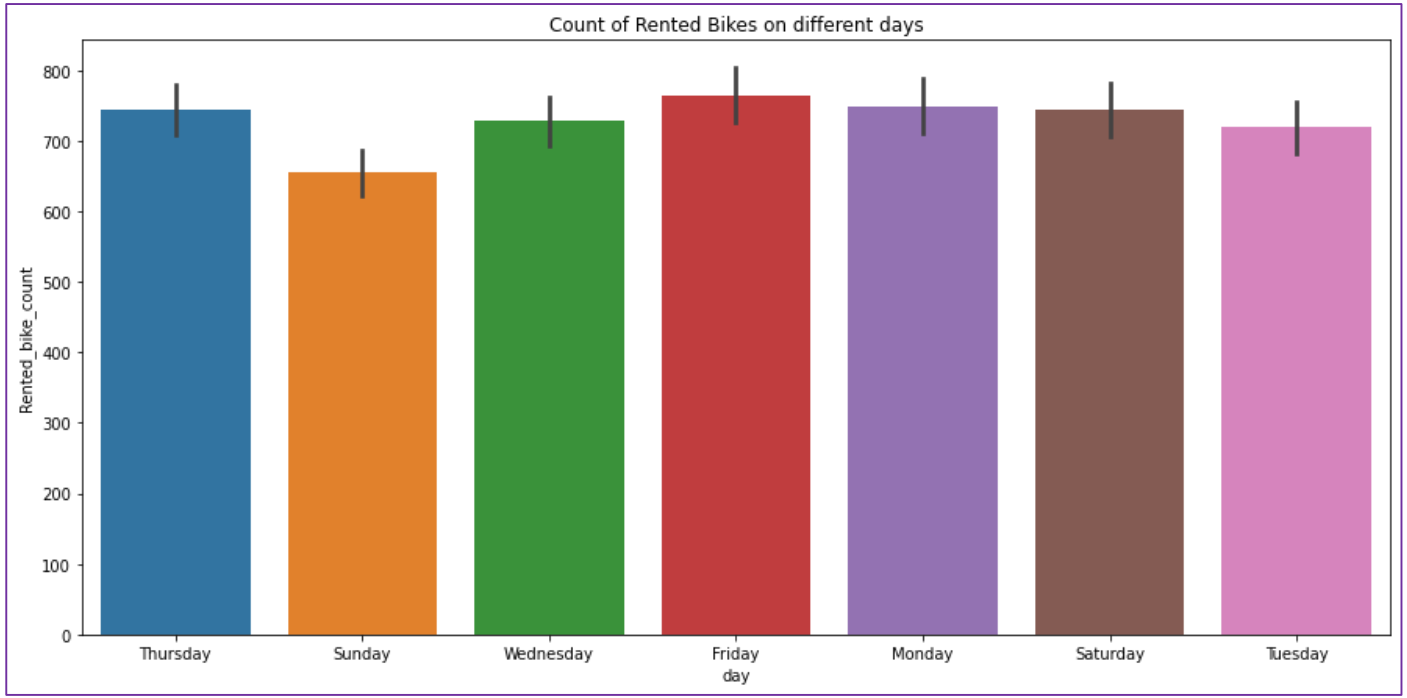




Holiday effect on Rented Bike Count :

- *Rented bike count has been usually higher on working days compared to holidays.*
- *Only in the months of* January, March *and* June, *there were more bikes rented on holidays.. Guess people love to go to vacations on these months and many like to travel on bikes too.*

Visualization :



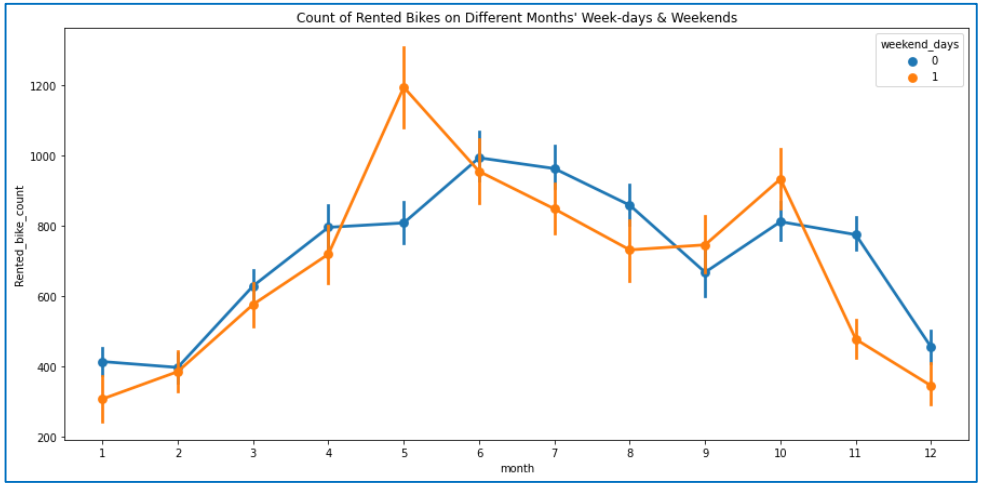
Rented Bike Count on different Days of the week :

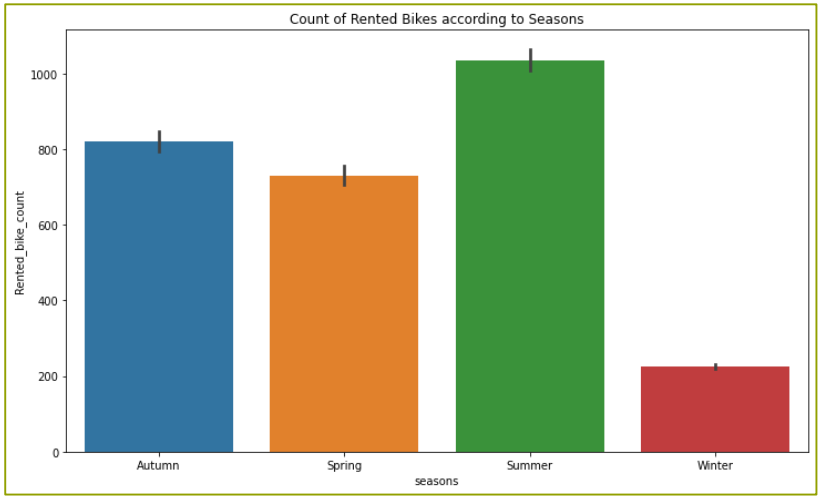
- *We can say that Rented bike count is more or less the same throughout the entire week.*
- *Though Sundays is the least busy day in the week.*

Visualization :

Bike rents on weekdays and weekends each month :

→ *Normally, bike rents numbers were better on weekdays throughout the year, except for* May *and* October.

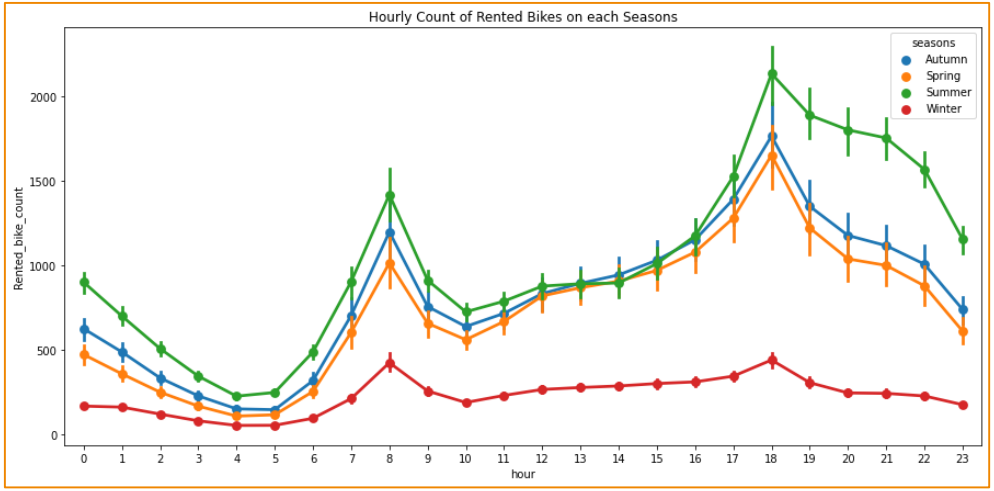


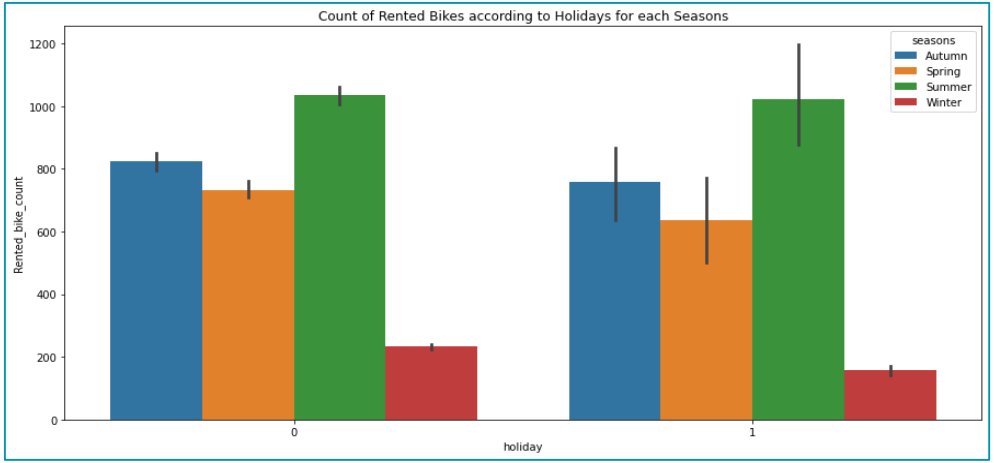


Seasonal Effect on Bike Rents :

- Summer *was most preferred for bike travels.*
- Winter *was least preferred by travellers.*

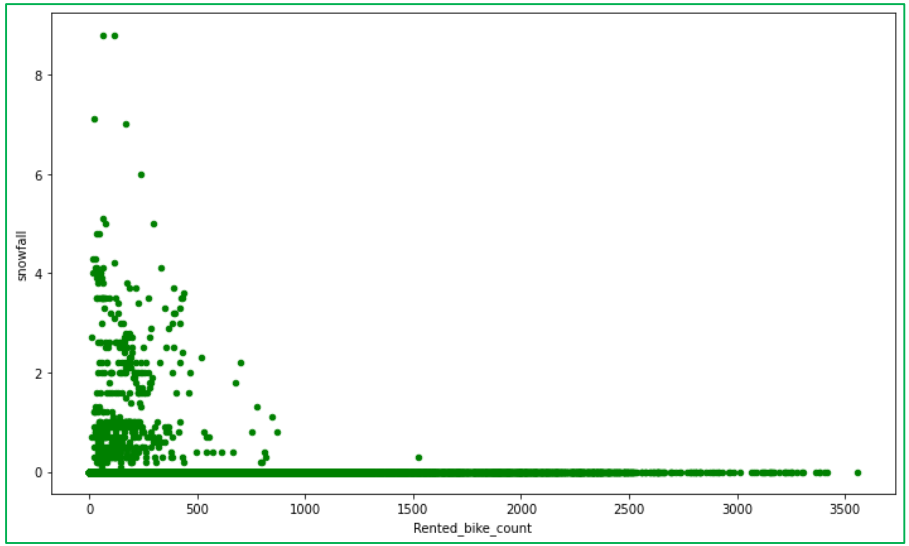
Visualization :



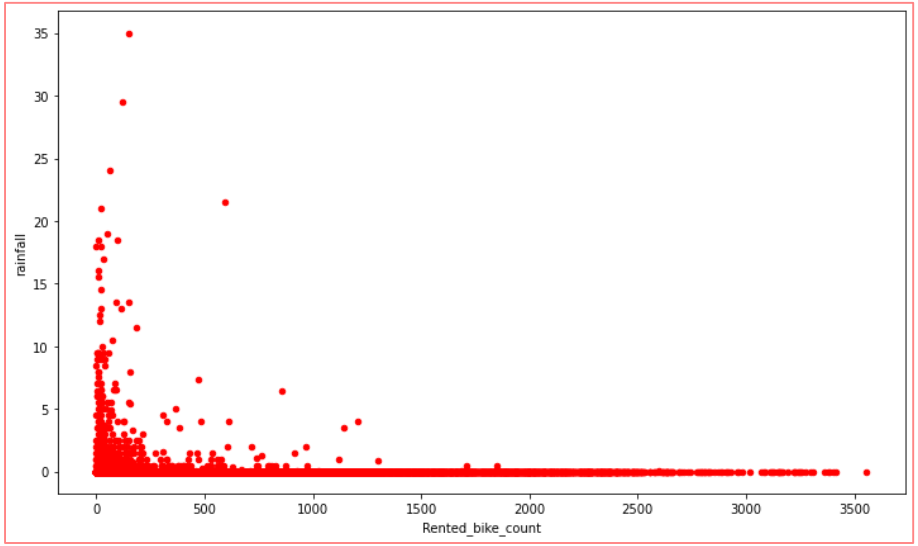


- It is evident that patterns and trends were similar for every seasons, only the numbers were different.
- The same thing happens for Holidays and non-holidays on every season.

Visualization :



Count vs Snowfall



Count vs Rainfall

No surprise! People do not like to ride bikes when it's raining or snowing outside. So, Rented Bike Count has a negative correlation with both Snowfall and Rainfall.

Data Preparation :

Transformation : To train most models, we need to do the encoding for the categorical columns. We take hour, seasons and month columns and perform one hot encoding on them.

After that we are left with 8465 rows and 47 columns in our dataframe including the dependent variable.

Data Scaling : Before feeding the data to ML models, we always need to make sure that all our features are scaled. If not, then model will get biased for certain variables with big values and importance for other variables will be taken less into consideration. We need to avoid that.

That's why we performed the *MinMaxScaler* on the features to get all the values shrank down between 0 and 1.

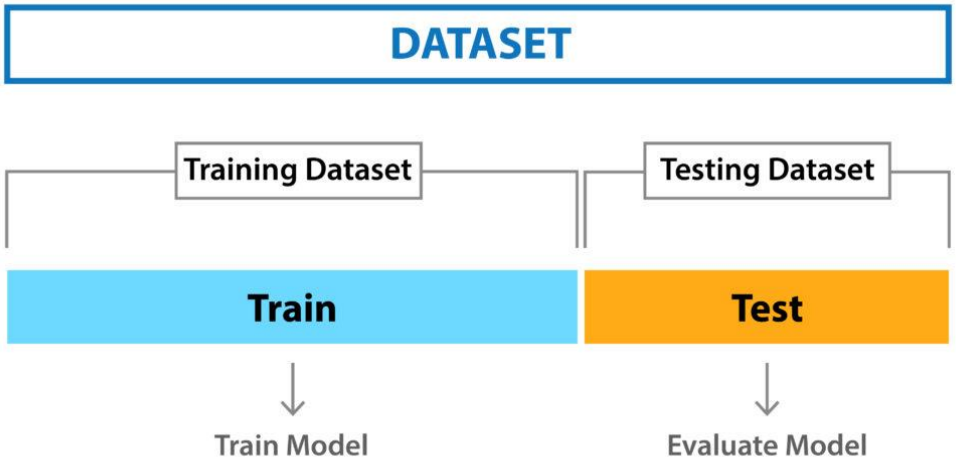
For the target variable, we did log10 transformation to avoid the outlier influence as much possible.

Data Preparation :

Train-Test Split : Train test split is a model validation procedure that allows you to simulate how a model would perform on new/unseen data.

This is a very important procedure for any machine learning technique.

We used sklearn's train_test_split method to divide the data. 80% of both independent and dependent variables were used for training and remaining 20% was left for evaluating as test data.



6772 observations used for training and 1693 observations for validation.

Applying Regression Models :

We applied the following baseline models and the results for the models were -

Model	R2 Score	Adjusted R2	MSE Value	RMSE Value
XGBoost Regressor	0.9	0.88	0.024	0.15
Gradient Boosting	0.88	0.88	0.026	0.16
Random Forest Regressor	0.86	0.86	0.032	0.18
Decision Tree Regression	0.79	0.79	0.048	0.22
Linear Regression	0.74	0.62	0.062	0.25
Ridge Regression	0.74	0.62	0.062	0.25
Elastic Net Regression	0.73	0.6	0.064	0.25
Knn Regressor	0.62	0.65	0.089	0.3
Lasso Regression	0.007	-0.26	0.237	0.48

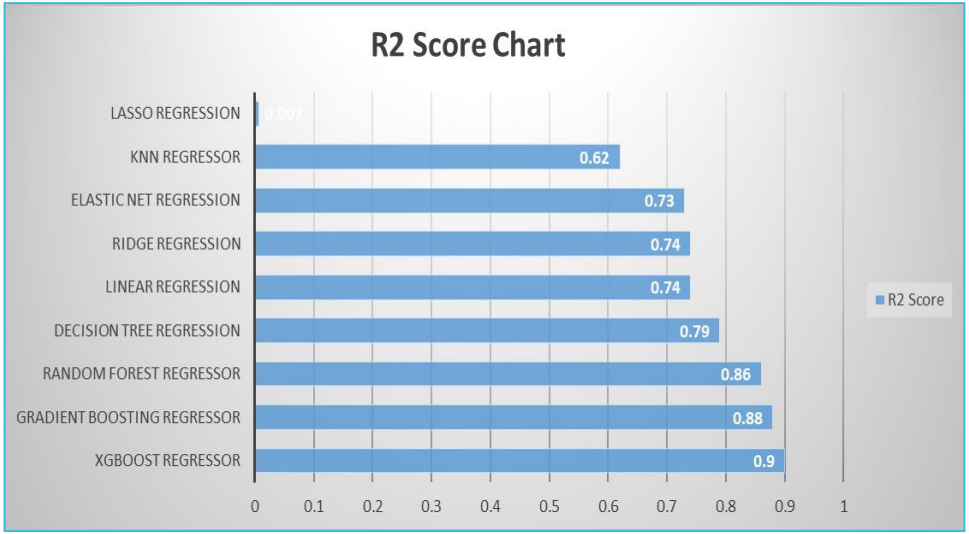
Metrics used for model evaluation :

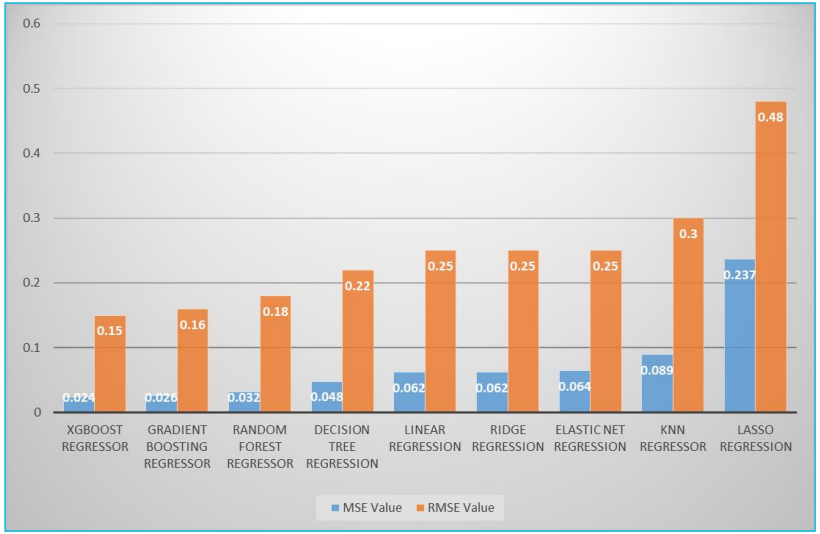
- R2 Score : It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset.
- MSE Value : Mean squared error (MSE) measures the amount of error in models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero.
- RMSE Value : Root mean square error or root mean square deviation shows how far predictions fall from measured true values using Euclidean distance.
- Adjusted R2 Score : The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable.

Scores(For Base Models) :

Best scores were acquired from **XGBoost** model with R2 score of **0.9** and MSE value of **0.024**

Gradient Boosting and **Random Forest** models also performed well.



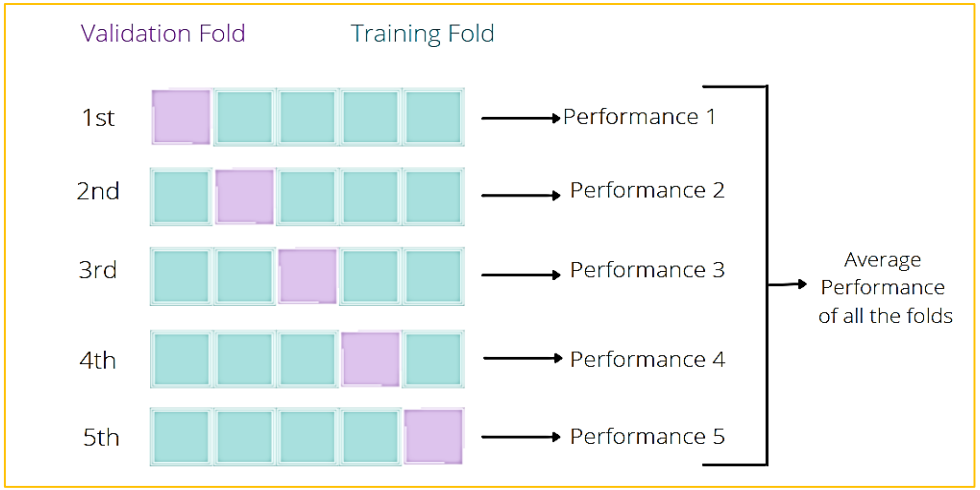


R2 Score for base models

MSE & RMSE Values

Data Preparation :

Cross Validation & Hyper parameter tuning : Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations.

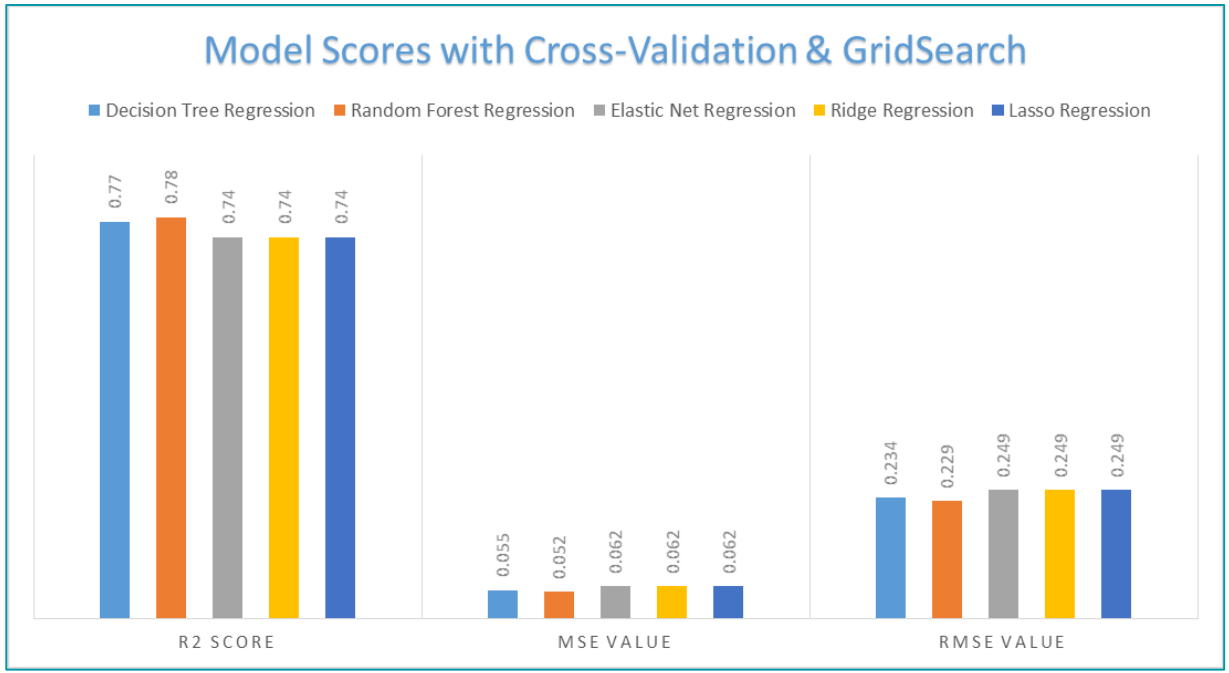


We used **k-fold cross validation**, where 'k' is the number of folds within the dataset.

GridSearchCV gives the best hyper parameters for a given model.

We combined these two methods to get the most generalized and well optimized model with the best scores possible.

Applying Cross Validation with GRidSearch :



Using K-fold Cross Validation, we made sure our data was well generalized.

GridSearch helped us to fine tune and find the best hyper-parameters for the all the different models.

We trained the models and find these results –

Random Forest model worked the best in this case. **R2 Score** for the model was **0.78, MSE** value was **0.052** and **RMSE** value was **0.229**

Conclusion :

With that, we have come to an end of our project. To summarize the entire process -

- We performed data cleaning and EDA. Used visualizations to better understand the relations between the various features that were present in our data.
- We come to an understanding of how different features impact the number of bikes that are being rented, the effect of weather conditions, temperature and holidays on bike rents. We also figured out which were the busiest hours of the day for business.
- In the pre-processing step, we removed multi-collinearity and unimportant features from the data. Encoding and feature scaling were done. So that the data can be passed through our ML models for training and validation.
- Finally models were trained to make predictions with the least margin of error possible. Then we chose the best model, which in our case was Xgboost.

With the assistance of our deployment-ready machine learning model and our analysis of the industry, the business will be able to understand when the demand for bikes will be high. Thus, they will be able to keep up with the supply in order to maximize the profits.

Thank You