Capstone Project

Seoul Bike Sharing Demand Prediction

By : Vijay N

Points to discuss

- Defining problem statement
- Data summary
- EDA
- Feature engineering
- Applying ML algorithms
- Comparing different ML models
- Conclusion

Defining problem statement

Rental bikes service is very crucial 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.

Therefore we have to predict the number of rental bikes required in each hour for smooth functioning of service.

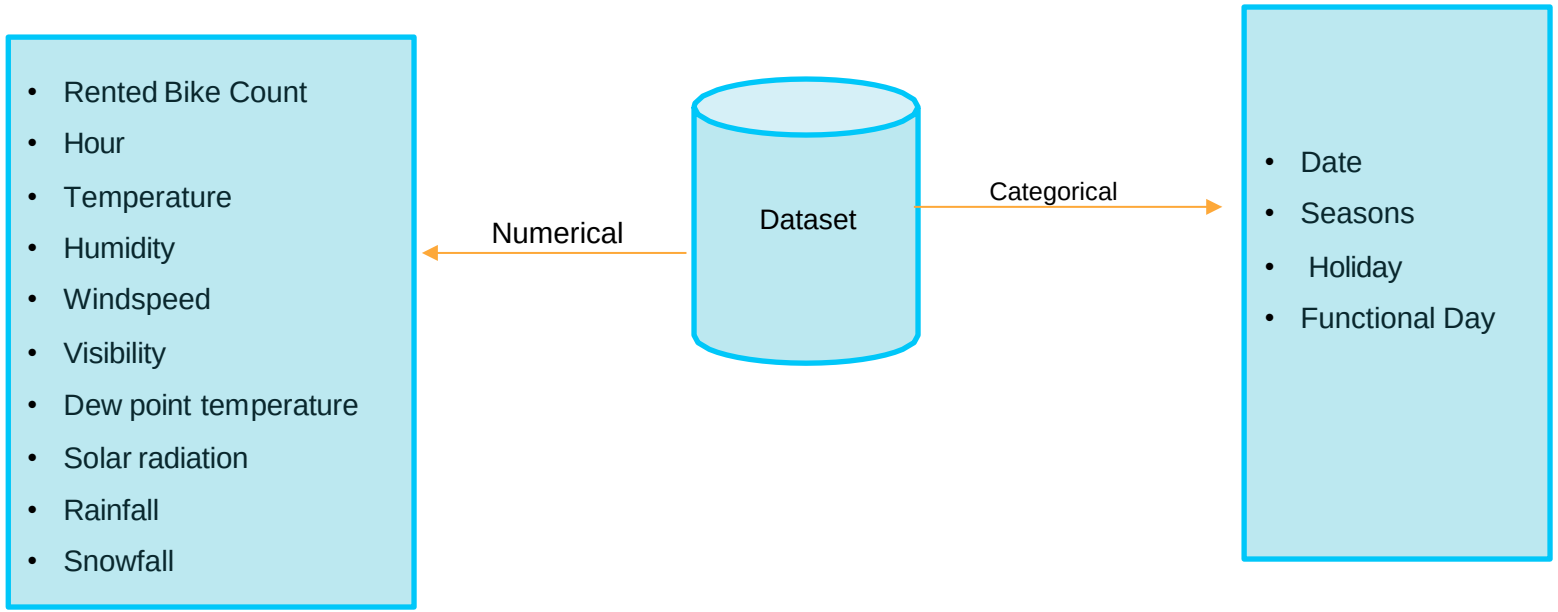
Data Summary

We are given dataset containing count of rental bikes from December 2017 to November 2018 for each day and each hour of day. Along with count of rental bikes there are following variables also present.

(1) Date : year-month-day
(2) Rented Bike count - Count of bikes rented at each hour
(3) Hour - Hour of the day
(4) Temperature-Temperature in Celsius
(5) Humidity - %
(6) Windspeed - m/s
(7) Visibility - 10m
(8) Dew point temperature - Celsius
(9) Solar radiation - MJ/m2
(10) Rainfall - mm
(11) Snowfall - cm
(12) Seasons - Winter, Spring, Summer, Autumn
(13) Holiday - Holiday/No holiday
(14) Functional Day - No (Non Functional Hours), Yes(Functional hours)

'Rented Bike count' is dependent variable.

Data Summary

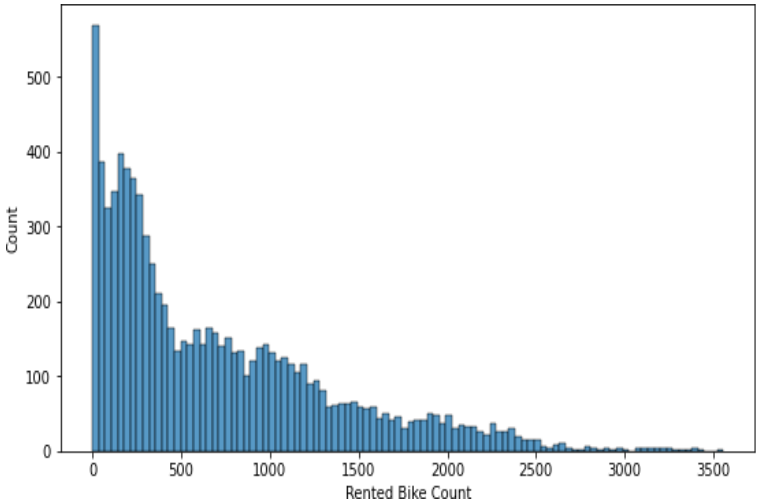


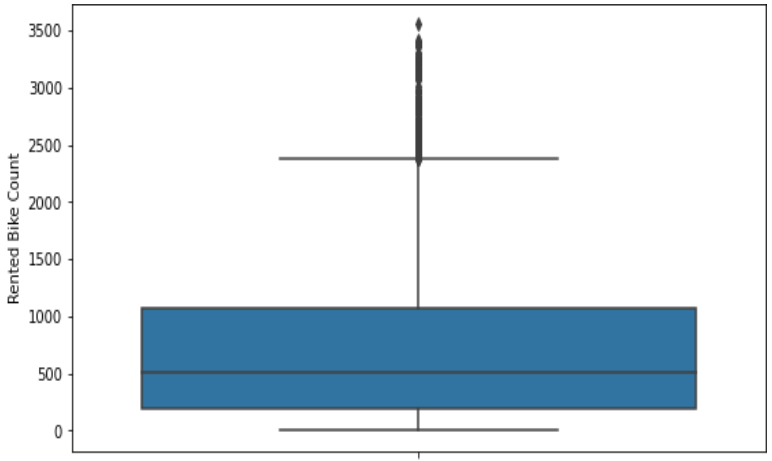
Dependent variable is 'Rented Bike Count' that describes the count of rented bikes for each hour.

EDA

Let us see how the values of ‘Rented Bike Count’ are distributed in given dataset.

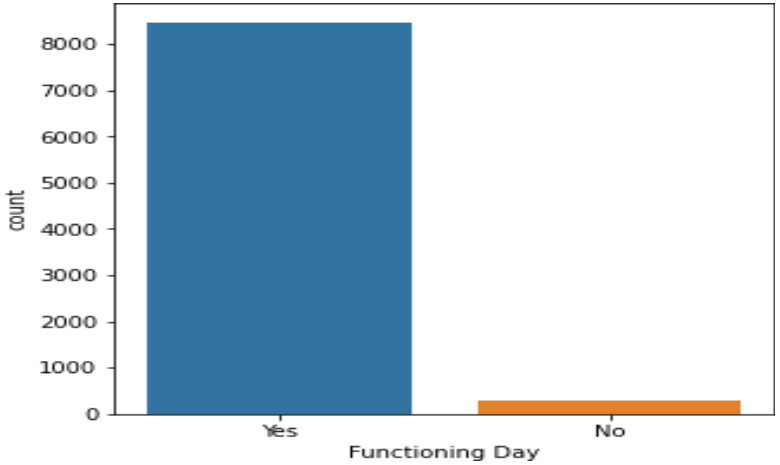
Distribution of values is highly positively skewed.

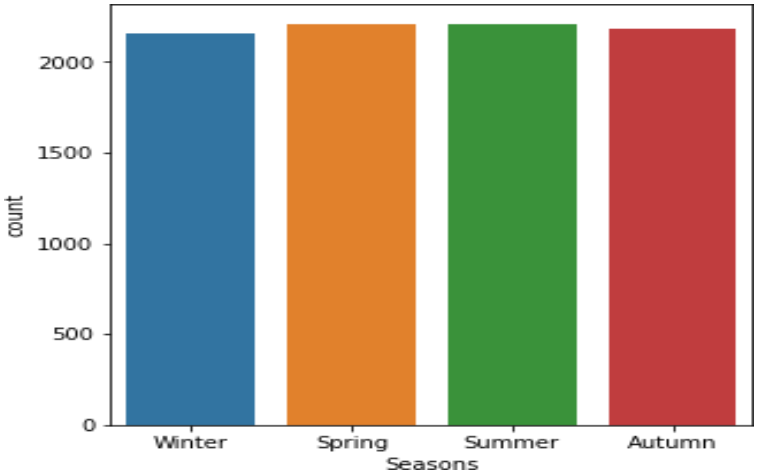


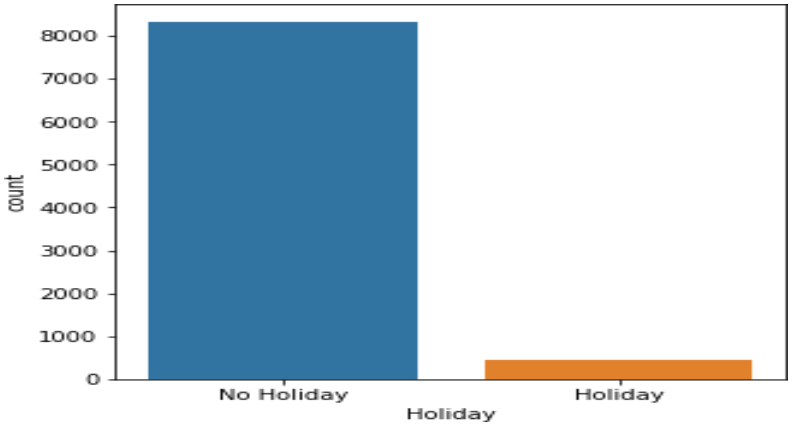


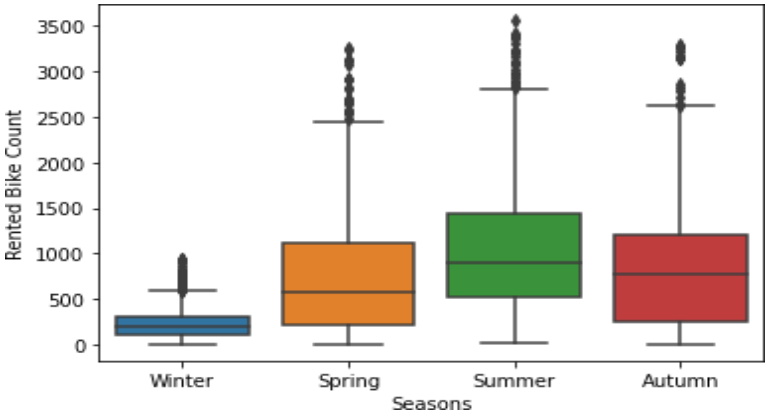
Count of values of categorical features.

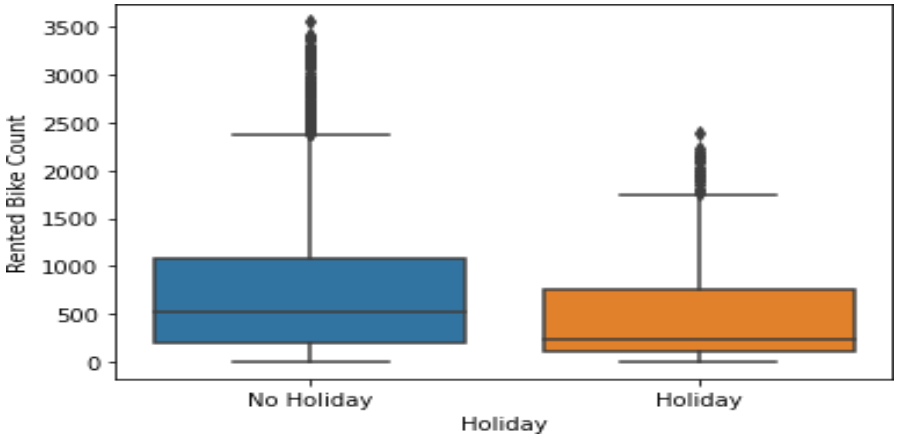
Functioning Day and Holiday have highly imbalanced count of values.



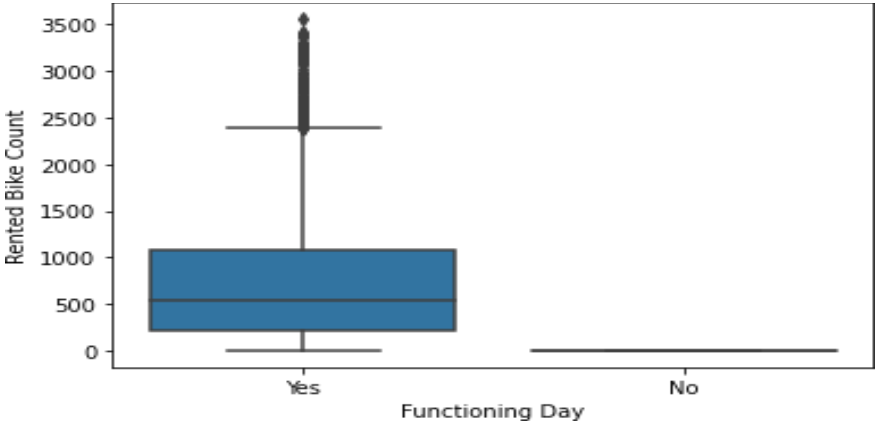




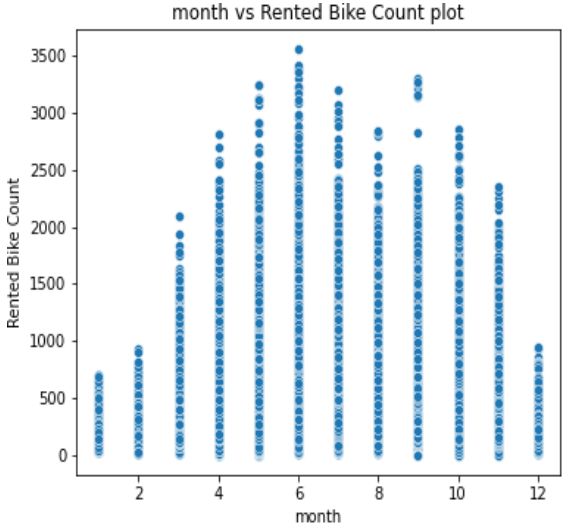


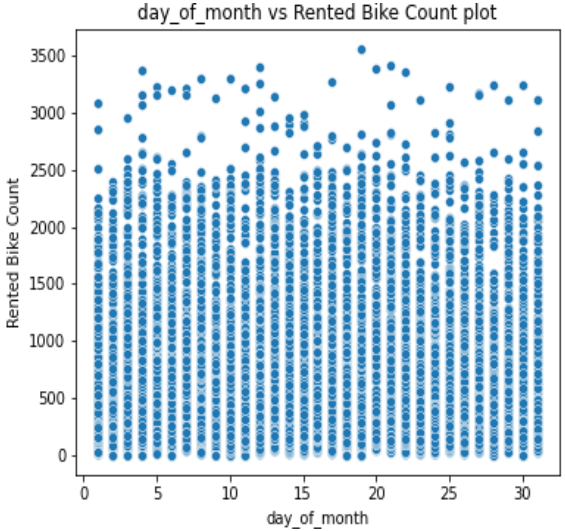


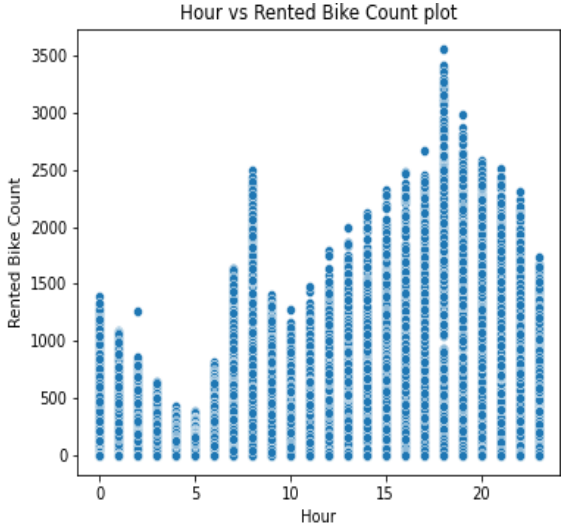
- Demand for rented bikes is low in winter seasons and high in summer season.
- Demand for rented bikes is low on holidays and non functional days.

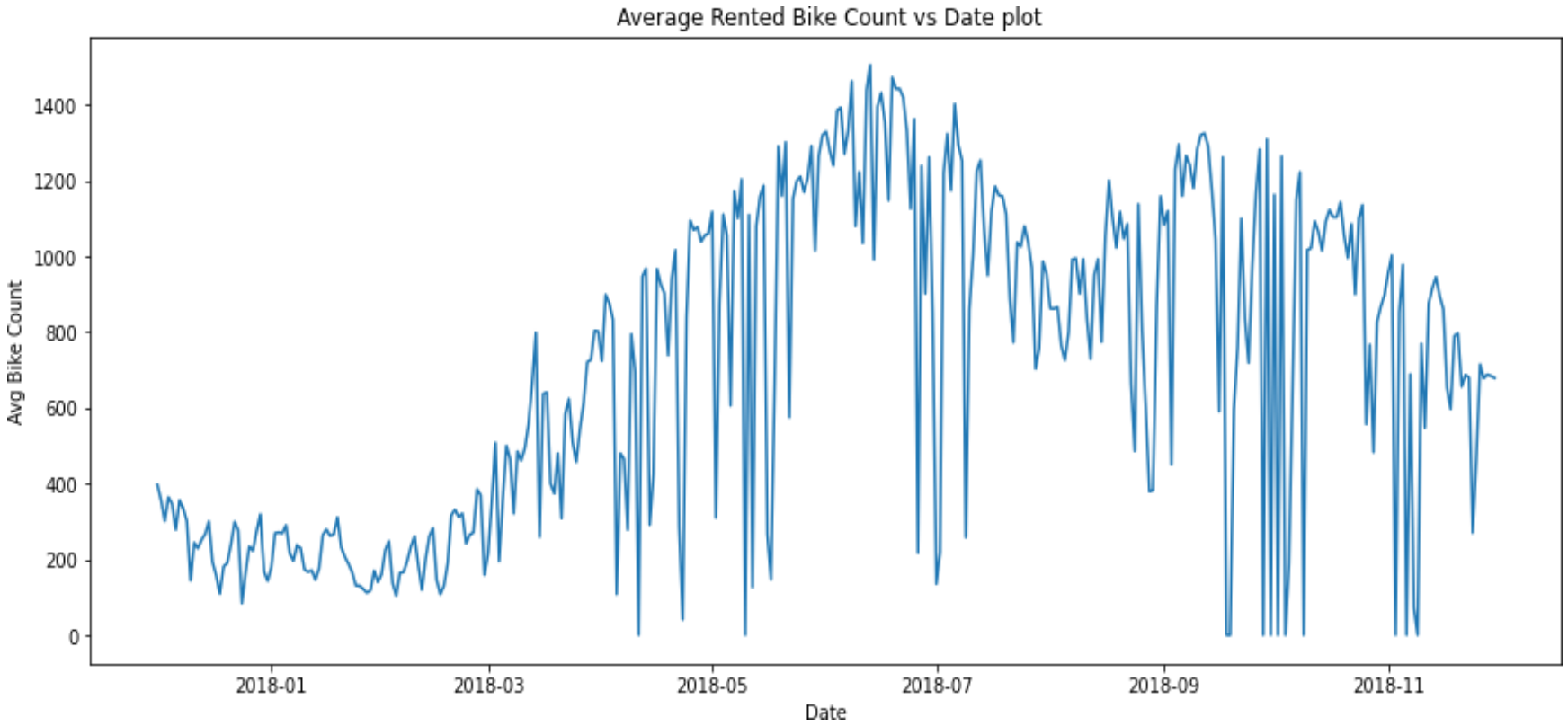


- Count of rented bike varies up to higher values from May to September.
- Within a month count of rented bikes is almost same.
- Demand for rented bikes increases during evening hours.

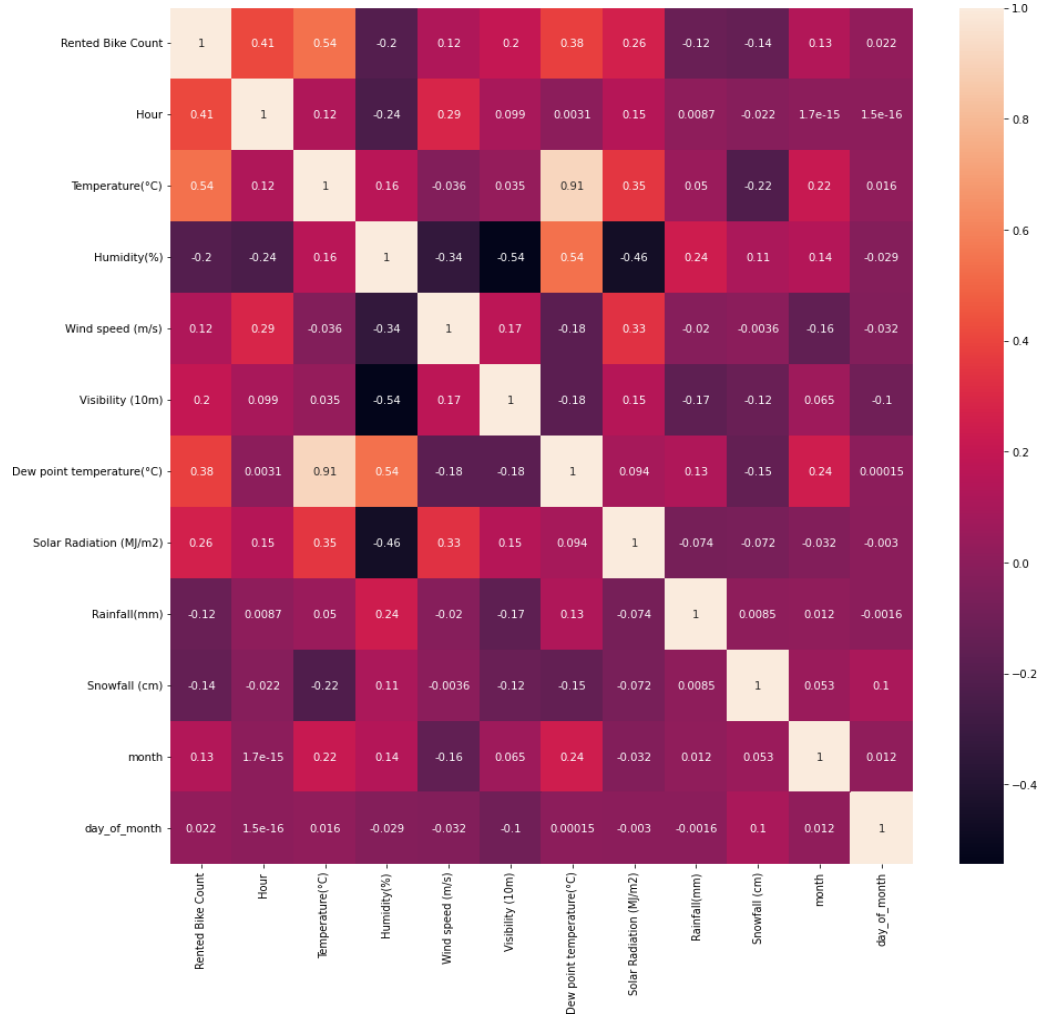








Average rented bike count is comparatively high from middle to end of year.



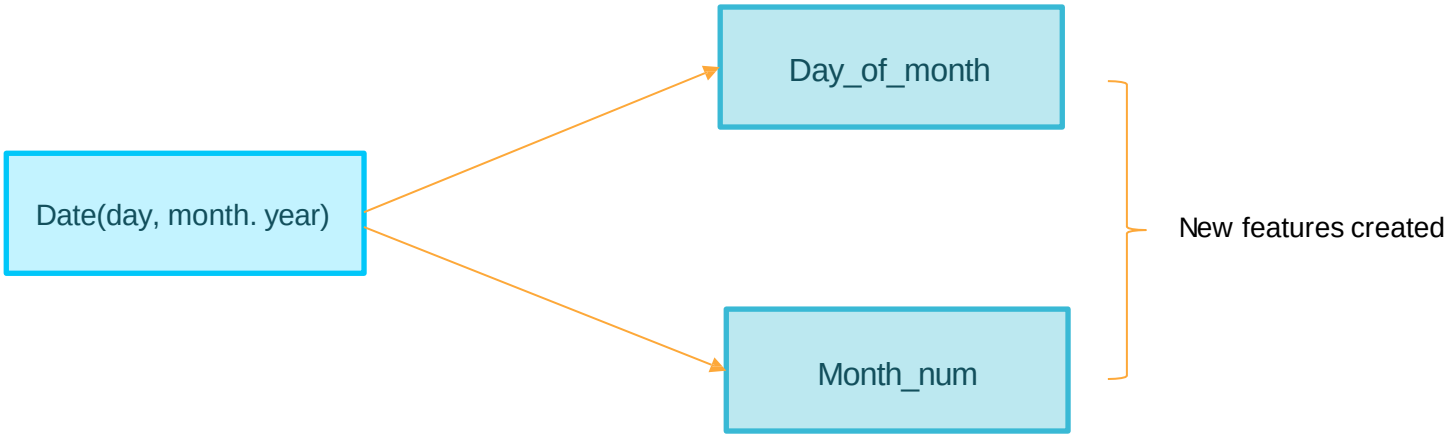
- Very high correlation between ‘Temperature’ and ‘Dew point temperature’ which is obvious.

- ‘Humidity’ is moderately correlated with ‘Solar Radiation’ and ‘Visibility’.

Feature Engineering

Feature engineering is the process of selecting, manipulating, and transforming initial variables into features that can be used in model training.

New features creation

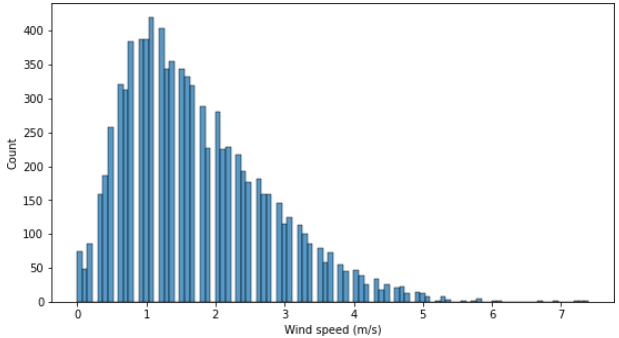


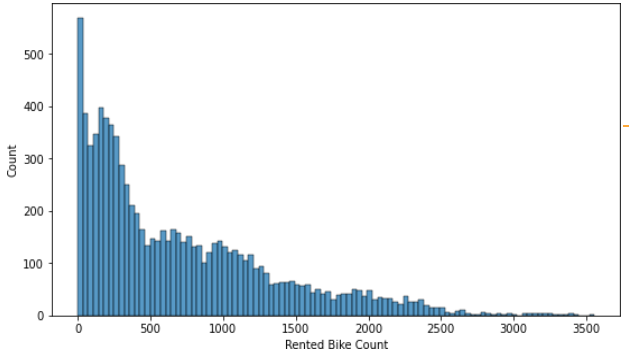
Features transformation

Some of our features are skewed so we can apply sqrt, log10 or reciprocal transformation on them to reduce skewness.

Here we have applied sqrt transformation.

Original Features

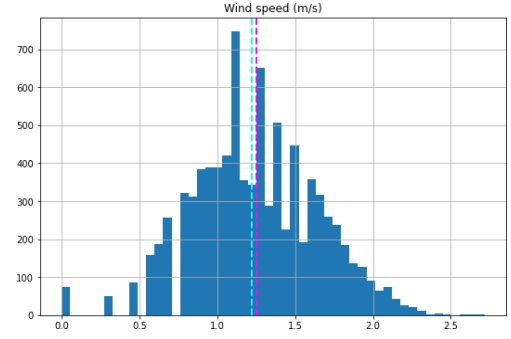


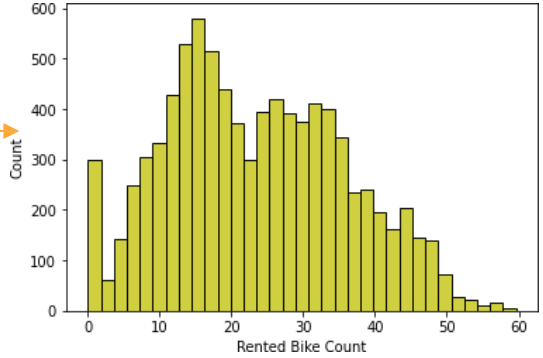


sqrt(wind speed)

sqrt(rented bike count)

Transformed Features





Features transformation

Due to presence of categorical features we cant feed data directly in ML algorithm. We need to transform categorical features that have strings datatype to numerical datatype. For which we have used One-hot encoding and label encoding for categorical features.

Seasons
Summer
Winter
Autumn
Spring

One hot encoding →

Summer	Winter	Autumn	Spring
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Similarly for ‘Holiday’ and ‘Functional Day’ feature.

Applying ML Algorithms

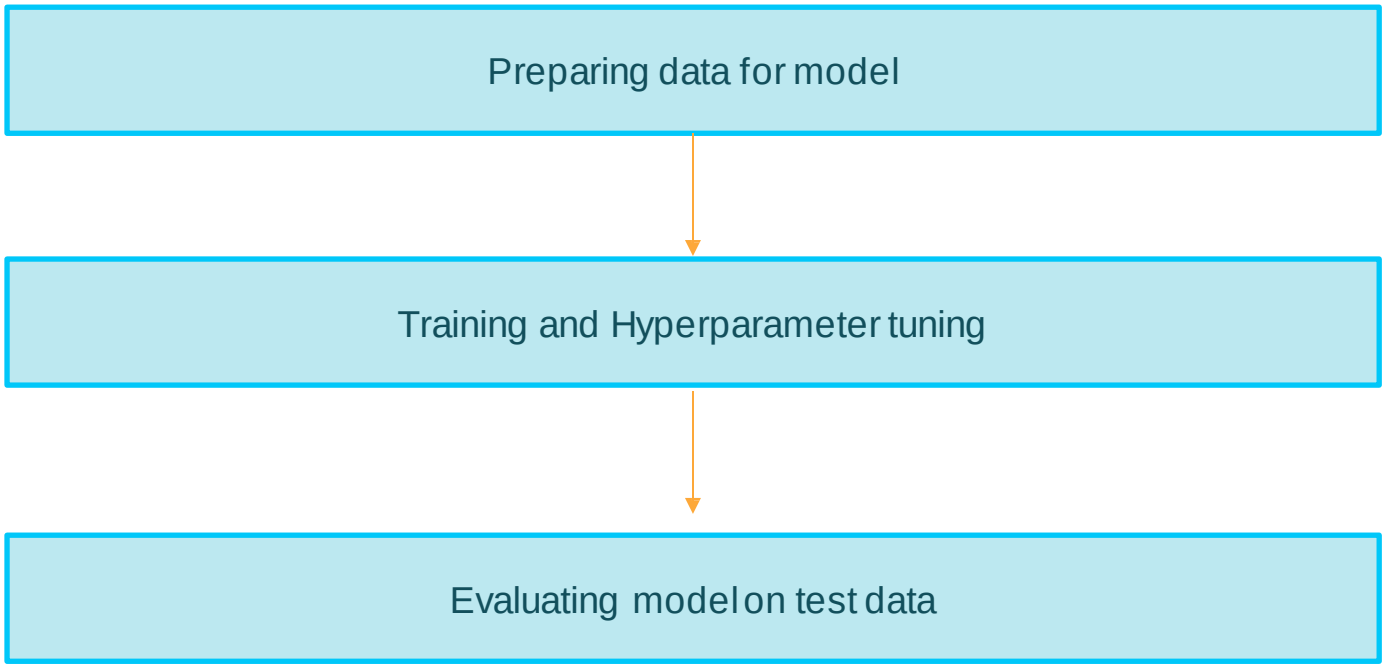
Since we have to predict the count of rented bikes required per hour. Hence, we have to use regression algorithms.

Algorithms that we will use are:

- Decision Tree
- Random Forest
- Linear Regression
- Lasso Regression
- Ridge Regression
- Elastic Net Regression

Applying ML Algorithms

Applying supervised ML algorithms have following steps



Training and hyperparameter tuning

We have trained the models and performed hyperparameter tuning using GridSearchCV and along with that we performed cross validation also.

Following are the optimal parameters of models for which we used GridSearchCV.

Model	Optimal parameters
Decision Tree	{'max_depth': 5, 'max_leaf_nodes': 20, 'min_samples_split': 3}
Random Forest	{'max_depth': 7, 'n_estimators': 80}
Lasso	{'alpha': 0.1}
Ridge	{'alpha': 1}
Elastic Net	{'alpha': 0.001, 'l1_ratio': 0.8}

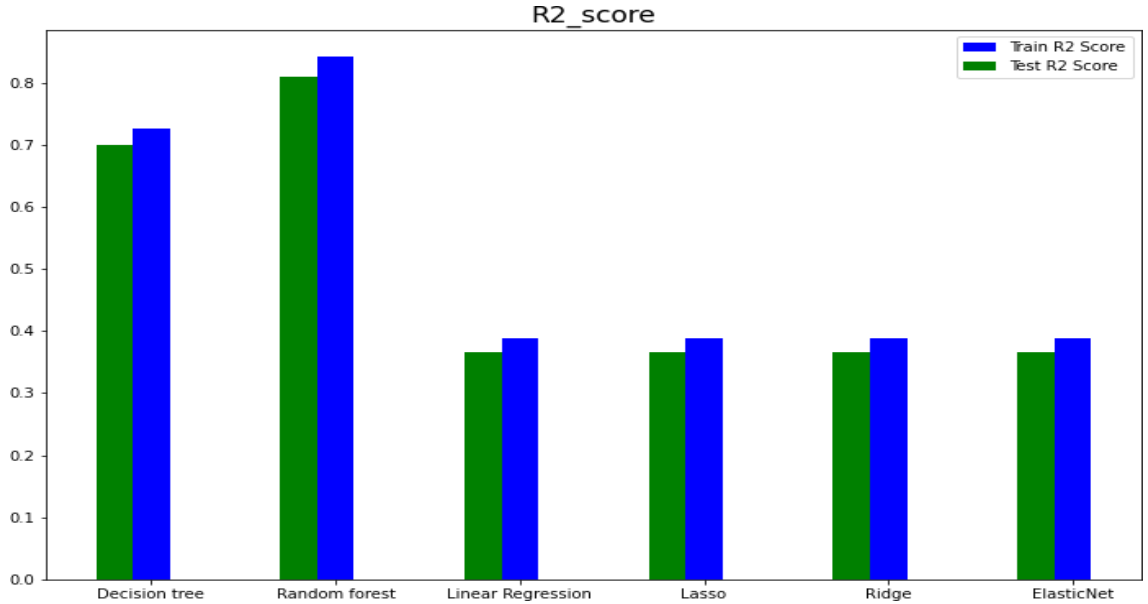
Evaluatingmodels

Model	Train data – MSE	Test data – MSE	Train data – R2-Score	Test data – R2-Score
Decision Tree	113611.96	125508.38	0.727	0.699
Random Forest	64549.49	78132.54	0.845	0.812
Linear Regression	185109.85	193993.06	0.38892	0.3660
Lasso	185110.79	193970.52	0.38813	0.3658
Ridge	185110.49	193971.57	0.38825	0.3661
Elastic Net	185111.11	193962.14	0.38822	0.3659

Comparing different ML Models

We evaluated different models on the basis of R2- Score metrics.

And comparing different models we get that Random Forest worked best in predicting the count of rented bikes as its R2-score is maximum from the tried models.



Conclusion

- In the given dataset there was no strong linear relation between dependent variable ‘Rented Bike Count’ and independent features. That’s why Linear regression model and its other regularization variant models didn’t performed well.

- Out of all models we apply Decision tree and Random forest model are most accurate. Reason for this are no specific relation between features and large data.

- Random Forest performed best as it is an ensemble model and result from multiple decision trees is average out to give the prediction.

Thank you