Capstone Project

Bike Sharing Demand Prediction

By:- Om Prakash Pradhan & Ruchika Nayak

Key Points

- ❑Introduction
- ❑Project Objectives
- ❑Data Summary
- ❑Methodology
- ❑Insights from EDA
- ❑Feature engineering
- ❑Data preparation for modelling
- ❑Model fitting and evaluation
- ❑Observations
- ❑Conclusion
- ❑Challenges

Introduction

- Bike sharing system is a shared transport service in which bicycles are made available for shared use to individuals on a short-term basis for a price or free.
- People use bike-share for various reasons. Some who would otherwise use their own bicycle have concerns about theft or vandalism, parking or storage, and maintenance.
- The typical bike-share has several defining characteristics and features, including station-based bikes and payment systems, membership and pass fees, and per-hour usage fees.

Project Objectives

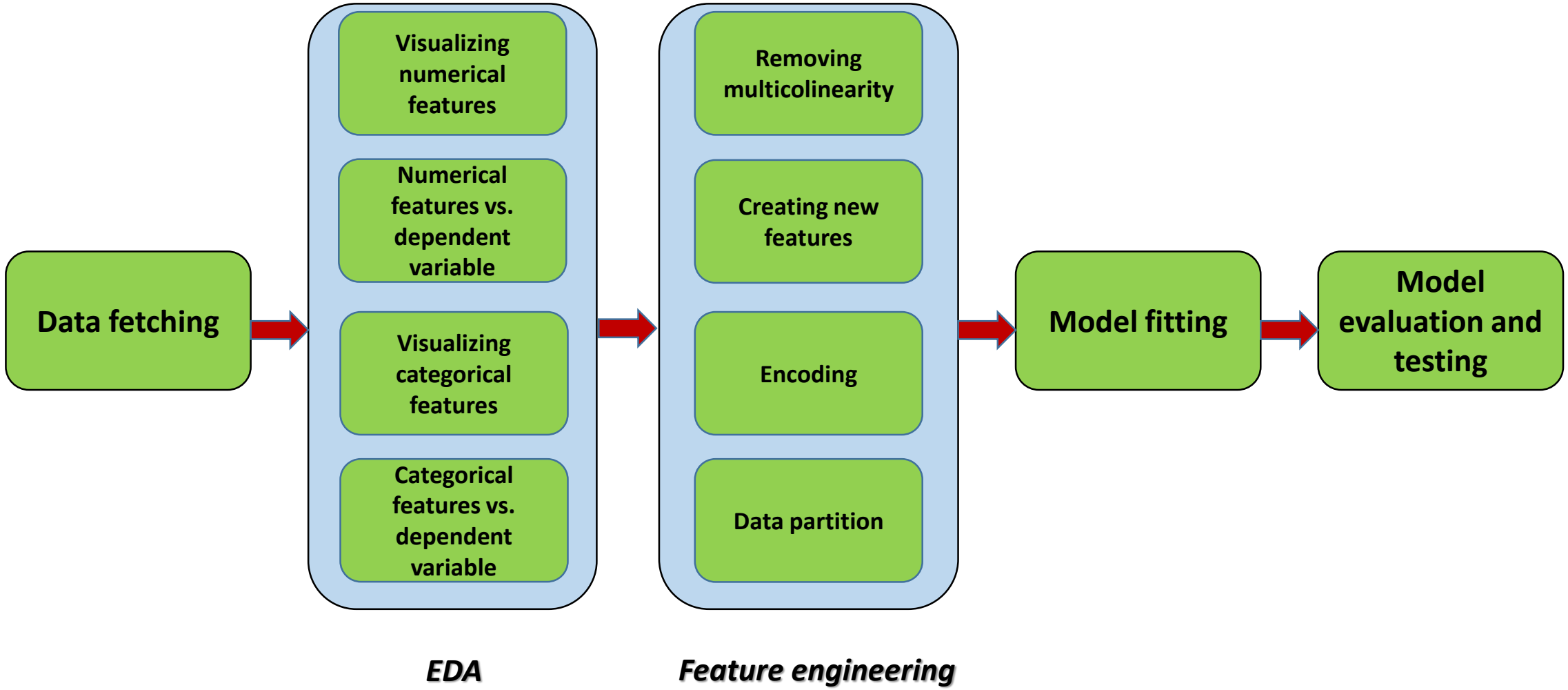
- To highlight the main variables/factors influencing rental bike count.
- To predict bike count required at each hour for the stable supply of rental bikes.
- To compare the various machine learning models and find out best model for the above task.

Data Summary

Index	Date	Rented bike count	Hour	Temperature	Humidity	Wind Speed	Visibility	Dew point Temperature	Solar Radiation	Rainfall	Snowfall	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

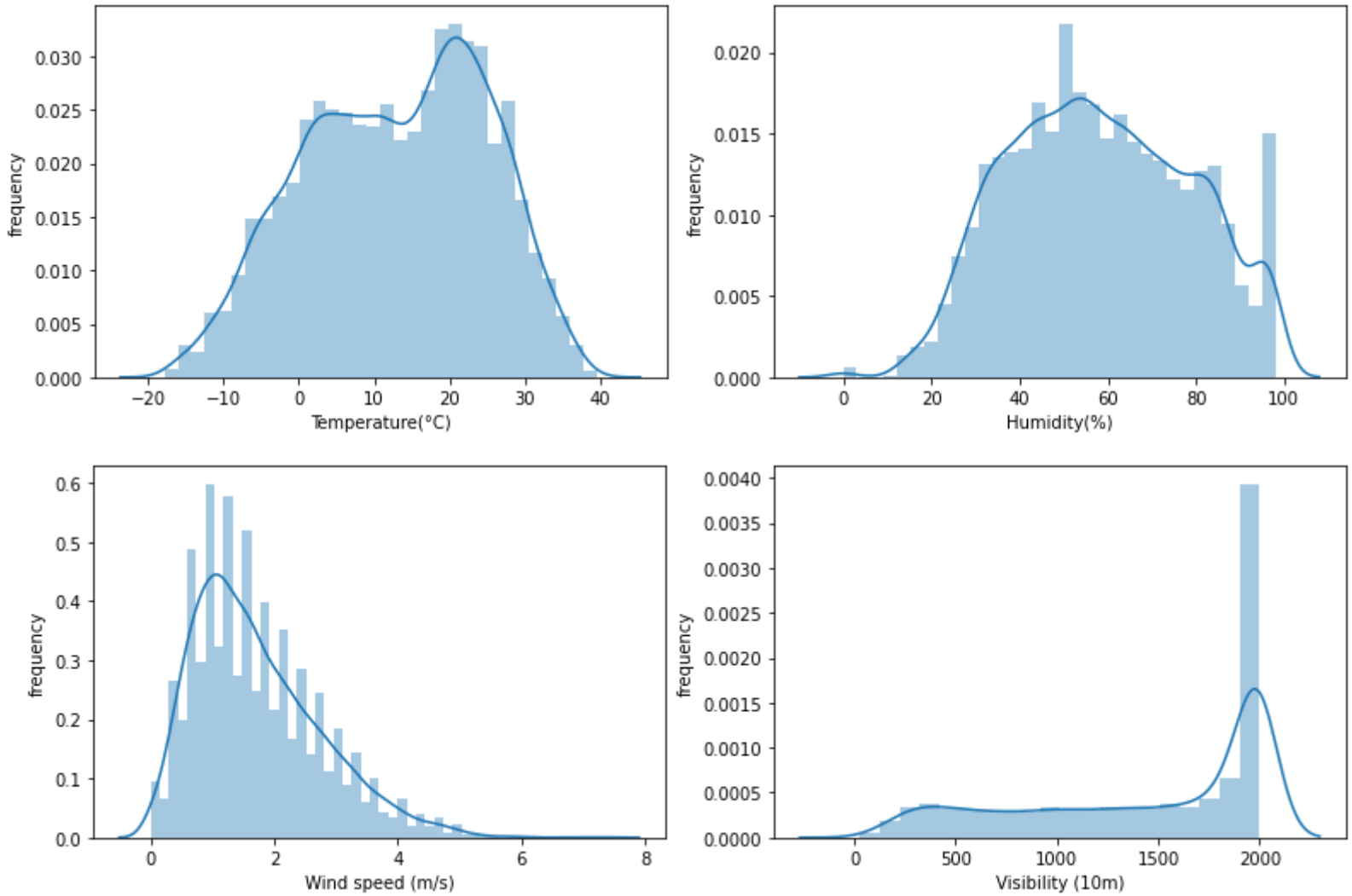
- The dataset contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), seasons, holiday, functioning day, the number of bikes rented per hour, and date information.
- Dataset comprises of total 8760 rows and 14 columns and there are no missing values and duplicate values.
- Out of all the features seasons, holiday and functioning day are categorical in nature.

Methodology



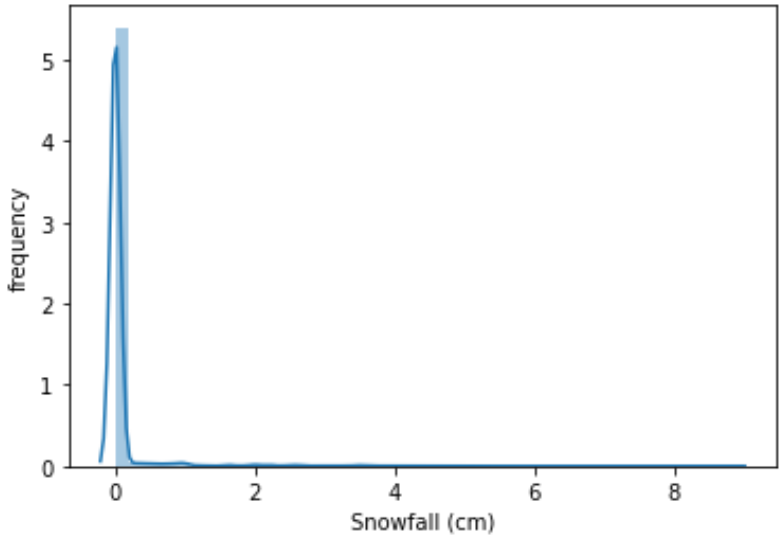
Insights from EDA

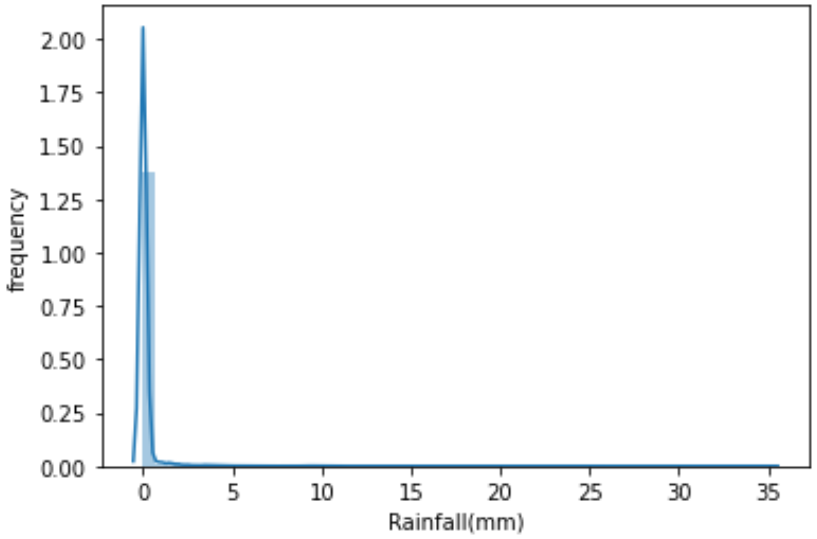
- The distribution of temperature, humidity and dew point temperature are nearly normal.
- The distribution of wind speed is slightly right skewed and visibility column is severely left skewed.

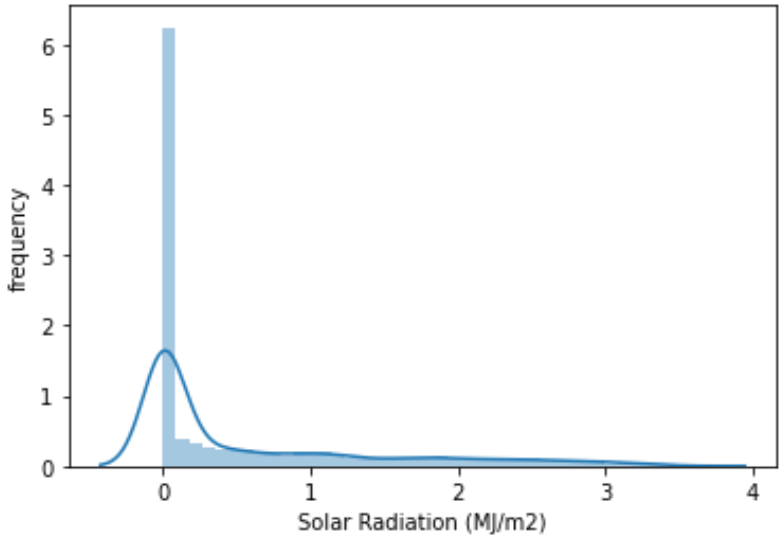


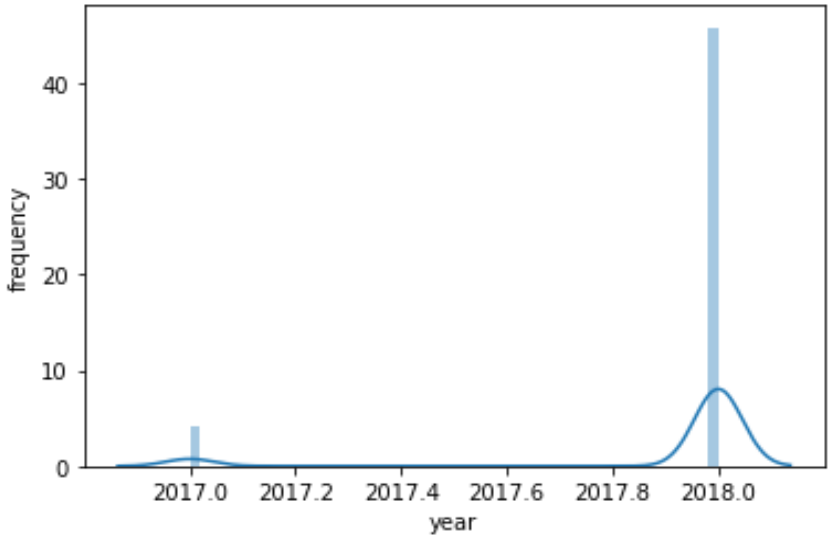
Insights from EDA

- Columns like snowfall, rainfall and solar radiation contains most of the values as zero.
- The dataset contains most of the data for 2018.



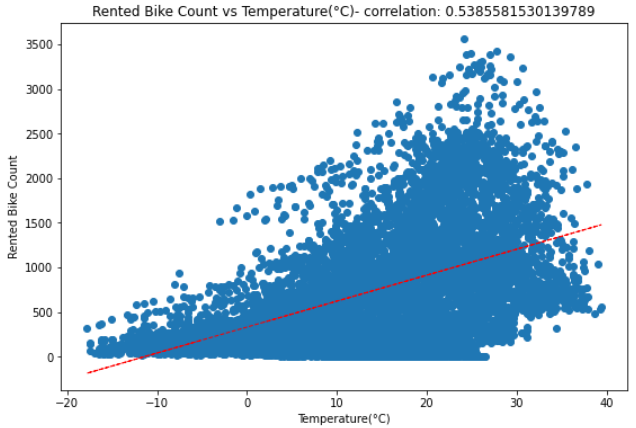


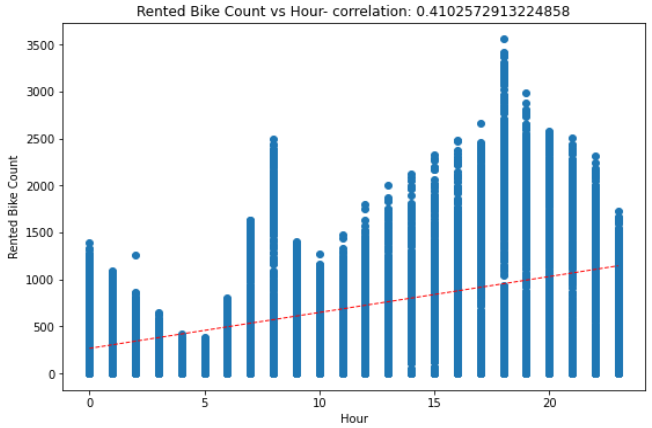


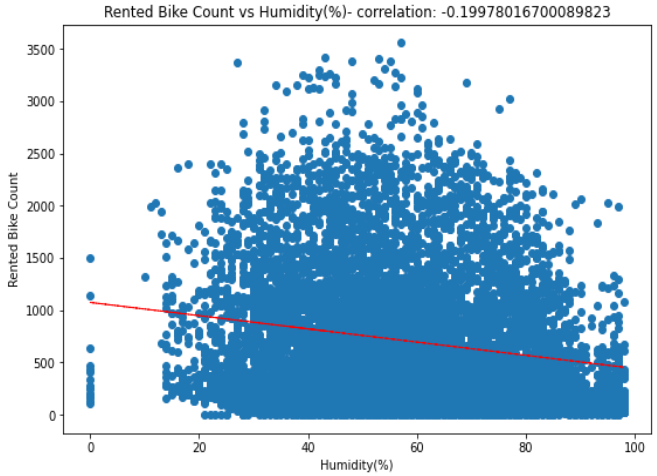


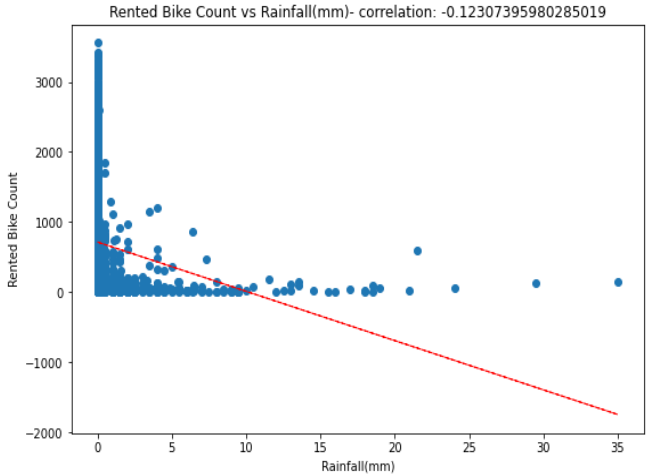
Insights from EDA(contd.)

- Features like hour, temperature and dew point temperature are highly correlated with dependent variable.
- Humidity, snowfall and rainfall are negatively correlated with dependent variable with very low correlation coefficient.



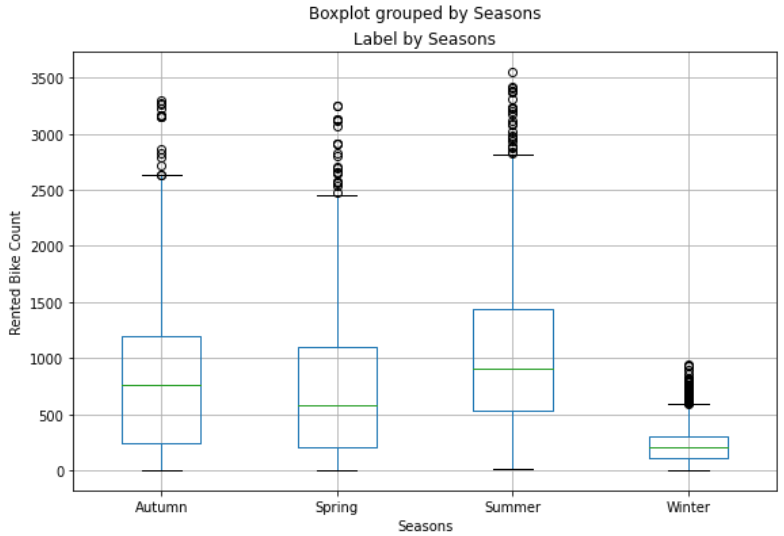


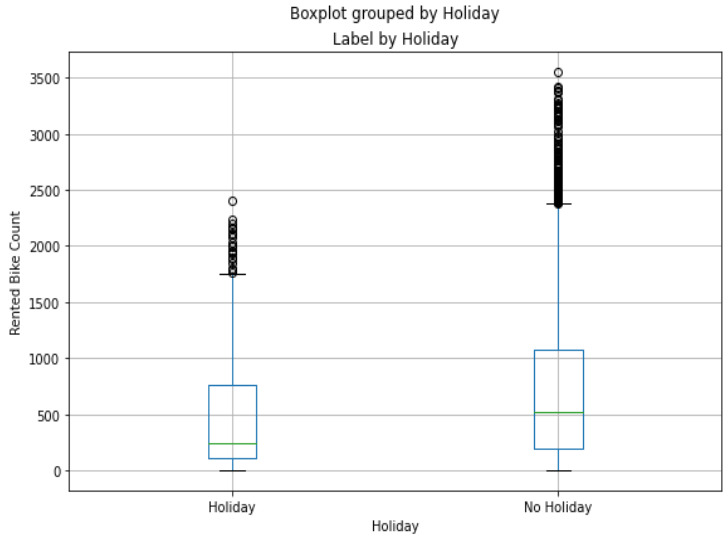


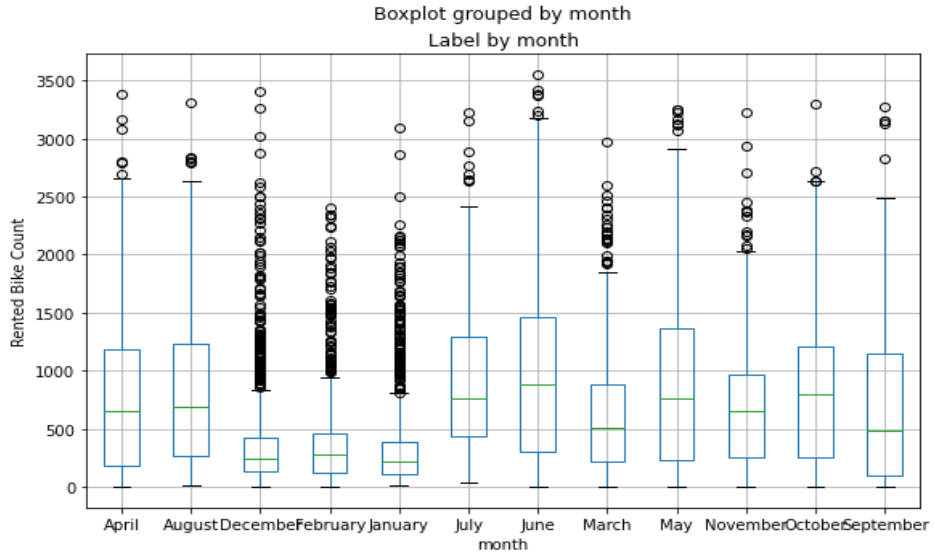


Insights from EDA(contd.)

- Bike renting is less in winter season.
- In holidays bike renting is slightly less.
- There are very few people renting bikes in the month of December, January and February.

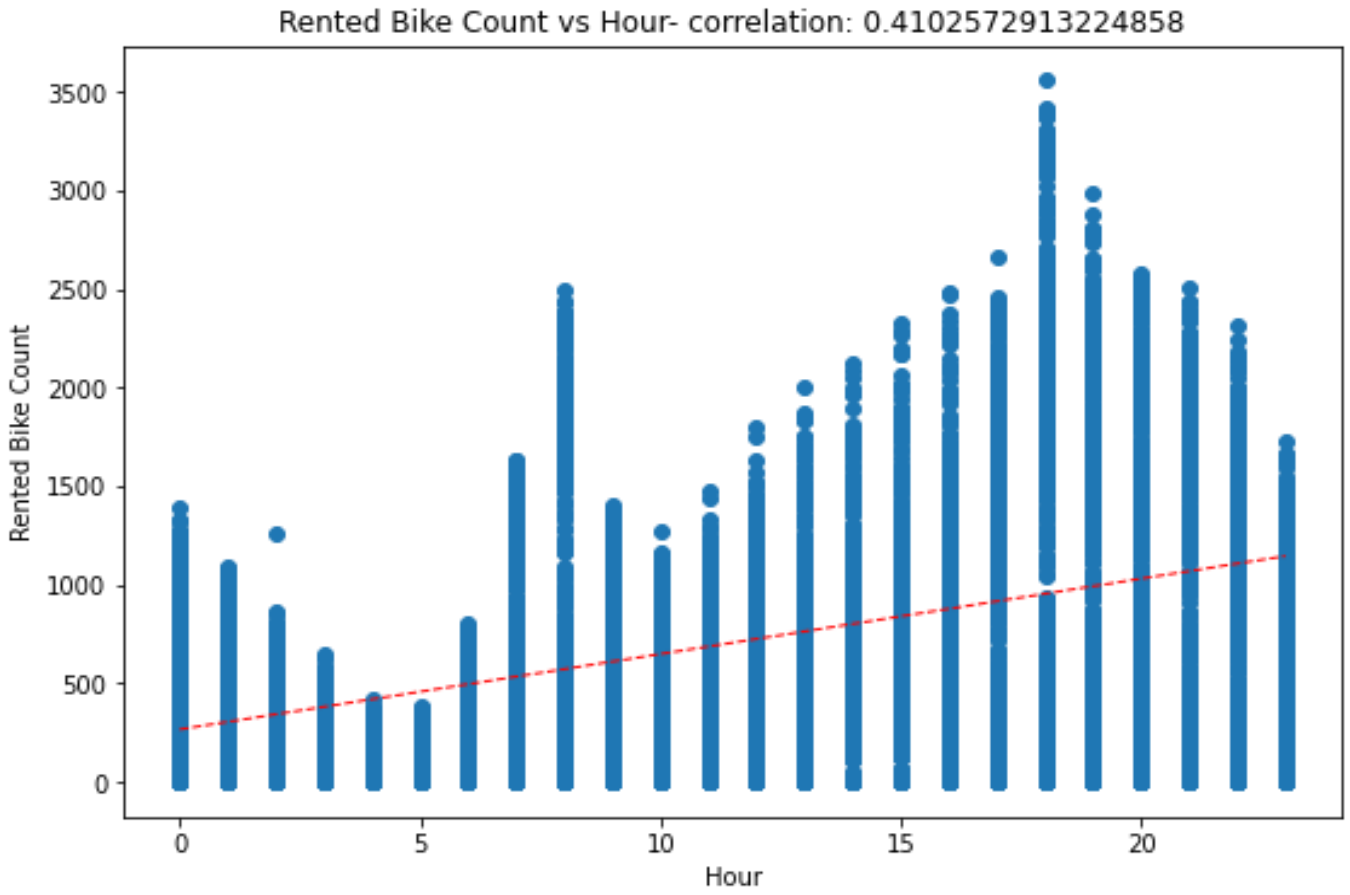






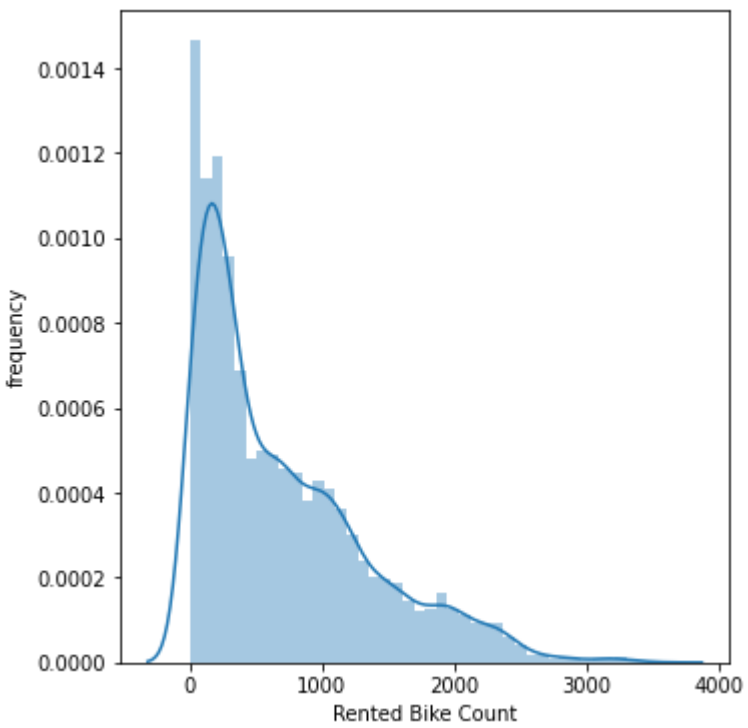
Insights from EDA(contd.)

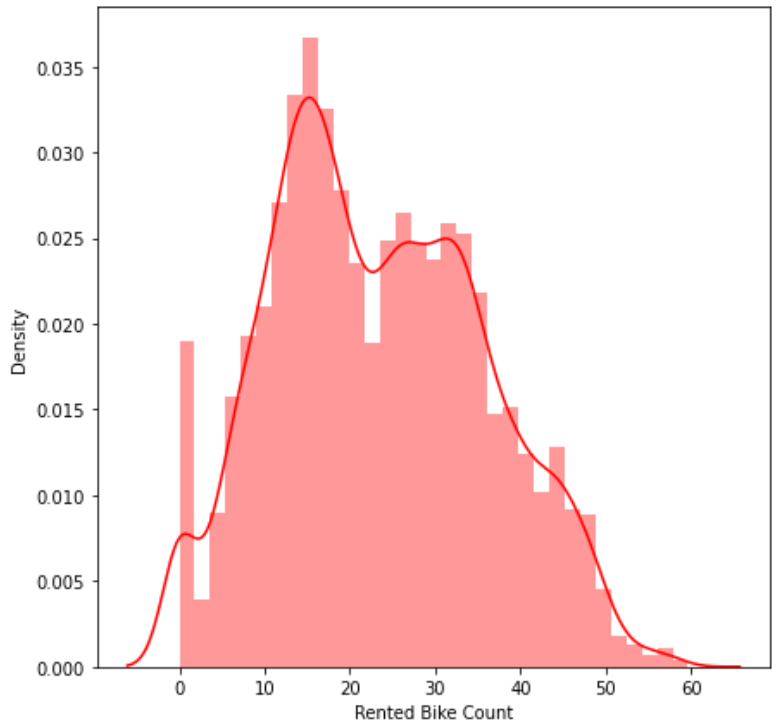
- The peak time for bike renting is during the working hours that is from morning 7am to 10am and from evening 4pm to 9pm.



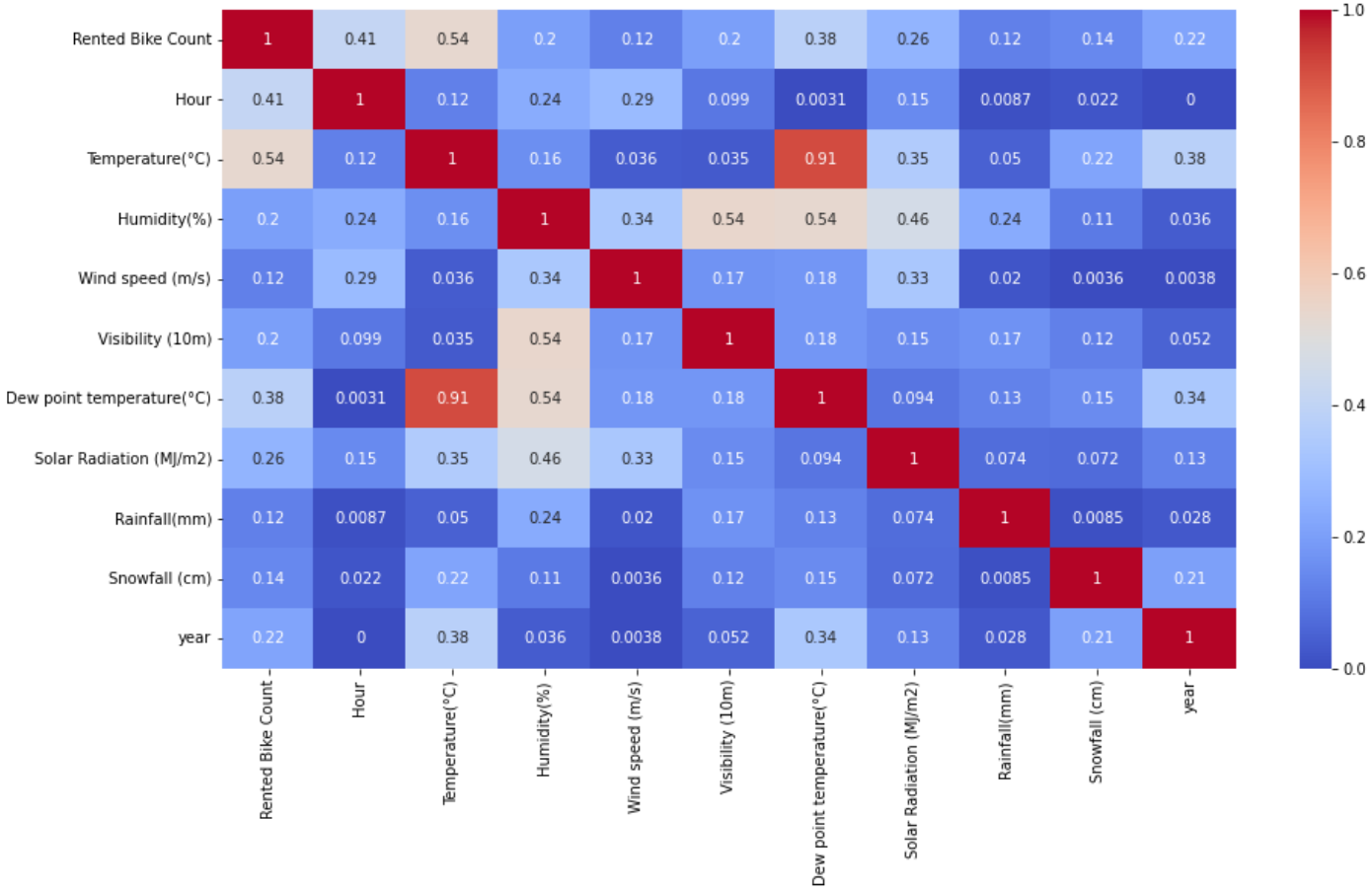
Dependent variable

- The distribution of rented bike count which is the dependent variable is right skewed.
- To make it normal we applied square root transformation on it, which will help to increase model accuracy.





Multicolinearity



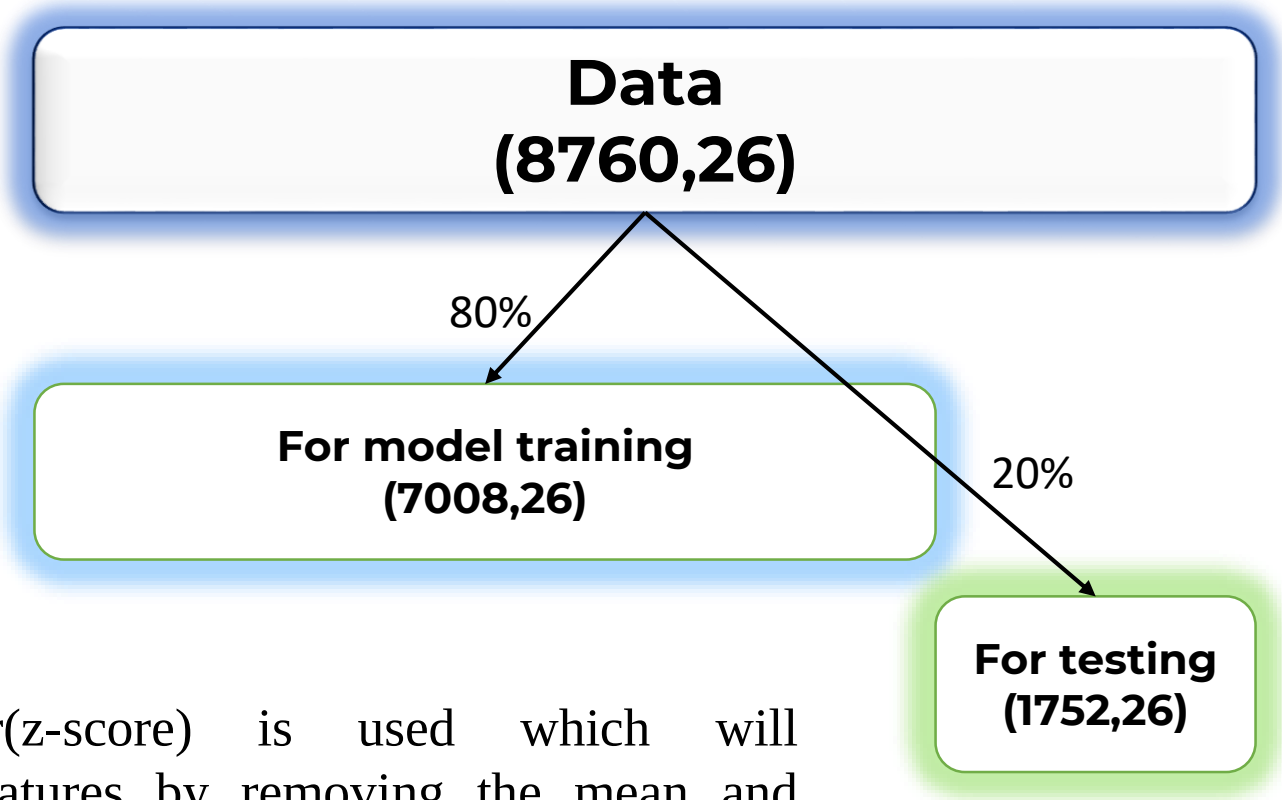
	variables	VIF
0	Hour	4.450547
1	Temperature(°C)	188.496720
2	Humidity(%)	186.877169
3	Wind speed (m/s)	4.811966
4	Visibility (10m)	10.313000
5	Dew point temperature(°C)	126.950456
6	Solar Radiation (MJ/m2)	2.888695
7	Rainfall(mm)	1.103251
8	Snowfall (cm)	1.127819
9	year	397.756221

- Temperature and dew point temperature are highly correlated.
- VIF value of year is too high. After removing year, VIF value of all the features are falling within the limit. So we have removed year while building the model for better accuracy.

Feature engineering

- We have extracted two different features month and year from date column.
- We have applied label encoding for holiday and functioning day features as it comprises of only two distinct labels.
- For seasons and month, we have applied one hot encoding as it contains more than two labels.
- As in snowfall and rainfall columns most of the values are zero and also correlation of them with the dependent variable is very less, we have not included them in our model building.
- We have applied square root transformation for dependent variable.

Data preparation for modelling

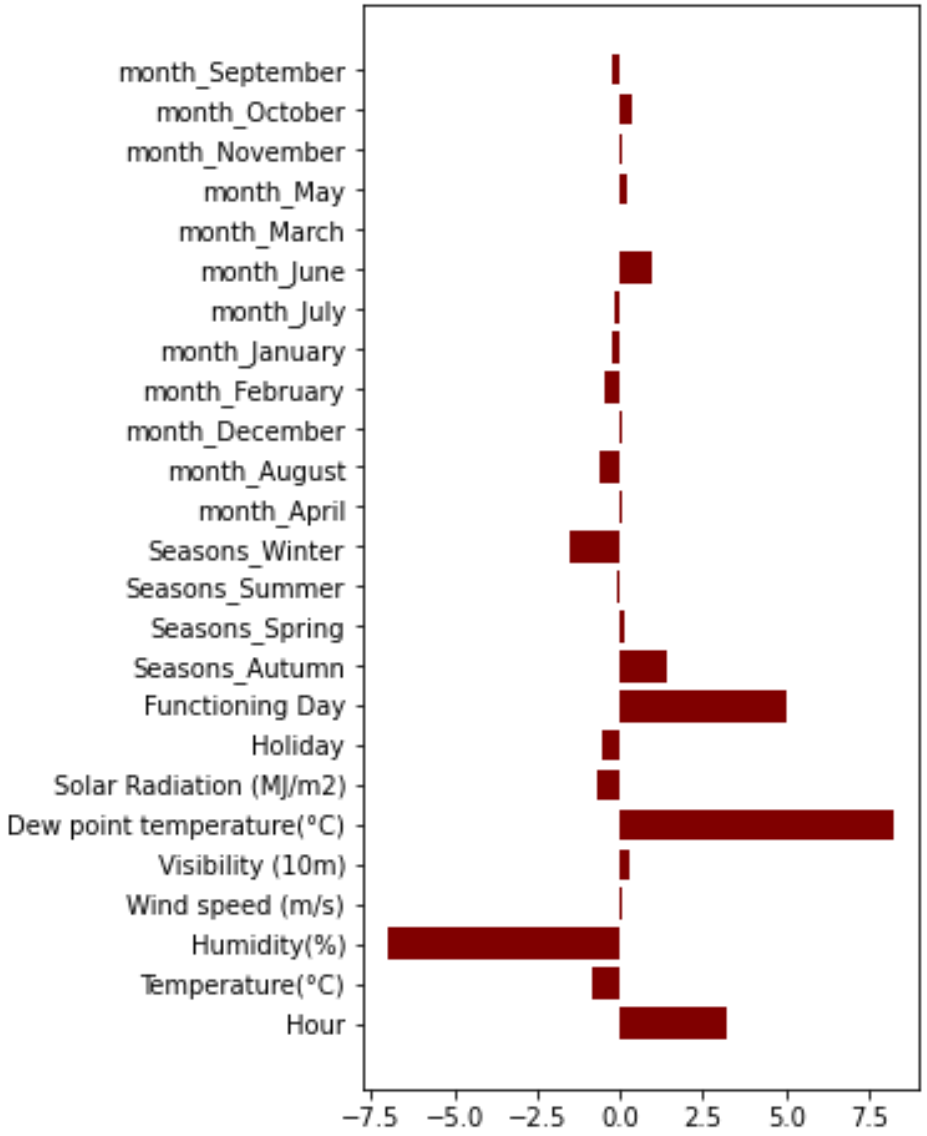


- StandardScalar(z-score) is used which will standardize features by removing the mean and scaling to unit variance.
- We have also used square root transformation for dependent variable.

Linear Regression

Linear Regression	MSE	RMSE	R2	Adjusted R2
Train	174034	417.1	0.58	0.58
Test	174523	417.7	0.58	0.57

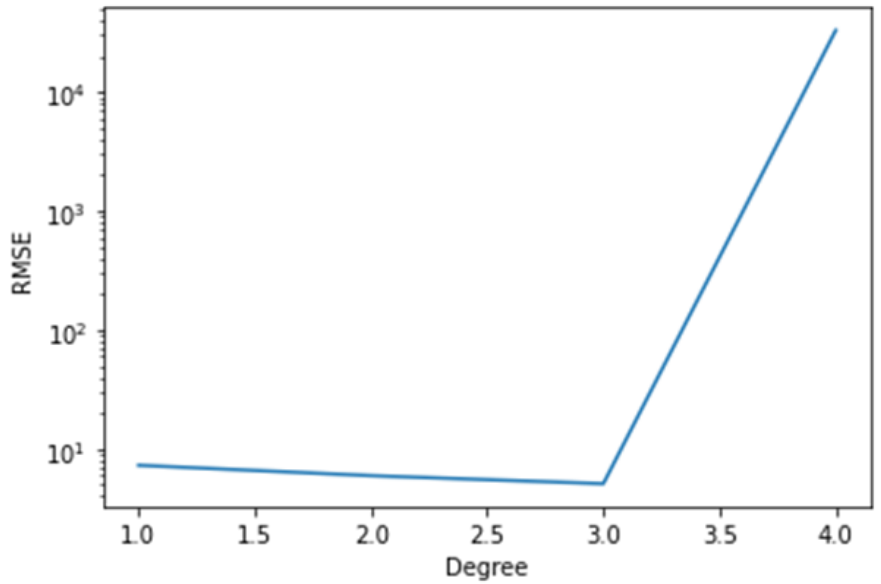
- Humidity, hour, dew point temperature and functioning day are most important features for linear regression, out of which humidity negatively affecting rented bike count.
- Accuracy of the model is very low and variance is also low.



Polynomial Regression

Polynomial Regression	MSE	RMSE	R2	Adjusted R2
Train	58983	242.86	0.85	0.80
Test	83783	289.45	0.79	0.78

- Optimal degree for polynomial regression is 3.
- Here the model is slightly overfitting the data.



Regularization

LASSO

	MSE	RMSE	R2	Adjusted R2
Train	67723	260.23	0.83	0.83
Test	77199	277.84	0.81	0.81

RIDGE

	MSE	RMSE	R2	Adjusted R2
Train	69940	264.46	0.83	0.83
Test	82636	287.46	0.80	0.79

ELASTIC NET

	MSE	RMSE	R2	Adjusted R2
Train	68512	261.74	0.83	0.83
Test	78342	279.89	0.81	0.80

After performing different regularization techniques, variance of the model is successfully reduced.

Decision tree regressor

- After performing GridsearchCV on decision tree regressor, the best parameters are as follows:

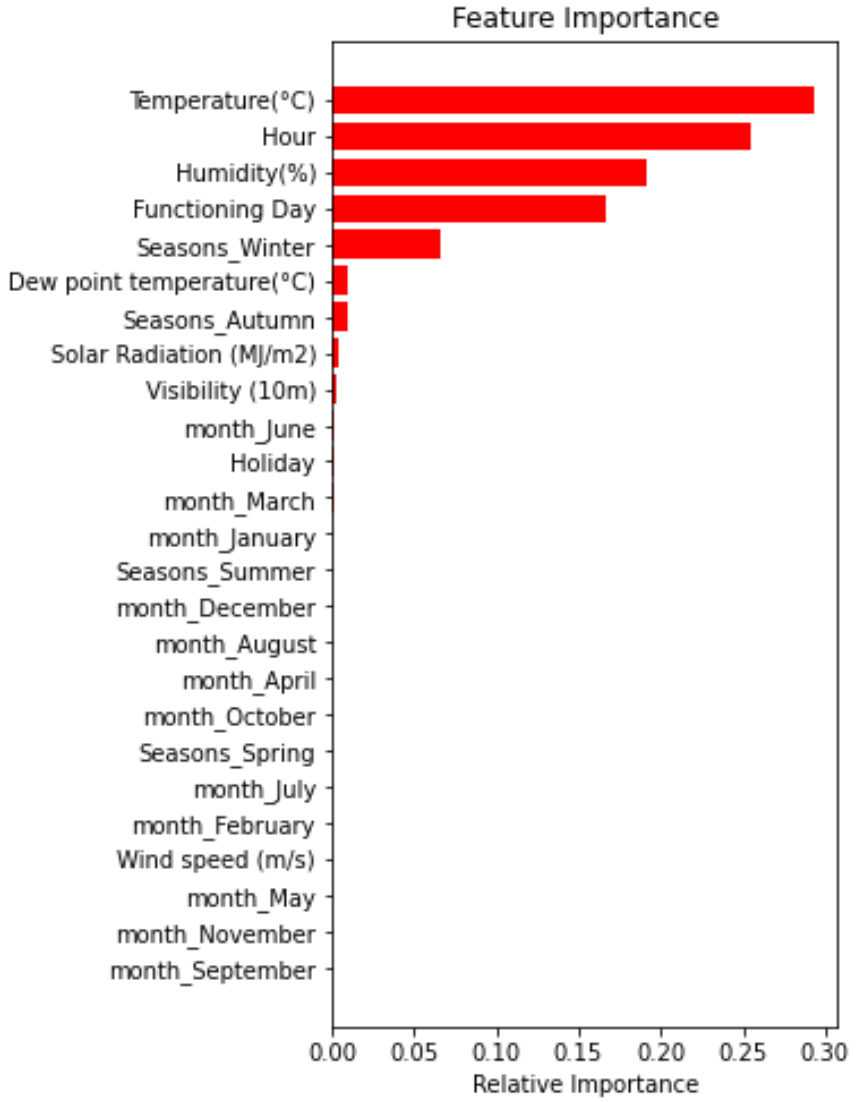
criterion	max_depth	max_leaf_nodes	min_samples_leaf	min_samples_split
MSE	8	100	20	10

- The performance matrix of the model with these parameters are as follows:

Decision tree	MSE	RMSE	R2	Adjusted R2
Train	78837	280.78	0.81	0.80
Test	89792	299.65	0.78	0.78

Decision tree regressor(contd.)

- Top 3 most important features are Temperature, Hour and Humidity.
- Surprisingly not a single month is given importance while fitting this model.



Ensemble models

Random forest

	MSE	RMSE	R2	Adjusted R2
Train	7874	88.73	0.98	0.98
Test	58255	241.36	0.85	0.85

Gradient boosting

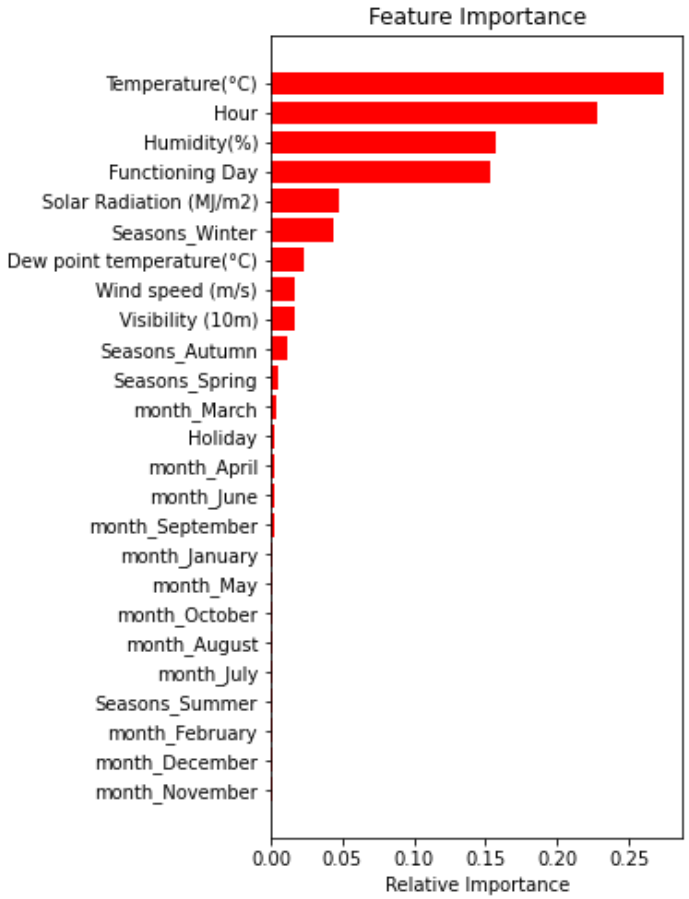
	MSE	RMSE	R2	Adjusted R2
Train	5888	76.73	0.98	0.98
Test	48235	219.62	0.88	0.88

XGBoost

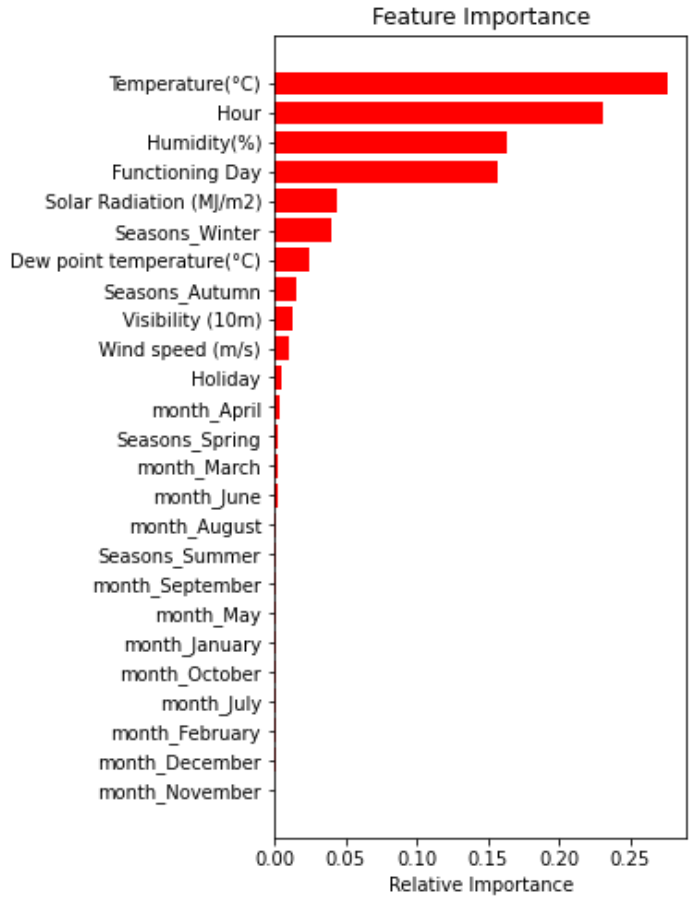
	MSE	RMSE	R2	Adjusted R2
Train	7319	85.55	0.98	0.98
Test	49867	223.31	0.88	0.87

- This is the result of three different ensemble models.
- All the three models are overfitting the data.

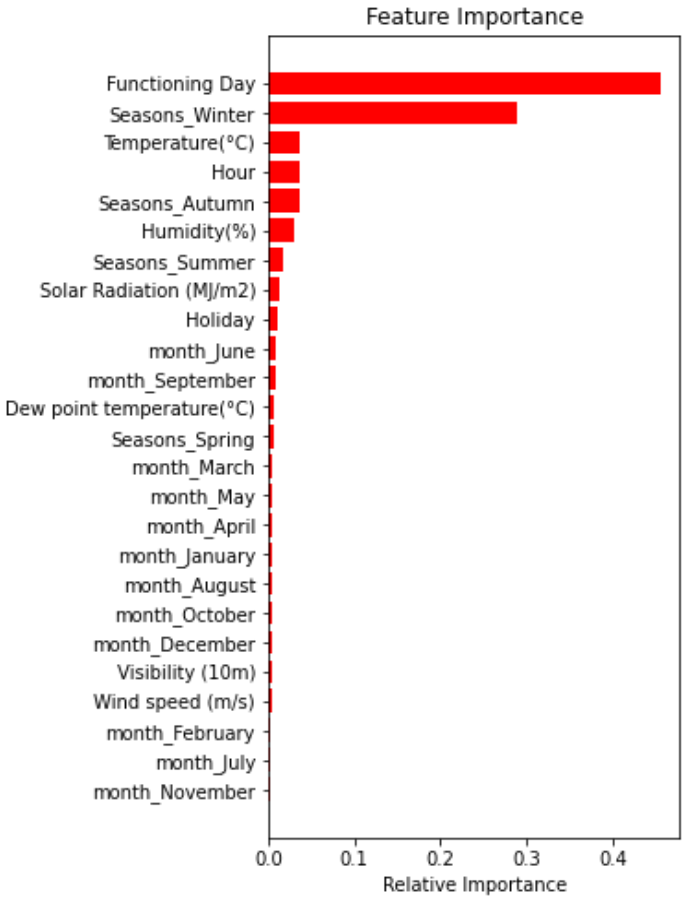
Ensemble models(Contd.)



Random forest



Gradient boosting



XGBoost

Observations

- With accuracy of 58% linear regression model has high bias.
- Non linear regression model(degree 3) with regularization gives good accuracy score of 83% on training data and 81% on testing data.
- Decision tree model gives accuracy score of 81% on training data whereas 78% accuracy on testing data.
- All the 3 different ensemble methods are overfitting the data with almost 10% difference between train and test accuracy.
- Out of all ensemble methods used gradient boosting and XGBoost giving slightly good result with training accuracy of 98% and testing accuracy of 88%.

Conclusion

- Temperature, hour and humidity are the three most important features given by all the models except XGBoost.
- Three degree polynomial regression model with regularization technique is giving very decent accuracy of 81% without overfitting the model.
- Also Gradient boosting is giving good accuracy of 88% on test data by slightly overfitting the data. The optimal parameters for gradient boosting are given by GridSearchCV are learning_rate = 0.03, max_depth = 8, n_estimators = 500, subsample = 0.5.
- XGBoost model certainly not improving the model performance beyond Gradient boosting.

Challenges

- Selecting most relevant features.
- Selecting relevant set of hyper parameters for tuning.
- Computation time during GridSearchCV.


Thank
you