Bike Renting

By

C Sai Amogh

Content:-

1 Introduction

1.1 Business problem

1.2 Metric for business problem

1.2.1 RMSE

1.2.2 R square

1.3 Data

2 Methodology

2.1 Statistical analysis and EDA

2.1.1 Dimensions and Structure of the data

2.1.2 Missing value analysis

2.1.3 Descriptive statistics

2.1.4 Outlier analysis(Box plot) and Treating the extreme outliers

2.1.5 Box plot after treating the extreme outliers

2.1.6 Density plots on continuous variables

2.1.7 Bar/Count plots on categorical variables

2.1.8 Bivariate analysis (EDA)

2.1.8.1 Time series plots(line plots)

2.1.8.2 Bar plot Scatter plots

2.1.8.3 Violin plot

2.1.8.4 Scatter plots and Multivariate Data analysis

2.1.9 Correlation analysis

3 Modeling

3.1 Machine learning Models

3.1.0 Train and Test Data

3.1.1 Linear Regression

3.1.2 Linear SVM

3.1.3 Decision Trees

3.1.4 Adaboost

3.1.5 Random forest

3.1.6 Extra Trees Regressor

3.1.7 XGBOOST

4 Results and Conclusion

5 Annexure code (PDF)

1 INTRODUCTION

1.1 Business Problem:- The objective is to Predict the bike rental count on daily based on the environmental and seasonal settings.

1.2 Metric for business problem

1.2.1 RMSR: - Measures the average of the squares of the errors or deviations—that is, the difference between the estimator and what is estimated and take the square root.

1.2.2 R squared: - Is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination.

1.3 Data

The details of data attributes in the dataset are as follows -

- instant: Record index
- dteday: Date
- season: Season (1:springer, 2:summer, 3:fall, 4:winter)
- yr: Year (0: 2011, 1:2012) mnth: Month (1 to 12)
- holiday: weather day is holiday or not (extracted fromHoliday Schedule)
- weekday: Day of the week
- workingday: If day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit: (extracted fromFreemeteo)
- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered
- Clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp: Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min),
- t_min=-8, t_max=+39 (only in hourly scale)
- atemp: Normalized feeling temperature in Celsius. The values are derived via(t-t_min)/(t_maxt_
- min), t_min=-16, t_max=+50 (only in hourly scale)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

2 METHODOLOGY

2.1 Statistical analysis and EDA

2.1.1 Dimensions and Structure of the data:-

There are around 731 rows and 16 variables, I renamed the following columns for easy identification

'yr':'year','mnth':'month','cnt':'count','dteday':'dateday'

2.1.2 Missing value analysis:-

There are no missing values in the given dataset

2.1.3 Descriptive analysis on continuous variables:-

The following where given to me in the problem statement:-

- temperature(temp): t_min=-8, t_max=+39 (only in hourly scale) Celsium before normalization
- atemp :- t_min=-16, t_max=+50 (only in hourly scale) Celsium before normalization
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)

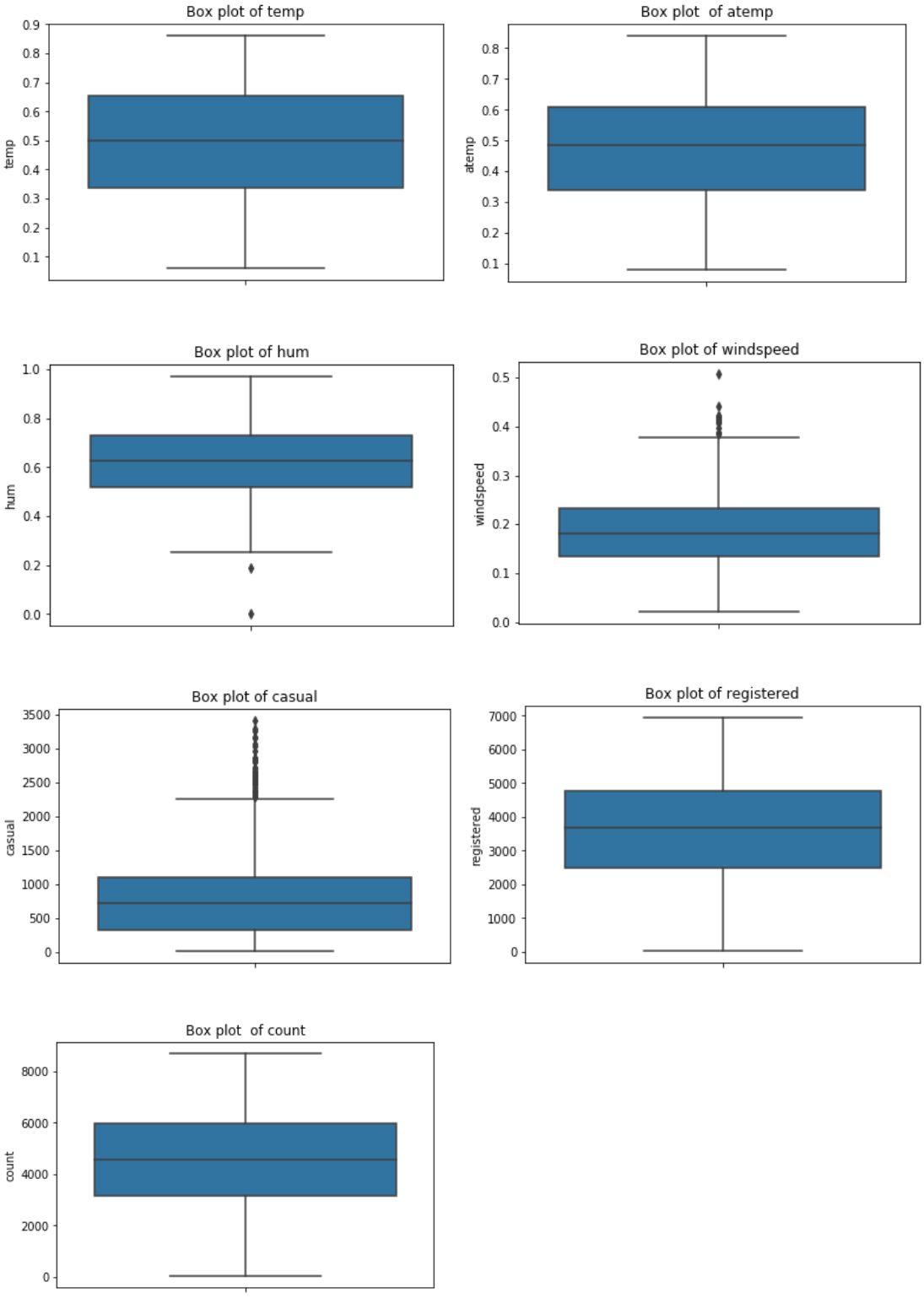
The main findings are as followers:-

	temp	atemp	hum	windspeed	Casual	registered	count
mean	0.495	0.474	0.627	0.190	848	3656	4504
std	0.183	0.162	0.142	0.077	686.6	1560.2	1937.2
min	0.059	0.079	0	0.022	2	20	22
25%	0.337	0.337	0.520	0.134	315	2497	3152
50%	0.498	0.486	0.626	0.180	713	3662	4548
75%	0.655	0.608	0.730	0.233	1096	4776	5956
max	0.861	0.840	0.972	0.507	3410	6946	8714

- Distribution of count is ranging from min (22) to max (8714)
- There might be outliers in humidity at (min to 25%) ,wind speed at (75% to max) ,casual at (75% to max)

2.1.4 Outlier analysis(Box plot) and Treating the extreme outliers

Box plots



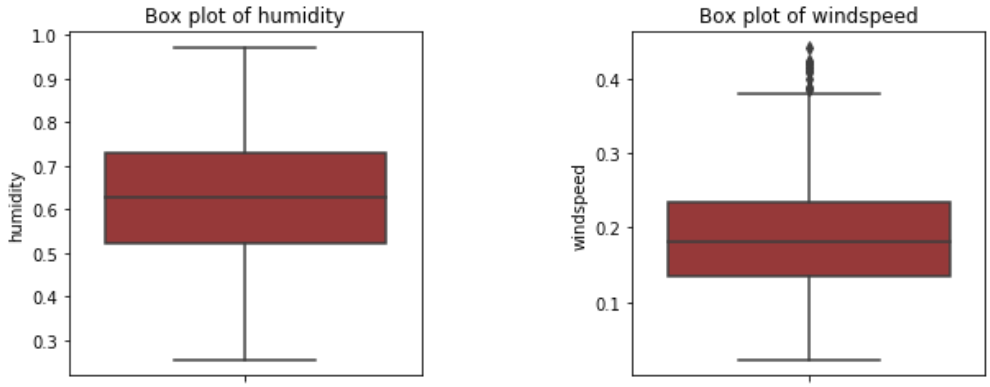
The main findings are as followers:-

- There are outliers in humidity, wind speed

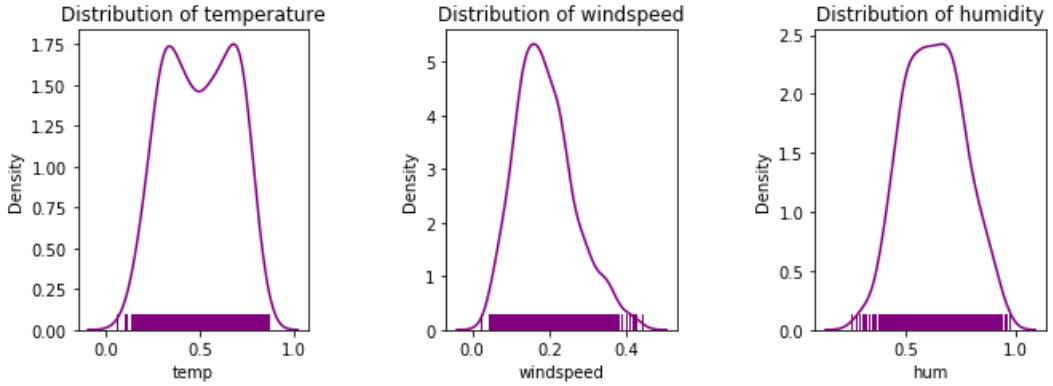
Treating the extreme outliers:-

- Humidity can’t be zero, so replaced the zero value based on mean value of 5 previous and 5 following days.
- One row values of humidity and wind speed have been inter changed as they don’t match the pattern.

2.1.5 Box plot after treating the extreme outliers:-



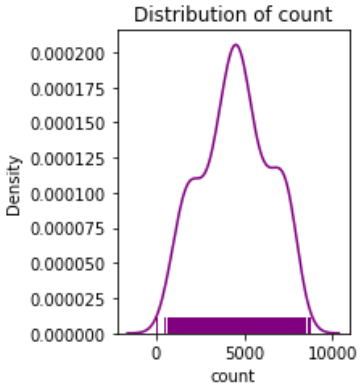
2.1.6 Density plots:-



If we compare the descriptive statistics and above density plots

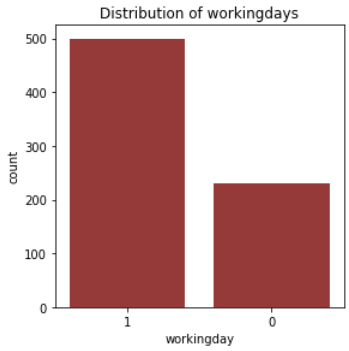
The main findings are as followers:-

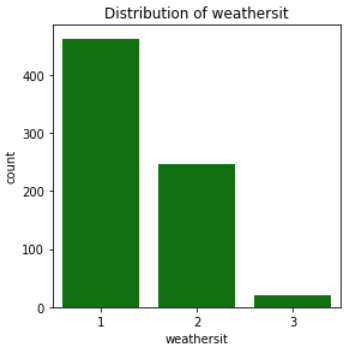
- Temp :- Distribution of temperature is lying between min is 0.059130 and max 0.861667 and mean 0.49
- humidity :- almost normally distributed
- wind speed :- slightly right skewed

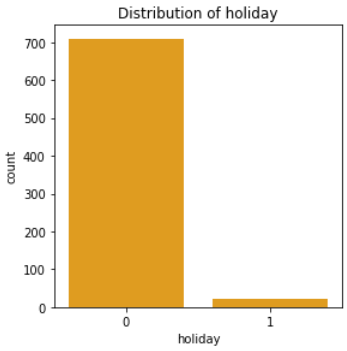


- We can see the Variation of dependent variable with a mean of 4504

2.1.7 Count/Bar plots on categorical variables







The main findings as followers:-

From working day count plot

- There are around 231 weekends or holidays (0) and rest are working days(1)

From count plot of holidays

- We have 21 days as holidays for two years (taken from holiday schedule)

From weather sit count plot

- Most of the days are 1 followed by 2,3

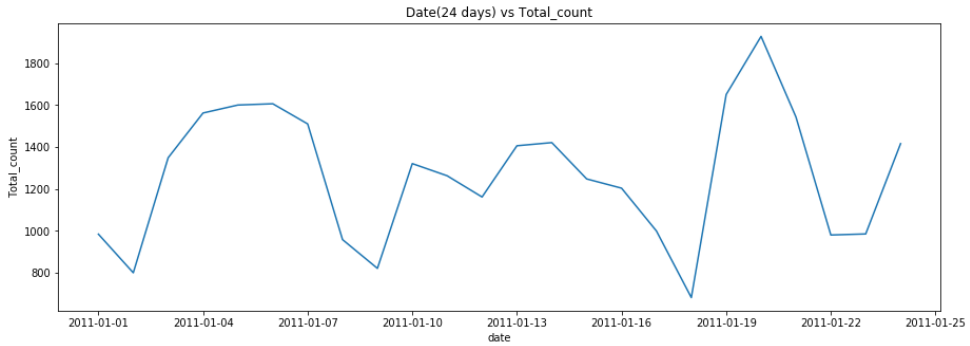
Information Given to us:-

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

2.1.8 Bivariate analysis (EDA):-

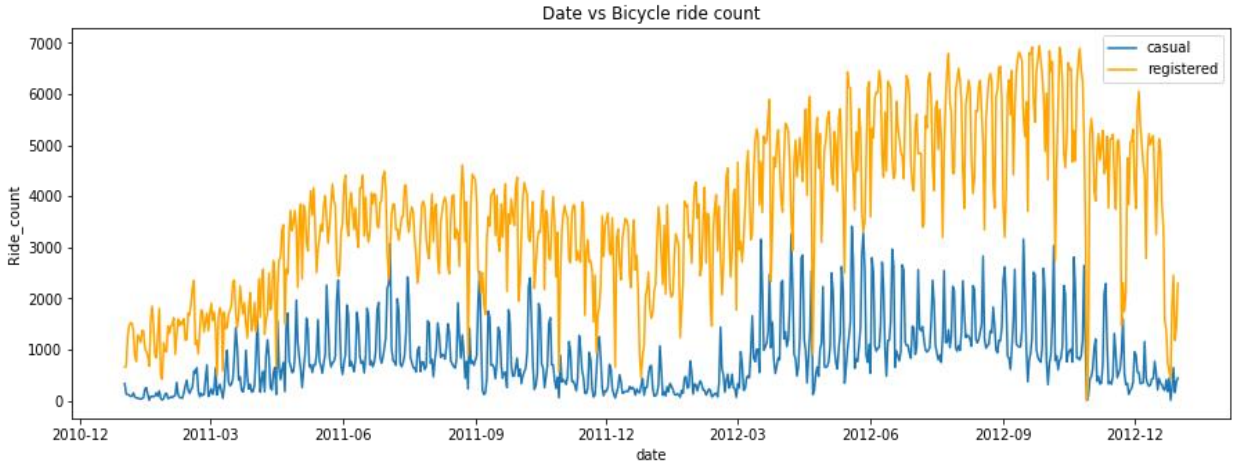
2.1.8.1 Time series plots or line plots

Date day (24 days) vs. Count



- From the above plot we can see the effect of weekends, the bicycle rides in weekends is less compared to weekdays

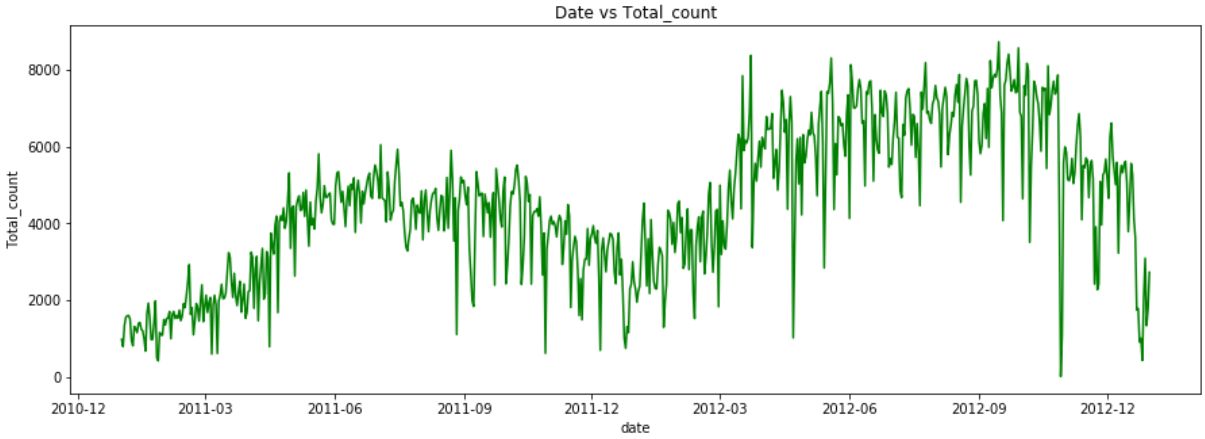
Date vs. Bicycle ride count of casual and registered



The main findings for the above plot are as below:-

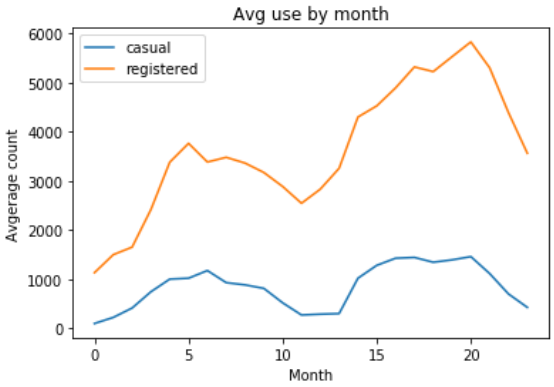
- we can see an upward trend in 2012 compared to 2011, That means there is an increase in rides in 2012 compared to 2011
- For casual users, rides taken are less between December to January compared to other months.
- Even for registered users, rides taken are less during December this may be due to holidays.

Date vs. count



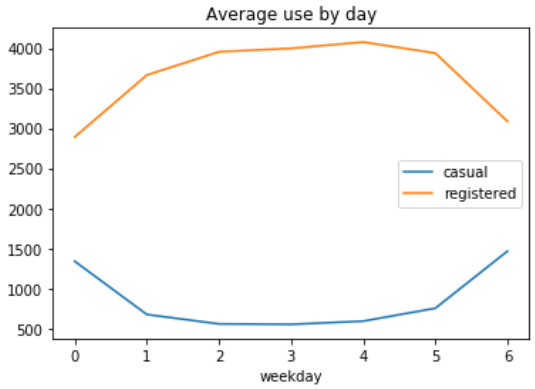
- (trend increased) in 2012 compared to 2011 and dip in months of December, January.

Average use by month:-



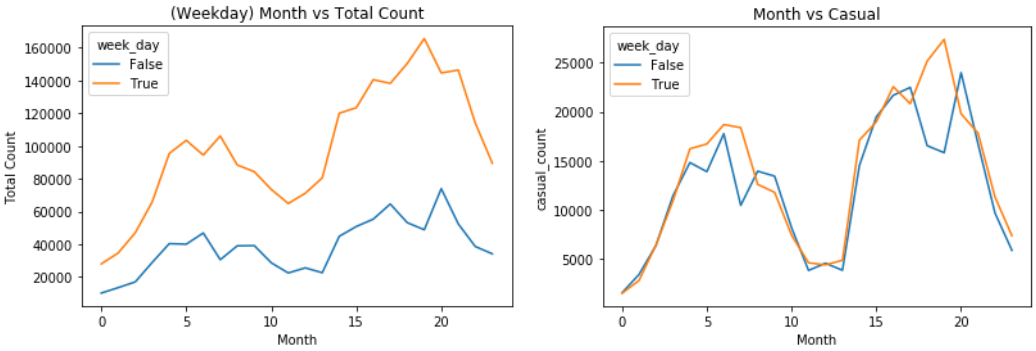
Above we can see the average use by month by casual and registered, we can see the average increase in trend.

Average use by day:-



From the above plot it's clear that registered users tend to ride more on weekdays (1-5) then week ends further casual riders, ride bicycle little more on weekend then weekdays and weekday will an important variable for prediction.

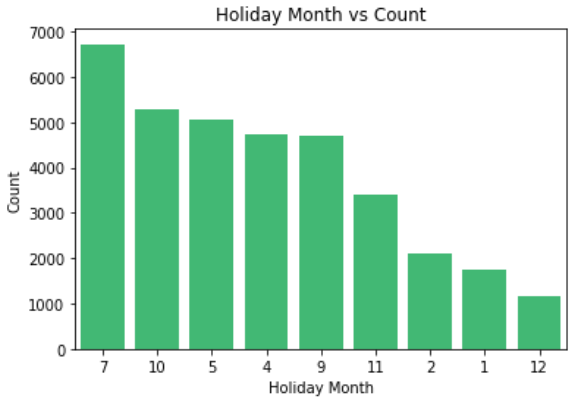
Month vs. total count by weekday and Month vs. casual by weekday



From the above plot we can see the count on weekday, weekend by month on total count and see the effect of weekday(1-5) and weekend(2 days) on casual riders, we see there are almost equal number of casual riders on weekend which of 2 days compared to weekday which of 5 days.

2.1.8.2 Bar plot

Holiday month vs. Count of riders:-



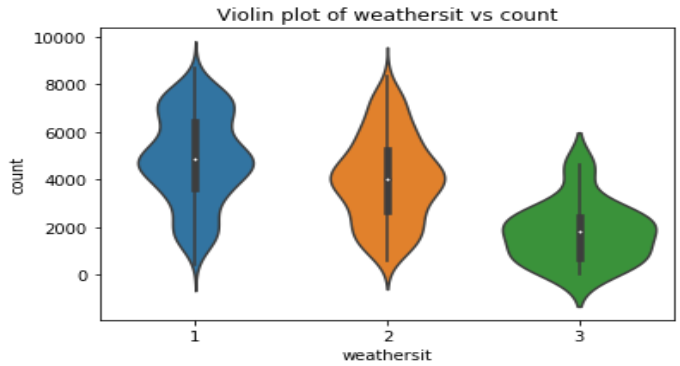
The main findings from the above bar chat is as follows:-

- 12 th month has the lowers riders on holidays, followed by January
- The highest rides on holidays is in the month on July

So variable holiday will be help full in prediction

2.1.8.3 Violin plots:-

Violin plot of Weather sit vs. Count

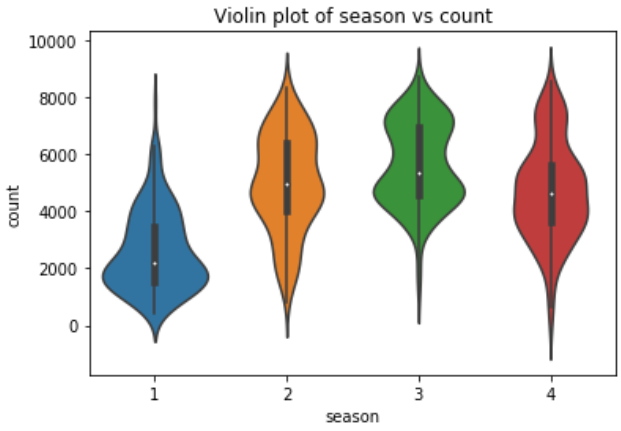


From the plot above we find the following:-

- We can see the distribution of data for each sub variable of weathersit
- Box plot shows the variation of median of the sub categories, 3 sub category has the lowest median, followed by 2nd sub category
- The 3 sub categories are skewed
- Total count is less for 3rd sub category of weather sit but dense around median.

The weather sit variable will be help full in prediction.

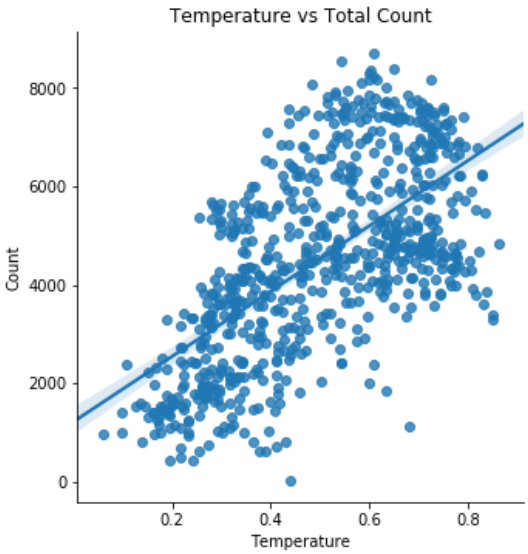
Violin plot of Season vs. Count



- Median and count is lowest for season one, season 3 has the highest median and more rides too.

2.1.8.4 Scatter plots and Multivariate Data analysis :-

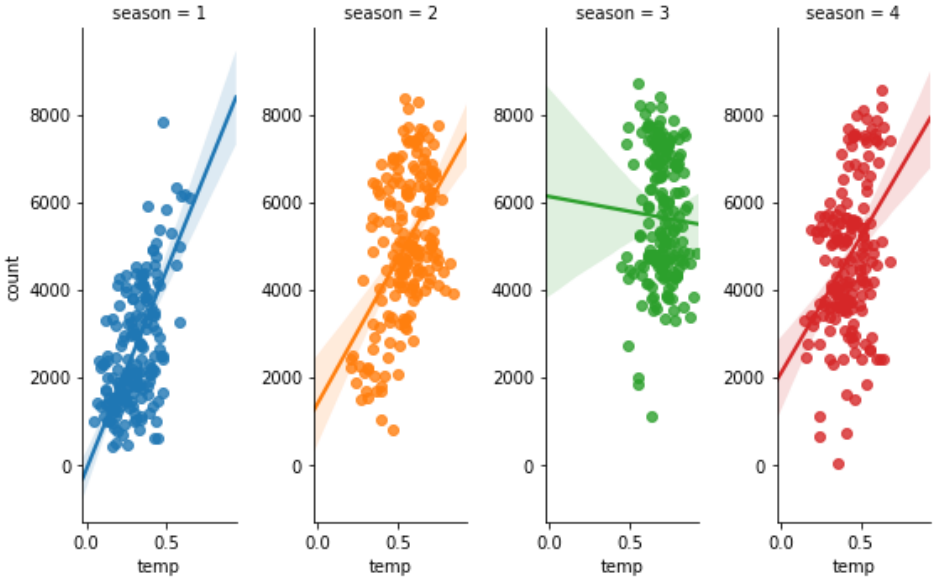
Temperature vs. Count



The main findings from the above plot are as followers

- Seems like there is an positive relation between temperature and total count

Multivariate analysis with season



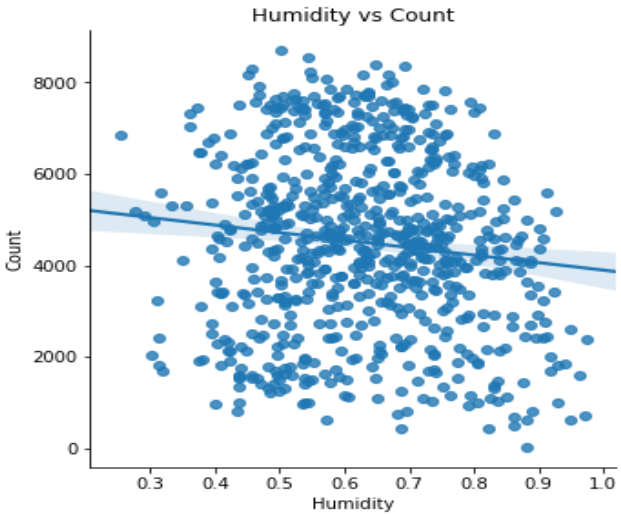
There is a very slight down ward trend and more people are riding in season 3, rest of the seasons have positive relation.

Mean of temperature by seasons

1 0.297748
2 0.544405
3 0.706309
4 0.422906

Season 3 has the highest mean temperature and people are riding around this temperature

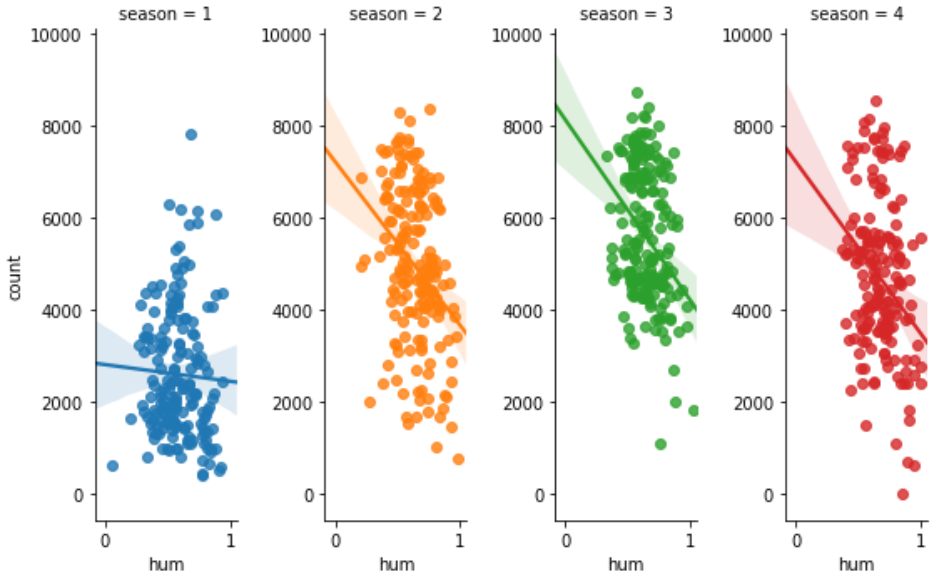
Humidity vs. Count:-



The main findings from the above plot are as followers

- Negative relationship between Humidity and total count

Multivariate analysis with season



- All the 4 seasons have a negative relation further we can see the change in humidity and c ount season wise.

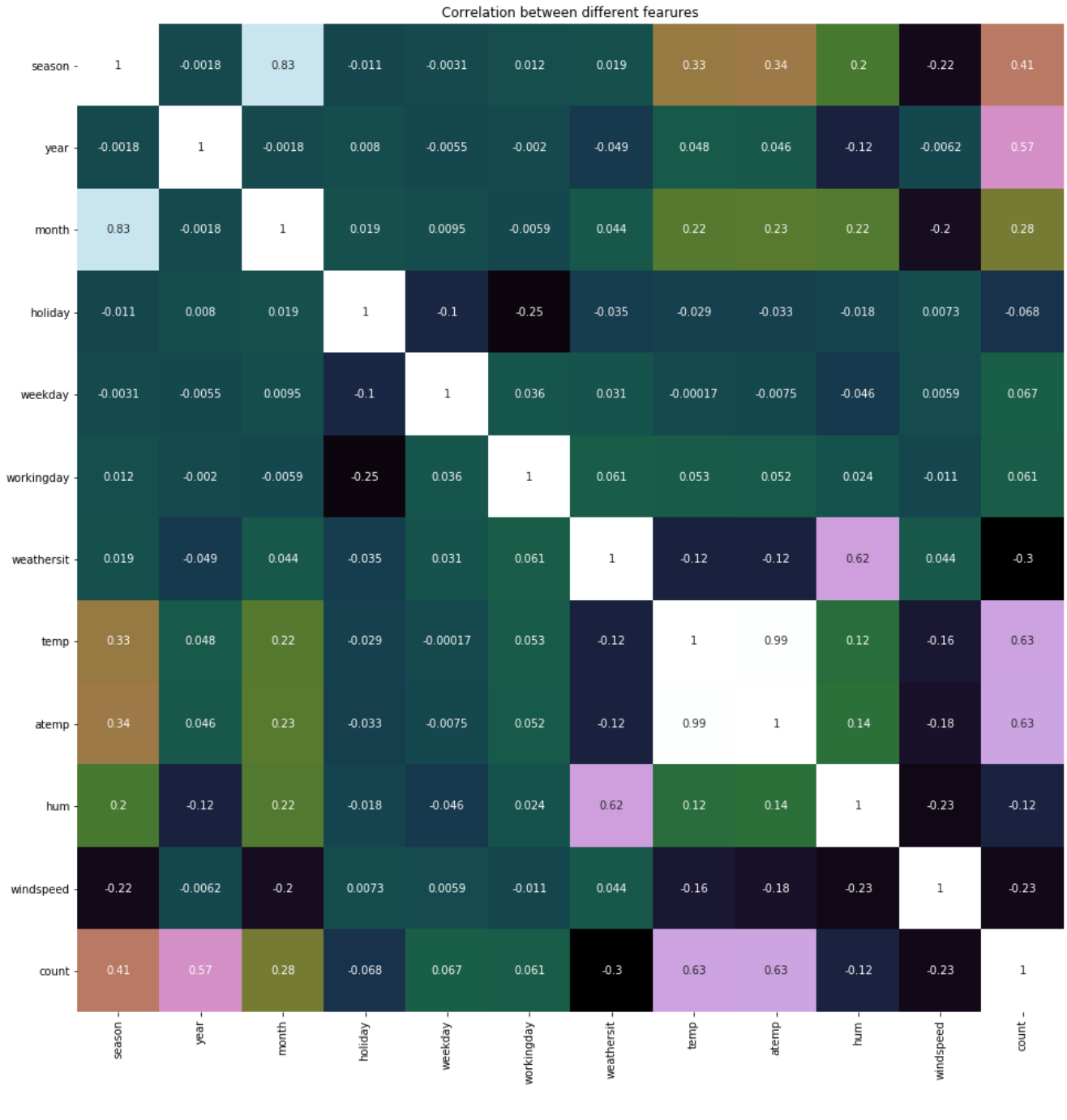
Wind speed:-

- Humidity and wind speed are opposite to each other

Temperature, humidity, wind speed, season are important variables for prediction.

2.1.9 **Correlation analysis :-**

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two variables



From the above plot the followings are the findings.

There is an strong correlation b/w temp and atemp

3 MODELING

3.1 Machine learning Methods

3.1.1 Training and Test set:- I have taken last 21 days as test and rest of the days or points as training set. Train set has 710 rows and Test set has 21 rows

3.1.2 Linear Regression:- a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables

Dummy variables:-

Created dummy variables for the following

- Weather sit
- Season
- Weekday

Following are the variables used for prediction of count:-

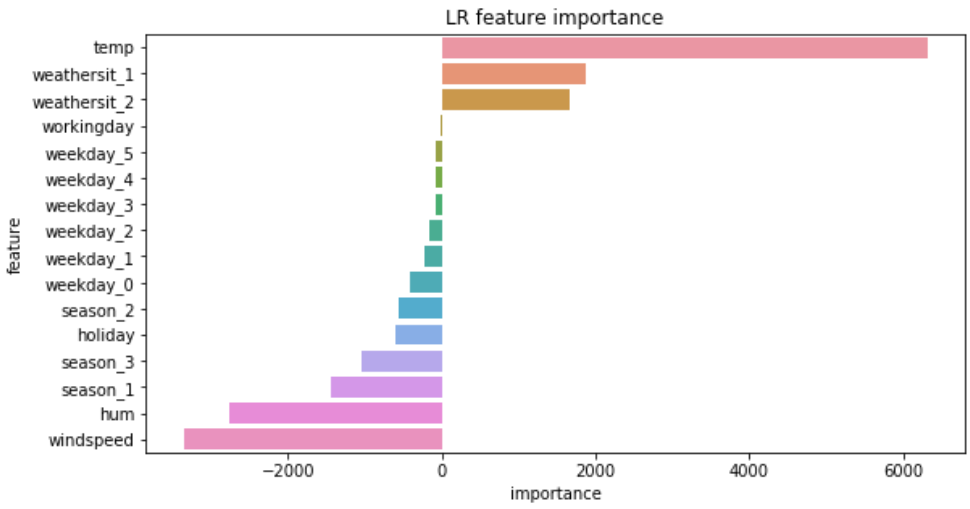
'temp', 'hum', 'windspeed','workingday','holiday','weathersit_1', 'weathersit_2', 'season_1', 'season_2','season_3', 'weekday_0', 'weekday_1', 'weekday_2', 'weekday_3' 'weekday_4', 'weekday_5'

Testing on test dataset the followings are the metrics outputs-

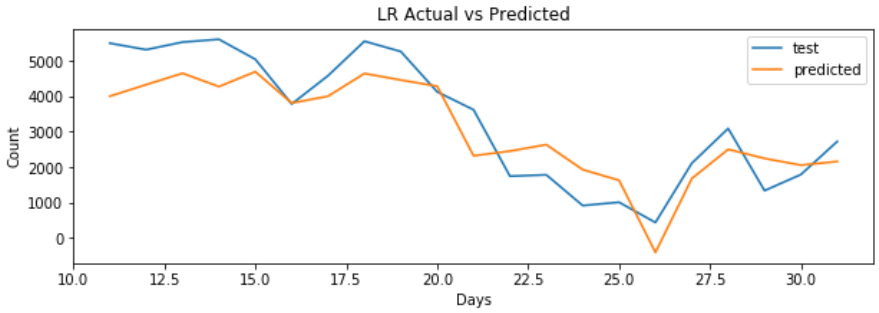
Root mean square error (RMSE) - 831

R squared - 0.78

Feature importance from the model is as followers:-



Actual (test set) vs. predicted plot:-



3.1.3 Linear SVM:- a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection.

Following are the variables used in prediction:-

'season', 'holiday', 'weekday','workingday', 'weathersit', 'temp', 'hum', 'windspeed'

Parameter turning:- This was done by grid search with cv equal to 3

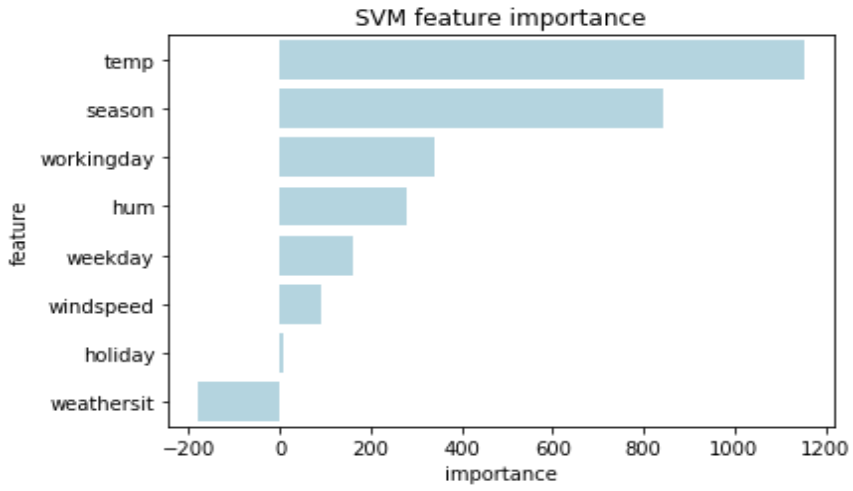
C = 1/alpha, by parameter turning C = 17

Testing on test dataset the followings are the outputs-

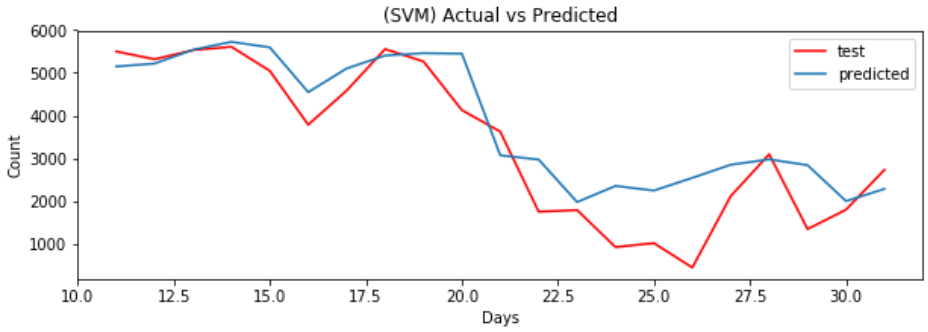
Root mean square error (RMSE) - 878

R squared - 0.75

Feature importance from the model is as followers:-



Actual (test set) vs. predicted plot:-



3.1.4 Decision Tree :- Decision tree builds regression or classification models in the form of a tree structure.

Hyper turning parameters:- This was done with the help of Grid search and cv = 3

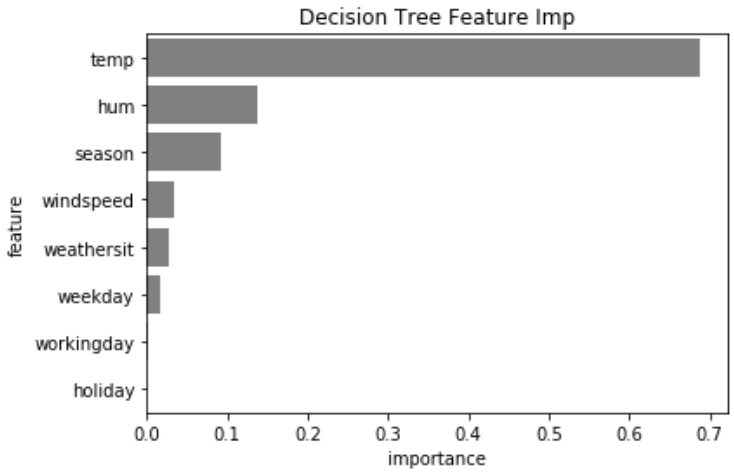
'max_depth': 6
'min_samples_split': 4

Testing on test dataset the followings outputs are from the metrics-

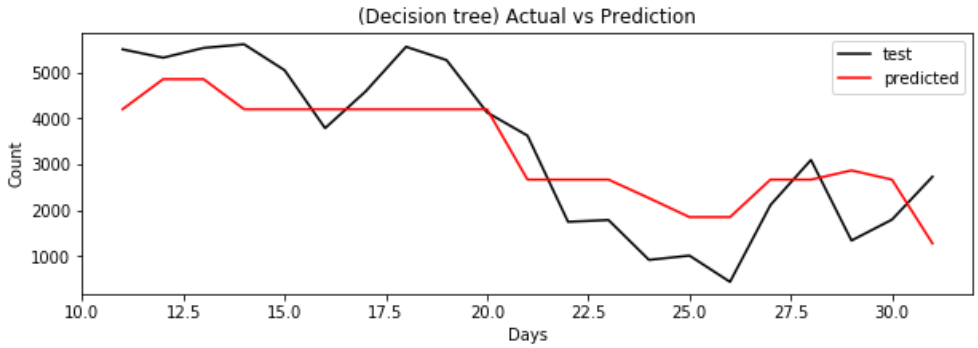
Root mean square error (RMSE) - 1002

R squared - 0.67

Feature importance from the model is as followers:-



Actual (test set) vs. predicted plot:-



3.1.5 Adaboost:- Adaptive Boosting, is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire, who won the 2003 Gödel Prize for their work. It can be used in conjunction with many other types of learning algorithms to improve performance.

Base estimator = Decision tree with max depth equal to 0.6

Tuned parameter as followers:-

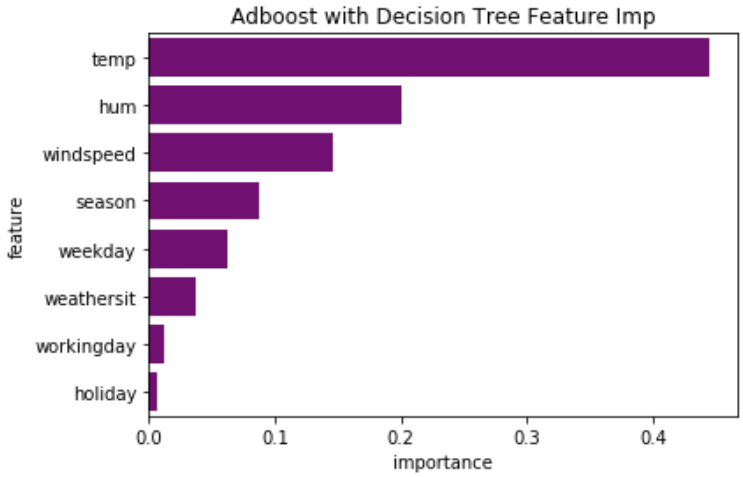
N estimators = 150

Testing on test dataset the followings outputs are from the metrics-

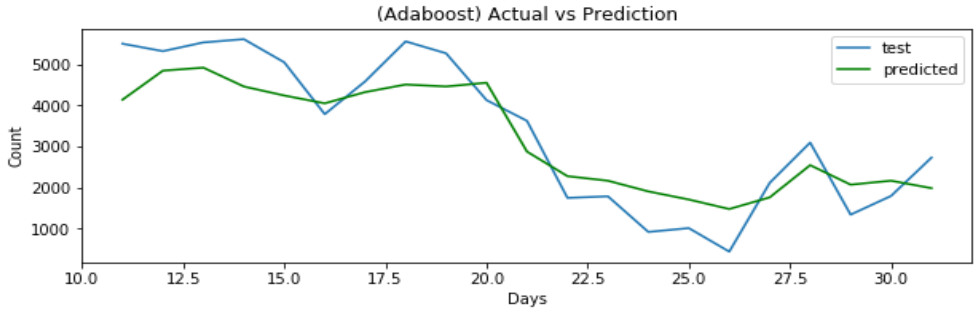
Root mean square error (RMSE) - 745

R squared - 0.82

Feature importance from the model is as followers:-



Actual (test set) vs. predicted plot:-



3.1.6 Random forest:- random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression).

Parameter turning by grid search is as followers:-

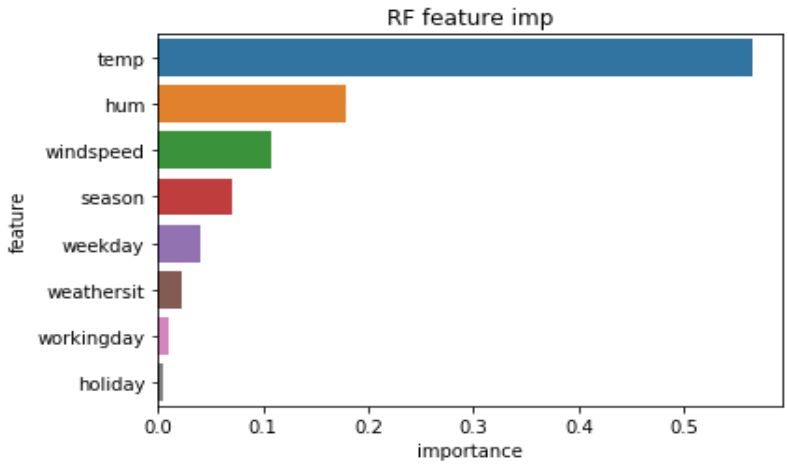
{'n_estimators': 250}
{'max_depth': 10}
{'min_samples_split': 4}
{'min_samples_leaf': 1}
{'max_leaf_nodes': None}
{'max_features': 'auto'}

Testing on test dataset the followings are the outputs from the metrics-

Root mean square error (RMSE) - 904

R squared - 0.73

Features imp:-



3.1.7 Extra tree regressor :- This class implements a meta estimator that fits a number of rando mized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and use averaging t o improve the predictive accuracy and control over-fitting.

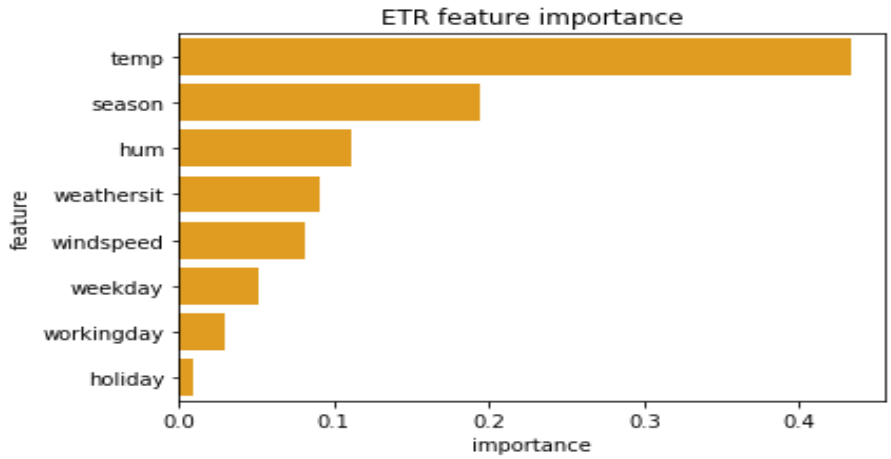
Parameter turning by grid search is as followers:-

{'n_estimators': 30}
{'max_depth': 12}
{'min_samples_split': 3}
{'min_samples_leaf': 1}
{'max_leaf_nodes': None}
{'max_features': 'auto'}

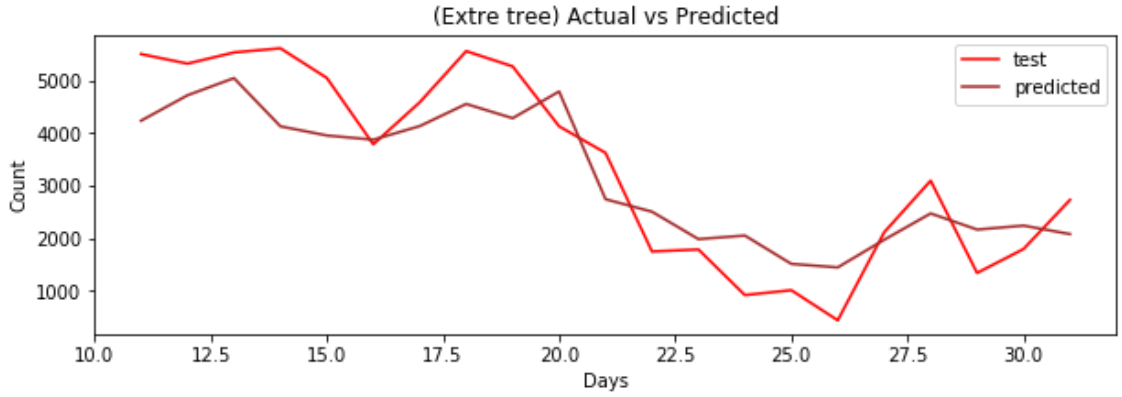
Testing on test dataset the followings are the outputs from the metrics-

Root mean square error (RMSE) - 850
R squared - 0.77

Features imp :-



Actual (test set) vs. predicted plot



3.1.8 XGBOOST:- XGBoost is an open-source software library which provides the gradient boo sting framework for Python, R, and Julia. It works on Linux, Windows, and macOS. From the pr oject description, it aims to provide a "Scalable, Portable and Distributed Gradient Boosting (GB M, GBRT, GBDT) Library".

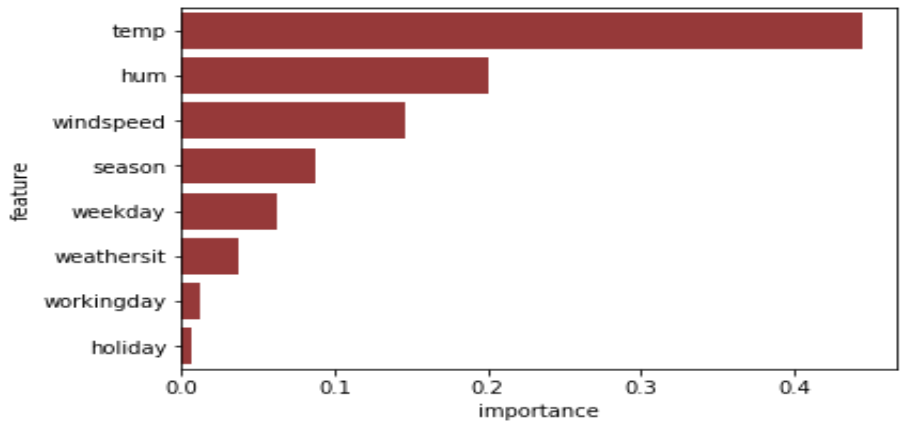
Parameter turning by grid search is as followers:-

{'n_estimators': 100}
{'max_depth': 3}
{'min_child_weight': 6}
{'gamma': 0.0}
{'subsample': 0.7}
{'colsample_bytree': 0.9}
{'reg_alpha': 0.05}
{'learning_rate': 0.1}

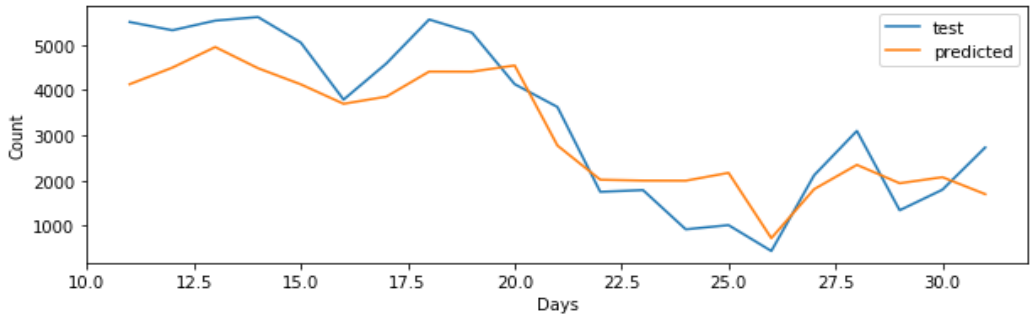
Testing on test dataset the followings are the outputs from the metrics-

Root mean square error (RMSE) - 798
R squared - 0.795

Features importance:-



Actual (test set) vs. predicted plot



4 Results and Conclusions

Results:-

Model	RMSR	R squared
Linear regression	831	0.77
Linear SVM	871	0.75
Decision Tree	1002	0.67
Adaboost	745	0.82
Random forest	904	0.73
Extra tress regressor	850	0.77
Xgboost	798	0.795

Conclusion:- From the above table we can conclude that Adaboost had performed well in predic ting the last 21 days count, followed by Xgboost model, further in the modeling stage feature im portance from the models had been shown in the form of bar plot.