Bike Rental

Sanket Agrawal

29th May, 2019

Contents

1 Introduction 2
1.1 Problem Statement 2
1.2 Data 2
2 Methodology 6
2.1 Pre-Processing 6
2.1.1 Uni variate Analysis 6
2.2 Missing value Analysis 8
2.3 Outlier Analysis 8
2.4 Impute missing values 10
2.5 Feature Selection 10
2.6 Model 11
3 Conclusion 13
3.1 Model evaluation 13
3.2 Model Selection 15
Appendix - A 16
Appendix - B 20
References 22

Chapter 1

Introduction

1.1 Problem Statement

The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings.

1.2 Data

Dataset Details: Dataset has a total number of twelve independent variables and three dependent variables and one index column. Out of twelve independent variables there are eight categorical variables and four continuous variables. The dataset has 731 data instances.

Attribute Information:

1 instant: Record index
2 dteday: Date
3 season: Season (1:spring, 2:summer, 3:fall, 4:winter)
4 yr: Year (0: 2011, 1:2012)
5 mnth: Month (1 to 12)
6 hr: Hour (0 to 23)
7 holiday: weather day is holiday or not (extracted from Holiday Schedule)
8 weekday: Day of the week working day: If day is neither weekend nor holiday is 1, otherwise is 0.
9 weathersit: (extracted from Freemeteo) 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
10 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)
11 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)
12 hum: Normalized humidity. The values are divided to 100 (max)

13 windspeed: Normalized wind speed. The values are divided to 67 (max)
14 casual: count of casual users
15 registered: count of registered users
16 cnt: count of total rental bikes including both casual and registered

Sample data: The table 1.1 shows the a sample instance of the data.

Table 1.1: Sample instance from the dataset

Instant	**Date day**	**Season**	**Year**	**Month**	**Holiday**	**Week day**	**Working day**
1	2011-01-01	1	0	1	0	6	0
Weather situation	**Temperature**	**Actual temp**	**Humidity**	**Wind speed**	**Casual**	**Registered**	**Total Count**
2	0.344167	0.363625	0.805833	0.160446	331	654	985

Data Characteristic: The table 1.2 shows the data characteristic such as count mean first, second, third quartile values etc.

Table 1.2: Columns and their statistical details.

	Season	Year	Month	Holiday	Week day	Working day	Weather situation
count	731	731	731	731	731	731	731
mean	2.49658	0.500684	6.519836	0.028728	2.997264	0.683995	1.395349
std	1.110807	0.500342	3.451913	0.167155	2.004787	0.465233	0.544894
min	1	0	1	0	0	0	1
25%	2	0	4	0	1	0	1
50%	3	1	7	0	3	1	1
75%	3	1	10	0	5	1	2
max	4	1	12	1	6	1	3
	temp	atemp	Humidity	Wind speed	Casual	Registered	Count
count	731	731	731	731	731	731	731
mean	0.495385	0.474354	0.627894	0.190486	848.176471	3656.172367	4504.348837
std	0.183051	0.162961	0.142429	0.077498	686.622488	1560.256377	1937.211452
min	0.05913	0.07907	0	0.022392	2	20	22
25%	0.337083	0.337842	0.52	0.13495	315.5	2497	3152
50%	0.498333	0.486733	0.626667	0.180975	713	3662	4548
75%	0.655417	0.608602	0.730209	0.233214	1096	4776.5	5956
max	0.861667	0.840896	0.9725	0.507463	3410	6946	8714

Unique Value Counts: The table 1.3 shows the Columns on the left side and the unique value counts on the right. This helped us in differentiating the variables in continuous and categorical.

Table 1.3: Columns with their value counts

Column Name	Value Counts
Temperature	499
Actual temperature	690
Humidity	595
Wind speed	650
Season	4
Year	2
Month	12
Holiday	2
Weekday	7
Working day	2
Weather situation	3
Casual	606
Registered	679
Count	696

Chapter 2

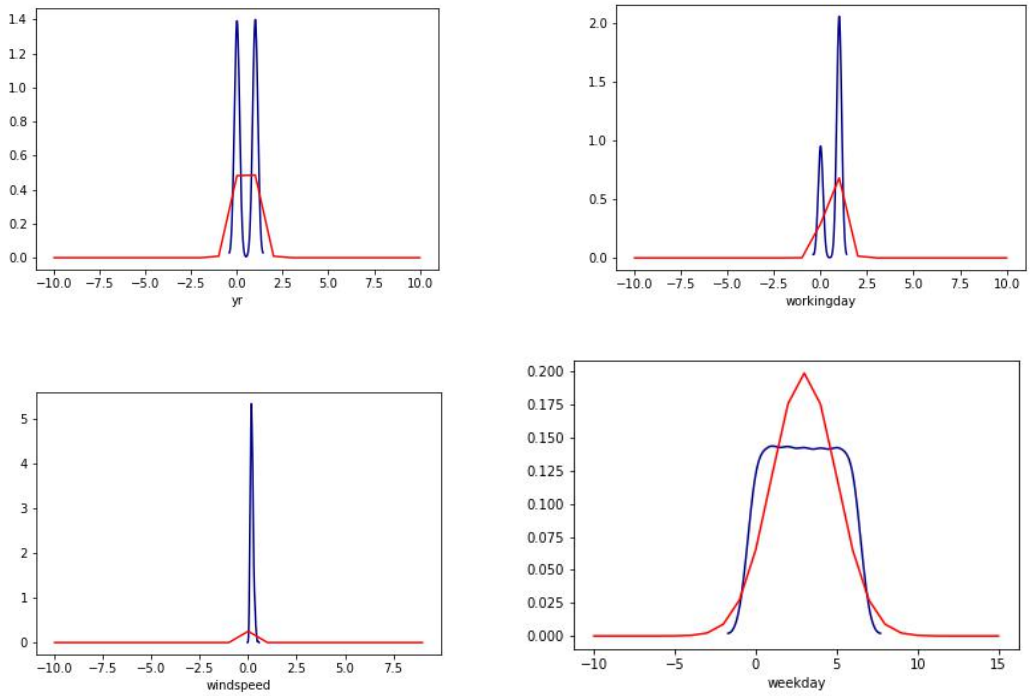
Methodology

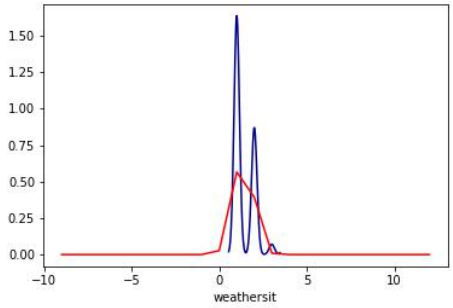
2.1 Pre-Processing

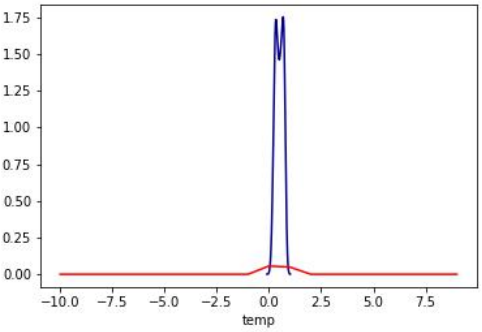
The categorical data is already in numerical form so we don't need to perform any sort of conversion to get the categorical codes out of text.

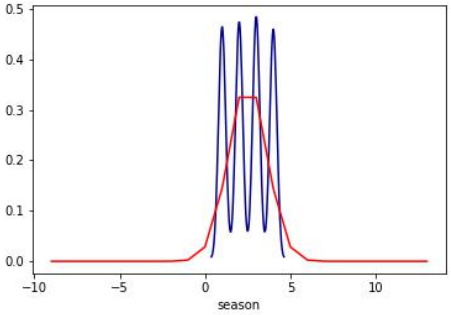
2.1.1 Uni variate Analysis

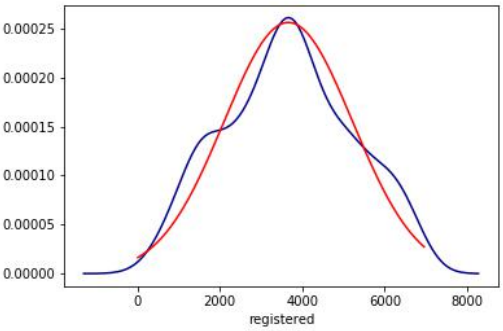
The figure 2.1 shows the data distribution of the variables (blue line) and the normal distribution curve (red line) with the variable mean and standard deviation.
We can see that the continuous variables follows the normal distribution. The error is considered to be normally distributed (residuals). So, checking if the variable is normally distributed or not helps us to verify the hypothesis.

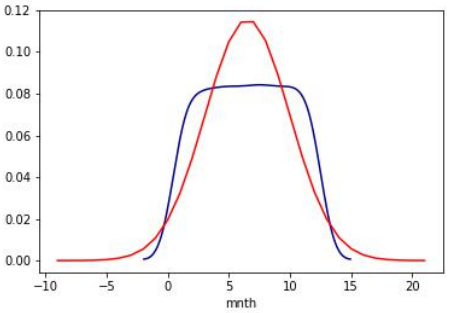


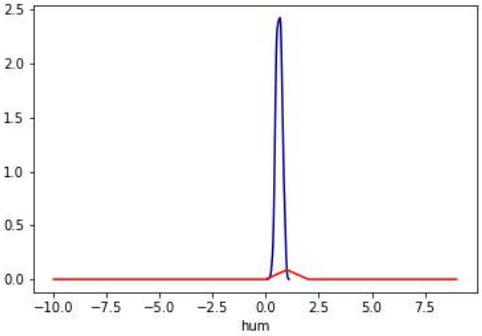

weathersit

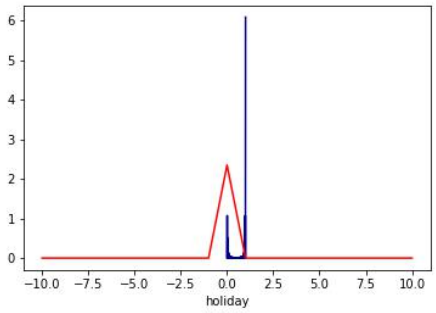

temp

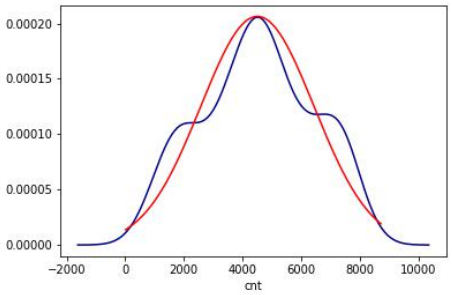

season

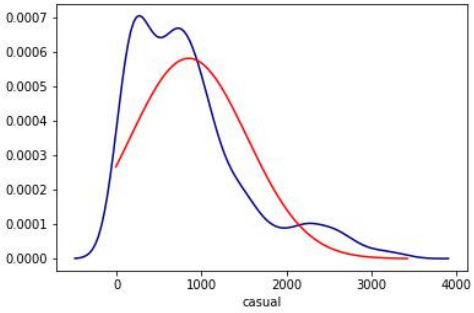

registered


mnth


hum


holiday


cnt


casual

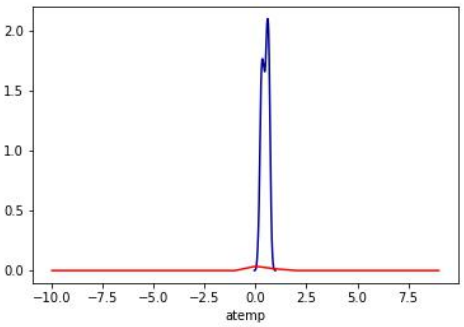

atemp

Figure 2.1: Probability Density Function of different Variables

2.2 Missing Value Analysis

The table 2.1 shows the variables and the missing value percentage. We can see the dataset doesn't has any missing values.

Table 2.1: Variables with their missing value percentage

Variables	Missing percentage
dteday	0
Temperature	0
Actual temperature	0
Humidity	0
Wind speed	0
Season	0
Year	0
Month	0
Holiday	0
Weekday	0
Working day	0
Weather situation	0
Casual	0
Registered	0
Count	0

2.3 Outlier Analysis

Outliers are those instances whose values doesn't t follows the regular pattern. Either these are the erroneous cases or they are the special cases which we need to consider separately. Here, in this case we considered some of the outliers as the erroneous cases and has been deleted and others as special case and thus remained as the part of the dataset. We

first choose to get the instances which has values less than the first quartile value - 1.5 times the IQR or greater than third quartile value + 1.5 times the IQR are considered as the outliers. Then, we plot these minimum and maximum boundaries for the variables wind speed and humidity. We can see in the figure 2.2 that there are two points outside the permissible limits. One of which having zero humidity is erroneous and other with very high wind speed is a special case.

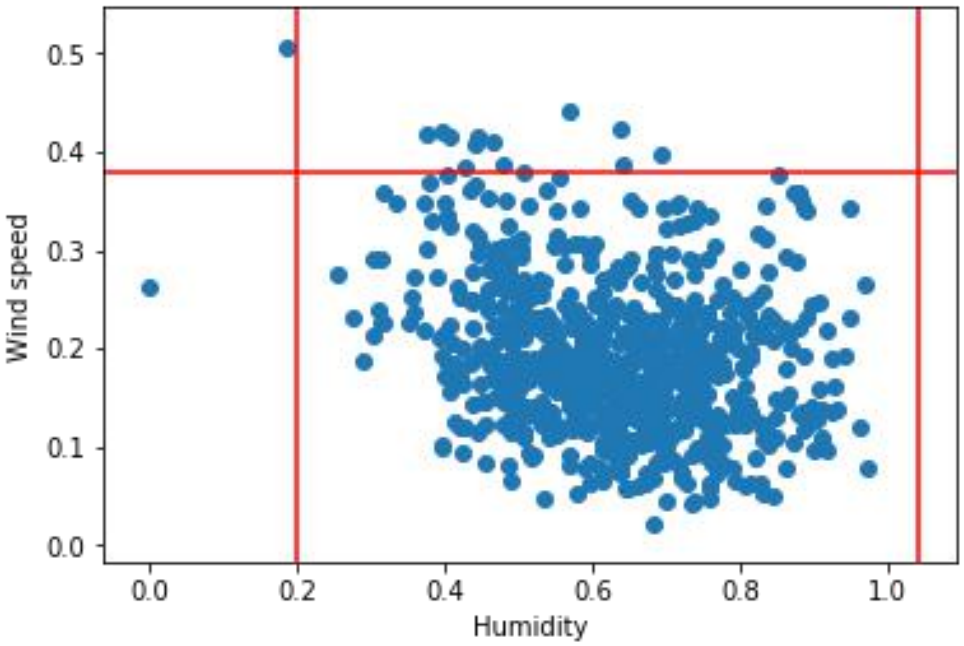


Figure 2.2: Showing the data distribution between humidity and wind speed.

Below is the information of each variable:

Table 2.2: Variables and outlier details

Variable Name	Maximum permissible value	Minimum permissible value	Outlier count
Temperature	1.132916	-0.140416	0
Actual temperature	1.01474125	-0.06829675	0
Humidity	1.04552125	0.20468725	2
Wind speed	0.38061125	-0.01244675	13

Figure 2.3 has the box plots for the Continuous variables showing t he outliers.

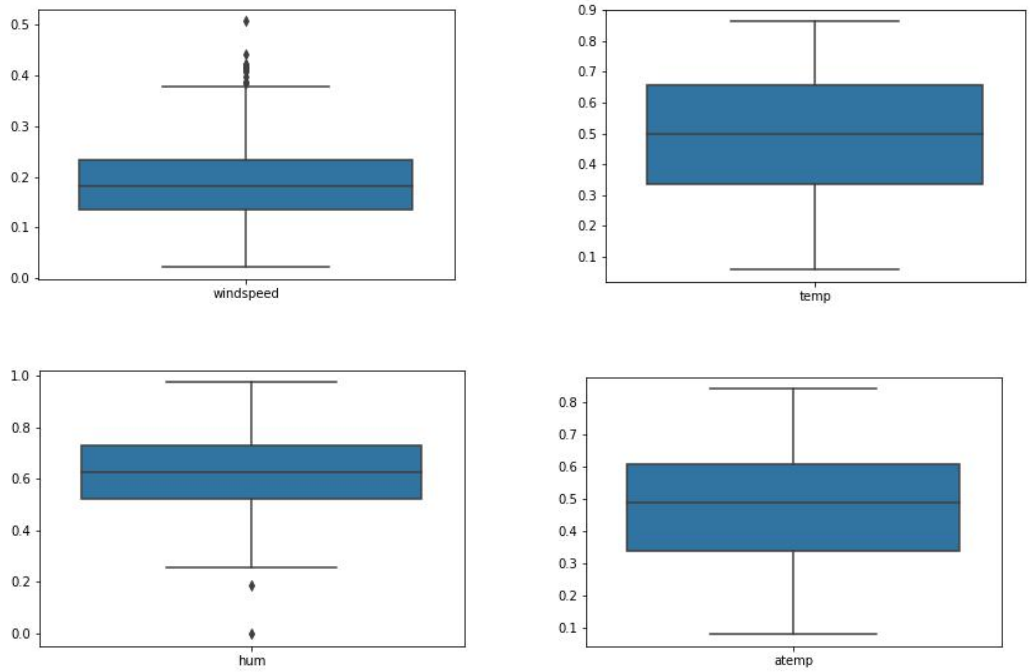


Figure 2.3: Box plots for different Variables

2.4 Impute Missing Value

There are no missing values in the dataset so no data imputation is required.

2.5 Feature Selection

Feature selection is an important and next step after the removal of outliers and imputation of NaNs. In feature selection we choose the independent variables which are really independent to each other and contributes maximum for predicting the dependent variable. For this purpose we have used three tests:

➢ Correlation matrix

The correlation matrix between all the continuous variables is used for feature selection. If the correlation value is very high then we drop one of the two variables. As the information incorporated by t hese variables is redundant. Thus, we choose to drop temp as it ha s correlation of 0.99 with atemp. Also, we are not removing the re gistered or cnt variables as these are dependent variables and will b e used in predictions (Figure 2.4).

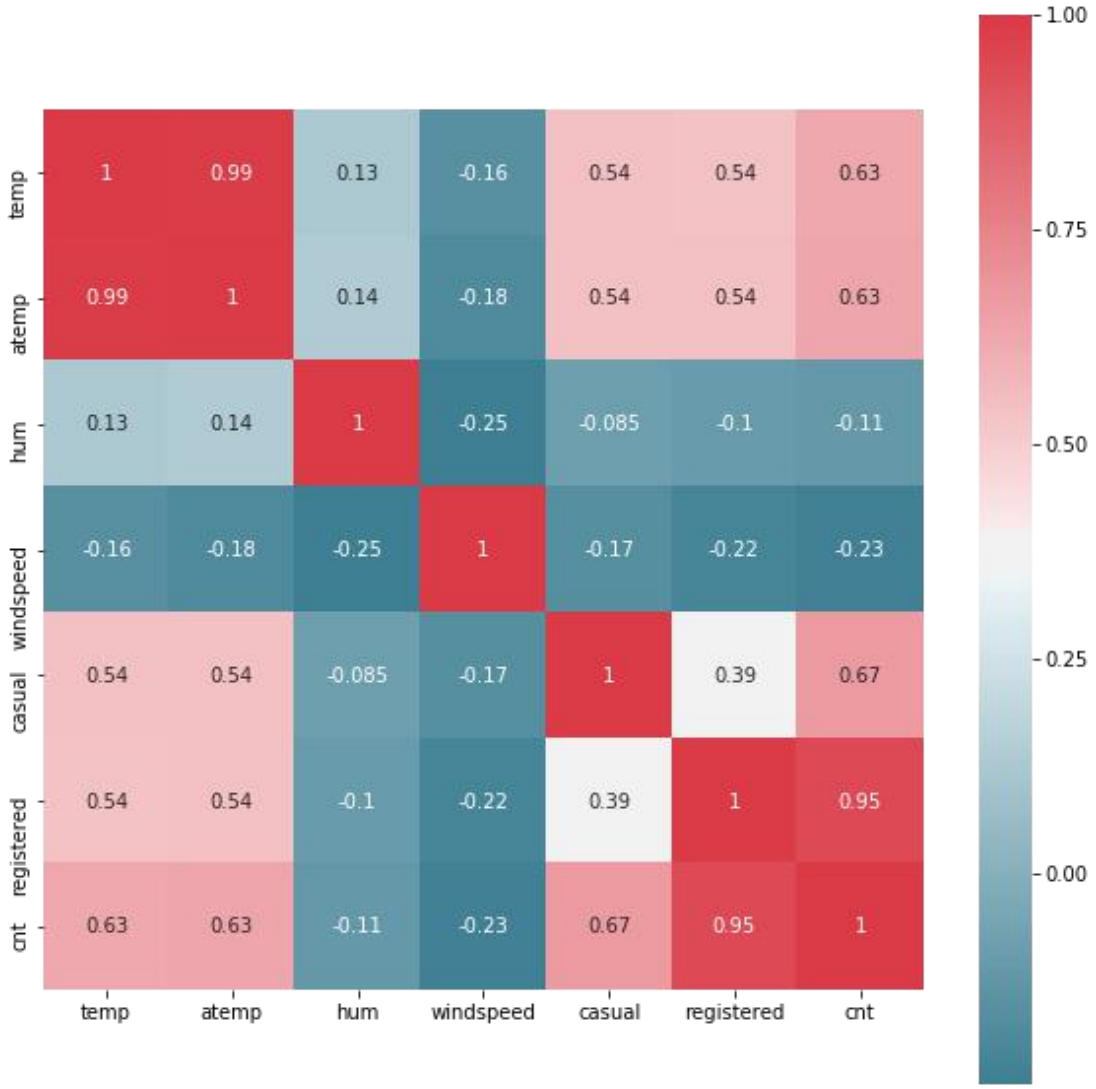


Figure 2.4: Correlation Matrix of Continuous Variables

Thus, we drop the **"temp"** variable from the dataset. Also, the **"dteday"** is dropped as the information incorporated by this variable is already stored in the form of day, year and month variables.

2.6 Model

We have used two regressive models for the prediction of the dependent variable bike rental count. For prediction, we choose Random forest regressor and XGboost regressor. Choosing the model is just not enough. Complex models like these has many parameters and choosing the right parameters is very important as this will prevent over-fitting and under-fitting and giving us the maximum accuracy. This configuration is applied to total count, casual, and registered counts separately. Then, the predictions of casual and registered are added. And, finally the average of the total count predictions and this sum of casual and registered

predictions is taken. This model has given us the proper balance of variance and bias in the model. After tuning the parameters we found the best suited parameters for the choose regression models are as follows:

First model i.e., we used the an ensembler, Random forest a bagging algorithm, for predicting the values. The configurations (python) used for it are as follows: Number of trees used to estimate the values as 25, max depth as 18, random state as 1, min samples leaf as 1, min samples split as 10.

Second model i.e., XGBoost, is a boosting algorithm. The configuration (python) used is as follows: colsample_bytree as 0.8, gamma as 0.3, max depth as 2, min child weight as 5, subsample as 0.8.

Table 2.3 shows the importance values given to each independent variable by the random forest regressor:

Table 2.3: Variables with their importance values by random forest regressor

Variable Name	Feature importance
Actual temperature	0.47364
Year	0.284287
Season	0.093701
Humidity	0.049912
Month	0.032165
Weather situation	0.026336
Wind speed	0.022859
Week day	0.013067
Holiday	0.002231
Working day	0.001802

Chapter 3

Conclusion

3.1 Model Evaluation

3.1.1 RMSE(Root Mean Squared Error)

RMSE or root mean squared error is an error metric for regression problems. It is calculated as the square root of squared difference of the predicted and target values. The lower the value the better is the performance.

Random Forrest Regressor:

Table 3.1: RMSE values by Random forest Regressor

	Casual	Registered	Total count(P1)	Casual+Registered(P2)	(P1+P2)/2
Training Dataset	179.02	358.61	**424.41**	**432.63**	**417.99**
Validation Dataset	319.67	564.31	**768.83**	**742.93**	**741.55**

Clearly, the loss in the fifth column is minimum. Because the average has reduced the overall variance and the bias of the predictions.

XG Boost Regressor:

Table 3.2: RMSE values by Random forest Regressor

	Casual	Registered	Total count(P1)	Casual+Registered(P2)	(P1+P2)/2
Training Dataset	228.07	435.30	**519.49**	**518.53**	**511.42**
Validation Dataset	304.13	521.76	**694.52**	**680.08**	**679.62**

Clearly, the loss in the fifth column is minimum. Because the average has reduced the overall variance and the bias of the predictions.

3.1.2 MAPE Score

MAPE or Mean absolute percentage error is a metric for regression models. It is calculated as the mean of absolute value of difference of true_value & predicted value divided by the true_value. Lower the value better is the performance.

Random Forrest Regressor:

Table 3.3: MAPE values by Random forest Regressor

	Casual	Registered	Total count(P1)	Casual+Registered(P2)	(P1+P2)/2
Training Dataset	55.53	27.96	**31.46**	**29.08**	**30.04**
Validation Dataset	38.64	16.15	**17.03**	**16.54**	**16.45**

Here, the training data has higher loss than the validation data. This is against the conventional loss. This can be explained as: the training data may have the special points and validation data may not have. Thus, increasing the training loss. We can still use the metrics by comparing the loss relatively. Column fourth indicates the over-fitting as in the fifth column our training loss increases and the validation loss decreased and column three represents the under-fitting.

XG Boost Regressor:

Table 3.4: MAPE values by XG boost Regressor

	Casual	Registered	Total count(P1)	Casual+Registered(P2)	(P1+P2)/2
Training Dataset	42.07	28.47	**26.23**	**26.90**	**26.27**
Validation Dataset	38.15	14.06	**15.02**	**14.20**	**14.41**

3.2 Model Selection

From the above tables we can infer that the XGBRegressor has the minimum loss after taking mean of the predictions P1 & P2. Although, the predictions P2 can also be used as they also have comparable loss.

Appendix A - Observations

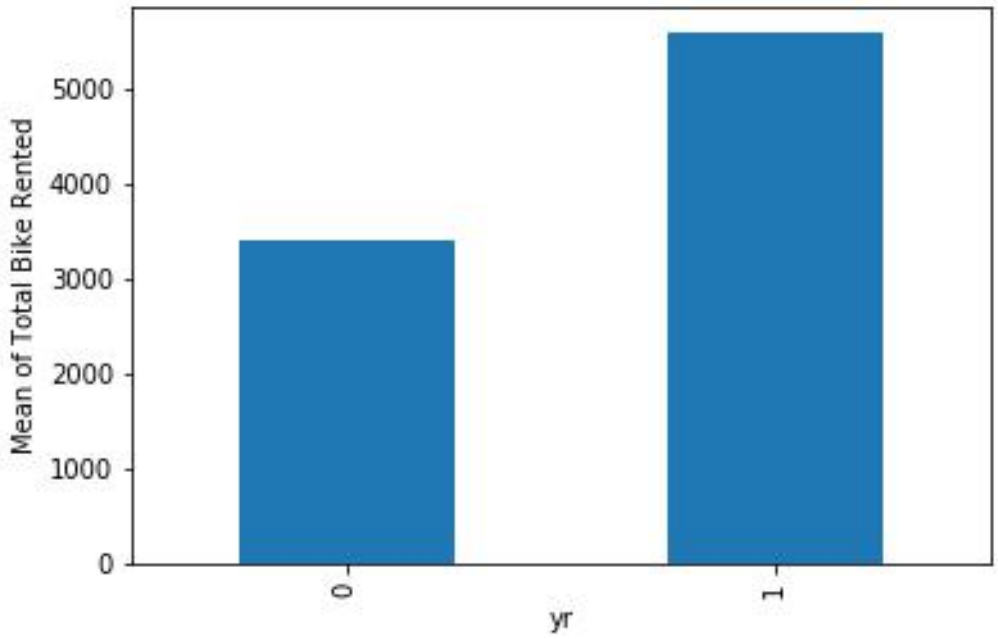


Figure 1: Showing the mean of total bikes rented and the Year.

Figure 1 shows that the more bikes are rented in the year 2012 than in the year 2011.

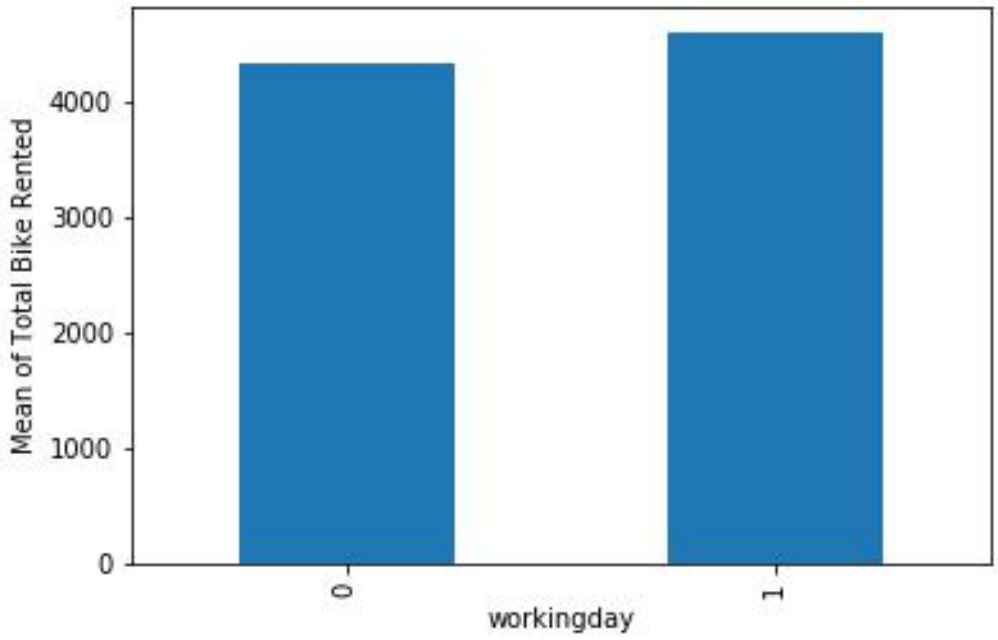


Figure 2: Showing the mean of total bikes rented and the working day.

Figure 2 shows that the more bikes are rented on the working days than on the holidays.

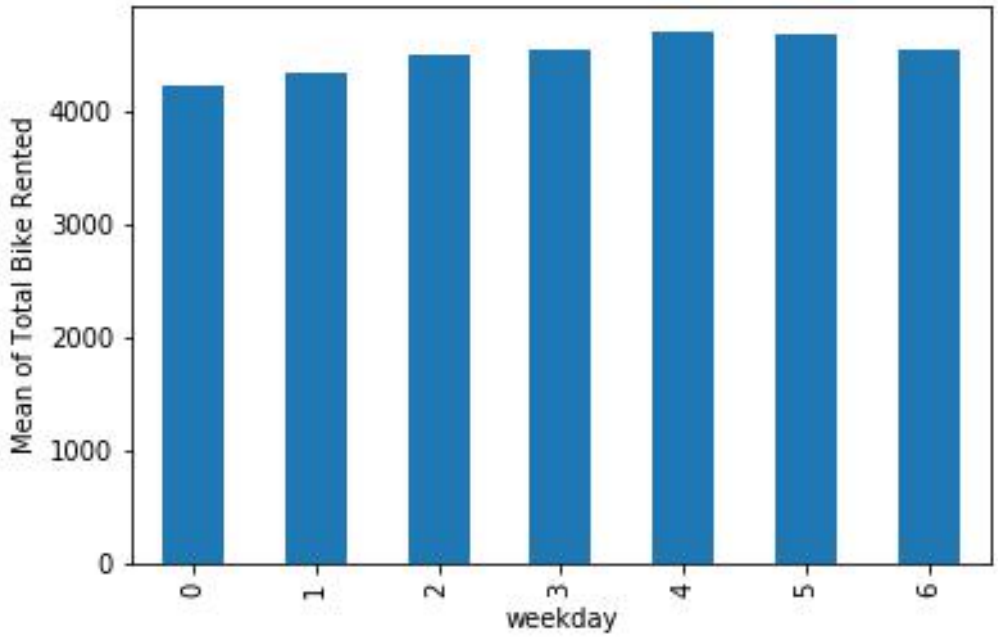


Figure 3: Showing the mean of total bikes rented and the week day.

Figure 3 shows that the bikes are rented in a similar manner.

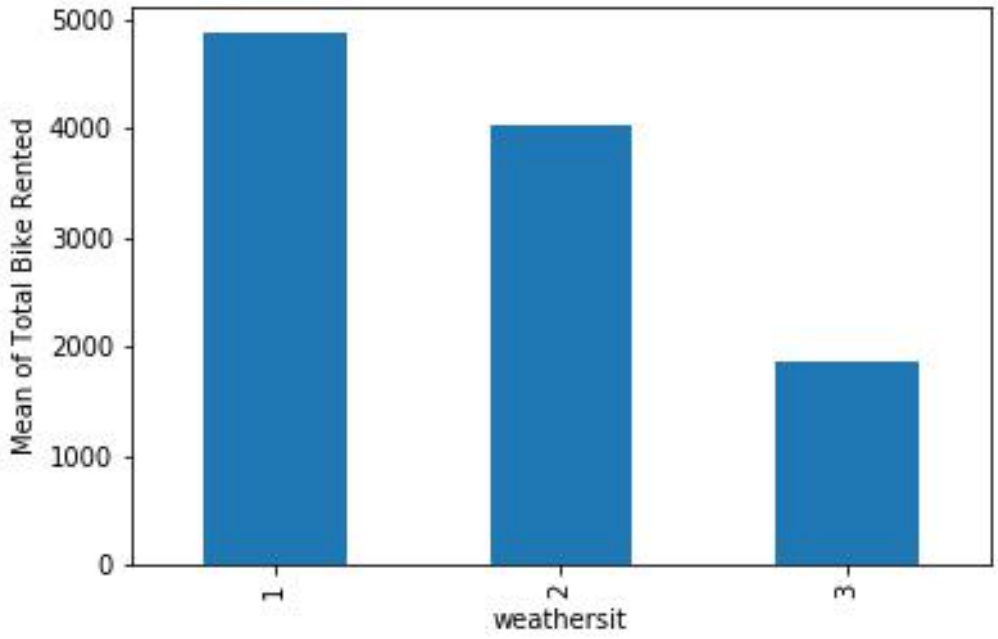


Figure 4: Showing the mean of total bikes rented and the weather situation.

Figure 4 shows that the more bikes are rented for weather condition 1 i.e., Clear, Few clouds, Partly cloudy, Partly cloudy.

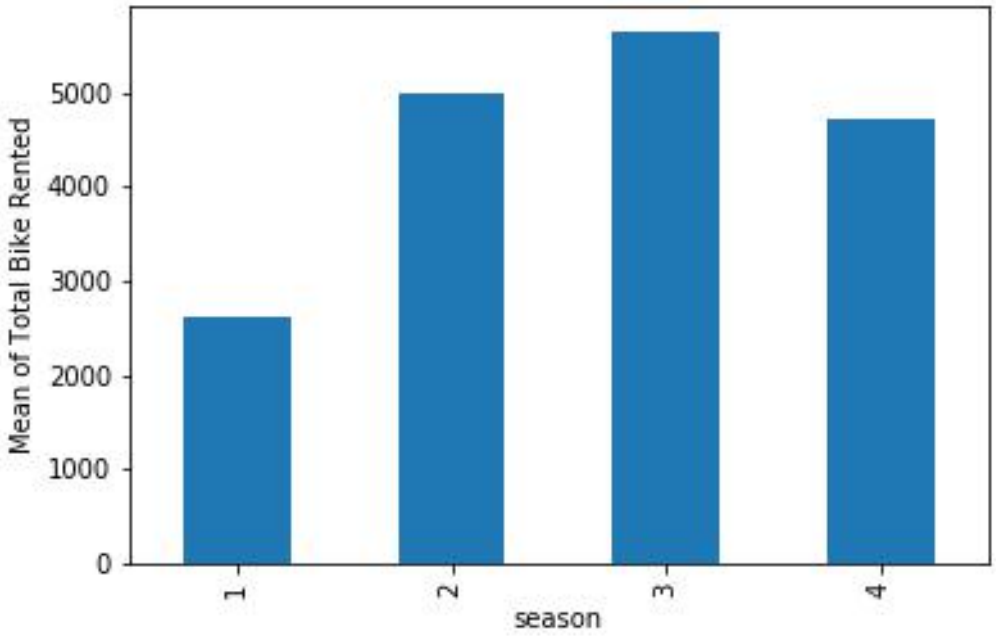


Figure 5: Showing the mean of total bikes rented and the Season.

Figure 5 shows that the more bikes are rented in fall.

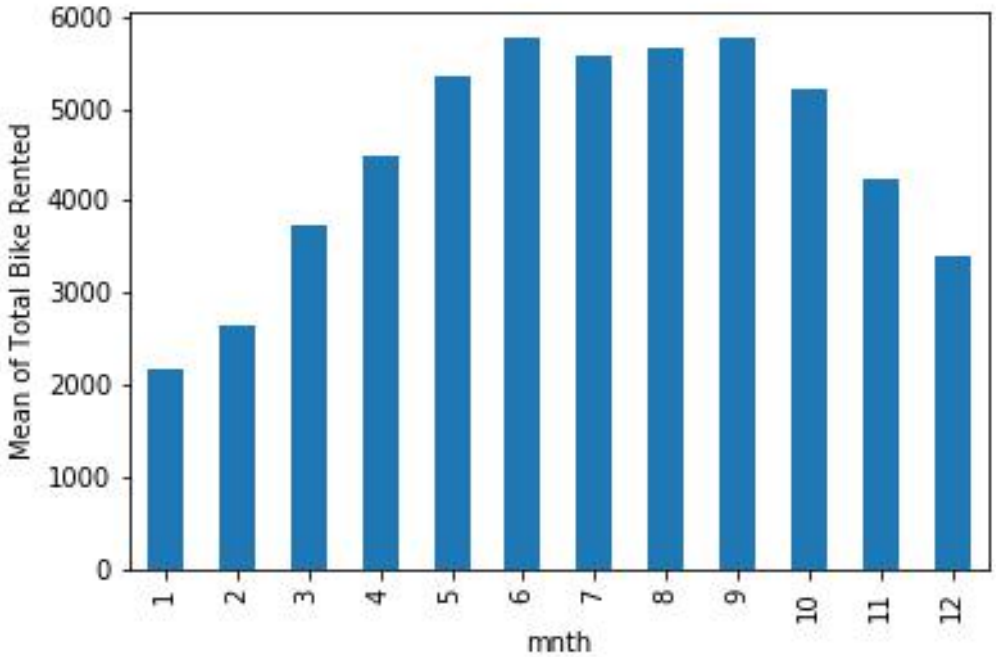


Figure 6: Showing the mean of total bikes rented and the Month.

Figure 6 shows that the more bikes are rented in the months of June to September.

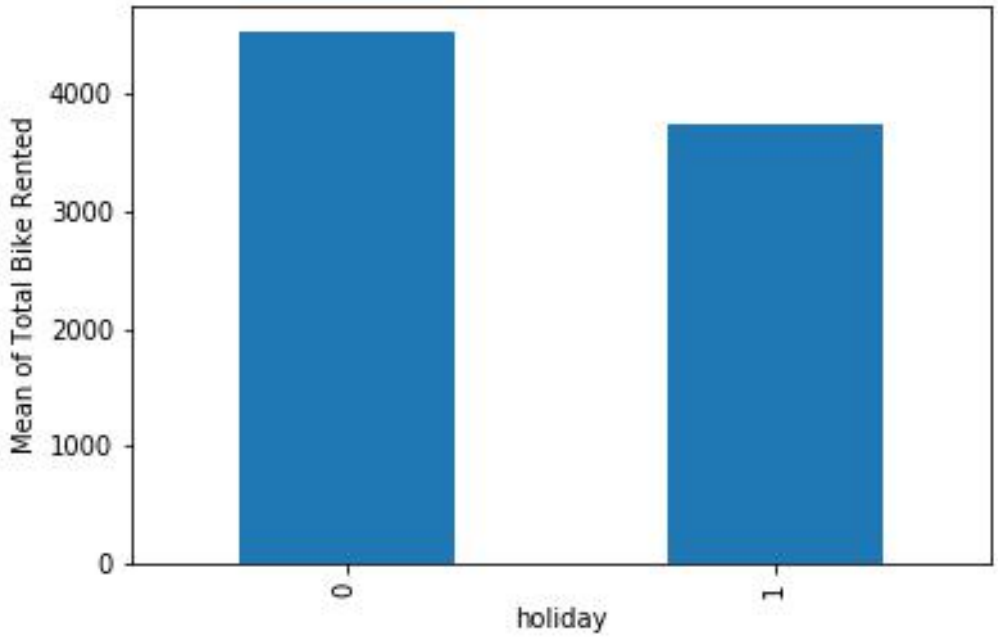


Figure 7: Showing the mean of total bikes rented and the Holiday.

Figure 7 shows that the more bikes are rented on working days.

Appendix B - R Results

A. RMSE(Root Mean Squared Error)

Random Forrest Regressor:

Table 1A: RMSE values by Random forest Regressor

	Casual	Registered	Total count(P1)	Casual+Registered(P2)	(P1+P2)/2
Training Dataset	135.67	259.06	**424.41**	**334.97**	**323.03**
Validation Dataset	269.26	526.45	**768.83**	**641.99**	**617.86**

XG Boost Regressor:

Table 2A: RMSE values by Random forest Regressor

	Casual	Registered	Total count(P1)	Casual+Registered(P2)	(P1+P2)/2
Training Dataset	193.13	375.64	**475.03**	**443.64**	**440.55**
Validation Dataset	259.90	510.88	**615.37**	**600.96**	**585.01**

B. MAPE Score

Random Forrest Regressor:

Table 1B: MAPE values by Random forest Regressor

	Casual	Registered	Total count(P1)	Casual+Registered(P2)	(P1+P2)/2
Training Dataset	34.95	22.65	**24.13**	**23.60**	**23.74**
Validation Dataset	56.49	16.61	**17.36**	**17.80**	**17.29**

XG Boost Regressor:

Table 2B: MAPE values by XG boost Regressor

	Casual	Registered	Total count(P1)	Casual+Registered(P2)	(P1+P2)/2
Training Dataset	30.62	20.94	**18.81**	**14.88**	**16.38**
Validation Dataset	35.25	15.10	**15.00**	**16.26**	**15.24**

References

Garrett Grolemund and Hadley Wickham. 2016. R for Data Science. ISBN: 9781491910399

Jake VanderPlas. 2016. Python Data Science Handbook: Essential Tools for Working with Data. ISBN: 9781491912058