Predicting Bike Rental Count

Roopak Krishna

9 December 2019

Content

Chapter 1

Introduction

1.1 Problem Description

The objective of this project is to predict the count of bike rentals based on the seasonal and environmental settings. By achieving this goal, it would be possible to help accommodate in managing the number of bikes required on a daily basis, and providing better services to its customer.

1.2 Data

Our task is to build a regression models which will be responsible for predicting bike counts based on certain variables. Given below is the sample data that is used.

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600

Fig.1 Data Overview

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
instant	1	731	366.00	211.17	366.00	366.00	271.32	1.00	731.00	730.00	0.00	-1.20	7.81
dteday*	2	731	366.00	211.17	366.00	366.00	271.32	1.00	731.00	730.00	0.00	-1.20	7.81
season	3	731	2.50	1.11	3.00	2.50	1.48	1.00	4.00	3.00	0.00	-1.35	0.04
yr	4	731	0.50	0.50	1.00	0.50	0.00	0.00	1.00	1.00	0.00	-2.00	0.02
mnth	5	731	6.52	3.45	7.00	6.52	4.45	1.00	12.00	11.00	-0.01	-1.21	0.13
holiday	6	731	0.03	0.17	0.00	0.00	0.00	0.00	1.00	1.00	5.63	29.75	0.01
weekday	7	731	3.00	2.00	3.00	3.00	2.97	0.00	6.00	6.00	0.00	-1.26	0.07
workingday	8	731	0.68	0.47	1.00	0.73	0.00	0.00	1.00	1.00	-0.79	-1.38	0.02
weathersit	9	731	1.40	0.54	1.00	1.33	0.00	1.00	3.00	2.00	0.95	-0.15	0.02
temp	10	731	0.50	0.18	0.50	0.50	0.23	0.06	0.86	0.80	-0.05	-1.12	0.01
atemp	11	731	0.47	0.16	0.49	0.48	0.20	0.08	0.84	0.76	-0.13	-0.99	0.01
hum	12	731	0.63	0.14	0.63	0.63	0.16	0.00	0.97	0.97	-0.07	-0.08	0.01
windspeed	13	731	0.19	0.08	0.18	0.19	0.07	0.02	0.51	0.49	0.67	0.39	0.00
casual	14	731	848.18	686.62	713.00	744.95	587.11	2.00	3410.00	3408.00	1.26	1.29	25.40
registered	15	731	3656.17	1560.26	3662.00	3641.72	1712.40	20.00	6946.00	6926.00	0.04	-0.72	57.71
cnt	16	731	4504.35	1937.21	4548.00	4517.19	2086.02	22.00	8714.00	8692.00	-0.05	-0.82	71.65

Fig.2 Data Stats

Chapter 2

Methodology

2.1 Pre-Processing

This Part is valuable to data science projects since it allows to get closer to the certainty that the future results will be *valid*, *correctly interpreted*, and *applicable* to the desired business contexts. Such level of certainty can be achieved only after raw data is validated and checked for anomalies, ensuring that the data set was collected without errors. EDA also helps to find insights that were not evident or worth investigating to business stakeholders and data scientists but can be very informative about a particular business.

This is called Exploratory Data Analysis (EDA) which helps to answer all these questions, ensuring the best outcomes for the project. It is an approach for summarizing, visualizing, and becoming intimately familiar with the important characteristics of a data set.

2.1.1 Missing Value

It is the first step to identify the rows that contains missing data which will lead to undesired results or create some errors if not removed.

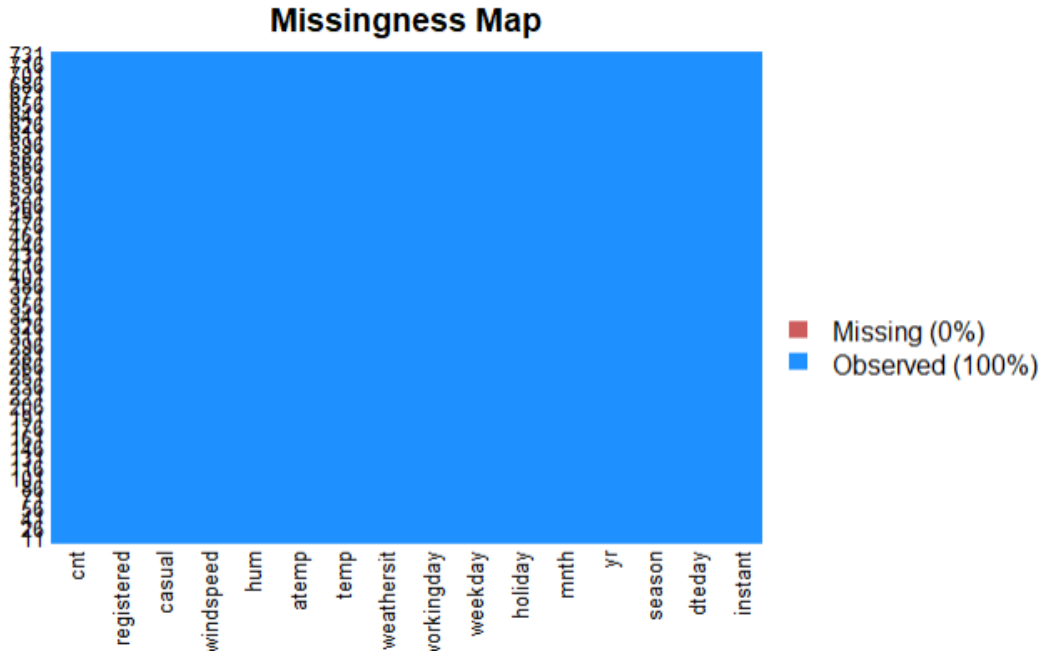


Fig.3 Missing Data Graph

We can observe from this missingness graph that is provided by library(Amelia) that there is no missing Data.

2.1.2 Outlier Analysis

Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample.

Detecting outliers is of major importance for almost any quantitative discipline (ie: Physics, Economy, Finance, Machine Learning, Cyber Security). In machine learning and in any quantitative discipline the quality of data is as important as the quality of a prediction

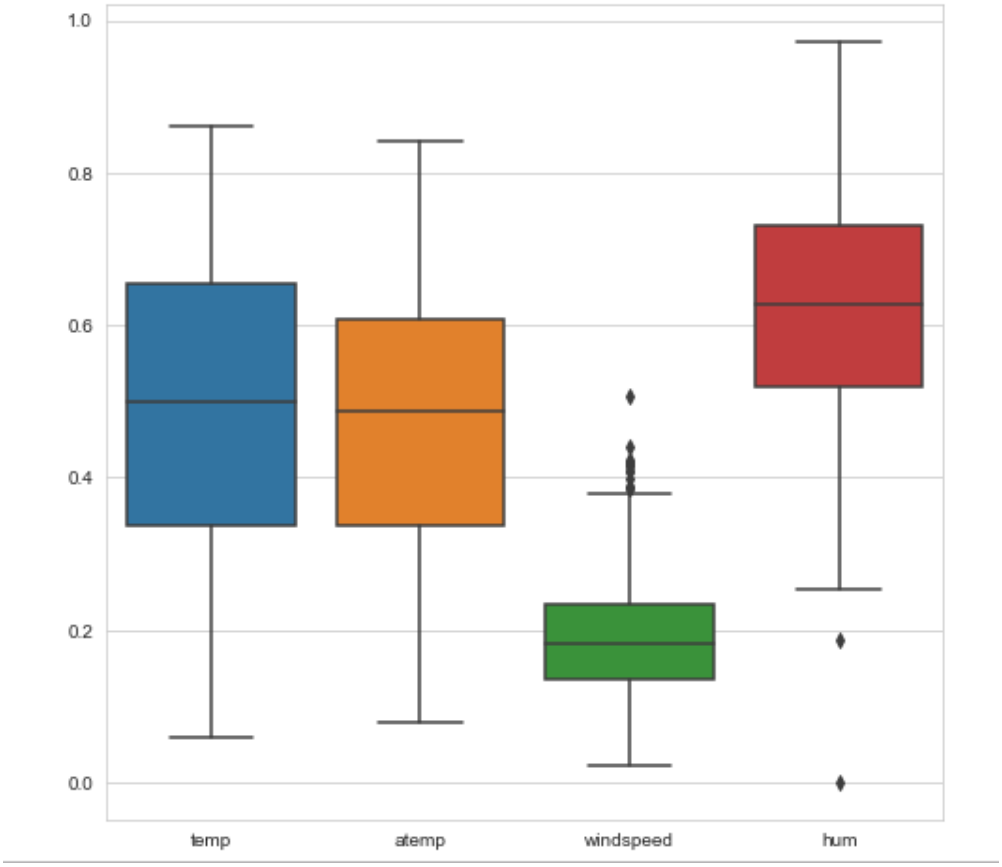
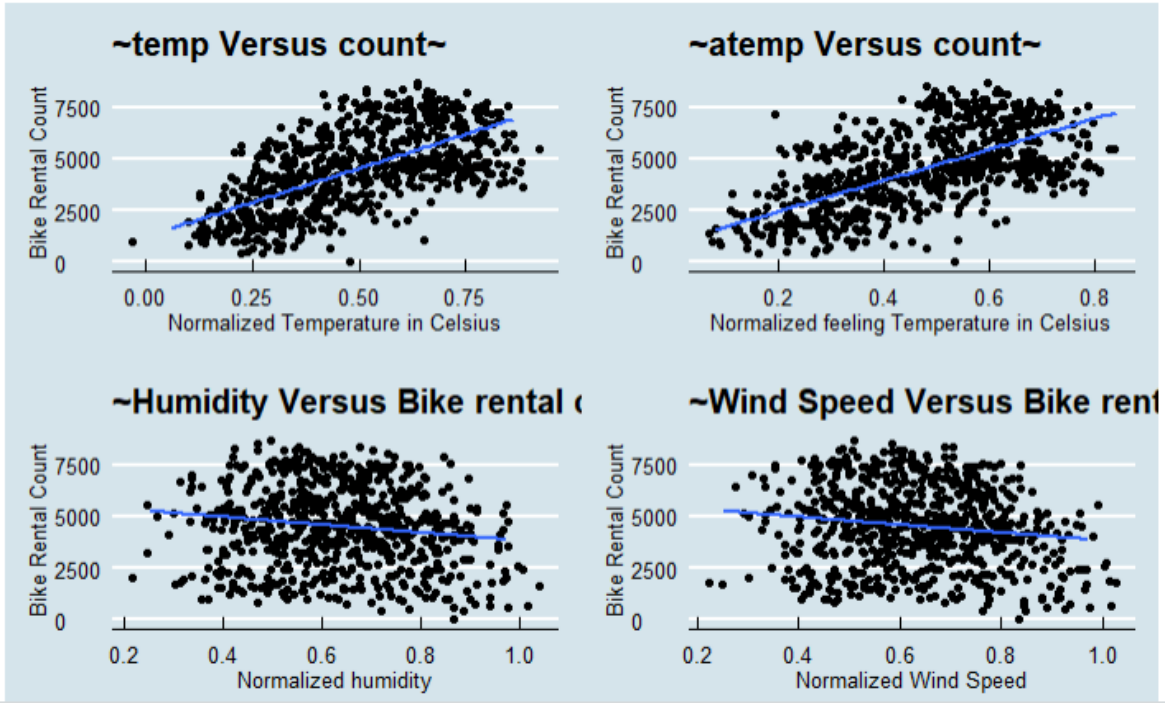


Fig.4 Boxplot for ouliers(python)

As we can see in Fig.4 that there are some outliers observed in windspeed and hum variables. We will further analyse these four variables for better understanding.

Fig.5 Temp, atemp, hum, windspeed against count



We can observe the distribution of temp, atemp, hum, windspeed variables and relation with the bike rental count. The slope that is blue line in each graph indicates the relationship with the bike rental count. Temp, atemp are directly proportional to the count and hum, windspeed are inversely proportional to the count of bike rental.

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
49	50	2011-02-19	Spring	2011	2	No holiday	6	No Working day	Clear	0.399167	0.391404	0.187917	0.507463	532	1103	1635
68	69	2011-03-10	Spring	2011	3	No holiday	4	Working day	Light rain, Cloudy	0.389091	0.385668	0.000000	0.261877	46	577	623

Fig.6 hum outliers

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
44	45	2011-02-14	Spring	2011	2	No holiday	1	Working day	Clear	0.415000	0.398350	0.375833	0.417908	208	1705	1913
93	94	2011-04-04	Summer	2011	4	No holiday	1	Working day	Clear	0.573333	0.542929	0.426250	0.385571	734	2381	3115
94	95	2011-04-05	Summer	2011	4	No holiday	2	Working day	Partly Cloudy, Mist	0.414167	0.398350	0.642083	0.388067	167	1628	1795
292	293	2011-10-20	Winter	2011	10	No holiday	4	Working day	Clear	0.475833	0.466525	0.636250	0.422275	471	3724	4195
382	383	2012-01-18	Spring	2012	1	No holiday	3	Working day	Clear	0.303333	0.275254	0.443333	0.415429	109	3267	3376
407	408	2012-02-12	Spring	2012	2	No holiday	0	No Working day	Clear	0.127500	0.101658	0.464583	0.409212	73	1456	1529
420	421	2012-02-25	Spring	2012	2	No holiday	6	No Working day	Clear	0.290833	0.255675	0.395833	0.421642	317	2415	2732
432	433	2012-03-08	Spring	2012	3	No holiday	4	Working day	Clear	0.527500	0.524604	0.567500	0.441563	486	4896	5382
433	434	2012-03-09	Spring	2012	3	No holiday	5	Working day	Partly Cloudy, Mist	0.410833	0.397083	0.407083	0.414800	447	4122	4569

Fig.7 Some Outliers of Windspeed

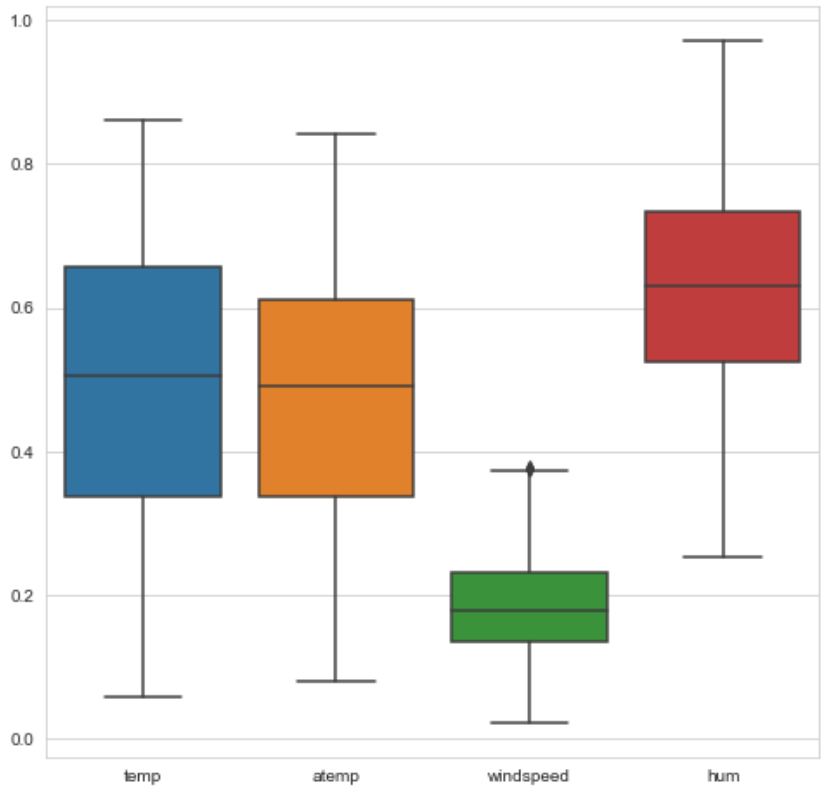


Fig.8 Boxplot after removing outliers

We can see the variation in boxplot graph after removing the outliers. It is important to detect, analyse and remove the outliers. As it will help in building robust model.

2.1.3 Multi Collinearity & Feature Selection

Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be *independent*. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

A key goal of regression analysis is to isolate the relationship between each independent variable and the dependent variable. The interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each 1 unit change in an independent variable when you *hold all of the other independent variables constant*.

When independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable *independently* because the independent variables tend to change in unison.

Here, I have used variance inflation factor (VIF) for a set of variables and exclude the highly correlated variables from the set through a stepwise procedure. This method can be used to deal with multicollinearity problems when you fit statistical models.

```
1 variables from the 7 input variables have collinearity problem:

atemp

After excluding the collinear variables, the linear correlation coefficients ranges between:
min correlation ( hum ~ instant ):  7.544427e-05
max correlation ( registered ~ instant ):  0.663711

---------- VIFs of the remained variables --------
   Variables      VIF
1    instant 2.207050
2       temp 2.254465
3        hum 1.249052
4  windspeed 1.128765
5     casual 1.589413
6 registered 3.152459
```

Fig.9 vifcor

We can observe that our data has multicollinearity problem. Variables temp, atemp are highly correlated. Thus they carry the same information so we will drop one variable that is atemp.

Further, I have implemented the correlation plot which is provided by library corrgram.

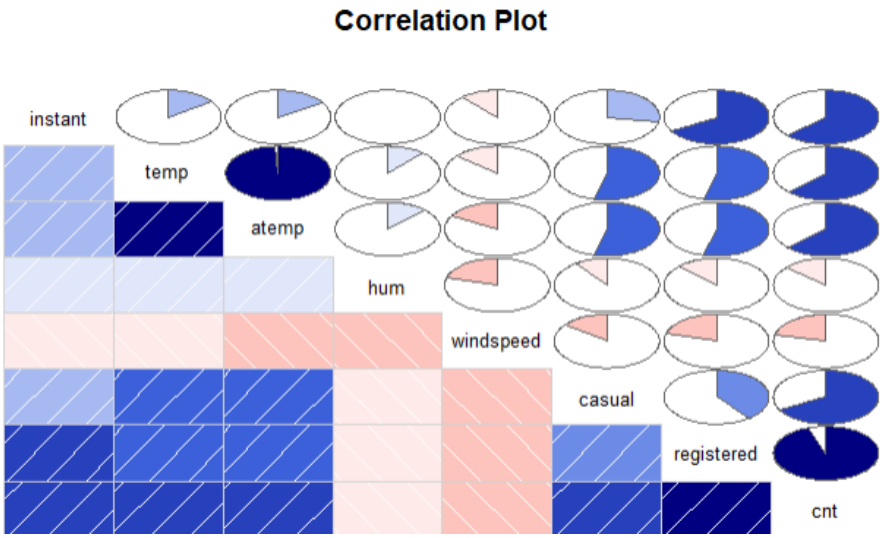


Fig.10 Correlation plot

From this plot, we can identify that temp, atemp are highly correlated that we have already analyised with the vifcor but there is one more variable hum that has not effect over our target variable cnt. So, hum is no use of us.

2.2 Modeling

2.2.1 Model Selection

First crucial step in modeling to know which algorithms we have to used based on our target variable.
The target variable can fall in either of the four categories:
1. Nominal
2. Ordinal
3. Interval
4. Ratio

The target variable in our model is a continuous variable i.e., Count of bike rentals(cnt). Hence the models that we choose are Linear Regression, Random Forest. The error metric chosen for the problem statement is Mean Absolute Error (MAE).

2.2.2 Multiple Linear Regression

Multiple linear regression is the most common form of linear regression analysis. Multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical.

```
Call:
lm(formula = cnt ~ . - casual - registered, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3984.3  -359.3    80.3   429.7  2921.4

Coefficients: (1 not defined because of singularities)
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               1873.51     270.36   6.930 1.20e-11 ***
seasonSummer                               804.06     195.52   4.112 4.52e-05 ***
seasonFall                                 758.95     231.66   3.276 0.001119 **
seasonWinter                              1425.88     197.84   7.207 1.91e-12 ***
yr2012                                    2003.53      64.96  30.843  < 2e-16 ***
mnth2                                      190.34     166.40   1.144 0.253187
mnth3                                      687.29     193.74   3.548 0.000422 ***
mnth4                                      777.21     278.15   2.794 0.005386 **
mnth5                                      930.33     300.60   3.095 0.002070 **
mnth6                                      717.56     317.97   2.257 0.024424 *
mnth7                                      321.22     348.16   0.923 0.356610
mnth8                                      679.56     337.56   2.013 0.044593 *
mnth9                                     1219.33     294.20   4.145 3.95e-05 ***
mnth10                                     795.15     272.60   2.917 0.003681 **
mnth11                                      90.41     255.06   0.354 0.723145
mnth12                                     109.79     203.99   0.538 0.590643
holidayHoliday                            -365.41     208.92  -1.749 0.080853 .
weekdayMonday                              199.74     122.51   1.630 0.103609
weekdayTuesday                             254.75     123.09   2.070 0.038957 *
weekdayWednesday                           334.16     121.28   2.755 0.006059 **
weekdayThursday                            327.83     121.24   2.704 0.007067 **
weekdayFriday                              415.03     120.25   3.451 0.000601 ***
weekdaySaturday                            368.79     118.22   3.119 0.001908 **
workingdayWorking Day                          NA         NA      NA       NA
weathersitMist, Cloudy                    -415.26      87.34  -4.754 2.55e-06 ***
weathersitLight Snow, Rain, Thunderstorm -1485.16     245.08  -6.060 2.54e-09 ***
temp                                      4069.12     474.07   8.583  < 2e-16 ***
hum                                      -1765.59     340.62  -5.183 3.07e-07 ***
windspeed                                -3260.74     485.57  -6.715 4.73e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 754.8 on 545 degrees of freedom
Multiple R-squared:  0.8494,    Adjusted R-squared:  0.8419
F-statistic: 113.8 on 27 and 545 DF,  p-value: < 2.2e-16
```

Fig.11 lm model

As you can see the Adjusted R-squared value, we can explain 84.19% of the data using our multiple linear regression model. By looking at the F-statistic and combined p-value we can reject the null hypothesis that target variable does not depend on any of the predictor variables.

2.2.3 Random Forest

Random forest is like bootstrapping algorithm with Decision tree (CART) model. Random forest tries to build multiple CART models with different samples and different initial variables. Here, we have given ntree value 500.

	%IncMSE	IncNodePurity
season	7.582560	15531874.8
yr	17.132960	109124156.8
mnth	15.020004	39391231.4
holiday	-1.626002	350910.5
weekday	14.054444	12379903.2
workingday	13.626769	11350171.2
weathersit	2.874658	1106110.5
temp	14.117907	115224150.3
hum	5.111791	5993490.3
windspeed	3.727803	3832883.7
casual	45.007948	441364140.7
registered	81.007983	1289956128.8

Fig.12 varimp

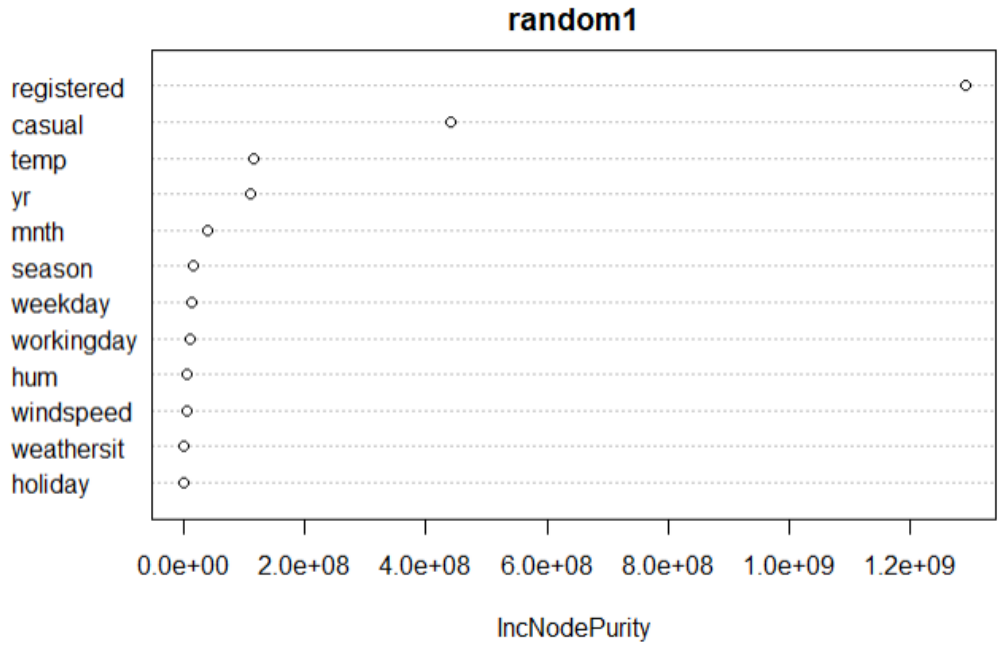


Fig.13 varimp plot

We will remove some variables that have low importance according to Fig.12 and Fig.13. Variables holiday, weathersit, windspeed have low importance. We will repeat this process again by tuning the parameters and take the final result of this model.

Chapter3

Conclusion

3.1 Model Evaluation

Now that we have a few models for predicting the target variable, we need to decide which one to choose.
There are several criteria that exist for evaluating and comparing models. We can compare the models using
any of the following criteria:
1. Predictive Performance
2. Interpretability
3. Computational Efficiency

Mean Absolute Percent Error(MAPE)

The mean absolute percentage error (MAPE) is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a loss function for regression problems in machine learning.

For linear regression: 0.6761
For Random Forest: 0.0330

Root Mean Square Error(RMSE)

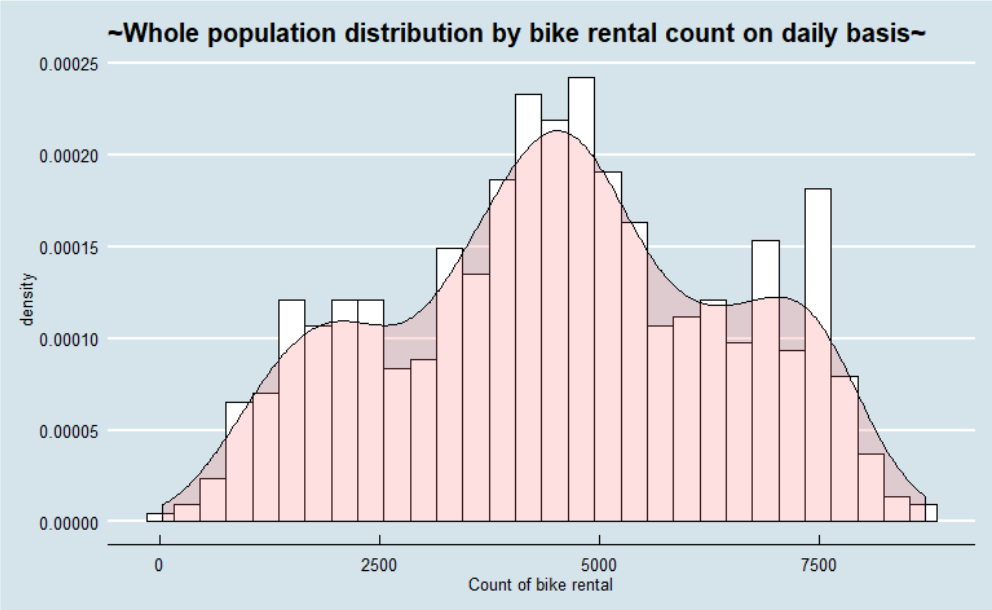
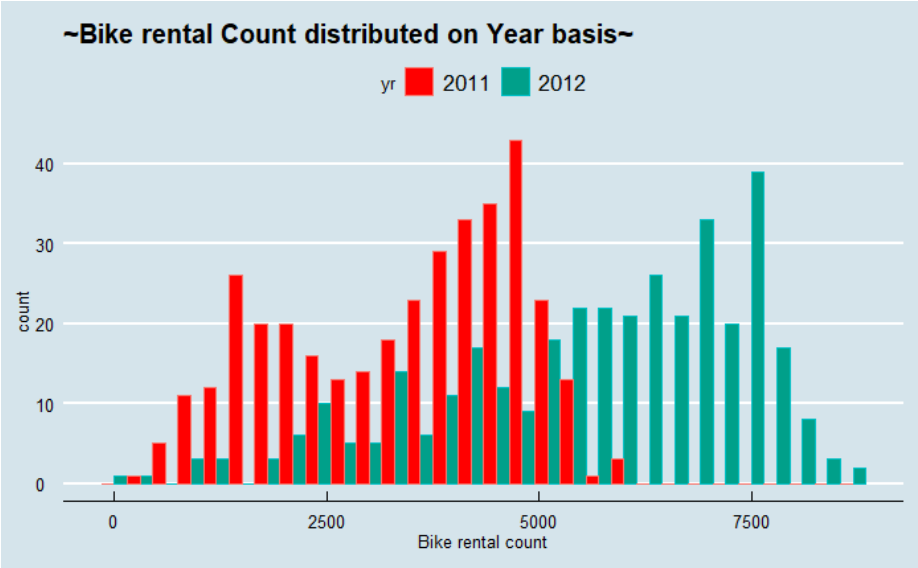
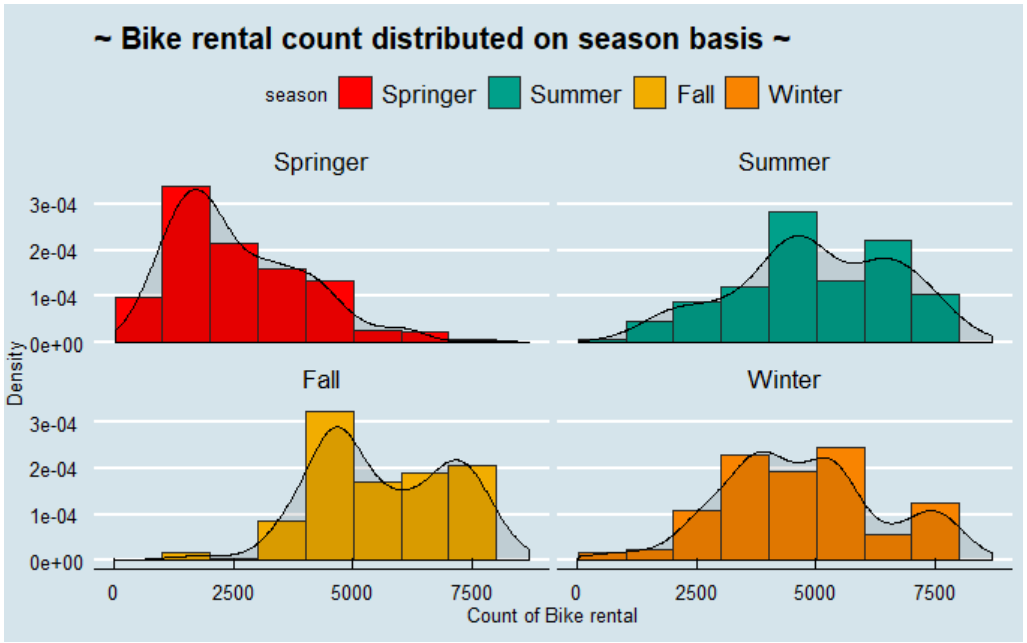
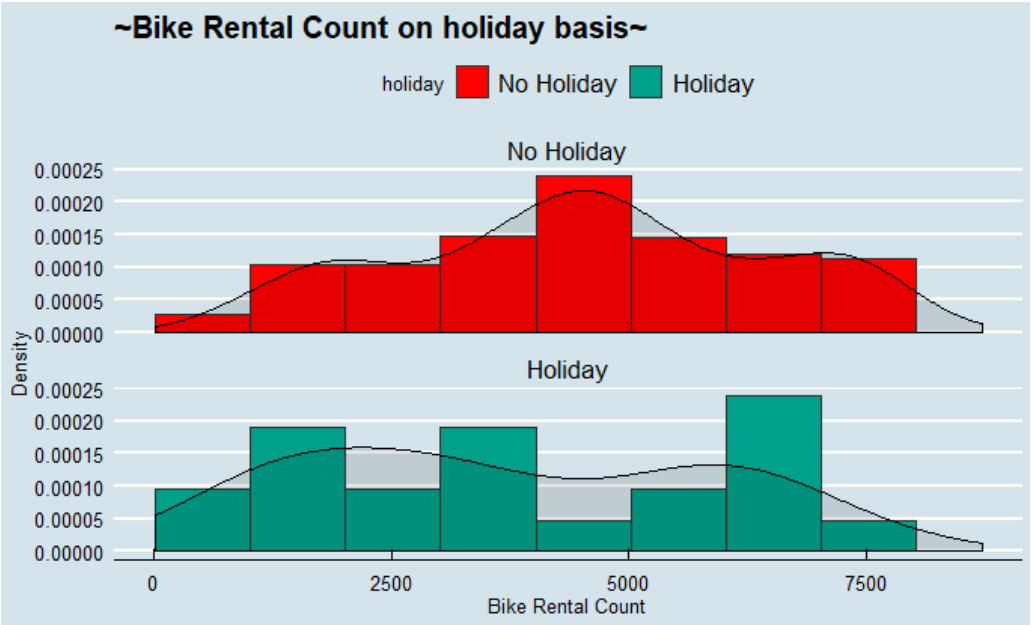
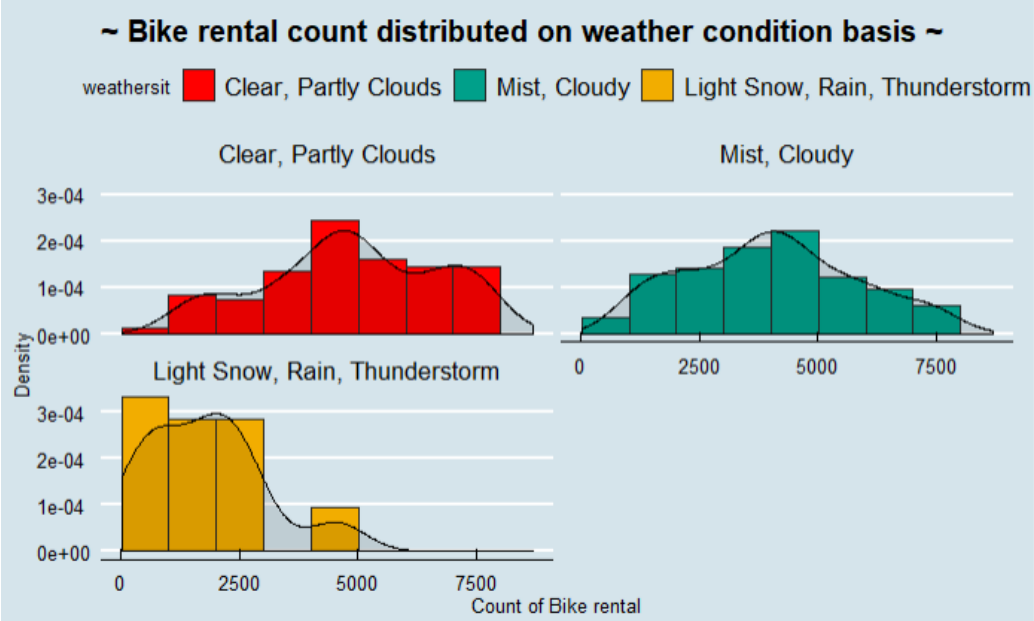
RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

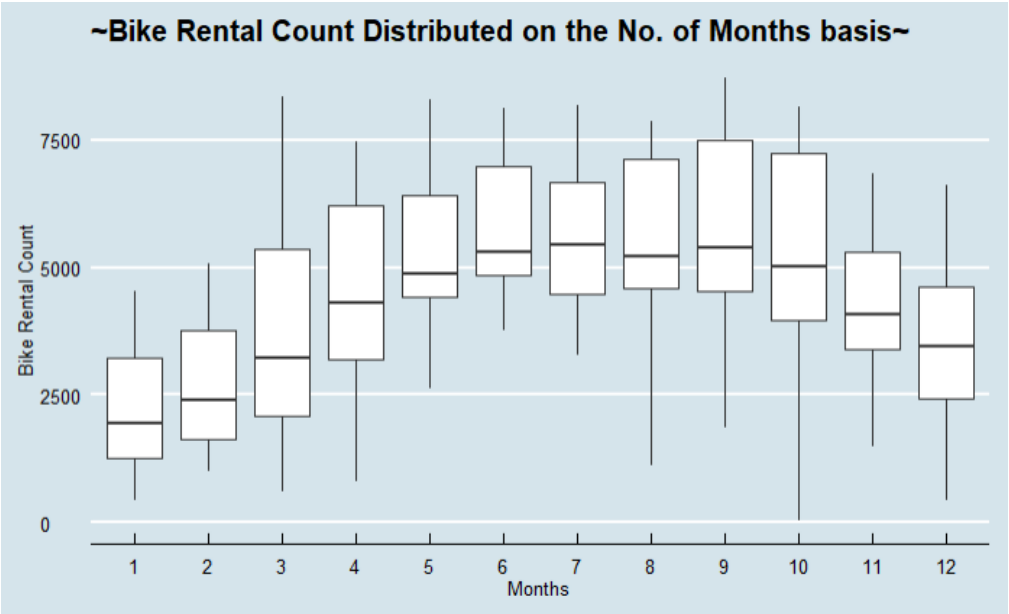
For linear regression: 2697.235
For Random Forest: 114.25

We can clearly observe that linear regression has performed poorly. So, Random forest fits well here.

Appendix A – R Code

Density Histogram of Bike rental count:

```
```{r echo=FALSE,warning=FALSE,message=FALSE,fig.width=10}
ggplot(new_bike,aes(cnt)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white")+
  geom_density(alpha=.2, fill="#FF6666") +
  theme_economist() +
  xlab("Count of bike rental")+
  ggtitle("~Whole population distribution by bike rental count on daily basis~")
```
```



Histogram of year against cnt:

```
```{r echo=FALSE,warning=FALSE,message=FALSE}
ggplot(new_bike, aes(x=cnt, color=yr,fill=yr)) +
  geom_histogram( position="dodge")+
  theme_economist() +
  scale_fill_manual(values=wes_palette(n=2,name="Darjeeling1"))+
  xlab("Bike rental count")+
  ggtitle("~Bike rental Count distributed on Year basis~")
```
```



Histogram of season against cnt:

```
```{r echo=FALSE,warning=FALSE,message=FALSE}
ggplot(new_bike,aes(cnt, fill=season)) +
  geom_histogram(aes(y=..density..),breaks=seq(20, 8800, by=1000), color="grey17") +
  geom_density(alpha=.1, fill="black")+
  facet_wrap(~season, ncol=2,scale="fixed") +
  theme_economist()  +
  scale_fill_manual(values=wes_palette(n=4,name="Darjeeling1")) +
   ylab("Density")+
  xlab("Count of Bike rental")+
  ggtitle("~ Bike rental count distributed on season basis ~")
```
```



Histogram of Holiday against cnt:

```
```{r echo=FALSE,warning=FALSE,message=FALSE}
ggplot(new_bike,aes(cnt,fill=holiday)) +
  geom_histogram(aes(y=..density..),breaks=seq(20,8800,by=1000),color="grey20")+
  geom_density(alpha=.1,fill="black")+
  facet_wrap(~holiday,ncol=1,scale="fixed")+
  theme_economist()+
  scale_fill_manual(values=wes_palette(n=2,name="Darjeeling1"))+
  ylab("Density")+
  xlab("Bike Rental Count")+
  ggtitle("~Bike Rental Count on holiday basis~")
```
```



Histogram of weathersit against cnt:

```
```{r echo=FALSE,warning=FALSE,message=FALSE}
ggplot(new_bike,aes(cnt, fill=weathersit)) +
  geom_histogram(aes(y=..density..),breaks=seq(20, 8800, by=1000), color="grey17") +
  geom_density(alpha=.1, fill="black")+
  facet_wrap(~weathersit, ncol=2,scale="fixed") +
  theme_economist()  +
  scale_fill_manual(values=wes_palette(n=4,name="Darjeeling1")) +
   ylab("Density")+
  xlab("Count of Bike rental")+
  ggtitle("~ Bike rental count distributed on weather condition basis ~")
```
```



Histogram of mnth against cnt:

```
```{r echo=FALSE,warning=FALSE,message=FALSE}
ggplot(new_bike, aes(x = mnth, y = cnt)) + geom_boxplot() +
  theme_economist() +
  xlab("Months")+
  ylab("Bike Rental Count")+
  ggtitle("~Bike Rental Count Distributed on the No. of Months basis~")
```
```



References

Book: R CookBook by Paul Teetor
Python for Data Analysis by Wes McKinney