Bike Renting Prediction

Venkata Subramanian R

03 January 2019

**Table of Contents:**

1. **Introduction**
   1. **Problem Statement-------------------------------------------------------3**
   2. **Data----------------------------------------------------------------------------3**
2. **Methodology**
   1. **Pre-Processing-------------------------------------------------------------5**
      1. **Missing Value Analysis----------------------------------------------5**
      2. **Outlier Analysis--------------------------------------------------------5**
      3. **Correlation Plot--------------------------------------------------------8**
      4. **Anova Testing----------------------------------------------------------9**
      5. **Feature Selection----------------------------------------------------11**
      6. **Modeling----------------------------------------------------------------12**
3. **Conclusion**
   1. **Model Evaluation--------------------------------------------------------14**
   2. **Model Selection----------------------------------------------------------14**
4. **Appendix---------------------------------------------------------------------------15**

**Chapter 1**

**Introduction**

**1.1 Problem Statement**

The objective of this project is to predict the bike rental count on daily based on the environmental and seasonal settings. Analyze the data and develop a model to predict the behavioural changes with environmental and seasonal settings.

**1.2 Data**

Our task is to build a regression model which will find the Count of the daily users of the bike for rental purpose. Below are some sample data which will predict the Bike Rental Count.

Table 1.1: Bike Rental Sample Data (Columns 1-9)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| instant | dteday | season | yr | mnth | holiday | weekday | workingday | weathersit |
| 1 | 1/1/2011 | 1 | 0 | 1 | 0 | 6 | 0 | 2 |
| 2 | 1/2/2011 | 1 | 0 | 1 | 0 | 0 | 0 | 2 |
| 3 | 1/3/2011 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| 4 | 1/4/2011 | 1 | 0 | 1 | 0 | 2 | 1 | 1 |
| 5 | 1/5/2011 | 1 | 0 | 1 | 0 | 3 | 1 | 1 |
| 6 | 1/6/2011 | 1 | 0 | 1 | 0 | 4 | 1 | 1 |
| 7 | 1/7/2011 | 1 | 0 | 1 | 0 | 5 | 1 | 2 |
| 8 | 1/8/2011 | 1 | 0 | 1 | 0 | 6 | 0 | 2 |
| 9 | 1/9/2011 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |

Table 1.2 Bike Rental Sample Data (Columns 10-16)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| temp | atemp | hum | windspeed | casual | registered | cnt |
| 0.344167 | 0.363625 | 0.805833 | 0.160446 | 331 | 654 | 985 |
| 0.363478 | 0.353739 | 0.696087 | 0.248539 | 131 | 670 | 801 |
| 0.196364 | 0.189405 | 0.437273 | 0.248309 | 120 | 1229 | 1349 |
| 0.2 | 0.212122 | 0.590435 | 0.160296 | 108 | 1454 | 1562 |
| 0.226957 | 0.22927 | 0.436957 | 0.1869 | 82 | 1518 | 1600 |
| 0.204348 | 0.233209 | 0.518261 | 0.089565 | 88 | 1518 | 1606 |
| 0.196522 | 0.208839 | 0.498696 | 0.168726 | 148 | 1362 | 1510 |
| 0.165 | 0.162254 | 0.535833 | 0.266804 | 68 | 891 | 959 |
| 0.138333 | 0.116175 | 0.434167 | 0.36195 | 54 | 768 | 822 |

**Chapter 2**

**Methodology**

**2.1** **Pre Processing**

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis.**

To Start with, we will compute the Missing Values

**2.1.1 Missing Value Analysis**

In the given data set, there are no missing values found. So no need to perform any imputation methods analysis.

**2.1.2 Outlier Analysis**

We have found that there are no missing values to impute. Now as the next step in Data Pre-Processing we need to detect the outliers that is present in the data set and remove it.

We can detect the outliers in the data set using Box Plot Method. In the below figure we have plotted the outliers in the employee absenteeism data set using Boxplot.

Fig 2.1

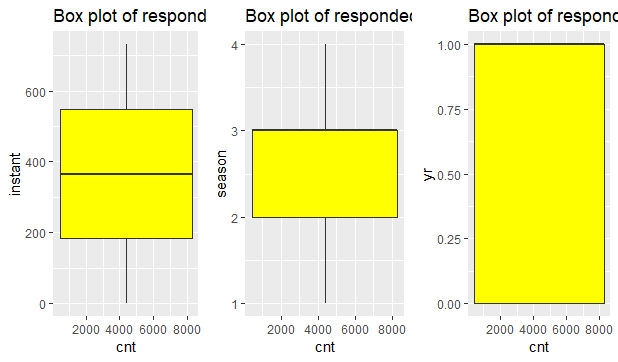


Fig 2.2

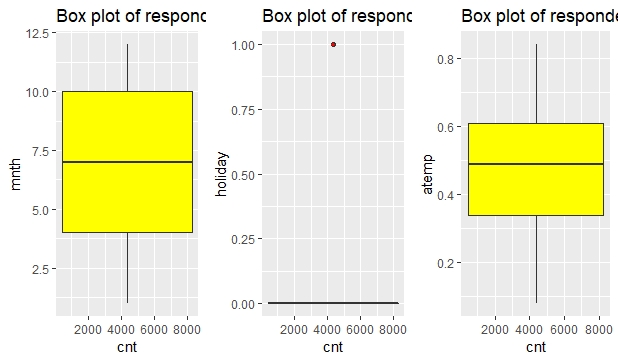


Fig 2.3

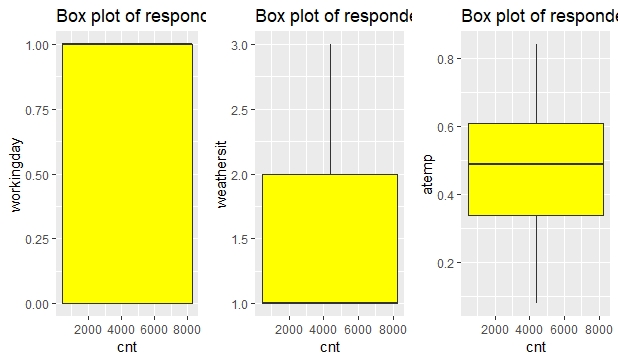
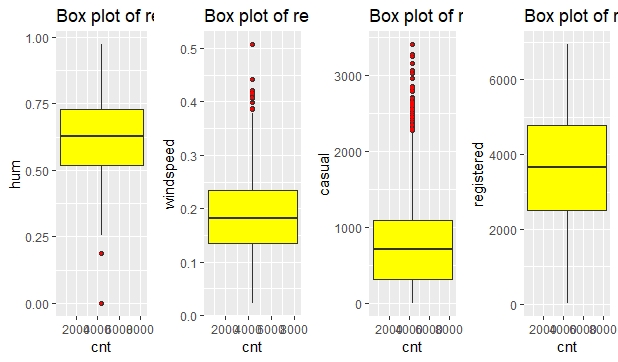


Fig 2.4



From the Figure 2.1,2.2,2.3 and 2.4 we can infer that we have some outliers in hum, windspeed and casual count variables.

We can remove the values of these variables and replace them with NA.

Once replaced with NA, we can impute the missing value using KNN Imputation Method

Now with this Pre-Processing Technique, we have removed the outliers from the data set.

After computing the missing value using KNN Imputation, we can see in Fig 2.5 that almost 90% of the outliers values are removed.

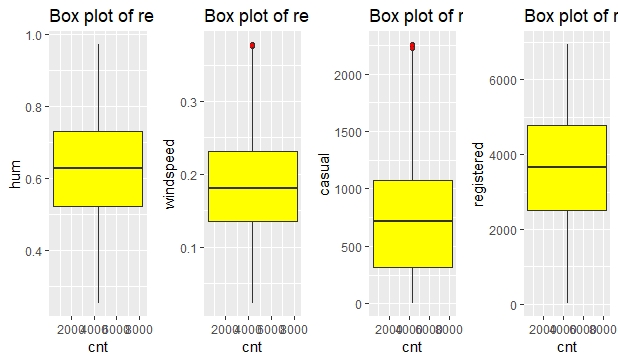


Fig 2.5

**2.1.3 Correlation Plot**

As the next step in Pre-Processing we need to find the multi- collinearity between the variables. If two variables contain the same information, then we can discard one variable while passing it to the model. So, with the help of Correlation Plot we can find the multi collinearity between the continuous variables.

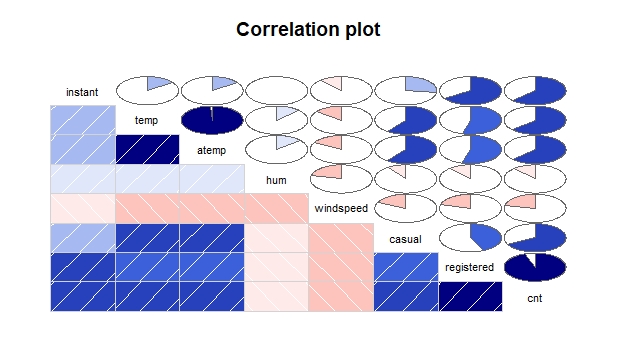


Fig 2.6

From Figure 2.6, we can infer that the variables temp and atemp are highly positively correlated with each other and Instant Variable is also highly correlated with few variables. As per our assumptions two independent variables should have less dependency between each other and high dependency to the tatget variable.

So, we can drop atemp from the dataset. Since Instant variable contains only the row number we can drop that variable from the data set. Also seeing the correlation plot, the variable hum contains much less information for the target variable. So we can drop that variable too.

From Correlation Plot Analysis we have dropped three variables **atemp,Instant and hum** from the data set.

**2.1.4 Anova Testing**

We have done Collinearity check between Continuous variables. Now we need to find the dependency between Target Variable(Continuous) and Categorical Variables. We can use Anova Testing to determine the dependency between these variables.

When executed the Anova Testing, we got the following observations.

[1] "season"

Df Sum Sq Mean Sq F value Pr(>F)

factor\_data[, i] 3 9.506e+08 316865289 128.8 <2e-16 \*\*\*

Residuals 727 1.789e+09 2460715

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

[1] "yr"

Df Sum Sq Mean Sq F value Pr(>F)

factor\_data[, i] 1 8.798e+08 879828893 344.9 <2e-16 \*\*\*

Residuals 729 1.860e+09 2551038

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

[1] "mnth"

Df Sum Sq Mean Sq F value Pr(>F)

factor\_data[, i] 11 1.070e+09 97290206 41.9 <2e-16 \*\*\*

Residuals 719 1.669e+09 2321757

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

[1] "holiday"

Df Sum Sq Mean Sq F value Pr(>F)

factor\_data[, i] 1 1.280e+07 12797494 3.421 0.0648 .

Residuals 729 2.727e+09 3740381

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

[1] "weekday"

Df Sum Sq Mean Sq F value Pr(>F)

factor\_data[, i] 6 1.766e+07 2943170 0.783 0.583

Residuals 724 2.722e+09 3759498

[1] "workingday"

Df Sum Sq Mean Sq F value Pr(>F)

factor\_data[, i] 1 1.025e+07 10246038 2.737 0.0985 .

Residuals 729 2.729e+09 3743881

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

[1] "weathersit"

Df Sum Sq Mean Sq F value Pr(>F)

factor\_data[, i] 2 2.716e+08 135822286 40.07 <2e-16 \*\*\*

Residuals 728 2.468e+09 3389960

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Based on the above observations, we need to remove the variables whose p-value is >0.05 since our significance value is 0.05.

On Interpreting the results, we can remove the variables **holiday, weekday and workingday** from our data set.

**2.1.5 Feature Selection**

Before Performing any type of modeling, we need to access the importance of each predictor variables in our data sets. There is every chance that all variables in our analysis set may not be as important as all in the problem of class prediction. Below we have used Random Forest method to Perform Feature Selection and got the following observations.

%IncMSE

dteday 0.000000

season 12.428415

yr 14.835223

mnth 18.867626

weathersit 15.108630

temp 23.601904

windspeed 7.220629

casual 36.577292

registered 50.056438

From the above information we can infer that **Casual and Registered** users variables are the most important predictors in our data set.

The variables **dteday,windspeed** contain much less information for deriving target variable and hence can be dropped from the data set. So we can do dimension reduction and drop these variables.

Since our Prediction is there should be less dependency between the independent variables in a dataset.

So, we need to find the correlation between the independent variables. We can have the threshold value as 0.8-0.9.

vifcor(bike\_data[,-10],th=0.8)

1 variables from the 9 input variables have collinearity problem:

season

After excluding the collinear variables, the linear correlation coefficients ranges between:

min correlation ( temp ~ weekday ): -0.0001699624

max correlation ( casual ~ temp ): 0.6007518

---------- VIFs of the remained variables --------

Variables VIF

1 yr 2.440600

2 mnth 1.276269

3 weekday 1.018092

4 workingday 2.925689

5 weathersit 1.248503

6 temp 2.498084

7 casual 3.807222

8 registered 5.211081

From the above observation we can find that season variable has higher threshold value. So, we can drop the variable from the data set.

**2.6 Modeling**

Now we have done Pre-Processing Techniques on the data set and our data is ready to develop a model.

**2.2.1 Model Selection**

We need to Predict the Count Variable which is a continuous variable. For dependent variable we have 4 categories:

* Nominal
* Ordinal
* Interval
* Ratio

If Nominal it will come under classification model. If Interval or Ratio it will come under regression model. Since our variable is a continuous one it will come under Regression Model.

So, we will start from the basic model and try to predict the Target variable.

We will first use Linear Regression Model to Predict the value of Target Variable

Summary of the Linear Regression Model is as below.

Call:

lm(formula = cnt ~ ., data = train\_data)

Residuals:

Min 1Q Median 3Q Max

-372.95 -120.47 -36.30 26.51 2076.23

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -163.27213 60.65950 -2.692 0.00732 \*\*

yr 112.07324 35.00346 3.202 0.00144 \*\*

mnth 0.24118 3.71162 0.065 0.94821

weathersit 0.99596 24.18092 0.041 0.96716

temp 51.12720 99.20814 0.515 0.60650

casual 1.15670 0.02923 39.577 < 2e-16 \*\*\*

registered 0.97792 0.01390 70.374 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 284.2 on 577 degrees of freedom

Multiple R-squared: 0.979, Adjusted R-squared: 0.9788

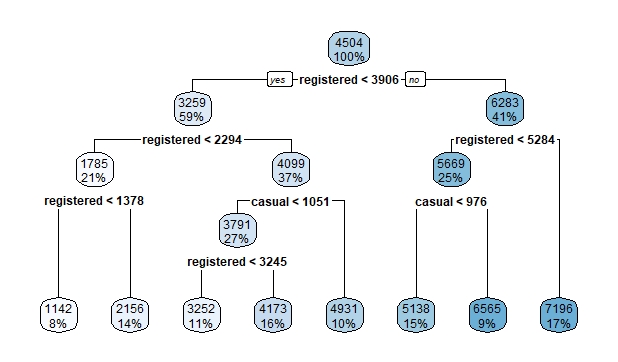
F-statistic: 4484 on 6 and 577 DF, p-value: < 2.2e-16

From the above observation on seeing the value of *Adjusted R-Squared,* we can see that the model can explain 97.8% of the data. From the above model summary, we can find the variables which carry more information to derive the target variable.

This Accuracy is much higher and this model is better to be choosed for predicting the target variables.

**Decision Tree Plot**

Fig 2.7



**Conclusion**

**3.1 Model Evaluation**

Now that we have developed a few models for predicting our target variable and we have to decide which one to choose. We will evaluate some error metrics and decide which one to choose.

We can evaluate the Predictive Performance of the model with real values of the target variables and calculating some average error measures.

**3.1.1 Mean Absolute Postive Error (MAPE)**

As we know that MSE and RMSE we need to use for Time series or a transactional data. Since here we are finding the count of Bike users. We will go with MAPE. MAPE will give the percentage of error in the test data.

So let us Calculate the MAPE for both Regression Model and Decision Tree Model.

**Linear Regression Model:**

For Linear Regression Model we get MAPE as **2.376633e-02**

**Decision Tree Model:**

For Decision Tree Model we get MAPE as **6.387914e-01**

For both the models MAPE value is very much less.

**3.2 Model Selection**

Based on the Above Observations, we can see that Linear Regression Model is more suitable since the error value is less and Accuracy is also very high. So we can freeze Linear Regression Model for our Data set.

**Appendix A - Extra Figures**

