Prediction of Bike Rental

Himanshu Bisht

Contents

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

Our task is to build regression models which will predict the bike rental count on daily based on the environmental and seasonal settings. Given below is the sample of the dataset that we are using to predict bike rental.

Number of attributes:

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)
hr: Hour (0 to 23)
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

Chapter 2

Methodology

2.1.0 Pre Processing

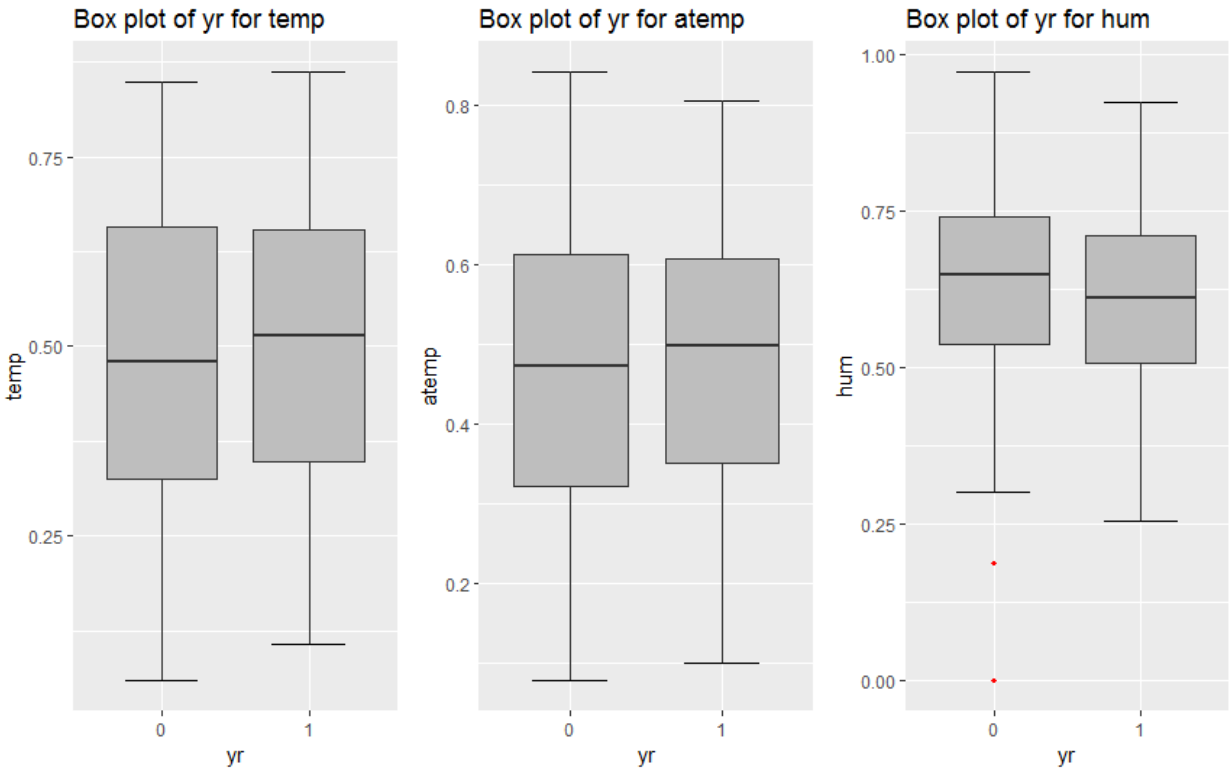
Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

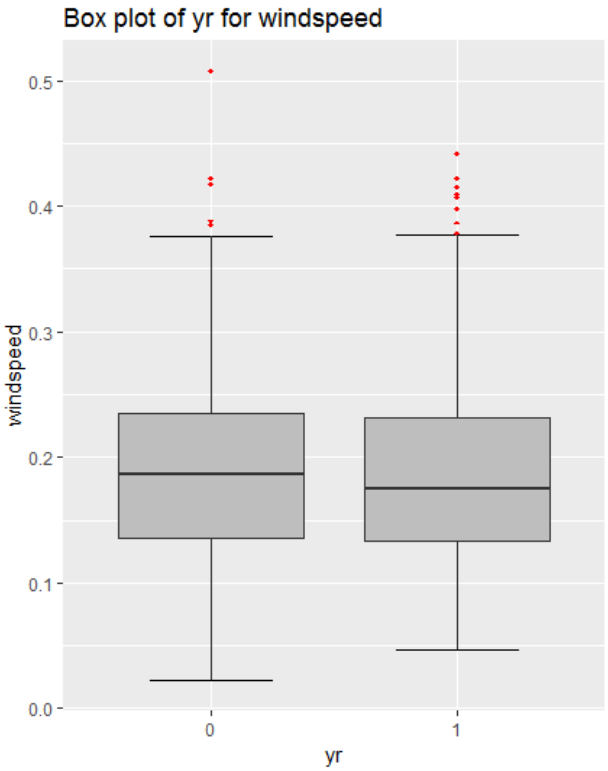
In Real world data are generally incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. Noisy: containing errors or outliers. Inconsistent: containing discrepancies in codes or names.

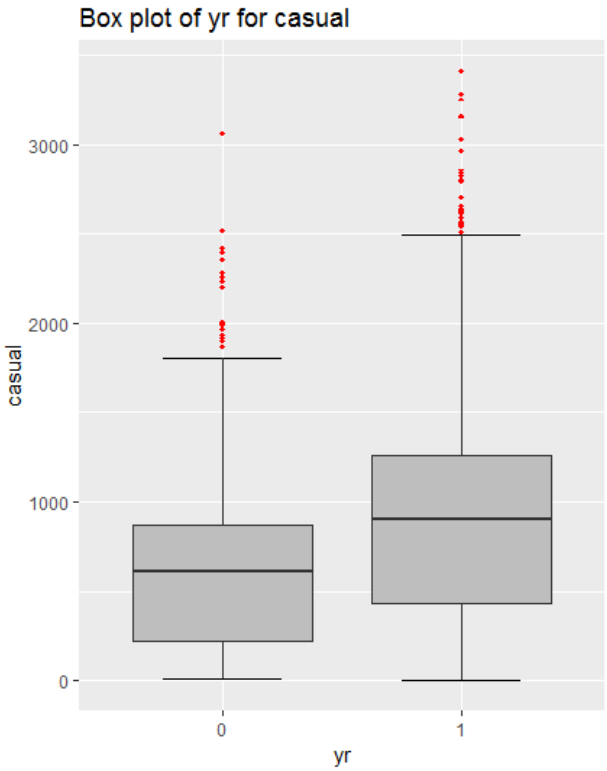
2.1.1 Outlier Analysis

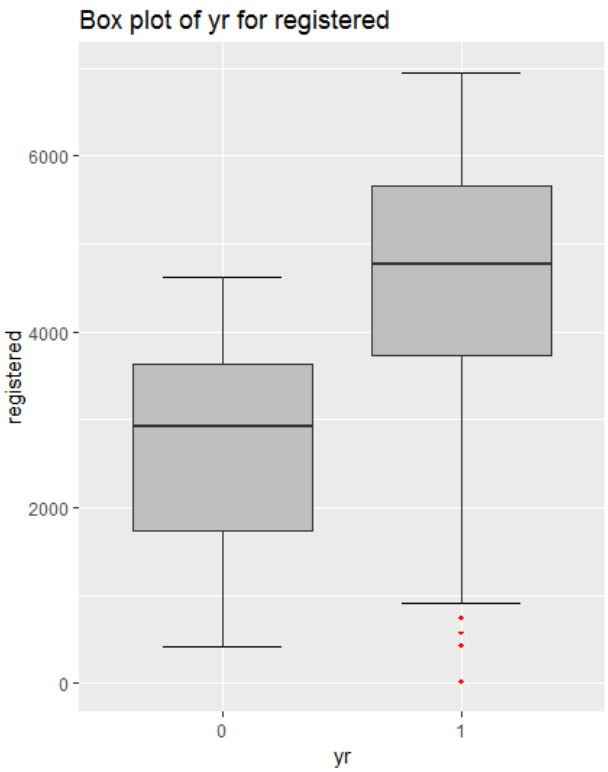
An *Outlier* is a rare chance of occurrence within a given data set. In Data Science, an *Outlier* is an observation point that is distant from other observations. An *Outlier* may be due to variability in the measurement or it may indicate experimental error.

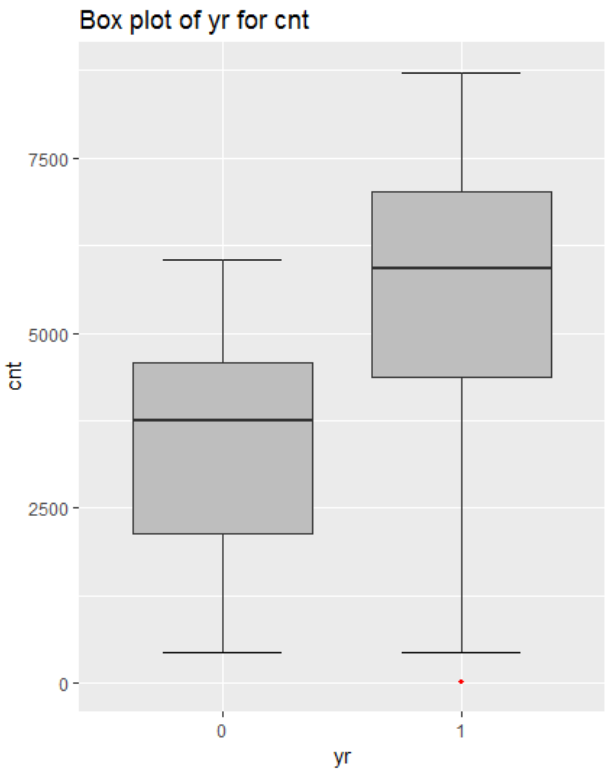
Now below are the boxplot of all the continous variables.




Box plot of yr for windspeed
windspeed
0.5
0.4
0.3
0.2
0.1
0.0
0
1
yr


Box plot of yr for casual
casual
3000
2000
1000
0
0
1
yr


Box plot of yr for registered
registered
6000
4000
2000
0
0
1
yr


Box plot of yr for cnt
cnt
7500
5000
2500
0
0
1
yr

NOTE – After this I have delete all the observations which are containing the Outliers from the dataset.

2.1.2 Missing Value Analysis

The concept of missing values is important to understand in order to successfully manage data. If the missing values are not handled properly by the researcher, then he/she may end up drawing an inaccurate inference about the data. Due to improper handling, the result obtained by the researcher will differ from ones where the missing values are present.

There are three methods which are mostly adopted to treat the missing values

1. Mean method
2. Median method
3. KNN method

Out of these three methods that method will be adopted for imputing values which will provide nearest value to the missing value.

NOTE – In my project after deleting the outliers from the dataset there is no missing value present in the dataset df.

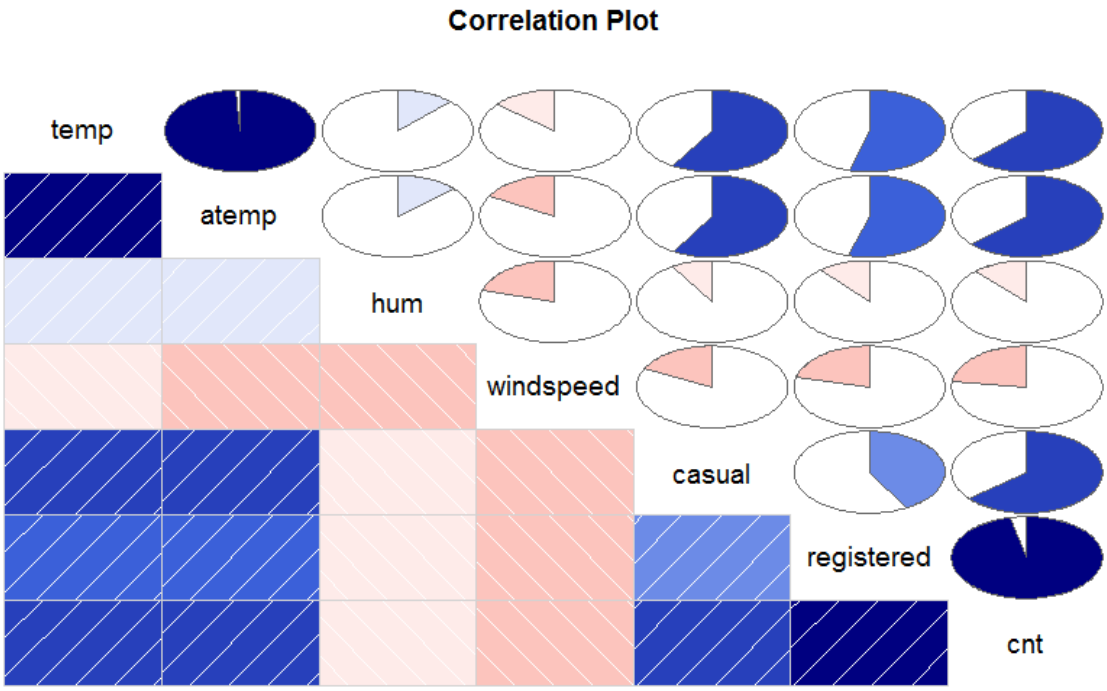
2.1.3 Feature Selection

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

NOTE – 1.In my project in case of continuous variables I am applying the Co relation plot which shows that the variables temp and atemp are highly co related to each other and also the variables registered and cnt are highly co related and variable hum is not carrying much information about the target variable.

So I am removing the variable hum and atemp form my dataset df.



2. In case of categorical variables I have used Chi Square test to check the Dependencies of variables with each other. From Chi Square test the following variables are co related with each other ;

season , weathersit, mnth
#weekday , workingday, holiday

So I have removed the mnth,weathersit,workingday,holiday,detday,yr variables from my dataset.

2.1.4 Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Feature scaling can vary your results a lot while using certain algorithms and have a minimal or no effect in others.

NOTE – 1.In my project I have used Normalisation which will convert the values of variable from 0 to 1.
2. Standardisation will only be use in case of uniformly distribution.
3. I have applied normalization only on casual, registered and cnt variables.

2.2 Modeling

2.2.0 Model Selection

Model selection is the process of choosing between different machine learning approaches - e.g. SVM, logistic regression, etc. - or choosing between different hyper parameters or sets of features for the same machine learning approach - e.g. deciding between the polynomial degrees/complexities for linear regression.

According to my problem statement i.e to predict the bike rental, as we can see it is clearly falls under the regression type. So for this I have used mainly three algorithm below:

1. Linear Regression
2. Decision Tree
3. Random Forest

2.2.1 Linear Regression

It is basically a statistical model. In this model instead of patterns we are dealing with weight, coefficient or numbers of each independent variable. It is only used for regression.

```
Residuals:
       Min         1Q     Median         3Q        Max
-1.276e-14 -5.160e-17  2.380e-17  1.025e-16  7.984e-16

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) -8.868e-16  1.448e-16 -6.125e+00 1.81e-09 ***
season2     -3.085e-16  9.911e-17 -3.113e+00  0.00196 **
season3     -2.290e-16  1.297e-16 -1.766e+00  0.07804 .
season4     -2.591e-16  8.651e-17 -2.995e+00  0.00288 **
weekday1    -6.846e-17  1.194e-16 -5.730e-01  0.56658
weekday2    -2.066e-16  1.272e-16 -1.624e+00  0.10491
weekday3    -2.324e-16  1.298e-16 -1.791e+00  0.07393 .
weekday4    -1.753e-16  1.252e-16 -1.400e+00  0.16198
weekday5    -1.535e-18  1.190e-16 -1.300e-02  0.98971
weekday6    -1.153e-17  1.125e-16 -1.030e-01  0.91840
temp        -4.114e-17  2.948e-16 -1.400e-01  0.88905
windspeed    6.736e-16  3.885e-16  1.734e+00  0.08353 .
casual       2.613e-01  2.255e-16  1.159e+15  < 2e-16 ***
registered   8.497e-01  1.810e-16  4.694e+15  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.087e-16 on 510 degrees of freedom
Multiple R-squared:      1,    Adjusted R-squared:      1
F-statistic: 5.534e+30 on 13 and 510 DF,  p-value: < 2.2e-16
```

NOTE – Out of all the variables season , casual , registered are important as their P values are less than 0.05 which means reject the null hypothesis.

2.2.2 Decision Tree

A Decision Tree is an algorithm used for supervised learning problems such as classification or regression. A decision tree or a classification tree is a tree in which each internal (nonleaf) node is labeled with an input feature.

```
library(rpart)
fit = rpart(cnt ~., data =train, method = "anova")
```

2.2.3 Random Forest

Random forest algorithm can also be used for problems such as classification and regression. The idea behind this algorithm is to bulid n no of trees to have more accuracy on dataset.

```
randomForest(formula = cnt ~ ., data = train, importance = TRUE,      ntree =
500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.0009051195
                    % Var explained: 98.22
```

3 Conclusion

3.1 Model Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. There are different types of error matrices which are used to evaluate the model. So we can use two error matrices in this project:

1. MAE(Mean Absolute Error)
2. MAPE(Mean Absolute Percentage Error)
3. RMSE(Root Mean Squared Error)
   Used for time series analysis
4. MSE(Mean Squared Error)

3.1.1 Mean Absolute Percentage Error (MAPE)

For this project I have used MAPE because it will provide error in the form of percentage which can be easy to understand and to evaluate the model performance.

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

where A_t is the actual value and F_t is the predicted value. The difference between A_t and F_t is divided by the actual value A_t again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n. Multiplying by 100% makes it a percentage error.

IMPORTANT POINTS

1. Now the errors in Linear regression model are

mae	mse	rmse	mape
4.826211e-16	2.960975e-31	5.441484e-16	1.656079e-15

2. Now the errors in Decision tree model are

mae	mse	rmse	mape
0.038216183	0.002398237	0.048971801	0.109472107

3. Now the errors from the random forest model are

```
        mae          mse         rmse         mape
0.0215441781 0.0009400145 0.0306596560 0.0709796807
```

CHECKING ACCURACY

a. Linear Regression model

   MAPE = 1.656079e-15

   Accuracy =

b. Decision tree model

   MAPE = 10.94

   Accuracy = 89.06

c. Random forest model

MAPE = 7.09

Accuracy = 92.91

4.1.2 Final Model Selection

Depending upon error and the accuracy we have to select a model for future data prediction. So for my project I will go with Linear Regression model.

4 Model Deployment

The concept of model deployment in data science refers to the application of a model for prediction using new data. Building a model is not generally the end the project. The model needs to be organized and presented in such a way that a customer can use it. Different methods to deploy a model in production system are following:

1. ONLINE METHOD

a. Data mining tools
-Revo Deploy R
-Orange
These tools will provide a framework or UI through which we can deploy our model at the cloud level. Based on these tools it will give UI which a non technical client can use and generate output out of it.

b. Programming languages such as JAVA, C, VB
Whenever we choose any environment such as R or Python then the library should be capable to sink with JAVA or C.

c. Database and SQL script such as TSQL, PL-SQL
Sinking the database with R or Python is also important.

d. PMML (Predictive Model Markup Language)

2. OFFLINE METHOD

Using Schedulers
In this the model should run every week and generate the output and send across to the mails.