Predicting Rental Bikes Count

-Deepak Dogra

Contents

Sno Name page no

1 Introduction 3

1.1 Problem Statement 3 1.2 Data 3

2 Methodology 5

2.1 Pre Processing 5

2.1.1 Outlier Analysis 6

2.1.2 Feature Selection 8

2.2 Modeling 14

2.2.1 Model Selection 14

2.2.2. Decision Tree 14

2.2.3 Linear Regression 16

2.2.4 Random Forest 18

3 Conclusion 19

3.2 Model selection 22

4 Reference 24

Chapter 1

Introduction

* 1. Problem Statement

The objective of this project is to predict the rental count of bike used by both registered and casual users on daily basis based on the seasonal and environmental factors.

* 1. Data

We are given a set of data containing details about various environmental parameters such as humidity, weather condition and also a particular day is a holiday or not etc . Using this details we have to prepare a data prediction model which can predict the count of bike that can go on rental on a particular day.Given below is a sample of data set that we will be using to train our model.

Table 1.1: Sample Data (Columns: 1-6)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| weekday | workingday | weathersit | temp | atemp | hum |
| 6 | 0 | 2 | 0.344167 | 0.363625 | 0.805833 |
| 0 | 0 | 2 | 0.363478 | 0.353739 | 0.696087 |
| 1 | 1 | 1 | 0.196364 | 0.189405 | 0.437273 |
| 2 | 1 | 1 | 0.2 | 0.212122 | 0.590435 |

Table 1.2: Sample Data (Columns: 7-12)

|  |  |  |  |
| --- | --- | --- | --- |
| Windspeed | casual | registered | cnt |
| 0.160446 | 331 | 654 | 985 |
| 0.248539 | 131 | 670 | 801 |
| 0.248309 | 120 | 1229 | 1349 |
| 0.160296 | 108 | 1454 | 1562 |

Table 1.3: Sample Data (Columns: 13-16)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| instant | dteday | season | yr | Mnth | holiday |
| 1 | 1/1/2011 | 1 | 0 | 1 | 0 |
| 2 | 1/2/2011 | 1 | 0 | 1 | 0 |
| 3 | 1/3/2011 | 1 | 0 | 1 | 0 |
| 4 | 1/4/2011 | 1 | 0 | 1 | 0 |

The observation regarding the dataset was as follows-

1)There are 12 independent variables namely-dteday,season,yr,mnth,holiday,weekday,workingday,weathersit,temp,atemp,humidity and wingspread.

2)There are three dependent variables namely- Casual, registered and cnt.

Chapter 2

Methodology

2.1 Pre Processing

Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

Various preprocessing techniques are applied to clean the data such as if there are missing values in the data then missing value analysis is done. Outlier analysis is done to remove the exceptional data sets from the data. Feature selection is done to remove the multicollinearity and to reduce the size of the dataset. Next step comes the feature scaling in which data is either normalized or standardized. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

2.1.1 Outlier Analysis

In statistics, 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; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses. Thus it is very important to remove an outlier from our data set. Outliers can have many anomalous causes. A physical apparatus for taking measurements may have suffered a transient malfunction. There may have been an error in data transmission or transcription. Outliers arise due to changes in system behavior, fraudulent behavior, human error, instrument error or simply through natural deviations in populations. A sample may have been contaminated with elements from outside the population being examined. Alternatively, an outlier could be the result of a flaw in the assumed theory, calling for further investigation by the researcher. Additionally, the pathological appearance of outliers of a certain form appears in a variety of datasets, indicating that the causative mechanism for the data might differ at the extreme end (King effect).

The first analysis to be done was outlier analysis on the data set and following box plot was plotted -

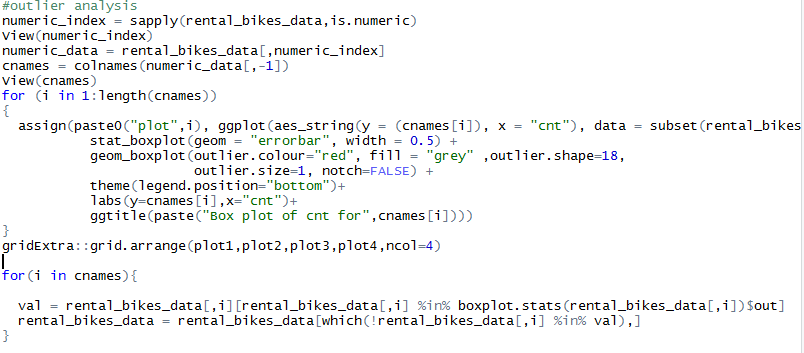
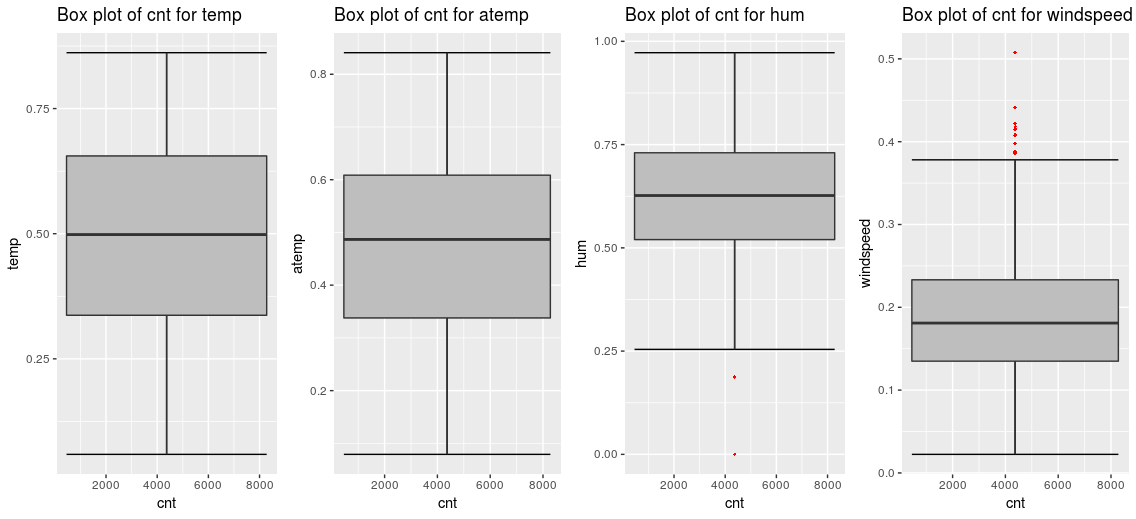


Fig 2.1 boxplots for the all 4 predictors



As clearly seen from the boxplots humidity and windspead has outliers.thus the outliers from both this variable are removed.

2.1.2 Feature Selection

The central premise when using a feature selection technique is that the data contains some features that are either *redundant* or *irrelevant*, and can thus be removed without incurring much loss of information. *Redundant* and *irrelevant* are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated.

Feature Selection is a very critical component in a Data Scientist’s workflow. When presented data with very high dimensionality, models usually choke because

1. **Training time** increases exponentially with number of features.
2. Models have increasing risk of **over fitting**with increasing number of features.

Feature Selection methods helps with these problems by reducing the dimensions without much loss of the total information. It also helps to make sense of the features and its importance.

In our data set we first carried out correlation analysis on categorical predictors so as to deal with the problem of multicollinearity as well as remove predictors which are supplying redundant information.

The correlation plot plotted is as given below-

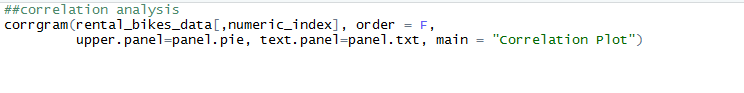
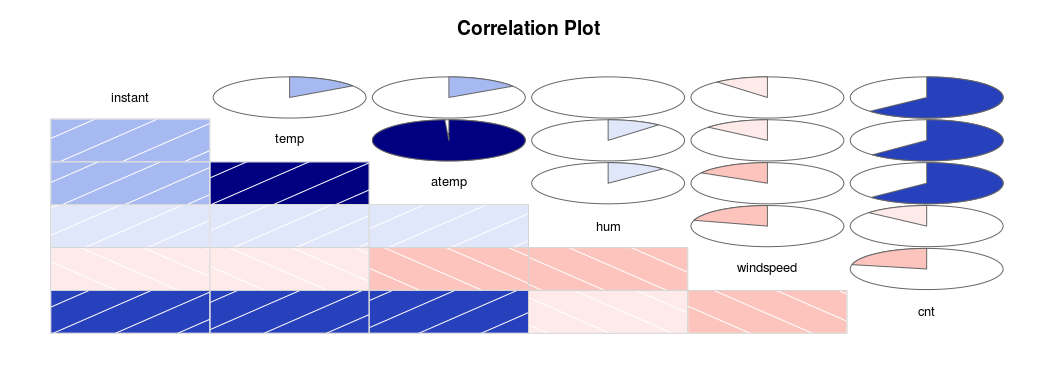


Fig 2.2 correlation plot



The intensity of blue color indicates how positively correlated the predictor is with the other variable while the intensity of pink color indicates how negatively correlated the predictor is with the other variable.

Studying the correlation plot we get the understanding that atemp predictor is highly positively related with the temp predictor . Thus we can drop one of these variable as both of them are carrying same information . Here the final decision was to drop atemp predictor.

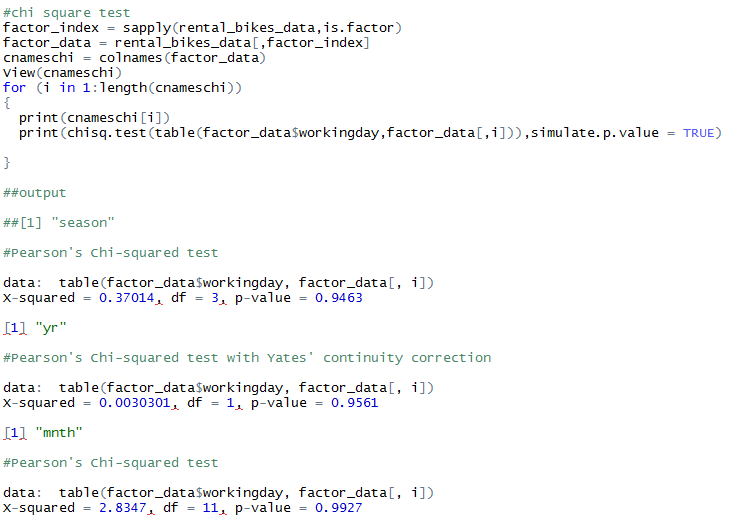
Chi –square test analysis

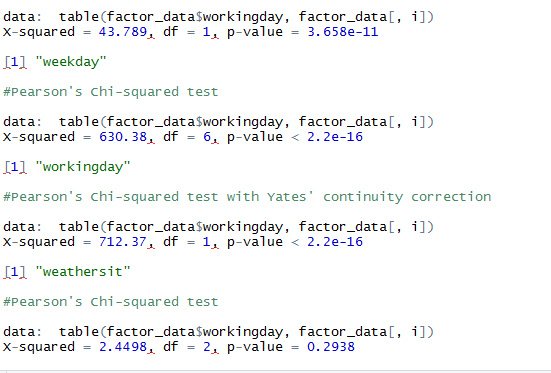
The next step was to carry out chi square test analysis. This analysis is carried out on categorical predictors. In this analysis two assumptions are made first, one is that there should be high dependency between dependent and independent variable .the second assumption is that there should be no dependency between two independent variables. After this two hypothesis are stated-

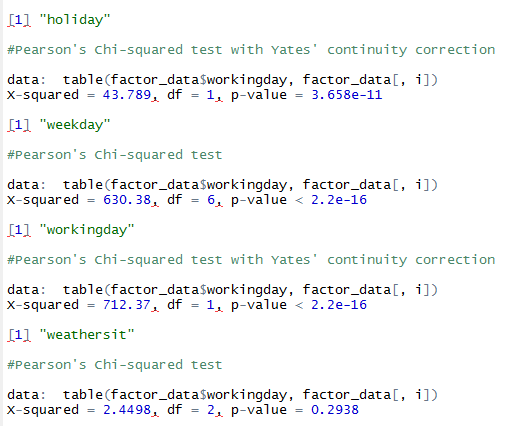
Null hypothesis- the two predictors are highly dependent

Alternate hypothesis- the two predictors are highly independent.

In this analysis we get p value ,if p value is less then 0.05(significant value) then we accept the null hypothesis and say that the two predictors are highly correlated .if p value is greater than 0.05, then we reject null hypothesis and say that the two predictors are independent.







In our test we found out that the holiday predictor , working day and week day accepts the null hypothesis and are highly correlated but as we wanted that the independent variables should be highly independent, thus the final decision was to drop holiday predictor from our dataset.

Thus following predictors were dropped from the data set before using the dataset to train the models.

1) instant- not significant for the analysis.

2) Atemp= shows high correlation with temp as indicated in correlation plot.

3) holiday- shows high dependency with weekday as well as with working day as seen in chi square test.

4) Casual and registered- they are dependent variables and not significant.

5)dteday- redundant information

2.2 Modeling

2.2.1 Model Selection

We have 8 independent variable and 1 dependent variable. As our dependent variable is continuous .we have to choose a regression model. The first model we chose for our study is decision tree.

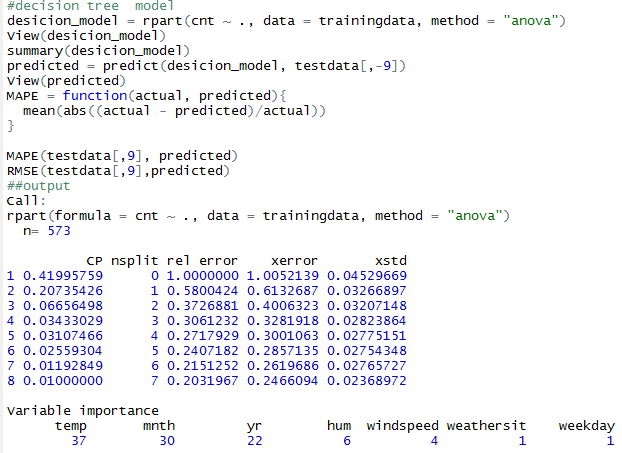
2.2.2 Decision Tree

Decision trees, as the name suggest, is a tree shaped visual representation of one can reach to a particular decision by laying down all options and their probability of occurrances.Decision trees are extremely easy to understand and interpret. At each node of the tree, one can interpret what would be the consequence of selecting that node or option.

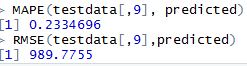
#decision tree model

desicion\_model = rpart(cnt ~ ., data = trainingdata, method = "anova"

The important observation was made when we looked at the summary of this model –



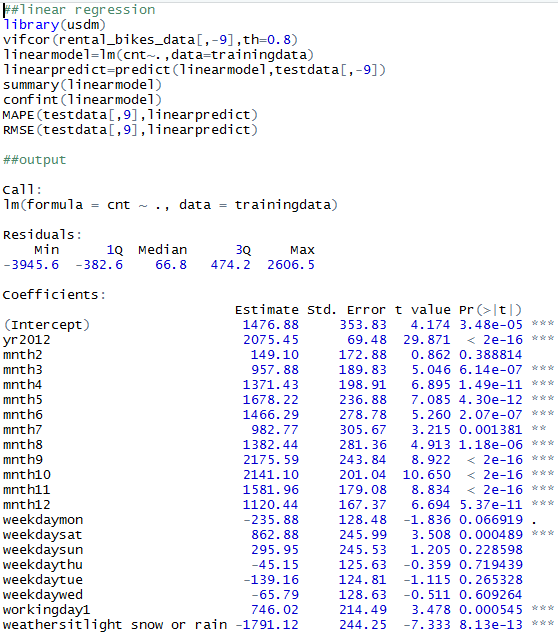
The temp predictor was found to be the most important variable and contributed the most in explaining our target variable whereas weathersit and weekday were least important.

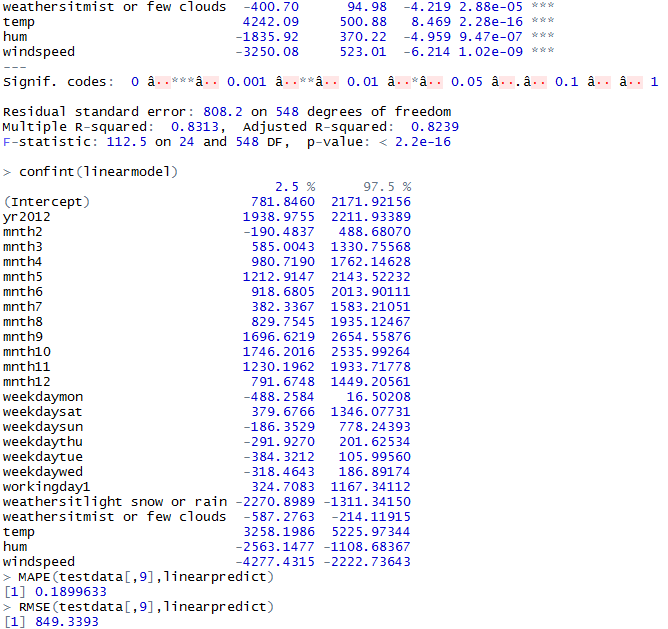


The MAPE is mean absolute percent error and the error is 23% for decision tree.

The RMSE is root mean square error and its value is 989.7755 for decision tree.

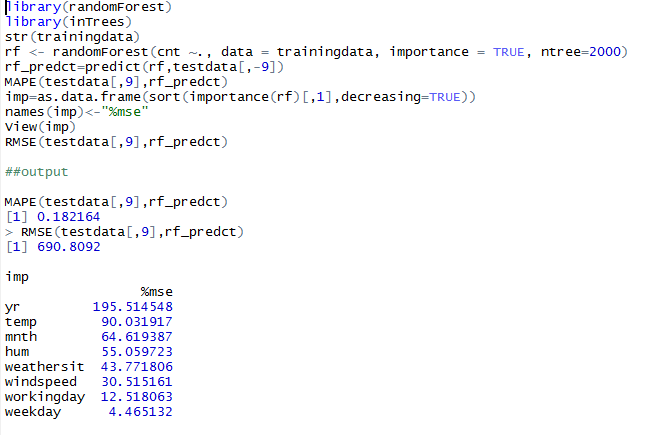
2.2.3 Linear Regression





The R-squared value shows that our model is 83%efficient in explaining the data .thus the model is very good.

2.2.4 Random Forest



Chapter 3

Conclusion

|  |  |  |  |
| --- | --- | --- | --- |
| name | RMSE | MAPE | ACCURACY |
| Decision tree | 989.77 | 23.34% | 77% |
| Linear Regression | 849.333 | 18.99% | 81% |
| Random Forest | 690.809 | 18.21% | 81.40% |

RMSE-

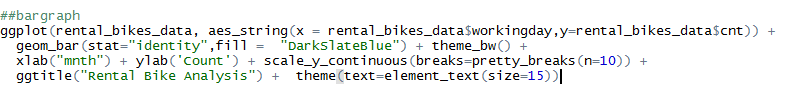
The **root-mean-square deviation (RMSD)** or **root-mean-square error (RMSE)** (or sometimes **root-mean-square*d* error**) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an [estimator](https://en.wikipedia.org/wiki/Estimator) and the values observed.

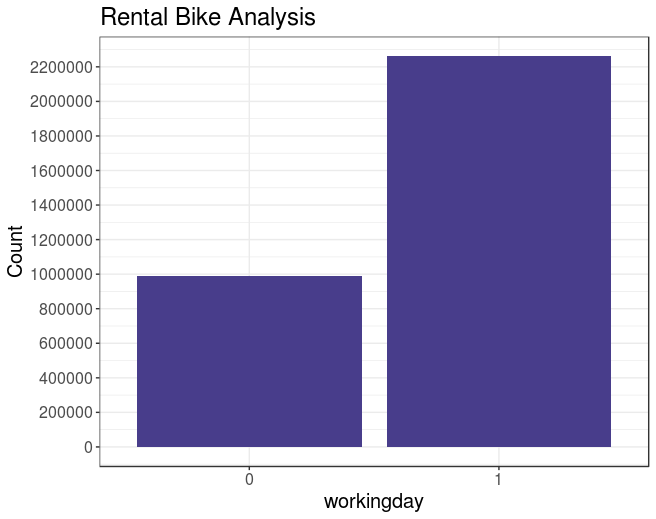
The RMSE value should be as less as possible and as we can see random forest has less RMSE value as compared to decision and linear regression.

MAPE-

The **mean absolute percentage error** (**MAPE**), also known as **mean absolute percentage deviation** (**MAPD**), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation.

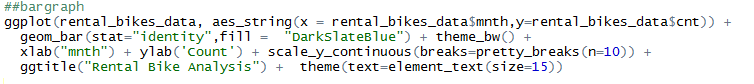
The MAPE value gives us the accuracy of the model and in our study we found out that Linear Regresssion and Random Forest have almost same accuracy of 81%

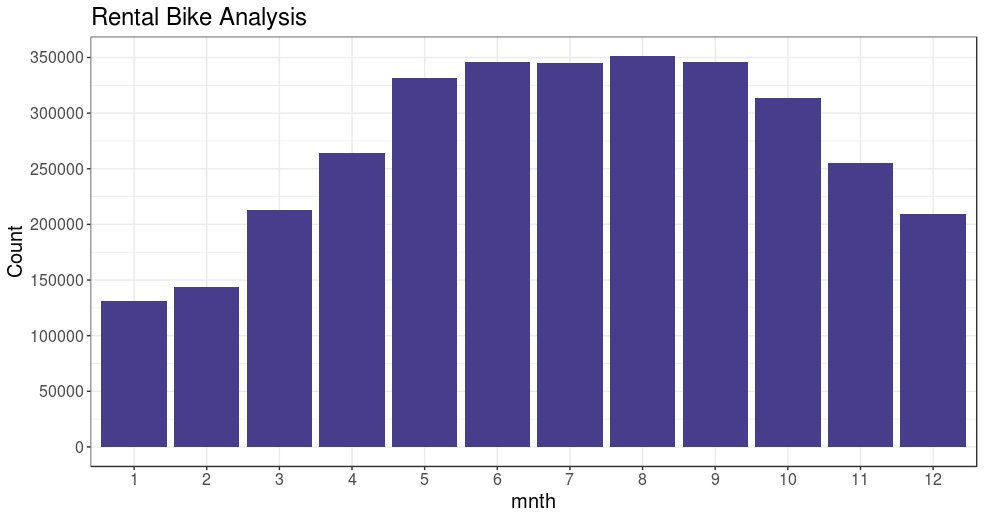




Looking at the bar graph we can observe that the bikes are more rented in the working day as compared to a non working day.

{\displaystyle {\mbox{M}}={\frac {1}{n}}\sum \_{t=1}^{n}\left|{\frac {A\_{t}-F\_{t}}{A\_{t}}}\right|,}





August is the month having peak count of bike renting as it has a very good weather having few clouds.

3.2 Model Selection

We can see that both linear regression and random forest give approximately same accuracy but the RMSE of random forest is less.Thus random forest is the best model that can be used for modeling this dataset.

References

1. Wikipedia

2. Edwisor

3. Youtube