Bike Rental Count– Project Report

Ambika Iragappa

30-09-2019

Contents

Chapter 1: Introduction

1.1 Problem Statement

The aim of this project is to predict the count of bike rentals based on the seasonal and environmental settings. By predicting the count, it would be possible to help accommodate in managing the number of bikes required on a daily basis and being prepared for high demand of bikes during peak periods.

1.2 Data

Given below is a sample of the data set that we are using to predict the number of bikes:

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

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit
1	2011-01-01	1	0	1	0	6	0	2
2	2011-01-02	1	0	1	0	0	0	2
3	2011-01-03	1	0	1	0	1	1	1
4	2011-01-04	1	0	1	0	2	1	1
5	2011-01-05	1	0	1	0	3	1	1

Table 1.2: Bike Count Sample Data (Columns: 10-16)

temp	atemp	hum	windspeed	casual	registered	cnt
0.3441670	0.3636250	0.805833	0.1604460	331	654	985
0.3634780	0.3537390	0.696087	0.2485390	131	670	801
0.1963640	0.1894050	0.437273	0.2483090	120	1229	1349
0.2000000	0.2121220	0.590435	0.1602960	108	1454	1562
0.2269570	0.2292700	0.436957	0.1869000	82	1518	1600

As we can see in the table below, we have the following 13 variables, using which we have to correctly predict the count of bikes:

Sl.No	Variables
1	Instant
2	Dteday
3	Season
4	Yr
5	Month
6	Holiday
7	Weekday
8	Workingday
9	Weathersit
10	Temp
11	Atemp
12	Hum
13	windspeed

Table 1.3: Predictor Variables

The details of variable present in the dataset are as follows - instant: Record index

dteday: Date

season: Season (1: spring, 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 from Holiday Schedule)

weekday: Day of the week

workingday: If day is neither weekend nor holiday is 1, otherwise is 0.

weathersit: (weather situation 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

temp: Normalized temperature in Celsius.

atemp: Normalized feeling temperature in Celsius.

hum: Normalized humidity.

windspeed: Normalized wind speed.

casual: count of casual users

registered: count of registered users

cnt: count of total rental bikes including both casual and registered

Chapter 2: Methodology

2.1 Pre-Processing

A predictive model requires that we look at the data before we start to create a model. However, in data mining, looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is known as Exploratory Data Analysis.

2.2 Exploratory data Analysis - Distribution of continuous variables

It can be observed from the below histograms is that temperature and feel temperature are normally distributed, whereas the variables windspeed and humidity are slightly skewed.
The skewness is likely because of the presence of outliers and extreme data in those variables.

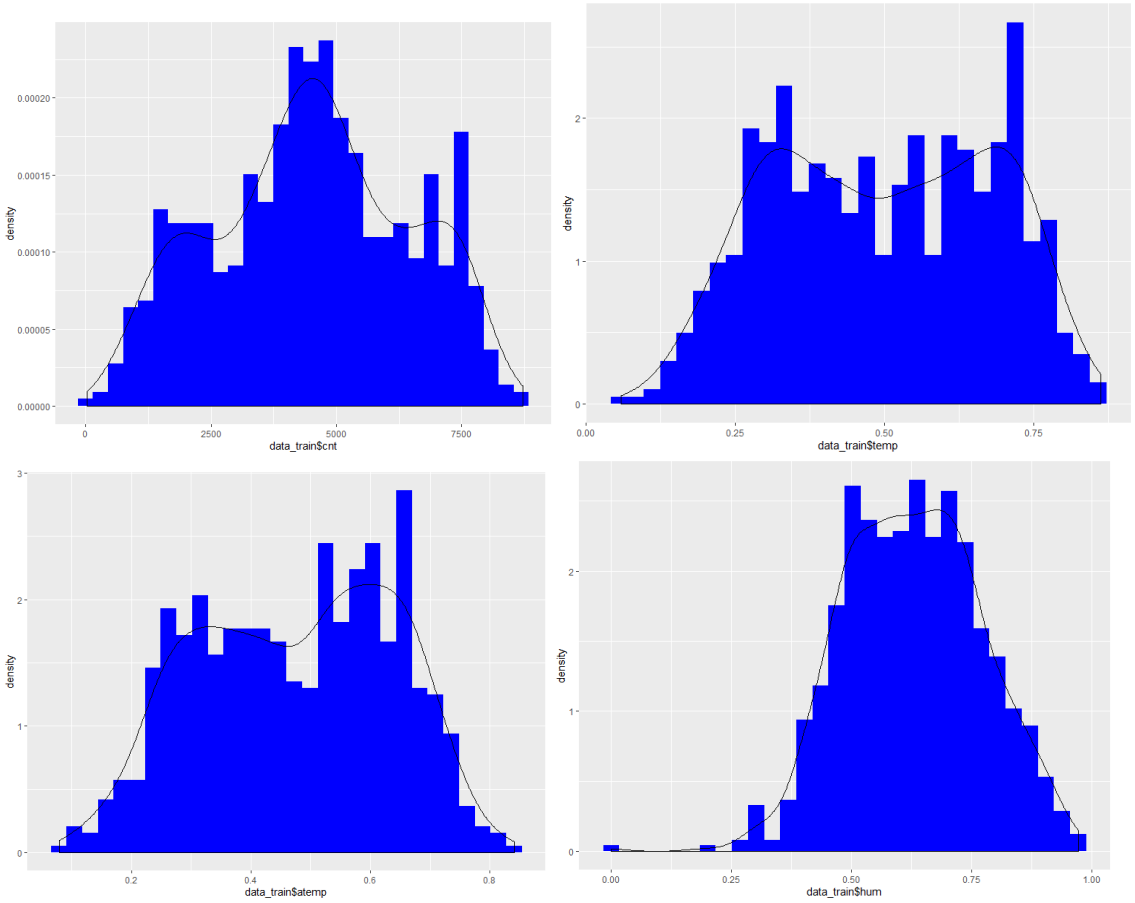

Fig 2.1: Distribution of continuous variables using Histograms

2.3 Relations of Variables against bike rental count – categorical variables

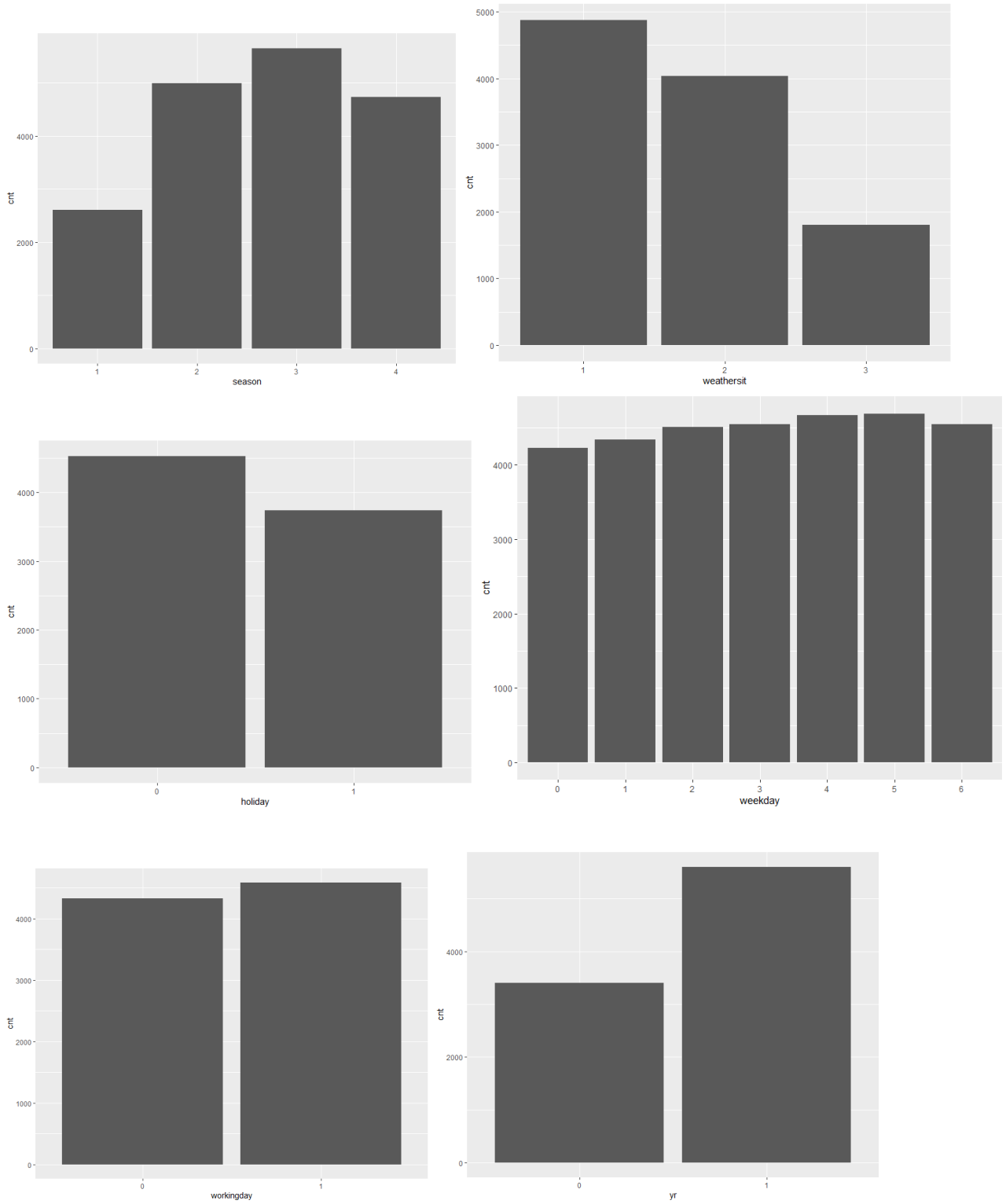

Fig 2.2: Distribution of categorical variables using bar plots

2.4 Missing Value Analysis

There are no missing values in the given dataset, so we need not apply any missing value imputation methods to fill the NAN's.

2.5 Detection of outliers:

Outliers are detected using boxplots. Below figure illustrates the boxplots for all the continuous variables. There are outliers in humidity and windspeed.

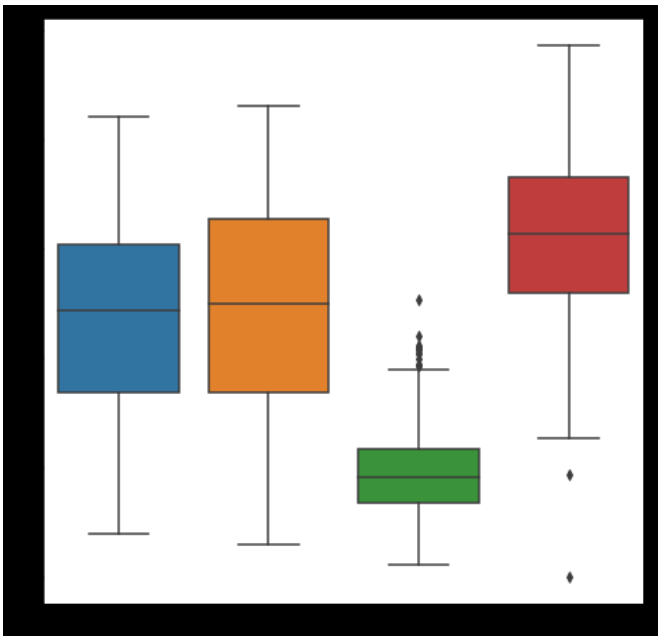

Fig 2.4: Boxplot of continuous variables

Outlier detection using IQR.

Inter Quartile Range (IQR) is calculated and the minimum and maximum value are calculated for the variables. Any value ranging outside the minimum and maximum value are discarded.

IQR = q75 - q25
min = q25 - (IQR*1.5)
max = q75 + (IQR*1.5)

2.6 Feature Selection

Feature Selection reduces the complexity of a model and makes it easier to interpret. It also reduces overfitting. Features are selected based on their scores in various statistical tests for their correlation with the outcome variable.
Correlation plot is used to find out if there is any multicollinearity between variables. The highly collinear variables are dropped and then the model is executed.

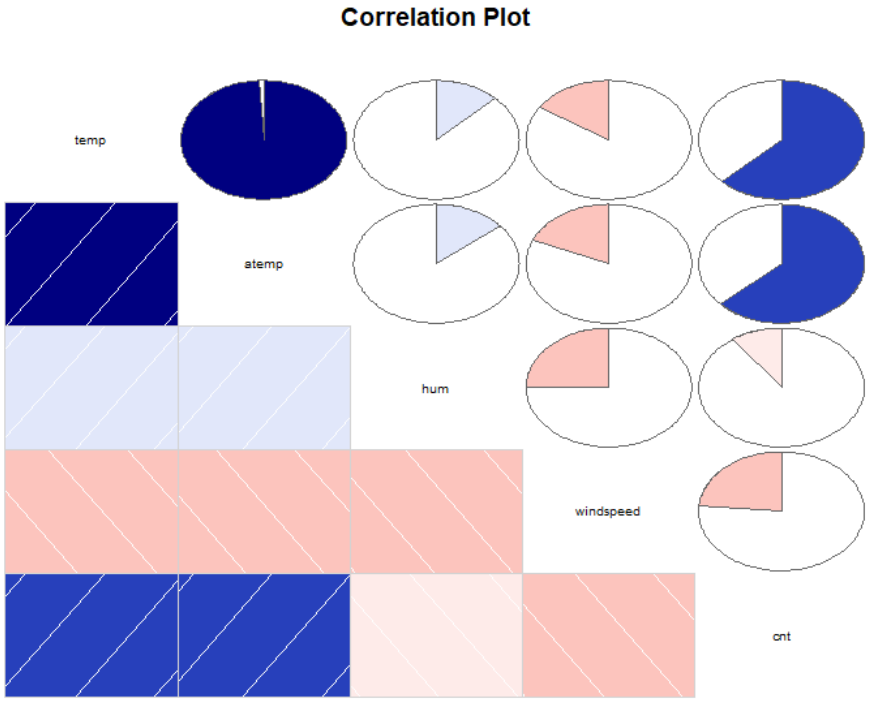

Fig 2.7: Correlation plot of all the variables

We can also use the VIF (Variance Inflation Factor) method to calculate the multicollinearity.

VIF = 1 / (1- r^2)

If the VIF <= 5, the variables have little or moderate linearity between the variables
If VIF > 5, the variables have high linearity between the variables.

Here in the diagram, it clearly shows that, the correlation between temp and atemp is very high so, we can drop atemp variable in order to overcome the multicollinearity.
Also, we can drop casual and registered variables since these are the result of count i.e., the sum of these two variables is count, the target variable. It doesn't make any sense or gives any value to the model, because that is the value we have to predict.

2. 7 Feature Scaling

Feature scaling includes two functions normalization and standardization. It is done to reduce the unwanted variation either within or between variables and to bring all of the variables into proportion with one another.
In the given dataset all numeric values are already present in normalized form. So, we need not perform any scaling methods here.

Chapter 3: Modelling

3.1 Model Selection

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

3.2 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.

```
lm_model = lm(cnt ~., data = train)

predictions_LR = predict(lm_model,test[,1:11])
```

```
Call:
lm(formula = cnt ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3765.5  -332.0    82.1   452.4  2873.4

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1502.77     280.78   5.352 1.35e-07 ***
season2       752.59     207.14   3.633 0.000310 ***
season3       458.96     256.00   1.793 0.073625 .
season4      1427.41     229.48   6.220 1.08e-09 ***
yr1          2036.77      68.76  29.621  < 2e-16 ***
mnth2         246.10     178.53   1.378 0.168689
mnth3         564.94     203.54   2.776 0.005725 **
mnth4         646.20     294.84   2.192 0.028878 *
mnth5         824.90     319.81   2.579 0.010192 *
mnth6         748.08     337.99   2.213 0.027341 *
mnth7         417.38     372.10   1.122 0.262550
mnth8         733.92     360.06   2.038 0.042061 *
mnth9        1348.72     322.76   4.179 3.48e-05 ***
mnth10        641.78     298.97   2.147 0.032318 *
mnth11         46.44     285.36   0.163 0.870802
mnth12         42.85     235.12   0.182 0.855471
holiday1     -908.80     213.59  -4.255 2.51e-05 ***
weekday1      230.35     128.06   1.799 0.072675 .
weekday2      169.11     126.74   1.334 0.182747
weekday3      396.48     128.32   3.090 0.002119 **
weekday4      381.25     122.97   3.100 0.002046 **
weekday5      441.65     128.56   3.435 0.000643 ***
weekday6      392.93     127.92   3.072 0.002248 **
workingday1       NA         NA      NA       NA
weathersit2  -424.39      92.46  -4.590 5.66e-06 ***
```

```
weathersit3 -1899.84     232.88  -8.158 2.97e-15 ***
temp         4790.38     495.36   9.671  < 2e-16 ***
hum         -1608.37     349.98  -4.596 5.52e-06 ***
windspeed   -3232.21     487.35  -6.632 8.88e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 758.8 on 483 degrees of freedom
Multiple R-squared:  0.8513,    Adjusted R-squared:  0.843
F-statistic: 102.4 on 27 and 483 DF,  p-value: < 2.2e-16
```

As we can see the Adjusted R-squared value, we can explain 84.30% 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. This model explains the data very well and is considered to be good.

Even after removing the non-significant variables, the accuracy, Adjusted R-squared and F- statistic do not change by much, hence the accuracy of this model is chosen to be final.
MAPE of this multiple linear regression model is 10.09%. Hence the accuracy of this model is 89.91%. This model performs very well for this test data.

3.3 Decision Tree:

```
fit = rpart(cnt ~ ., data = train, method = "anova")
predictions_DT = predict(fit, test[,-11])
```

A decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.

Using decision tree, we can predict the value of bike count. The MAPE for this decision tree is 13.73%. Hence the accuracy for this model is 86.27%.

3.4 Random Forest:

```
RF_model = randomForest(cnt ~ ., train, importance = TRUE, ntree = 200)

predictions_RF = predict(RF_model, test[,-11])
```

Using Classification for prediction analysis in this case is not normal, though it can be done. The number of decision trees used for prediction in the forest is 500. Using random forest, the MAPE was found to be 9.80%. Hence the accuracy is 90.20%.

Chapter 4: Conclusion

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

In our case of Bike count prediction data, Interpretability and Computation Efficiency, do not hold much significance. Therefore, we will use Predictive performance as the criteria to compare and evaluate models.

Predictive performance can be measured by comparing Predictions of the models with real values of the target variables and calculating some average error measure.

4.1 Mean Absolute Percentage Error (MAPE)

MAPE is one of the error measures used to calculate the predictive performance of the model. We will apply this measure to our models that we have generated in the previous section.

```
MAPE <- function (y,  yhat)
{
  mean (abs (y  - yhat) / y) * 100
}
```

Linear Regression: MAPE = 10.09%

Decision Tree: MAPE = 13.73%.

Random Forest: MAPE = 9.80%

Based on the above error metrics, Random Forest is the better model for our analysis. Hence Random Forest is chosen as the model for prediction of bike rental count.