Occupancy Detection - Craig Hardie

Craig Hardie

Occupancy Detection

Using the Occupancy Detection dataset from https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+ I am going to determine the most relevant factors in determining whether a room is occupied.

Setup

First the project requires the following libraries to run the project corrplot, MASS, and randomForest. The following code will install the packages if required and load them ready for use in the project.

```
knitr::opts_chunk$set(echo = TRUE)
```

```
#install the packages if required
if(!require(corrplot)) install.packages(corrplot)
if(!require(MASS)) install.packages(MASS)
if(!require(randomForest)) install.packages(randomForest)
#load the libraries
library(corrplot)
library(MASS)
library(randomForest)
```

Data Preparation

The dataset is already split into test and two training sets as it is delivered, so that will be the split I will use for my project. It is required to convert all the datasets using the command data.matrix so it is all numeric values and then I concatenated the two training sets together using the rbind command.

```
training<-read.csv("data/datatraining.txt")
training<-data.matrix(training)

trainframe<-data.frame(training)
test<-read.csv("data/datatest.txt")
test2<-read.csv("data/datatest2.txt")
test<-data.matrix(test)
test2<-data.matrix(test2)
testframe1<-data.frame(test)
testframe2<-data.frame(test2)

testframe<-rbind(testframe1,testframe2)
```

Exploratory Analysis

Running the command sum(is.na(trainframe)) on the trainframe and again same command on the testframe shows there are no null values and we can get the number of rows in the test and training sets with the nrow and ncol commands and the column names.

```
sum(is.na(trainframe))
```

```
## [1] 0
```

```
sum(is.na(testframe))
```

```
## [1] 0
```

```
nrow(testframe)
```

```
## [1] 12417
```

```
nrow(trainframe)
```

```
## [1] 8143
```

```
ncol(trainframe)
```

```
## [1] 7
```

```
colnames(trainframe)
```

```
## [1] "date"          "Temperature"   "Humidity"      "Light"
## [5] "CO2"           "HumidityRatio" "Occupancy"
```

Running the head(trainframe) command will show the first rows of the training data which would suggest that Occupancy would be the target data of our experiements.

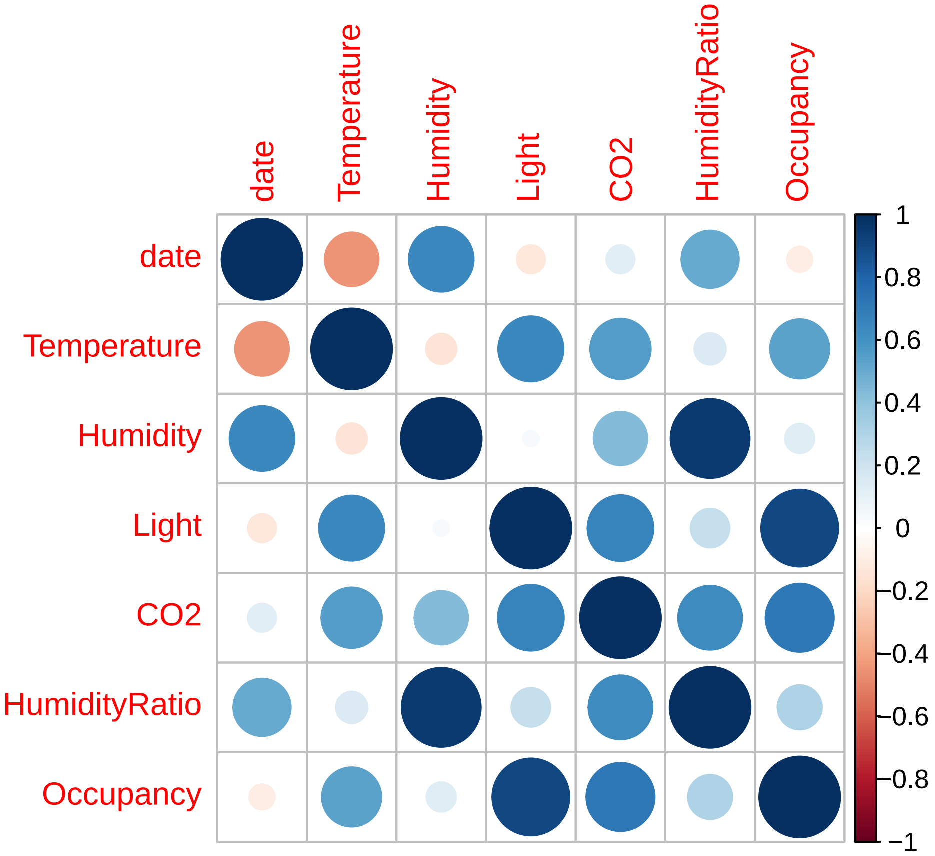
```
head(trainframe)
```

```
##   date Temperature Humidity Light    CO2 HumidityRatio Occupancy
## 1    1       23.18  27.2720 426.0 721.25   0.004792988         1
## 2    2       23.15  27.2675 429.5 714.00   0.004783441         1
## 3    3       23.15  27.2450 426.0 713.50   0.004779464         1
## 4    4       23.15  27.2000 426.0 708.25   0.004771509         1
## 5    5       23.10  27.2000 426.0 704.50   0.004756993         1
## 6    6       23.10  27.2000 419.0 701.00   0.004756993         1
```

Running the following commands on the training data will produce a correlation matrix which would suggest that the most likely significators of a room being occupied is Light with CO2 and Temperature could also be of significance.

```
cortf<-cor(trainframe)

corrplot(cortf)
```



Supervised Learning Experiment - Predicting Occupancy

Linear Determinant Analysis

The experiment I intend to carry out is to determine the likelihood of a room being occupied based on some other factors and determine the most important factors in predicting this.
To this end I will be carrying out Group Means that there is a far greater descrepancy for Occupancy based on Light than any other with CO2 also having a deciding factor in predicting the Occupancy. This confirms what the exploratory analysis had suggested might be the case.

```
lda.fit<-lda(Occupancy ~ ., data=trainframe)
lda.fit
```

```
## Call:
## lda(Occupancy ~ ., data = trainframe)
##
## Prior probabilities of groups:
##         0         1
## 0.7876704 0.2123296
##
## Group means:
##       date Temperature Humidity     Light      CO2 HumidityRatio
## 0 4191.795    20.33493 25.34968  27.77644  490.3203   0.003729632
## 1 3627.600    21.67319 27.14794 459.85435 1037.7048   0.004355428
##
## Coefficients of linear discriminants:
```

```
##                          LD1
## date          -2.062643e-04
## Temperature   -8.306103e-01
## Humidity      -6.592929e-02
## Light          1.314403e-02
## CO2            2.420417e-03
## HumidityRatio  6.379053e+02
```

```
lda.predict<-predict(lda.fit, testframe)
table(lda.predict$class , testframe$Occupancy)
```

```
##
##        0    1
##   0 9274    8
##   1  122 3013
```

```
mean(table(lda.predict$class == testframe$Occupancy))
```

```
## [1] 6208.5
```

Random Forest

I am going to use Random Forest analysis to carry out a similar analysis to I did with LDA and determine what the deciding factors could be in whether a room is occupied.

```
rt.fit<-randomForest(Occupancy ~., data=trainframe)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
conf<-rt.fit$confusion
conf

trainResults<-predict(rt.fit, trainframe)
table(Predictions=trainResults, Actual=trainframe$Occupancy)

testResults<-predict(rt.fit, testframe)
table(Predictions=testResults, Actual=testframe$Occupancy)
```

```
importance(rt.fit)
```

```
##               IncNodePurity
## date               90.78584
## Temperature       134.95304
## Humidity           25.43220
## Light             726.90511
## CO2               326.22745
## HumidityRatio      54.21730
```

This confirms what the LDA analysis suggested which is that Light and Co2 are highly significant in determining whether a room is occupied with Light being the most significant factor. Temparature is also highly significant in this analysis.

The outcomes of the LDA and Random Forest have largely confirmed what the initial exploration suggested which is that Light is the most significant predictor and that Co2 and Temparature are also significant.