Comparing Room Occupancy Classifiers

Suha Niyas
sniyas

Due Wed, Apr 28, at 8:00PM (Pittsburgh time)

Contents

```r
set.seed(151)
library("knitr")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
library("MASS")
library("klaR")
library("tree")
library("rpart")
library("rpart.plot")
```

```r
occupancy_train <- readr::read_csv("http://stat.cmu.edu/~gordonw/occupancy_train.csv")
occupancy_test <- readr::read_csv("http://stat.cmu.edu/~gordonw/occupancy_test.csv")
```

Introduction

Knowing whether a room is occupied or not is a useful piece of knowledge in many different situations, whether in a serious emergency like a fire occurring in the building or just knowing if the person living in the room is still alive or even away on vacation. In this project, we hope to use different classifiers and methods to determine if a room is occupied or not. We will base the room occupancy based on different characteristics about the room being occupied and the time the room is being occupied.

Exploratory Data Analysis

Overview/Background of Dataset and Variables

The following predictor values are:

- Temperature: room temperature in degrees C
- Humidity: room relative humidity, in percent
- CO2: room's carbon dioxide in ppm
- Hour: hour of the day, from 0 to 23

and the response variable to predict is:

- Occupancy: binary, 0 for not occupied, 1 for occupied status

EDA on Response Variable

```
table(occupancy_train$Occupancy)
```

```
##
##    0    1
## 4497 1203
```

```
prop.table(table(occupancy_train$Occupancy))
```

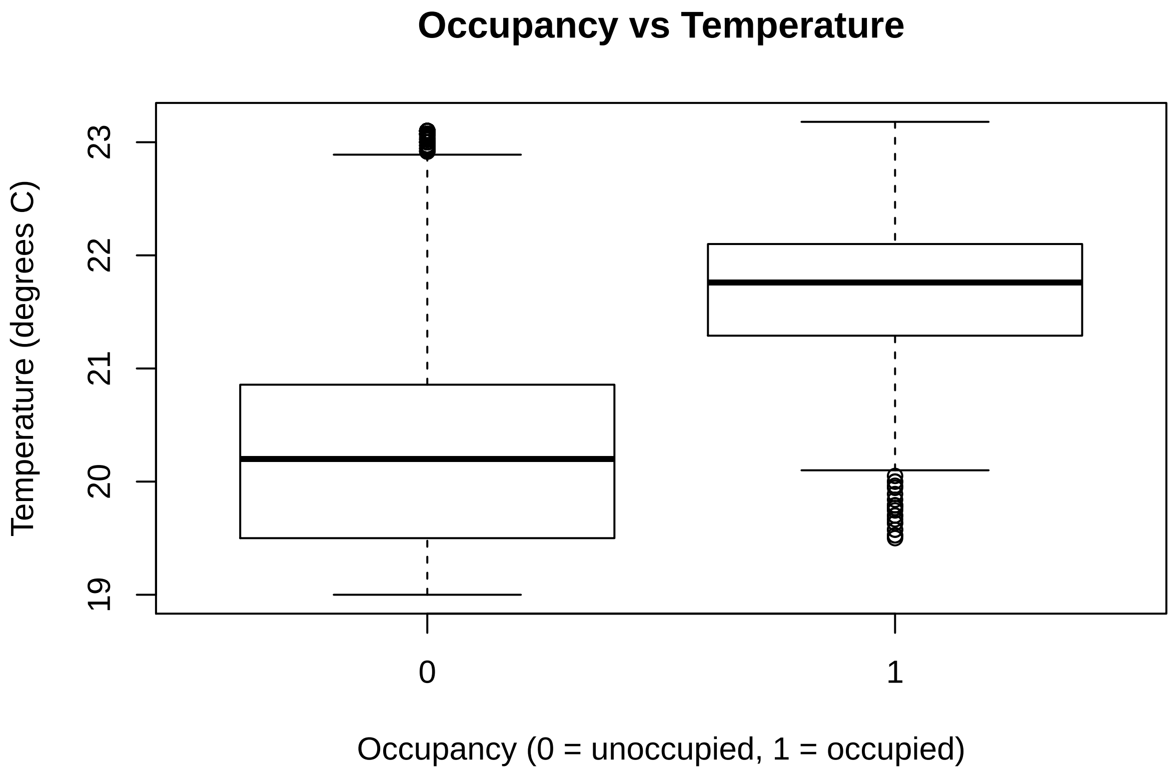
```
##
##         0         1
## 0.7889474 0.2110526
```

In the training dataset, there were 5700 observations with 4497 of the rooms not being occupied and 1203 of the rooms being occupied, which is 78.89474% and 78.89474% of the dresses, respectively.

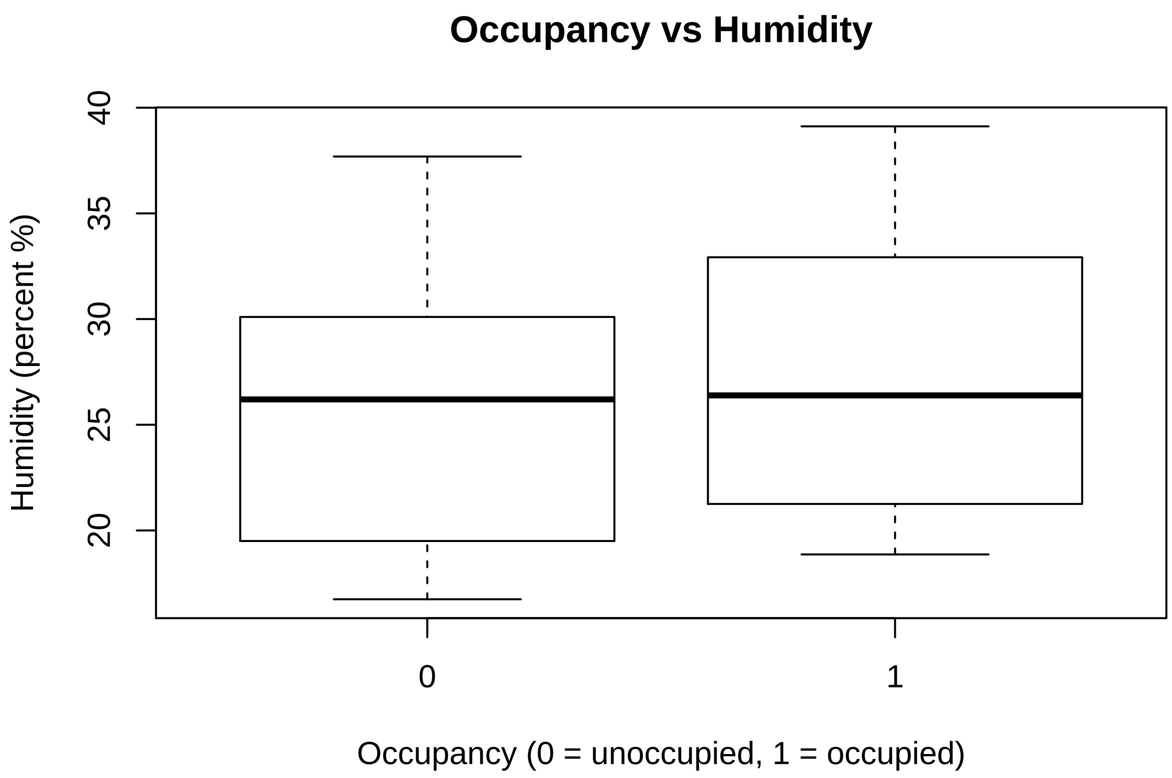
EDA on Relationship between Response Variable and Each Quantitative Variable

We will now visualize the relationships between the response variable, `Recommendation`, and the categorical quantitative predictor variables by looking at the conditional proportions in bar plots.

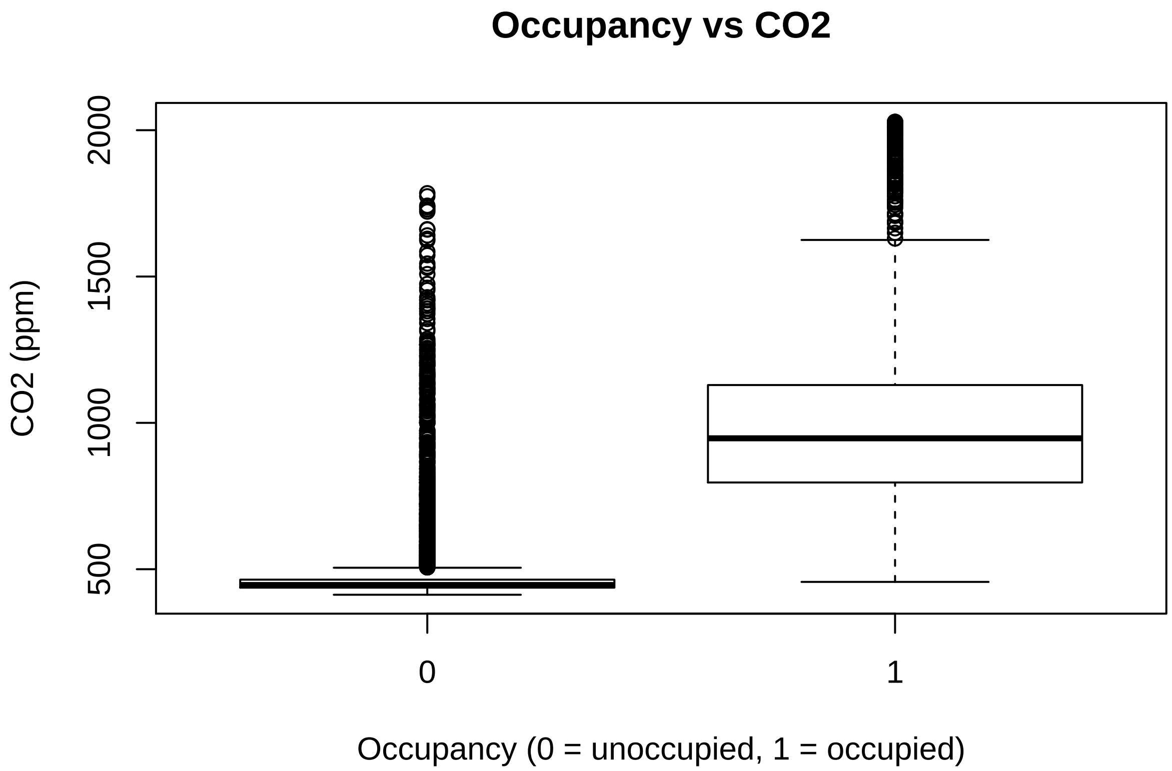
```
boxplot(Temperature ~ Occupancy,
        main = "Occupancy vs Temperature",
        xlab = "Occupancy (0 = unoccupied, 1 = occupied)",
        ylab = "Temperature (degrees C)",
        data = occupancy_train)
```



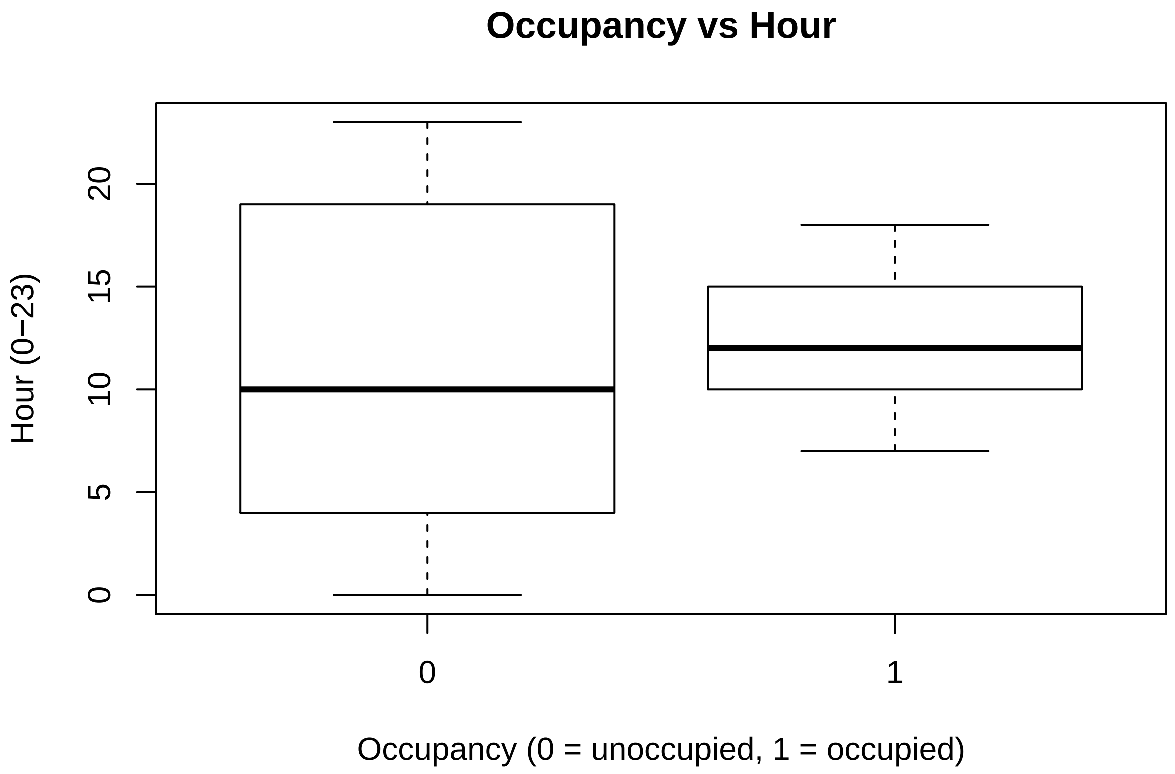
```r
boxplot(Humidity ~ Occupancy,
        main = "Occupancy vs Humidity",
        xlab = "Occupancy (0 = unoccupied, 1 = occupied)",
        ylab = "Humidity (percent %)",
        data = occupancy_train)
```



```
boxplot(CO2 ~ Occupancy,
        main="Occupancy vs CO2",
        xlab = "Occupancy (0 = unoccupied, 1 = occupied)",
        ylab = "CO2 (ppm)",
        data = occupancy_train)
```



```r
boxplot(Hour ~ Occupancy,
        main="Occupancy vs Hour",
        xlab = "Occupancy (0 = unoccupied, 1 = occupied)",
        ylab = "Hour (0-23)",
        data = occupancy_train)
```



In the boxplots above, we notice that if the boxplots show differences between the rooms that are occupied and aren't occupied, we will have evidence of a relationship and a variable useful in the classifiers. We note that the temperature is typically higher in an occupied room. We also notice that the median is quite similar for both unoccupied and occupied rooms as well as the spread of humidity. We note that ppm of CO2 is higher in occupied rooms and lower in unoccupied rooms. This boxplot also displays many outliers for both types of occupancy that are all greater than the maximum value. The median hours for both occupied and unoccupied rooms are slightly similar, but the spread of the occupied rooms is smaller than the spread of the unoccupied rooms.

Modeling

Now, we will build the classifiers for predicting the occupancy of a room. The four classifiers that will be used are linear discriminant analysis (LDA), quadratic discriminant analysis (LDA), classification trees, and binary logistic regression.

We split the observations randomly into training and test sets to make sure that the models aren't overfitting to the sample. By utilizing the same training observations and assessing the same test observations, all four classifier models were constructed.

LDA

We will exclude our categorical values from the LDA and only use quantative predictor values.

The LDA classifier on the training data is built as:

```r
occupancy.lda <- lda(Occupancy ~ Temperature + Humidity + CO2 + Hour,
                     data = occupancy_train)
```

The performance of the LDA classifier on the test data is:

```r
occupancy.lda.predict <- predict(occupancy.lda, as.data.frame(occupancy_test))
```

```r
table(occupancy.lda.predict$class, occupancy_test$Occupancy)
```

```
## 
##        0    1
##   0 1844  111
##   1   73  415
```

```r
lda.overall.error = (73+111)/(1844+73+111+415)
lda.overall.error
```

```
## [1] 0.07531723
```

```r
lda.0.error = 73/(1844+73)
lda.0.error
```

```
## [1] 0.03808033
```

```r
lda.1.error = 111/(111+415)
lda.1.error
```

```
## [1] 0.2110266
```

Gathering results from the test dataset, the overall error rate from the LDA was 0.07531723. The error rate for unoccupied rooms of 0.03808033 was lower than the error rate for the occupied rooms of 0.2110266.

QDA

The QDA classifier on the training data is built as:

```r
occupancy.qda <- qda(Occupancy~Temperature + Humidity + CO2 + Hour,
                     data=occupancy_train)
```

The performance of the QDA classifier on the test data is:

```r
occupancy.qda.predict <- predict(occupancy.qda, as.data.frame(occupancy_test))
```

```r
table(occupancy.qda.predict$class, occupancy_test$Occupancy)
```

```
## 
##        0    1
##   0 1832   81
##   1   85  445
```

```r
qda.overall.error = (85+81)/(1832+85+81+445)
qda.overall.error
```

```
## [1] 0.06794924
```

```r
qda.0.error = 85/(1832+85)
qda.0.error
```

```
## [1] 0.04434011
```

```
qda.1.error = 81/(81+445)
qda.1.error
```

```
## [1] 0.1539924
```

Gathering results from the test dataset, the overall error rate from the QDA was 0.06794924. The error rate for unoccupied rooms of 0.04434011 was lower than the error rate for the occupied rooms of 0.1539924.

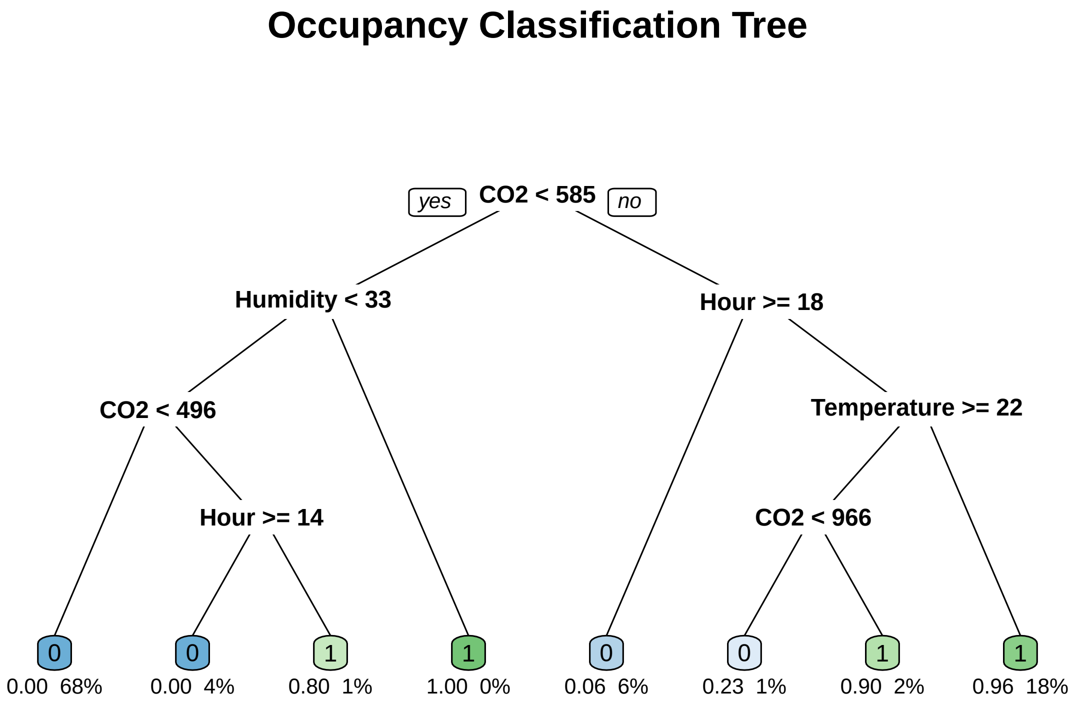
The overall error rate from the LDA was higher than the overall error rate from the QDA. The error rate for unoccupied rooms from the LDA was lower than the error rate for unoccupied rooms from the QDA. Also, the error rate for occupied rooms from the LDA was higher than the error rate for occupied rooms from the QDA. These results could show that compared to the LDA, the QDA may have overfitted for the unoccupied rooms, explaining the higher error rate.

Classification Tree

The Classification Tree on the training data is built as:

```
occupancy.tree <- rpart(Occupancy~Temperature + Humidity + CO2 + Hour, data=occupancy_train,
                        method = "class")

rpart.plot(occupancy.tree,
           type = 0,
           clip.right.labs = FALSE,
           branch = 0.1,
           under = TRUE,
           main = "Occupancy Classification Tree")
```



We notice that the classification tree above selected the predictor variable CO2 to classify occupancy.

The performance of the tree on the test data is as shown:

```
occupancy.tree.predict <- predict(occupancy.tree,as.data.frame(occupancy_test),type="class")
```

```
table(occupancy.tree.predict, occupancy_test$Occupancy)
```

```
##
## occupancy.tree.predict    0    1
##                      0 1883   15
##                      1   34  511
```

```
tree.overall.error = (15+34)/(1883+34+15+511)
tree.overall.error
```

```
## [1] 0.02005731
```

```
tree.0.error = 34/(1883+34)
tree.0.error
```

```
## [1] 0.01773605
```

```
tree.1.error = 15/(15+511)
tree.1.error
```

```
## [1] 0.02851711
```

Gathering results from the test dataset, the overall error rate from the tree was 0.02005731 The error rate for unoccupied rooms of 0.01773605 was lower than the error rate for the occupied rooms of 0.02851711. All of the error rates from the tree were lower than the error

rates from LDA and QDA. These results could show that the tree may have overfitted for both occupied and unoccupied rooms.

Binary Logistic Regression

The binary logistic regression on the test data is built:

```
occupancy.blr <- glm(factor(Occupancy) ~ Temperature + Humidity + CO2 + Hour,
                           data = occupancy_train,
                           family = binomial(link = "logit"))
```

```
occupancy.blr.prob <- predict(occupancy.blr, as.data.frame(occupancy_test),
                                     type = "response")
```

The performance of the binary logistic regression on the test data is as shown:

```
levels(factor(occupancy_test$Occupancy))
```

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

```
occupancy.blr.predict <-ifelse(occupancy.blr.prob > 0.5,"1","0")
```

```
table(occupancy.blr.predict, occupancy_test$Occupancy)
```

```
##
## occupancy.blr.predict    0    1
##                     0 1849   96
##                     1   68  430
```

```
blr.overall.error = (68+96)/(1849+68+96+430)
blr.overall.error
```

```
## [1] 0.06713058
```

```
blr.0.error = 68/(1849+68)
blr.0.error
```

```
## [1] 0.03547209
```

```
blr.1.error = 96/(96+430)
blr.1.error
```

```
## [1] 0.1825095
```

Gathering results from the test dataset, the overall error rate from the binary logistic regression was 0.06713058 The error rate for unoccupied rooms of 0.03547209 was lower than the error rate for the occupied rooms of 0.1825095. The error rate for the occupied rooms from the binary logistic regression was lower than the error rates from LDA, QDA, and tree for the occupied rooms.

The error rate for unoccupied rooms from binary logistic regression is higher than the error rate from the classification tree but lower than error rates from LDA and QDA. The overall error rate from binary logistic regression is higher than the overall error rate from the classification tree, around the same from QDA, and lower than the overall error rate from the LDA.

Discussion

Testing all four classifiers, we discovered the classification tree was the best due to it having the lowest overall error rate, the lowest error rate for unoccupied rooms, and the lowest error rate for occupied rooms. We also notice that the error rates for the unoccupied rooms were lower from all classifier models than the error rates for the occupied rooms. The binary logistic regression model and the QDA model both had similar error rates, but the binary logistic regression model performed slightly better than the QDA model. The worst classifier was the LDA model due to its high error rates. We would recommend using the classification tree, but there are limitations to the tree model such as the lower error rates possibly hinting towards an issue of overfitting.

For the future, having more predictor variables in the dataset, such as the room size for estimated occupancy, would be useful for the problem. With more predictor variables, we can create better classifiers to test for error rates and determine room occupancy. We can use this updated data and results for situations like emergencies and help people know which rooms are occupied during a serious situation.