Building and Assessing Occupancy Classifiers

Steffi Chern
steffic

Contents

Introduction

As technology advancements allow us to remotely receive information regarding room conditions, we can now explore how the different conditions are related to whether or not a room is occupied. Accurately predicting the occupancy of a room is important when a fire comes (to evacuate people at a faster rate), or checking the availability of hotel rooms. In this paper, we explore the potential room conditions that predict the occupancy of rooms by training classification models for classifying occupancy.

Exploratory Data Analysis

Background of Variables

In this research, we collect our sample of occupancy data from "Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models," by Luis M. Candanedo and Véronique Feldheim, in 2016. To best predict if a room is occupied or not, we first summarize the available 4 predictors (Temperature, Humidity, CO2, and Hour), as well as the response variable (Occupancy), as shown below:

- Temperature: the room temperature (in °C).
- Humidity: the room's relative humidity (in %)
- CO2: the room's CO2 level (in ppm)
- Hour: the hour of a day (from 0 to 23)
- Occupancy: 0 if the room is not occupied, 1 if the room is occupied (binary)

EDA on Occupancy

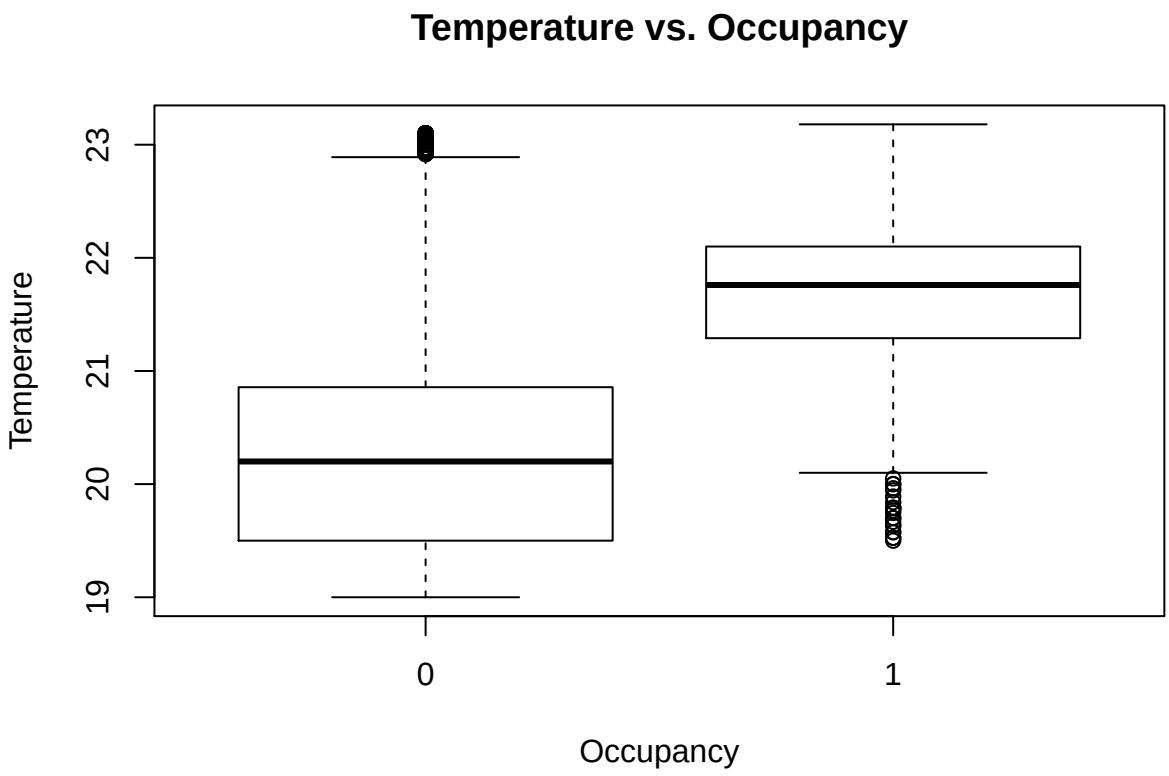
In our training dataset, we have a total of 5700 observations. There are 4497 of them (consists of 78.9%) that are not occupied, while there are 1203 of them (consists of 21.1%) that are occupied, as shown below:

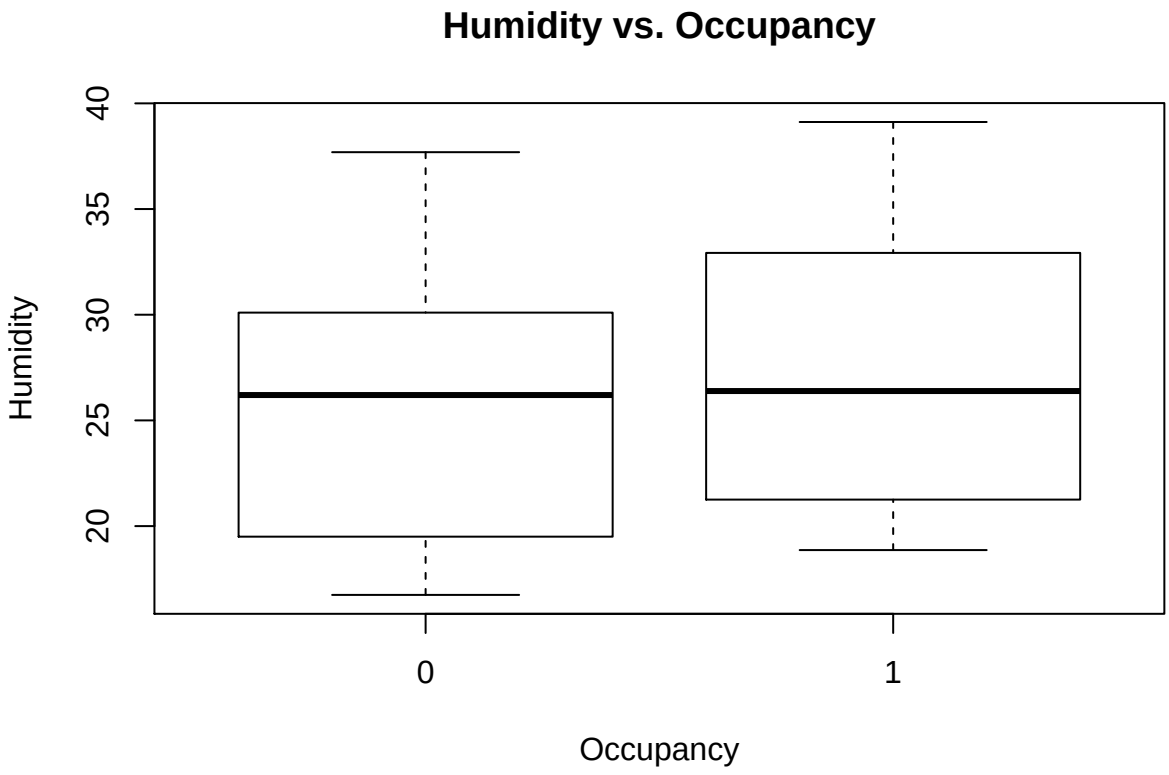
```
##
##    0    1
## 4497 1203
```

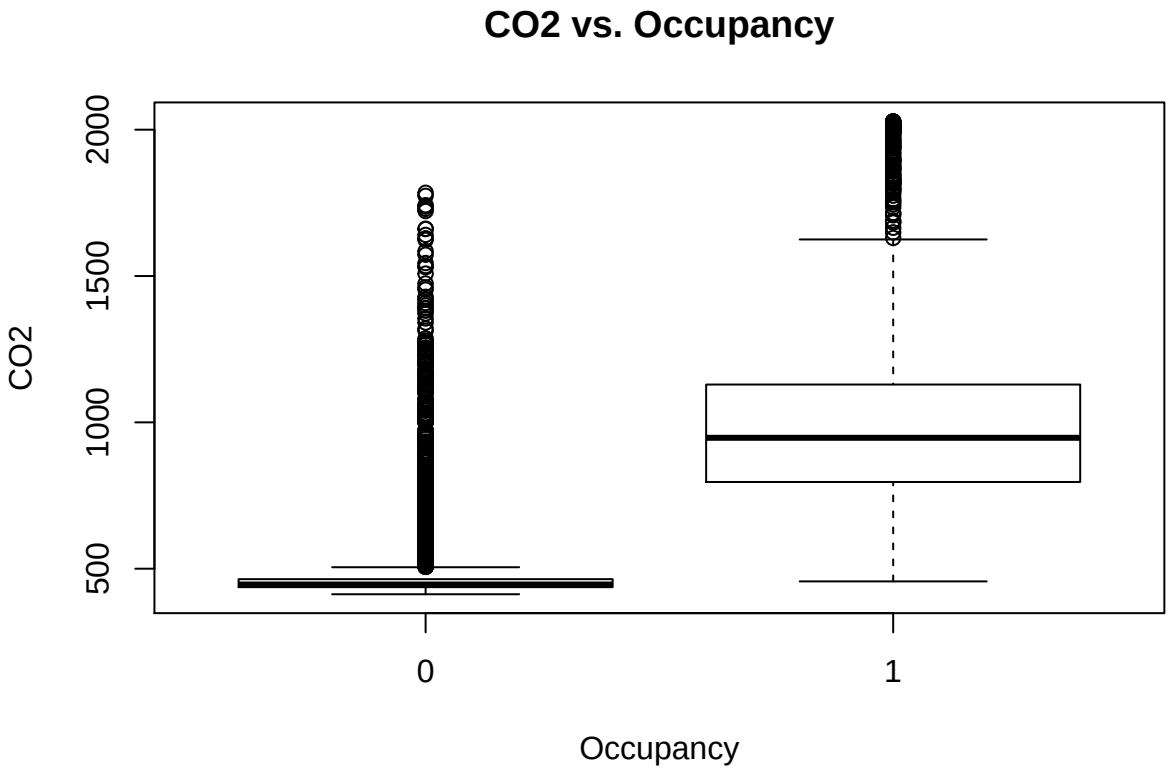
```
##
##         0         1
## 0.7889474 0.2110526
```

EDA on Occupancy vs. Quantiative Predictors

The quantitiative predictors we have in the occupancy sample are **Temperature**, **Humidity**, and **CO2**. We display the boxplots of each quantitative predictor against the response variable **Occupancy** below:


Temperature vs. Occupancy
Temperature
19
20
21
22
23
0
1
Occupancy


Humidity vs. Occupancy
Humidity
20
25
30
35
40
0
1
Occupancy

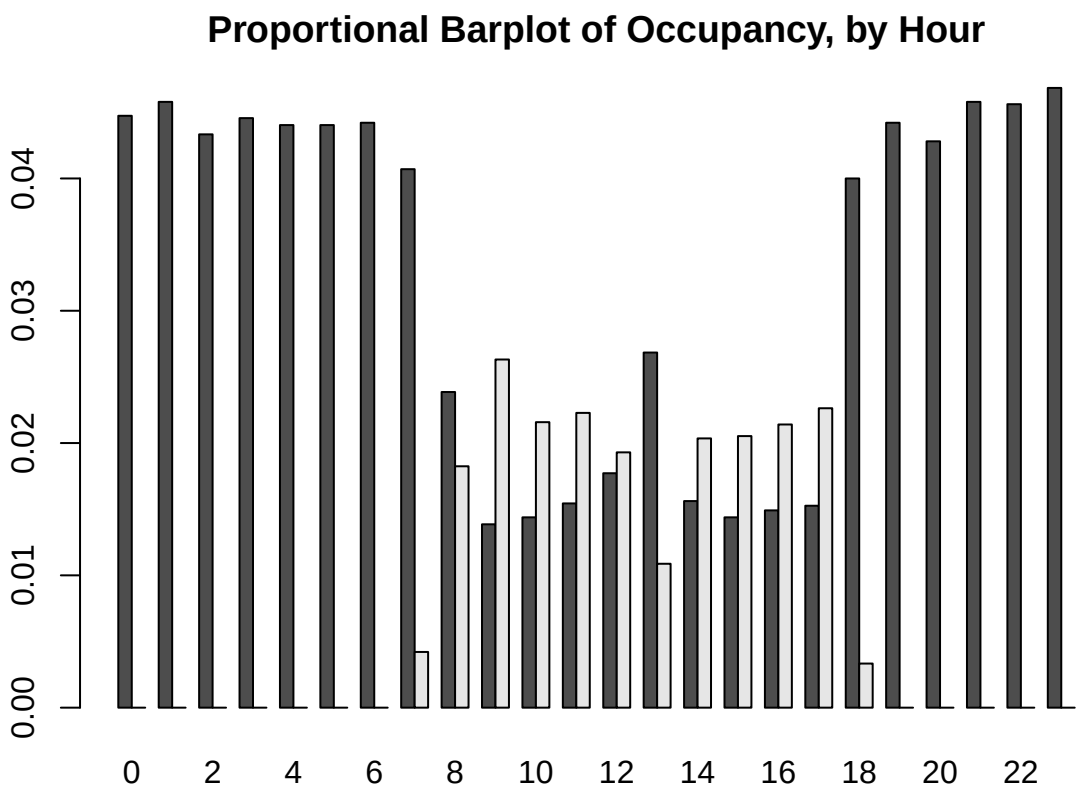


We notice a clear difference in the boxplots for **Temperature** and **Occupancy**, as well as **CO2** and **Occupancy**. It seems to be that rooms that are occupied tend to have higher temperature and higher CO2 levels (in contrast, rooms that are empty tend to have lower room temperature and CO2 levels). However, there isn't much difference in humidity between whether or not a room is occupied (the medians are very close).

EDA on Occupancy vs. Categorical Predictors

The only categorical predictor we have in this sample is **Hour**. Therefore, we showcase the EDA between **Hour** and **Occupancy** by a table of conditional proportions and a proportional barplot, as shown below:

```
##
##                0           1           2           3           4
##   0 0.044736842 0.045789474 0.043333333 0.044561404 0.044035088
##   1 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
##
##                5           6           7           8           9
##   0 0.044035088 0.044210526 0.040701754 0.023859649 0.013859649
##   1 0.000000000 0.000000000 0.004210526 0.018245614 0.026315789
##
##               10          11          12          13          14
##   0 0.014385965 0.015438596 0.017719298 0.026842105 0.015614035
##   1 0.021578947 0.022280702 0.019298246 0.010877193 0.020350877
##
##               15          16          17          18          19
##   0 0.014385965 0.014912281 0.015263158 0.040000000 0.044210526
##   1 0.020526316 0.021403509 0.022631579 0.003333333 0.000000000
##
##               20          21          22          23
##   0 0.042807018 0.045789474 0.045614035 0.046842105
##   1 0.000000000 0.000000000 0.000000000 0.000000000
```

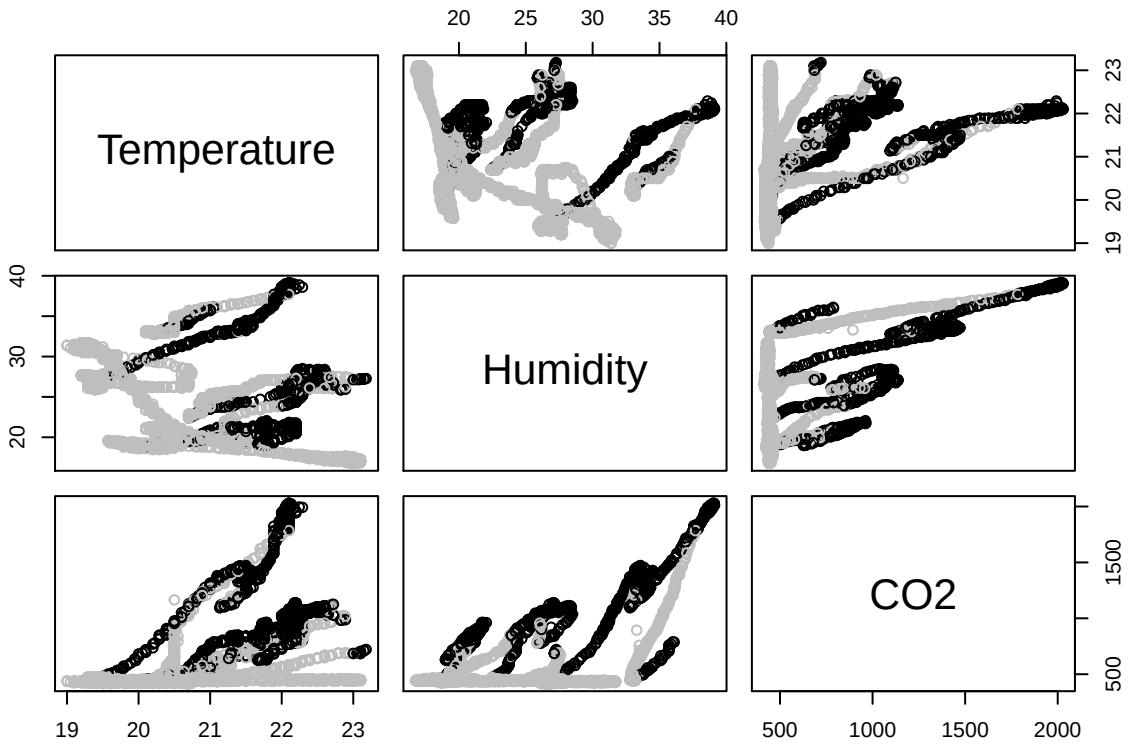


Note: The bars in black indicate room unoccupied, whereas grey bars indicate room occupied.

We see from above that during early morning and late night hours, rooms are more likely to be unoccupied.

EDA on Classification Pairs

To decide which pairs of quantitative predictors should be used to classify **Occupancy**, we examine the bivariate pairs plot below:



Note: The black circles on the plots indicate occupancy=1 (occupied), while the grey circles on the plot indicate occupancy=0 (unoccupied).

From the pairs plot above, we observe that for all combinations, the black circles do not overlap much with the grey circles (the separations are relatively clear), which indicates that they may all be useful combinations to classify the response variable.

Modeling

We start by building and examining the 4 different binary classifiers (namely LDA, QDA, classification tree, and binary logistic regression) on the occupancy training data to classify the binary response variable (**Occupancy**). Moreover, the occupancy test data is used to assess the 4 models.

Linear Discriminant Analysis (LDA)

We build the LDA model using the quantitative predictors (**Temperature**, **Humidity**, and **CO2**), and observe how the LDA classifier performs on the occupancy test data, as follows:

```
##
##        0    1
##   0 1842  116
##   1   75  410
```

After testing the LDA model with our occupancy test data, we obtain an overall error rate of (75+116)/2443 = 0.078. Specifically, the error rate for classifying when a room is unoccupied is 75/(1842+75) = 0.039, while the error rate for classifying when a room is occupied is 116/(116+410) = 0.221.

We see that the LDA model does best at classifying when a room is unoccupied.

Quadratic Discriminant Analysis (QDA)

We build the QDA model using the same quantitative predictors (**Temperature**, **Humidity**, and **CO2**), and observe how the QDA classifier performs on the occupancy test data, as follows:

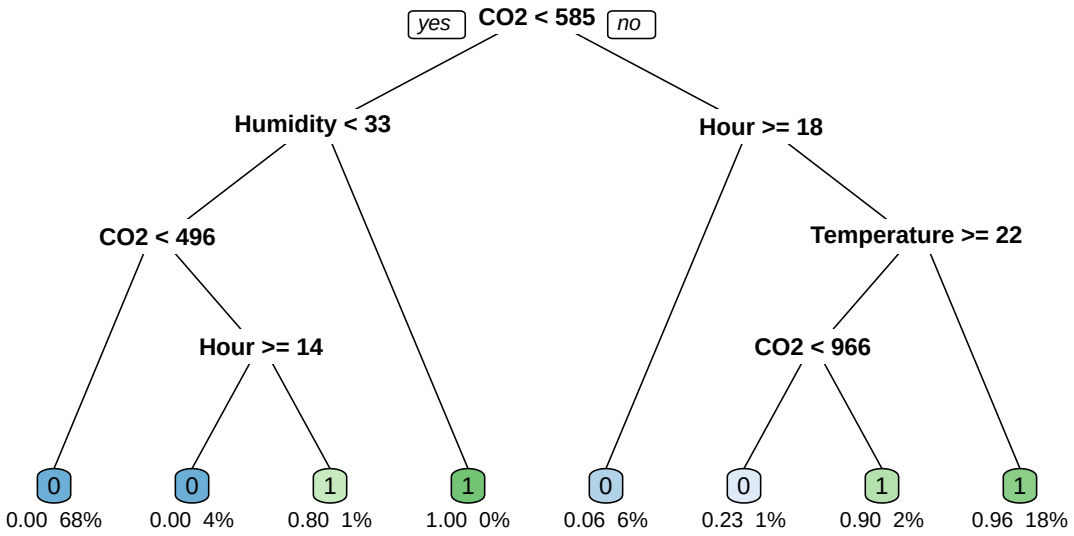
```
##
##        0    1
##   0 1834   98
##   1   83  428
```

For the QDA model, we obtain an overall error rate of (83+98)/2443 = 0.074. Specifically, the error rate for classifying when a room is unoccupied is 83/(1834+83) = 0.043, while the error rate for classifying when a room is occupied is 98/(98+428) = 0.186.

The QDA model performed better than the LDA model, as we can see from the decrease in the overall error rate (0.074 < 0.078). The QDA model also performed better when classifying occupied rooms. However, the QDA performed slightly worse than the LDA model when classifying unoccupied rooms.

Classification Trees

For classification tree, we can build it using both quantitative predictors (**Temperature**, **Humidity**, and **CO2**), and the categorical predictor (**Hour**). Again, we build the classfication tree with the occupancy training data, and observe how it performs on the occupancy test data.




```
##
## occupancy_tree_pred    0    1
##                   0 1883   15
##                   1   34  511
```

The classification tree has an overall error rate of (34+15)/2443 = 0.02. Moreover, the error rate for classifying an unoccupied room is 34/(1883+34) = 0.018, and the error rate for classifying an occupied room is 15/(15+511) = 0.029.

The classficiation tree model performed better than both the LDA and QDA model overall, as well as for identifying unoccupied and occupied rooms.

Binary Logistic Regression

For our last classifier–Binary Logistic Regression, we again build it using both the quantitative predictors (**Temperature**, **Humidity**, and **CO2**), and the categorical predictor (**Hour**). We fit the Binary Logistic Regression on the occupancy training data, and analyze the confusion matrix from the occupancy test data.

```
##
## occupancy_logit_pred    0    1
##                    0 1871   44
##                    1   46  482
```

The binary logistic regression model (with threshold probability = 0.5) has an overall error rate of (46+44)/2443 = 0.037. In addition, the error rate for classifying an unoccupied room is 46/(1871+46) = 0.024, and the error rate for classifying an occupied room is 44/(44+482) = 0.084.

Compared with the previous three classifiers, we observe that the binary logistic regression model performed better than LDA and QDA, but slightly worse than the classficiation tree. We notice the same result with identifying unoccupied and occupied rooms as well.

Final Recommendation on Classifiers

Based on the results above, we see that the classfication tree outperformed the other 3 classifiers. The binary logistic regression performed slightly worse than the classification tree, but the model still seems to be a good one. LDA and QDA performed the worst out of the 4 classifiers overall.

Both LDA and QDA were better at identifying unoccupied rooms than occupied rooms. However, the classfication tree and the binary logistic regression were better at identifying occupied rooms than unoccupied rooms.

After careful evaluations, we decide that the classification tree is the best classifier since it has the lowest overall error rate and the lowest error rates in classifying unoccupied and occupied rooms. The binary logistic regression serve as the secondary recommendation when choosing classfiers since it has the second lowest overall error rate.

Discussion

From what we tested above, we can conclude that our classifier models are all relatively good at classifying whether or not a room is occupied. Again, we recommend the classification tree model due to the fact that it has the lowest overall error rate, and does best at identifying unoccupied and occupied rooms. There doesn't seem to have any obvious overfitting problems in the models, thus it is safe to use the classfication tree.

Other possible predictors that could be taken into consideration are the intensity of light, or whether an air-conditioner/heater is on/off. By intuition, these factors seem to differ when a room is occupied or not. Therefore, future projects should include these factors to make more accurate predictions on room occupancy.