Classification Methods for Occupation of Rooms

Madhumita Krishnan |Applied Data Science|12/07/2018

Introduction

This project deals with Supervised Classification methods such as Logistic Regression and Random Forests. In this project the occupancy of a room based on several factors was determined. These factors include Temperature, Humidity, Light, Co2. Different classification models were trained and tested to determine which model predicts the response more accurately and these results were validated using confusion matrices.

Dataset

The UCI machine learning repository was the source of the Occupancy Detection Dataset. It consisted of a training data set and two testing datasets. The training dataset consisted of around 8143 observations of 7 features which included

- Date time year-month-day hour:minute:second
- Temperature, in Celsius
- Relative Humidity, %
- Light, in Lux
- CO2, in ppm
- Humidity Ratio, in kg-water-vapor/kg-air
- Occupancy, 0 or 1, 0 for not occupied, 1 for occupied status

The two testing datasets consisted of the same attributes as above. Occupancy was used as the binary response variable in the classification methods. The rest of the attributes provided were used as predictors.

Link to the Dataset: https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

Exploratory Data Analysis

The relationship between occupancy and the attributes was visualized using the ggplot function in R. Some of the graphs are shown below.

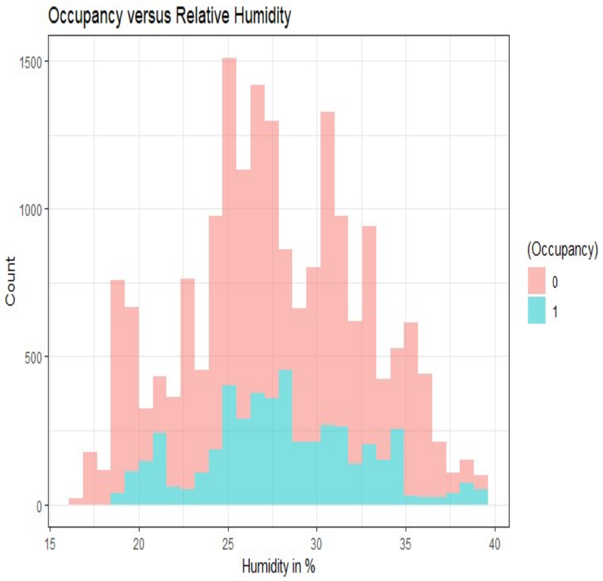


Figure1: Plot for Occupancy versus Humidity

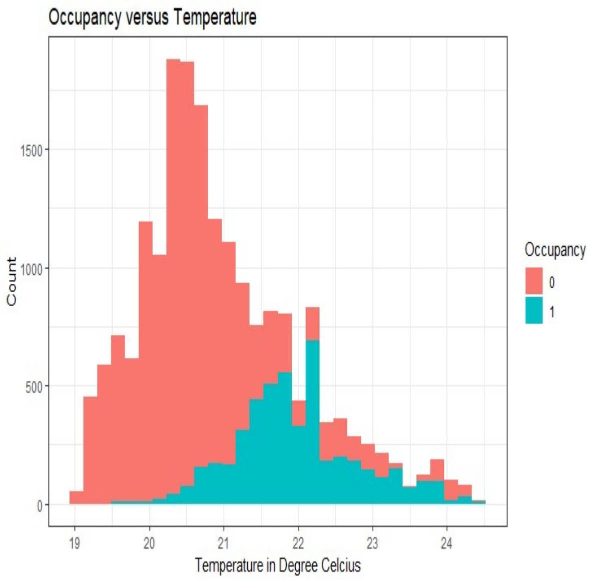


Figure 2: Plot for Occupancy versus Temperature

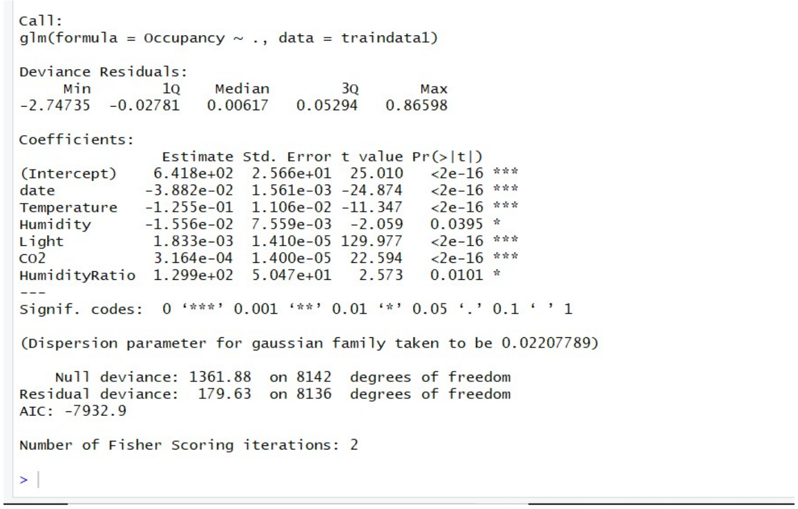
The above figures show the Occupancy Count versus the Relative Humidity and the Occupancy Count versus the Temperature in degree Celsius. It can be seen from Figure 1 that occupancy variation with respect to the humidity is maximum at the middle. On the other hand we can see from Figure 2 that as the Temperature level increases the Occupancy Count also increases and its maximum at the middle. Similarly we can generate the Occupancy count versus the other attributes to get a visual representation.

Model Selection and Validation

The first consideration for modelling was the selection of predictors for the model. For this the generalized linear model was fit at first and a summary was found as seen in Figure 3 which showed us that all the predictors provided were statistically significant from the p-value. To further validate the variable selection, stepwise model selection method was used and found that all the predictors were significant for the models.

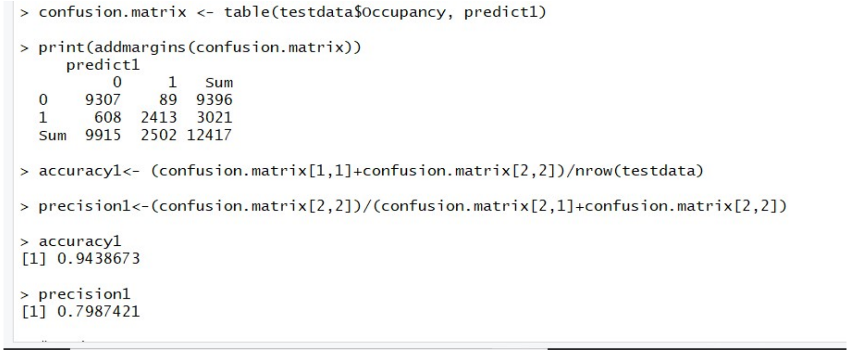
The next step in the project was model selection. Two major models which were implemented under Supervised Classification were Logistic Regression and Random Forests. For Logistic Regression the training data was used to fit the model with Occupancy as the response variable and all the other features as predictors. The two test datasets were combined, and the prediction was done using the test data provided.

Similarly, for Random Forests, the training data was used to fit the model and the Random Forest library was used for this purpose. The model was tested on the combined test dataset.



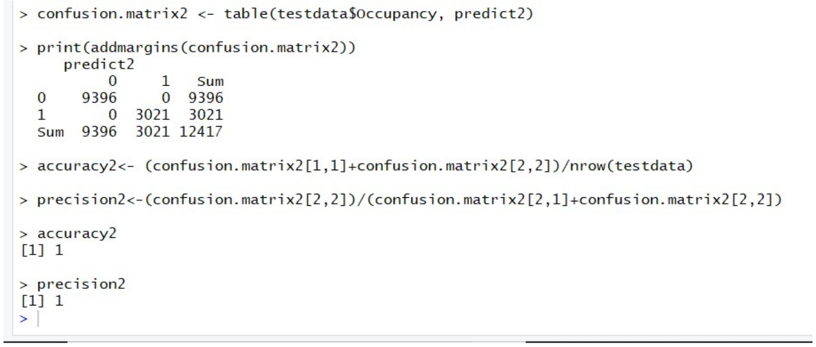
```
Call:
glm(formula = Occupancy ~ ., data = traindata1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.74735  -0.02781   0.00617   0.05294   0.86598

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.418e+02  2.566e+01  25.010   <2e-16 ***
date          -3.882e-02  1.561e-03 -24.874   <2e-16 ***
Temperature   -1.255e-01  1.106e-02 -11.347   <2e-16 ***
Humidity      -1.556e-02  7.559e-03  -2.059   0.0395 *
Light          1.833e-03  1.410e-05 129.977   <2e-16 ***
CO2            3.164e-04  1.400e-05  22.594   <2e-16 ***
HumidityRatio  1.299e+02  5.047e+01   2.573   0.0101 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.02207789)

    Null deviance: 1361.88  on 8142  degrees of freedom
Residual deviance:  179.63  on 8136  degrees of freedom
AIC: -7932.9

Number of Fisher Scoring iterations: 2

>
```

Figure 3: Summary of GLM model



```
> confusion.matrix <- table(testdata$Occupancy, predict1)

> print(addmargins(confusion.matrix))
     predict1
          0     1   Sum
  0    9307    89  9396
  1     608  2413  3021
  Sum  9915  2502 12417

> accuracy1<- (confusion.matrix[1,1]+confusion.matrix[2,2])/nrow(testdata)

> precision1<-(confusion.matrix[2,2])/(confusion.matrix[2,1]+confusion.matrix[2,2])

> accuracy1
[1] 0.9438673

> precision1
[1] 0.7987421
```

Figure 4: Confusion Matrix for Logistic Regression



```
> confusion.matrix2 <- table(testdata$Occupancy, predict2)

> print(addmargins(confusion.matrix2))
     predict2
          0     1   Sum
  0    9396     0  9396
  1       0  3021  3021
  Sum  9396  3021 12417

> accuracy2<- (confusion.matrix2[1,1]+confusion.matrix2[2,2])/nrow(testdata)

> precision2<-(confusion.matrix2[2,2])/(confusion.matrix2[2,1]+confusion.matrix2[2,2])

> accuracy2
[1] 1

> precision2
[1] 1
>
```

Figure 5: Confusion Matrix for Random Forests

For Model Validation the Confusion Matrix was used. The first Confusion Matrix was created for Logistic Regression as it can be seen in Figure 4. This matrix shows the number of observations under Occupancy that were classified correctly and the ones that were not. The accuracy and precision among other characteristics of the model can be determined from the confusion matrix. Here the accuracy is 94.38% which is high accuracy for the model. We can also see that the precision is not too high for this model.

Similarly Confusion Matrix was created for Random Forests by comparing the actual occupancy values with the ones predicted and the Random Forests can be seen to have 100% accuracy and 100% precision in prediction.

Conclusion

Some other models that can be tested in the future include support vector machines. Other factors could also help predict room occupancy such as rent and income. From the above models it can be concluded that Random Forests is the most ideal model for classification of room occupancy in terms of precision and accuracy.

Reference

Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Luis M. Candanedo, VÃ©ronique Feldheim. Energy and Buildings. Volume 112, 15 January 2016, Pages 28-39.