Room Occupancy Detection

By Aishwarya Srivastava

INTRODUCTION:

An experiment was conducted where time stamped pictures were taken, and it was found that Temperature, Humidity, Light and CO2 play a major role in determining the occupancy of a room. A multivariate time series dataset was obtained which has 7 attributes: Date, Temperature, Humidity, Light, CO2, HumidityRatio and Occupancy. Supervised classification is used to predict the occupancy of rooms taking into consideration the above attributes.

Following steps have been taken for building a model:

- Checking for NA values.
- Exploratory Data Analysis
- Variable selection using Stepwise AIC, Best Subset selection and Domain knowledge
- Model Building using Logistic Regression, KNN and Random Forest for classification.
- Validation using the Confusion Matrix considering Accuracy, Precision, Recall.

DATASET:

Dataset has been obtained from UCI Machine Learning Repository. Three datasets are provided for training and testing. Training dataset contains 8143 rows for training and two testing dataset contains 2665 and 9752 rows with 7 variables in all three datasets. The variables in the dataset are as follows:

- Date [Format: year-month-day hour:min:sec]
- Temperature; in Celsius
- Humidity
- Light; in Lux
- CO2; in ppm
- HumidityRatio; derived from temperature and humidity, in kgwater-vapor/kg-air
- Occupancy; 0 or 1, 0 for the room not occupied and 1 for room occupied

Here our response variable is Occupancy which is a binary attribute.

EXPLORATORY DATA ANALYSIS:

To gain more insight of the dataset, Exploratory Data Analysis was performed. Below are the graphs for EDA.

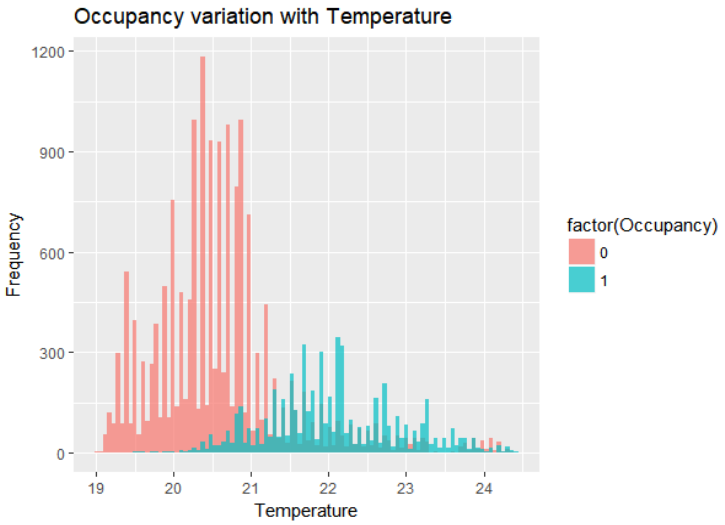


Figure 1a: We observe that at low temperature the occupancy is mostly 0, as the temperature reaches an optimal level between 21°C to 24°C the occupancy is 1.

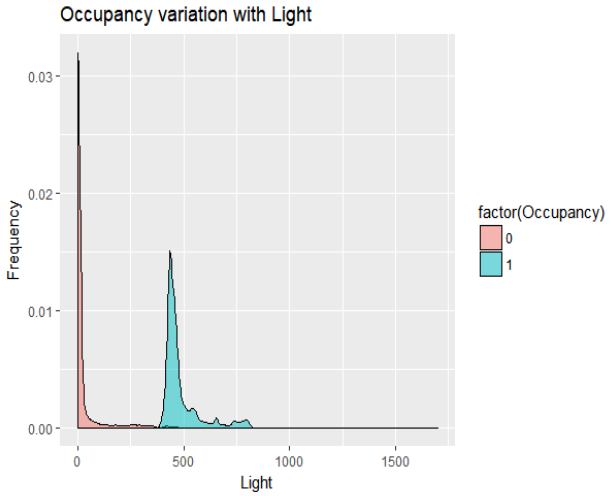


Figure1b: We observe that in low light the occupancy is 0, while when the light is about 450 lux the occupancy is 1.

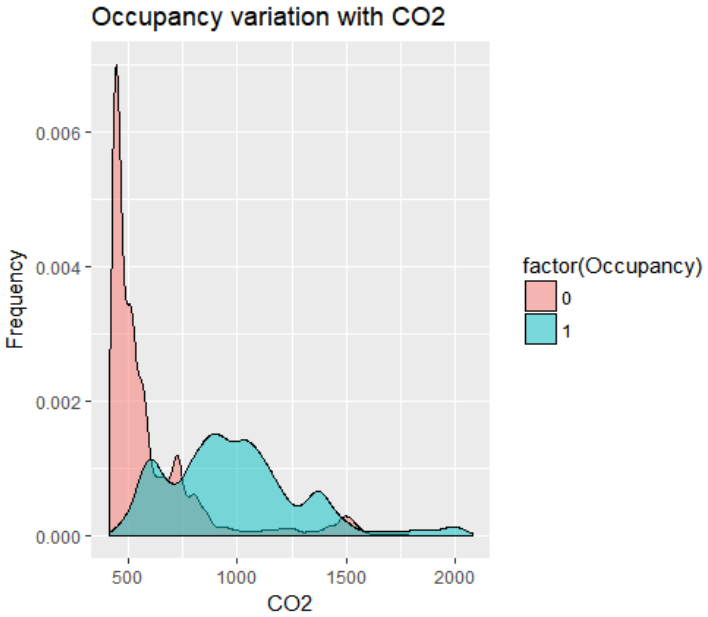


Figure 1c: Occupancy varies with the amount of CO2, and the occupancy is minimum at a lower CO2 level.

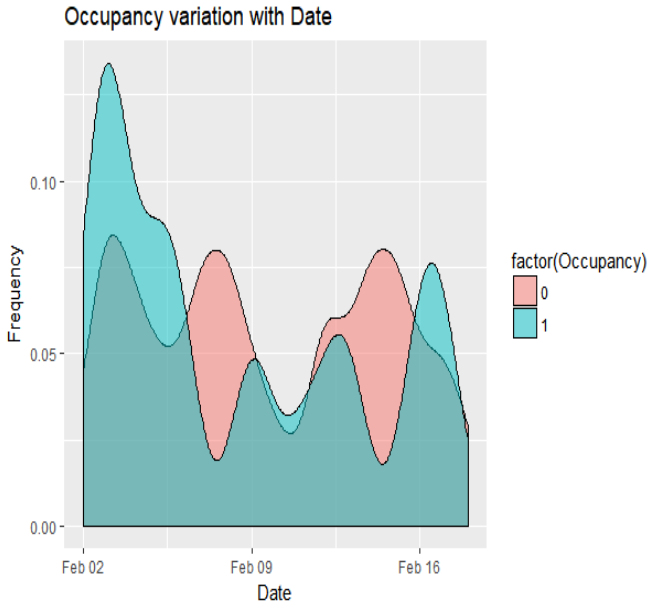


Figure 1d: Variation of occupancy with respect to days and is maximum during the weekends i.e. 7th Feb 2015 and 14th Feb 2015.

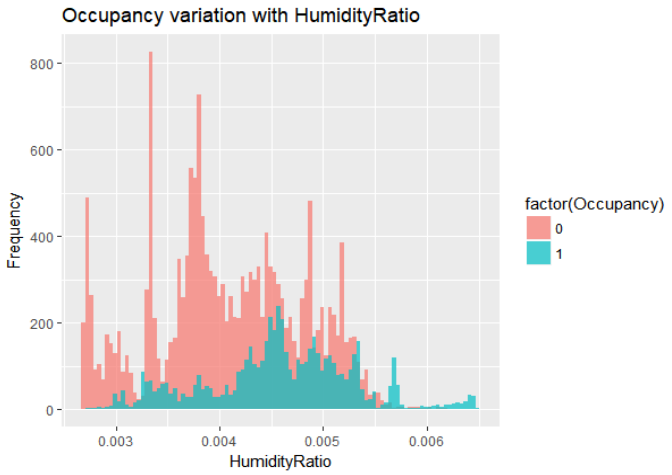


Figure 1e: This graph shows the variation of Occupancy with the Humidityratio in percentage.

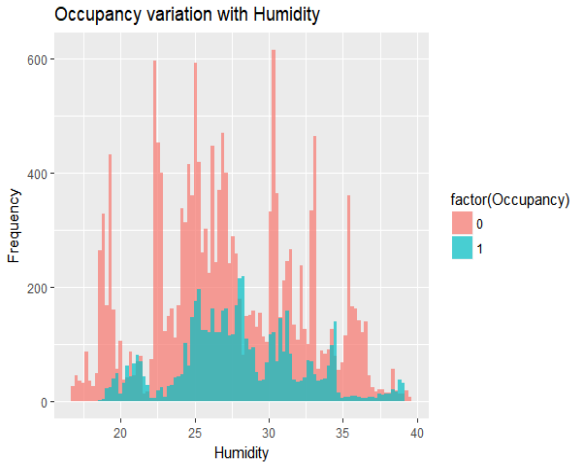


Figure 1f: This histogram shows the variation of Occupancy with the humidity. Not much variation is visible.

MODEL SELECTION AND VARIATION:

After performing EDA, we select variables which have a significant effect on the model by applying Stepwise AIC and Best Subset Selection.

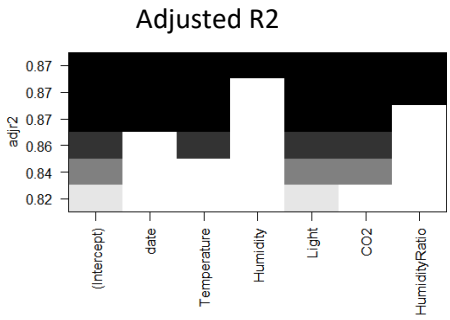
Stepwise AIC yields the below results, with the final model being the one with all six predictors.

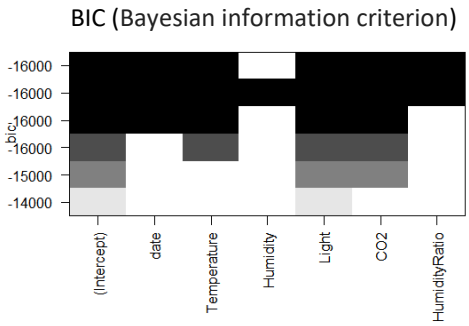
```
116  step = stepAIC(full,trace=FALSE)
117  step$anova
118
119
117:1  (Untitled)
Console  Terminal
~/

Initial Model:
Occupancy ~ date + Temperature + Humidity + Light + CO2 + HumidityRatio

Final Model:
Occupancy ~ date + Temperature + Humidity + Light + CO2 + HumidityRatio


  Step Df Deviance Resid. Df Resid. Dev       AIC
1                       8136   179.6257 -7932.885
```

Figure 2: Code snippet with the output after applying stepwise AIC for variable selection.

Applying Best Subset Selection to verify the results by AIC. Select the model on basis of R square value and BIC. The black block below shows if the variable is included in the model and white means they are not included. R square value of 0.87 includes a model with all the parameters while Minimum BIC value of -16000 includes a model with all parameter except Humidity.

Humidity and Humidityratio are relatively the same terms therefore we include just the humidity ratio. Thus, on basis of Best Subset Selection and domain knowledge, we select Date, Temperature, Humidityratio, Light and CO2 as the predictors.





Now, applying Logistic Regression, K nearest neighbor and Random Forest classification algorithms and validating the results with the Confusion matrix by considering the Precision, Accuracy, Recall and the Misclassification Rate.

```
> confusion_matrix_forest <- confusionMatrix(random_forest
> confusion_matrix_forest
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1693    0
         1    0  972

               Accuracy : 1
                 95% CI : (0.9986, 1)
    No Information Rate : 0.6353
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.6353
         Detection Rate : 0.6353
   Detection Prevalence : 0.6353
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

Fig3a: Random Forest

```
          Reference
Prediction    0    1
         0 1641   52
         1   26  946

               Accuracy : 0.9707
                 95% CI : (0.9636, 0.9768)
    No Information Rate : 0.6255
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9372
 Mcnemar's Test P-Value : 0.004645

            Sensitivity : 0.9479
            Specificity : 0.9844
         Pos Pred Value : 0.9733
         Neg Pred Value : 0.9693
             Prevalence : 0.3745
         Detection Rate : 0.3550
   Detection Prevalence : 0.3647
      Balanced Accuracy : 0.9661

       'Positive' Class : 1
```

Fig3b: KNN algorithm

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1639    3
         1   54  969

               Accuracy : 0.9786
                 95% CI : (0.9724, 0.9838)
    No Information Rate : 0.6353
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9544
 Mcnemar's Test P-Value : 3.528e-11

            Sensitivity : 0.9681
            Specificity : 0.9969
         Pos Pred Value : 0.9982
         Neg Pred Value : 0.9472
             Prevalence : 0.6353
         Detection Rate : 0.6150
   Detection Prevalence : 0.6161
      Balanced Accuracy : 0.9825

       'Positive' Class : 0
```

Fig3c: Logistic Regression

Random Forest Classification algorithm has an accuracy of 100%, as the dataset has been made to fit this model instead of model being fitted to a dataset. The dataset provided on UCI is already cleaned and processed.

We can validate the results from the three algorithms using a Confusion Matrix, which takes into consideration only 5 predictors namely: Time, CO2, Light, Humidityratio and Temperature.

MODEL	ACCURACY%	RECALL%	PRECISION%
Logistic Regression	97.86	96.81	99.82
K Nearest Neighbor	97.07	94.79	97.33
Random Forest	100	100	100

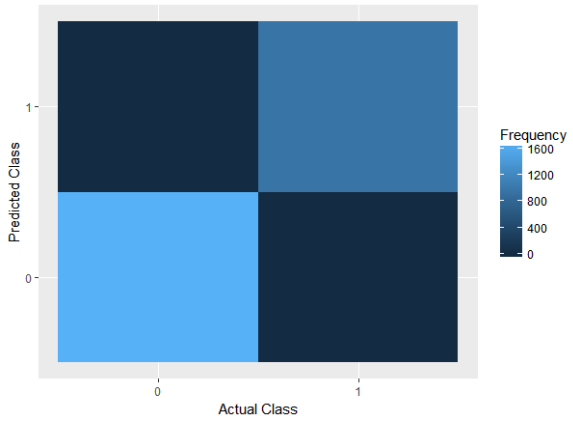


Fig 4: This is a confusion matrix for the best model i.e. Random Forest Classification algorithm which yields an accuracy of 100% and a recall of 100%

CONCLUSION:

We can draw an inference of the occupancy of a room by considering external factors which would in turn help in energy conservation and design the rooms accordingly. We effectively predicted the occupancy of room using Random Forest Classification Algorithm considering 5 predictors: Date, Humidityratio, CO2, Temperature and Light. The accuracy of the model and sensitivity is 100% using Random Forest. Thus, Random Forest proves to be an appropriate model instead of KNN and Logistic Regression. We learn that if we have a dataset with these five predictors we can effectively predict the occupancy of a room.

REFERENCES:

1) 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.