Room Occupancy Prediction

(1) A literature review on this problem.

Room occupancy prediction is useful in energy saving. “Experimental measurements reported that energy savings was 37% and between 29% and 80% when occupancy data was used as an input for HVAC control algorithms” (qtd. in Candanedo and Feldheim). Intuitively, higher accuracy rate results in higher energy saving and hence the goal for room occupancy prediction is to find the best learning models for the prediction while maintaining the adequate indoor comfort standards.

The paper “Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models” by Luis M. Candanedo and Véronique Feldheim does a comprehensive research about the accuracy difference of determining occupancy in buildings using various approaches and algorithms. They first introduce and summarize the existing works in occupancy modeling. Knowing that the performance of some classification models, such as LDAs, RFs, and GBM, has not been evaluated in the previous work, they then decided to conduct their own modeling using CART, RF, GBM, and LDA, with temperature, humanity, light, CO_2, humanity ratio, number of seconds from midnight for each day, and a week indicator as variables. Finally, they found out that it is also possible to obtain high accuracies in the determination of occupancy with RF, CART and LDA models. The correlation between variables is also an important determinant that can alter the accuracy rate. That is, high accuracies (around 97%) were found when using only two predictors (temperature and light, light and CO_2 and light and humidity, light and humidity ratio) with the LDA models.

Candanedo and Feldheim first introduce previous work in occupancy modeling by analyzing various types of approach and algorithm used by each work, and compare their corresponding accuracy rate. One occupancy model was based on stochastic approach such as Bayesian statistics to deal with the data collected from digital video cameras, passive infrared detection, CO_2 sensors, and successfully reduced the average error from 70% to 11%. Other occupancy models either used time series to identify differences between weekend and weekdays, or relied on both multivariate Gaussian distribution and Agent Based Model (ABM) to predict occupancy. Another approach, on the other hand, used ABM and added noise to the data to simulate the situation that may come from short-period occupancy. A model had wireless sensor network to collect data and used EnergyPlus to estimate the energy consumption.

Candanedo and Feldheim then focus on the works that have real time occupancy detection, which can be conducted by wide range of methods including pattern discovery model for the prediction of user behavior, hidden Markov models, Support Vector Machines (SVM), Neutral Networks (NN), relative information gain (RIG), PIR sensor, video camera, passive-infrared sensors with the extension of Kalman filter, wireless sensor network, and Latent Dirichlet Allocation. The corresponding accuracy rates from the above models in the previous works are ranged from 73% to 80%. In particular, the real time occupancy detection using decision tree had 97.9% reported accuracy when using data from a passive infrared motion sensor. However, a decrease in the accuracy was found when adding to many sensor readings that have potential to cause overfitting according to the authors of that paper.

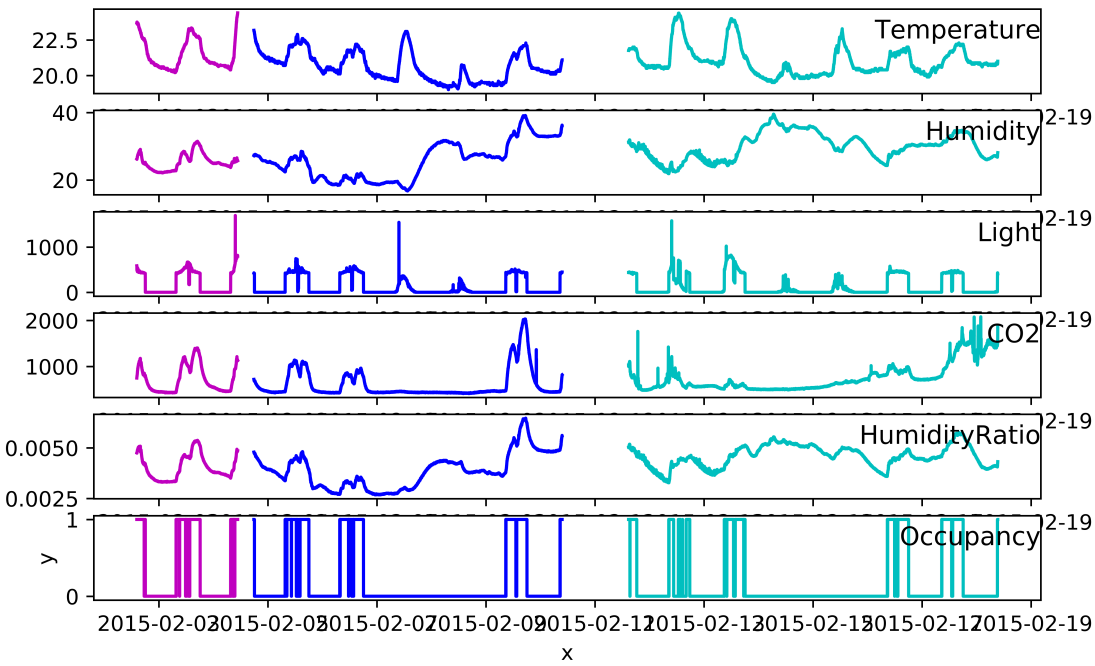
Candanedo and Feldheim then construct their own work using sensors with high resolution to collect data and R programming language to carry out the

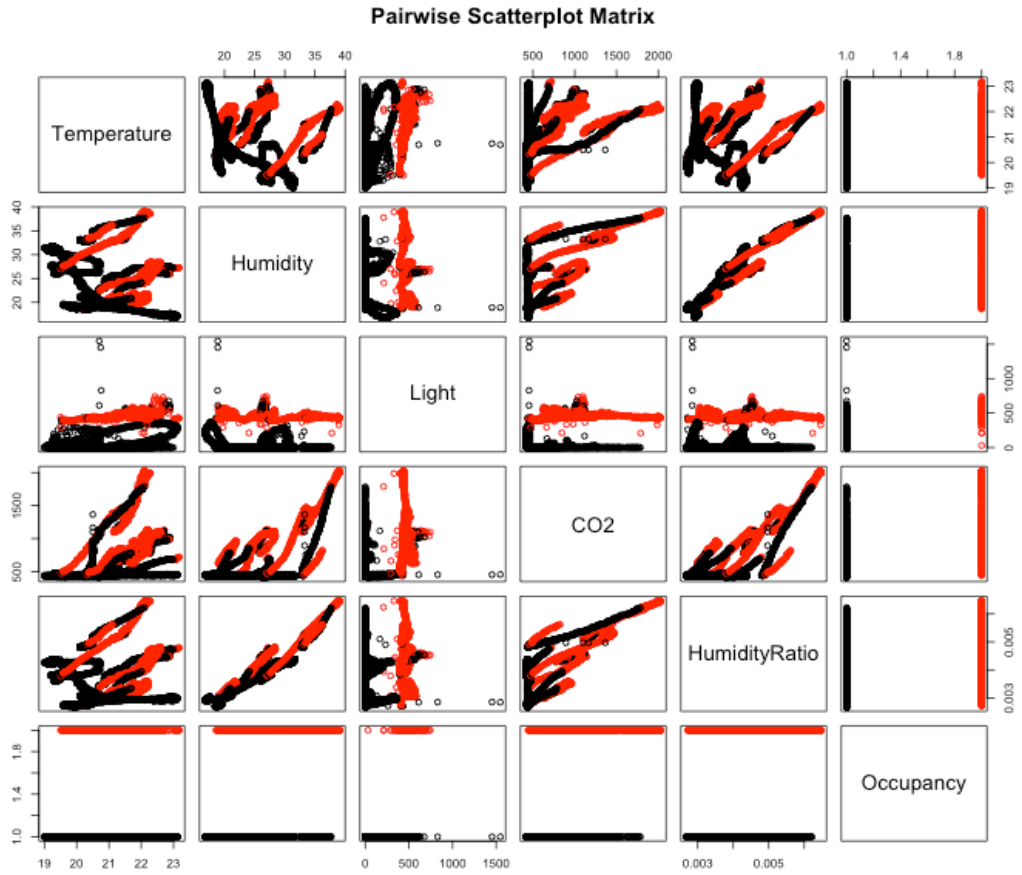
calculations. They computed the covariance matrix and visualized it as a pairs plot. The result shows that there are correlations between natural phenomena such as CO2, temperature, and humidity. Nevertheless, the pair combinations of temperature and light, humidity and light, CO2 and light, humidity ratio and light, show no evidence of correlation from the plot. Therefore, they are good candidates for training the classification models. Candanedo and Feldheim use one training set and two test sets. The training set was taken with the door closed during occupied status. One of the test sets was taken with the door opened and the other one with the door closed. The trained and tested statistical models in their work are CART, RF, GBM and LDA. The accuracy rate then varies from different models and number of predictors used with some interesting correspondences. For example, Random Forest models with the different parameter combinations seem to have a much larger accuracy in the training set than in the test sets. Number of trees used should be no less than 200 to stably control the OOB rate. The maximum accuracies for both tests are all from LDA model with adequate parameters used. That is, LDA models seem to provide a consistent accuracy prediction for all the cases in the training and test sets. This is especially true when the parameters used in the model are uncorrelated.

Work Cited

Candanedo Ibarra, Luis & Feldheim, Veronique. (2015). Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Energy and Buildings. 112. 10.1016/j.enbuild.2015.11.071.

(2) Explore the data through visualization tools and exploratory statistics.





The red data correspond to the occupied status and black to the not occupied status

Correlation Matrix

	Temperature	Humidity	Light	CO2	HumidityRatio
Temperature		-0.14175931	0.64994184	0.5598938	0.1517616
Humidity	-0.1417593		0.03782794	0.4390228	0.9551981
Light	0.6499418	0.03782794		0.6640221	0.2304202
CO2	0.5598938	0.43902276	0.66402206		0.6265559
HumidityRatio	0.1517616	0.95519808	0.23042021	0.6265559	

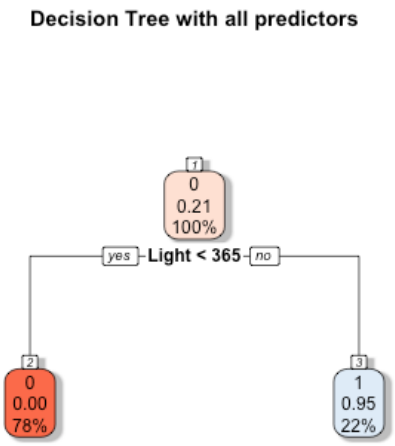
(3) Try and compare different classifiers. (set seed = 1234)

The trained and tested models in this project are Bagging, Classification Decision Tree, Linear Discriminant Analysis (LDA), and Support Vector Machines (SVM).

Here is the accuracy table for the above models using all the predictors (T = Temperature, H = Humidity, Light = L, CO2, and HumidityRatio = HR).

Bagging: mtry = 5

Classification Decision Tree:.



R1={X | LIGHT < 365}, R2={X | LIGHT >= 365}

LDA: prior probabilities: P (Occupancy = 0) = 0.788, P (Occupancy = 1) = 0.212

SVM: gemma = 0.2 (default), penalty parameter (cost) = 1 (default)

Summary of the results:

	Parameters	Training Accuracy (%)	Testing Accuracy (%)
Bagging	T, H, L, CO2, HR	99.34	96.95
Decision Tree	T, H, L, CO2, HR	98.78	99.31
LDA	T, H, L, CO2, HR	98.78	98.76
SVM	T, H, L, CO2, HR	98.88	95.31

(4) Fine tuning one or more classifiers to achieve higher prediction performance.

Since the accuracy performance for bagging is not perfect, which means that the predictor size m can be better than choosing p-1, where p is the number of all variables. Therefore, I will use Random Forest (RF) instead of bagging and use tuning method to find the best predictor size m for the model.

Here is the result for the tuning result:

mtry	Accuracy	Kappa
2	0.9872437	0.9725809
3	0.9877435	0.9736546
4	0.9876184	0.9733847
5	0.9866180	0.9712133

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 3.

Hence, the RF model hits the best accuracy when the predictor size m = 3.

Finally, the training accuracy for bagging before the tuning is 99.34% and is 99.39% after the tuning using RF; the testing accuracy for bagging before the tuning is 96.95% and is 97.30% after the tuning using RF.

Similarly, since the accuracy performance for SVM is not good when using default gamma value and penalty parameter (cost), and the fact that SVM is quite sensitive to the choice of parameters, it is reasonable and common to tune the game and cost in SVM model to find the best values for them that minimize the classification error.

Here is the short summary of using tune.svm() in R to tune the parameters:

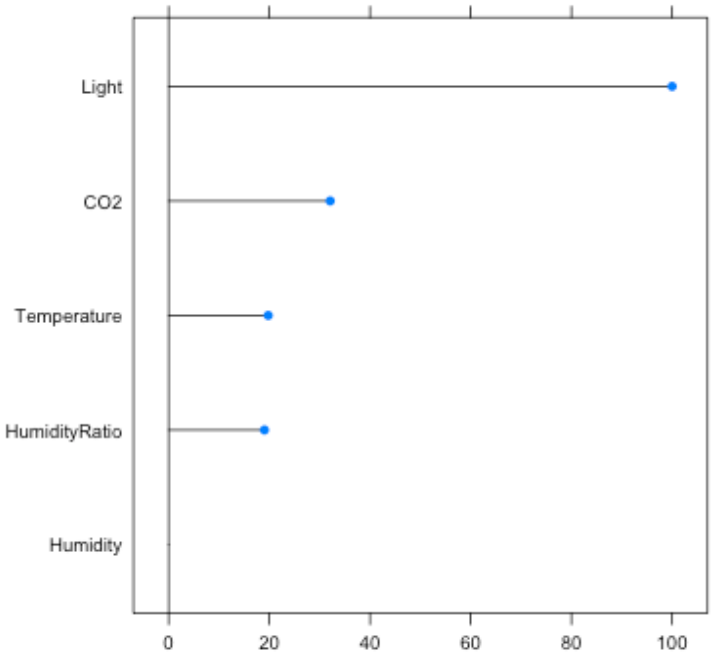
Parameter tuning of 'svm':

sampling method: 10-fold cross validation

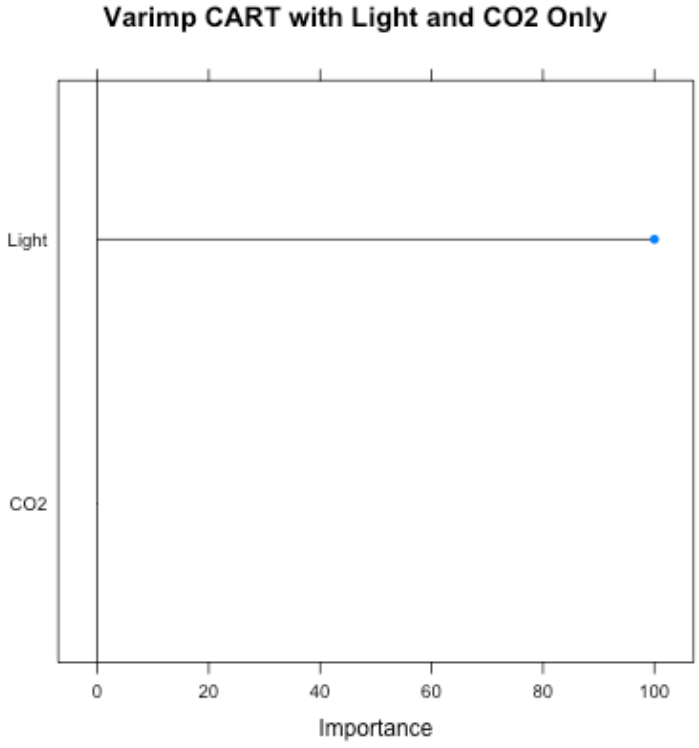
best parameters:

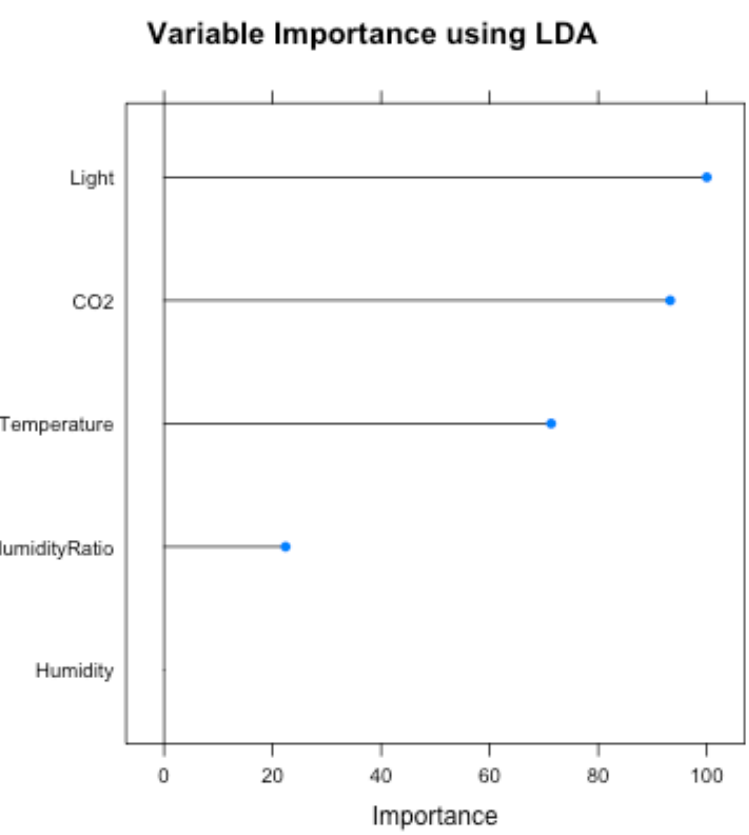
gamma	cost
0.08	0.5

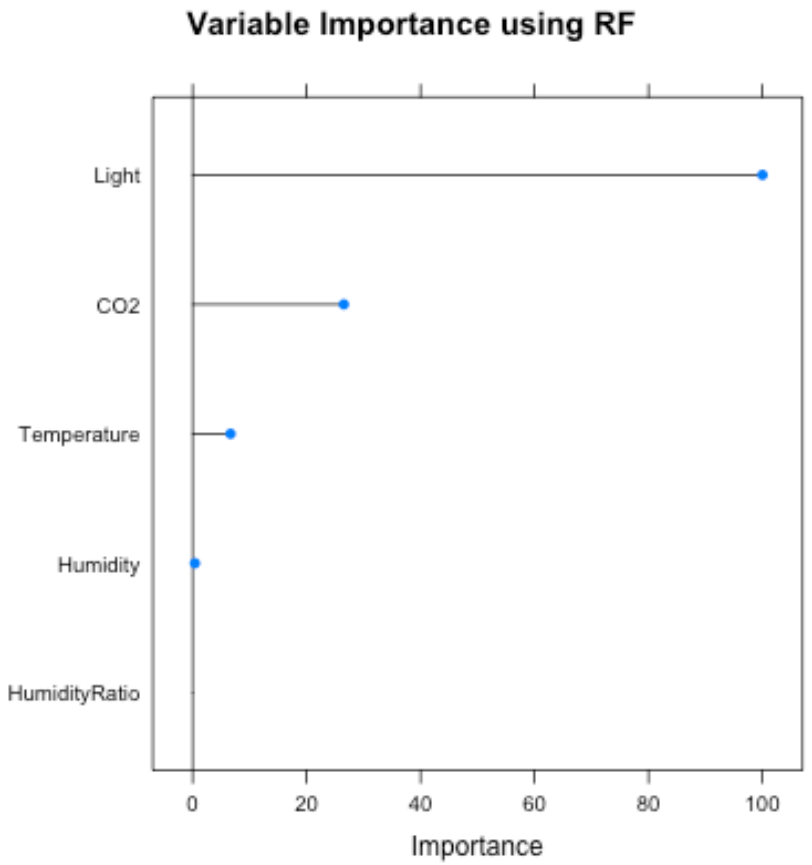
best performance: 0.02064149

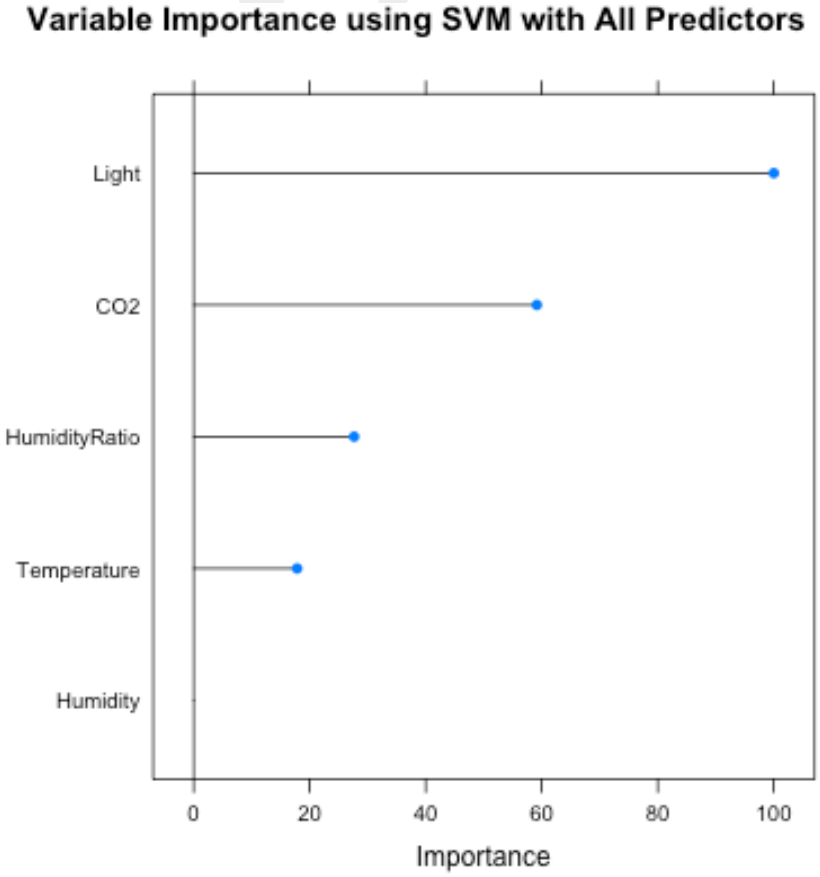


VARIABLE IMPORTANCE USING DECISION TREE WITH ALL PREDICTORS




Variable Importance using LDA
Light
CO2
Temperature
HumidityRatio
Humidity
0
20
40
60
80
100
Importance


Variable Importance using RF
Light
CO2
Temperature
Humidity
HumidityRatio
0
20
40
60
80
100
Importance


Variable Importance using SVM with All Predictors
Light
CO2
HumidityRatio
Temperature
Humidity
0
20
40
60
80
100
Importance

Therefore, the training accuracy for SVM before the tuning is 98.88% and is 95.31% after the tuning; the testing accuracy for SVM before the tuning is 98.87% and is 98.39% after the tuning.

The following table shows the summary of accuracy rate before and after the tuning:

	Training Accuracy (%) Before Tuning	Training Accuracy (%) After Tuning	Testing Accuracy (%) Before Tuning	Testing Accuracy (%) After Tuning
Bagging -> RF	99.34	99.39	96.95	97.30
SVM	98.88	98.87	95.31	98.39

(5) Conclusion

I am satisfied with all of the methods I used in this project. In particular, the significant improvement in accuracy when applying the tuning parameters for SVM is really impressive. Generally speaking, none of the methods I used in the project have relatively low performance even without applying tuning parameters or changing the formula. That is, the overfitting problem does not occur obviously even if I include all features in the prediction formula, mainly because the variable light is almost uncorrelated with the rest of the variables according to the pairwise correlation matrix, which makes the variable becomes an exceptional predictor compared to others.

Moreover, by looking at the plots of relative variable importance for each of the model, we can easily see that light has significant importance compared to others. After all, as long as light is included in the formula computed in the chosen model, a good accuracy rate of predicting the response variable occupancy will be expected.