Room Occupancy Prediction Problem

Table Of Contents

1. Introduction
2. Literature Review
3. Exploratory Analysis and Data Visualization
4. Method Performance Summary
5. Results Comparison
6. Conclusions

Introduction

Globally, greenhouse gas emissions are the single largest contributor to climate change. The most prominent and common gas emitted is Carbon Dioxide, the bi-product of most forms of energy production, industrial processes, and transportation vehicles. In every industry, organizations are constantly seeking ways to cut down on carbon dioxide emissions, namely to save money on their electricity bills. In the electricity bill of the typical urban office of today, 70% of the overall consumption and cost stems from the use of the air conditioning system, which runs most of the day to heat and cool the building to a perfect working temperature. Although these systems can be controlled manually, most systems of today are automated to prevent human error and optimize energy usage. Ideally, the systems should run only when necessary (when office workers are in the building and the air requires conditioning). This paper will document the research and experiments conducted to build models to automatically predict whether a room is occupied or not, based on attributes such as room temperature, humidity and light.

Literature Review

We began our experiment conducting background research on the topic, focused on papers written on the topic directly. The most relevant pieces are summarized below:

Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models - Luis M. Candanedo and Veronique Feldheim

The report first summarizes previous research and experiment results, from which the authors drew a few key conclusions: Poor sensor calibration caused inaccurate results, too many features often causes a decrease in model accuracy, the date/time variable had never been employed in classification, and the performance of classification models such as LDA, RF, and GBM had never been tested. For these reasons, the authors conducted an additional experiment,

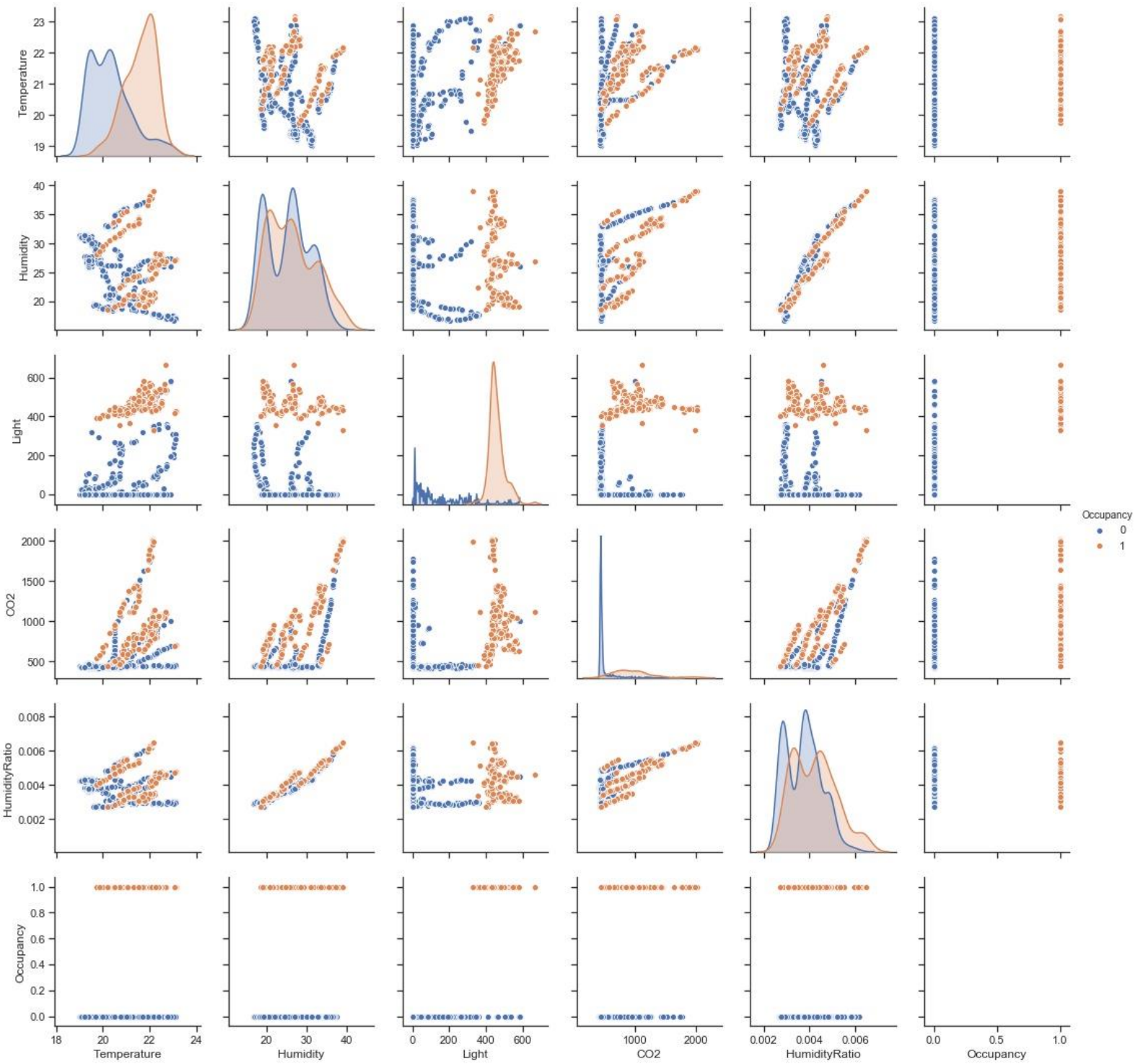
hoping to accurately predict the occupancy in an office room using data from light, temperature, humidity, and CO2 sensors. Their best results were found training Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), and Random Forest (RF) models, which produced accuracies between 90-95%. Timestamp information was often included in the models as well, and often produced more accurate results. Notably, the LDA model was able to estimate occupancy of about 84% using just one predictor variable (temperature).

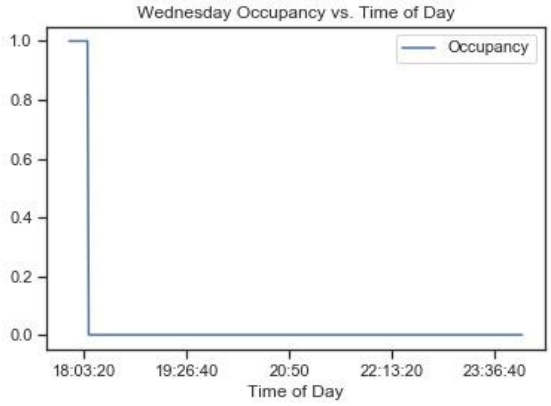
Occupancy Measurement In Commercial Office Buildings For Demand-Driven Control Applications- A Survey And Detection System Evaluation

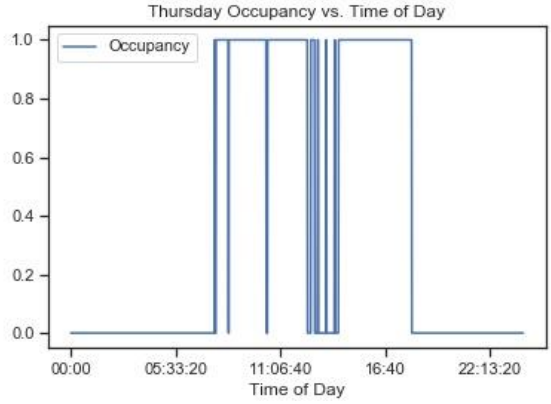
This report evaluated the accuracy of current methods for detecting occupancy within a building. Office buildings represent the largest in floor area in most developed nations, and therefore consume a large amount of energy. This report evaluated the accuracy of several occupancy detection systems including: CO2-based, passive infrared, ultrasonic, image, sound, electromagnetic signal, energy measurement, computer activity, sensor fusion, and (experimental factor) in-chair sensors. Their conclusions were limited by the stochastic behavior of humans in an office environment. The research did, however, conclude that ventilation systems which can be dynamically varied as opposed to constant ventilation did improve overall building energy performance. In-chair sensors were proven to be a reliable measure of building occupancy, more so than existing methods of determining accuracy.

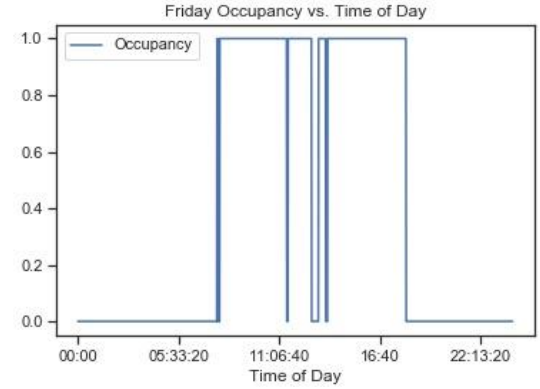
Exploratory Analysis and Data Visualization

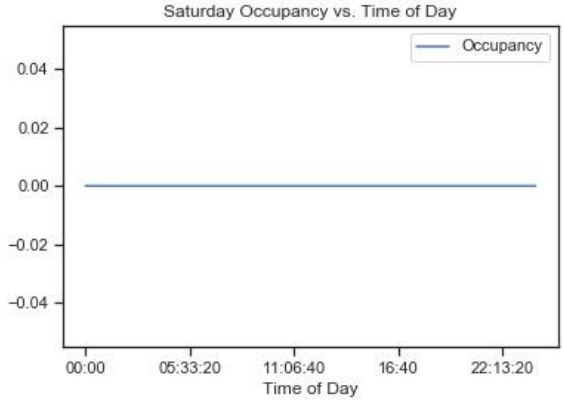
For consistency, clarity, and visibility in this section, we will first display the various figures, and then discuss them in full at the bottom of this section.

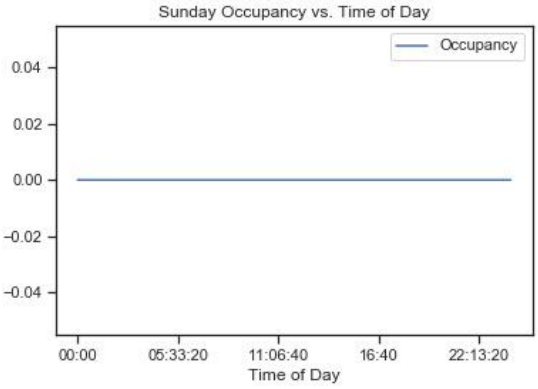

Temperature
Humidity
Light
CO2
HumidityRatio
Occupancy
Occupancy
0
1

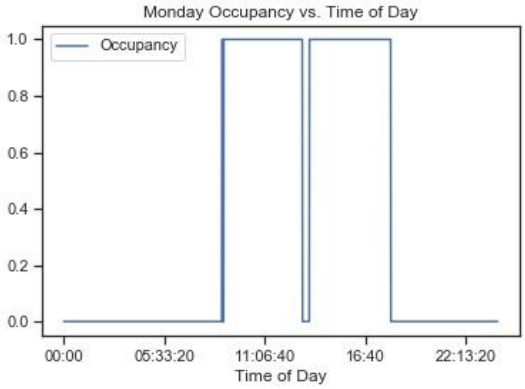

Wednesday Occupancy vs. Time of Day
Occupancy
1.0
0.8
0.6
0.4
0.2
0.0
18:03:20
19:26:40
20:50
22:13:20
23:36:40
Time of Day

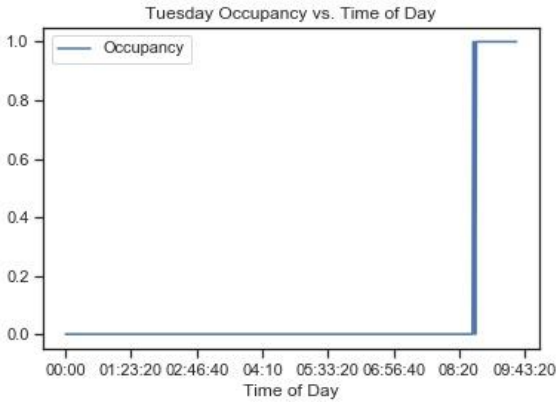

Thursday Occupancy vs. Time of Day
Occupancy
1.0
0.8
0.6
0.4
0.2
0.0
00:00
05:33:20
11:06:40
16:40
22:13:20
Time of Day


Friday Occupancy vs. Time of Day
Occupancy
1.0
0.8
0.6
0.4
0.2
0.0
00:00
05:33:20
11:06:40
16:40
22:13:20
Time of Day


Saturday Occupancy vs. Time of Day
Occupancy
0.04
0.02
0.00
−0.02
−0.04
00:00
05:33:20
11:06:40
16:40
22:13:20
Time of Day


Sunday Occupancy vs. Time of Day
Occupancy
0.04
0.02
0.00
−0.02
−0.04
00:00
05:33:20
11:06:40
16:40
22:13:20
Time of Day


Monday Occupancy vs. Time of Day
Occupancy
1.0
0.8
0.6
0.4
0.2
0.0
00:00
05:33:20
11:06:40
16:40
22:13:20
Time of Day


Tuesday Occupancy vs. Time of Day
Occupancy
1.0
0.8
0.6
0.4
0.2
0.0
00:00
01:23:20
02:46:40
04:10
05:33:20
06:56:40
08:20
09:43:20
Time of Day

First, we will discuss the scatterplot matrix as displayed on page 3. In terms of methodology, at first we attempted to plot the entire dataset, but we found that this data became incredibly cluttered with noise, and was not useful in acquiring any information from the dataset. So, instead of using the entire set, we randomly sampled one thousand entries without replacement, and plotted those pairwise using python's seaborn package instead, using color to denote occupancy (blue for unoccupied, orange for occupied). From this we were able to see a number of interesting phenomena that may be worth investigating.

The first, and possibly most striking is the relationship that light and occupancy have across all scatter plots that include light. A distinct clustering effect is visible in each plot as light increases, which intuitively makes sense; people are likely to turn the lights on as they enter the room, and turn them off when they exit. This shows us that light might be a good predictor variable in order to determine occupancy.

Some other relationships and correlations of interest are that temperature and occupancy don't seem to depend on each other all that much based on this visualization, and that CO2 has a positive correlation with both humidity and temperature. These positive correlations may lead us to the ability to perform some dimensionality reduction, however, further investigation is required for this.

The other figures that we created were a time series of occupancy for each day of the week (created with the python package pandas), so that we can see how different days of the week may impact the occupancy of a room. We immediately see that there are some strange properties in the graphs of Tuesday and Wednesday, in that their patterns and scales are drastically different from the other days, however, upon further investigation of the data set, we find that Wednesday afternoon is when the data collection began, and Tuesday morning is when it ended, thus explaining our initial anomaly.

Beyond this, the differences between the other weekdays appear merely superficial. While there seems to be more spurts of unoccupied-ness in the afternoon of Thursday than the afternoons of Friday or Monday, this is most likely just noise, as when reflecting on the nature of these days, it doesn't seem likely to have a true causation for this pattern.

Lastly, this office seems to not have anyone working on the weekends, as there are no occupied entries for either of these days of the week, so this is probably a very good predictor of whether or not the office is occupied or not.

Method Performance Summary

In this section, we first give a brief background of three algorithms we want to compare and why we chose them. Then we summarize our implementation of these three algorithms and the parameter we chose for each algorithm.

Linear Discriminant Analysis - "finds a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification." (LDA Wiki)

Summary: In order to find the prior probabilities, we used the training data priors choice. The prior probability for 1 occupancy was 0.2123 and for 0 occupancy was 0.7877. Then, we calculated the column means and covariance matrices of each occupancy for the columns Temperature, Humidity, Light, CO2, and HumidityRatio. Using the covariance matrices, we calculated the pooled covariance matrix. We then calculated the alpha and beta values for each occupancy. With the alpha and beta values, we were able to calculate linear discriminant functions for each occupancy. We then classified the observation to the occupancy with highest function value. We, then, checked to see how many predictions we got correct. We ended up with 98.5% accuracy.

Random Forest - "ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees." (Random Forest Wiki)

Summary: The reason why we want to use random forest is that at each split in the tree, the algorithm is restricted to consider only a small subset from available predictors. This minor trick can moderate the effect of a strong predictor when we split the data, which is a serious drawback of bagging. Therefore, random forest overcome this correlation problem and other predictors will have more chances when splitting.

Since we are given a validation set, we use the validation set to tune the hyperparameter "mtry", which is the number of predictors for each subset. We tried different value of mtry from 1 to 5 and find out that 1 is the best value. So we let mtry=1 and test on the testing set. The accuracy we finally got is around 97%.

Support Vector Classifier - **"**supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis." (Support Vector Machine Wiki)

Summary: Given the training data, SVM will output an optimal hyperplane that can divide the points into two sides where each side contains a class. In this research, we use the linear kernel for SVM. In order to determine the hyperparameter "C", which is the constant of the regularization term, we chose 7 different values ranging from 0.01 to 100 and used validation set to determine the best C. The result we got was C=10. Finally, we generated a SVM model using C=10 and tested with the testing set. We ended up with 98.9% accuracy.

Results Comparison

In this research, we used the classification accuracy as the metrics. Classification accuracy is defined as the percentage of correct prediction for a certain classifier. For each method, we first trained the classifier using the training data and then used the validation set to gain the proper parameter which minimizes the training error. Then, we calculated the classification accuracy by running each classifier on the test data and measuring the percentage of correct predictions. We repeated this training and testing procedure three times for each classifier and took the average of these three results as the final classification accuracy.

Our results were as follows:

- 98.5% accuracy for Linear Discriminant Analysis
- 97% accuracy for Random Forest Classifier
- 98.9% accuracy for Support Vector Classifier

We are satisfied with the results of all three models. Although all three of our methods produced accurate results, we are most satisfied with the Support Vector Classifier method because it resulted in the highest accuracy. Due to the optimal margin gap between separating hyperplanes, SVC is robust and performs better on predictions. The SVC accuracy was achieved through effective predictions of both occupancy and non-occupancy. Linear Discriminant Analysis was also very successful at accurate predictions of occupancy and non-occupancy. Although also a strong method, the Random Forest Classifier achieved fantastic accuracy predicting occupancy, but fell short predicting non-occupancy leading to a lower overall result (displayed above).

Conclusions

Our research project objective was to build a model to automatically predict whether a room is occupied or not, based on attributes such as room temperature, humidity and light. Based on our literature review of existing research and the results of our exploratory data analysis, we selected three classification models to fine-tune and test. All three models produced accurate

results, but we were most satisfied with the results of the Support Vector Classifier. This model has the potential to increase the efficiency of HVAC and lighting systems within office buildings and ensure that these systems are not running extraneously. This will lead to energy reductions throughout the industry and could have lasting mitigating effects on climate change.

Works Cited

1. 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. Timilehin, Labeodan & Zeiler, Wim & Boxem, Gert & Zhao, Yang. (2015). Occupancy Measurement In Commercial Office Buildings For Demand-Driven Control Applications- A Survey And Detection System Evaluation. Energy and Buildings. 93. 10.1016/j.enbuild.2015.02.028.

3. "Support-Vector Machine." *Wikipedia*, Wikimedia Foundation, 22 May 2019, en.wikipedia.org/wiki/Support-vector_machine.

4. "Linear Discriminant Analysis." *Wikipedia*, Wikimedia Foundation, 30 Apr. 2019, en.wikipedia.org/wiki/Linear_discriminant_analysis.

5. "Random Forest." *Wikipedia*, Wikimedia Foundation, 30 May 2019, en.wikipedia.org/wiki/Random_forest.