Brandon Tsui

A14630205

6/9/2019

Math 189

Final Project

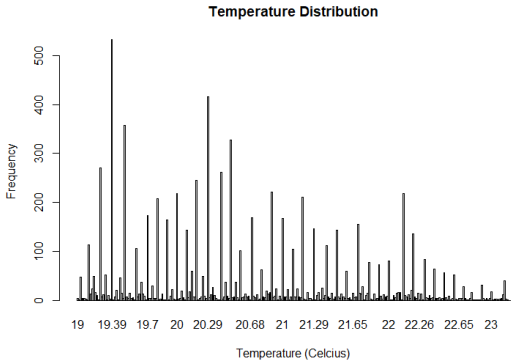
Introduction (1):

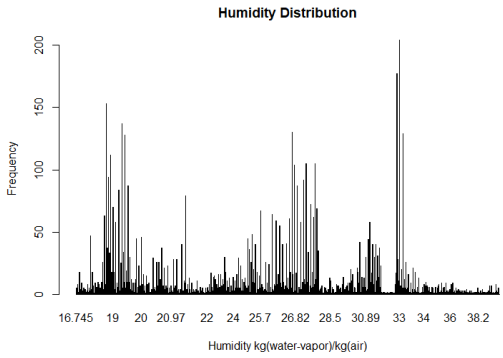
In this project, I will be discussing methods for predicting the occupancy of a room based on the date, temperature, humidity, light, CO_2, and humidity ratio. The purpose of doing so is to find a cheaper, more effective way of detecting room occupancy in order to save money by being able to better regulate HVAC system and lighting in non-residential buildings. While currently, cameras can be used to solve the same problem, using sensors for the aforementioned variables would be cheaper with the improvements in sensors and computing power, as well as better for privacy purposes. These methods are expected to work due to the nature of the features. Temperature, humidity, and CO_2 are all affected the human presence based on bodily functions such as breathing and natural body heat exfoliation while date and light are affected by human behavior. Time could be a very important variable as rooms are less likely to be occupied at night during off-work hours, and light levels are completely controlled by humans choosing to turn lights on or off which is typically dependent on occupancy to begin with. While not all variables will be used in the model trainings, I will be testing various methods such as Linear Discriminant Analysis, Classification Trees, and Random Forests in order to find a model which minimizes error and produces the most accurate predictions.

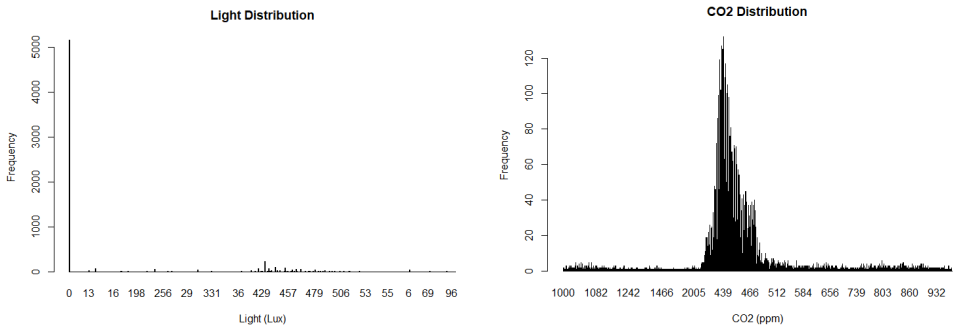
Exploratory Data Analysis (2):

In this section, I will be using basic statistical tools and visualization techniques in order to get an introductory look at each variable individually as well as see any correlations between variables to inform the actual process of feature and model selection.

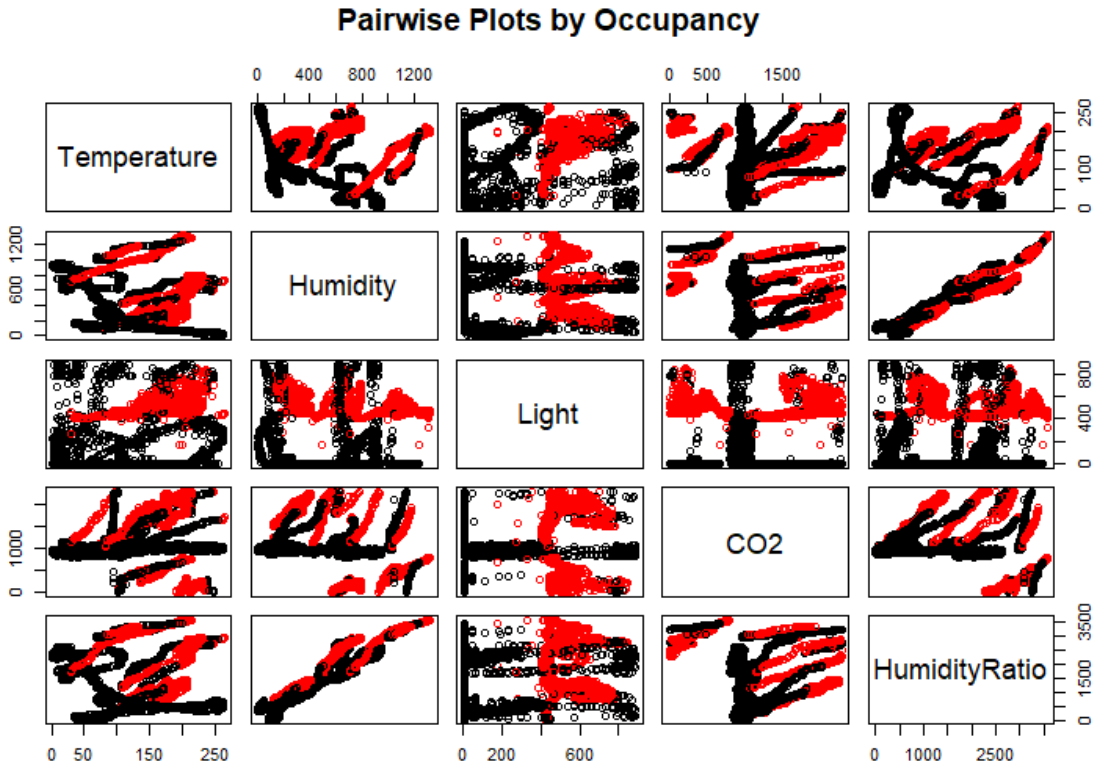
Histograms of each variable:



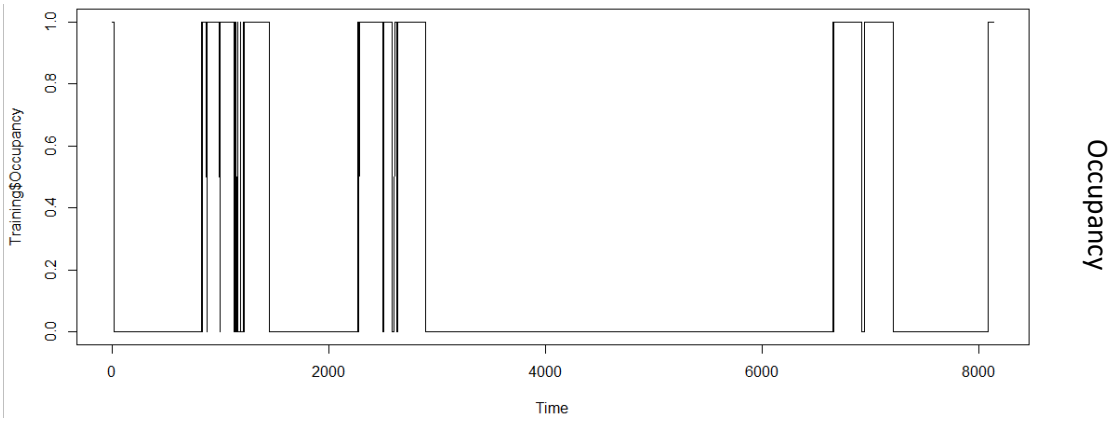


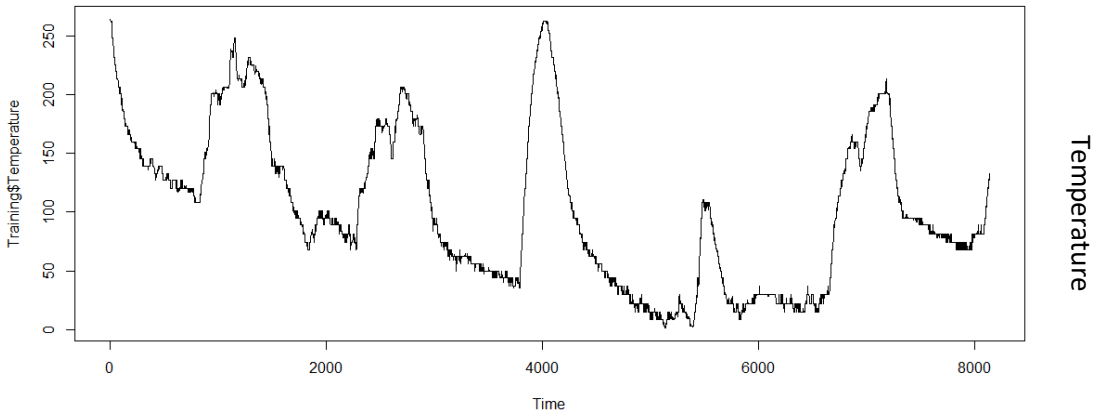


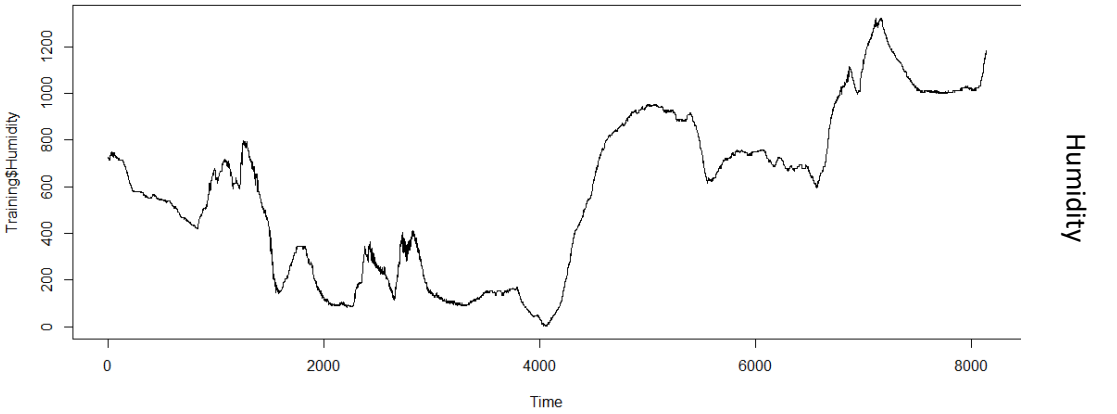
Looking at the individual distributions of each feature we can see a few trends. Although these histograms do not compare the variables to occupancy status, we can infer meaning based on some of the patterns shown. For example, looking at the temperature distribution, we see certain temperatures that occur at a much higher frequency than others. It is very possible that these constants correlate with no occupancy because when no one is in the building, there will be no temperature variation. The humidity distribution follows the same logic and inference although not as distinct as temperature because it is not as highly regulated by humans like temperature is. The light distribution is highly concentrated around 0. Looking closer at the data, we see that every single row in which Light is 0, Occupancy is also 0. This is nearly a hundred percent correlation except for the fact that there 1254 instances in which light is not 0 but Occupancy is one. This is expected due people leaving the lights on even when no one is left. Regardless, we can see that light is an extremely important variable in determining occupancy.

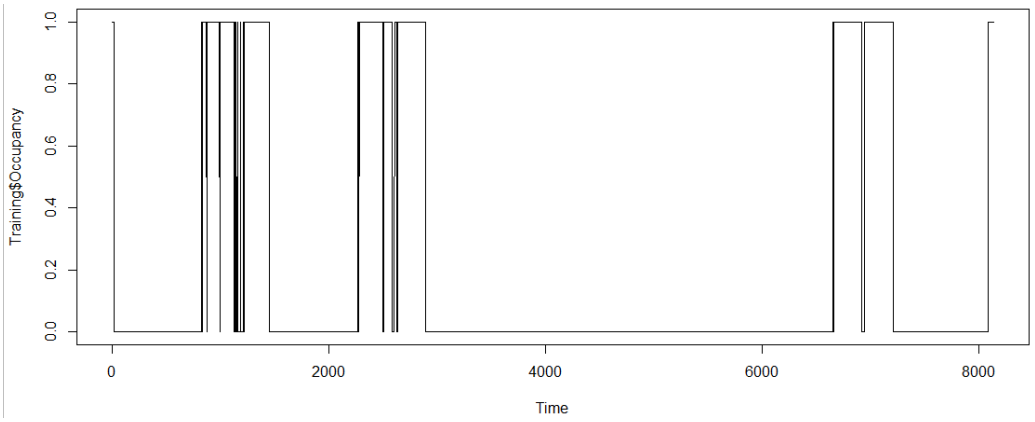


The pairwise plots above are very important and informative for this analysis. I compared each variable with one another and plotted each data point by occupancy with 0 in black and 1 in red. This allows us to see not only correlation between variables but also identify patterns relating to occupancy. Immediately, the most striking trend is the strong correlation between humidity and humidity ratio which makes sense because humidity ratio is based off humidity and is dependent on it. This information is irrelevant. What is more striking is the relationship between light and the other variables. There are distinct areas in which there is occupancy seen in each plot with little overlap meaning that for a certain light range and temperature range for instance, there is high likelihood of occupancy. These plots also confirm that when light is 0, occupancy is always 0.

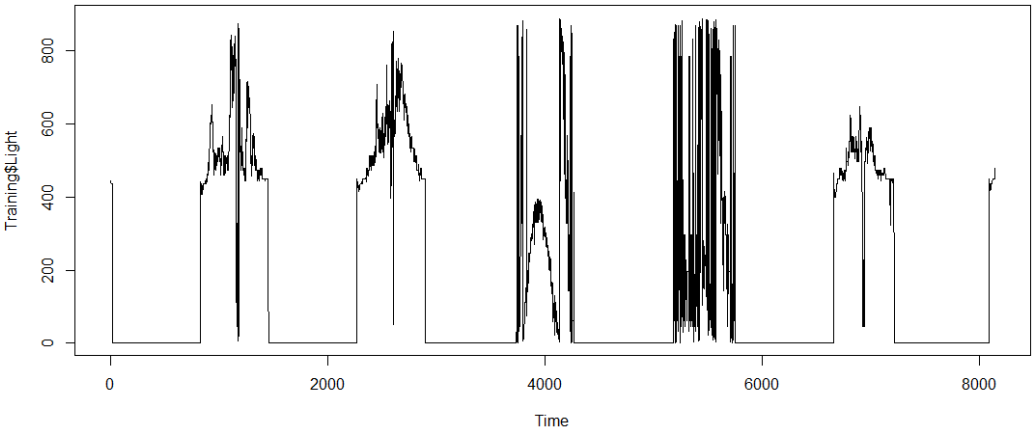




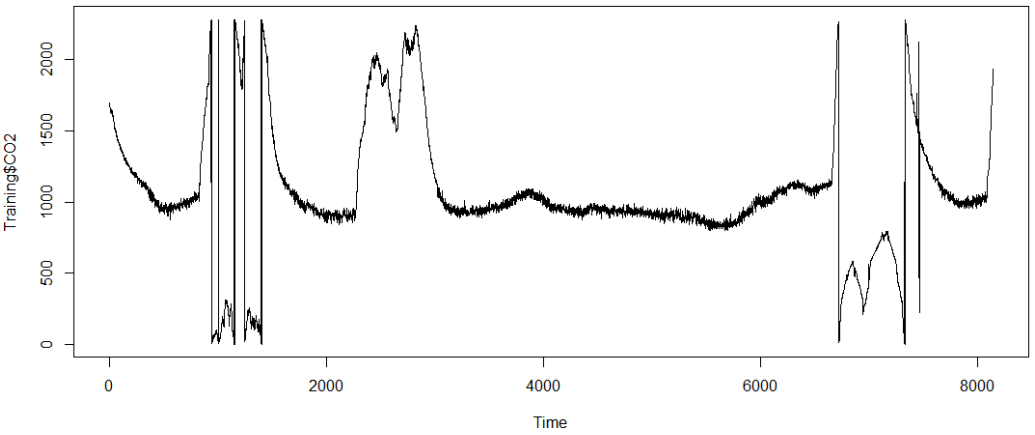




Occupancy



Light



CO2

One important variable we haven't looked at yet is time. Here we have each variable (Except HumidityRatio because it is so closely linked to humidity) tracked versus time (Every minute from 2015-02-04 17:51:00 to 2015-02-10 09:33:00). We compare it to occupancy over time to identify trends. Looking at temperature, we see distinct spikes when the room is occupied although there are similar spikes even when the room is not occupied. This could be a result of preset temperatures for the building. Similarly, we see spikes in humidity accompanying occupancy although over the humidity trended upwards which may have to do with the natural weather at the time. We see the same trend with light levels with very strange spikes even with no occupancy at the time. CO_2 seems to have correlation with occupancy as there are drastic

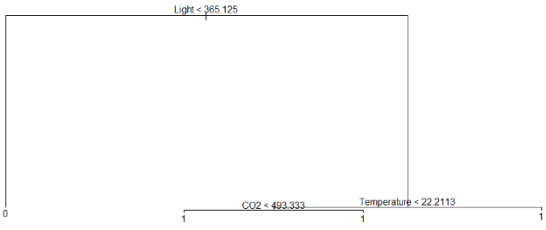
differences from what seems like a generally consistent trendline when unoccupied. However, since the CO_2 levels go up and down so drastically, it may be hard to use CO_2in our methods.

Overall through this EDA we see trends between each of our variables and occupancy, solidifying that we will find a method that yields good results. One big idea that will be explored further in the methods is the importance of light as a highly explanatory variable. For methods such as LDA, we will look heavily at light going forward.

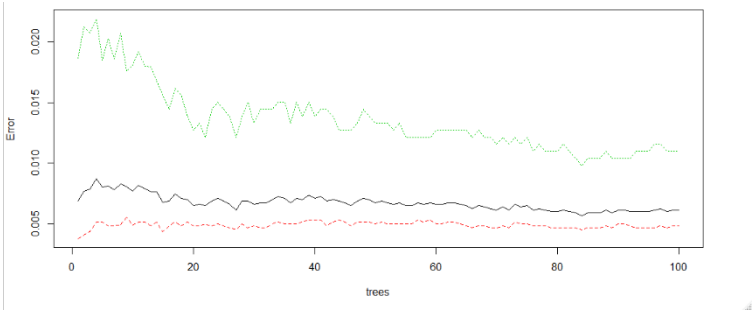
Methods (3):

LDA: The first method I tried was Linear Discriminant Analysis on all the variables except for date because the data was not numeric. I selected prior size based on the relative sample size of the training data. This was the most logical choice because 78.77% of the training data was categorized as not occupied(0) and there is no reason to believe that prior size should be equal. Common sense tells us that humans occupy buildings less than 50% of the time based on average work hours so this is consistent with our selection to make prior size relative to the training set. After selecting priors, I proceeded normally with the LDA, calculating sample mean vectors group by occupancy as well as covariance matrices. From the pooled sample covariance matrix, I found alpha and beta for each subset and used these to create the respective linear discriminant functions and applied it to each row of the Test data, classify occupancy based on the highest result from each function. The results yielded accurate predictions 98.5% of the time which is already excellent.

Classification Tree: The second method I tried was using a classification tree on all the data except for date because there were too many levels when using date as part of the training data. Looking at the initial tree, it is clear that the tree classifies occupancy solely on the light level with 0 if light is below 365.125 and 1 if above. The leaves on the right side of the tree are all 1. Based on this extremely simple classification model we get an accurate prediction 99.3% of the time which is extremely high considering the method.



Random Forest with Bagging: Last, I tried random forest with bagging. Again, date was left out for the above reason. I tried various different variations of the test, mostly changing values for mtry in order to get the highest accuracy. With mtry equal to the number of features – 1 (4), the accuracy on the Test dataset was 97%. With mtry equal to the square root of the number of columns (2), the accuracy on the Test dataset was slightly higher at 97.45%. I also adjusted the number of trees, but no changes resulted in any significant improvement in error.



Improving Methods (4):

Based on the results from the previous section, I decided to attempt to improve the LDA method and the simple classification tree since these methods resulted in higher initial accuracy rates on the Test data, and I already tested different variations of the random forest method with no significant improvements.

Improving LDA:

Since we know that changing priors will not do anything to affect classification, we will check if decreasing the number of features yields better results. From the EDA, we know that light is an important feature so we will test combinations of the other features with light. After testing nearly every combination of variables with light included, while most did not result in significant results, one combination did. While some of these options could be better because of lower computing power (i.e. they would need less sensor to accomplish a relatively high prediction rate), it was a combination of Temperature, Humidity, Light, and CO2 that yielded a better result. Here is just a sample of some of the LDA's I ran:

Temperature + Humidity + Light + CO2 + HumidityRatio = 98.5%

Temperature + Light = 98.15%

Humidity + Light = 97.1%

CO2 + Light = 97.3%

HumidityRatio + Light = 96.8%

Temperature + Humidity + Light = 98%

Temperature + Humidity + Light + CO2 = 98.6%

Temperature + Humidity + Light + HumidityRatio = 98.1%

Improving Classification Tree:

Unfortunately, after testing many iterations of trees by training it on different sets of variables, there were no improvements in accuracy. In fact, every variation simply led to the exact same base tree with only two leaves based on the value of light. In theory, if we could make a more detailed tree using more variables, that differentiate occupancy when light is above 365, we could yield better results. However, this might result in overfitting of the training set.

This is most likely the reason the classification tree made in R only goes so far because the program works against overfitting.

Conclusion (5):

While I was able to find a slight improvement in the LDA method, using a simple classification tree based on light still proved to be extremely effective with an accurate prediction 99.3% of the time. The problem with this model, however, is that more misclassifications are Type I errors, meaning that the model predicts occupancy when occupancy is 0. Like mentioned above, this model could be improved. If we do a closer analysis of the data in which light is above the threshold, we may be able to get more accurate results. One adjustment I would make is including time as a variable. In the dataset's current format, time cannot be utilized, however, if we separate it and simplify the data down to certain time thresholds per day such as just the hour/time of day, we might get better results. These methods would account for the strange spikes we saw in the light vs time graph which make up weekends and times at night with no occupants in which light levels were high, likely as a result of automatic building lighting systems. Overall, after all the analysis done, it turns out that one of the best solutions was extremely simple. The classification tree outperforms its competitors because as we saw earlier, light is highly correlated with occupancy. Other methods were affected by all the other variables which did not have as significant an impact as light did, so the model based solely on it proved to be simpler and more efficient. One further application that this work could be used for is adjusting the models so that it not only predicts occupancy or not, but also the number of people in the building. While obviously this cannot be done on the given data, if we collect more information on the number of people when a room is occupied, we could use methods such as regression trees in order to get accurate predictions of numeric values as opposed to just 0 or 1. This could also be useful for energy saving purposes because the impact of one person versus many is huge in terms of how much light and HVAC needs to be adjusted.

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. Abade, Bruno, et al. “A Non-Intrusive Approach for Indoor Occupancy Detection in Smart Environments.” *Sensors*, vol. 18, no. 11, 2018, p. 3953., doi:10.3390/s18113953.