Spencer Palladino
Stat 4250
Final Project

**Introduction (1)**

The optimization of room utility is something that should theoretically become more efficient and easier to manage given new emerging technologies. Even systems like online room reservations should help to create a space with a higher proportion of occupancy. If a room can stay filled a higher percentage of the time, we could consider the room to have a higher operational efficiency. Many room reservations systems, whether used to make a reservation the day of or in advance, have predetermined time slots with specific length that do not always go fulfilled given the variability of how long someone will actually need a room. A situation like this would then create a lower operational efficiency. In order to analyze situations in which we would like to predict whether or not a room is occupied, we will be using data collected by Luis Candanedo and Veronique Feldheim that they used in their own analysis for Occupancy using the following factors: Temperature (in Celsius), Humidity(%), Light (in Lux), $CO_2$(in ppm), and Humidity Ratio(in kgwater/kgair).

**General Visualization (2)**

In this project I will attempt to use several classification methods to create my own effective model for predicting whether or not a room is occupied. To prepare for analyzing data, I believe it is important to visualize the data to see general correlations between variables. In **Figure 1**, we have a scatterplot matrix of our factors and have separated the observations by whether or not the room is occupied or not (purple is occupied, blue is not). There are a few interesting observations here that pertain to our analysis. First, Humidity and Humidity Ratio have a clear positive linear correlation. To further test this, I made a correlation plot in **Figure 2** to better visualize the relationships between all variables. Beyond that, Light and $CO_2$ in their respective pair plots appear nearly completely separated for when the room is occupied and unoccupied. This is apparent from the fact that in the plots, there is little to no overlap in the area under the

blue and purple curves. Before we even do any further analysis, I hypothesize that Light and $CO_2$ will be valuable predictors for room occupancy.

*Figure 1*

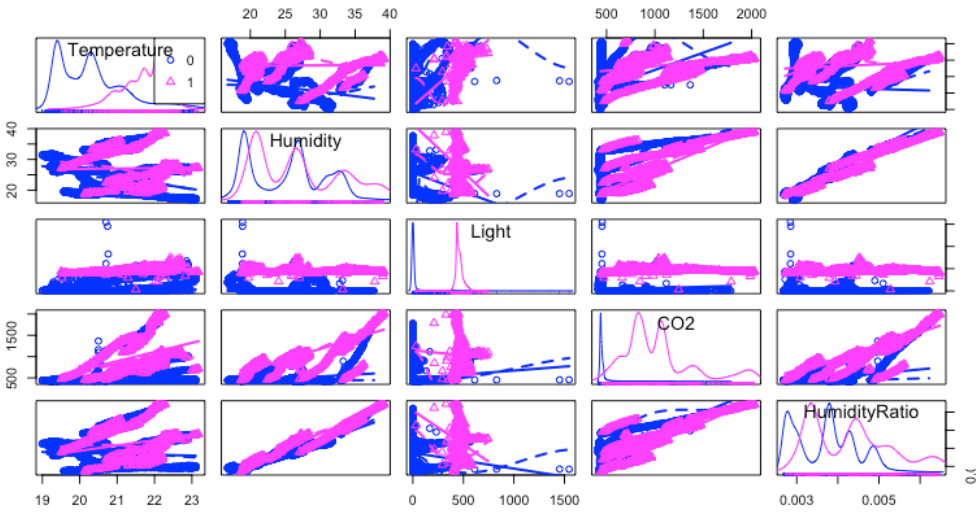
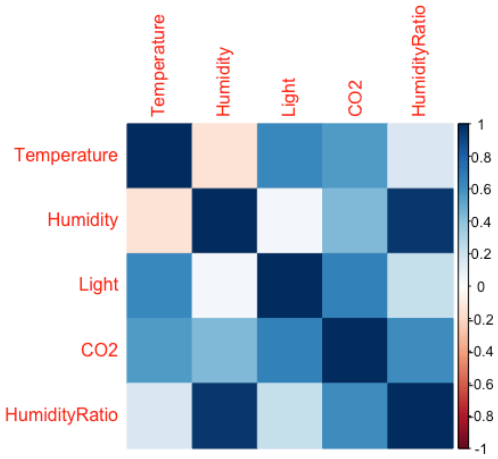



*Figure 2*




**Data Analysis/Visualization Methodologies (3)**

*Logistic Regression (3A)*

Given that occupancy is measured in the data with 1 being equal to occupied and 0 for not being occupied, our first methodology will be a logistic regression model because in this case we have a qualitative binary response variable. Our full model will include Temperature, Humidity, Light, $CO_2$, and Humidity Ratio as factors. Using our training set, our model is:

*Occupancy = 8.695 – 0.904(Temp)+.310(Humidty)+0.0206(Light)+0.0064($CO_2$)-2259(HumidityRatio)*

Two of our explanatory variables, Light and $CO_2$ had significantly low p-values, however none of our other factors did. In order to combat this, we can try to search for new models. In multivariate logistic regression, Akaike Information Criterion (AIC) is a great estimator for the overall quality of a model by expressing the relative amount of information lost. Thus, we can try to find models with lower AIC's (a lower AIC indicates a better model) than our full model. In

the full model, the AIC = 968.12. To test for better models, I used both forward and backward stepwise selection. In both scenarios, the same model was produced that **did not** include the factor Humidity. The new model calculated is:

*Occupancy = 18.73 – 1.373(Temp)+ 0.0261(Light)+0.0063($CO_2$)-259(HumidityRatio)*

This model is very similar to the previous with just a slightly lower AIC of 966.7. Each factor has a fascinatingly close coefficient to that of the full model which can most likely be explain by how close of a relationship HumidityRatio and Humidity had. Because of their closeness, the model does not significantly change despite our new model having a lower AIC. Without Humidity, however, we see that the p-values of our new model are very significant for all remaining factors except HumidityRatio and even HumidityRatio has a p-value of .069, which would be significant if I wanted to use alpha = .1.

It should also be noted that based on our coefficients from our model, it can be determined that when the temperature and humidity ratio go up, the probability that the room is occupied go down. When light and $CO_2$ levels are higher, the probability that the room is occupied increases. To visually look at our model, we can look at **Figure 3.** In the top left, our Residual/ Fitted plot has two distinct lines. Because we predict a probability for a variable taking values 0 or 1. If Occupancy is 0, then we always predict more because the value is 0, and residuals have to be negative and if occupied the value is 1, then we underestimated, and residuals will be positive. I do however have a slight concern for the just how curved each line is. To me that indicates that there might be a higher order term with our model. Our normal QQ plot in the top right of **Figure 3** may also indicate normality issues because of the S shape of the plot. Lastly, I want to mention the bottom right does indicate an outlier as point 3832 is out Cook's Distance. Because there are so many points, I don't think it should significantly affect our model and in turn affect our Test data conclusions.

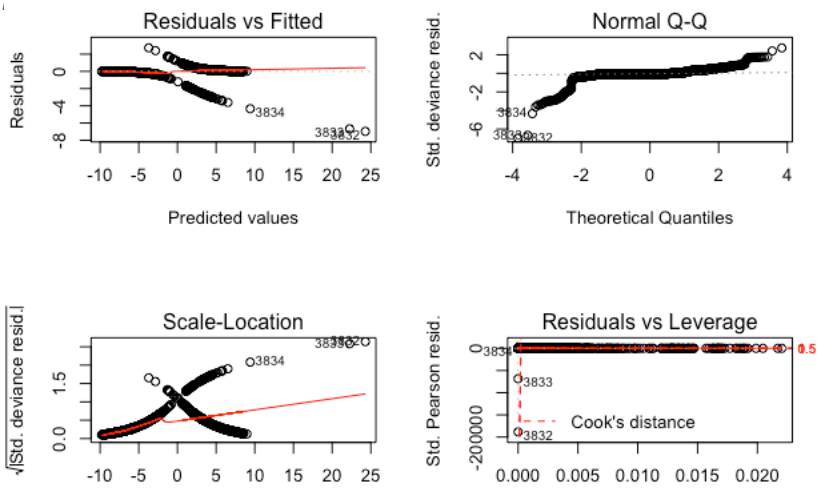
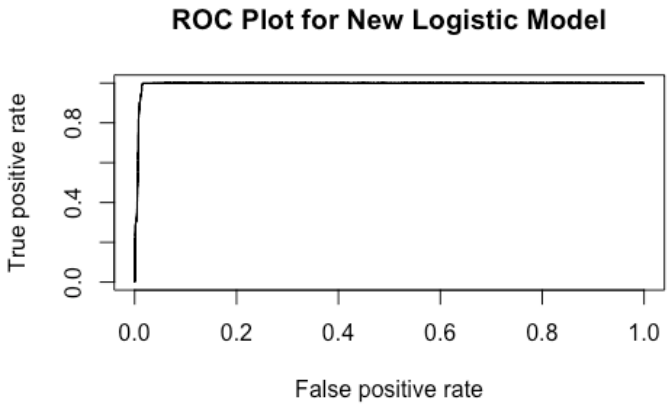



*Figure 3*

Along with AIC and visually looking at the model plots, I believe that AUC (Area Under Curve) is clear and crucial characteristic for predicting how well our model is going to be able to predict whether the room is occupied. Thus, I made a ROC plot (**Figure 4**) to help visualize the AUC of our new logistic regression model. I calculated an AUC= 0.9945762 using our Test data. The closer the AUC is to 1, the better our model is for classifying our response variable (Occupancy in this case). Because our AUC is so close to 1, our model should be considered a good predictor of Occupancy.

*Figure 4*



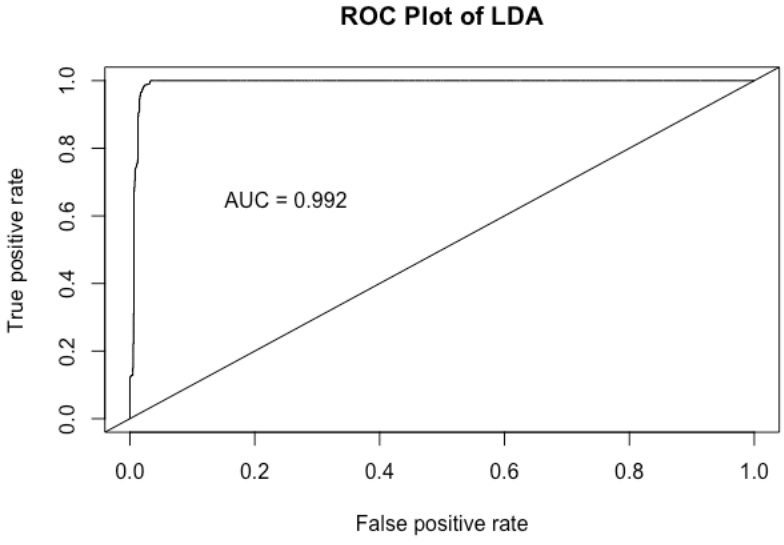

*Linear Discriminant Analysis (3B)*

Another methodology I felt was necessary for finding good model to predict Occupancy was Linear Discriminant Analysis (LDA). LDA is specifically used when there is a known qualitative response variable and given that our problem is based around using data to classify rooms as either occupied or not occupied, LDA is an appropriate method of analysis. Using the training set, we could determine the linear discriminant function and make a ROC plot (**Figure 5)** and get the AUC with test set. The linear discriminant function:

*D(x) = -0.409(Temperature) + 0.0123(Light) + 0.00233(CO2)- 106.09(HumidityRatio)*

The AUC with both the full model used previously and our new model we made through stepwise procedures had an AUC of 0.992.

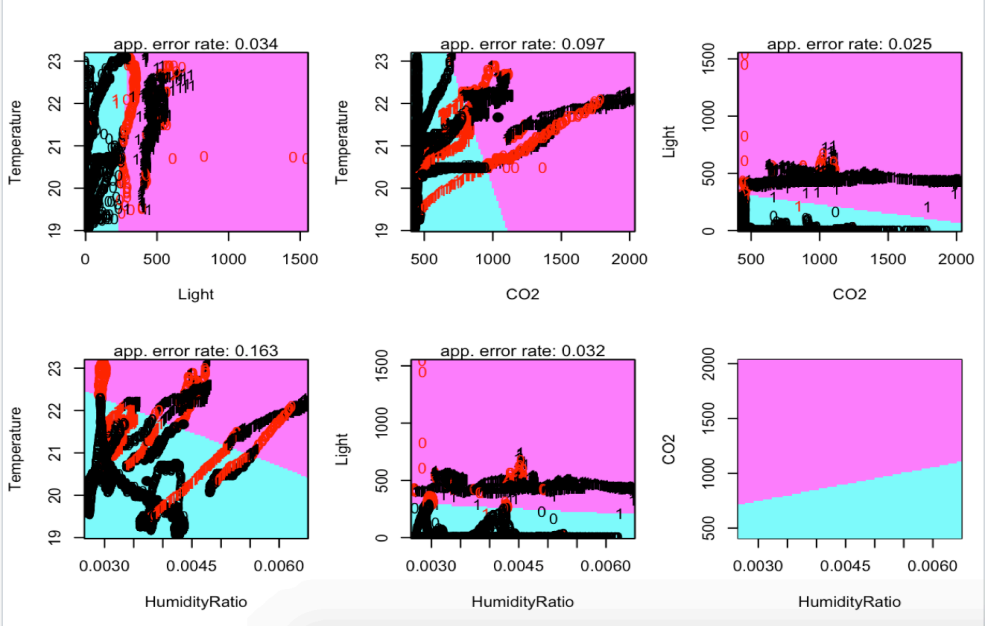
*Figure 5*




Along with the ROC plot, we could use our training LDA to fit a scatterplot matrix for

our test set for how each pair of explanatory variables predicts the Occupancy of a room (**Figure 6**). The pink areas are where the room is occupied, and the teal areas are where the room is not occupied. As shown above the individual scatter plots, each pair of variables has a relatively low error rate. Specifically, the pairs of Temp/Light and Light/HumidityRatio appear to be able to predict occupancy of the room with significance. (Unfortunately, R was glitching out on the last plot)

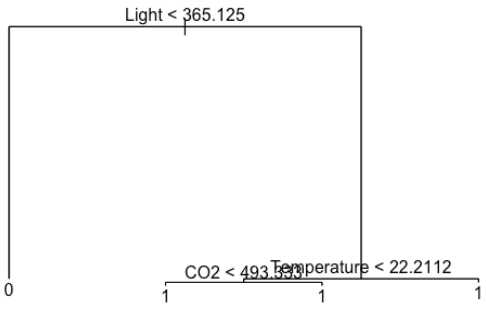
***Figure 6***




*Tree Based Methods (3C)*

Tree methods are a great way to model the data in such a way that it can be easily visualized and interpreted from a tree like plot. As an example, we can use our training data to form a single regression overall tree plot (**Figure 7**). To interpret the plot, we would start from the top and say that if light is less than 365.125, than the room is unoccupied. If it is greater, then we go down the ladder and look to see if temperature is greater than 22.2 degrees.

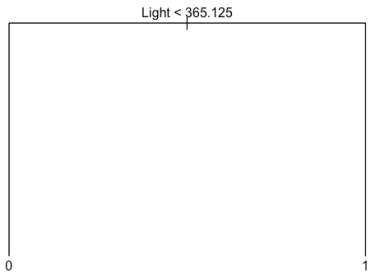
***Figure 7***



If it is, then the room is most likely occupied. If not, we can make a decision based on $C0_2$ levels. This particular figure is a little confusing because of my own programs techinical difficulties but the 1 on the bottom left under the CO2 branch should be a 0. In our training set, this only resulted in a missclassifcation situation 1.216% of the time which is very low and a solid RMSE = .07. We can then use the validation set to decide our hyper-parameters, meaning we will use the data to prune our tree and see if we can be better off with less branches. Using the elbow method of **Figure 8**, we should be able to minimize the mean squared error of the tree and keep the tree as simple as possible at size 2. **Figure 9** depicts what a tree of size 2 would look like and in this case it will be interpreted as if light is greater than 365.125 than the room is occupied and not if light is smaller. This smaller tree was then applied to the test data and resulted in a very low RMSE= .066 and had a 0.6% misclasification rate. It also benefits still from being incredibly easy to interpret. If you look all the way back at **Figure 1**, the information we learned here matches up with what we observed from the Light pair plot because light specifically had almost no overlap of occupied and unoccupied rooms.

***Figure 8***


***Figure 9***




A slightly more in depth tree method than single regression trees is random forests. This method starts with the entire model and recursively divides it into smaller regions very similar to a single regression tree in **Figure 7**. However, the random forest method constructs a bunch of these decision trees and aggregates the result from all the trees into

final classification parameters. This also helps decrease overfitting the data which will result in a lower RMSE. Similar to simple regression tree method, we start with the training set and use the Validation set to select our hyper-parameters to optimize our results. Our full model has a RMSE=0.066 with 500 trees using the training set. Tuning our parameter, with the Validation Set we can find that we hit a minimum RMSE 201 trees (the validation set had a generally higher RMSE). This can be visualized in **Figure 10**. Then applying this parameter to the test set, we ended up with a RMSE = 0.004 using 201 trees which is excellent. In order to compare this model with my logistic regression model and LDA model, I was able to compute the AUC for my random forest model. **Figure 11** depicts the ROC plot for my Random Forest model and it has an AUC= .99544.

*Figure 10*

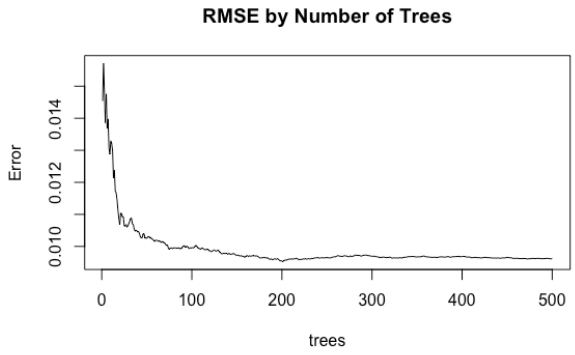
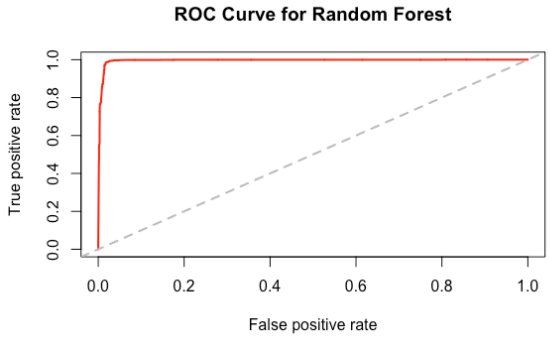



*Figure 11*




**Conclusion (4)**

In this project, I used methodologies such as linear regression, LDA, and tree based methods which all included data anaylses and visualizations to understand how the factors Temperature Humidity, Light, $CO_2$, and Humidity Ratio can be used to predict the Occupancy of a room. Because of the strong relationship between Humidity and Humidity Ratio, I found my models to be stronger when they avoided the redundancy of having both variables. I determined Humidity Ratio was the better variable to keep through step wise logistic regression and each models AIC characteristic. Light appeared to be the strongest and best predictor of Occupancy. This was able to be noticed from my general visualization of the Light pair plot in **Figure 1**, the low p-values in my logistic regression plot, and my single tree plot which was optimized having only light as a

predictor. We can also compare each of my models using the AUC of each ROC plot which I found for each Test set. My linear regression model had an AUC = 0.9945762, my LDA model had one= 0.992, and my random forest model had an AUC = .99544. Each model has well over .99 which means that they should all be excellent in predicting whether or not the room is occupied. Technically, my logistic regression model had a slight edge but it was not by a very large model, so I would not want to outright conclude that it is the best model. If I was to do further research into this subject, I may be able to find small improvements using a QDA model or by testing higher degree terms in my logistic regression model.

References:

Candanedo, Luis M., and Véronique Feldheim. “Accurate Occupancy Detection of an Office Room from Light, Temperature, Humidity and CO 2 Measurements Using Statistical Learning Models.” *Energy and Buildings*, vol. 112, 2016, pp. 28–39., doi:10.1016/j.enbuild.2015.11.071.

I used many powerpoint slides and code provided by Dr. Ke in class and on eLC.