Rhonald Reside

DS670 - Assignment 11

Midterm Draft

**Discussion and Conclusion**

I find it easiest to visualize my data after it was loaded into Tableau. I created line graphs, heat maps and averages to help me build a story from the data. I discovered a few things after this step using Tableau. First, I found that there was a data gap in from June 8, 2014 to August 1, 2014. I could see this once I created a line graph with temperature, dew point and humidity. It was a very interesting visual find. I suspect that since this this data gap occurred throughout all the variable, there was most likely the possibility of data storage issue or they all ran out of battery at the same time.

The next discovery I found was the warmest day and the coldest day in the data set. I found that August 2, 2014 was the warmest day measured at 80.60° F with a humidity of 37 and a dew point of 11. The coldest day occurred March 11, 2014 with a humidity of 84 and a dew point of -4. I decided to create a heat map that represented the high temperature all the days. I also did the same for the cold and created heat map of the coldest temperature day

Next I loaded all the data in R. By doing so, I found one of the variables to be incomplete. The visibility variable only ranged from February 2014 to June 2014. I therefore am deciding not to use this variable. At this point it would not matter because visibility variable does not really affect relative humidity.

I feel needed to apply complex statistic or machine learning to the project. R programming language made this step easier. I decided to first split my data into two data sets. I created a training data set and a test data set. This is an important step because you can run your statistics on your training data set. I hoped at this point that the statistics I have a very low error rate so that I can beat my competitive article. Just as a reminder, my competitive article has an error rate lower then 5%. The other point I would like to mention is that no matter how low your error rate is on your training data set, you expect your testing data set error rate to be slightly higher or at least very close to your training data set error rate. If your error rate for your testing data set is lower, then your training data set is not a good representation of your total data set.

I plan to use logistic regression as part of my analysis. Even though we can see there is a relationship with temperature, dew point and humidity when graphing in Tableau, I would like to see a statistical number that show this correlation. So when in R programming, I use the library ISLR for my statistical analysis. I plan to use temperature as my Y-Variable, also known as my dependent variable. I label my variables as dew point and humidity; we would also know this as independent variables.

From the ISLR library, I use the glm() function for the logistic regression model. I first use a binary response to tell me if my temperature is a hot day or a cold day. I use 65° as the decision variable. If the temperature is above 65°, then my algorithm will lable the temperature hot. If it is lower then 65°, then the temperature is labeled cold. By labeling my temperature inputs, the glm() function will create a confusion matrix. A confusion matrix is a table used to describe the classification performance of my model of my training data set. In this case, it shows us the accuracy of our classifier of 65°.

The generalized linear model us useful for predicting an outcome from a binary response from a data set. It is sometimes called discriminant function analysis because its assumptions are less restrictive. Our formula for the data set is

logistic\_model = glm ( Yresults ~ tmpm\_Fahrenheit + Dewptm + hum + Pressurem + wdir ,   
 data = data1 ,   
 family = binomial   
 )

In the formula above, we are calling our function a logistic\_model. Inside our glm() function, we have our independent variable and dependent variables, our data set, and binary response. Again, Yresults is our temperature. Our independent variable is the rest of our variables , , ,, and (i.e. temperature, dew point, humidity, pressure and wind direction.) Data is our training data set. Family equals binomial is our binary response to 65°.

We fit our data to model the glm() function. As an outout, we see the variable in the usual way and a receive a binomial error distribution. The estimators of the coefficient show us weight influences in a positive manner while displacement is negative. The goal is to calculate the predicted probability of temperature for the specified values of our predictors, i.e. dew point, humidity, pressure and wind direction.

I predict that there will be a strong correlation with humidity and dew point and weather. My hypothesis is that our output will agree with our competitive article. I think the glm() function will have lower error rate then our competitive article therefore making our model a better choice for relative humidity correlation. Our competitive article believes that you can lower your error rate when fitting the data without the help of computers. I believe otherwise, that it is very necessary to use computers to predict. Using computers will reduce your error.