Airport Delay Preliminary Analysis

Graeme Smith

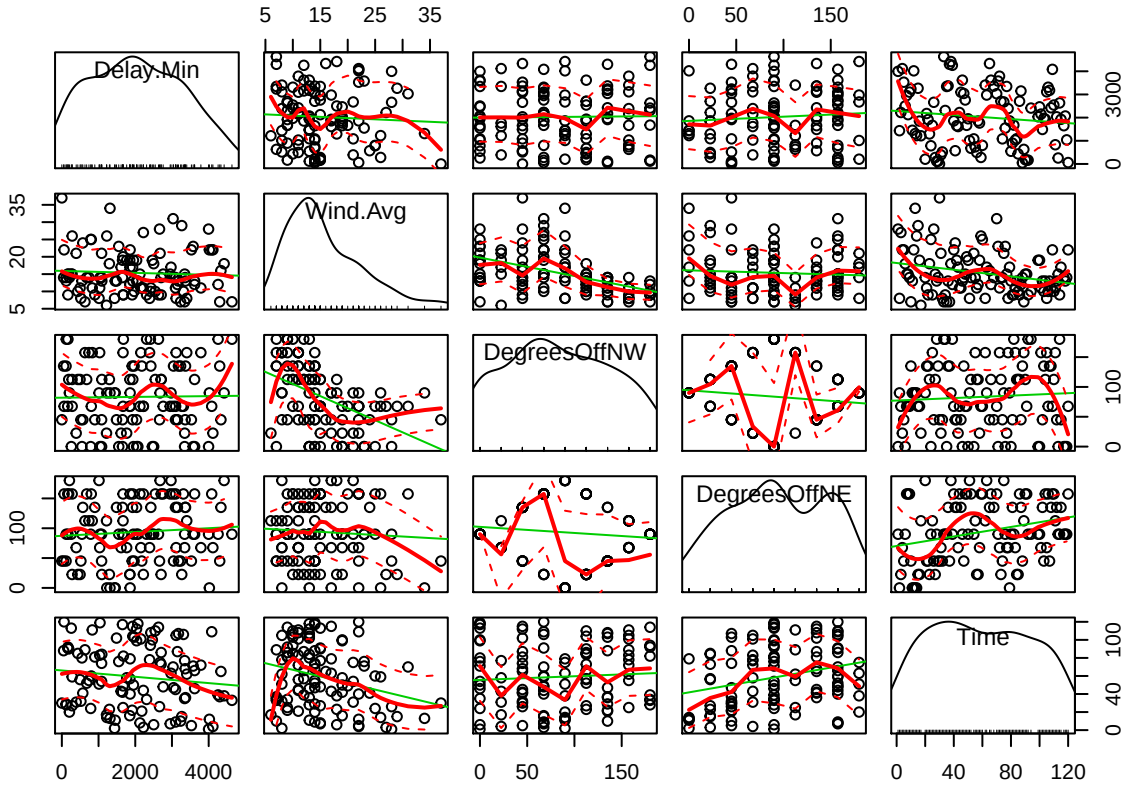
Tuesday, April 21, 2015

This is a very quick preliminary analysis on the data provided by Gerry to see if it shows any predictive value. The data consists of 120 daily data points taken from Newark airport over the spring of 2012. The raw data consisted of **Delay (in minutes)**, **Wind Speed Range** and **Wind Direction**.

Before doing my own analysis I tried running the original model in the Genie application. I admit I wasn't familiar with it, so may have made some mistakes. I tried using the *Validate* function to do *10 fold Cross-Validation* and *Leave One Out Cross Validation*. These both simulate out-of-sample performance by dividing the data into multiple training and testing sets and testing on the sets not used in training. The accuracy in the two cases was 26.7% and 32.5%, below random chance.

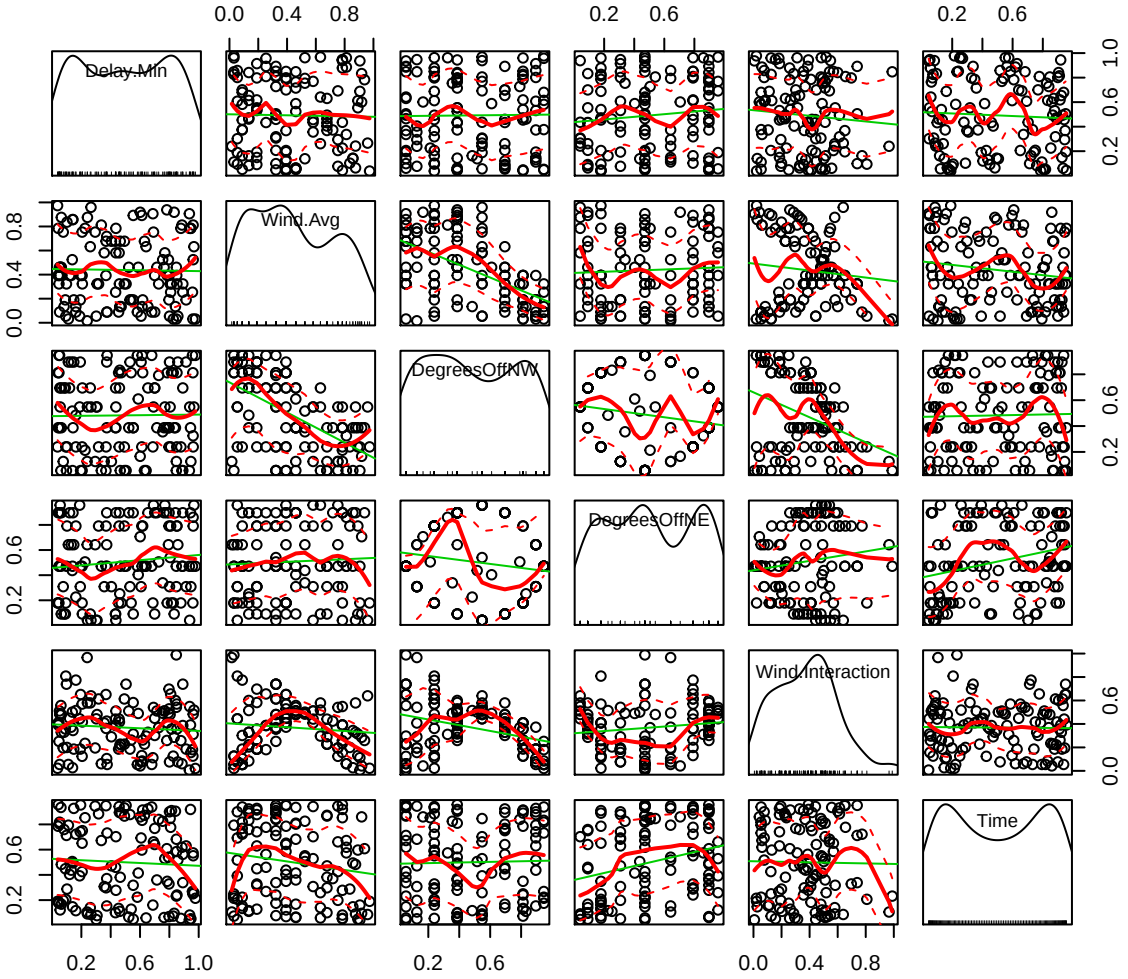
In order to initially do a visual inspection of the data I choose to look at the continuous variables rather than the categorized versions since this made visual features more obvious. I also separated the Wind Direction variable into two axis variables, **Degrees Off NW** and **Degrees Off NE**. I choose these axes because the wind strength seemed to pickup around the NW axis. I also added a **Time** variable since the time period chosen went from the end of winter through to the beginning of summer and it is quite possible that Winds or Delays may be worse during winter meaning possible correlations between these could the result of a third seasonal variable.

Chart 1. Variable Interactions



The Diagonal in the *Chart 1* shows the distributions of the data, the circles show the actual data points and the green and red lines are the best fit slope and smoothed best fit. The top line is the most important since it shows the interaction of each variable with the **Delay**. The first two things that stand out is the the **Wind (Avg)** is unevenly distributed and heavily skewed. Also there appears to be a strong correlation between **Wind (Avg)** and **Degrees Off NW**. So I did some preprocessing of the data to make the data approximately linearly distributed between 0 and 1. I also created a new variable that captured the interaction between the **Wind (Avg)** and **Degrees Off NW**. Low values would correspond to high NW Winds, medium values to low wind and high values to medium or high SE winds (although there were no high SE winds).

Chart 2. Normalised and Pre−processed Variable Interactions



Visually inspecting the first row of *Chart 2*, there appeared to be little interaction between any of the predictors and **Delay**. To verify this I also took the correlations between the variables and the delay and tested for significance.

Normally a P-Value under 0.05 or 0.01 would be used to test for significance. In this case all the P-Values round to 1.00, suggesting no correlation between the variables and **Delay**.

Next I tried running some fairly powerful predictive models, including *Random Forests*, *Support Vector Machines*, *C5.0* and *Linear Regression*. These generally were slightly more accurate than purely random, having Accuracy scores of between around 32% and 43%. In *Tables 2, 3* and *4* are the statistics for the

	Correlation	PValue
Wind.Avg	-0.02	1.00
DegreesOffNW	0.01	1.00
DegreesOffNE	0.11	1.00
Time	-0.05	1.00
Wind.Interaction	-0.08	1.00

Table 1: Correlation with Delay in Minutes

Random Forest Model. Although it was one of the better performing models with accuracy of 41%, the P-Value of 0.32 does not show statistical significance at the standard 0.01 or 0.05 levels.

	High_Delay	Low_Delay	Normal
High_Delay	6	13	8
Low_Delay	16	25	16
Normal	10	8	18

Table 2: Confusion Matrix for Random Forest Model

	x
Accuracy	0.41
Kappa	0.09
Accuracy Lower	0.32
Accuracy Upper	0.50
Accuracy Null	0.38
Accuracy PValue	0.32
Mcnemar PValue	0.36

Table 3: Overall Statistics for Random Forest Model

Lastly, the one thing I didn't get around to doing was creating my own Bayesian Network for doing cross-validation testing due to time constraints, and my belief that it wouldn't add anything more at this stage.

In conclusion I would say that with the current limited data it is not possible to say whether predicting flight delays is feasible or not. Going forward I would say priority should be given to showing feasibility and creating a realistic proof of concept model. This would require a much larger set of data indicators, and across a much longer time-frame, preferably spanning years to compensate for possible seasonal components.