PM Concentration Modelling - Week 4
Jonathan Levine, June 1st 2020

I re-tested the performance of the models on the **Whitehorse** dataset based on an adjusted-R^2 value.

$$\text{adj-R2} = 1 - \frac{(1 - R^2)(N - 1)}{N - K - 1}$$

Where N is the sample size and K is the number of independent variables (features).

From the correlation matrix (Figure 1) Gary suggested, we can see how linearly independent the features in the dataset are. From Figure 1, we can see that each has very low correlation.

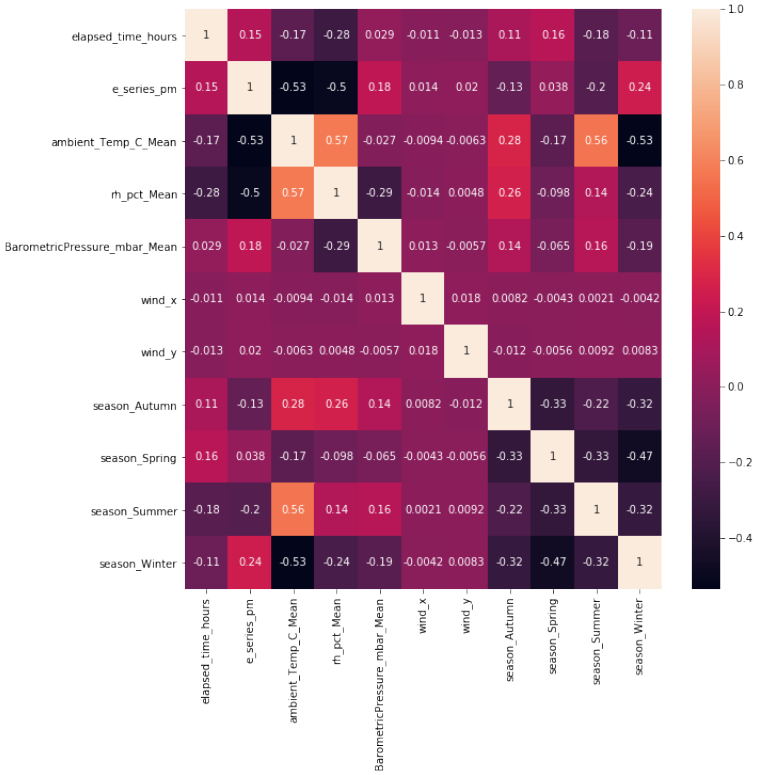

Figure 1: Correlation Matrix

And so this week I added recursive feature elimination techniques along with cross-validation to the models to select which features could be eliminated. With this, I saw

some improvement on the model accuracy. I also implemented a Random Forest Regression technique, which is showing a lot of promise.

For the first model I employed was an **ElasticNet** regression and split the whitehorse data into two separate datasets. One dataset covering the first half of the year, and the second dataset covering the second half of the year. I trained the model on the first half, and the results were: MAE: 1.5167 ugm3, R2: 0.7768 and adj-R2: 0.6053. This model was then applied on the second half of the dataset, and the results were: MAE: 1.8712 ugm3, R2: 0.8279 and adj-R2: 0.6849. The plots for the second half of the dataset are shown below,

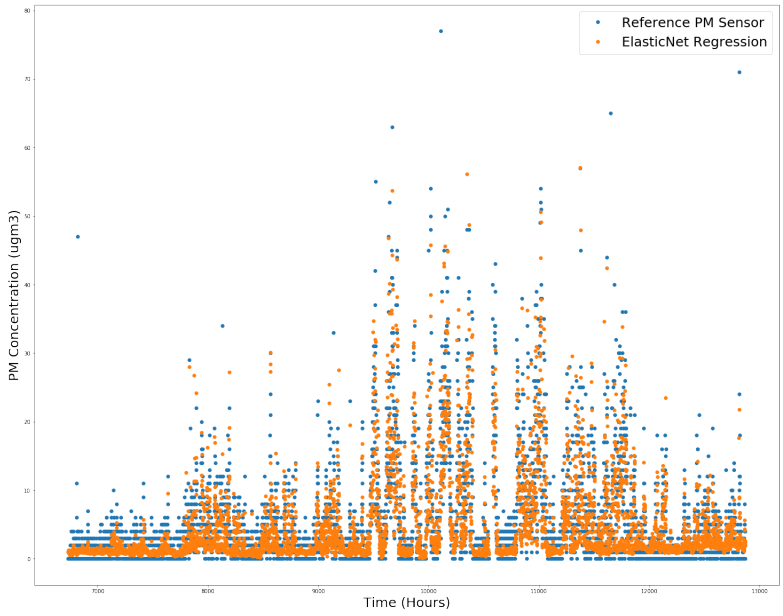

Figure 2: Elastic Net Regression Time Series Plot

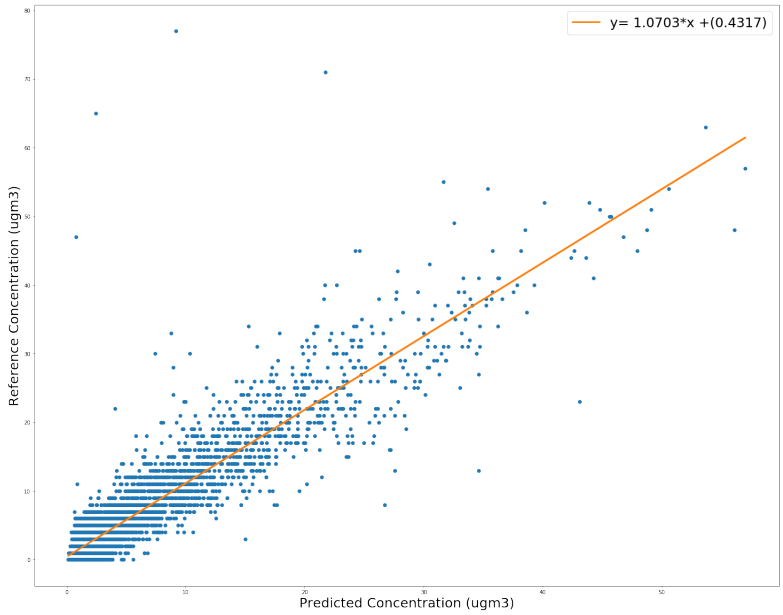

Figure 3: Reference vs. Predicted PM Concentration (ElasticNet)

The second model I employed was a **Decision Tree** regression. For this model I split the whitehorse data into two parts, the first 90% of the year I used for training, and the last 10% of the year I used for testing. On the training set the results were: MAE: 1.2611 ugm3, R2:0.8942, adj-R2: 0.7995, and on the testing set, the results were: MAE: 1.9604 ugm3, R2: 0.7047, and adj-R2: 0.4924. The plots for the testing set are shown below.

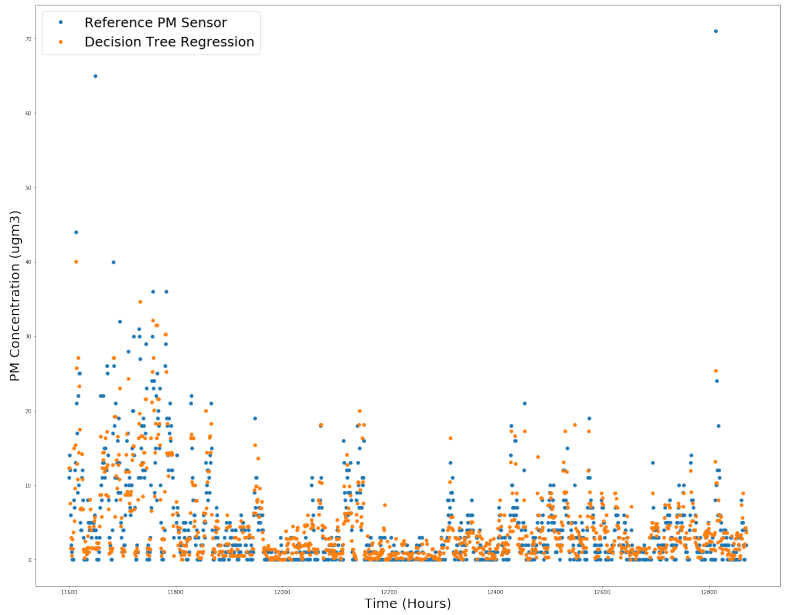

Figure 4: Decision Tree Regression Time Series Plot

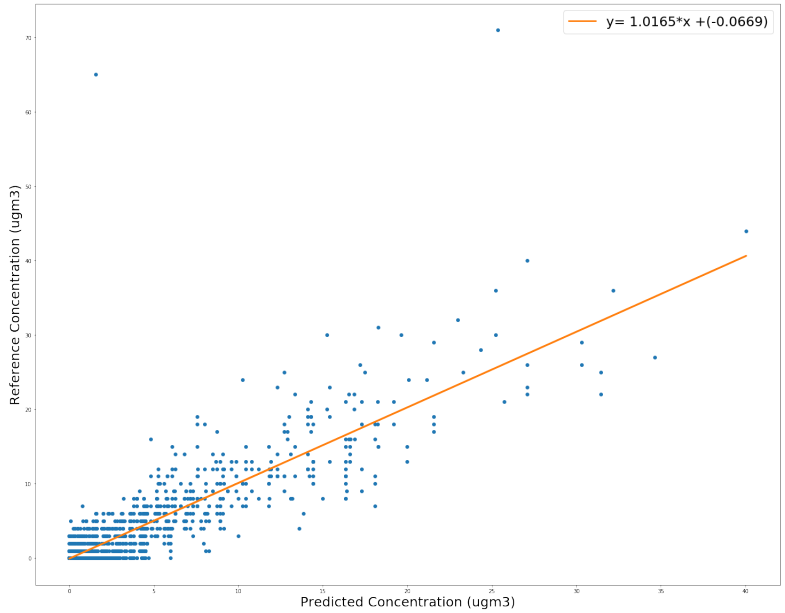

Figure 5: Reference vs. Predicted PM Concentration (Decision Tree Regression)

Note that the low adj-R2 value can be attributed to a small sample size. We can see that as the sample very large, adj-R2 approaches 1.

We can see this if I reverse the training and testing sets so that the model is only trained

on the first 10% of the Whitehorse data and tested on the last 90%. The R2 and adj-R2 values become 0.7730 and 0.5972 respectively.

An improvement on the Decision Tree regression is an ensemble approach, **Random Forest** regression. This modelling approach is quite powerful. When trained on the first 10% of the Whitehorse dataset, the model yields an MAE: 0.9991, R2: 0.9445, and adj-R2: 0.8912. Then when tested on the last 90%, the model yielded an MAE: 1.7494, R2: 0.7916, adj-R2: 0.6263, with plots,

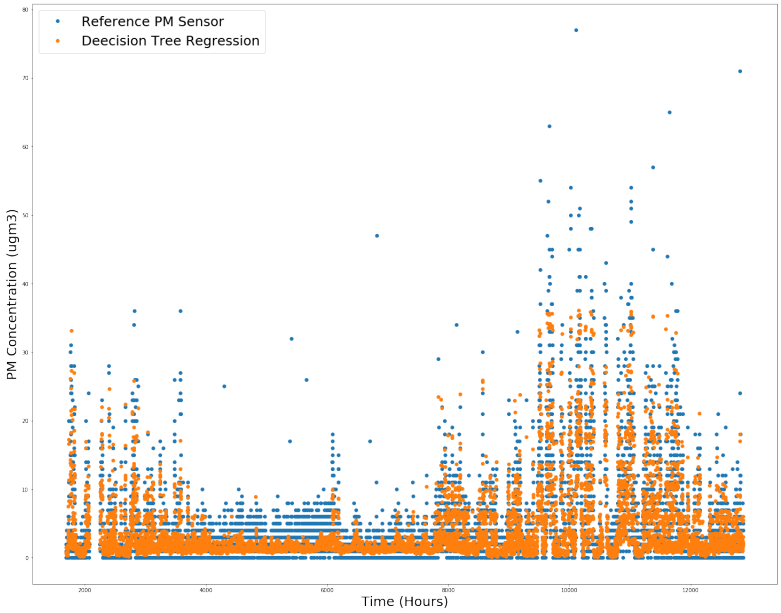

Figure 6: Random Forest Regression Time Series Plot

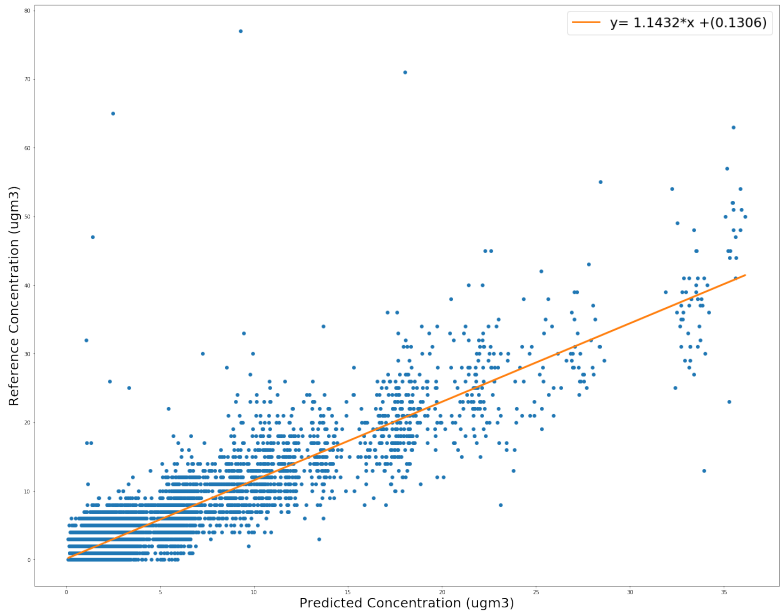

Figure 7: Reference vs. Predicted PM Concentration (Random Forest Regression)

With more data and further tuning of parameters, I believe that the Random Forest regression should perform even better. In the (Lim et Al) paper Joyce sent me, they apply more sophisticated boosting and ensemble techniques and they are able to reach cross-validated R2 scores of 0.8.

Going forward I believe that the analysis on the Whitehorse data is done. I will apply these modelling techniques to the Airpointer and Egg datasets. I plan to have most of the analysis done by Wednesday.