NOAA GPU Homework

Station: 352170

Name: DZHAMBEJTY

Problem 1: Temperature Prediction:

Model: Random Forest

- 1000 trees
- min_leaf_size = 10
- Train MSE: 15.82
- Val MSE: 27.64
- Test MSE: 32.94

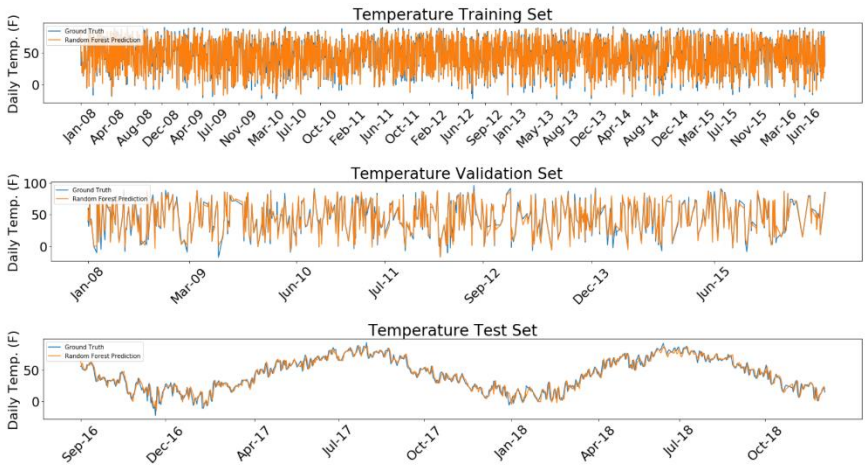

Figure 1: Temperature Prediction

Analysis:

Because the random forests place such large importance on the previous day's temperature: temp_lag1 (fig. 2), we are not gaining much information from the other variables and they are not useful features. Looking in closer at the test data it appears that the model is learning a delay model (appendix fig. 8) where it essentially predicts the last day's temperature. When that is the case, we are not truly learning anything, except that it is better to guess yesterday's temperature. In terms of improvement, there are a few things that come to mind. First, it appears that there is an underlying yearly seasonality in the data. Therefore, we may be able to give it some intuition about where we are in the year based on moving average data ranging from 1 week to 6 months. Additionally, we can look at dimensionality reduction where we remove some of the less important features in order to reduce the number of splits that don't provide any meaningful information.

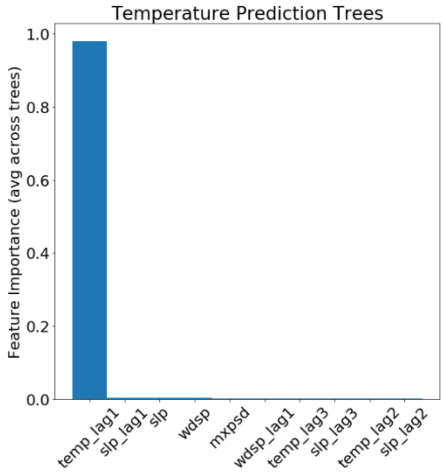

Figure 2: Model Feature Importance

Problem 2A: Wind speed Prediction

Model: Random Forest

- 1000 trees
- min_leaf_size = 10
- Base:
  - Train MSE: 5.47
  - Val MSE: 10.14
  - Test MSE: 7.79
- Attempted Improvement:
  - Train MSE: 5.17
  - Val MSE: 8.86
  - Test MSE: 7.91

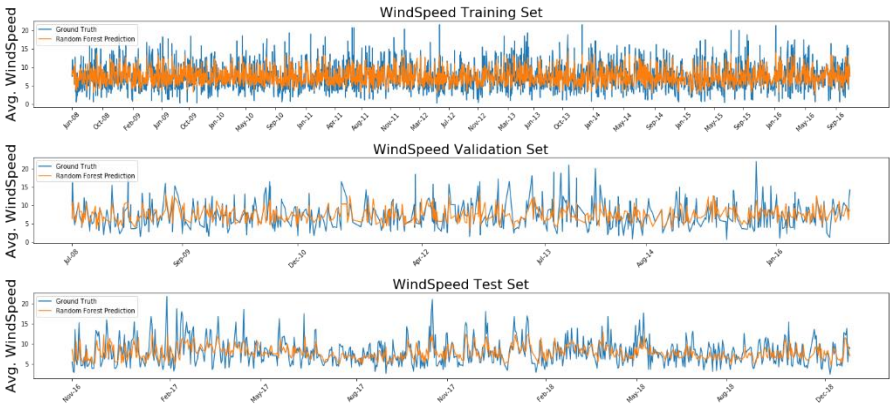

Figure 3: WindSpeed Improved Model Prediction

Part B

The Wind speed data has much higher variance than the temperature dataset, and if there is underlying seasonality or periodicity in the data, appears to be masked underneath the noise. As a whole we can see that there is a better distribution of feature importance (appendix fig. 9). By adding moving averages of 2,3,5,10,30,90, and 180 days as feature data the performance of the model was not improved but it yielded more consistent performance between train, validation, and test sets. Intuitively, the shorter-term moving averages acted as a form of low pass filtering, providing the model a smoothed input signal, and the longer term averages in theory can provide the model a method to infer underlying seasonality given the ratios of relative moving averages to each other. An FFT of the

sample data shows that there is no prominent seasonality or periodicity (appendix fig. 11). However, it can be seen that the longer-term moving averages rank among the top 10 features in terms of importance (appendix fig. 10), indicating that they provide the model with some useful context.

Problem 3: Time Series Decomposition of Temperature Time Series

Model: Random Forest

Base:

- 100 trees
- min_leaf_size = 10
- Train MSE: 16.19
- Val MSE: 30.92
- Test MSE: 28.00

Decomposed:

- 100 trees
- min_leaf_size = 10
- Train MSE: 17.27
- Val MSE: 26.52
- Test MSE: 26.43

In time series analysis, one of the underlying assumptions for the analysis is that the time series is stationary. This is also a critical assumption for the ML models since we are assuming that the distribution is consistent over time. For a given series to be considered stationary, the mean and standard deviation of the series must remain consistent over time. One of the tools used to evaluate stationarity is the Augmented Dickey Fuller test which provides a probability that a given time series is non-stationary. The lower the probability, the high the likelihood that the time series is stationary. The temperature data from my weather station proved to be stationary with a probability of 8e-4 (global warming be damned!... I kid I kid). Traditionally, the time series signal can be decomposed into linear trend, seasonality, and residual components. The trend is subtracted to center the series mean at zero while the seasonal components are removed after Fourier Analysis. This leaves the residual component which can broadly be interpreted as the noise in the data. Regression methods such as ARIMA and Exponential Smoothing are then used to predict this univariate residual component.

Figure 4: Part 3 - Temperature Baseline Model Prediction

During my research for this assignment I had come across a statement that random forests struggle with trends[1]. Although the time series is stationary, could shorter term trends lead to the inaccuracies in the prediction? Therefore, I decided to investigate if random forest models could be trained to regress the trend, seasonal, and residual components individually, and then be combined to improve overall performance. In addition to the decomposition, models were provided with moving average feature data (same averages as part 2B) and were trained on the back 80% of the dataset. Training on the last 80% was motivated by the fact that more extreme temperatures ranges were experienced in the last 20% of the dataset, which led to value clipped by the random forest models. This is because regression trees cannot extrapolate beyond the ranges of the data seen during training. Therefore, testing was conducted on the first 20% of the time series. For this implementation, all models used 100 trees and minimum leaf sizes of 10. It should be noted that these two adjustments during training time improved the baseline model performance compared the model used in problem 1.

The decomposed model provided some interesting results and insight into the capabilities and expressiveness of the model. In contrast to the statements about the pitfalls of the random forest models, they were able to nearly perfectly regress the seasonal and trend components.

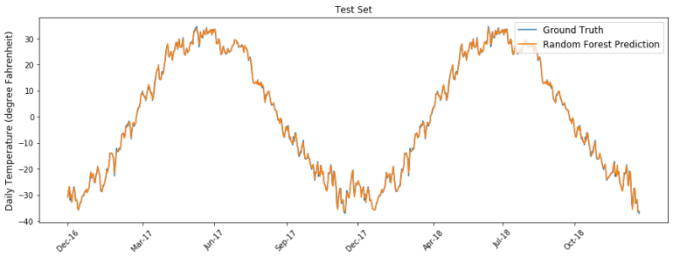

Figure 5: Seasonal Component Model Prediction

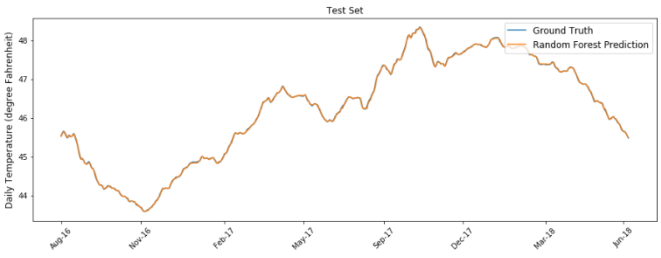

Figure 6: Trend Component Model Prediction

Ultimately, the decomposed model had incrementally better performance than the baseline model (table 1). This leads me to conclude that random forest models implicitly calculate the trend and seasonal component of a time series and the error in prediction is broadly caused by the residual component.

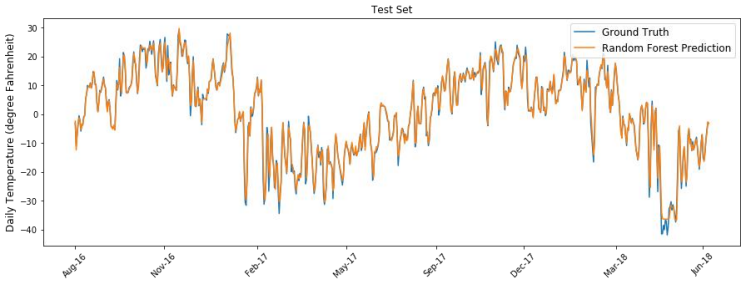

Figure 7: Residual Component Model Prediction

Table 1: Model Performance Comparison

	Baseline Model	Trend Comp.	Seasonal Comp.	Residual Comp.	Aggregate Model
Train MSE	16.19	0.0009	0.572	16.71	17.27
Validation MSE	30.92	0.0016	0.716	25.42	26.52
Test MSE	28.00	0.0012	0.592	26.46	26.43

REFERENCES

1. https://www.statworx.com/at/blog/time-series-forecasting-with-random-forest/

APPENDIX

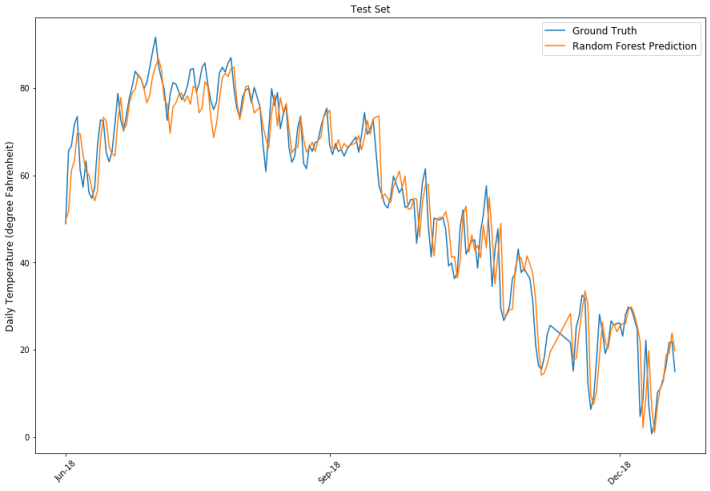

Figure 8: Problem 1 Temperature Model Prediction

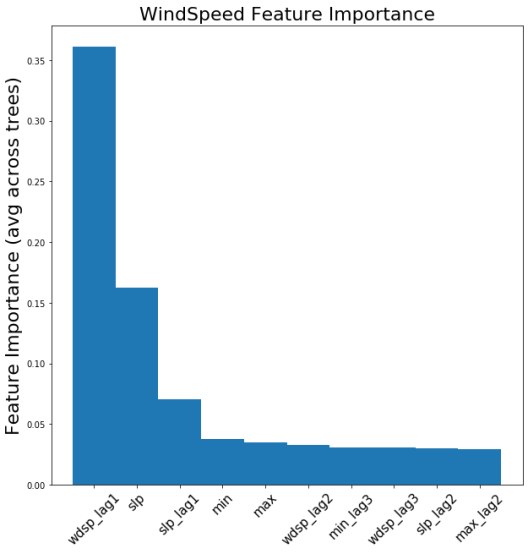

Figure 9: WindSpeed Base Model Feature Importance

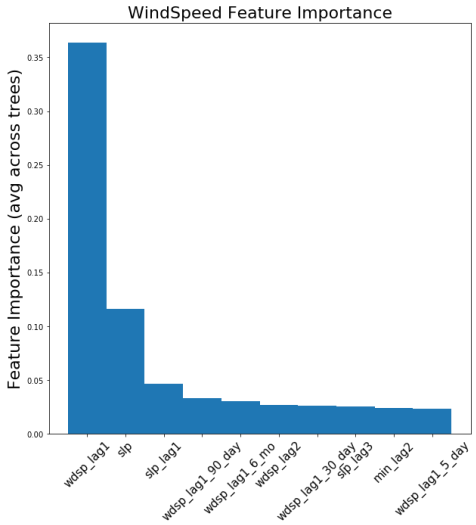

Figure 10: Windspeed Improved Model Feature Importance

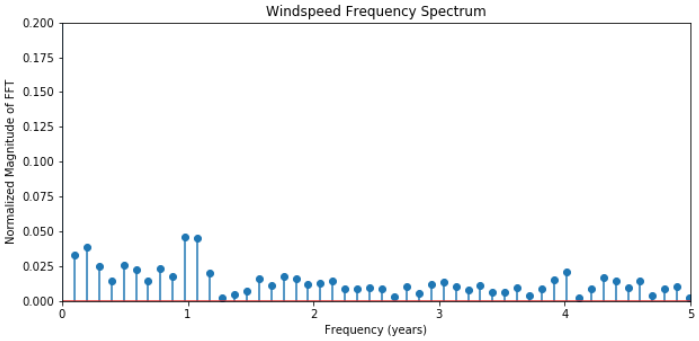

Figure 11: Windspeed Frequency Spectrum