NOAA HW_ECE 228_SP 2020 Hangquan Zhao A53320603

Problem 1

	usaf	wban	name	country	state	call	lat	lon	elev	begin	end
5946	712220	99999	DEASE LAKE (AUT)	CA	None	CWKX	58.433	-130.033	+0802.0	19930829	20190326

Figure 1：Original position of station in Problem1

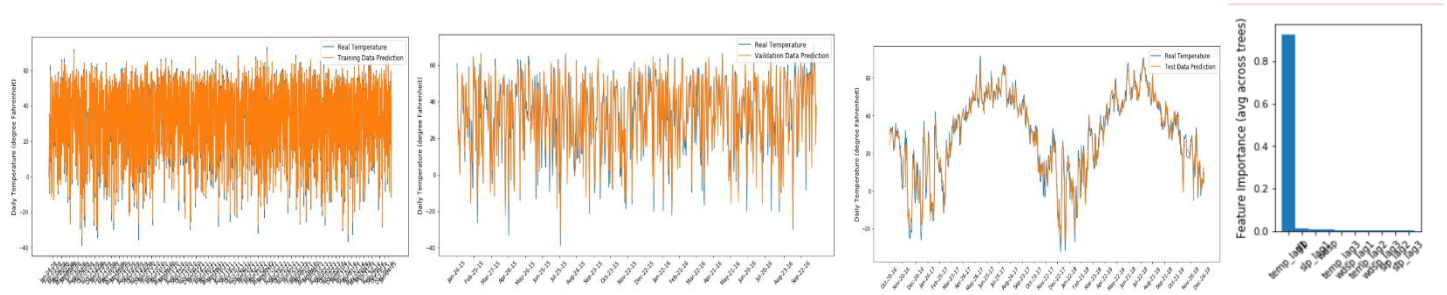


Figure 2：Results of original position

The training predictions, validation predictions, test predictions and feature importances of location in Figure 1 are shown in Figure 2 above. And we calculate the Mean Absolute Error of training set, validation set and test set are 1.4 degrees, 3.89 degrees and 3.92 degrees.

	usaf	wban	name	country	state	call	lat	lon	elev	begin	end
19261	085220	99999	FUNCHAL	PO	None	None	32.633	-16.9	+0056.0	19310101	20190401

Figure 3：New position of station (in the sea) in Problem1（Random Seed 5）

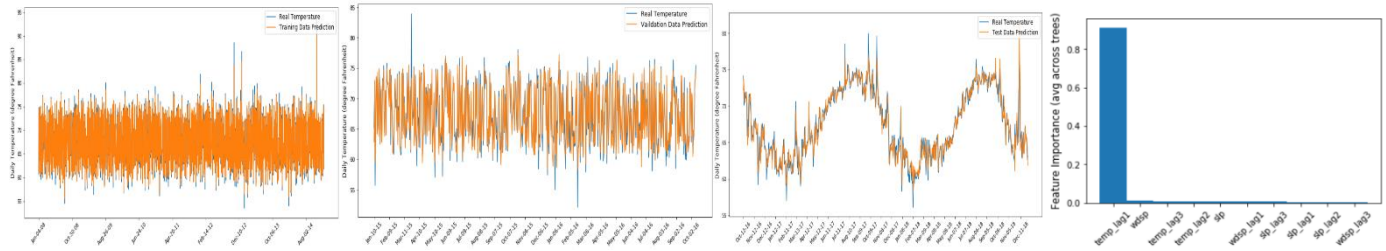


Figure 4：Results of new position

The training, validation, test predictions and feature importances of new location in Figure 3 are shown in Figure 4 above. And we calculate the Mean Absolute Error of training set, validation set and test set are 0.4 degrees, 1.04 degrees and 1.03 degrees.

Downsides and suggesting improvements of the model:

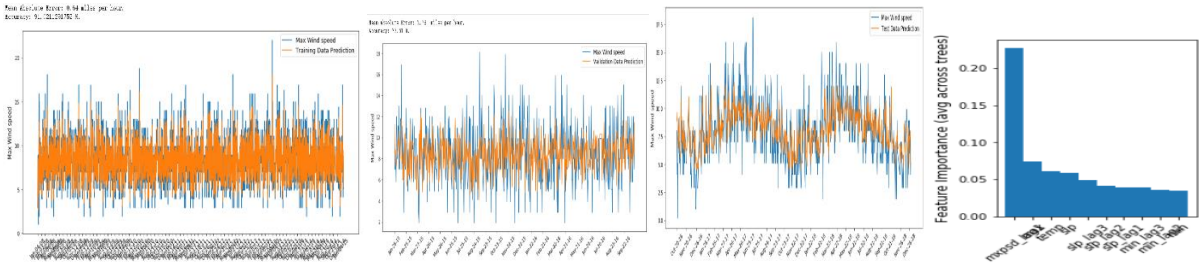
Downsides:

1. The main limitation of random forest this project is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions. In general, these algorithms are fast to train, but quite slow to create predictions once they are trained. A more accurate prediction requires more trees, which results in a slower model.
2. In some Data, the Mean Absolute Error is a little large and not good enough.

Improvement:

3. we use more (high-quality) data and feature engineering to improve the Accuracy and decrease the Errors. For example, we can increase random tree numbers, max_depth, min_samples_split, min_samples_leaf, bootstrap to improve the prediction Accuracy, but it also increases the calculation timing. Therefore, it needs tradeoff.

Problem 2



Here we also use the new position of station (in the sea) to predict the **max wind speed**.

Figure 5: Wind results of new position

The training predictions, validation predictions, test predictions and feature importances of new location in Figure 3 are shown in Figure 5 above. And we calculate the Mean Absolute Error of training set, validation set and test set are 0.64 m/s, 1.79 m/s and 1.84 m/s.

B. Based on your suggestions in Problem 1, try to improve your predictions for random forest. Briefly state what you did. Why do you think it did or didn't improve the results?

Time Consuming:

Because we have known the most five important features for the Temp, so we can just use the most important five features to train the Temp data and predict the result.

We can see in the temperature part, the most five important features are temp_lag1 slp slp_lag1 wdsp temp_lag3. After Dropping other features, the time of training data is as follow, which improves a lot.

	Training	Training data Prediction	Testing data Prediction	Validation data Prediction
Original time	18.2 s	0.61 s	0.245 s	0.20 s
Error		1.4 degrees	3.92 degrees	3.89 degrees
Optimal time	5.84 s	0.4 s	0.156 s	0.20 s
The final Errors		1.61 degrees	4.48 degrees	4.32 degrees

Increase Accuracy:

We have the following arguments for the random tree regressor
n_estimators = number of trees in the forest
max_features = max number of features considered for splitting a node
max_depth = max number of levels in each decision tree
min_samples_split = min number of data points placed in a node before the node is split
min_samples_leaf = min number of data points allowed in a leaf node
bootstrap = method for sampling data points (with or without replacement)
And our task is to choose the best value for every parameter to get the lowest errors, so I use random search first and then grid search, After we training the training temperature, we get the Mean Absolute Error: 3.87 degrees.(original one is 3.89) Which improves 0.5%. The final Result is as follow.

	Optimal Result	Default
Validation Data	3.87 degrees	3.89 degrees.
Test Data	3.91 degrees.	3.92 degrees.

Problem 3 (*Implement an another ML model for timeseries prediction*)

Because we use Random Forest Model here which is based Decision Tree. So I implement Decision Tree Model which definitely will work here for timeseries prediction and see what happens next, we use the following function to achieve decision tree model training.

```
clf = tree.DecisionTreeRegressor()
rf = clf.fit(x_train, y_train)
```

and the result is as follow:

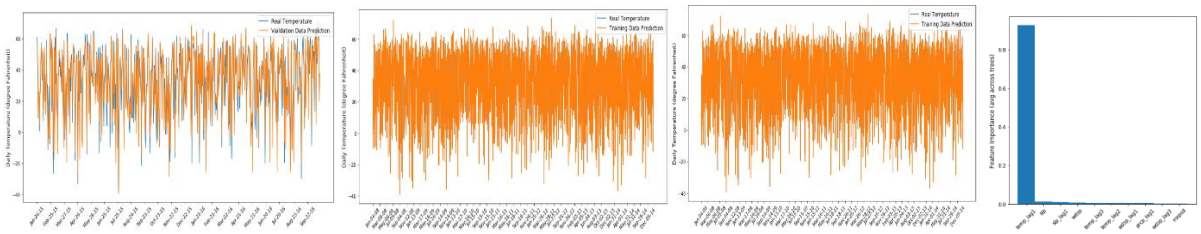


Figure 7: Temperature results of new position using DT model

In problem 1, the Mean Absolute Error of training set, validation set and test set are 1.4 degrees, 3.89 degrees and 3.92 degrees. But now the result of Decision Tree is 0 degrees, 5.29 degrees, 5.92 degrees. The results of validation data and test data is a little bit worse than Random Forest Model, but it's much faster and the model fitting time is only 0.036 s now. So that tells us in some simple situation, we can use simplified RF model which is DT model and the results won't be much worse. The new training predictions, validation predictions, test predictions and feature importances are shown in Figure 7 below.