Optimized Random Forest

1990s, All Stations

Optimization Results:

- Best GRID search hyperparameters are: {'max_depth': None, 'max_features': 50, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
- Best GRID search score is: 0.6319824753559693
- Best RANDOM search hyperparameters are: {'criterion': 'gini', 'max_depth': 70, 'max_features': 52, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 250}
- Best RANDOM search score is: 0.6352683461117197

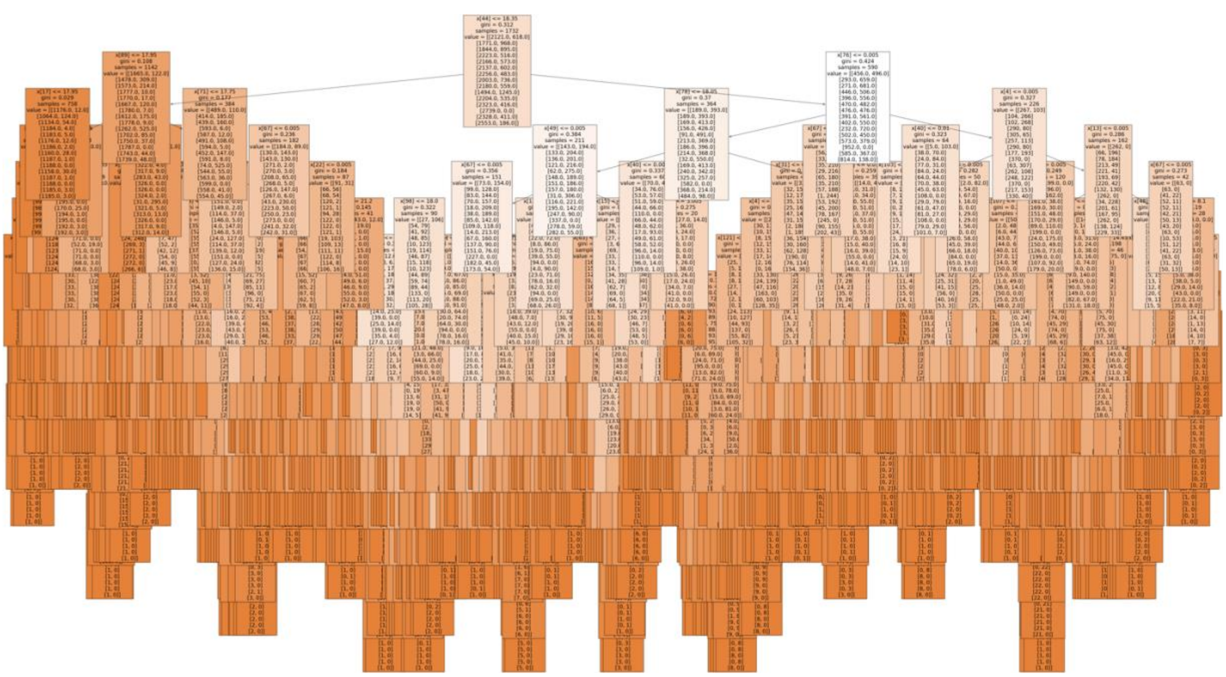
Paramters Used

- clf3 = RandomForestClassifier(n_estimators = 250, max_depth=70, max_features=52, min_samples_leaf=1, min_samples_split=2, criterion = 'gini')

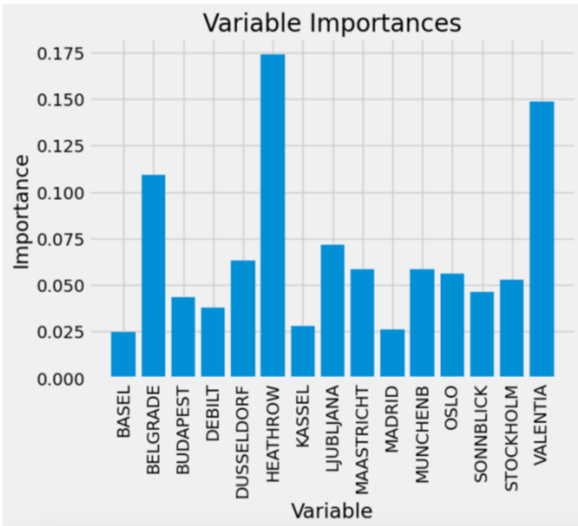
Pre-optimization Accuracy: 59.1%

Post-optimization Accuracy: 67.3%

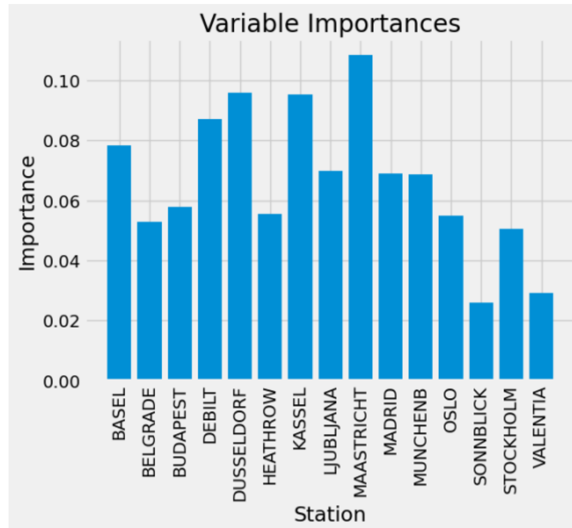
All Stations, Optimized:



Importances Optimized:



Importances Pre-Optimization:



Heathrow, All Years

Results:

- Best GRID search hyperparameters are: {'max_depth': 3, 'max_features': 15, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 10}
- Best GRID search score is: 1.0
- Best RANDOM search hyperparameters are: {'criterion': 'gini', 'max_depth': 9, 'max_features': 18, 'min_samples_leaf': 1, 'min_samples_split': 9, 'n_estimators': 24}
- Best RANDOM search score is: 1.0

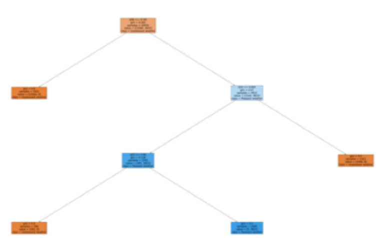
Parameters Run:

clf3 = RandomForestClassifier(n_estimators = 24, max_depth=9, max_features=18, min_samples_leaf=1, min_samples_split=2, criterion = 'gini')

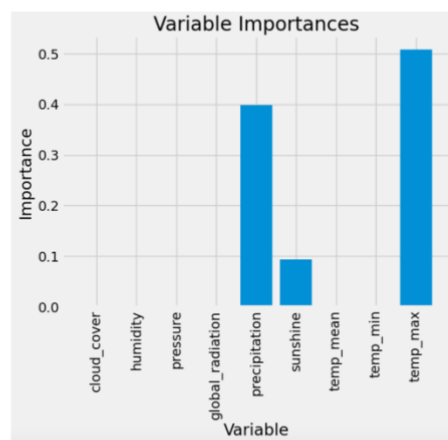
Pre-optimization Accuracy: N/A

Post-optimization Accuracy: 100% [Accurate?]

Forest:



Variable importances optimized:



Something is clearly wrong here. Trying a different station that matches one I already analyzed.

Maastricht, All Years

Results:

- Best GRID search hyperparameters are: {'max_depth': 10, 'max_features': 50, 'min_samples_leaf': 3, 'min_samples_split': 3, 'n_estimators': 100}
- Best GRID search score is: 0.8580640053267059
- Best RANDOM search hyperparameters are: {'criterion': 'entropy', 'max_depth': 11, 'max_features': 38, 'min_samples_leaf': 3, 'min_samples_split': 7, 'n_estimators': 180}
- Best RANDOM search score is: 0.8577733820599426

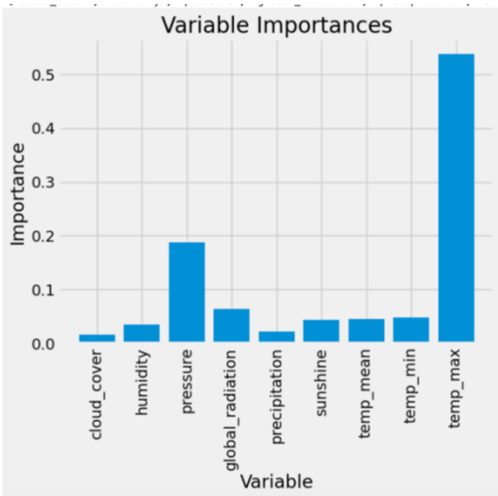
Parameters Run:

- clf3 = RandomForestClassifier(n_estimators = 100, max_depth=11, max_features=38, min_samples_leaf=3, min_samples_split=3, criterion = 'entropy')

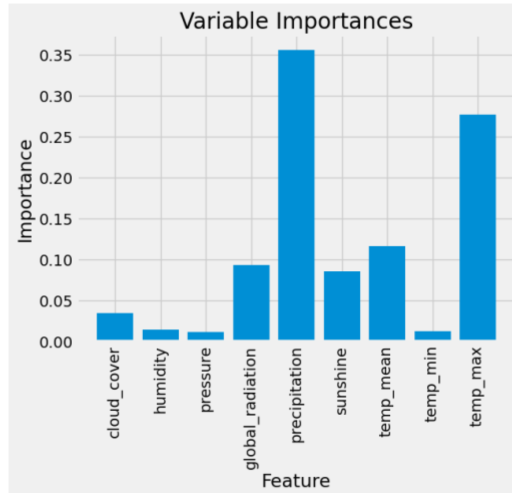
Pre-optimization accuracy: 100% [accurate?]

Post-optimization accuracy: 85.3%

Optimized importances:



Pre-optimization:



Random Forest Optimization: Observations

Optimization changes the results significantly. In the original model, the most important weather stations were Maastricht, Kassel, and Dusseldorf. In the optimized model, Heathrow, Valentia, and Belgrade led the way by a significant margin.

Optimization also shifted variable importances for types of weather observation. Temp max is the new leader, with pressure as a distant second. Precipitation is still a leader in the Heathrow observation. Still, given the absence of any other parameters, I would prefer to evaluate errors before drawing any conclusions based on that data.

CNN Optimization

Parameter	**2.2**	**2.4**
Epochs	30	47
Batch Size	16	460
N_Hidden	256	8
N_Classes	len(y_train[0])	15
Dropout		0.7296
Dropout Rate		0.1912
Kernel		1
Layers1		1
Layers2		2
Activation		Softsign
Neurons		61
Normalization		0.771
Optimizer		Adadelta

Post-Optimization

Accuracy: 61.2%

Loss: 1.83

Confusion matrix:

180/180 ━━━━━━━━━━━━━━━━━━━━ **1s** 3ms/step

Pred True	BASEL	BELGRADE	MAASTRICHT	VALENTIA
BASEL	3205	464	7	6
BELGRADE	799	293	0	0
BUDAPEST	177	37	0	0
DEBILT	56	26	0	0
DUSSELDORF	20	9	0	0
HEATHROW	72	10	0	0
KASSEL	10	1	0	0
LJUBLJANA	61	0	0	0
MAASTRICHT	9	0	0	0
MADRID	445	13	0	0
MUNCHENB	8	0	0	0
OSLO	5	0	0	0
STOCKHOLM	4	0	0	0
VALENTIA	1	0	0	0

Pre-Optimization:

Accuracy: 10.4%

Loss: 40.7

Confusion matrix:

```
144/144 ━━━━━━━━━━━━━━━━━━━━ 1s 8ms/step
Pred        BASEL  BELGRADE  BUDAPEST  DEBILT  DUSSELDORF  HEATHROW  KASSEL  \
True
BASEL           3       204        39      15          39        48     141
BELGRADE        0       118        19       0           0         0       0
BUDAPEST        0        10         5       0           0         0       0
DEBILT          0         1         3       0           0         0       0
DUSSELDORF      0         0         0       0           0         0       0
HEATHROW        0         0         0       1           0         1       0
KASSEL          0         2         0       0           0         0       0
LJUBLJANA       0         3         0       0           0         0       0
MAASTRICHT      0         0         0       0           0         0       0
MADRID          0         3         0       0           0         1       0
MUNCHENB        0         0         1       0           0         0       0
OSLO            0         0         0       0           0         0       0
STOCKHOLM       0         1         0       0           0         0       0
VALENTIA        0         0         0       0           0         0       0

Pred        LJUBLJANA  MAASTRICHT  MADRID  MUNCHENB  OSLO  SONNBLICK  \
True
BASEL              15          24    1365         2    87         63
BELGRADE            3           0     716         1     0          0
BUDAPEST            1           0     177         0     0          0
DEBILT              0           0      68         0     0          0
DUSSELDORF          0           0      25         0     0          0
HEATHROW            0           0      68         0     0          0
KASSEL              0           0       5         0     0          0
LJUBLJANA           0           0      27         0     0          0
MAASTRICHT          0           0       7         1     0          0
MADRID              0           0     342         0     0          2
MUNCHENB            0           0       2         1     0          0
OSLO                0           0       6         0     0          0
STOCKHOLM           0           0       0         0     0          0
VALENTIA            0           0       1         0     0          0

Pred        STOCKHOLM  VALENTIA
True
BASEL             102       821
BELGRADE            0         0
BUDAPEST            0         0
DEBILT              0         0
DUSSELDORF          0         0
HEATHROW            0         0
KASSEL              0         0
LJUBLJANA           0         0
MAASTRICHT          0         0
MADRID              0         0
MUNCHENB            0         0
OSLO                0         0
STOCKHOLM           0         0
VALENTIA            0         0
```

Analysis

Optimization significantly improved the accuracy of the model and reduced loss, but the model is now unable to recognize all 15 stations. It may be overfitting the data to Basel, though I would need to speak with a data scientist to confirm.

Iteration

Step 1: Selecting Components

I have already begun breaking the data down by weather station. I would like to continue that process as part of the iteration phase, looking at each station overall and then by decade.

If time and resources allow, I would then like to broaden the scope slightly and look at regional commonalities, grouping stations by type of climate (mountain locations, subtropical locations, etc.)

Once I found the most impactful types of measurements, I would look at the top 5 to 10, depending on results, across all stations.

Step 2: Choosing Models

At this point, I would use Random Forest models to run the per-station analyses, primarily because has generated more accurate results thus far.It also provides more interpretable results, allowing us to demonstrate with greater certainty where variable interactions come from. However, there are still problems with the optimization, which I would need to solve before investing significant additional resources.

CNN can run more complex analyses, but its lack of transparency and tendency to overfit make it a less appropriate choice at this phase. We may decide to implement it later, when we have a better idea of impactful weather data and need to identify more complex relationships.

Step 3: Recommended Variables

Based on the results of this analysis, I would recommend that Air Ambulance focus on the precipitation and maximum temperature variables. However, these optimizations have focused on finding patterns associated with pleasant weather, which may not be sufficient when looking at air travel safety. I would also recommend Air Ambulance track wind speed and cloud cover, which impact the trajectory of the aircraft.