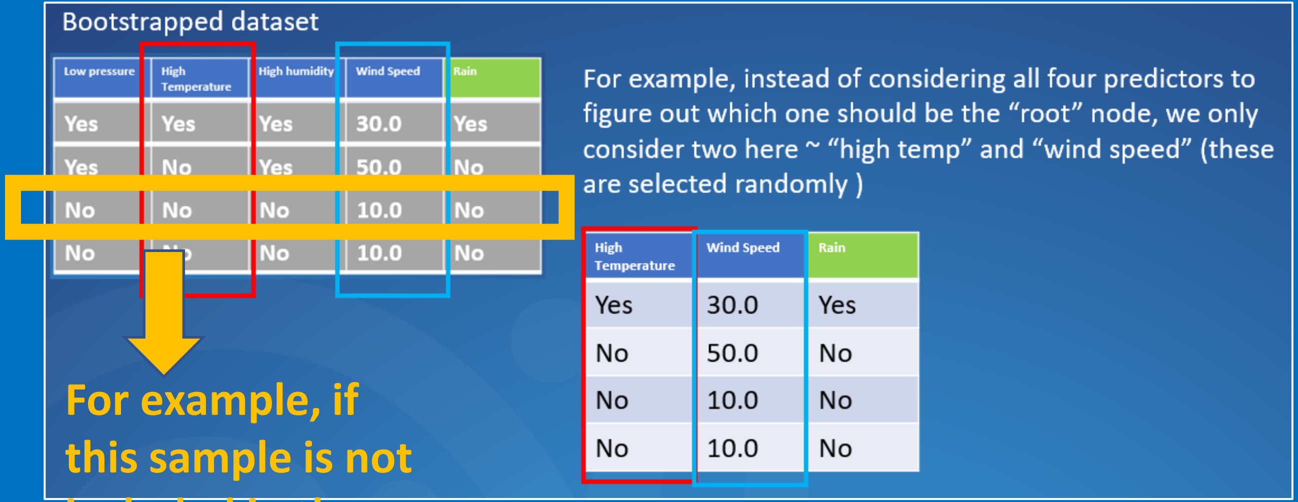
Random Forest

how to evaluate RF

So how can we know if RF is good



Bootstrapped dataset

Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

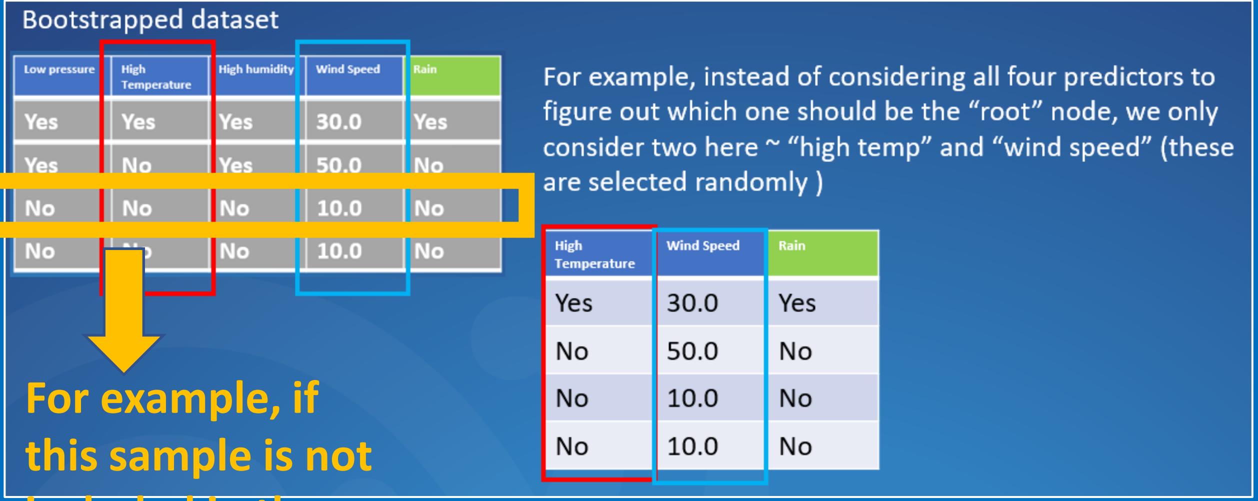
For example, instead of considering all four predictors to figure out which one should be the "root" node, we only consider two here ~ "high temp" and "wind speed" (these are selected randomly)

High Temperature	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

For example, if this sample is not included in the bootstrapped dataset

When we create the bootstrapped dataset, some original data are not included in the bootstrapped dataset

So how can we know if RF is good



Bootstrapped dataset

Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

For example, instead of considering all four predictors to figure out which one should be the “root” node, we only consider two here ~ “high temp” and “wind speed” (these are selected randomly)

High Temperature	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

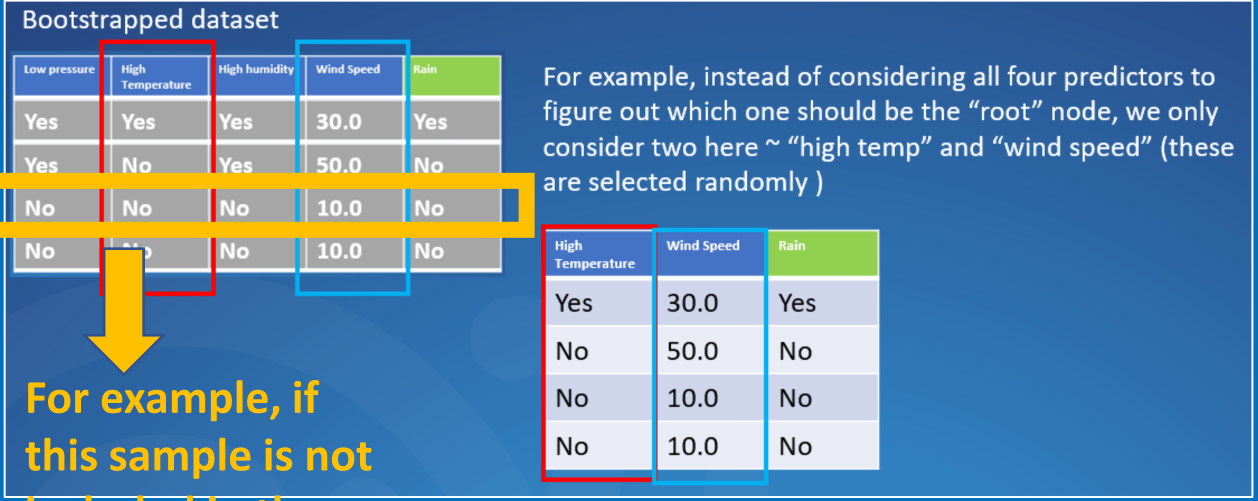
For example, if this sample is not included in the bootstrapped dataset

When we create the bootstrapped dataset, some original data are not included in the bootstrapped dataset

Those “missed” dataset is called “out-of-bag” dataset

When the sample size is big, there might be many “out-of-bag” dataset

So how can we know if RF is good



Bootstrapped dataset

Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

For example, instead of considering all four predictors to figure out which one should be the “root” node, we only consider two here ~ “high temp” and “wind speed” (these are selected randomly)

High Temperature	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

For example, if this sample is not included in the bootstrapped dataset

⇩

Those “missed” dataset is called “out-of-bag” dataset

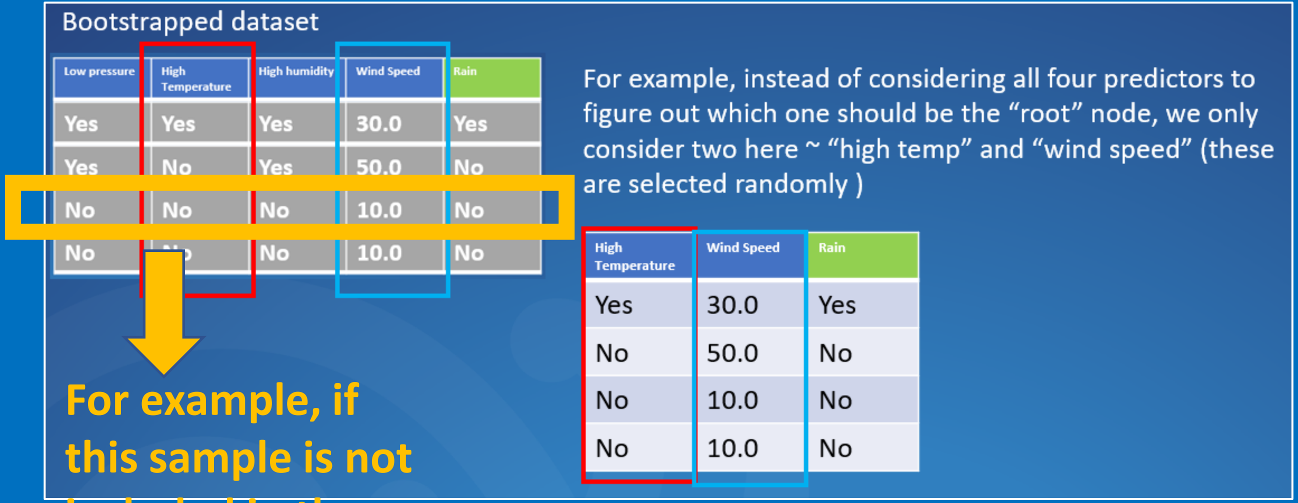
When the sample size is big, there might be many “out-of-bag” dataset

⇩

Since the “out-of-bag” dataset is not used to create the tree, we can use it to test the tree and see if the tree gets the right prediction

When we create the bootstrapped dataset, some original data are not included in the bootstrapped dataset

So how can we know if RF is good



Bootstrapped dataset

Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

For example, instead of considering all four predictors to figure out which one should be the "root" node, we only consider two here ~ "high temp" and "wind speed" (these are selected randomly)

High Temperature	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

For example, if this sample is not included in the bootstrapped dataset

When we create the bootstrapped dataset, some original data are not included in the bootstrapped dataset

⇩

Those "missed" dataset is called "out-of-bag" dataset

When the sample size is big, there might be many "out-of-bag" dataset

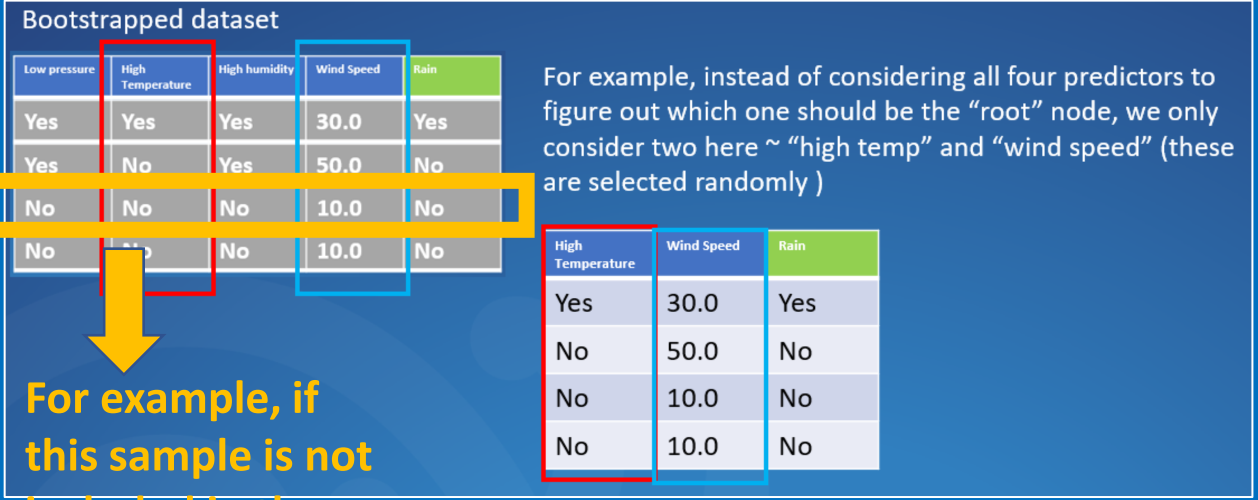
⇩

Since the "out-of-bag" dataset is not used to create the tree, we can use it to test the tree and see if the tree gets the right prediction

⇨

We can search all the trees without this dataset, and get all the predictions

So how can we know if RF is good



Bootstrapped dataset

Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

For example, instead of considering all four predictors to figure out which one should be the "root" node, we only consider two here ~ "high temp" and "wind speed" (these are selected randomly)

High Temperature	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

When we create the bootstrapped dataset, some original data are not included in the bootstrapped dataset

For example, if this sample is not included in the bootstrapped dataset

Those "missed" dataset is called "out-of-bag" dataset

When the sample size is big, there might be many "out-of-bag" dataset

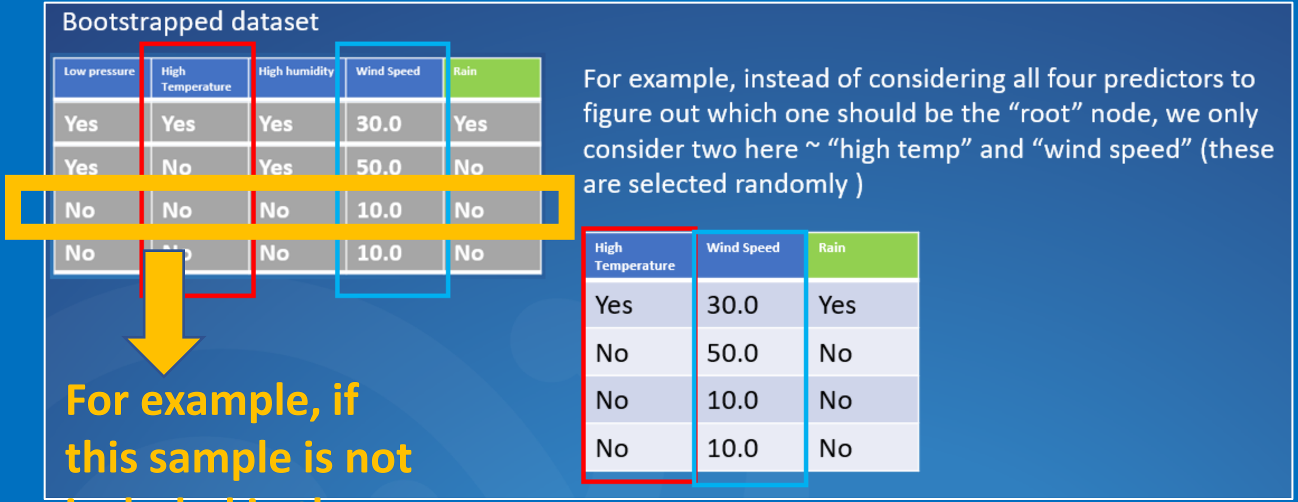
Since the "out-of-bag" dataset is not used to create the tree, we can use it to test the tree and see if the tree gets the right prediction

We can search all the trees without this dataset, and get all the predictions

The out-of-bag dataset will give us the prediction "YES", which does not match the dataset observation "NO"

Rain: YES	Rain: NO
11	7

So how can we know if RF is good



Bootstrapped dataset

Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

For example, instead of considering all four predictors to figure out which one should be the "root" node, we only consider two here ~ "high temp" and "wind speed" (these are selected randomly)

High Temperature	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

For example, if this sample is not included in the bootstrapped dataset

When we create the bootstrapped dataset, some original data are not included in the bootstrapped dataset

Those "missed" dataset is called "out-of-bag" dataset

When the sample size is big, there might be many "out-of-bag" dataset

Since the "out-of-bag" dataset is not used to create the tree, we can use it to test the tree and see if the tree gets the right prediction

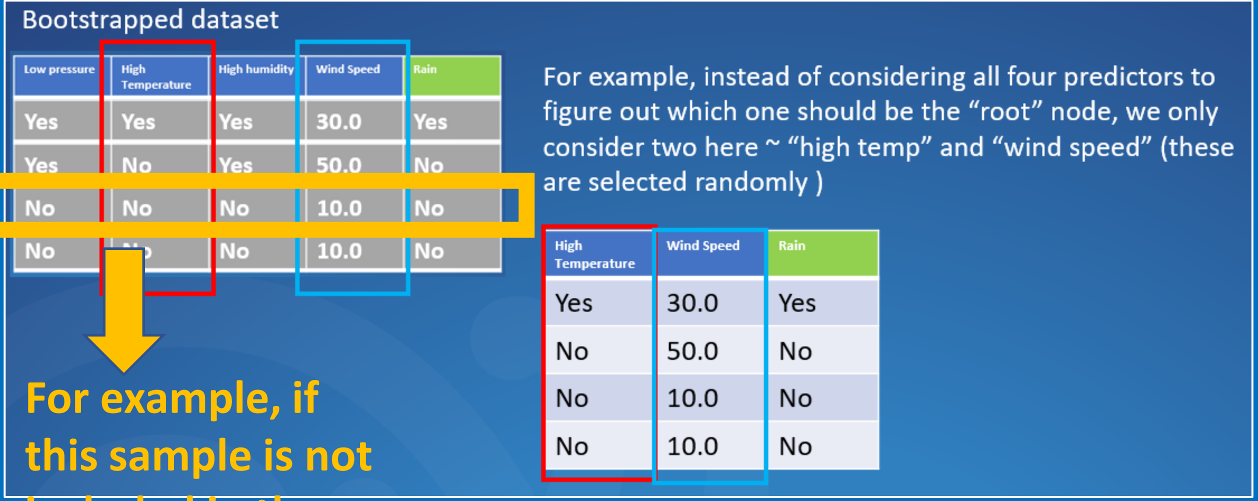
We can search all the trees without this dataset, and get all the predictions

Rain: YES	Rain: NO
11	7

The out-of-bag dataset will give us the prediction "YES", which does not match the dataset observation "NO"

We repeat this process for all "out-of-bag" samples for all the trees

So how can we know if RF is good



Bootstrapped dataset

Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

For example, instead of considering all four predictors to figure out which one should be the "root" node, we only consider two here ~ "high temp" and "wind speed" (these are selected randomly)

High Temperature	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

When we create the bootstrapped dataset, some original data are not included in the bootstrapped dataset

For example, if this sample is not included in the bootstrapped dataset

Those "missed" dataset is called "out-of-bag" dataset

When the sample size is big, there might be many "out-of-bag" dataset

Since the "out-of-bag" dataset is not used to create the tree, we can use it to test the tree and see if the tree gets the right prediction

We can search all the trees without this dataset, and get all the predictions

Rain: YES	Rain: NO
11	7

The out-of-bag dataset will give us the prediction "YES", which does not match the dataset observation "NO"

We repeat this process for all "out-of-bag" samples for all the trees

Ultimately, we can measure how accurate the RF is based on the proportion of how many "out-of-bag" dataset gets the correct prediction