NS/BH XRB classification : Random Forest

Data Imputation / Feature Importance / Feature Identification

Recap

- NS/BH classification
  - Classifier – Random Forest
  - Data Imputation – Median / Correlation
  - Result -
    - Total : 460 filtered observations ,
    - Ambiguous Class (pred prob < 0.8) : 42
    - Wrong classification : 0
  - Feature Importance
    - Feature inference was not clear

Aim

- Improve – Accuracy
  - Reduce ambiguity
- Better feature importance study

Improvement scope

- Classifier tuning
- Imputation Method
- Feature Importance

Imputation Method

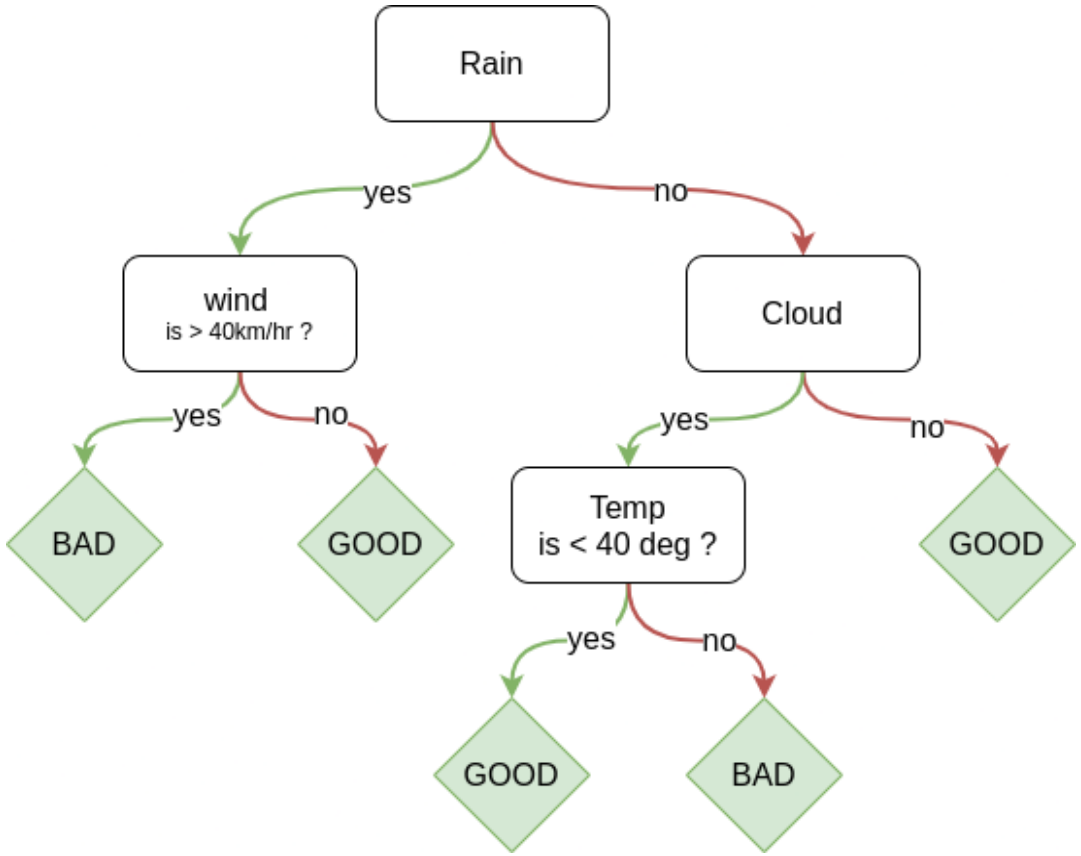
- Previously Used
  - Median imputation
  - Correlation based imputation
- New method
  - Using Random forest

Random Forest

Dataset

	Rain	wind	temp	cloud	Class
Day 1	yes	54	23	yes	BAD
Day 2	yes	20	20	yes	GOOD
Day 3	No	23	23	yes	GOOD
Day 4	No	18	43	no	BAD

Decision Tree

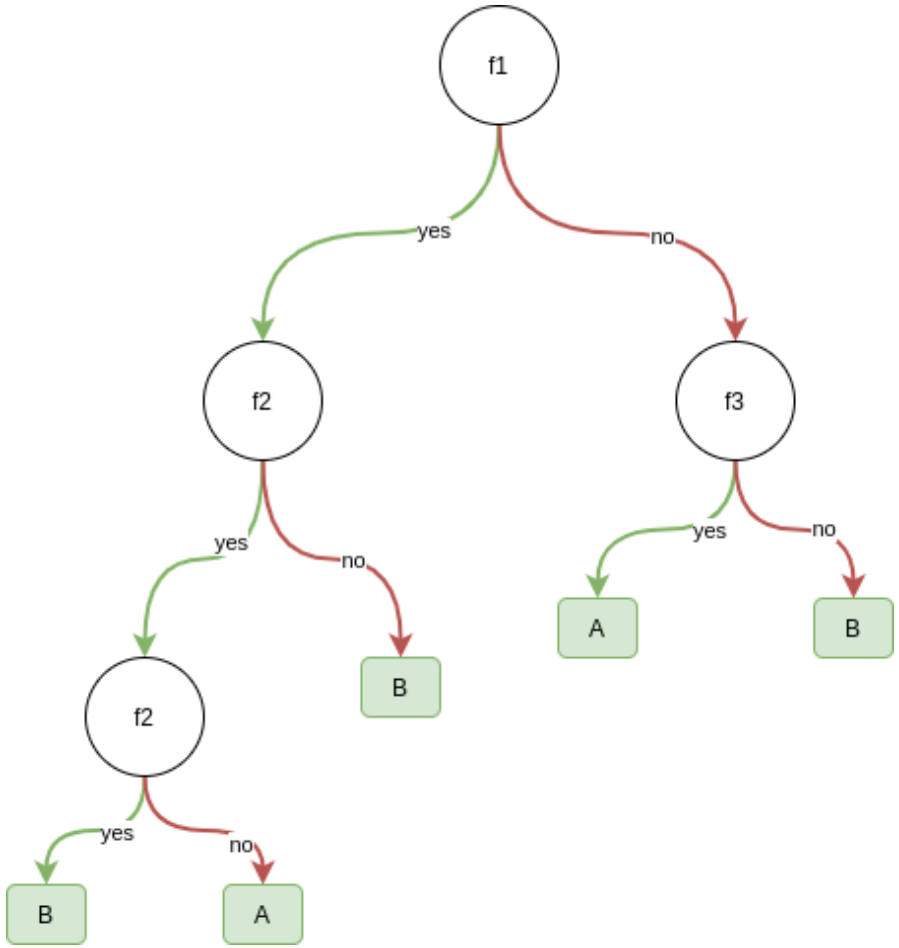


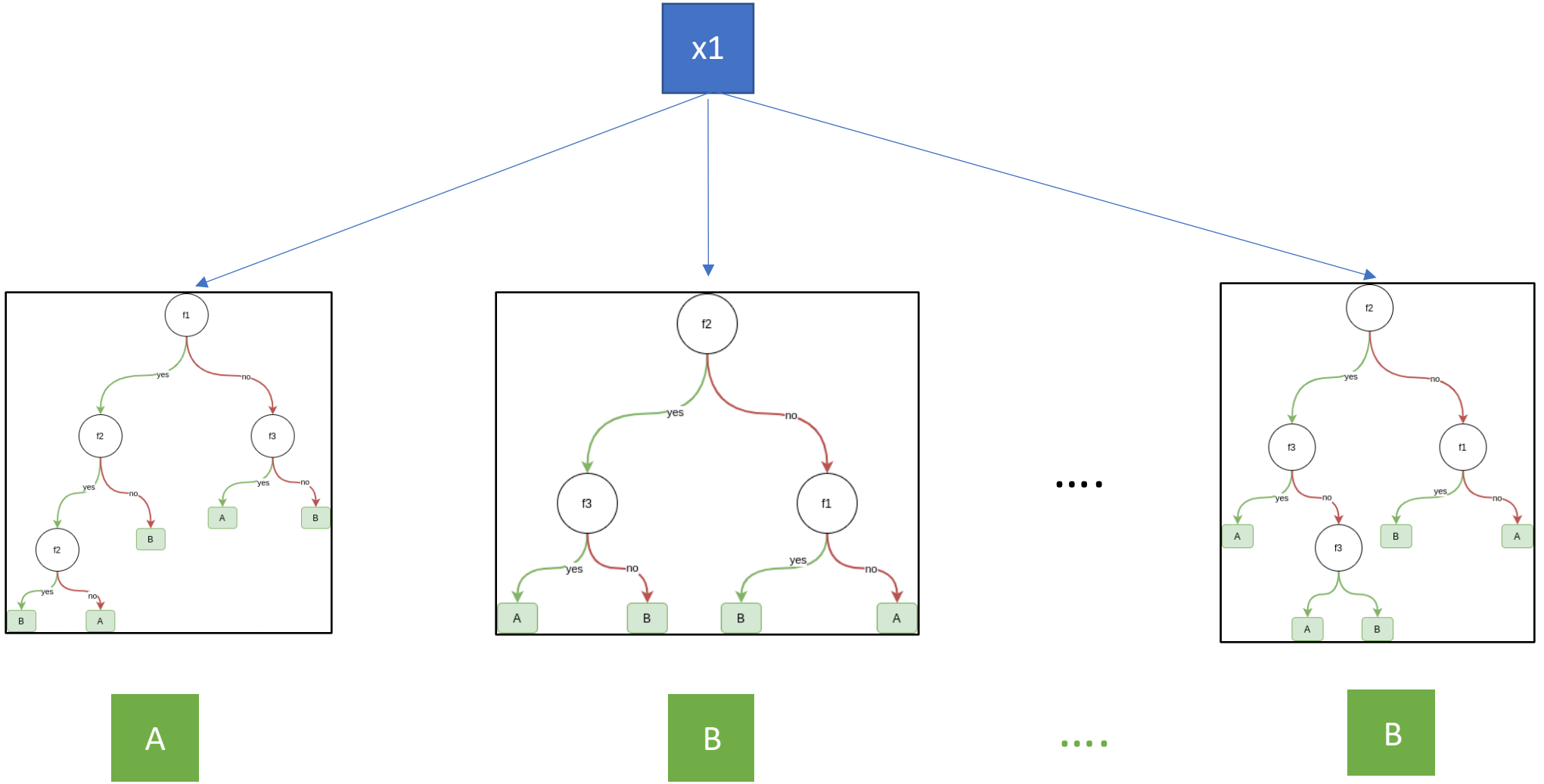
Random Forest

Dataset

	f1	f2	..	fn	class
x1					A
x2	---				A
x3			----		B
..					A
..					B
xn					B

Decision Tree





Max voted class is assigned to X

Imputation With RF

Find Closest example

Use that example to fill missing value

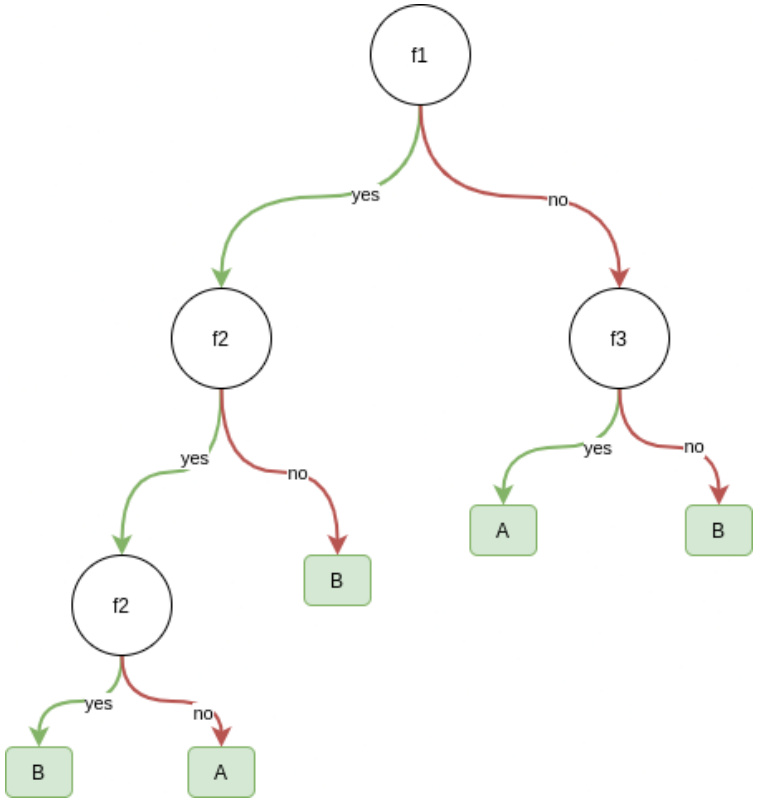
Imputation With RF

Find Closest example

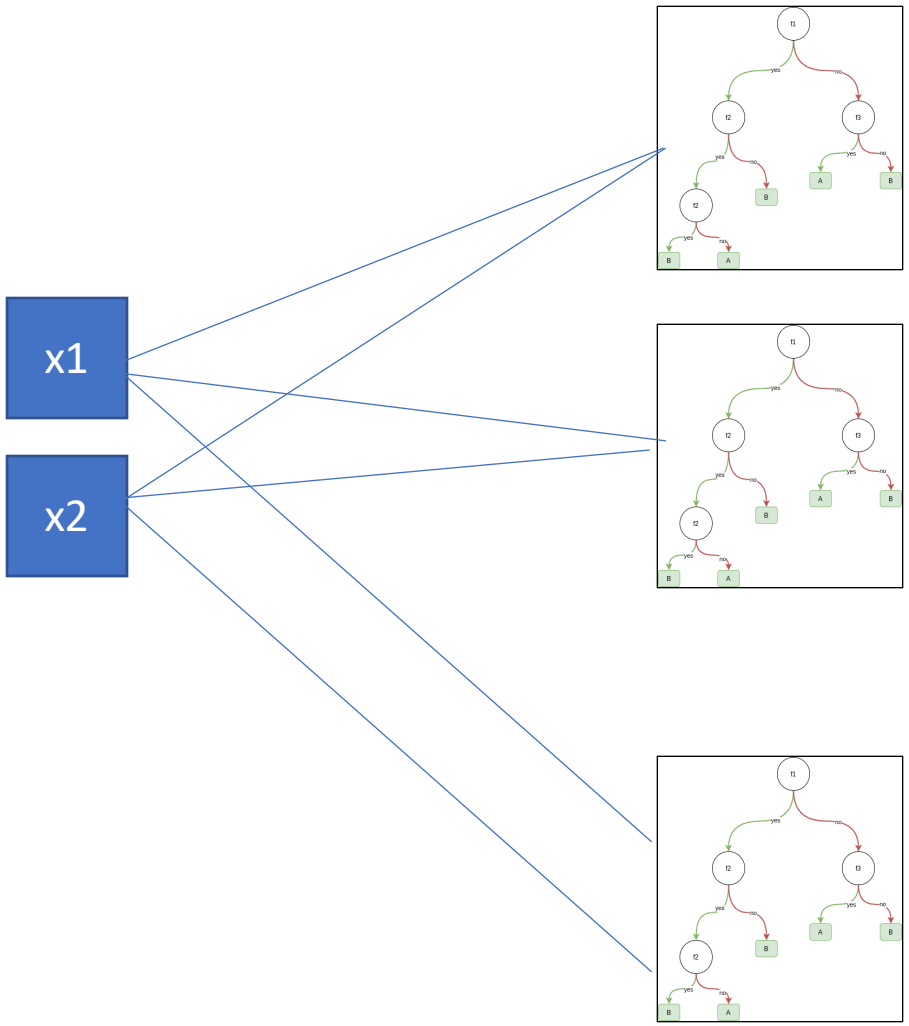
HOW ?

x1

x2



Imputation With RF



Count trees. / total trees

Proximity value for (x1,x2)

Imputation With RF

Find Closest example

Proximity Matrix

	x1	x2	x3	...	xn
x1	1	0.2	0.8	--	0.99
x2	0.2	1	0.4	--	0.1
x3	0.8	0.4	1	--	0.3
..	--	--	--	1	--
xn	0.99	0.1	0.3	--	1

	f1	f2	..	fn	class
x1		2.3			A
x2	---	NAN			A
x3		1.2	----		B
..					A
..					B
xn		8.6			B

Imputation With RF

Fill missing values with feature median

Calculate Proximity Matrix

Use this to update missing value

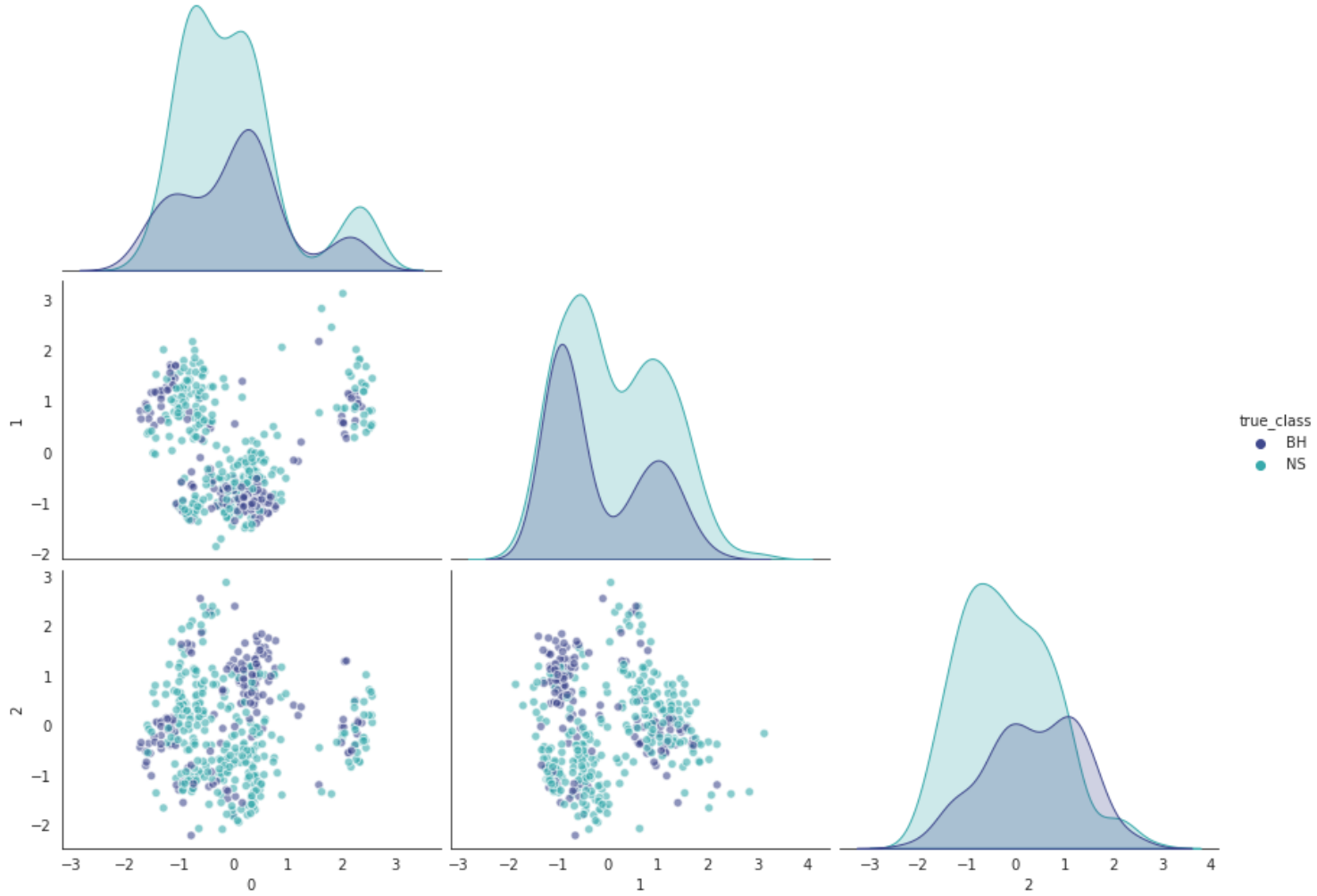
	f1	f2	..	fn	class
x1		2.3			A
x2	---	NAN			A
x3		1.2	----		B
..					A
..					B
xn		8.6			B

	x1	x2	x3	...	xn
x1	1	0.2	0.8	--	0.99
x2	0.2	1	0.4	--	0.1
x3	0.8	0.4	1	--	0.3
..	--	--	--	1	--
xn	0.99	0.1	0.3	--	1

$$f2(x2) = \frac{f2(x2) * p(x1,x2) + f2(x3) * p(x2,x3) + .. + f2(xn) * p(x2,xn)}{p(x1,x2) + p(x2,x3) + .. + p(x2,xn)}$$

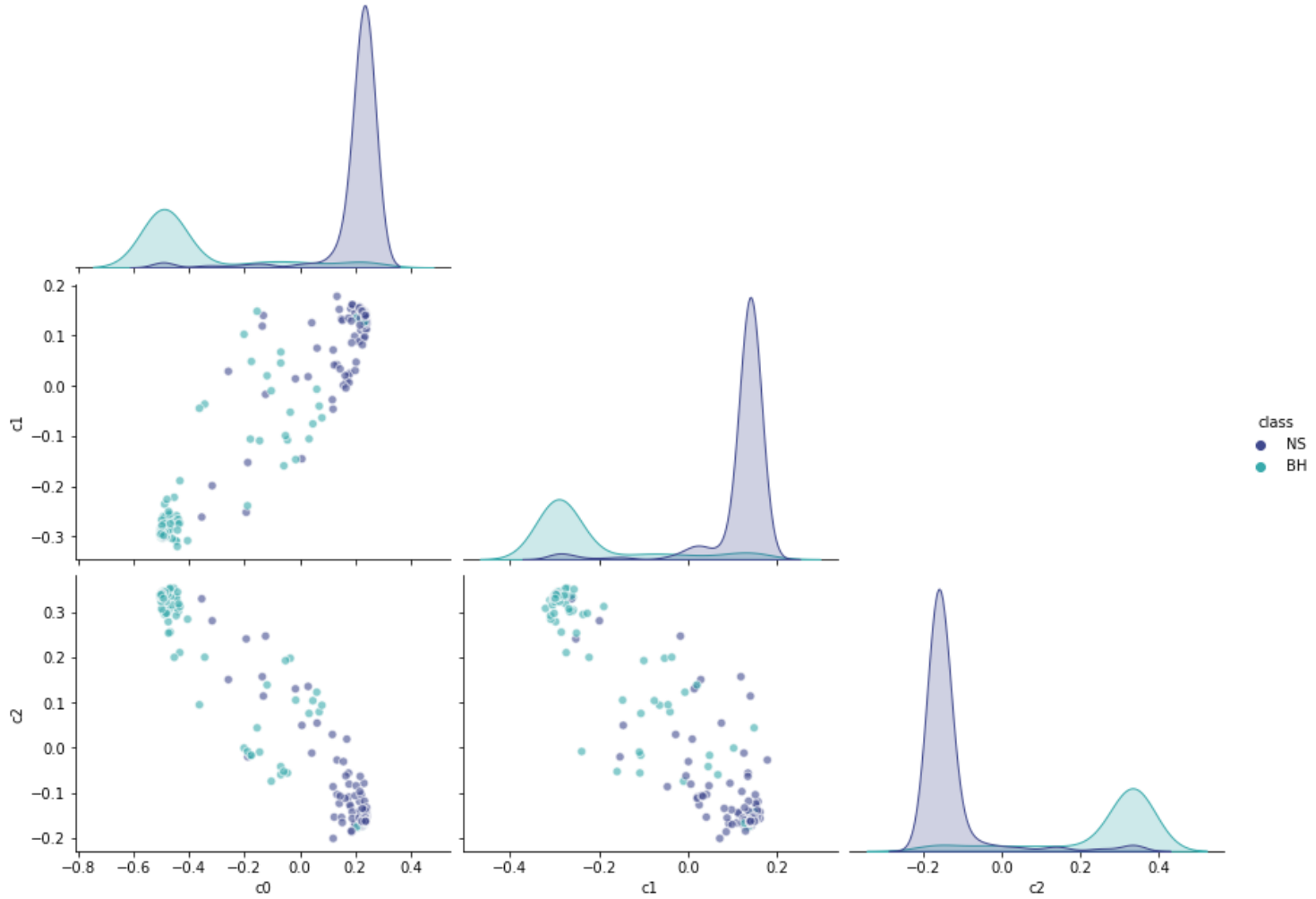
Result

Cluster Using correlation



Result

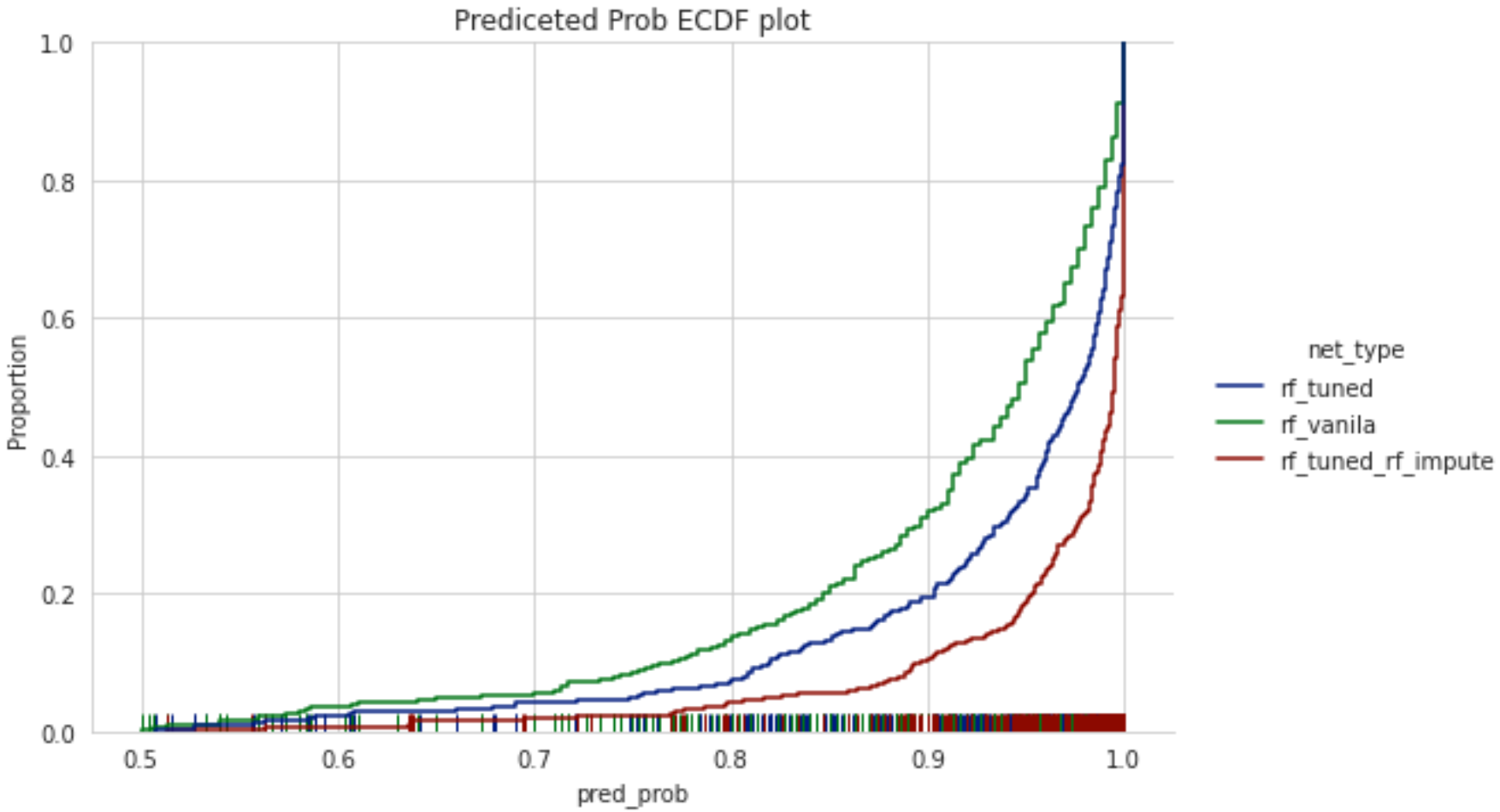
Cluster Using Proximity Matrix



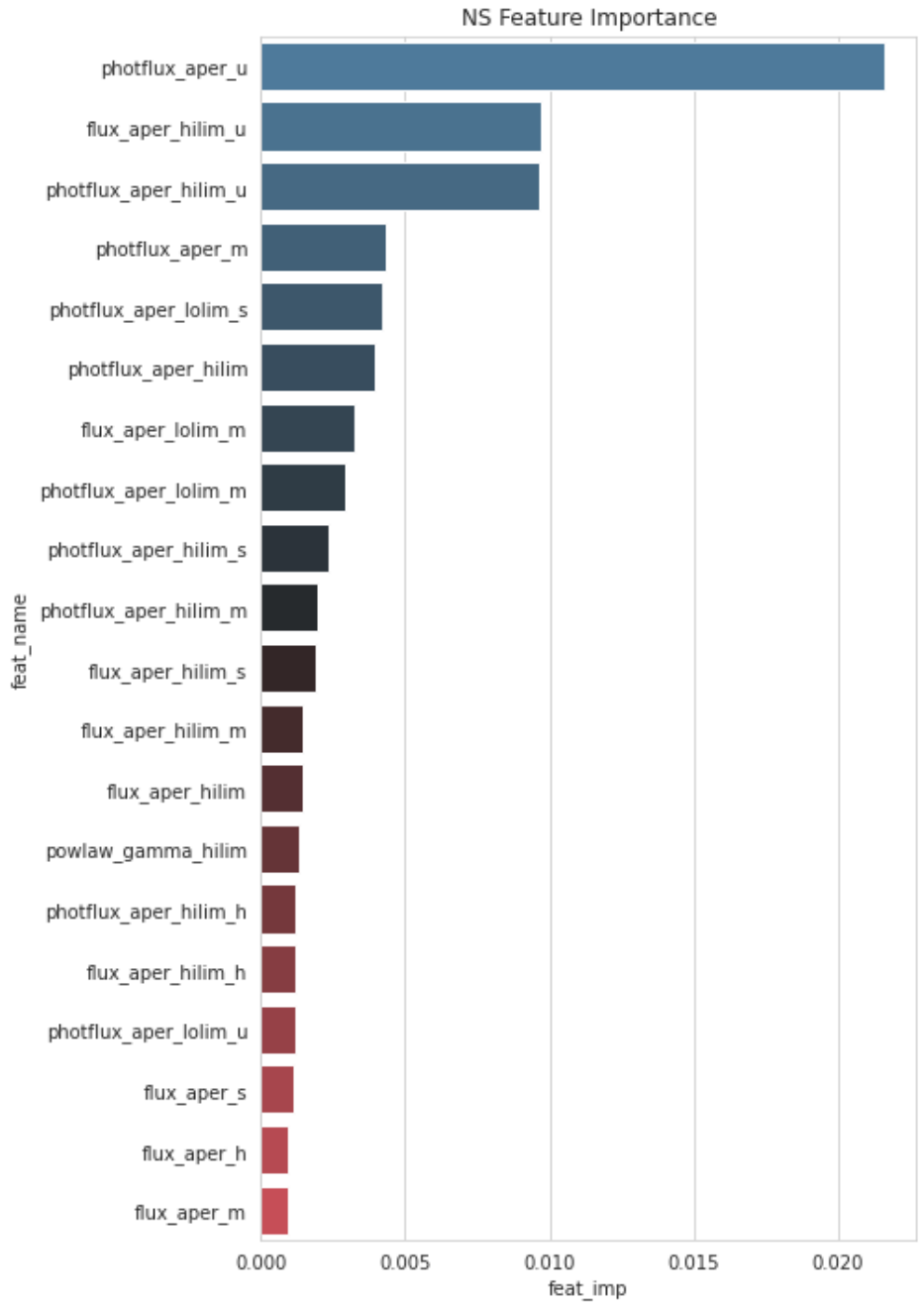
Result

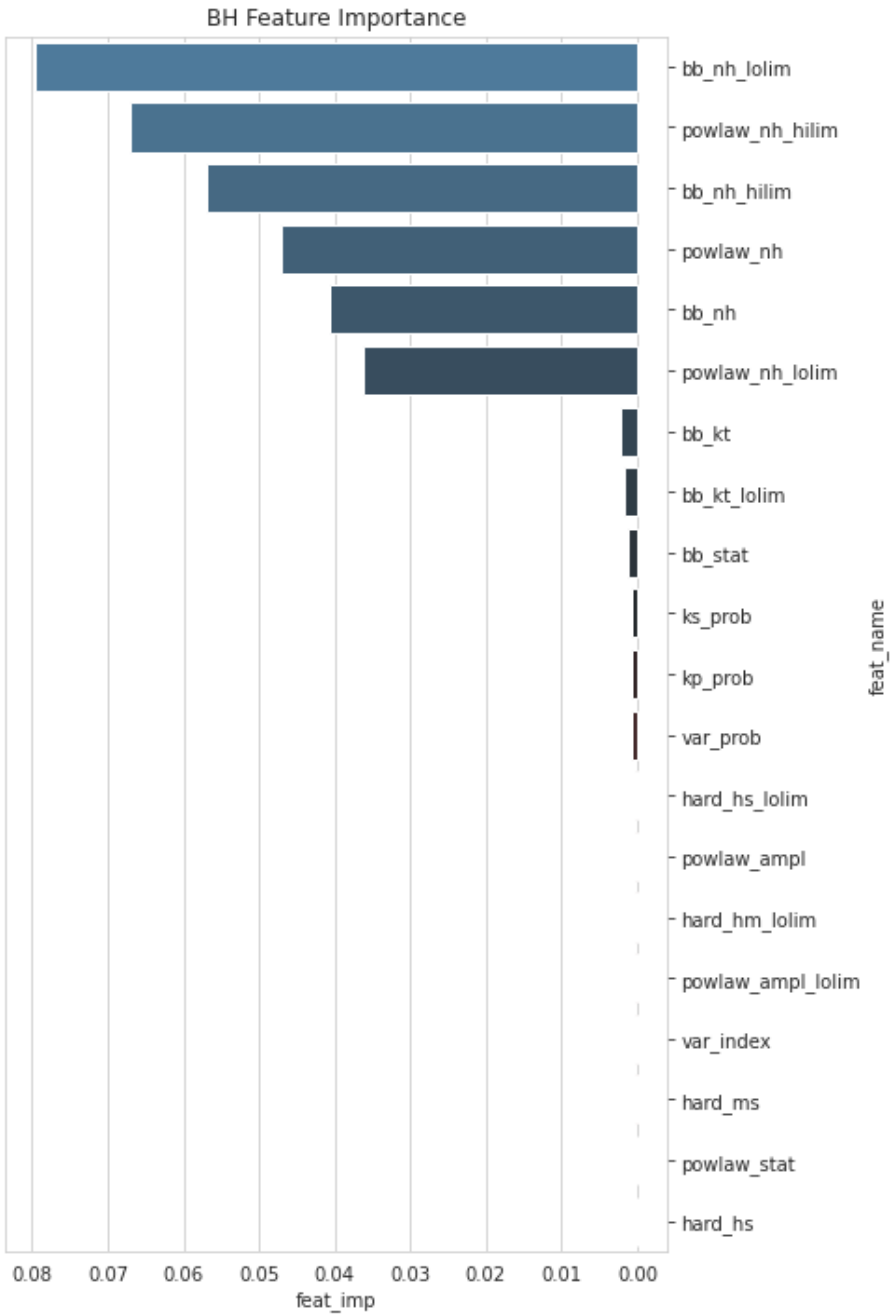
Using correlation Imputation	Using RF Imputation
Total Predictions – 460 Ambiguous – 42 Wrong – 0	Total Predictions – 460 Ambiguous – 21 Wrong – 0

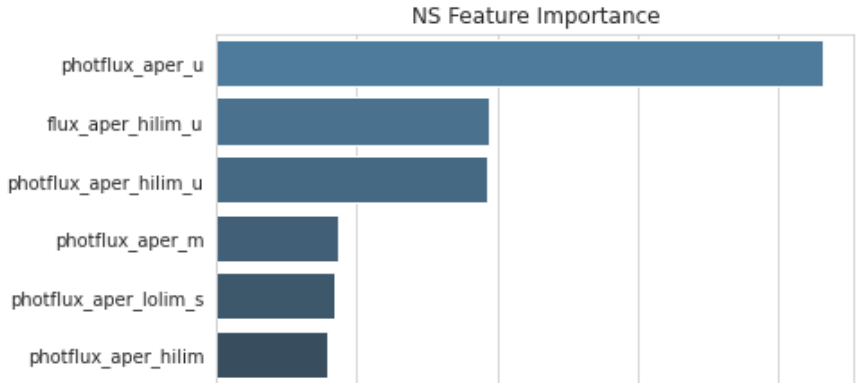
Result

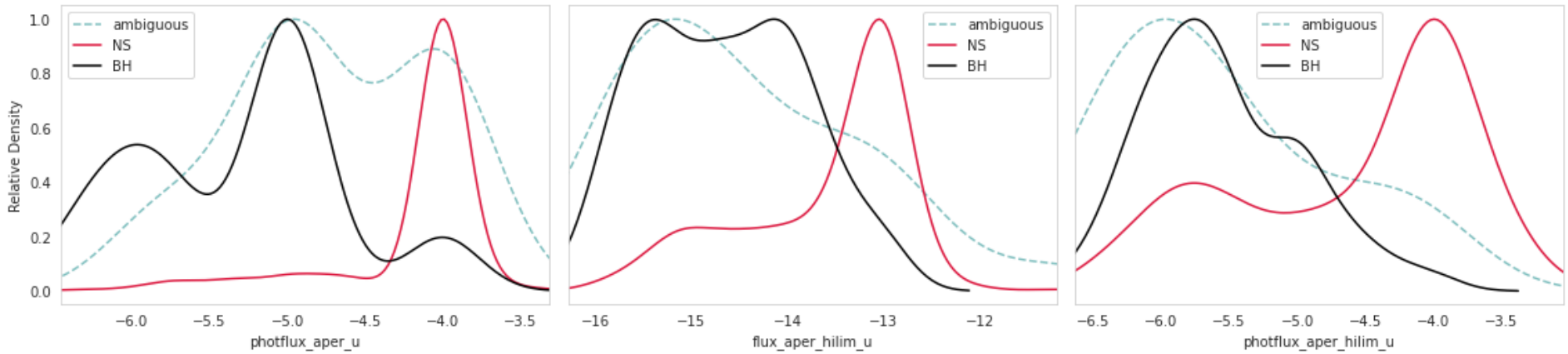

Prediceted Prob ECDF plot
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.5
0.6
0.7
0.8
0.9
1.0
pred_prob
net_type
rf_tuned
rf_vanila
rf_tuned_rf_impute

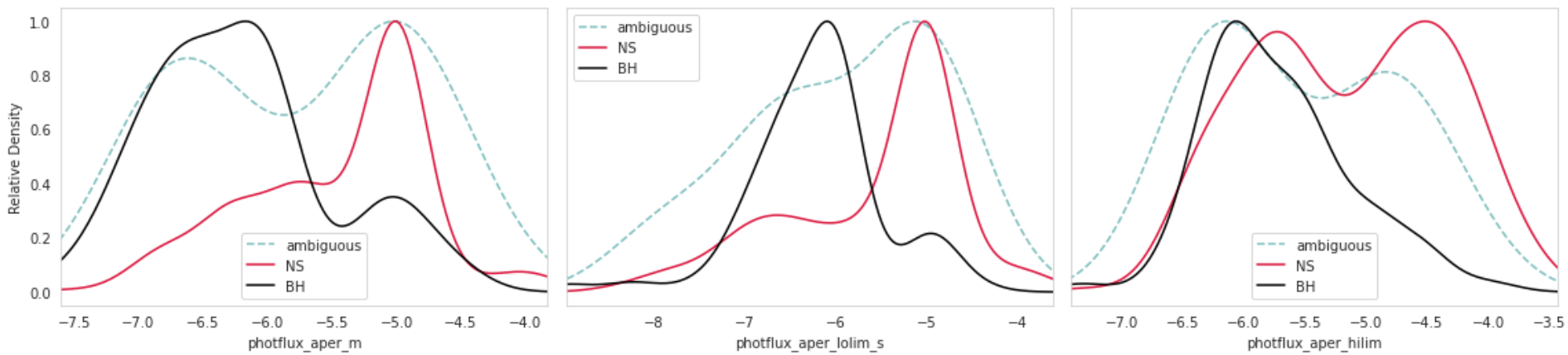
Feature Importance

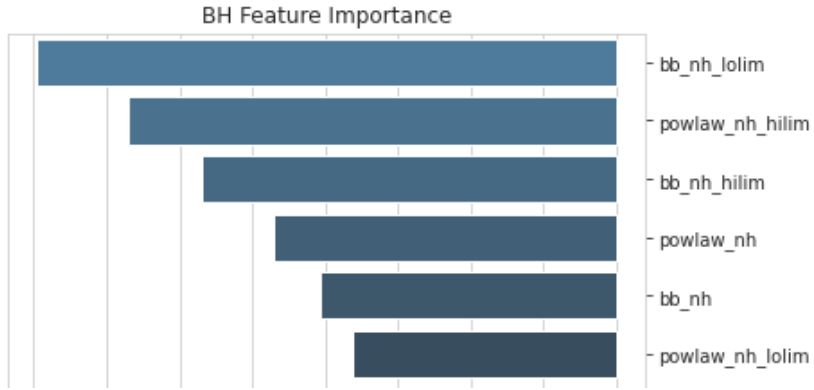


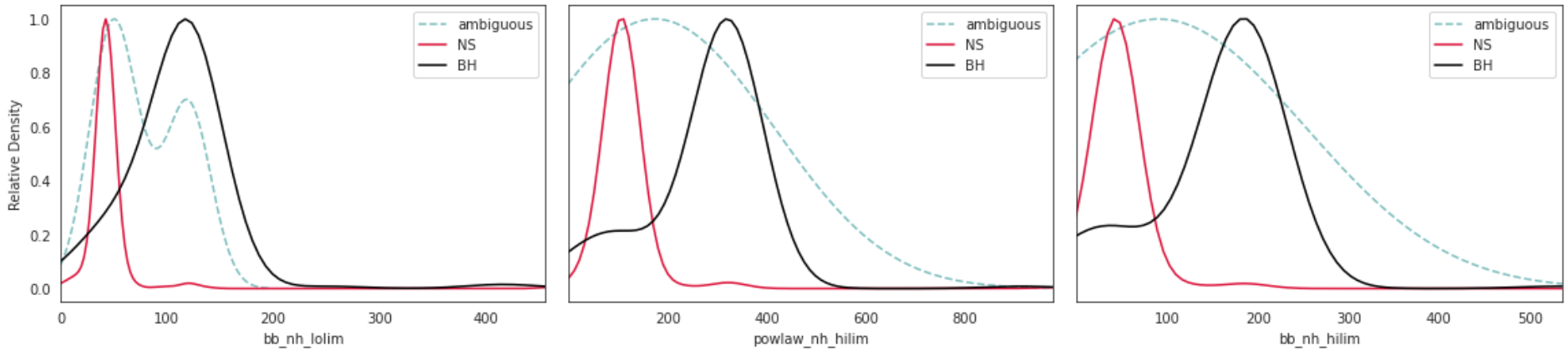


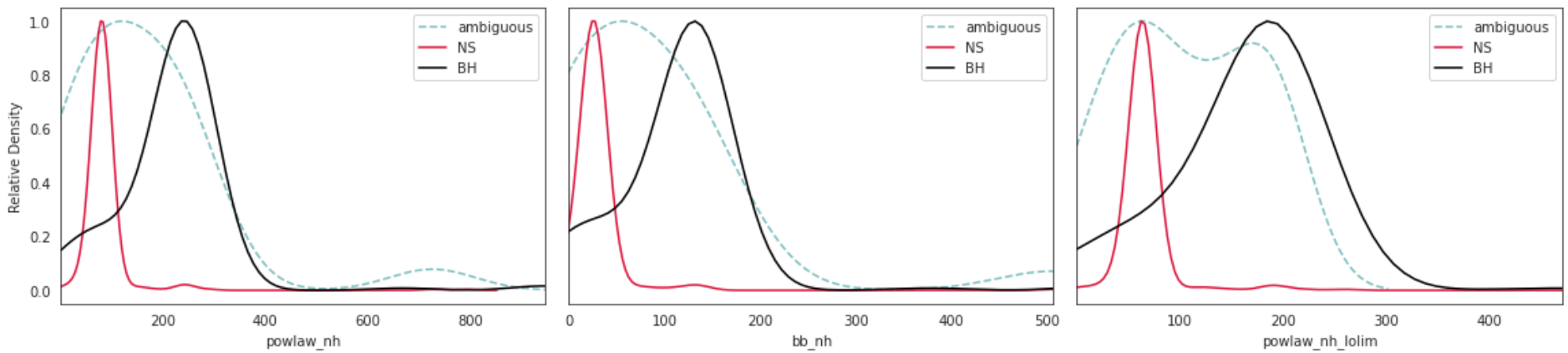

NS Feature Importance
photflux_aper_u
flux_aper_hilim_u
photflux_aper_hilim_u
photflux_aper_m
photflux_aper_lolim_s
photflux_aper_hilim


ambiguous
NS
BH
Relative Density
photflux_aper_u
flux_aper_hilim_u
photflux_aper_hilim_u


ambiguous
NS
BH
Relative Density
photflux_aper_m
photflux_aper_lolim_s
photflux_aper_hilim


BH Feature Importance
bb_nh_lolim
powlaw_nh_hilim
bb_nh_hilim
powlaw_nh
bb_nh
powlaw_nh_lolim


Relative Density
ambiguous
NS
BH
bb_nh_lolim
powlaw_nh_hilim
bb_nh_hilim


Relative Density
ambiguous
NS
BH
powlaw_nh
bb_nh
powlaw_nh_lolim