Pyridine Phosphination - Best Model Feature Analysis

Table of Contents

Published Model

5th Place with 5 Features

Table:

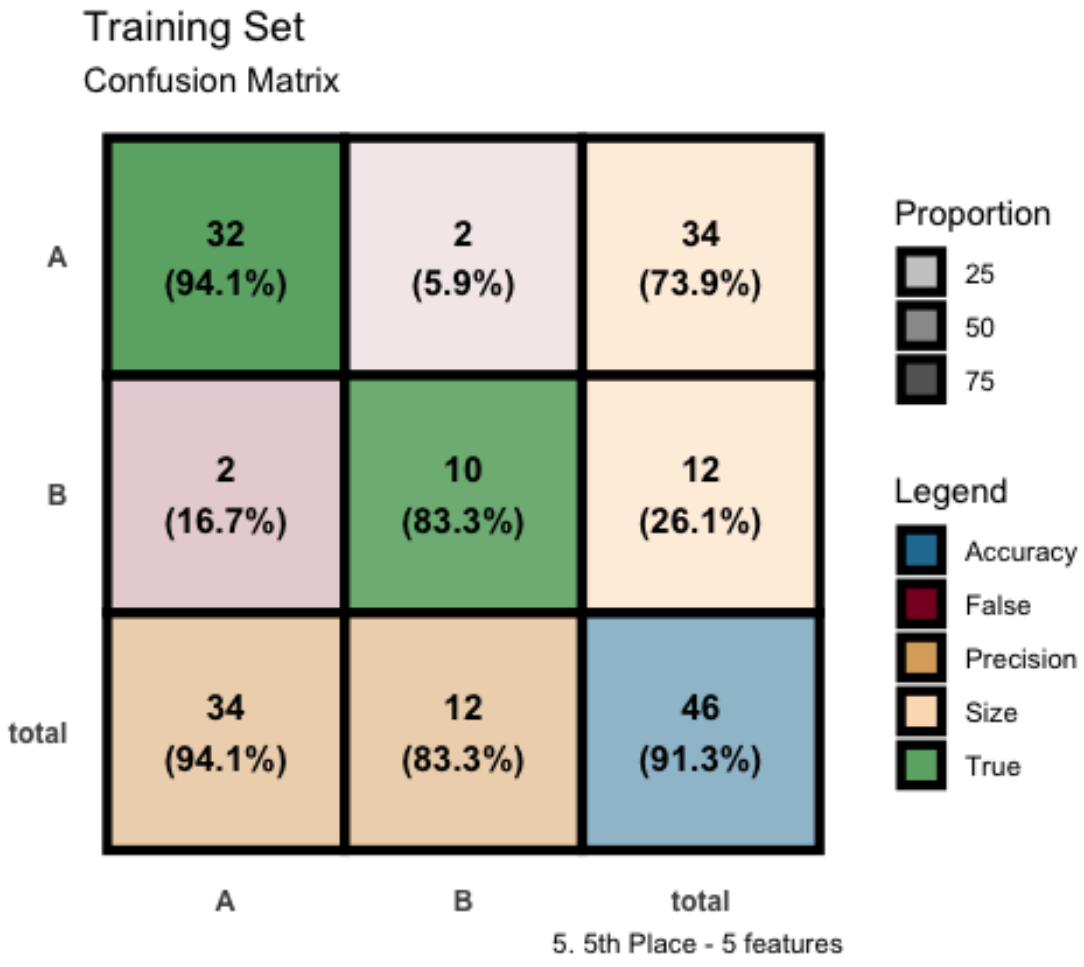
Full Model Stats - Overall Accuracy and Pseudo-R2

Accuracy	McFadden_R2
91.3%	0.578

Model Coefficients

	x
(Intercept)	38.099048
Avg_NPA_SM	48.412036
fr_Ar_N1	-31.903705
fr_Ar_N2	-60.673746
fr_Ar_N3	-66.149054
fr_aryl_methyl1	-4.592317
fr_aryl_methyl2	-37.006653
fr_benzene1	-1.970290
fr_benzene2	-77.887220
fr_halogen1	1.607589
fr_halogen2	2.929693
fr_halogen3	2.595277
fr_halogen4	-2.087011

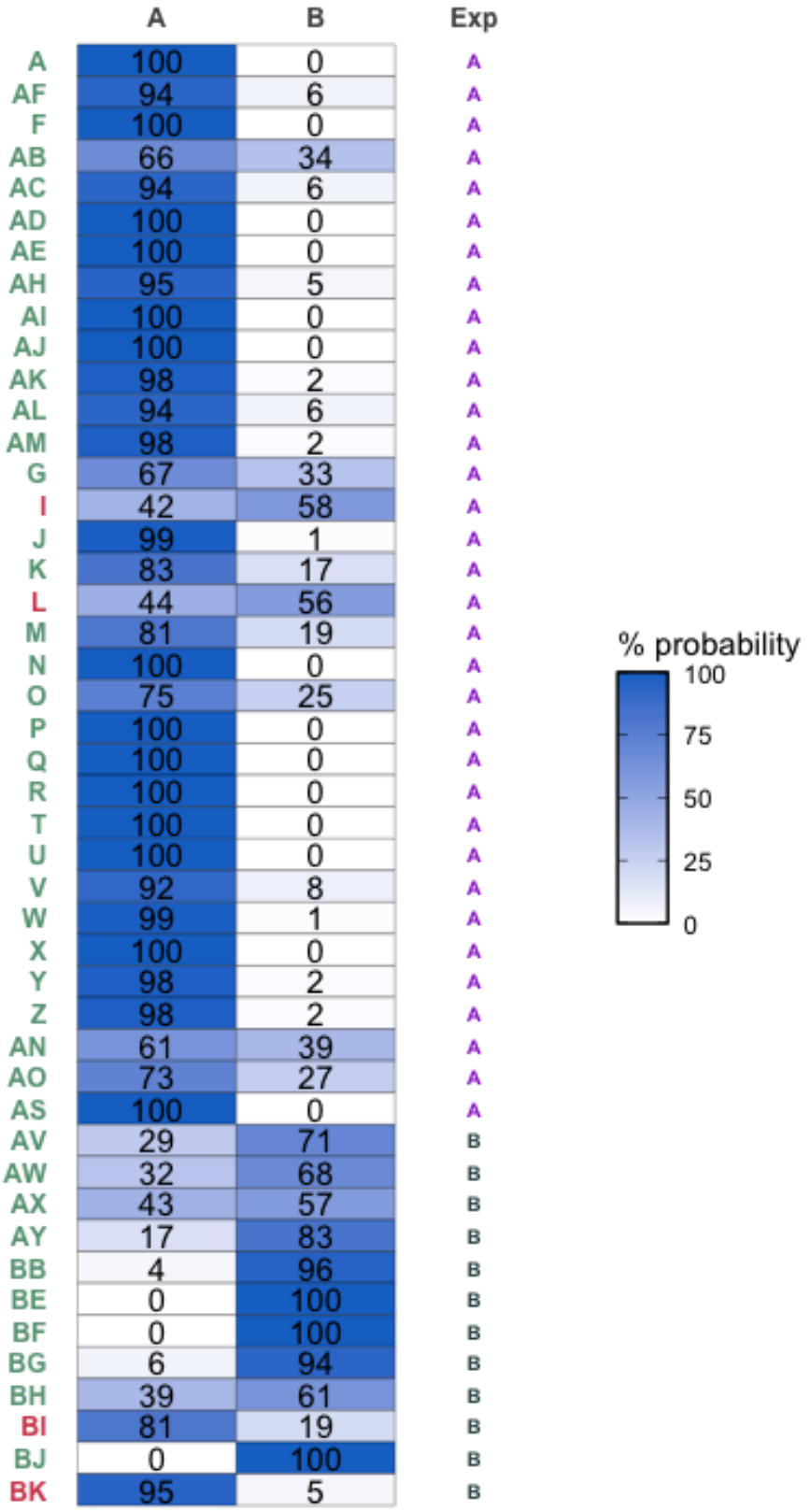
	x
fr_halogen5	2.187832
fr_halogen6	-2.311938



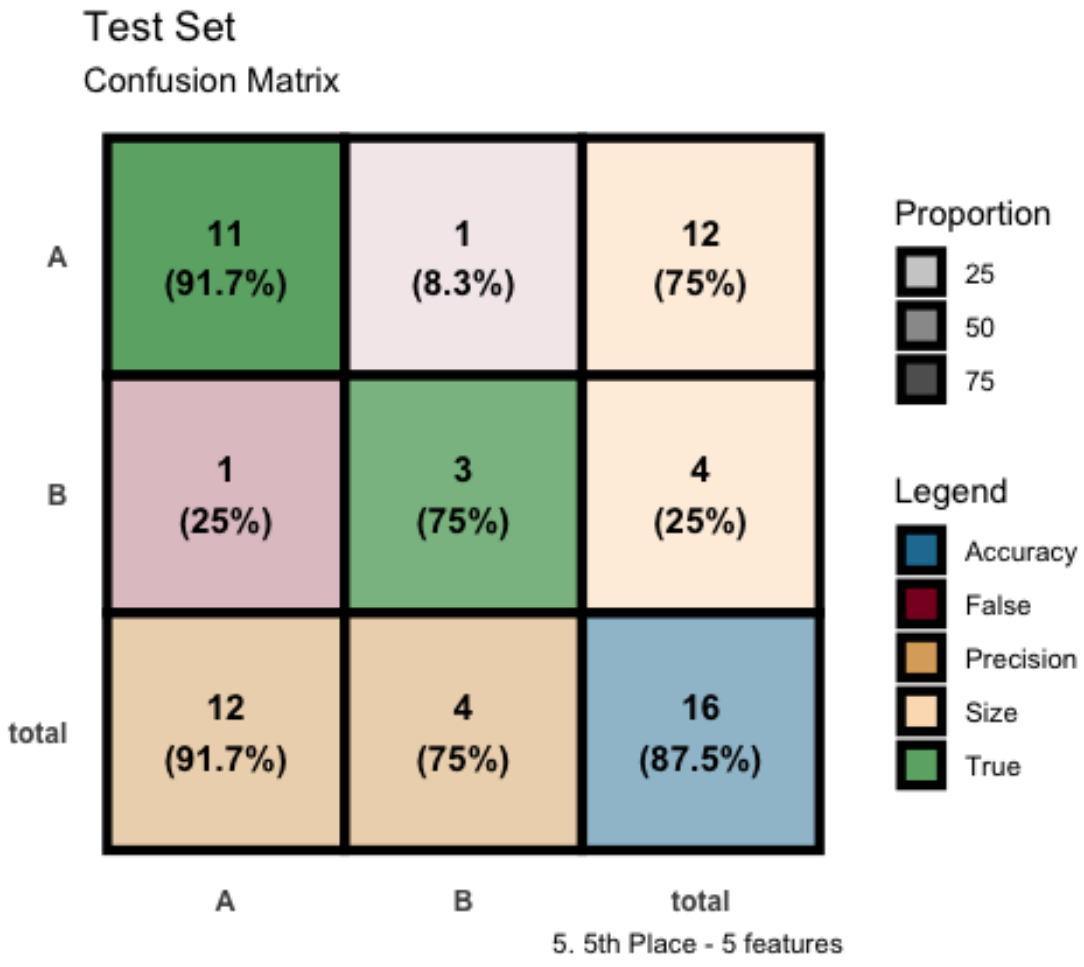
5. 5th Place - 5 features

Training Set

Probability Heatmap



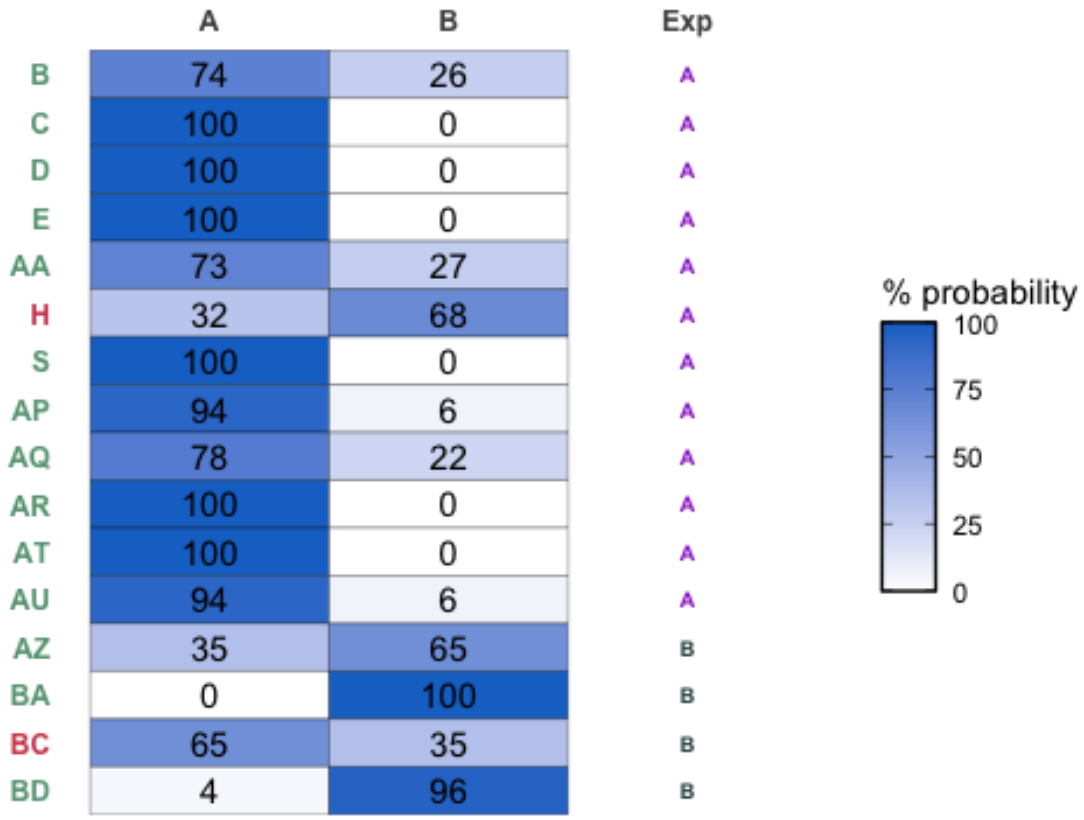
5. 5th Place - 5 features



5. 5th Place - 5 features

Test Set

Probability Heatmap



5. 5th Place - 5 features

Model Analysis - Variable Impact

1. Avg_NPA_SM Removed

formula

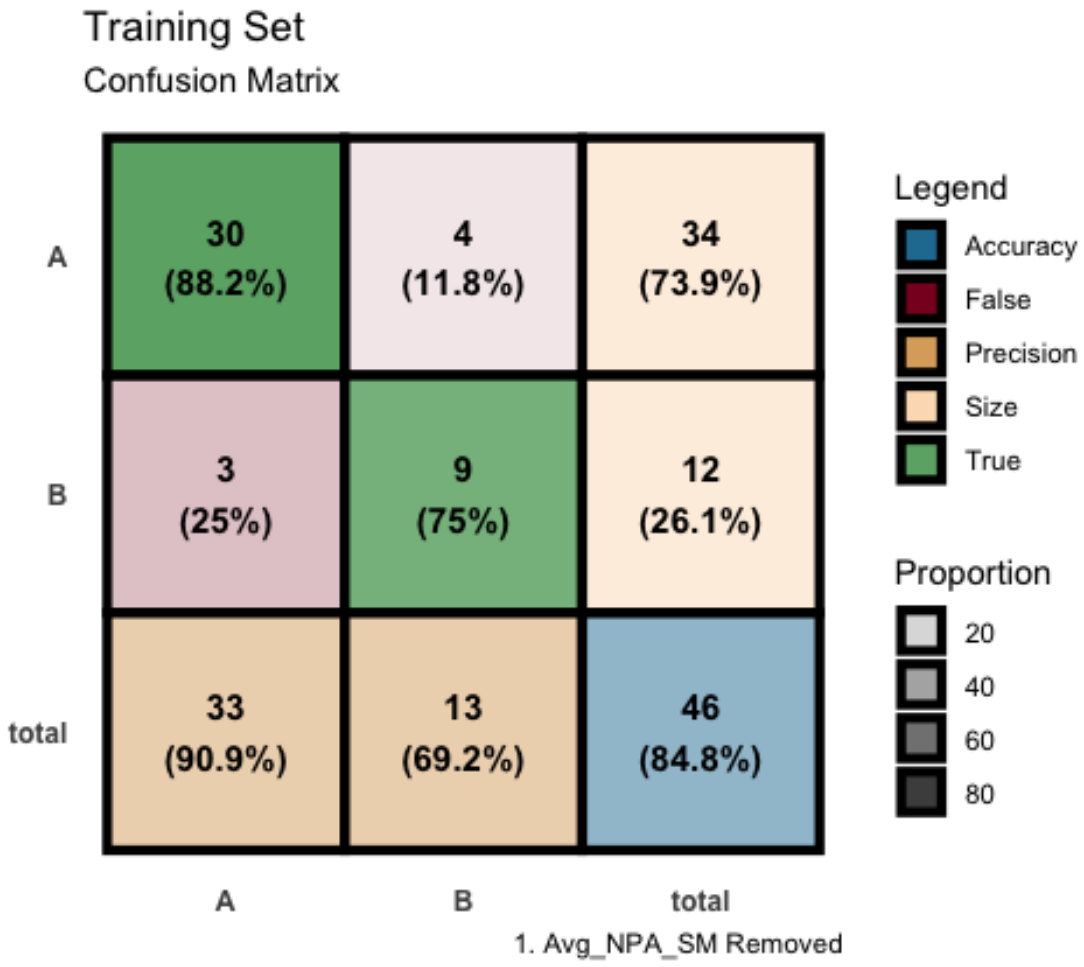
```
class ~ fr_Ar_N + fr_aryl_methyl + fr_benzene + fr_halogen
```

Full Model Stats - Overall Accuracy and Pseudo-R2

Accuracy	McFadden_R2
84.78%	0.52

Model Coefficients

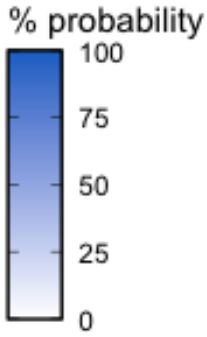
	x
(Intercept)	14.653960
fr_Ar_N1	-14.641780
fr_Ar_N2	-28.886527
fr_Ar_N3	-30.760596
fr_aryl_methyl1	-3.720066
fr_aryl_methyl2	-16.691481
fr_benzene1	-2.364301
fr_benzene2	-35.077472
fr_halogen1	2.342569
fr_halogen2	4.212291
fr_halogen3	1.848104
fr_halogen4	-1.081430
fr_halogen5	2.352174
fr_halogen6	-9.373352


Training Set
Confusion Matrix
A
B
total
30
(88.2%)
4
(11.8%)
34
(73.9%)
3
(25%)
9
(75%)
12
(26.1%)
33
(90.9%)
13
(69.2%)
46
(84.8%)
A
B
total
1. Avg_NPA_SM Removed
Legend
Accuracy
False
Precision
Size
True
Proportion
20
40
60
80

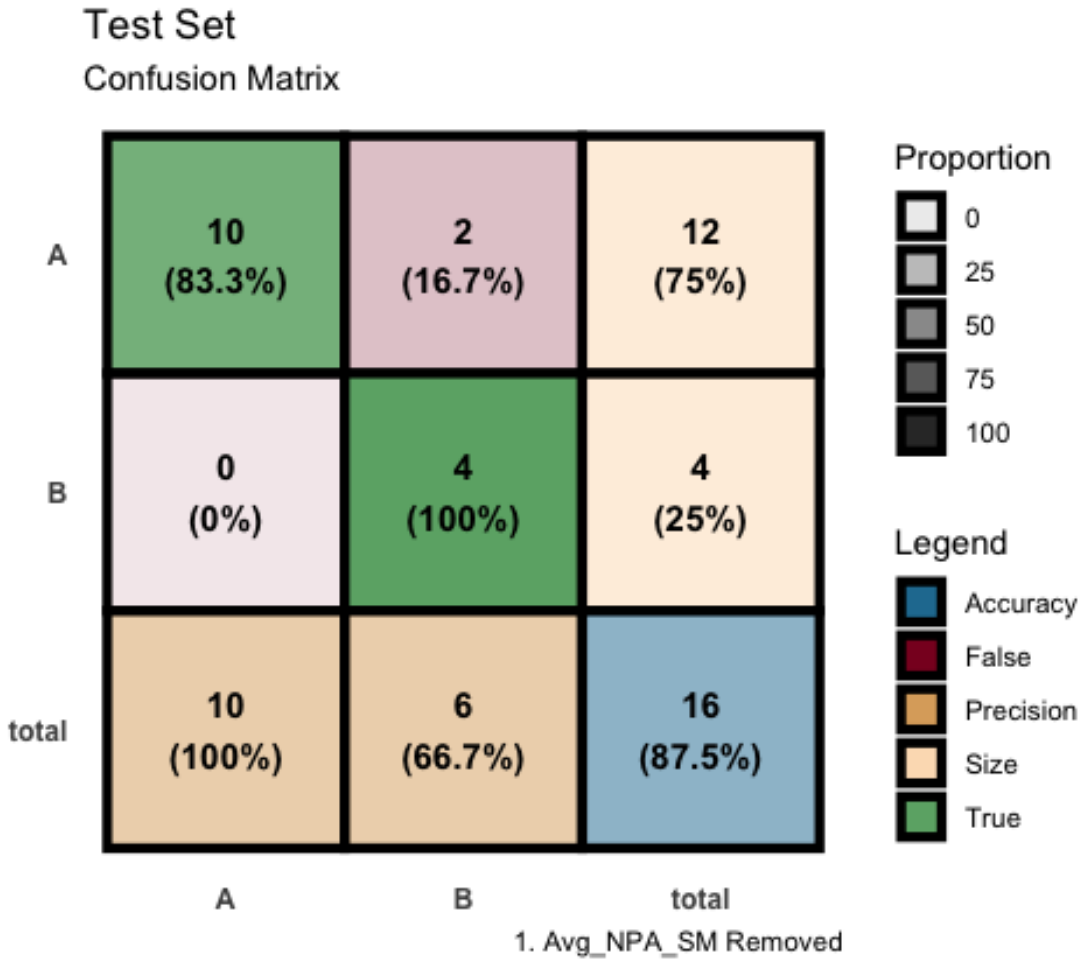
Training Set

Probability Heatmap

	A	B	Exp
A	100	0	A
AF	80	20	A
F	100	0	A
AB	50	50	A
AC	91	9	A
AD	100	0	A
AE	100	0	A
AH	80	20	A
AI	100	0	A
AJ	100	0	A
AK	98	2	A
AL	87	13	A
AM	98	2	A
G	50	50	A
I	50	50	A
J	100	0	A
K	87	13	A
L	50	50	A
M	91	9	A
N	100	0	A
O	91	9	A
P	100	0	A
Q	100	0	A
R	100	0	A
T	100	0	A
U	100	0	A
V	91	9	A
W	98	2	A
X	100	0	A
Y	98	2	A
Z	98	2	A
AN	50	50	A
AO	50	50	A
AS	100	0	A
AV	50	50	B
AW	50	50	B
AX	80	20	B
AY	13	87	B
BB	9	91	B
BE	0	100	B
BF	0	100	B
BG	13	87	B
BH	50	50	B
BI	50	50	B
BJ	9	91	B
BK	91	9	B

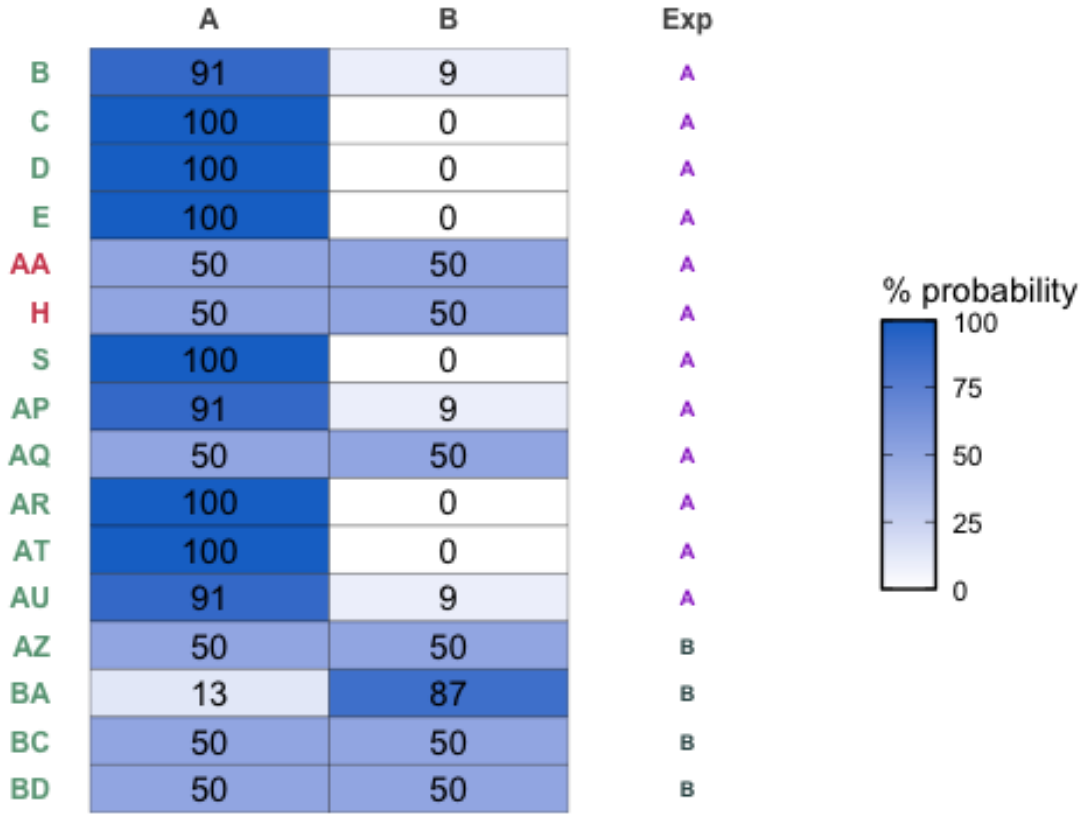


1. Avg_NPA_SM Removed


Test Set
Confusion Matrix
10
(83.3%)
2
(16.7%)
12
(75%)
0
(0%)
4
(100%)
4
(25%)
10
(100%)
6
(66.7%)
16
(87.5%)
A
B
total
A
B
total
1. Avg_NPA_SM Removed
Proportion
0
25
50
75
100
Legend
Accuracy
False
Precision
Size
True

Test Set

Probability Heatmap



1. Avg_NPA_SM Removed

2. fr_Ar_N Removed

formula

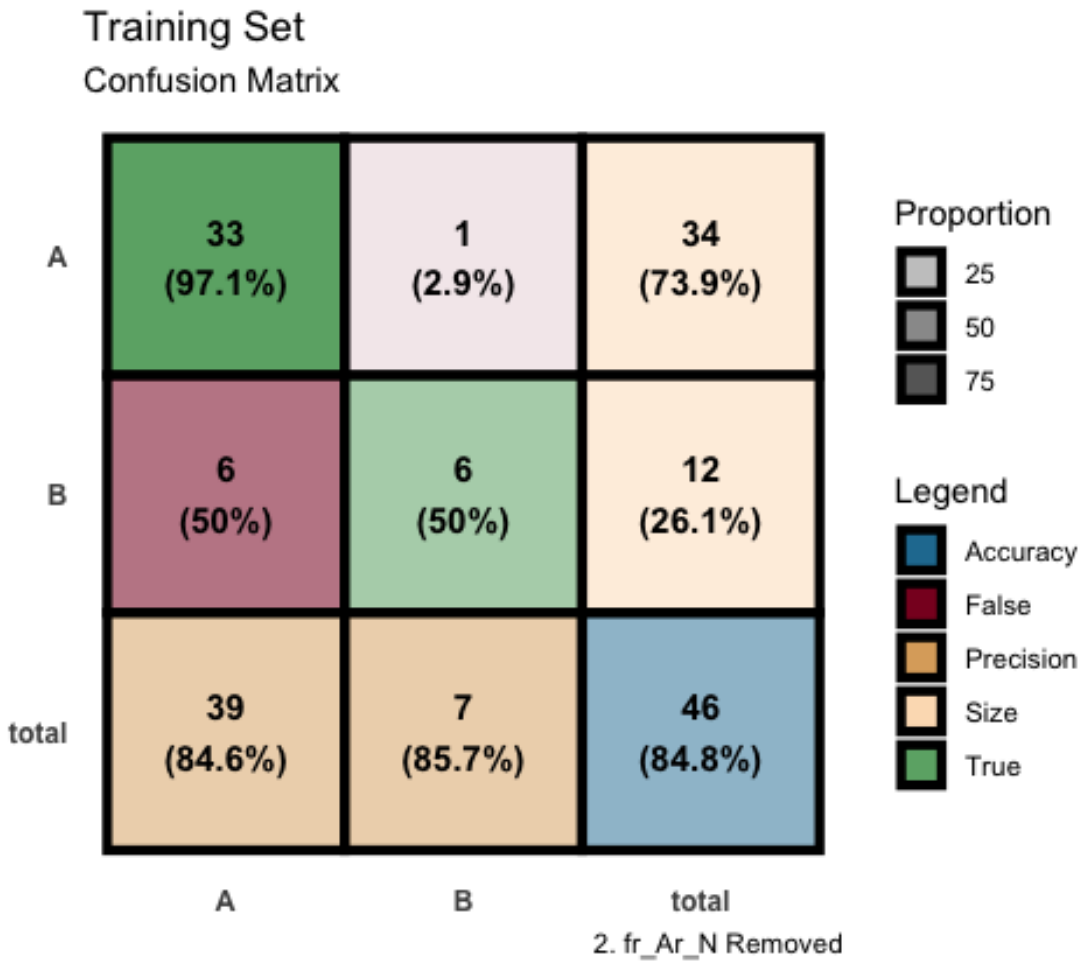
`class ~ Avg_NPA_SM + fr_aryl_methyl + fr_benzene + fr_halogen`

Full Model Stats - Overall Accuracy and Pseudo-R2

Accuracy	McFadden_R2
84.78%	0.328

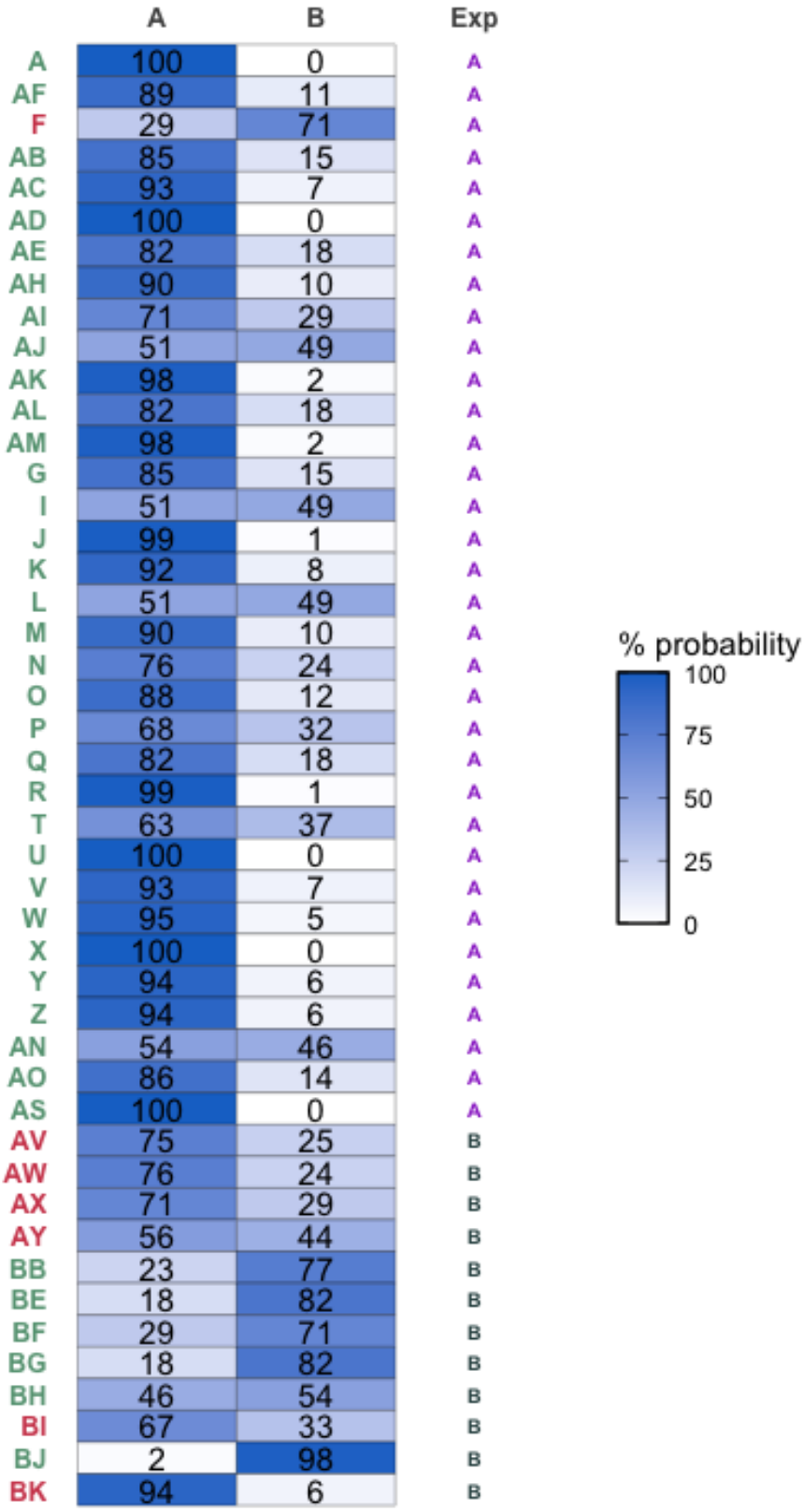
Model Coefficients

	x
(Intercept)	0.9188853
Avg_NPA_SM	18.4857033
fr_aryl_methyl1	-2.7592175
fr_aryl_methyl2	-12.9215911
fr_benzene1	-0.8918341
fr_benzene2	-32.6158024
fr_halogen1	2.0595415
fr_halogen2	3.2075153
fr_halogen3	1.5695469
fr_halogen4	-8.4512972
fr_halogen5	2.4217225
fr_halogen6	-4.8502196

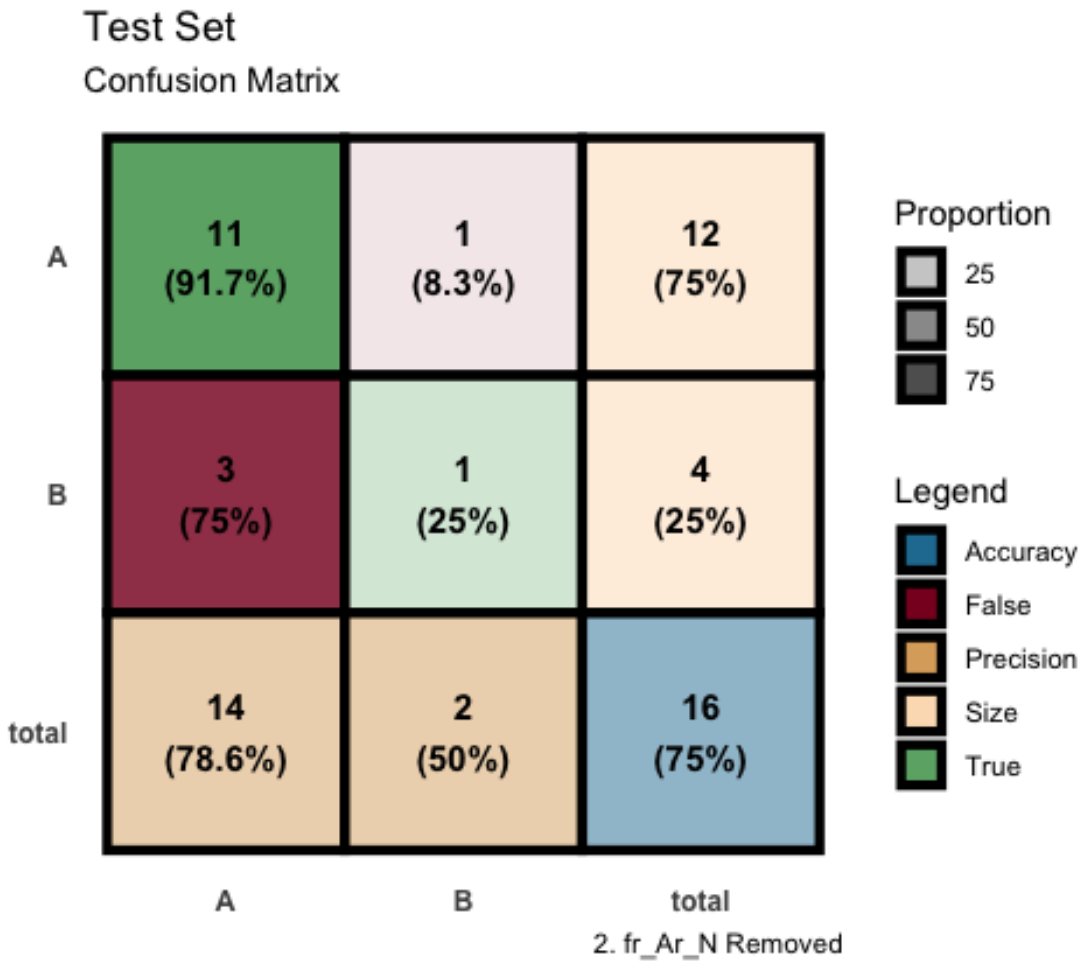

Training Set
Confusion Matrix
A
B
total
33
(97.1%)
1
(2.9%)
34
(73.9%)
6
(50%)
6
(50%)
12
(26.1%)
39
(84.6%)
7
(85.7%)
46
(84.8%)
A
B
total
2. fr_Ar_N Removed
Proportion
25
50
75
Legend
Accuracy
False
Precision
Size
True

Training Set

Probability Heatmap

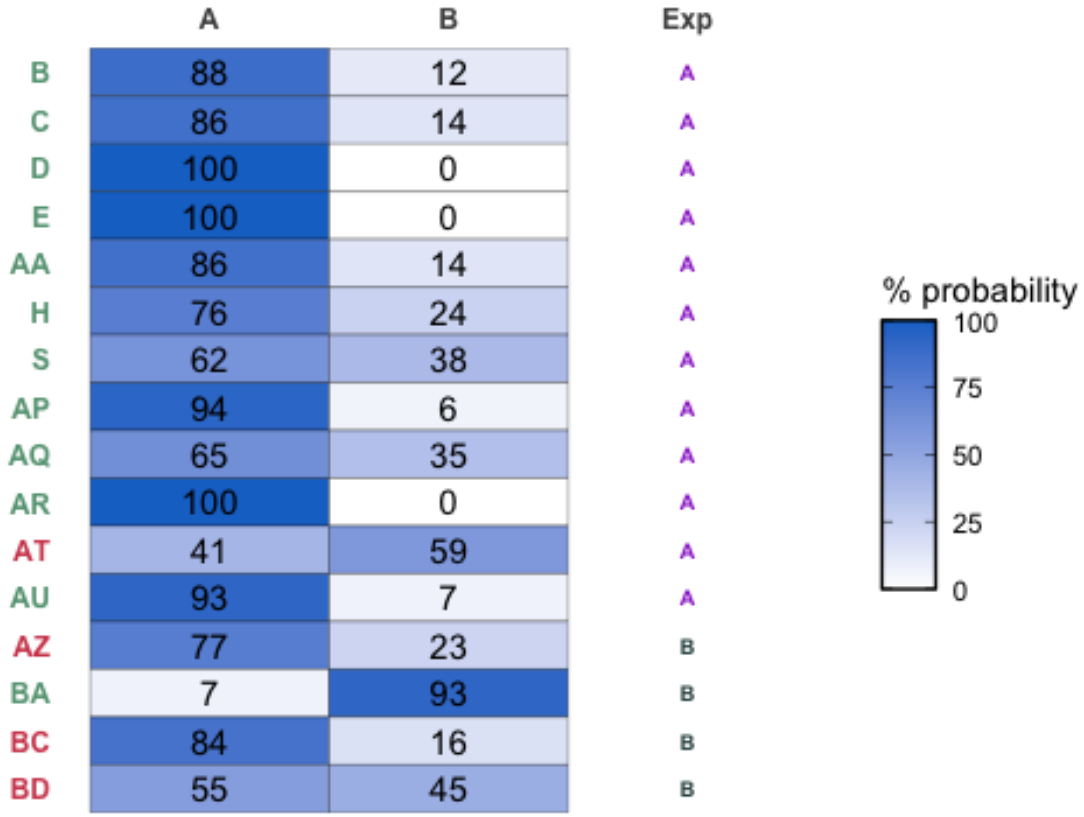


2. fr_Ar_N Removed


Test Set
Confusion Matrix
11
(91.7%)
1
(8.3%)
12
(75%)
3
(75%)
1
(25%)
4
(25%)
14
(78.6%)
2
(50%)
16
(75%)
A
B
total
A
B
total
2. fr_Ar_N Removed
Proportion
25
50
75
Legend
Accuracy
False
Precision
Size
True

Test Set

Probability Heatmap



2. fr_Ar_N Removed

3. fr_aryl_methyl Removed

formula

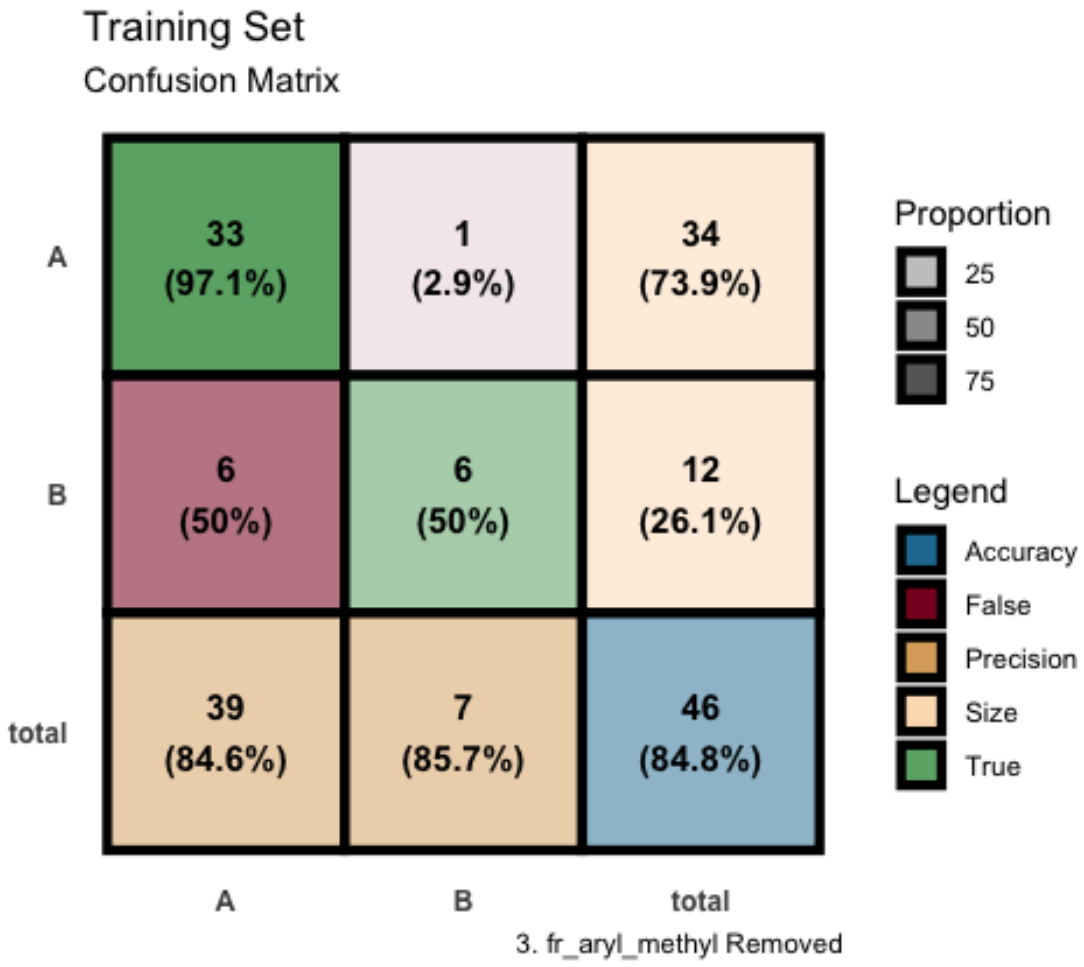
`class ~ Avg_NPA_SM + fr_Ar_N + fr_benzene + fr_halogen`

Full Model Stats - Overall Accuracy and Pseudo-R2

Accuracy	McFadden_R2
84.78%	0.36

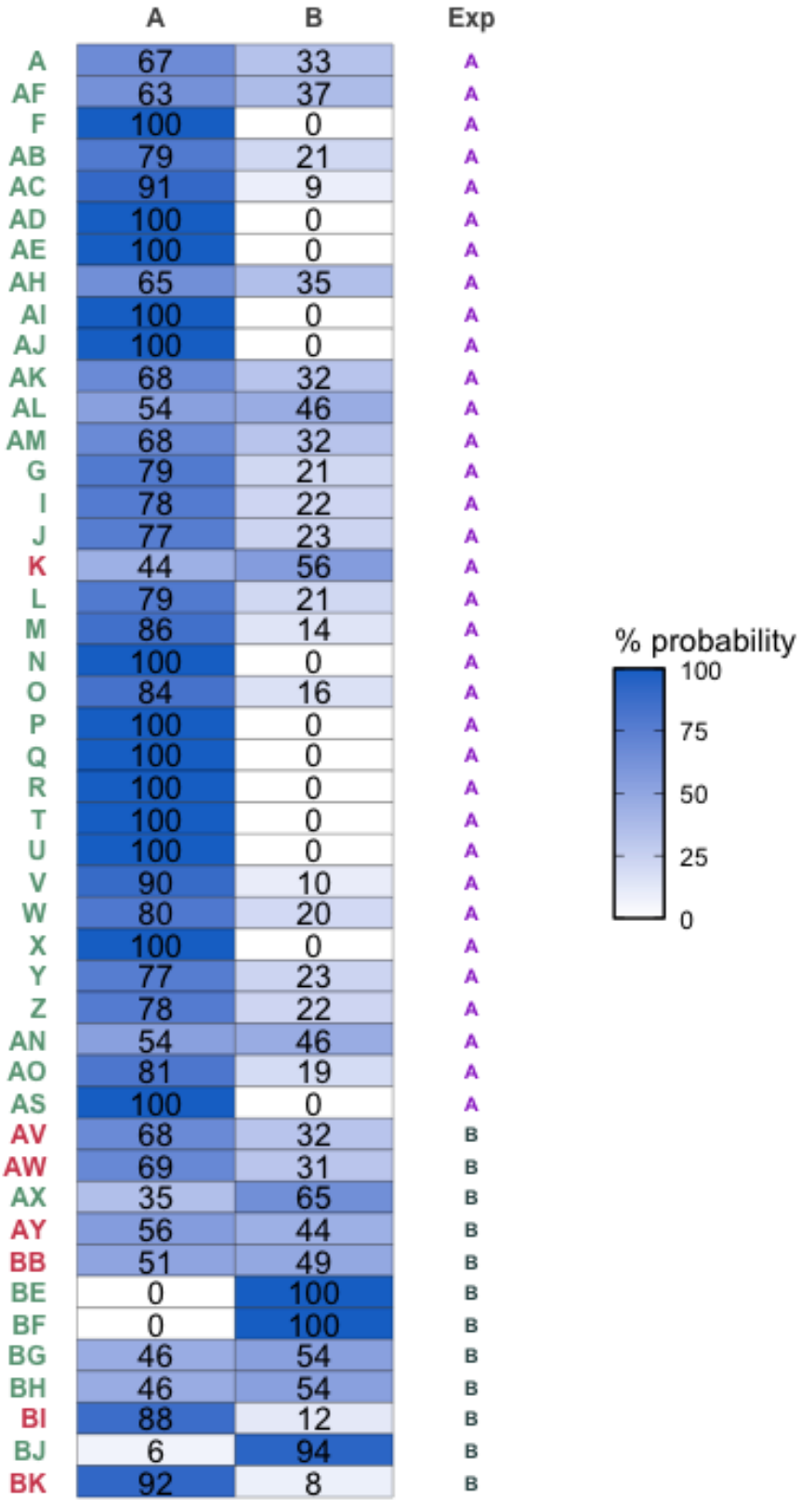
Model Coefficients

	x
(Intercept)	18.3982999
Avg_NPA_SM	18.5844081
fr_Ar_N1	-17.1017267
fr_Ar_N2	-29.0614501
fr_Ar_N3	-32.9558398
fr_benzene1	-0.9195207
fr_benzene2	-27.8437909
fr_halogen1	0.4496324
fr_halogen2	1.4880482
fr_halogen3	1.1965587
fr_halogen4	-1.7308296
fr_halogen5	2.0846872
fr_halogen6	-9.0103971

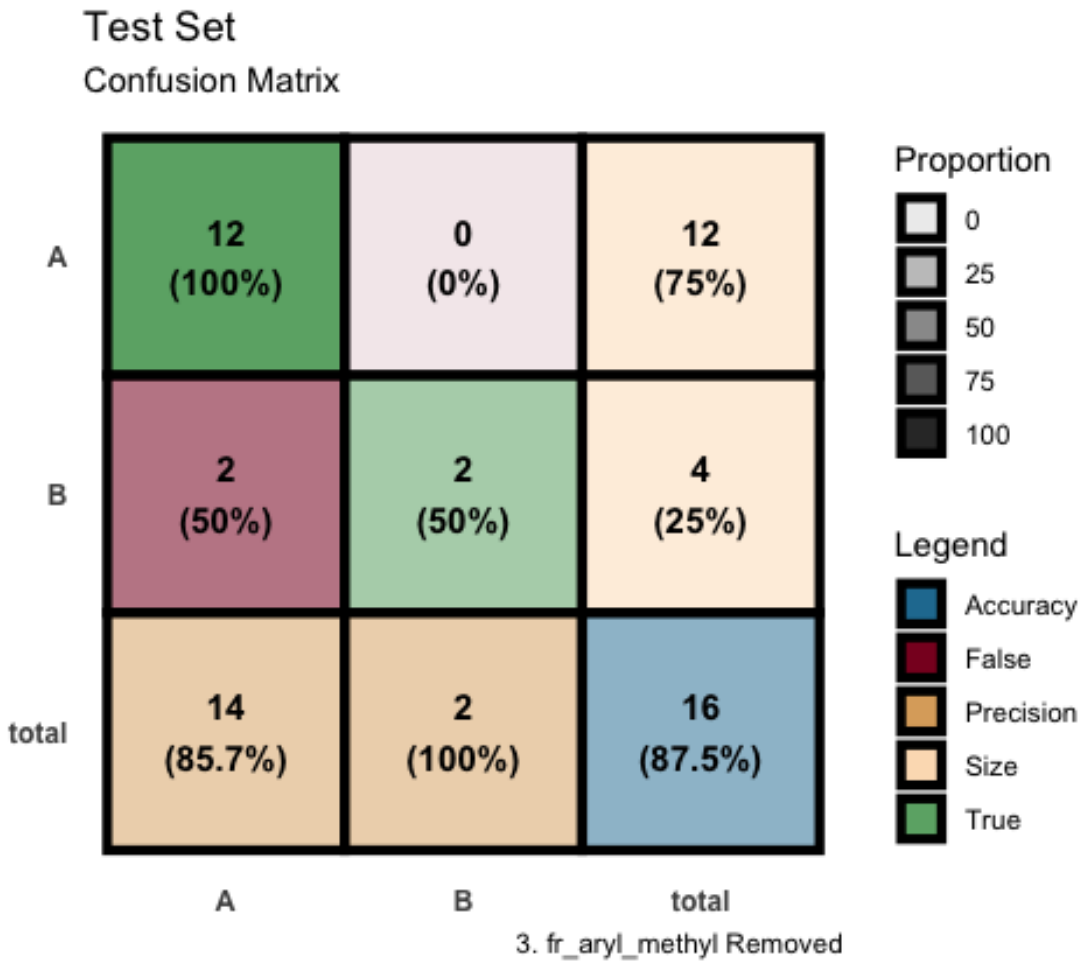

Training Set
Confusion Matrix
33
(97.1%)
1
(2.9%)
34
(73.9%)
6
(50%)
6
(50%)
12
(26.1%)
39
(84.6%)
7
(85.7%)
46
(84.8%)
A
B
total
A
B
total
3. fr_aryl_methyl Removed
Proportion
25
50
75
Legend
Accuracy
False
Precision
Size
True

Training Set

Probability Heatmap

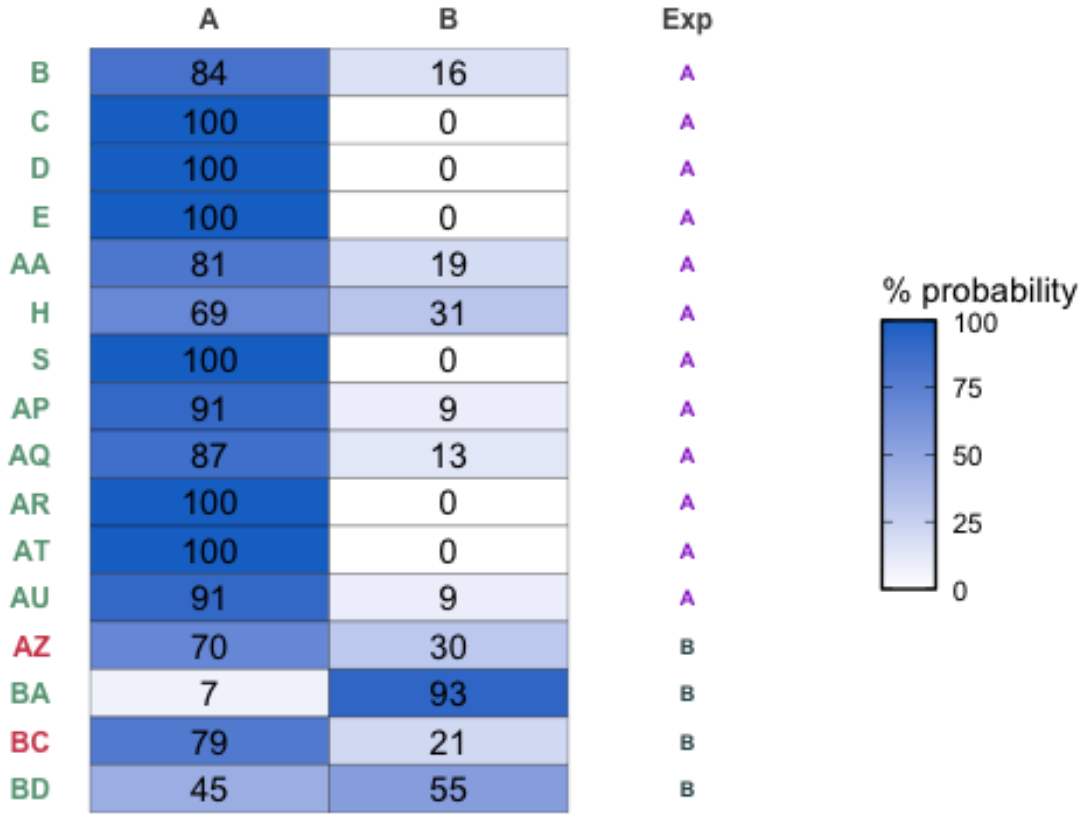


3. fr_aryl_methyl Removed


Test Set
Confusion Matrix
A
B
total
12
(100%)
0
(0%)
12
(75%)
2
(50%)
2
(50%)
4
(25%)
14
(85.7%)
2
(100%)
16
(87.5%)
A
B
total
3. fr_aryl_methyl Removed
Proportion
0
25
50
75
100
Legend
Accuracy
False
Precision
Size
True

Test Set

Probability Heatmap



3. fr_aryl_methyl Removed

4. fr_benzene Removed

formula

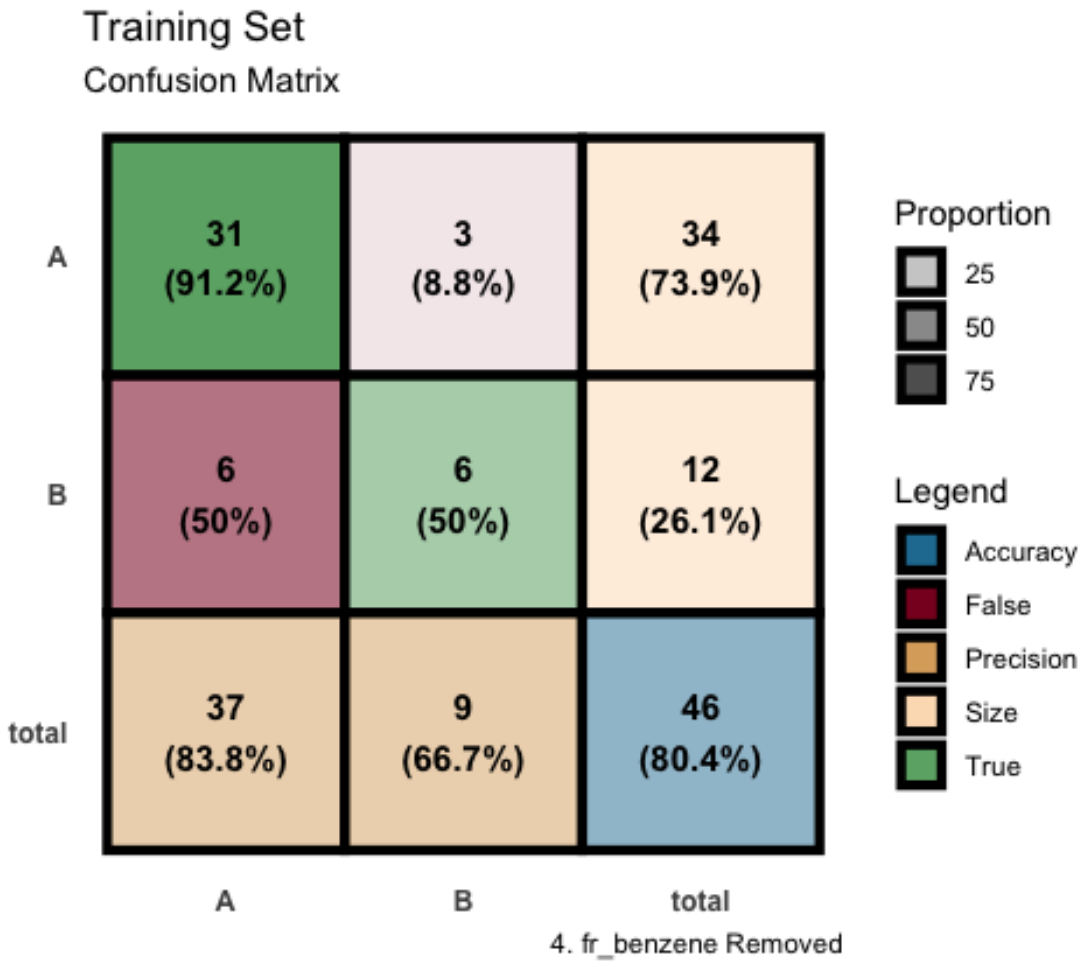
`class ~ Avg_NPA_SM + fr_Ar_N + fr_aryl_methyl + fr_halogen`

Full Model Stats - Overall Accuracy and Pseudo-R2

Accuracy	McFadden_R2
80.43%	0.404

Model Coefficients

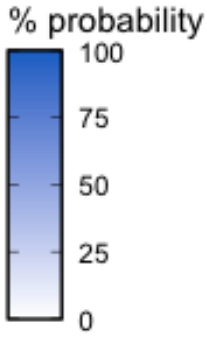
	x
(Intercept)	15.3364807
Avg_NPA_SM	19.3241422
fr_Ar_N1	-13.7390649
fr_Ar_N2	-26.0021302
fr_Ar_N3	-34.6453926
fr_aryl_methyl1	-2.5523230
fr_aryl_methyl2	-17.8132254
fr_halogen1	1.0049517
fr_halogen2	0.4974903
fr_halogen3	0.2314174
fr_halogen4	-2.3972048
fr_halogen5	0.9617090
fr_halogen6	-8.0304964


Training Set
Confusion Matrix
31
(91.2%)
3
(8.8%)
34
(73.9%)
6
(50%)
6
(50%)
12
(26.1%)
37
(83.8%)
9
(66.7%)
46
(80.4%)
A
B
total
A
B
total
4. fr_benzene Removed
Proportion
25
50
75
Legend
Accuracy
False
Precision
Size
True

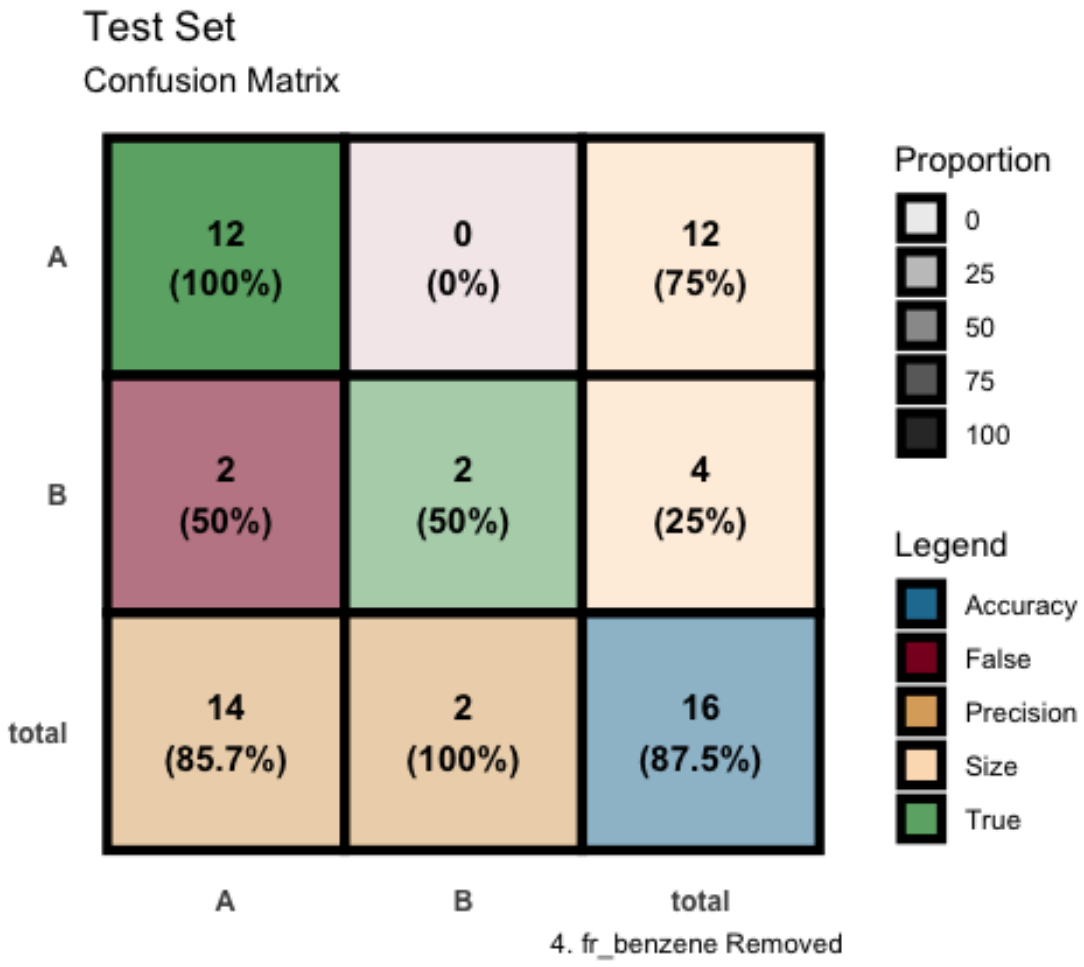
Training Set

Probability Heatmap

	A	B	Exp
A	100	0	A
AF	91	9	A
F	100	0	A
AB	76	24	A
AC	76	24	A
AD	31	69	A
AE	100	0	A
AH	92	8	A
AI	100	0	A
AJ	100	0	A
AK	96	4	A
AL	93	7	A
AM	96	4	A
G	76	24	A
I	40	60	A
J	93	7	A
K	95	5	A
L	41	59	A
M	66	34	A
N	100	0	A
O	63	37	A
P	100	0	A
Q	100	0	A
R	100	0	A
T	100	0	A
U	100	0	A
V	75	25	A
W	90	10	A
X	100	0	A
Y	89	11	A
Z	89	11	A
AN	55	45	A
AO	78	22	A
AS	65	35	A
AV	63	37	B
AW	64	36	B
AX	75	25	B
AY	74	26	B
BB	32	68	B
BE	0	100	B
BF	0	100	B
BG	42	58	B
BH	46	54	B
BI	58	42	B
BJ	2	98	B
BK	78	22	B

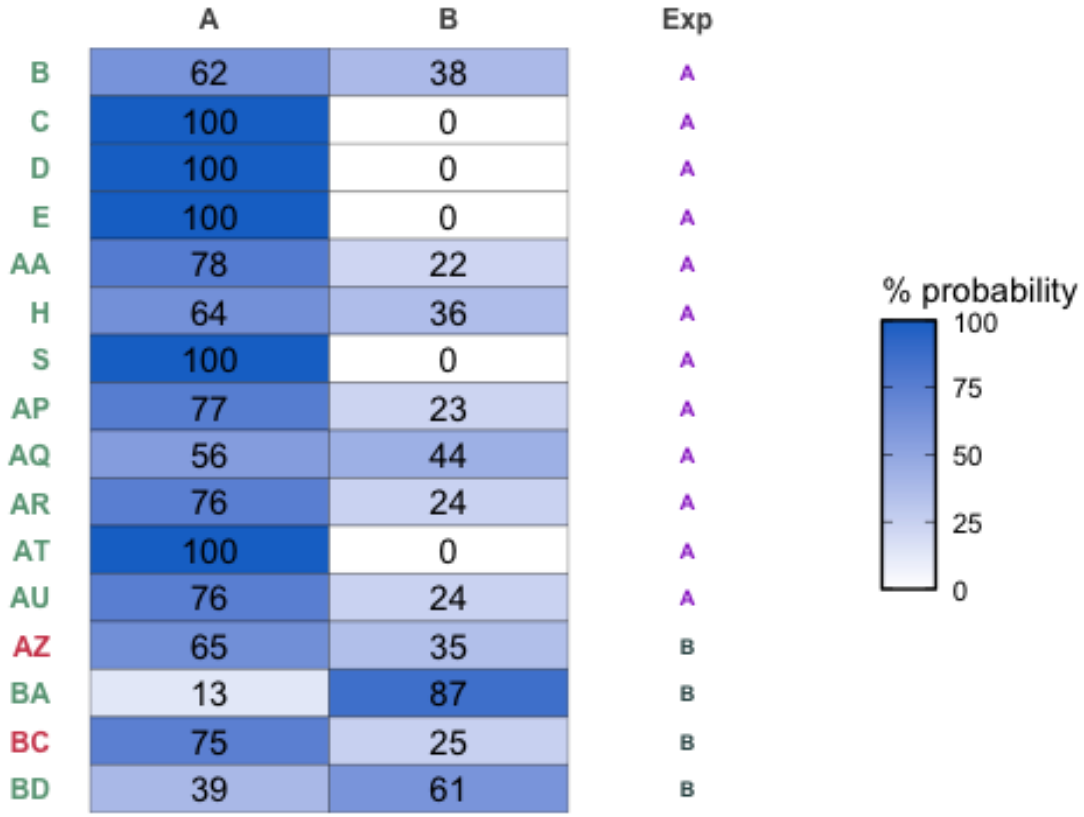


4. fr_benzene Removed


Test Set
Confusion Matrix
A
B
total
12
(100%)
0
(0%)
12
(75%)
2
(50%)
2
(50%)
4
(25%)
14
(85.7%)
2
(100%)
16
(87.5%)
A
B
total
4. fr_benzene Removed
Proportion
0
25
50
75
100
Legend
Accuracy
False
Precision
Size
True

Test Set

Probability Heatmap



4. fr_benzene Removed

5. fr_halogen Removed

formula

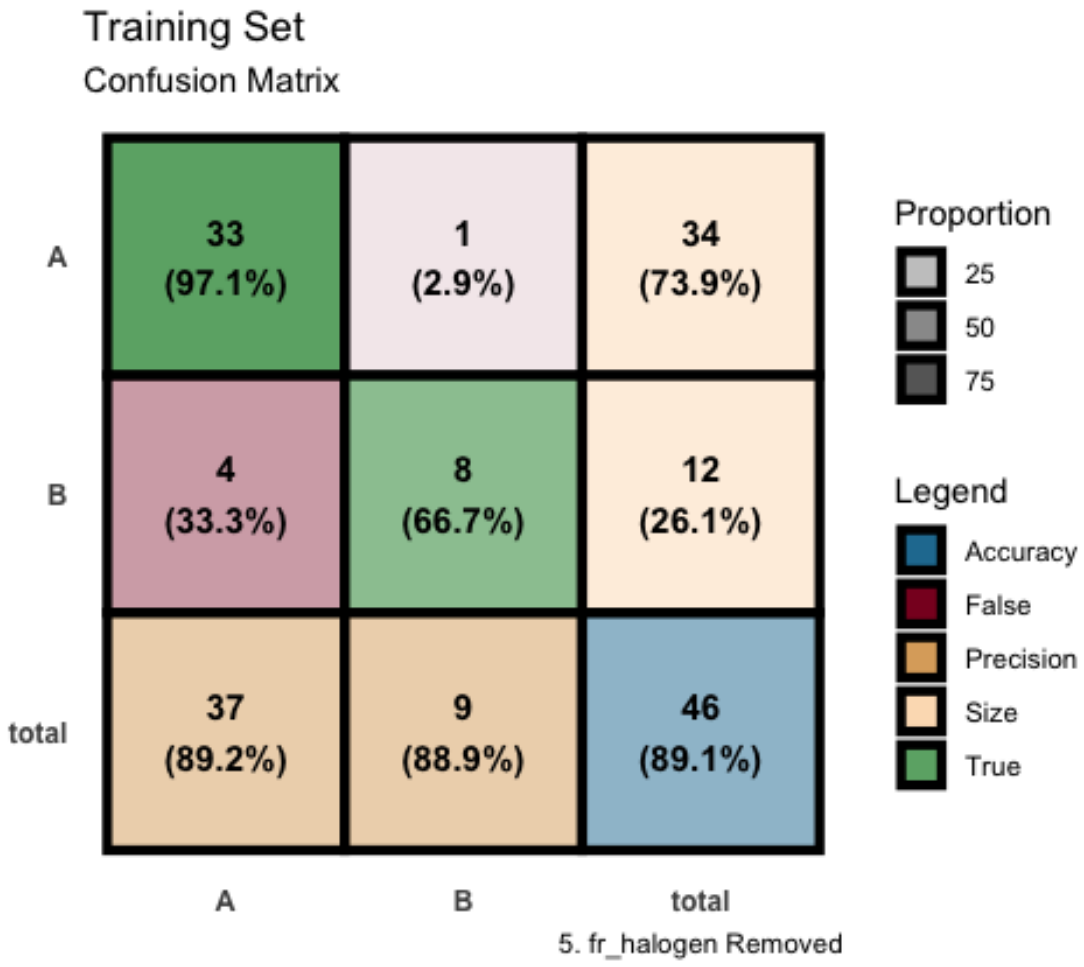
`class ~ Avg_NPA_SM + fr_Ar_N + fr_aryl_methyl + fr_benzene`

Full Model Stats - Overall Accuracy and Pseudo-R2

Accuracy	McFadden_R2
89.13%	0.512

Model Coefficients

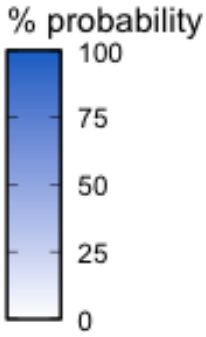
	x
(Intercept)	43.957001
Avg_NPA_SM	49.387661
fr_Ar_N1	-37.097092
fr_Ar_N2	-69.144375
fr_Ar_N3	-69.553894
fr_aryl_methyl1	-3.779621
fr_aryl_methyl2	-51.951799
fr_benzene1	-1.330087
fr_benzene2	-68.569342


Training Set
Confusion Matrix
A
B
total
33
(97.1%)
1
(2.9%)
34
(73.9%)
4
(33.3%)
8
(66.7%)
12
(26.1%)
37
(89.2%)
9
(88.9%)
46
(89.1%)
A
B
total
5. fr_halogen Removed
Proportion
25
50
75
Legend
Accuracy
False
Precision
Size
True

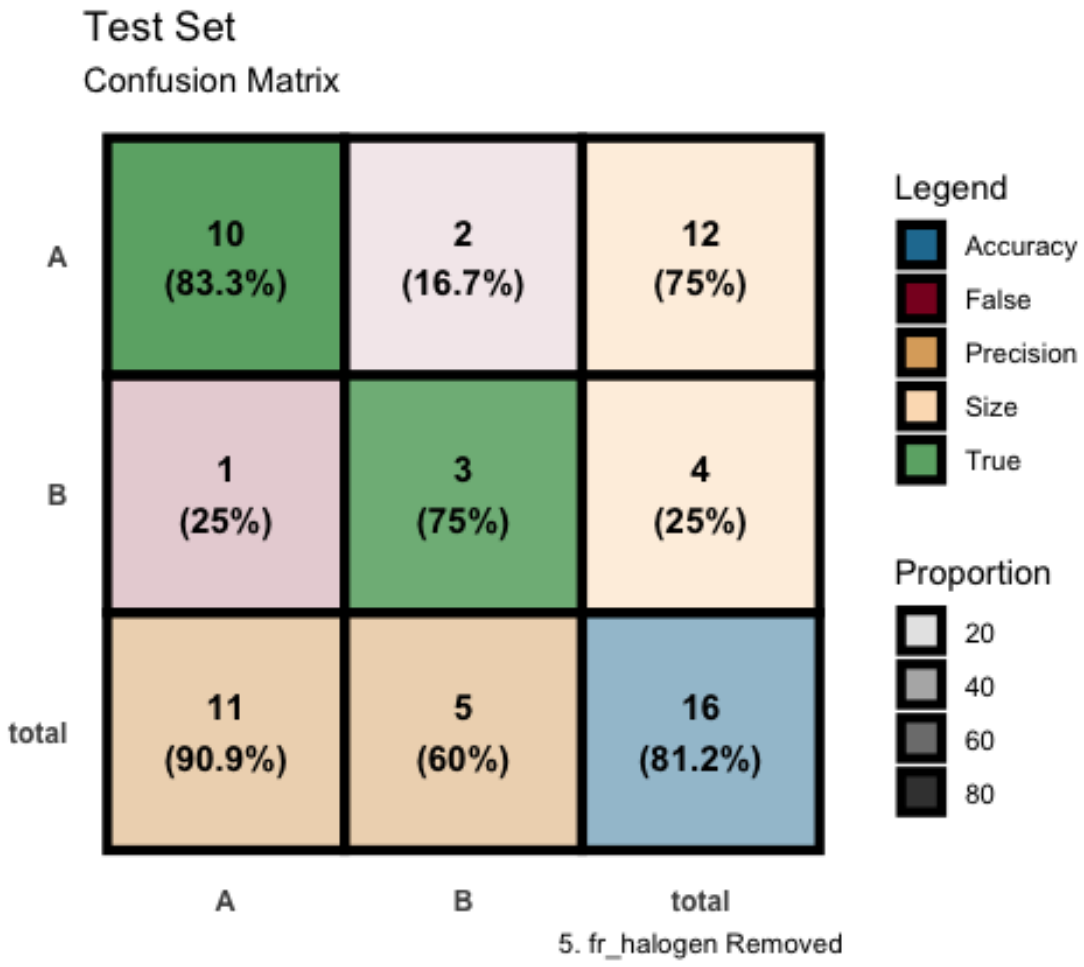
Training Set

Probability Heatmap

	A	B	Exp
A	100	0	A
AF	95	5	A
F	100	0	A
AB	54	46	A
AC	82	18	A
AD	100	0	A
AE	100	0	A
AH	96	4	A
AI	100	0	A
AJ	100	0	A
AK	92	8	A
AL	97	3	A
AM	92	8	A
G	54	46	A
I	53	47	A
J	93	7	A
K	95	5	A
L	55	45	A
M	57	43	A
N	100	0	A
O	47	53	A
P	100	0	A
Q	100	0	A
R	100	0	A
T	100	0	A
U	99	1	A
V	79	21	A
W	98	2	A
X	100	0	A
Y	98	2	A
Z	98	2	A
AN	81	19	A
AO	62	38	A
AS	100	0	A
AV	19	81	B
AW	22	78	B
AX	48	52	B
AY	61	39	B
BB	10	90	B
BE	0	100	B
BF	0	100	B
BG	27	73	B
BH	63	37	B
BI	87	13	B
BJ	0	100	B
BK	86	14	B

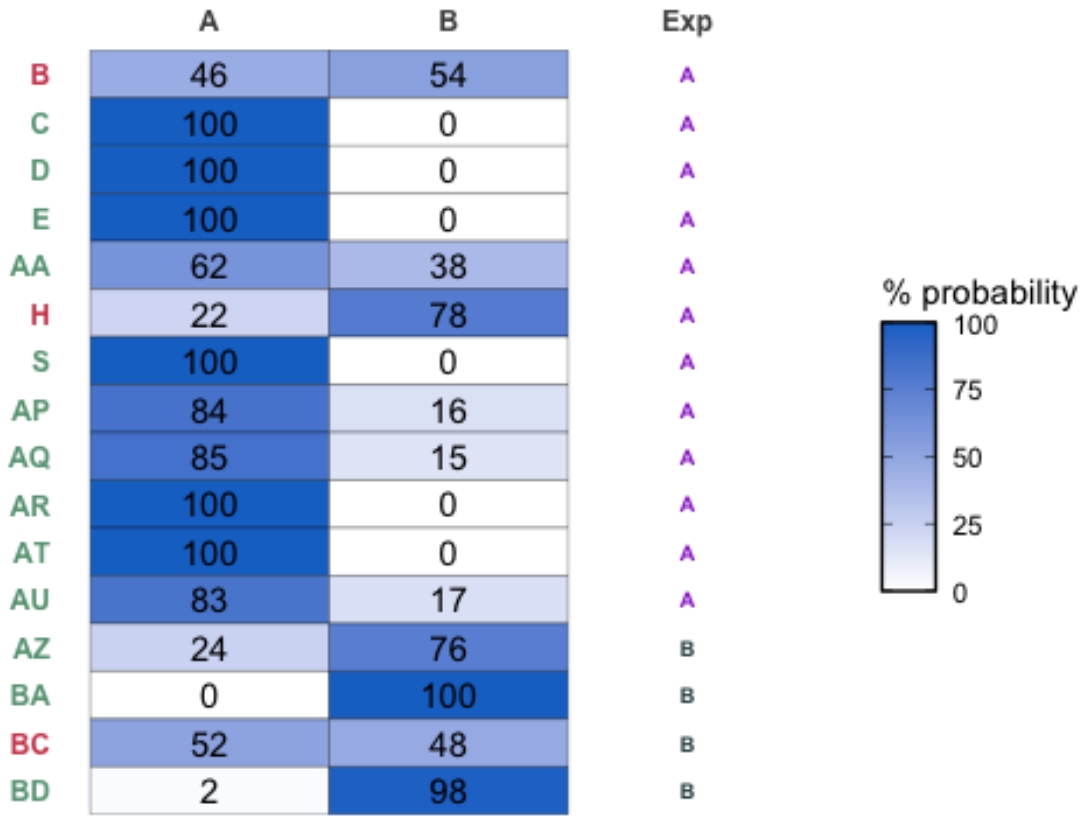


5. fr_halogen Removed


Test Set
Confusion Matrix
A
B
total
10
(83.3%)
2
(16.7%)
12
(75%)
1
(25%)
3
(75%)
4
(25%)
11
(90.9%)
5
(60%)
16
(81.2%)
A
B
total
5. fr_halogen Removed
Legend
Accuracy
False
Precision
Size
True
Proportion
20
40
60
80

Test Set

Probability Heatmap



5. fr_halogen Removed

Section Results Summary

	Training Accuracy	Test Accuracy
5. 5th Place with 5 features	91.30	87.5
1. Avg_NPA_SM Removed	84.78	87.5
2. fr_Ar_N Removed	84.78	75.0
3. fr_aryl_methyl Removed	84.78	87.5
4. fr_benzene Removed	80.43	87.5
5. fr_halogen Removed	89.13	81.2