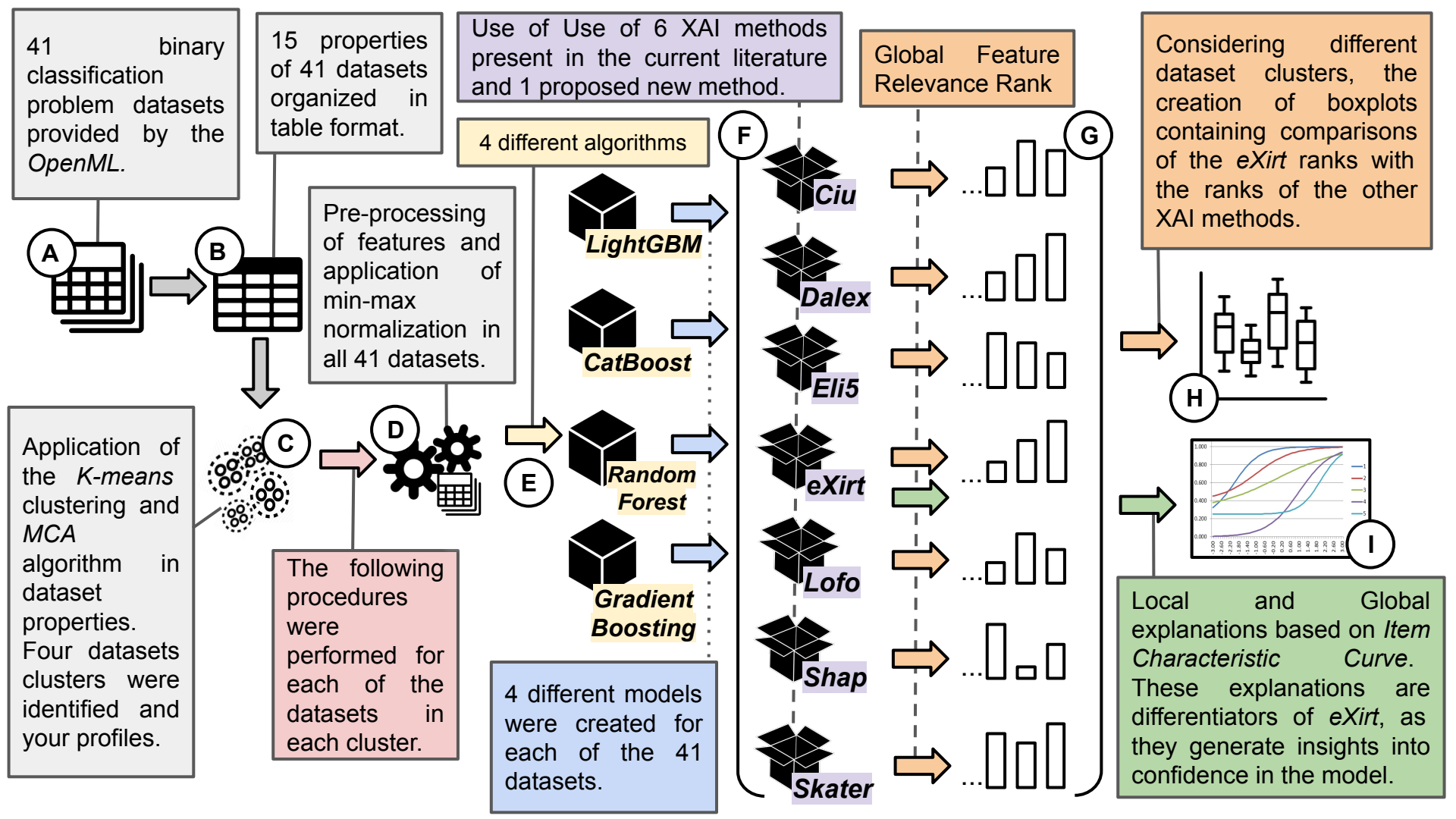
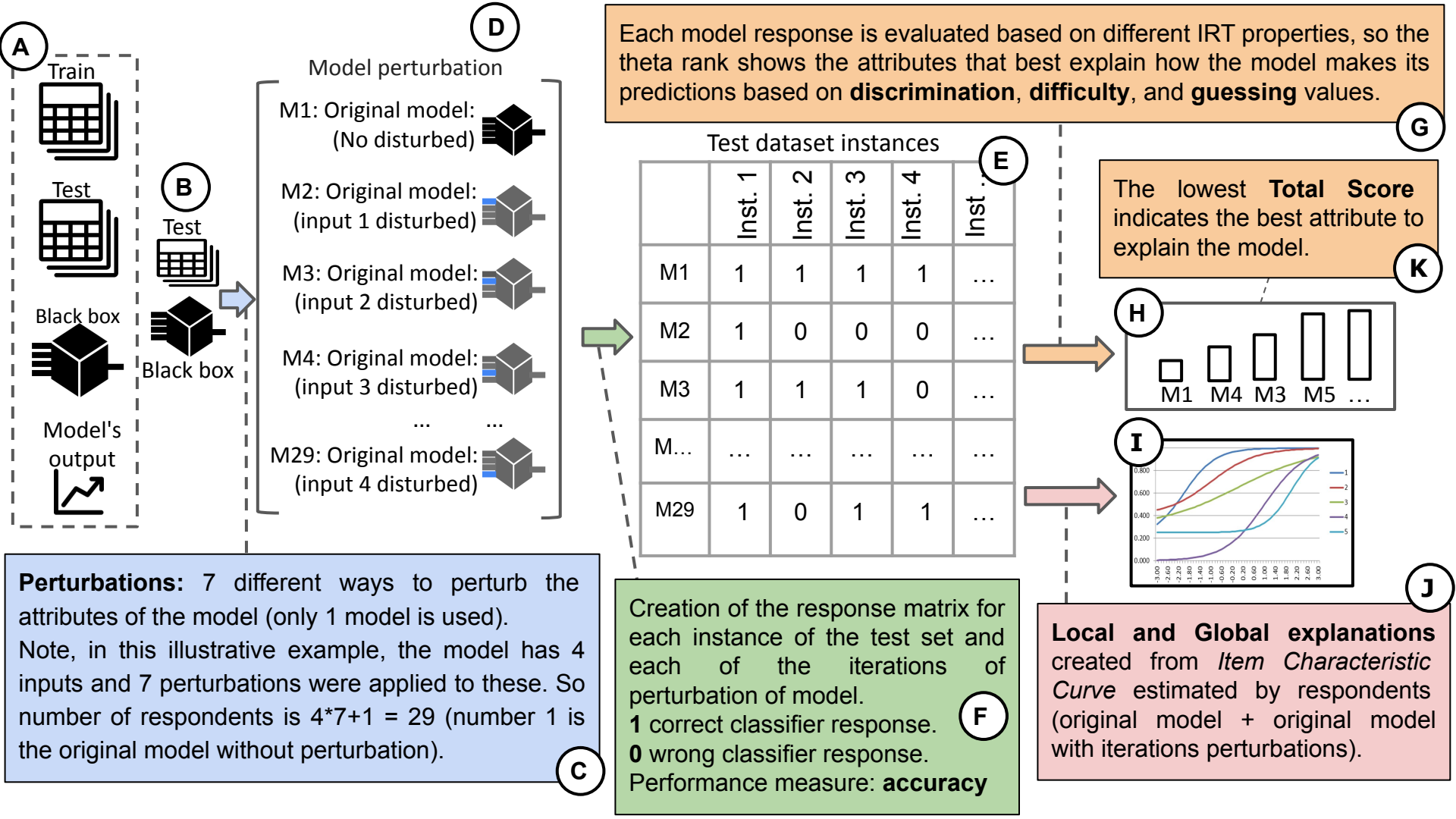
Supplementary Material Based on Illustrations

Paper: Explanations Based on Item Response Theory (*eXirt*): A Model-Specific Method to Explain Tree-Ensemble Model in Trust Perspective


41 binary classification problem datasets provided by the *OpenML*.
15 properties of 41 datasets organized in table format.
Use of Use of 6 XAI methods present in the current literature and 1 proposed new method.
Global Feature Relevance Rank
Considering different dataset clusters, the creation of boxplots containing comparisons of the *eXirt* ranks with the ranks of the other XAI methods.
A
B
Pre-processing of features and application of min-max normalization in all 41 datasets.
4 different algorithms
F
G
LightGBM
CatBoost
Random Forest
Gradient Boosting
Ciu
Dalex
Eli5
eXirt
Lofo
Shap
Skater
H
C
D
E
Application of the *K-means* clustering and *MCA* algorithm in dataset properties. Four datasets clusters were identified and your profiles.
The following procedures were performed for each of the datasets in each cluster.
4 different models were created for each of the 41 datasets.
I
Local and Global explanations based on *Item Characteristic Curve*. These explanations are differentiators of *eXirt*, as they generate insights into confidence in the model.


A
Train
Test
Black box
Model's output
B
Test
Black box
D
Model perturbation
M1: Original model: (No disturbed)
M2: Original model: (input 1 disturbed)
M3: Original model: (input 2 disturbed)
M4: Original model: (input 3 disturbed)
...
M29: Original model: (input 4 disturbed)
Each model response is evaluated based on different IRT properties, so the theta rank shows the attributes that best explain how the model makes its predictions based on **discrimination**, **difficulty**, and **guessing** values.
G
Test dataset instances
E
Inst. 1
Inst. 2
Inst. 3
Inst. 4
Inst. ..
M1 1 1 1 1 …
M2 1 0 0 0 …
M3 1 1 1 0 …
M… … … … … …
M29 1 0 1 1 …
The lowest **Total Score** indicates the best attribute to explain the model.
K
H
M1 M4 M3 M5 …
I
J
Perturbations: 7 different ways to perturb the attributes of the model (only 1 model is used).
Note, in this illustrative example, the model has 4 inputs and 7 perturbations were applied to these. So number of respondents is 4*7+1 = 29 (number 1 is the original model without perturbation).
C
Creation of the response matrix for each instance of the test set and each of the iterations of perturbation of model.
1 correct classifier response.
0 wrong classifier response.
Performance measure: **accuracy**
F
Local and Global explanations created from *Item Characteristic Curve* estimated by respondents (original model + original model with iterations perturbations).

Relation of terms IRT and ML

IRT: Questions
ML: Dataset instances

	Instance 1	Instance 2	Instance 3	Instance …	Instance n
Model 1	1	1	1	…	0
Model 2	0	1	1	…	1
Model 3	0	1	0	..	1
Model 4	1	1	0	..	0
Model 5	1	0	0	…	1
Model 6	1	0	0	…	0

IRT: Individuals
ML: Models

IRT: Responses from individuals
ML: Models responses

IRT: Estimation of parameter item

IRT: Item parameters

	Discrimination	Difficulty	Guessing
Inst. 1	0.9	1.3	0.0
Inst. 2	0.5	1.4	0.0
Inst. 3	0.4	1.1	0.2
…	…	...	…
Inst. n	0.3	0.1	0.4

IRT: Θ estimation

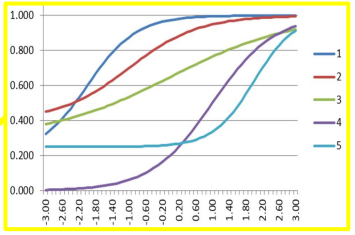
IRT: Individuals
ML: Models

IRT: Θ

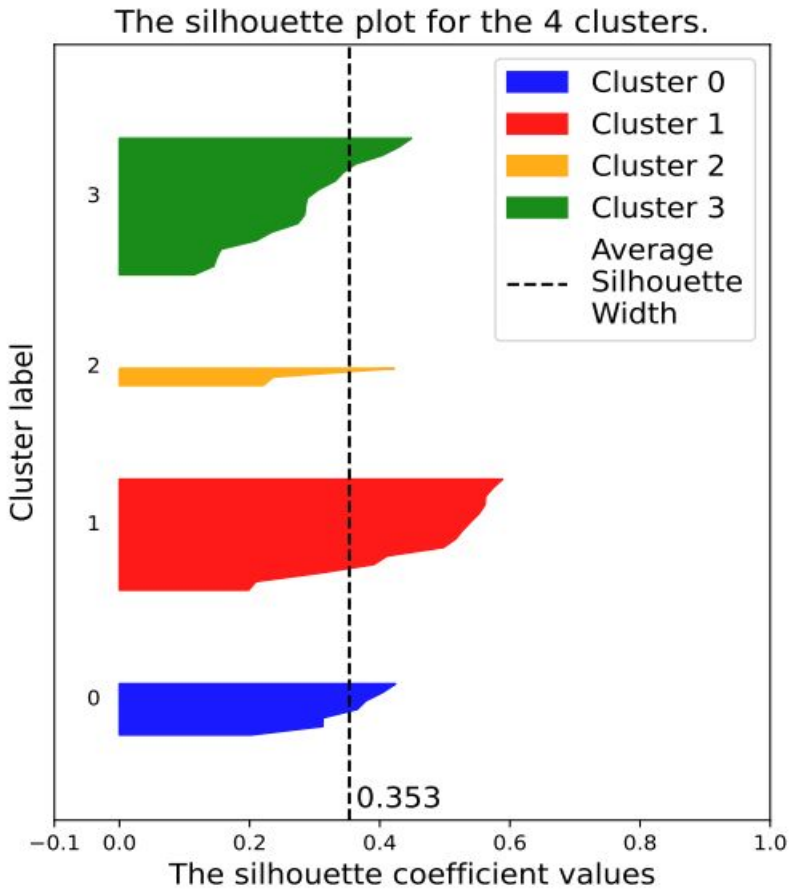
	Score
Model 2	21.5
Model 1	14.3
Model 6	13.4
Model 4	9.2
Model 5	4.4
Model 3	-1.1

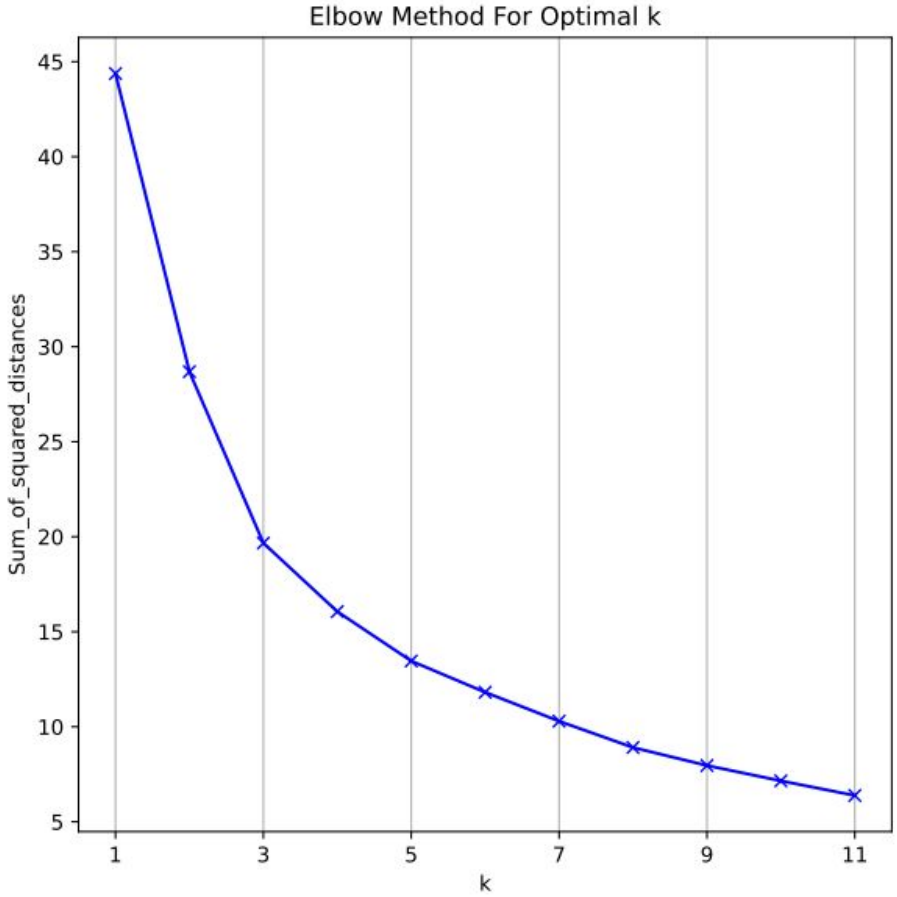
IRT: ICC plot

IRT: ICC result

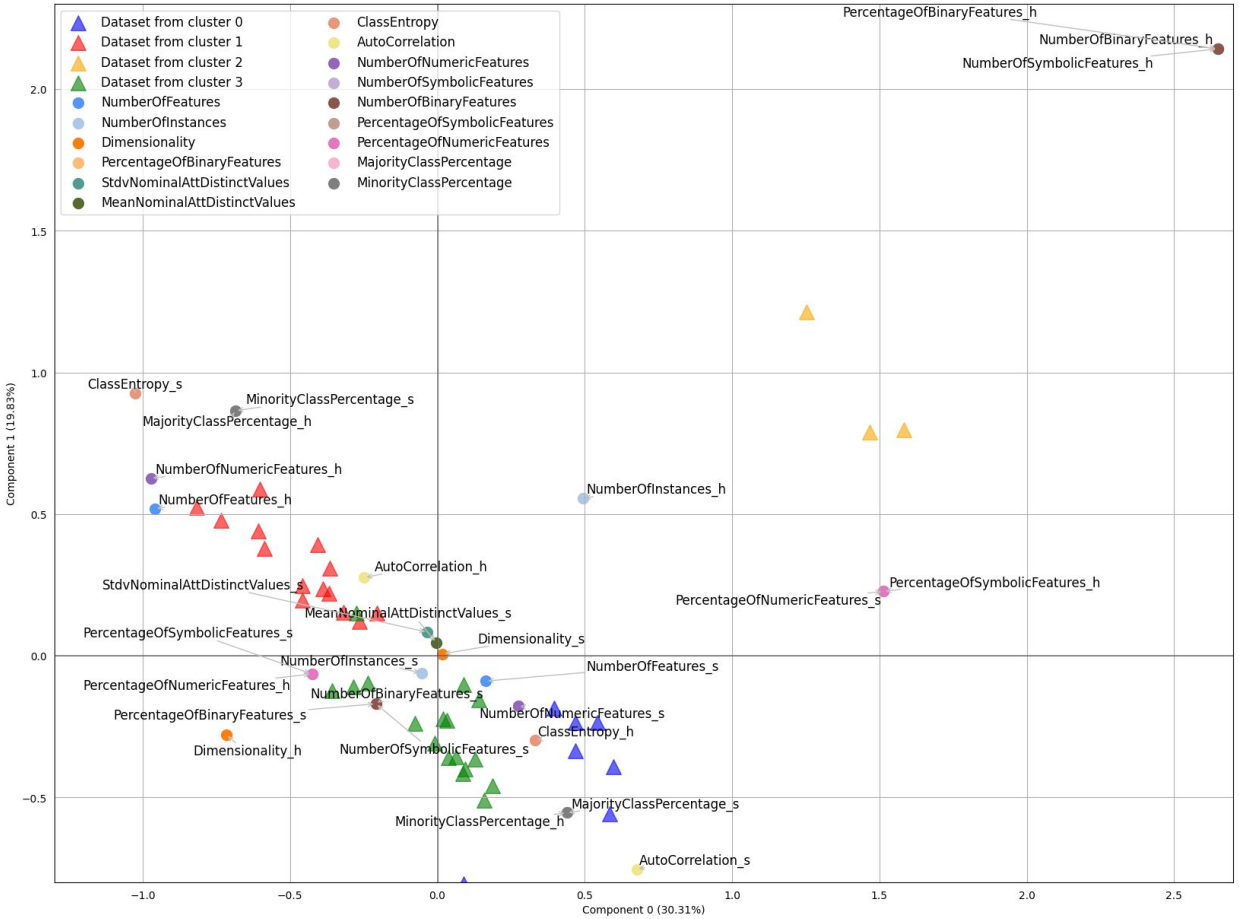


K-size identification of clustering using Silhouette and Elbow



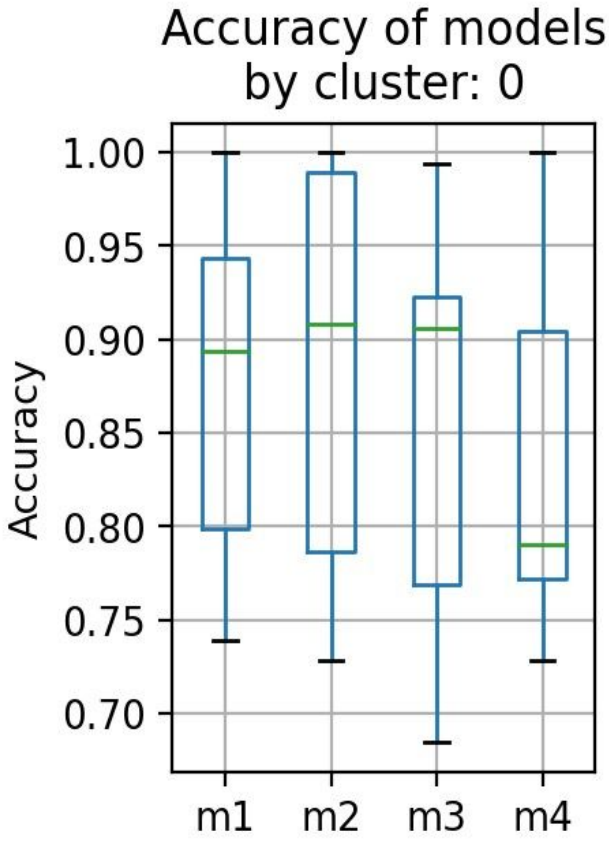


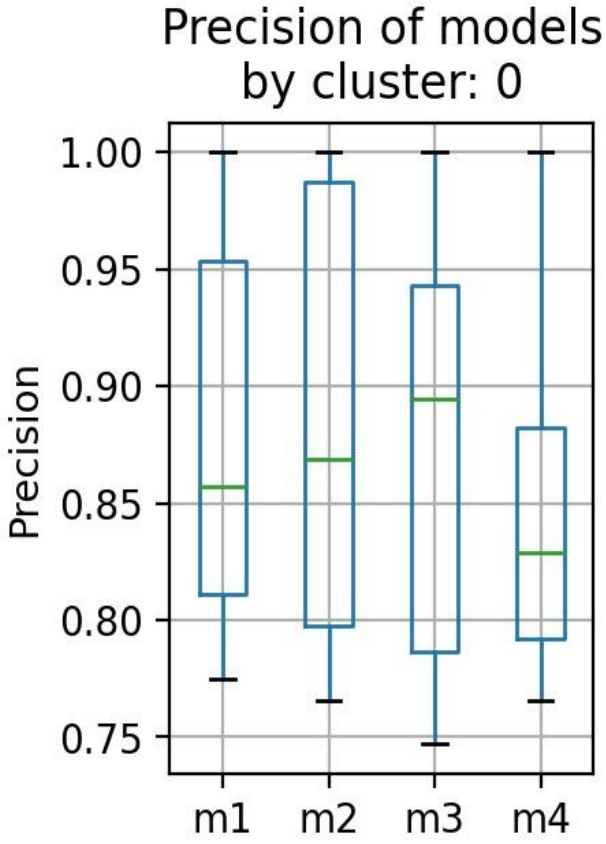
Profiles of datasets identified through Multiple Correspondence Analysis

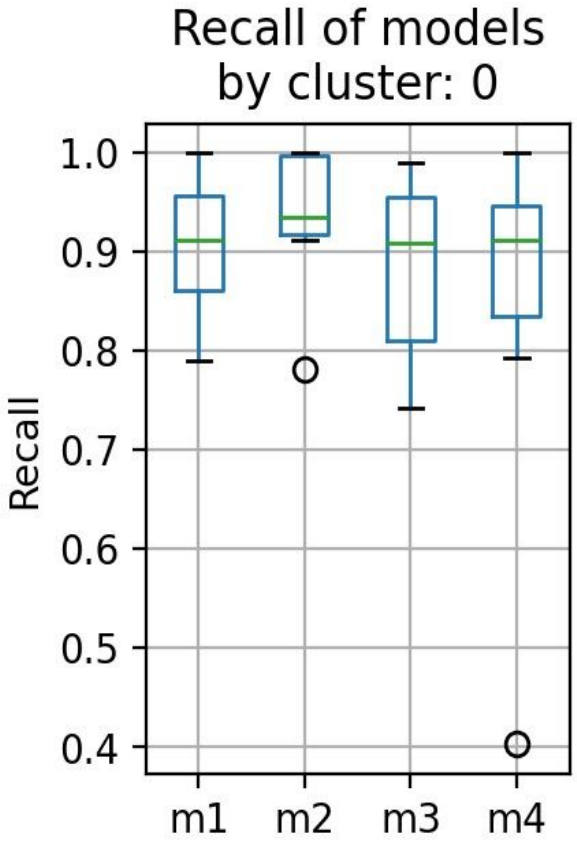


Cluster 0

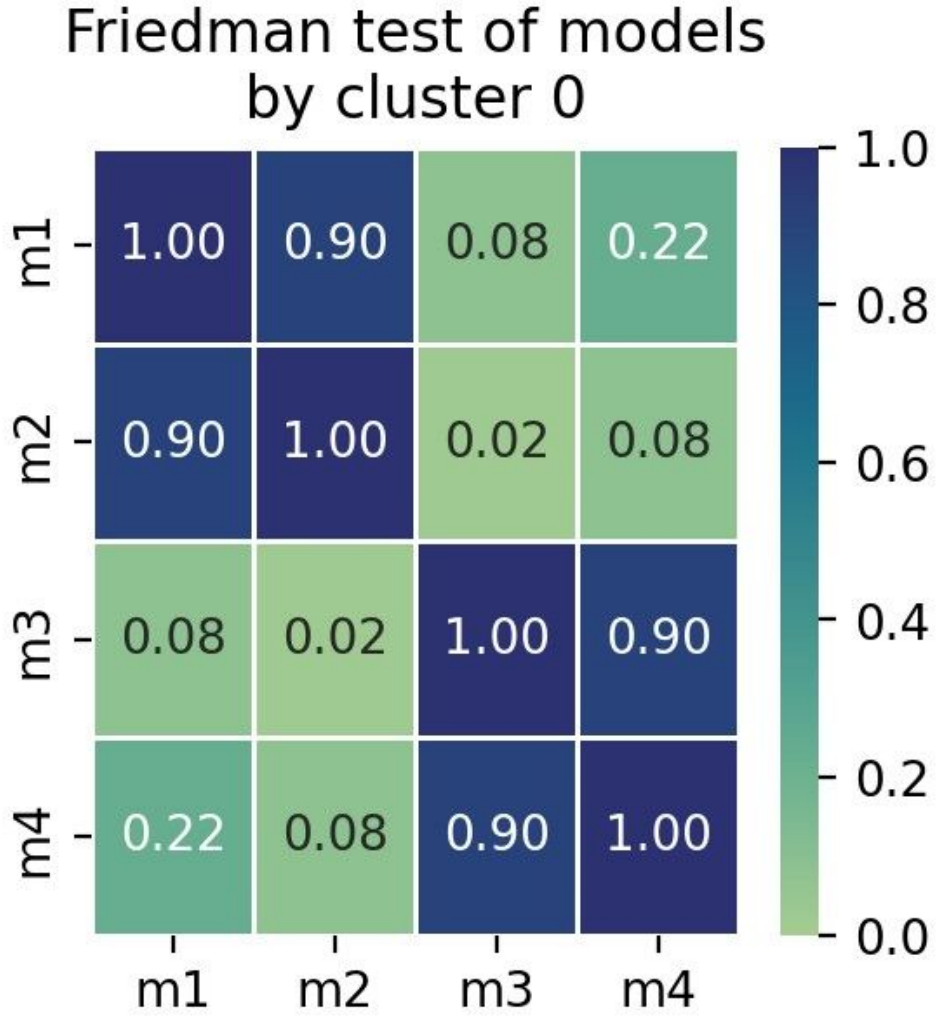
Data regarding the performance of models M1 to M4 based on cluster 0


Accuracy of models
by cluster: 0
Accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0.70
m1
m2
m3
m4

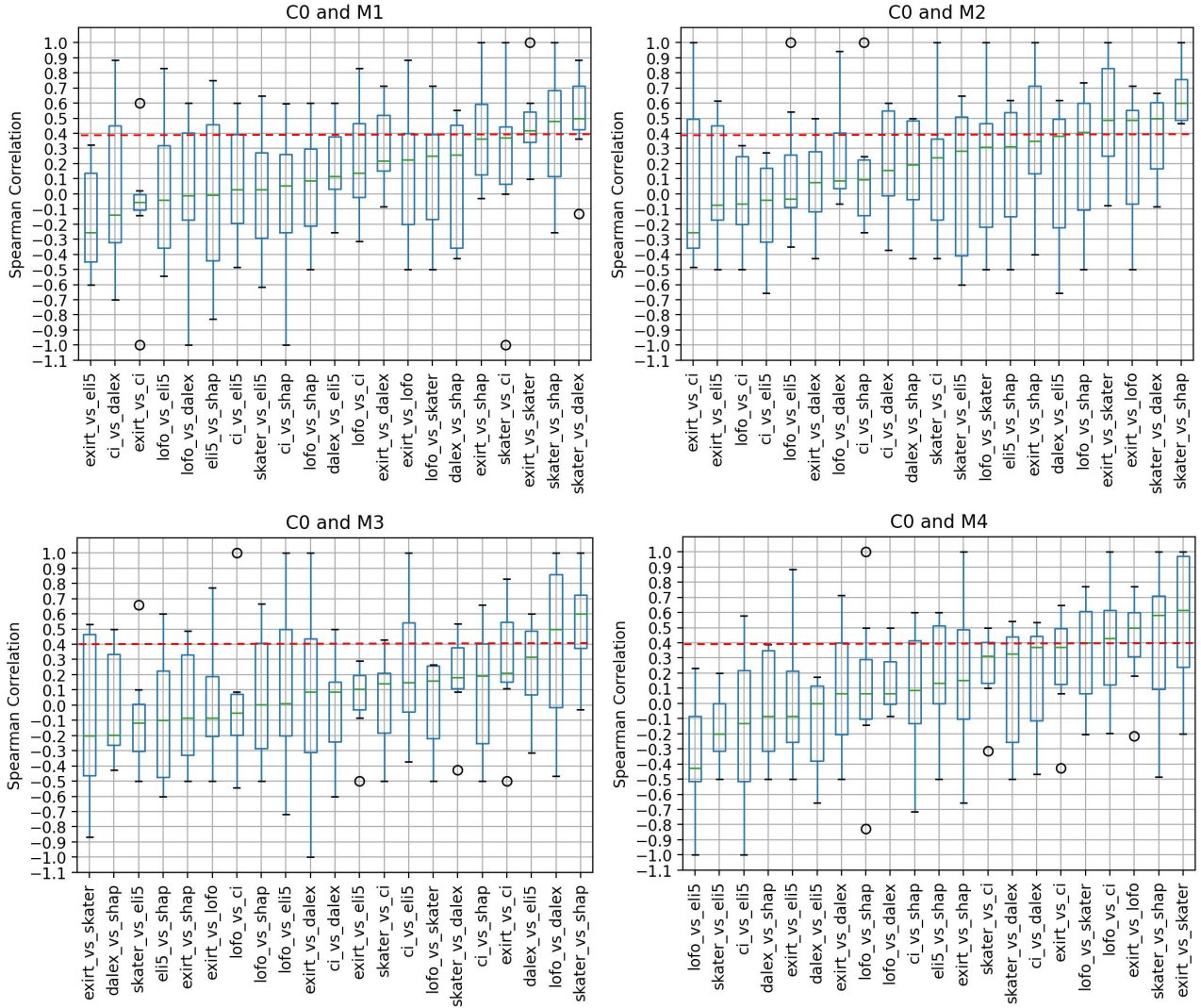

Precision of models
by cluster: 0
Precision
1.00
0.95
0.90
0.85
0.80
0.75
m1
m2
m3
m4


Recall of models
by cluster: 0
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
m1
m2
m3
m4

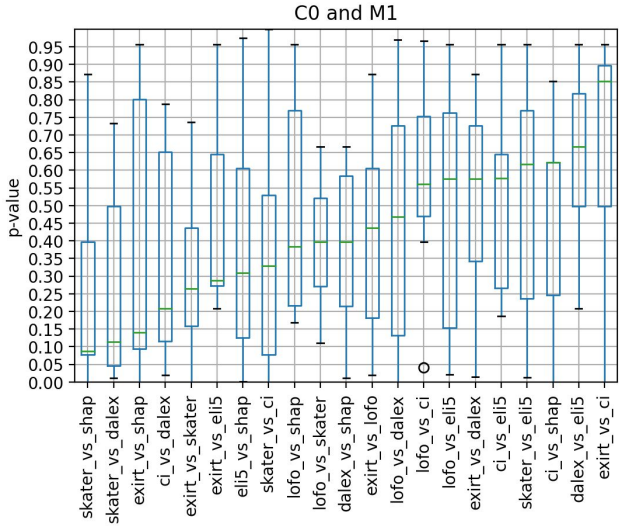
Friedman test result - Models by Cluster 0

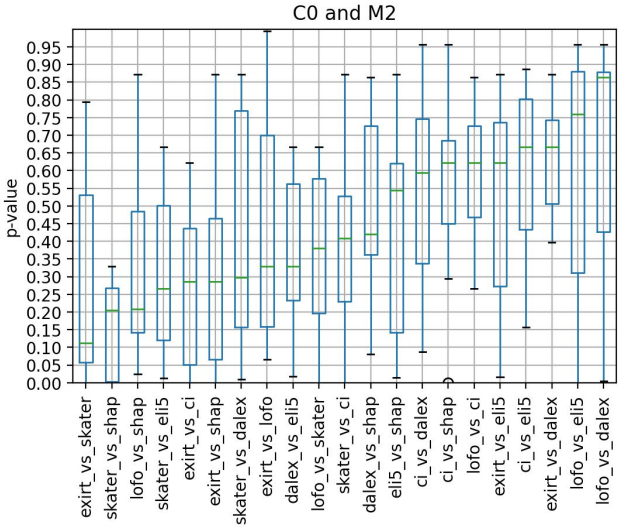

Friedman test of models
by cluster 0
m1
m2
m3
m4
1.00
0.90
0.08
0.22
0.90
1.00
0.02
0.08
0.08
0.02
1.00
0.90
0.22
0.08
0.90
1.00
m1
m2
m3
m4
1.0
0.8
0.6
0.4
0.2
0.0

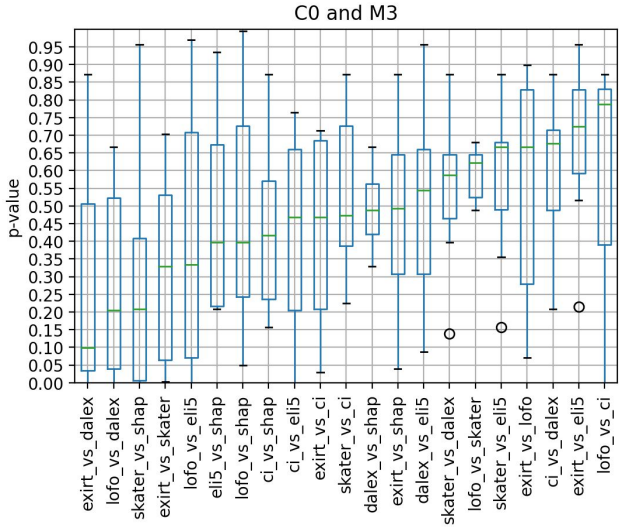
All pair spearman comparisons of feature relevance ranks for models (M1 to M4) based on cluster 0

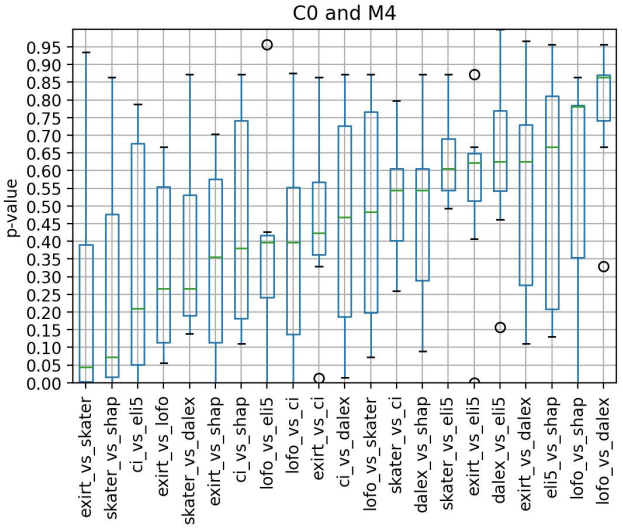

C0 and M1
Spearman Correlation
exirt_vs_eli5
ci_vs_dalex
exirt_vs_ci
lofo_vs_eli5
lofo_vs_dalex
eli5_vs_shap
ci_vs_eli5
skater_vs_eli5
ci_vs_shap
lofo_vs_shap
dalex_vs_eli5
lofo_vs_ci
exirt_vs_dalex
exirt_vs_lofo
lofo_vs_skater
dalex_vs_shap
exirt_vs_shap
skater_vs_ci
exirt_vs_skater
skater_vs_shap
skater_vs_dalex
C0 and M2
Spearman Correlation
exirt_vs_ci
exirt_vs_eli5
lofo_vs_ci
ci_vs_eli5
lofo_vs_eli5
exirt_vs_dalex
lofo_vs_dalex
ci_vs_shap
ci_vs_dalex
dalex_vs_shap
skater_vs_ci
skater_vs_eli5
lofo_vs_skater
eli5_vs_shap
exirt_vs_shap
dalex_vs_eli5
lofo_vs_shap
exirt_vs_skater
exirt_vs_lofo
skater_vs_dalex
skater_vs_shap
C0 and M3
Spearman Correlation
exirt_vs_skater
dalex_vs_shap
skater_vs_eli5
eli5_vs_shap
exirt_vs_shap
exirt_vs_lofo
lofo_vs_ci
lofo_vs_shap
lofo_vs_eli5
exirt_vs_dalex
ci_vs_dalex
exirt_vs_eli5
skater_vs_ci
ci_vs_eli5
lofo_vs_skater
skater_vs_dalex
ci_vs_shap
exirt_vs_ci
dalex_vs_eli5
lofo_vs_dalex
skater_vs_shap
C0 and M4
Spearman Correlation
lofo_vs_eli5
skater_vs_eli5
ci_vs_eli5
dalex_vs_shap
exirt_vs_eli5
dalex_vs_eli5
exirt_vs_dalex
lofo_vs_shap
lofo_vs_dalex
ci_vs_shap
eli5_vs_shap
exirt_vs_shap
skater_vs_ci
skater_vs_dalex
ci_vs_dalex
exirt_vs_ci
lofo_vs_skater
lofo_vs_ci
exirt_vs_lofo
skater_vs_shap
exirt_vs_skater

All pair *p-values* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 0

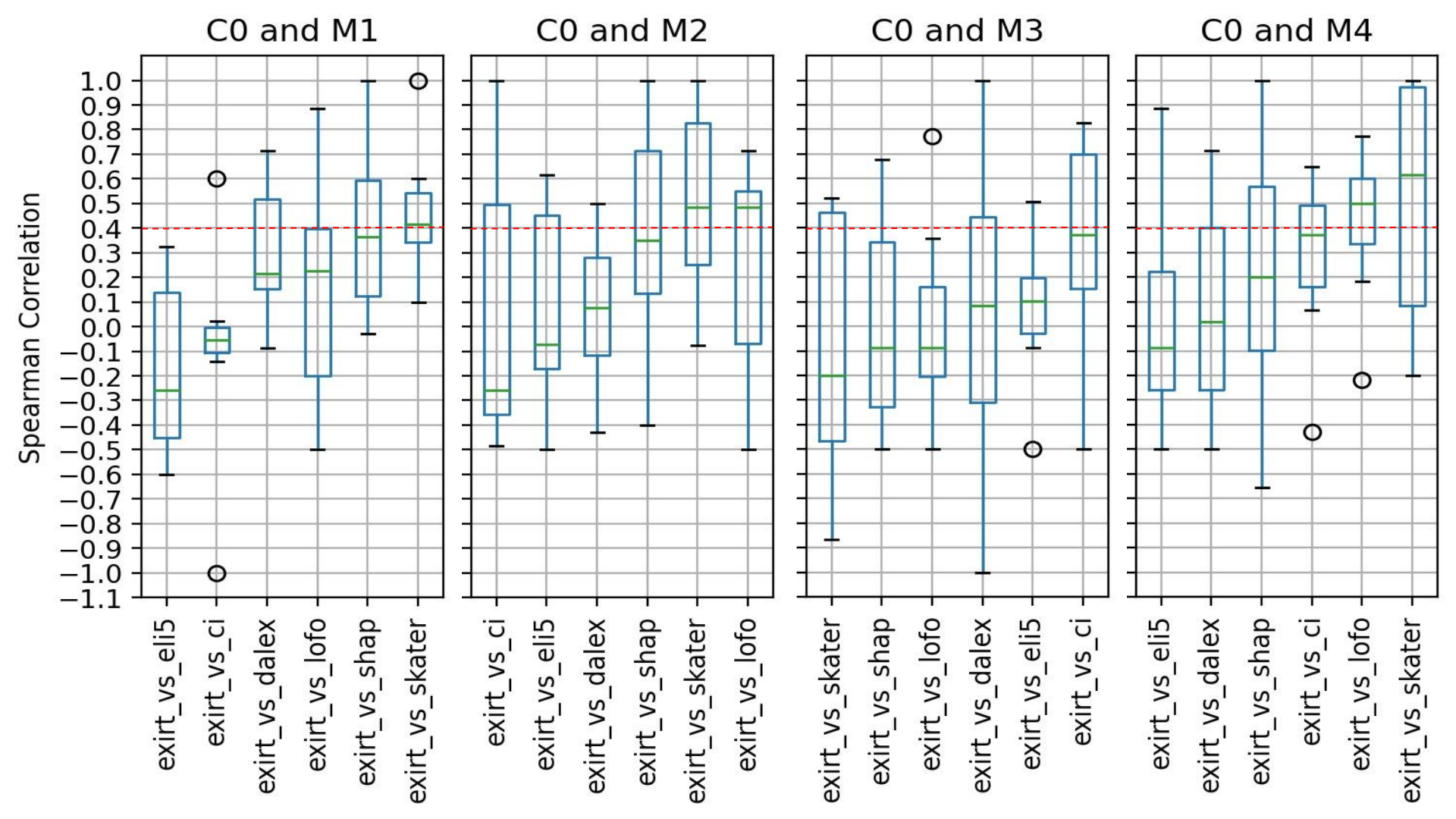

C0 and M1
p-value
skater_vs_shap
skater_vs_dalex
exirt_vs_shap
ci_vs_dalex
exirt_vs_skater
exirt_vs_eli5
eli5_vs_shap
skater_vs_ci
lofo_vs_shap
lofo_vs_skater
dalex_vs_shap
exirt_vs_lofo
lofo_vs_dalex
lofo_vs_ci
lofo_vs_eli5
exirt_vs_dalex
ci_vs_eli5
skater_vs_eli5
ci_vs_shap
dalex_vs_eli5
exirt_vs_ci


C0 and M2
p-value
exirt_vs_skater
skater_vs_shap
lofo_vs_shap
skater_vs_eli5
exirt_vs_ci
exirt_vs_shap
skater_vs_dalex
exirt_vs_lofo
dalex_vs_eli5
lofo_vs_skater
skater_vs_ci
dalex_vs_shap
eli5_vs_shap
ci_vs_dalex
ci_vs_shap
lofo_vs_ci
exirt_vs_eli5
ci_vs_eli5
exirt_vs_dalex
lofo_vs_eli5
lofo_vs_dalex

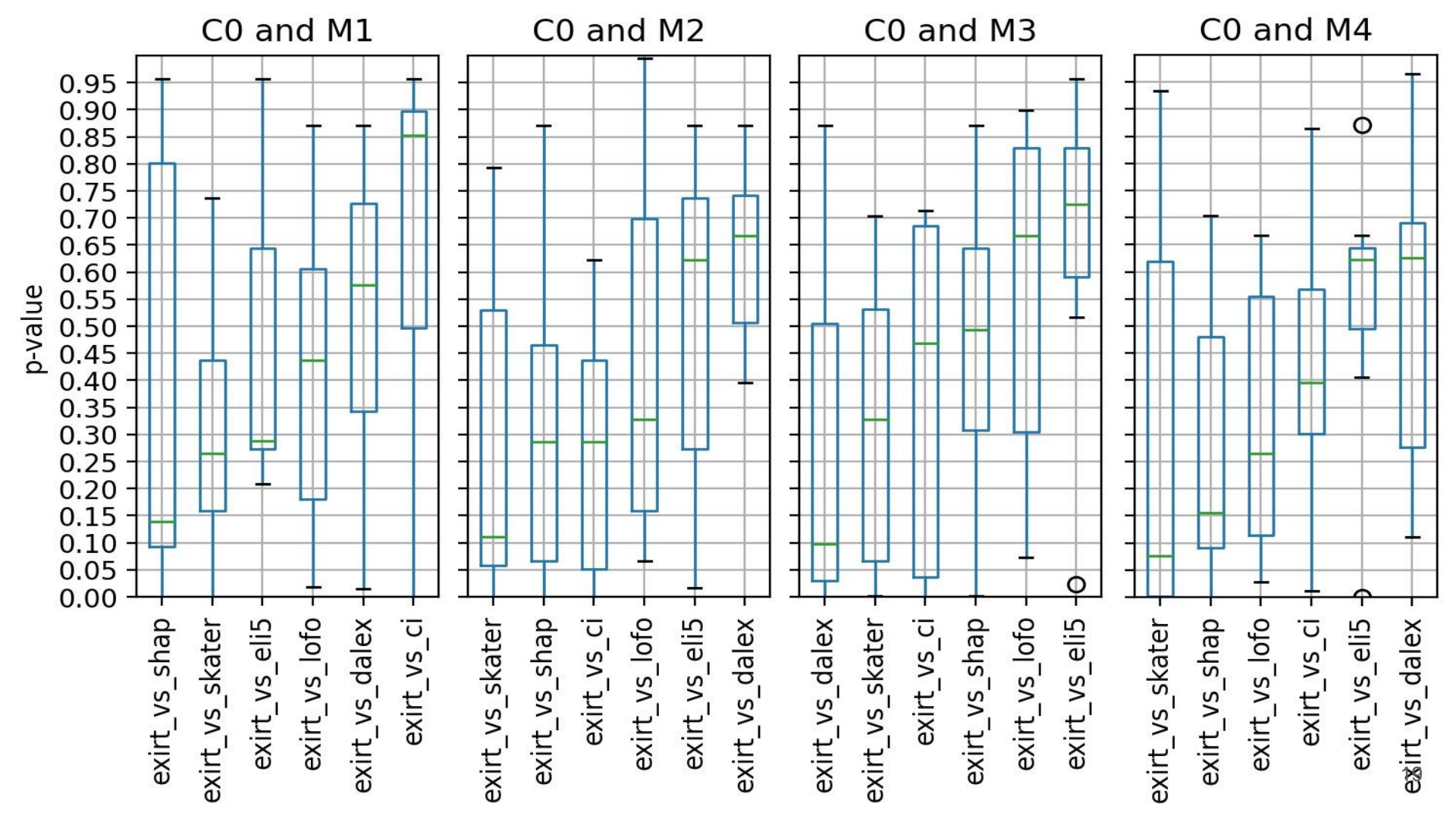

C0 and M3
p-value
exirt_vs_dalex
lofo_vs_dalex
skater_vs_shap
exirt_vs_skater
lofo_vs_eli5
eli5_vs_shap
lofo_vs_shap
ci_vs_shap
ci_vs_eli5
exirt_vs_ci
skater_vs_ci
dalex_vs_shap
exirt_vs_shap
dalex_vs_eli5
skater_vs_dalex
lofo_vs_skater
skater_vs_eli5
exirt_vs_lofo
ci_vs_dalex
exirt_vs_eli5
lofo_vs_ci


C0 and M4
p-value
exirt_vs_skater
skater_vs_shap
ci_vs_eli5
exirt_vs_lofo
skater_vs_dalex
exirt_vs_shap
ci_vs_shap
lofo_vs_eli5
lofo_vs_ci
exirt_vs_ci
ci_vs_dalex
lofo_vs_skater
skater_vs_ci
dalex_vs_shap
skater_vs_eli5
exirt_vs_eli5
dalex_vs_eli5
exirt_vs_dalex
eli5_vs_shap
lofo_vs_shap
lofo_vs_dalex

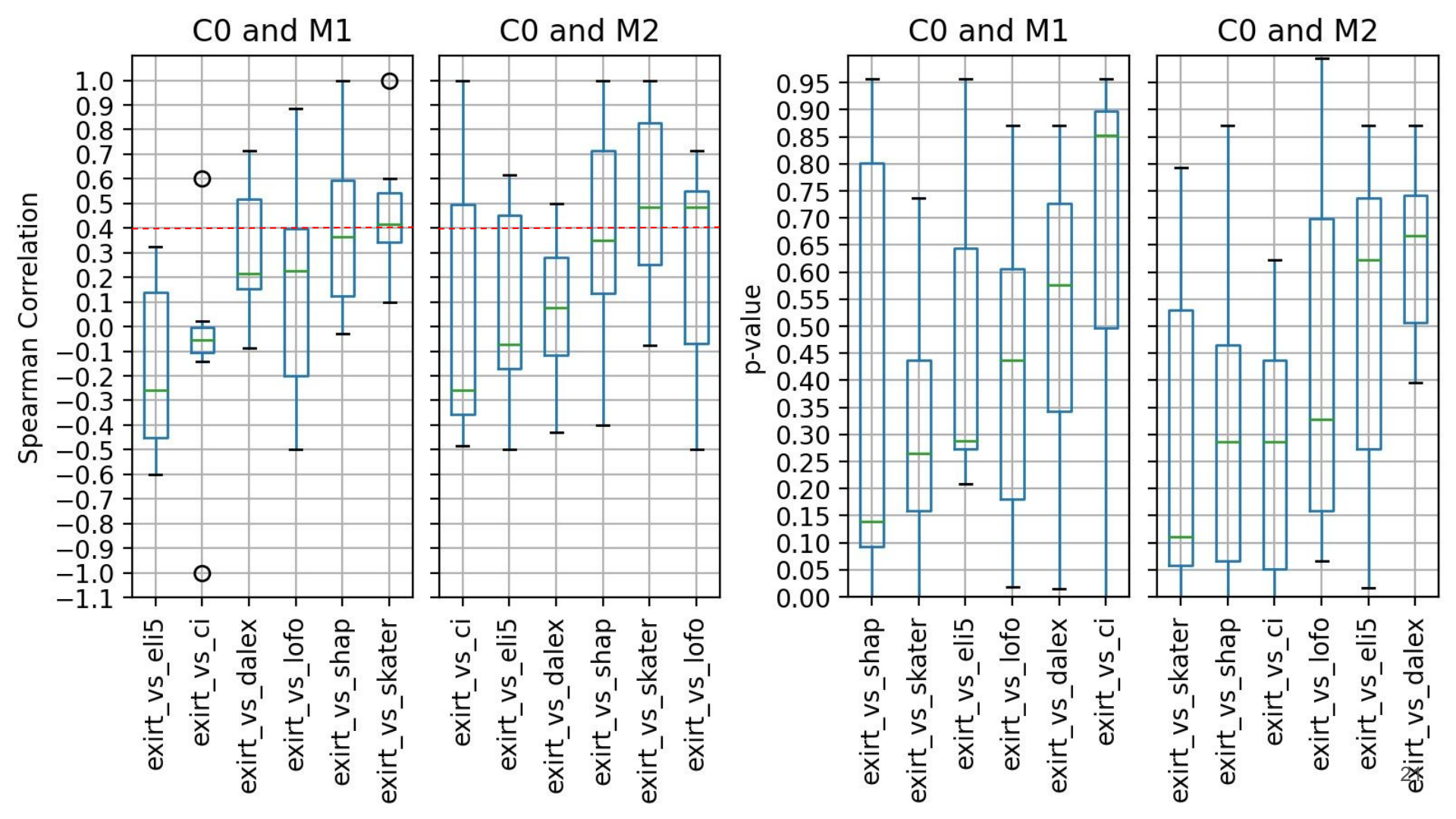
Only *eXirt* spearman comparisons of feature relevance ranks for models (M1 to M4) based on cluster 0


C0 and M1
C0 and M2
C0 and M3
C0 and M4
Spearman Correlation
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
−1.1
exirt_vs_eli5
exirt_vs_ci
exirt_vs_dalex
exirt_vs_lofo
exirt_vs_shap
exirt_vs_skater
exirt_vs_ci
exirt_vs_eli5
exirt_vs_dalex
exirt_vs_shap
exirt_vs_skater
exirt_vs_lofo
exirt_vs_skater
exirt_vs_shap
exirt_vs_lofo
exirt_vs_dalex
exirt_vs_eli5
exirt_vs_ci
exirt_vs_eli5
exirt_vs_dalex
exirt_vs_shap
exirt_vs_ci
exirt_vs_lofo
exirt_vs_skater

Only *eXirt p-values* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 0

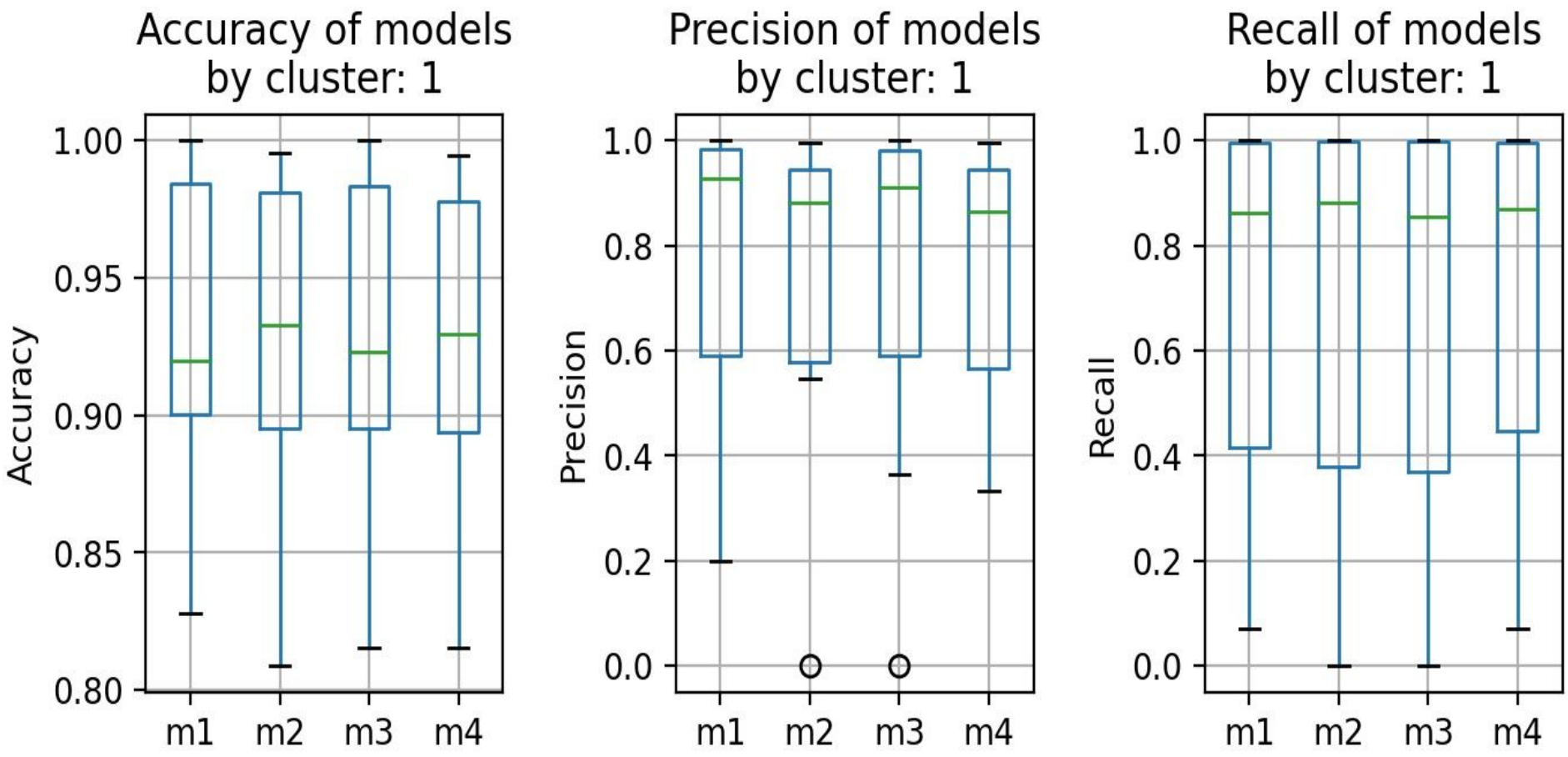

C0 and M1
C0 and M2
C0 and M3
C0 and M4
p-value
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
exirt_vs_shap
exirt_vs_skater
exirt_vs_eli5
exirt_vs_lofo
exirt_vs_dalex
exirt_vs_ci
exirt_vs_skater
exirt_vs_shap
exirt_vs_ci
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_dalex
exirt_vs_dalex
exirt_vs_skater
exirt_vs_ci
exirt_vs_shap
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_skater
exirt_vs_shap
exirt_vs_lofo
exirt_vs_ci
exirt_vs_eli5
exirt_vs_dalex

Only *eXirt* comparisons of feature relevance ranks for models (M1 and M2) based on cluster 0

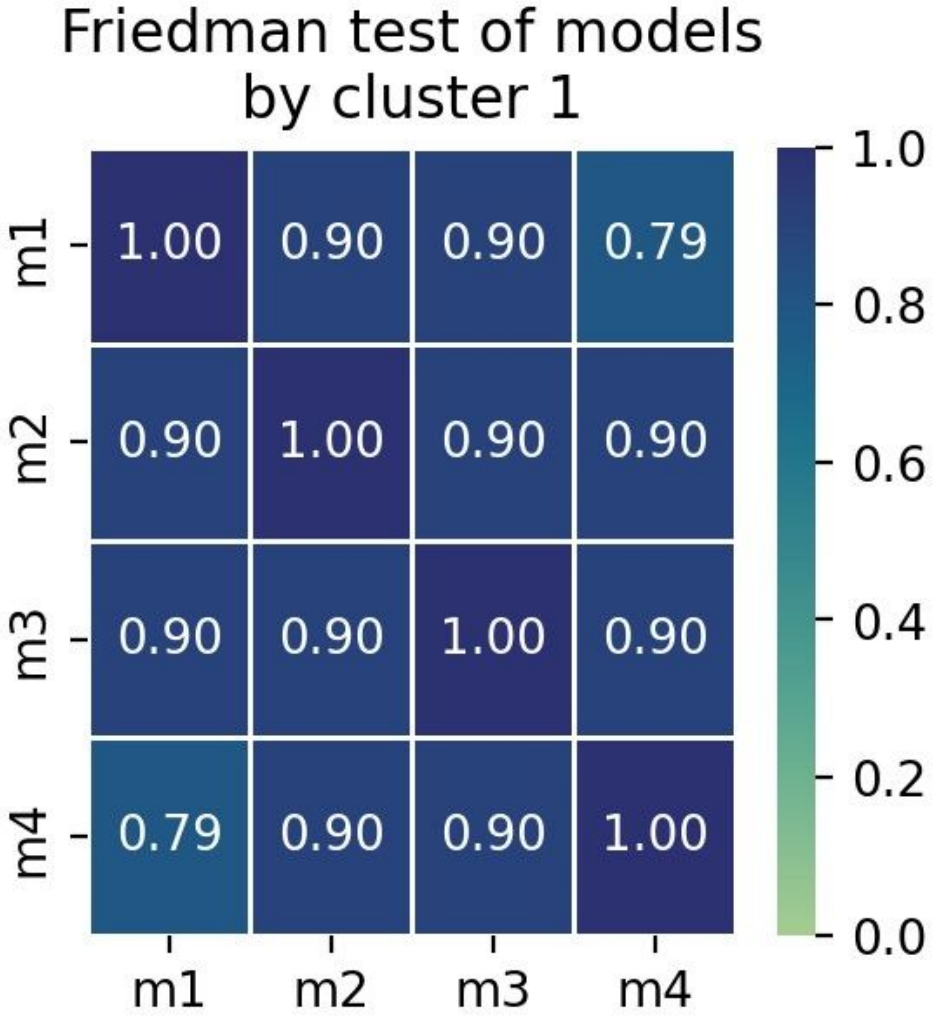

C0 and M1
C0 and M2
Spearman Correlation
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
−1.1
exirt_vs_eli5
exirt_vs_ci
exirt_vs_dalex
exirt_vs_lofo
exirt_vs_shap
exirt_vs_skater
exirt_vs_ci
exirt_vs_eli5
exirt_vs_dalex
exirt_vs_shap
exirt_vs_skater
exirt_vs_lofo
C0 and M1
C0 and M2
p-value
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
exirt_vs_shap
exirt_vs_skater
exirt_vs_eli5
exirt_vs_lofo
exirt_vs_dalex
exirt_vs_ci
exirt_vs_skater
exirt_vs_shap
exirt_vs_ci
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_dalex

Cluster 1

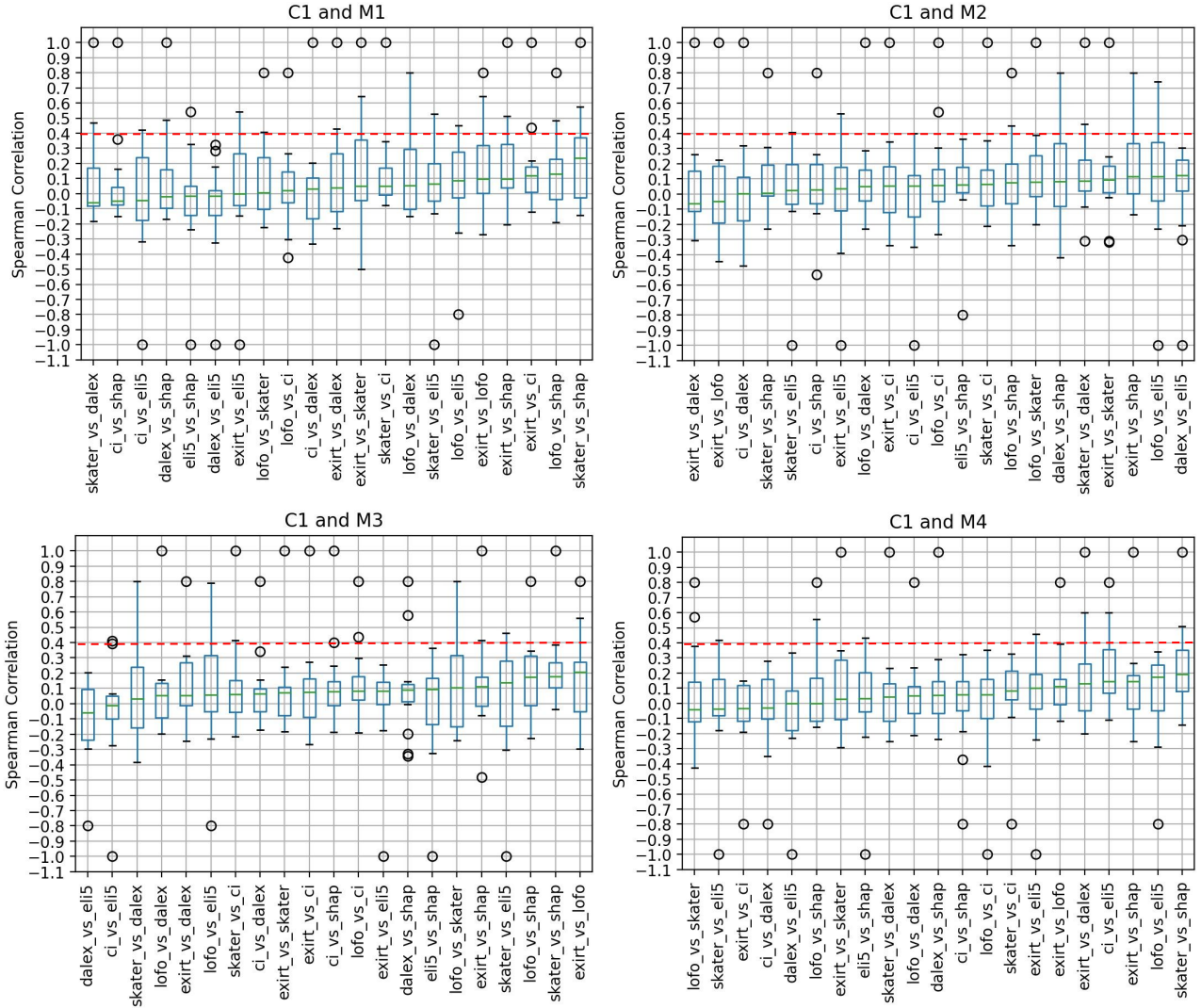
Data regarding the performance of models M1 to M4 based on cluster 1


Accuracy of models by cluster: 1
Accuracy
1.00
0.95
0.90
0.85
0.80
m1
m2
m3
m4
Precision of models by cluster: 1
Precision
1.0
0.8
0.6
0.4
0.2
0.0
m1
m2
m3
m4
Recall of models by cluster: 1
Recall
1.0
0.8
0.6
0.4
0.2
0.0
m1
m2
m3
m4

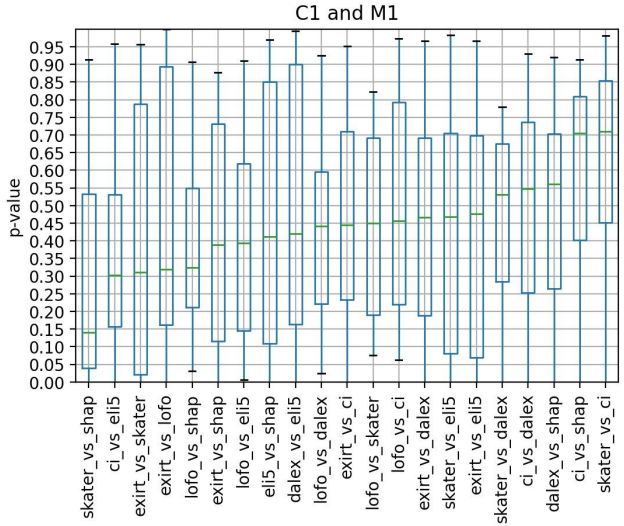
Friedman test result - Models by Cluster 1

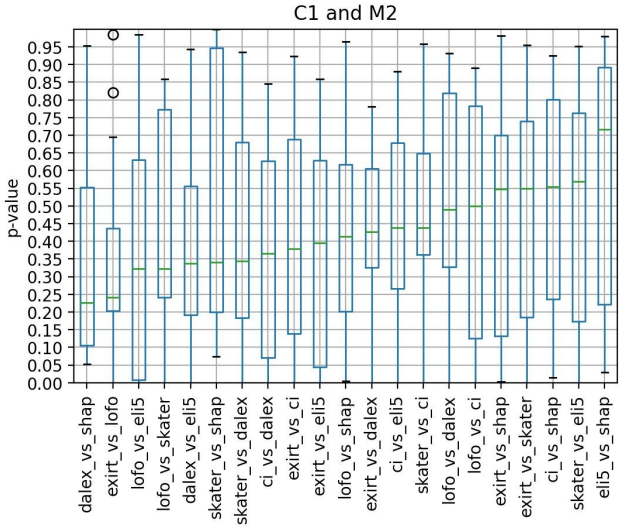

Friedman test of models
by cluster 1
m1
m2
m3
m4
1.00
0.90
0.90
0.79
0.90
1.00
0.90
0.90
0.90
0.90
1.00
0.90
0.79
0.90
0.90
1.00
m1
m2
m3
m4
1.0
0.8
0.6
0.4
0.2
0.0

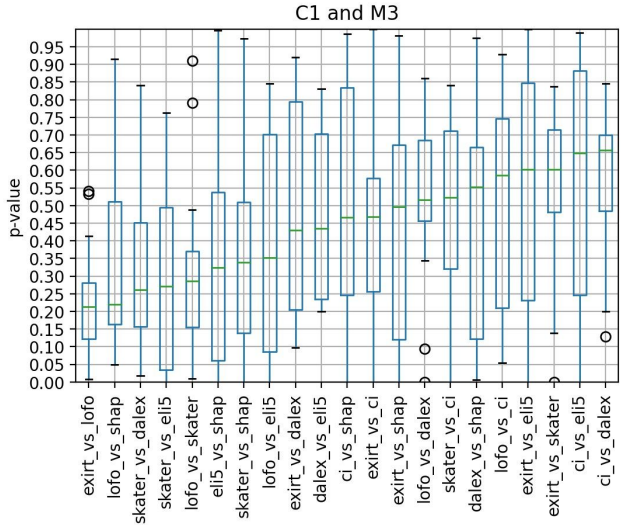
All pair spearman comparisons of feature relevance ranks for models (M1 to M4) based on cluster 1

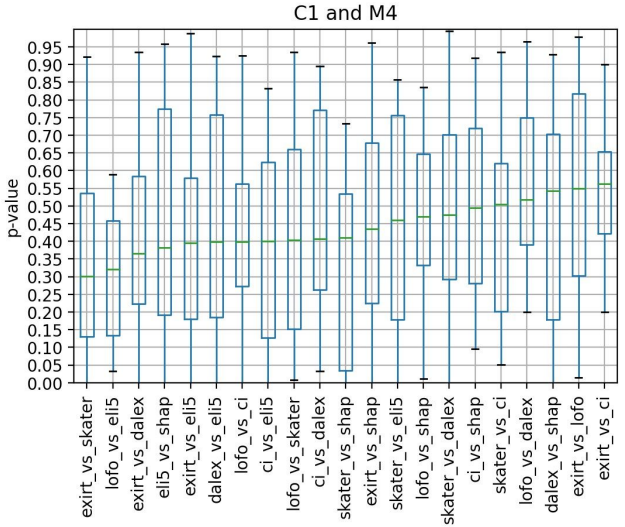

C1 and M1
Spearman Correlation
skater_vs_dalex
ci_vs_shap
ci_vs_eli5
dalex_vs_shap
eli5_vs_shap
dalex_vs_eli5
exirt_vs_eli5
lofo_vs_skater
lofo_vs_ci
ci_vs_dalex
exirt_vs_dalex
exirt_vs_skater
skater_vs_ci
lofo_vs_dalex
skater_vs_eli5
lofo_vs_eli5
exirt_vs_lofo
exirt_vs_shap
exirt_vs_ci
lofo_vs_shap
skater_vs_shap
C1 and M2
Spearman Correlation
exirt_vs_dalex
exirt_vs_lofo
ci_vs_dalex
skater_vs_shap
skater_vs_eli5
ci_vs_shap
exirt_vs_eli5
lofo_vs_dalex
exirt_vs_ci
ci_vs_eli5
lofo_vs_ci
eli5_vs_shap
skater_vs_ci
lofo_vs_shap
lofo_vs_skater
dalex_vs_shap
skater_vs_dalex
exirt_vs_skater
exirt_vs_shap
lofo_vs_eli5
dalex_vs_eli5
C1 and M3
Spearman Correlation
dalex_vs_eli5
ci_vs_eli5
skater_vs_dalex
lofo_vs_dalex
exirt_vs_dalex
lofo_vs_eli5
skater_vs_ci
ci_vs_dalex
exirt_vs_skater
exirt_vs_ci
ci_vs_shap
lofo_vs_ci
exirt_vs_eli5
dalex_vs_shap
eli5_vs_shap
lofo_vs_skater
exirt_vs_shap
skater_vs_eli5
lofo_vs_shap
skater_vs_shap
exirt_vs_lofo
C1 and M4
Spearman Correlation
lofo_vs_skater
skater_vs_eli5
exirt_vs_ci
ci_vs_dalex
dalex_vs_eli5
lofo_vs_shap
exirt_vs_skater
eli5_vs_shap
skater_vs_dalex
lofo_vs_dalex
dalex_vs_shap
ci_vs_shap
lofo_vs_ci
skater_vs_ci
exirt_vs_eli5
exirt_vs_lofo
exirt_vs_dalex
ci_vs_eli5
exirt_vs_shap
lofo_vs_eli5
skater_vs_shap

All pair *p-values* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 1

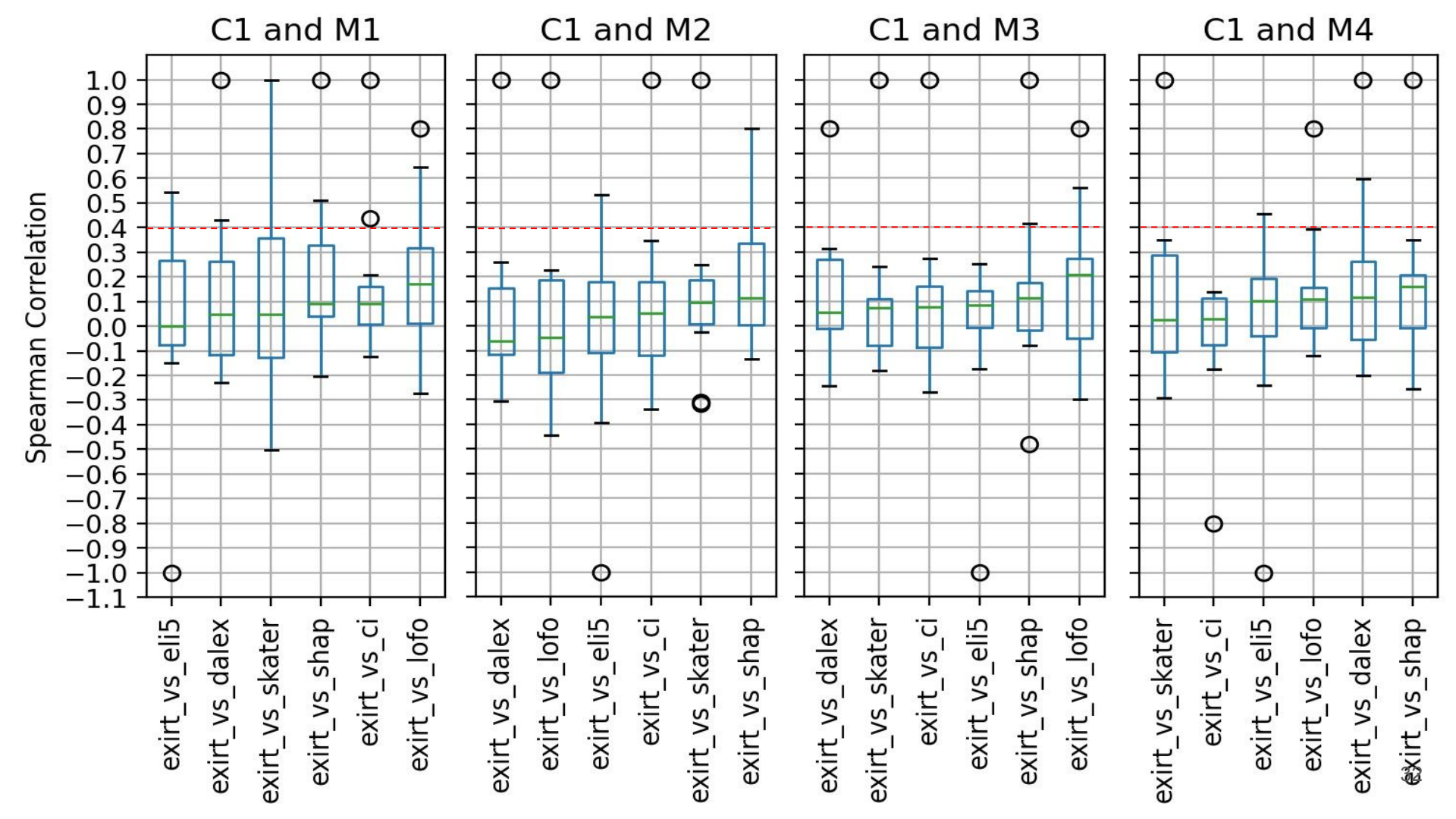

C1 and M1
p-value
skater_vs_shap
ci_vs_eli5
exirt_vs_skater
exirt_vs_lofo
lofo_vs_shap
exirt_vs_shap
lofo_vs_eli5
eli5_vs_shap
dalex_vs_eli5
lofo_vs_dalex
exirt_vs_ci
lofo_vs_skater
lofo_vs_ci
exirt_vs_dalex
skater_vs_eli5
exirt_vs_eli5
skater_vs_dalex
ci_vs_dalex
dalex_vs_shap
ci_vs_shap
skater_vs_ci


C1 and M2
p-value
dalex_vs_shap
exirt_vs_lofo
lofo_vs_eli5
lofo_vs_skater
dalex_vs_eli5
skater_vs_shap
skater_vs_dalex
ci_vs_dalex
exirt_vs_ci
exirt_vs_eli5
lofo_vs_shap
exirt_vs_dalex
ci_vs_eli5
skater_vs_ci
lofo_vs_dalex
lofo_vs_ci
exirt_vs_shap
exirt_vs_skater
ci_vs_shap
skater_vs_eli5
eli5_vs_shap

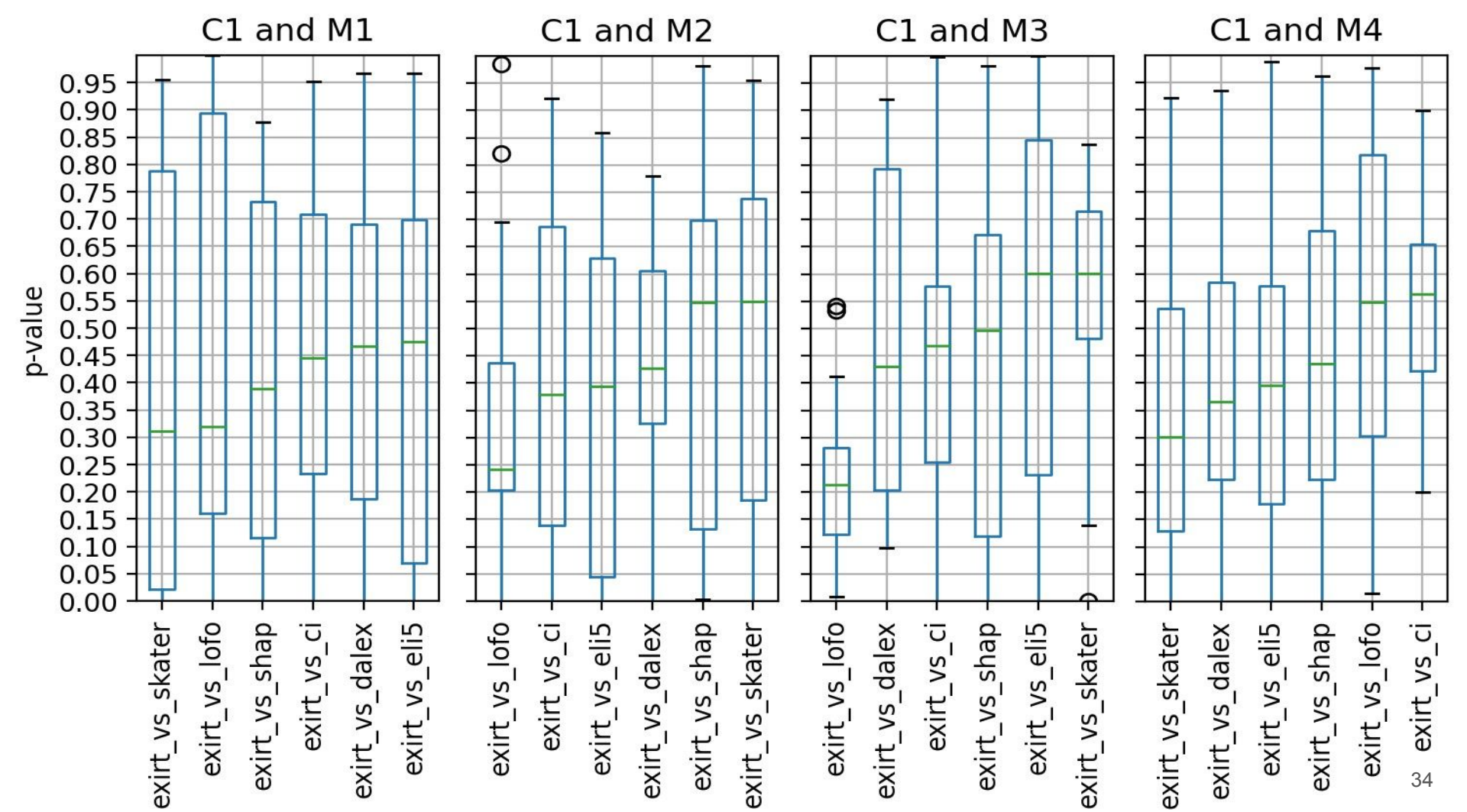

C1 and M3
p-value
exirt_vs_lofo
lofo_vs_shap
skater_vs_dalex
skater_vs_eli5
lofo_vs_skater
eli5_vs_shap
skater_vs_shap
lofo_vs_eli5
exirt_vs_dalex
dalex_vs_eli5
ci_vs_shap
exirt_vs_ci
exirt_vs_shap
lofo_vs_dalex
skater_vs_ci
dalex_vs_shap
lofo_vs_ci
exirt_vs_eli5
exirt_vs_skater
ci_vs_eli5
ci_vs_dalex


C1 and M4
p-value
exirt_vs_skater
lofo_vs_eli5
exirt_vs_dalex
eli5_vs_shap
exirt_vs_eli5
dalex_vs_eli5
lofo_vs_ci
ci_vs_eli5
lofo_vs_skater
ci_vs_dalex
skater_vs_shap
exirt_vs_shap
skater_vs_eli5
lofo_vs_shap
skater_vs_dalex
ci_vs_shap
skater_vs_ci
lofo_vs_dalex
dalex_vs_shap
exirt_vs_lofo
exirt_vs_ci

Only *eXirt* spearman comparisons of feature relevance ranks for models (M1 to M4) based on cluster 1

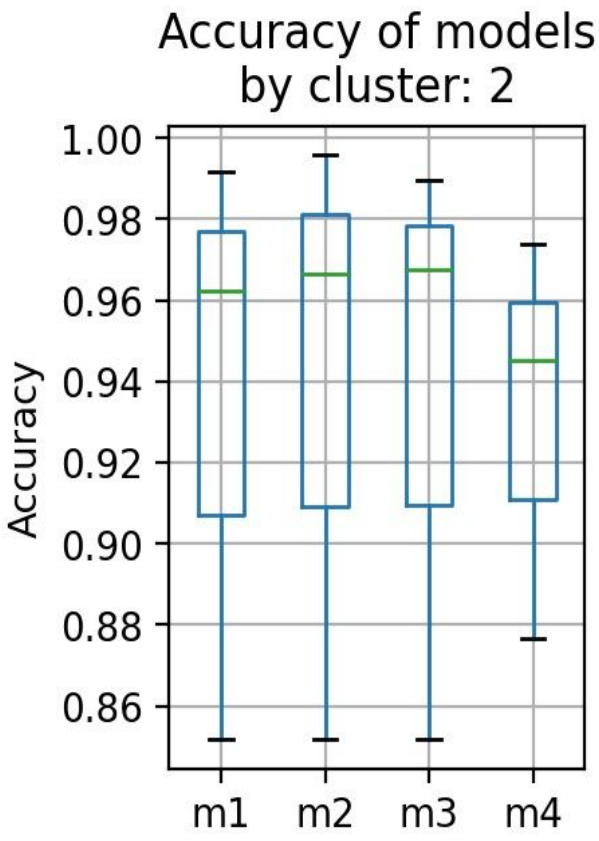

C1 and M1
C1 and M2
C1 and M3
C1 and M4
Spearman Correlation
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
−1.1
exirt_vs_eli5
exirt_vs_dalex
exirt_vs_skater
exirt_vs_shap
exirt_vs_ci
exirt_vs_lofo
exirt_vs_dalex
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_ci
exirt_vs_skater
exirt_vs_shap
exirt_vs_dalex
exirt_vs_skater
exirt_vs_ci
exirt_vs_eli5
exirt_vs_shap
exirt_vs_lofo
exirt_vs_skater
exirt_vs_ci
exirt_vs_eli5
exirt_vs_lofo
exirt_vs_dalex
exirt_vs_shap

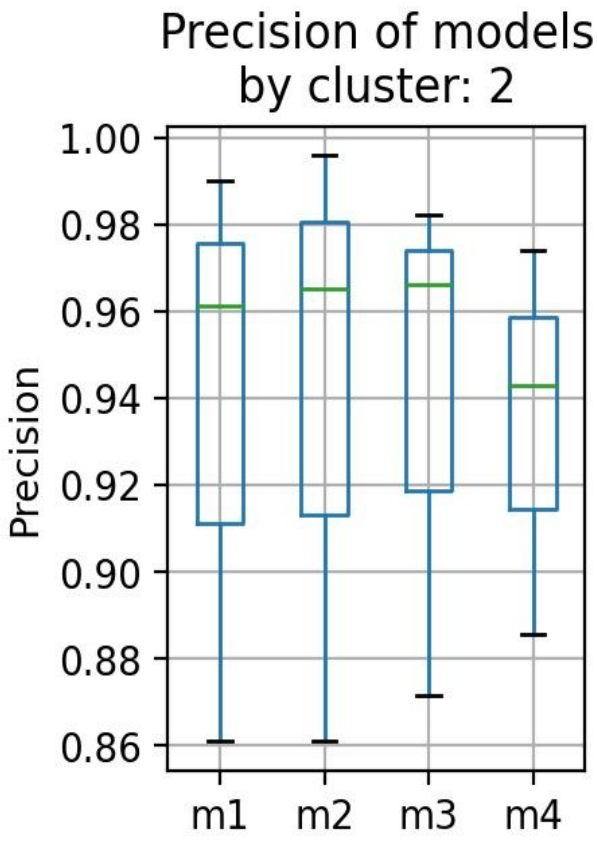
Only *eXirt p-values* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 1

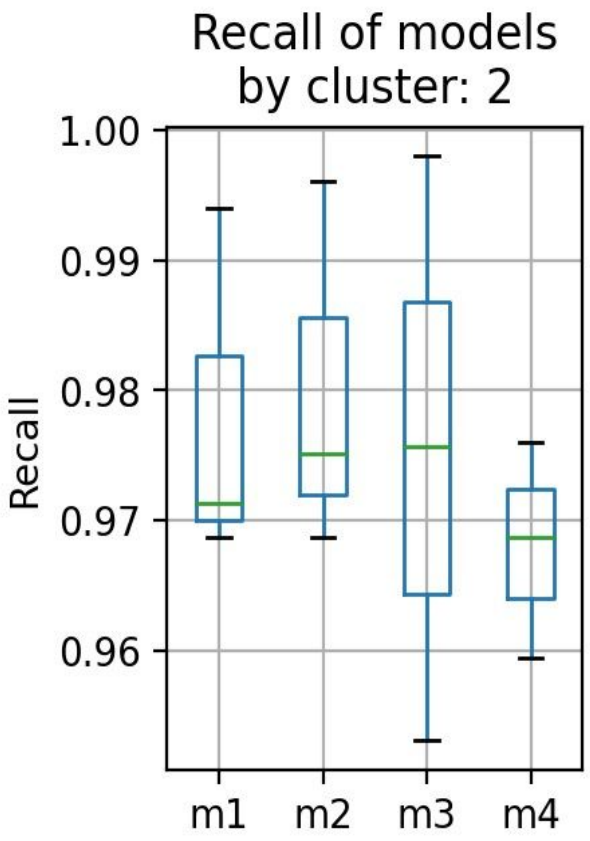

C1 and M1
C1 and M2
C1 and M3
C1 and M4
p-value
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
exirt_vs_skater
exirt_vs_lofo
exirt_vs_shap
exirt_vs_ci
exirt_vs_dalex
exirt_vs_eli5
exirt_vs_lofo
exirt_vs_ci
exirt_vs_eli5
exirt_vs_dalex
exirt_vs_shap
exirt_vs_skater
exirt_vs_lofo
exirt_vs_dalex
exirt_vs_ci
exirt_vs_shap
exirt_vs_eli5
exirt_vs_skater
exirt_vs_skater
exirt_vs_dalex
exirt_vs_eli5
exirt_vs_shap
exirt_vs_lofo
exirt_vs_ci

Cluster 2

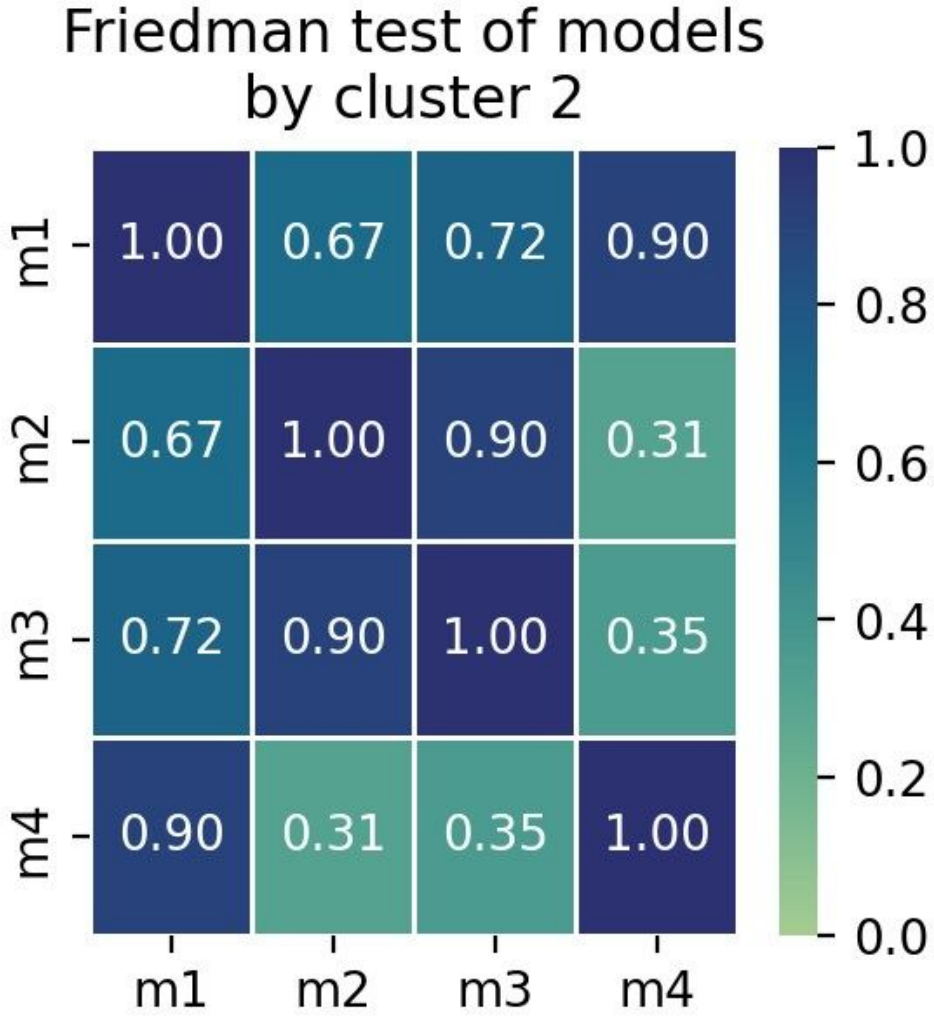
Data regarding the performance of models M1 to M4 based on cluster 2


Accuracy of models
by cluster: 2
Accuracy
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
m1
m2
m3
m4

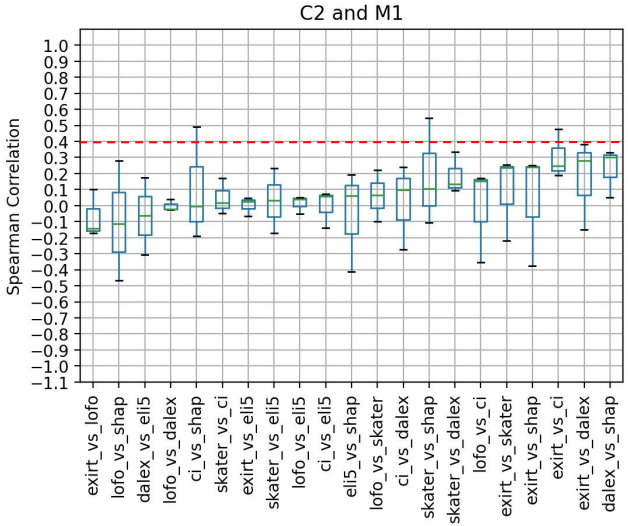

Precision of models
by cluster: 2
Precision
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
m1
m2
m3
m4

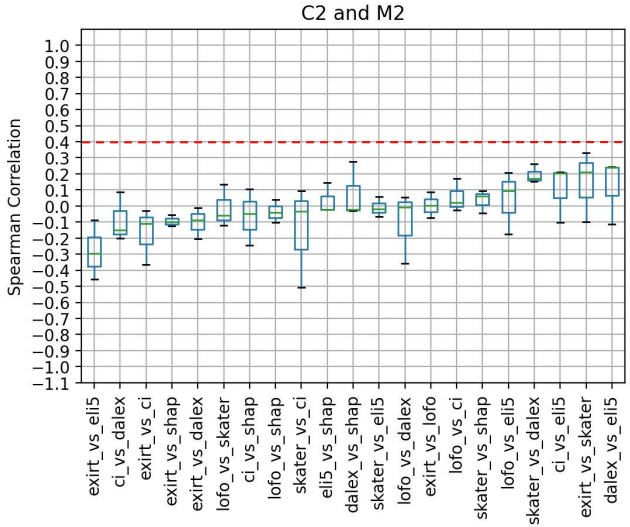

Recall of models
by cluster: 2
Recall
1.00
0.99
0.98
0.97
0.96
m1
m2
m3
m4

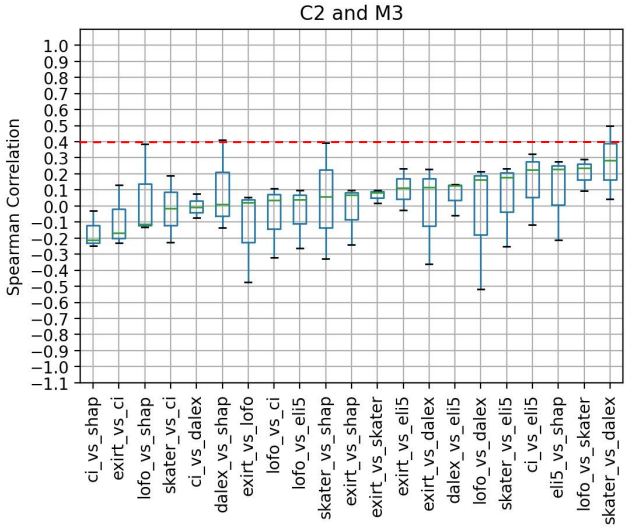
Friedman test result - Models by Cluster 2

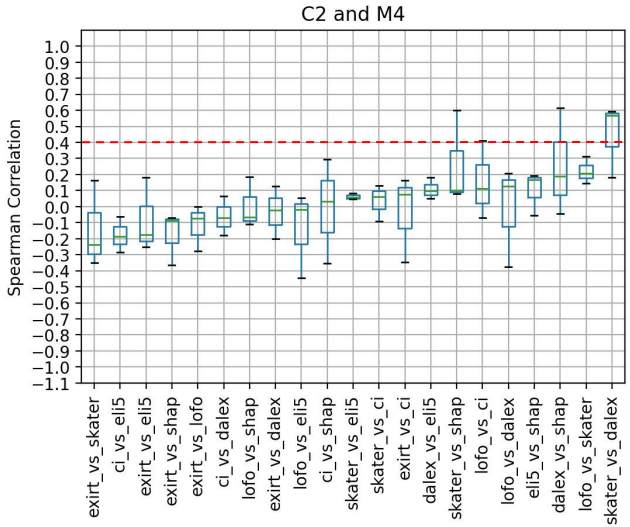

Friedman test of models
by cluster 2
m1
m2
m3
m4
1.00
0.67
0.72
0.90
0.67
1.00
0.90
0.31
0.72
0.90
1.00
0.35
0.90
0.31
0.35
1.00
m1
m2
m3
m4
1.0
0.8
0.6
0.4
0.2
0.0

All pair spearman comparisons of feature relevance ranks for models (M1 to M4) based on cluster 2

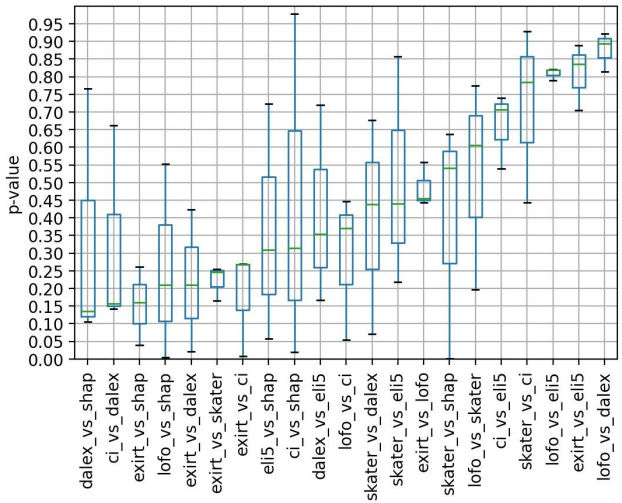

C2 and M1
Spearman Correlation
exirt_vs_lofo
lofo_vs_shap
dalex_vs_eli5
lofo_vs_dalex
ci_vs_shap
skater_vs_ci
exirt_vs_eli5
skater_vs_eli5
lofo_vs_eli5
ci_vs_eli5
eli5_vs_shap
lofo_vs_skater
ci_vs_dalex
skater_vs_shap
skater_vs_dalex
lofo_vs_ci
exirt_vs_skater
exirt_vs_shap
exirt_vs_ci
exirt_vs_dalex
dalex_vs_shap

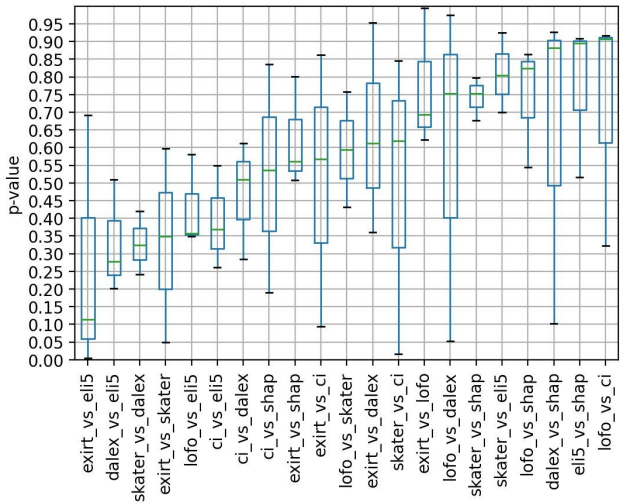

C2 and M2
Spearman Correlation
exirt_vs_eli5
ci_vs_dalex
exirt_vs_ci
exirt_vs_shap
exirt_vs_dalex
lofo_vs_skater
ci_vs_shap
lofo_vs_shap
skater_vs_ci
eli5_vs_shap
dalex_vs_shap
skater_vs_eli5
lofo_vs_dalex
exirt_vs_lofo
lofo_vs_ci
skater_vs_shap
lofo_vs_eli5
skater_vs_dalex
ci_vs_eli5
exirt_vs_skater
dalex_vs_eli5

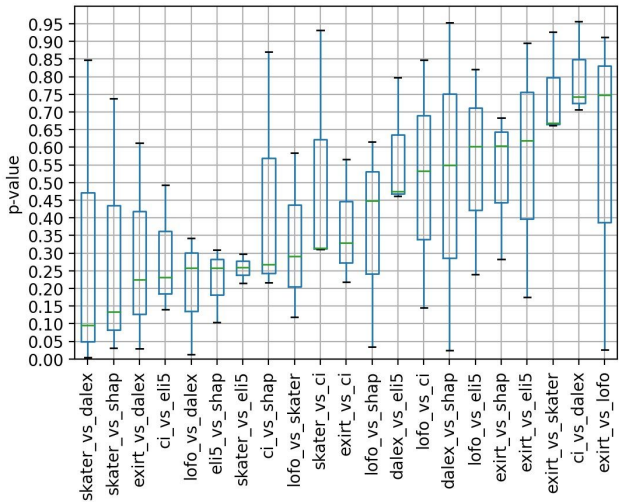

C2 and M3
Spearman Correlation
ci_vs_shap
exirt_vs_ci
lofo_vs_shap
skater_vs_ci
ci_vs_dalex
dalex_vs_shap
exirt_vs_lofo
lofo_vs_ci
lofo_vs_eli5
skater_vs_shap
exirt_vs_shap
exirt_vs_skater
exirt_vs_eli5
exirt_vs_dalex
dalex_vs_eli5
lofo_vs_dalex
skater_vs_eli5
ci_vs_eli5
eli5_vs_shap
lofo_vs_skater
skater_vs_dalex

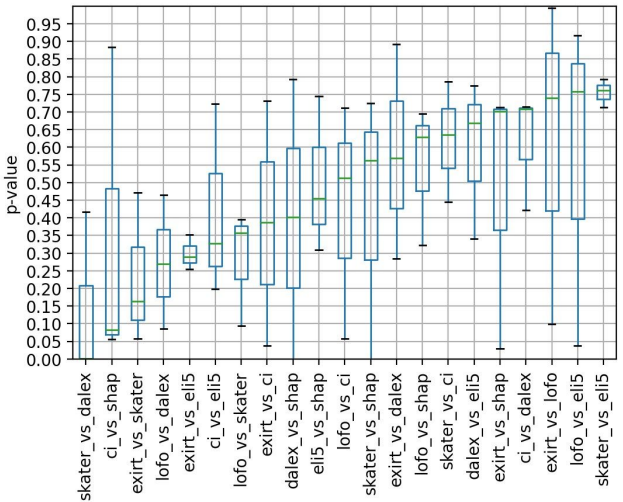

C2 and M4
Spearman Correlation
exirt_vs_skater
ci_vs_eli5
exirt_vs_eli5
exirt_vs_shap
exirt_vs_lofo
ci_vs_dalex
lofo_vs_shap
exirt_vs_dalex
lofo_vs_eli5
ci_vs_shap
skater_vs_eli5
skater_vs_ci
exirt_vs_ci
dalex_vs_eli5
skater_vs_shap
lofo_vs_ci
lofo_vs_dalex
eli5_vs_shap
dalex_vs_shap
lofo_vs_skater
skater_vs_dalex

All pair *p-values* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 2

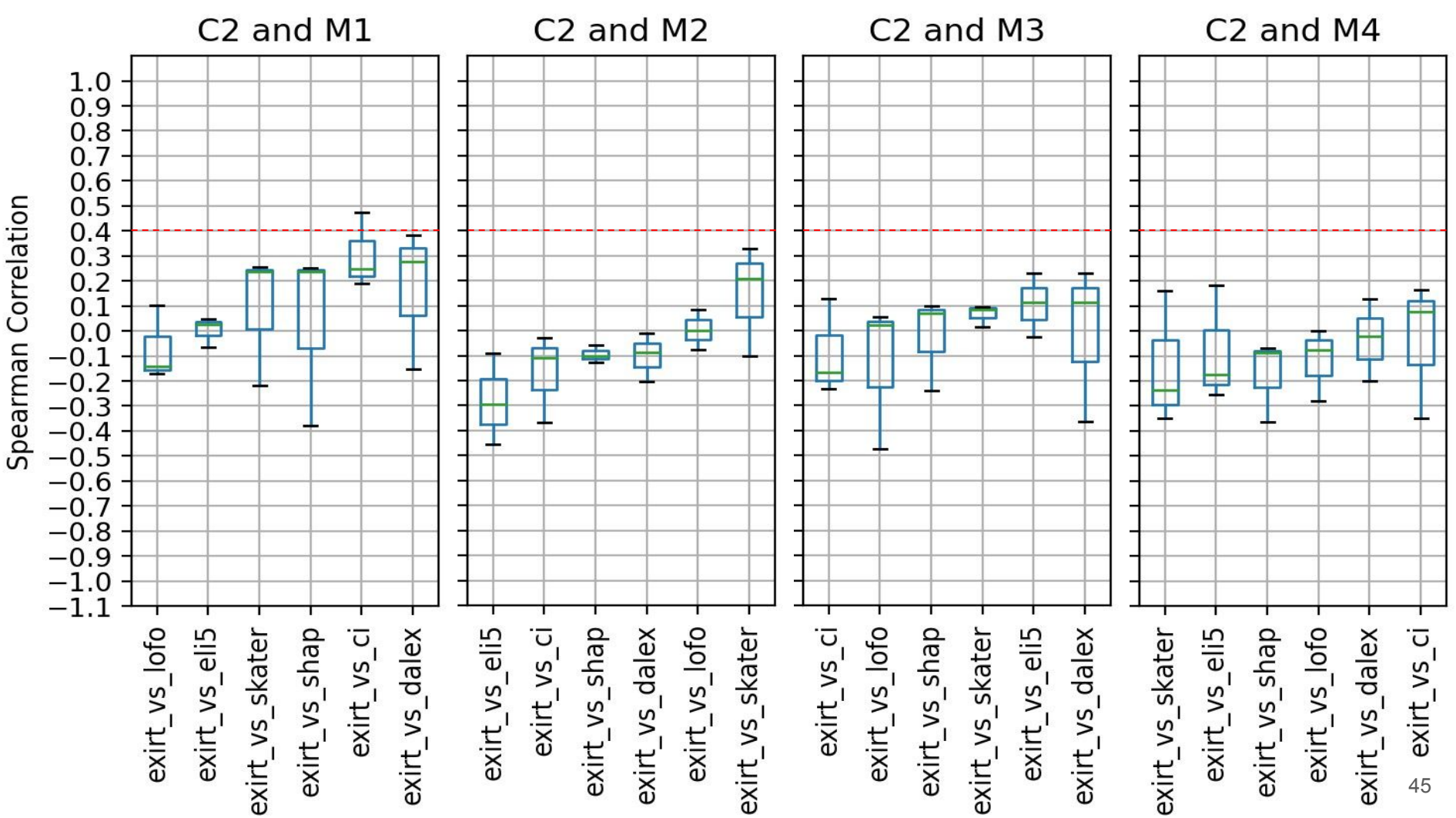

C2 and M1
p-value
dalex_vs_shap
ci_vs_dalex
exirt_vs_shap
lofo_vs_shap
exirt_vs_dalex
exirt_vs_skater
exirt_vs_ci
eli5_vs_shap
ci_vs_shap
dalex_vs_eli5
lofo_vs_ci
skater_vs_dalex
skater_vs_eli5
exirt_vs_lofo
skater_vs_shap
lofo_vs_skater
ci_vs_eli5
skater_vs_ci
lofo_vs_eli5
exirt_vs_eli5
lofo_vs_dalex


C2 and M2
p-value
exirt_vs_eli5
dalex_vs_eli5
skater_vs_dalex
exirt_vs_skater
lofo_vs_eli5
ci_vs_eli5
ci_vs_dalex
ci_vs_shap
exirt_vs_shap
exirt_vs_ci
lofo_vs_skater
exirt_vs_dalex
skater_vs_ci
exirt_vs_lofo
lofo_vs_dalex
skater_vs_shap
skater_vs_eli5
lofo_vs_shap
dalex_vs_shap
eli5_vs_shap
lofo_vs_ci

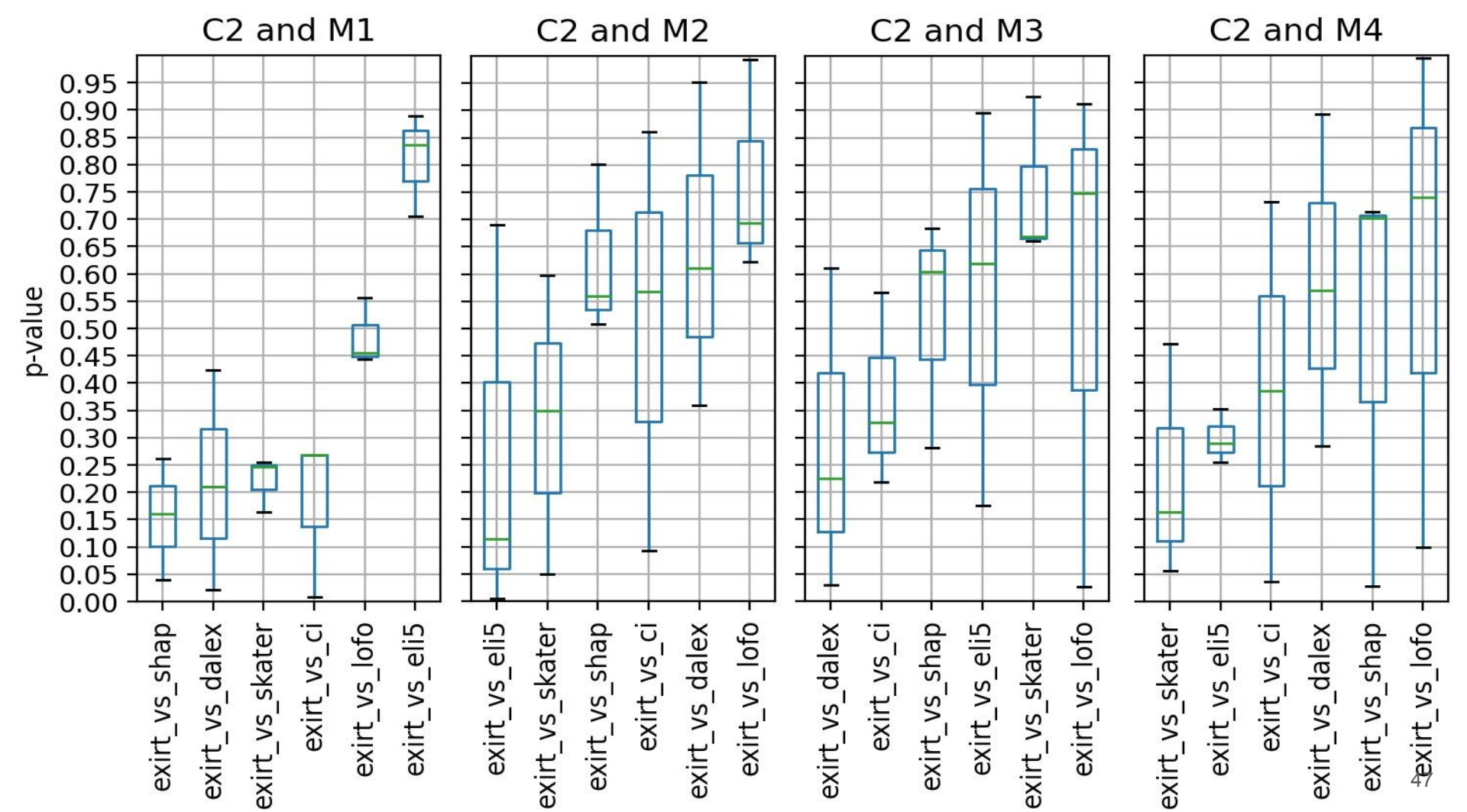

C2 and M3
p-value
skater_vs_dalex
skater_vs_shap
exirt_vs_dalex
ci_vs_eli5
lofo_vs_dalex
eli5_vs_shap
skater_vs_eli5
ci_vs_shap
lofo_vs_skater
skater_vs_ci
exirt_vs_ci
lofo_vs_shap
dalex_vs_eli5
lofo_vs_ci
dalex_vs_shap
lofo_vs_eli5
exirt_vs_shap
exirt_vs_eli5
exirt_vs_skater
ci_vs_dalex
exirt_vs_lofo


C2 and M4
p-value
skater_vs_dalex
ci_vs_shap
exirt_vs_skater
lofo_vs_dalex
exirt_vs_eli5
ci_vs_eli5
lofo_vs_skater
exirt_vs_ci
dalex_vs_shap
eli5_vs_shap
lofo_vs_ci
skater_vs_shap
exirt_vs_dalex
lofo_vs_shap
skater_vs_ci
dalex_vs_eli5
exirt_vs_shap
ci_vs_dalex
exirt_vs_lofo
lofo_vs_eli5
skater_vs_eli5

Only *eXirt* spearman comparisons of feature relevance ranks for models (M1 to M4) based on cluster 2

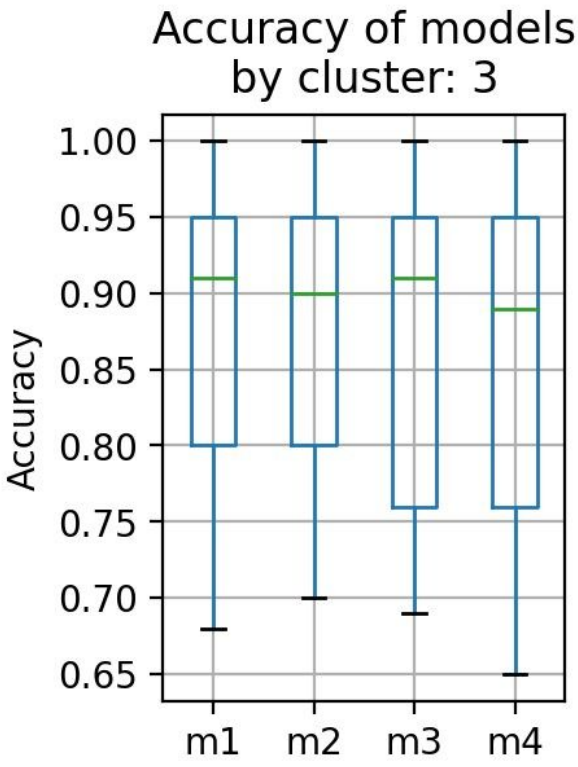

C2 and M1
C2 and M2
C2 and M3
C2 and M4
Spearman Correlation
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
−1.1
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_skater
exirt_vs_shap
exirt_vs_ci
exirt_vs_dalex
exirt_vs_eli5
exirt_vs_ci
exirt_vs_shap
exirt_vs_dalex
exirt_vs_lofo
exirt_vs_skater
exirt_vs_ci
exirt_vs_lofo
exirt_vs_shap
exirt_vs_skater
exirt_vs_eli5
exirt_vs_dalex
exirt_vs_skater
exirt_vs_eli5
exirt_vs_shap
exirt_vs_lofo
exirt_vs_dalex
exirt_vs_ci

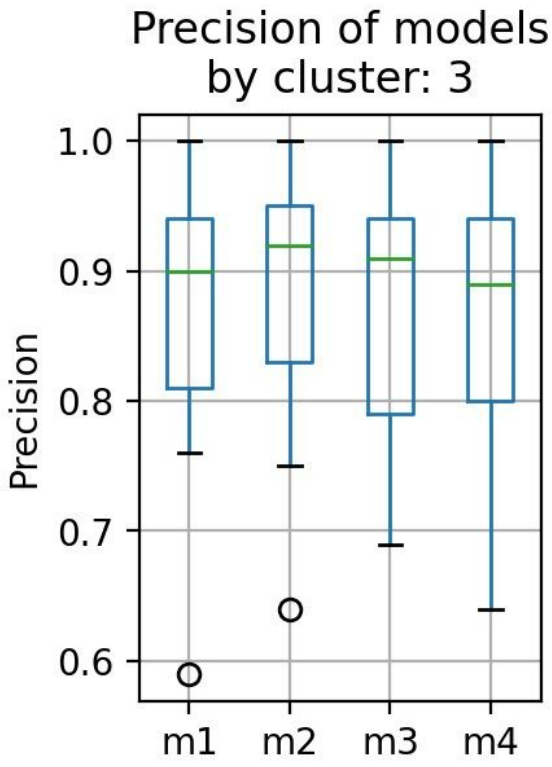
Only *eXirt p-values* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 2

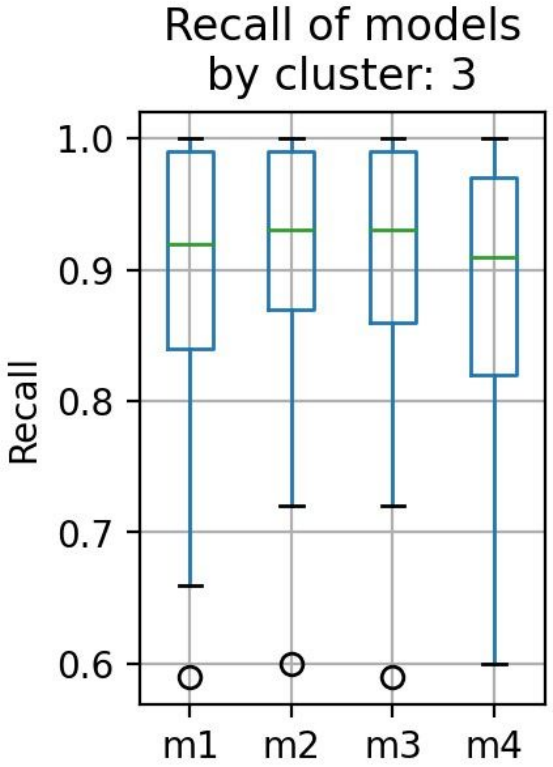

C2 and M1
C2 and M2
C2 and M3
C2 and M4
p-value
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
exirt_vs_shap
exirt_vs_dalex
exirt_vs_skater
exirt_vs_ci
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_eli5
exirt_vs_skater
exirt_vs_shap
exirt_vs_ci
exirt_vs_dalex
exirt_vs_lofo
exirt_vs_dalex
exirt_vs_ci
exirt_vs_shap
exirt_vs_eli5
exirt_vs_skater
exirt_vs_lofo
exirt_vs_skater
exirt_vs_eli5
exirt_vs_ci
exirt_vs_dalex
exirt_vs_shap
exirt_vs_lofo

Cluster 3

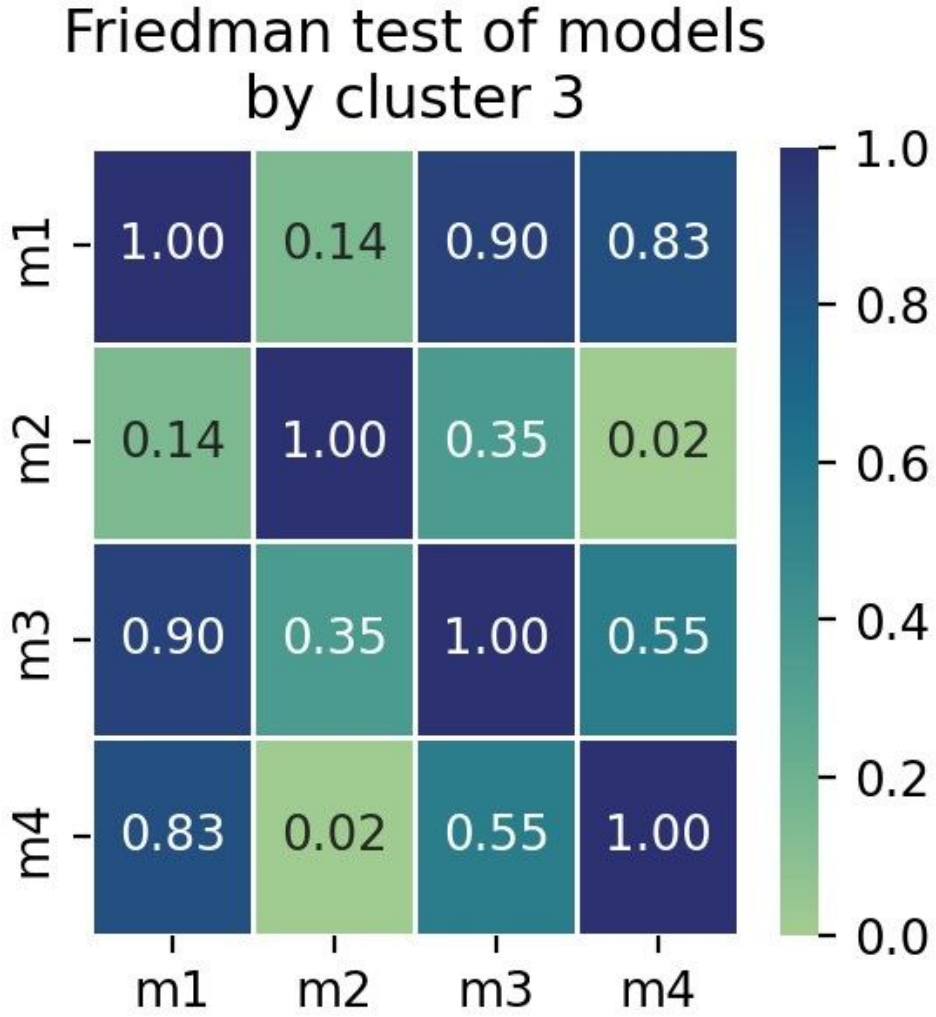
Data regarding the performance of models M1 to M4 based on cluster 3


Accuracy of models
by cluster: 3
Accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
m1
m2
m3
m4

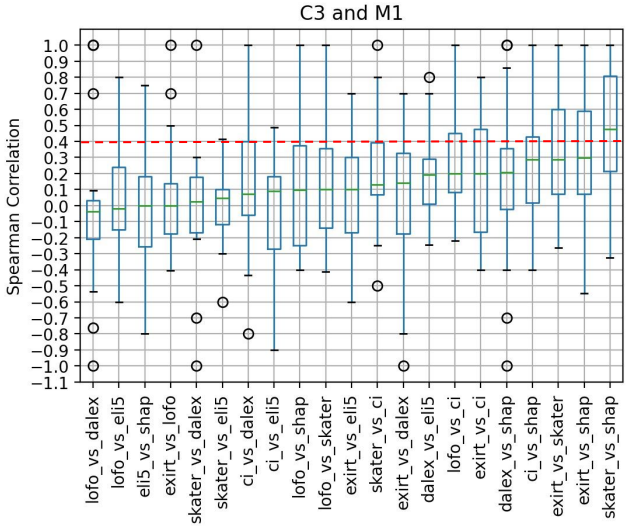

Precision of models
by cluster: 3
Precision
1.0
0.9
0.8
0.7
0.6
m1
m2
m3
m4

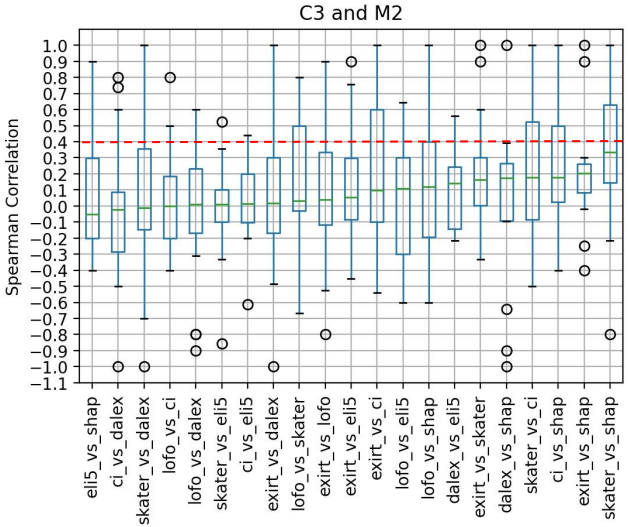

Recall of models
by cluster: 3
Recall
1.0
0.9
0.8
0.7
0.6
m1
m2
m3
m4

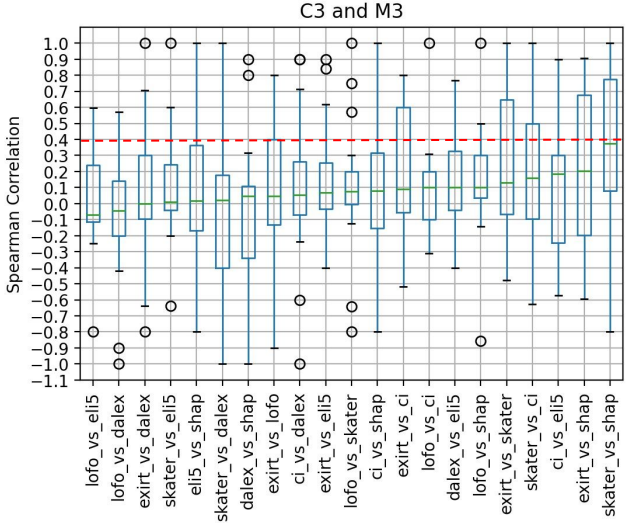
Friedman test result - Models by Cluster 3

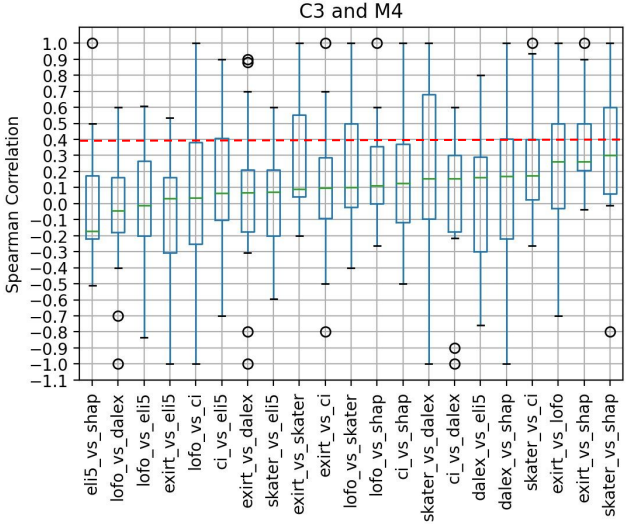

Friedman test of models
by cluster 3
m1
m2
m3
m4
1.00
0.14
0.90
0.83
0.14
1.00
0.35
0.02
0.90
0.35
1.00
0.55
0.83
0.02
0.55
1.00
m1
m2
m3
m4
1.0
0.8
0.6
0.4
0.2
0.0

All pair spearman comparisons of feature relevance ranks for models (M1 to M4) based on cluster 3

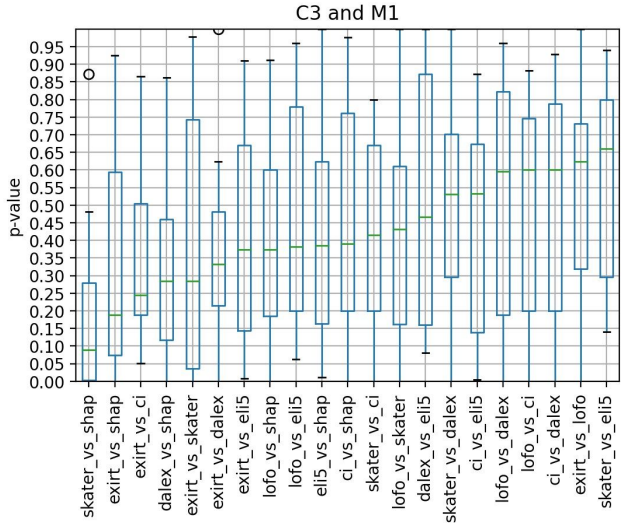

C3 and M1
Spearman Correlation
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0 −0.1 −0.2 −0.3 −0.4 −0.5 −0.6 −0.7 −0.8 −0.9 −1.0 −1.1
lofo_vs_dalex
lofo_vs_eli5
eli5_vs_shap
exirt_vs_lofo
skater_vs_dalex
skater_vs_eli5
ci_vs_dalex
ci_vs_eli5
lofo_vs_shap
lofo_vs_skater
exirt_vs_eli5
skater_vs_ci
exirt_vs_dalex
dalex_vs_eli5
lofo_vs_ci
exirt_vs_ci
dalex_vs_shap
ci_vs_shap
exirt_vs_skater
exirt_vs_shap
skater_vs_shap

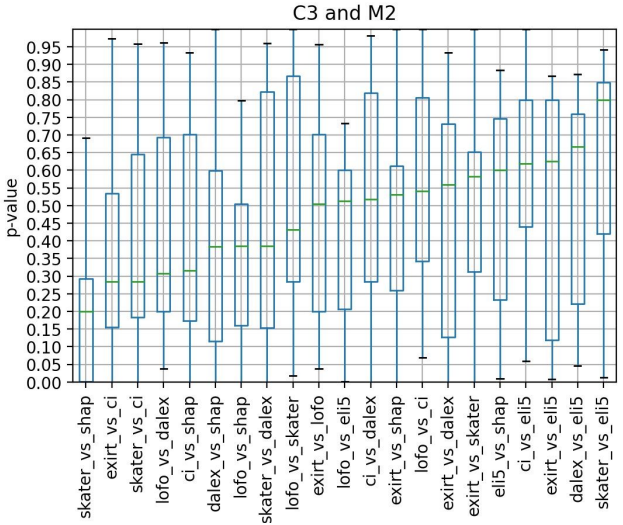

C3 and M2
Spearman Correlation
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0 −0.1 −0.2 −0.3 −0.4 −0.5 −0.6 −0.7 −0.8 −0.9 −1.0 −1.1
eli5_vs_shap
ci_vs_dalex
skater_vs_dalex
lofo_vs_ci
lofo_vs_dalex
skater_vs_eli5
ci_vs_eli5
exirt_vs_dalex
lofo_vs_skater
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_ci
lofo_vs_eli5
lofo_vs_shap
dalex_vs_eli5
exirt_vs_skater
dalex_vs_shap
skater_vs_ci
ci_vs_shap
exirt_vs_shap
skater_vs_shap

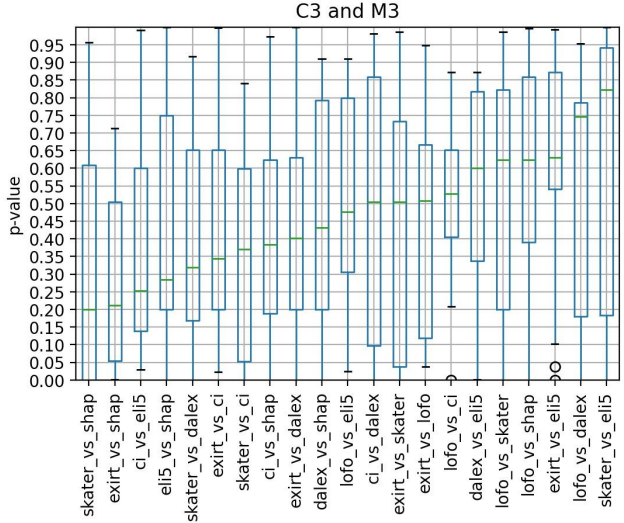

C3 and M3
Spearman Correlation
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0 −0.1 −0.2 −0.3 −0.4 −0.5 −0.6 −0.7 −0.8 −0.9 −1.0 −1.1
lofo_vs_eli5
lofo_vs_dalex
exirt_vs_dalex
skater_vs_eli5
eli5_vs_shap
skater_vs_dalex
dalex_vs_shap
exirt_vs_lofo
ci_vs_dalex
exirt_vs_eli5
lofo_vs_skater
ci_vs_shap
exirt_vs_ci
lofo_vs_ci
dalex_vs_eli5
lofo_vs_shap
exirt_vs_skater
skater_vs_ci
ci_vs_eli5
exirt_vs_shap
skater_vs_shap

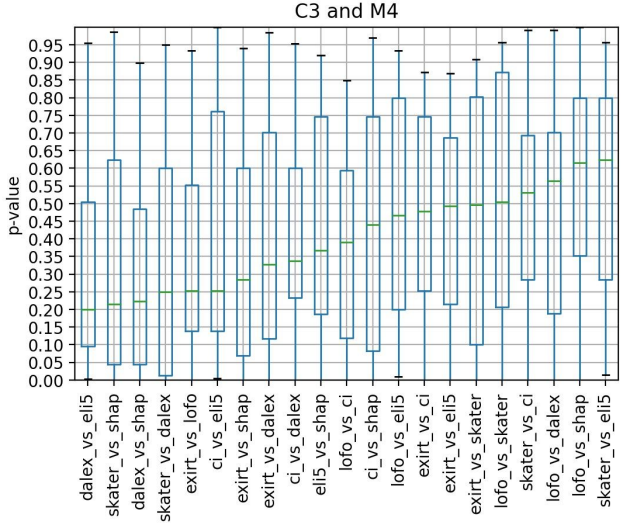

C3 and M4
Spearman Correlation
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0 −0.1 −0.2 −0.3 −0.4 −0.5 −0.6 −0.7 −0.8 −0.9 −1.0 −1.1
eli5_vs_shap
lofo_vs_dalex
lofo_vs_eli5
exirt_vs_eli5
lofo_vs_ci
ci_vs_eli5
exirt_vs_dalex
skater_vs_eli5
exirt_vs_skater
exirt_vs_ci
lofo_vs_skater
lofo_vs_shap
ci_vs_shap
skater_vs_dalex
ci_vs_dalex
dalex_vs_eli5
dalex_vs_shap
skater_vs_ci
exirt_vs_lofo
exirt_vs_shap
skater_vs_shap

All pair *p-values* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 3

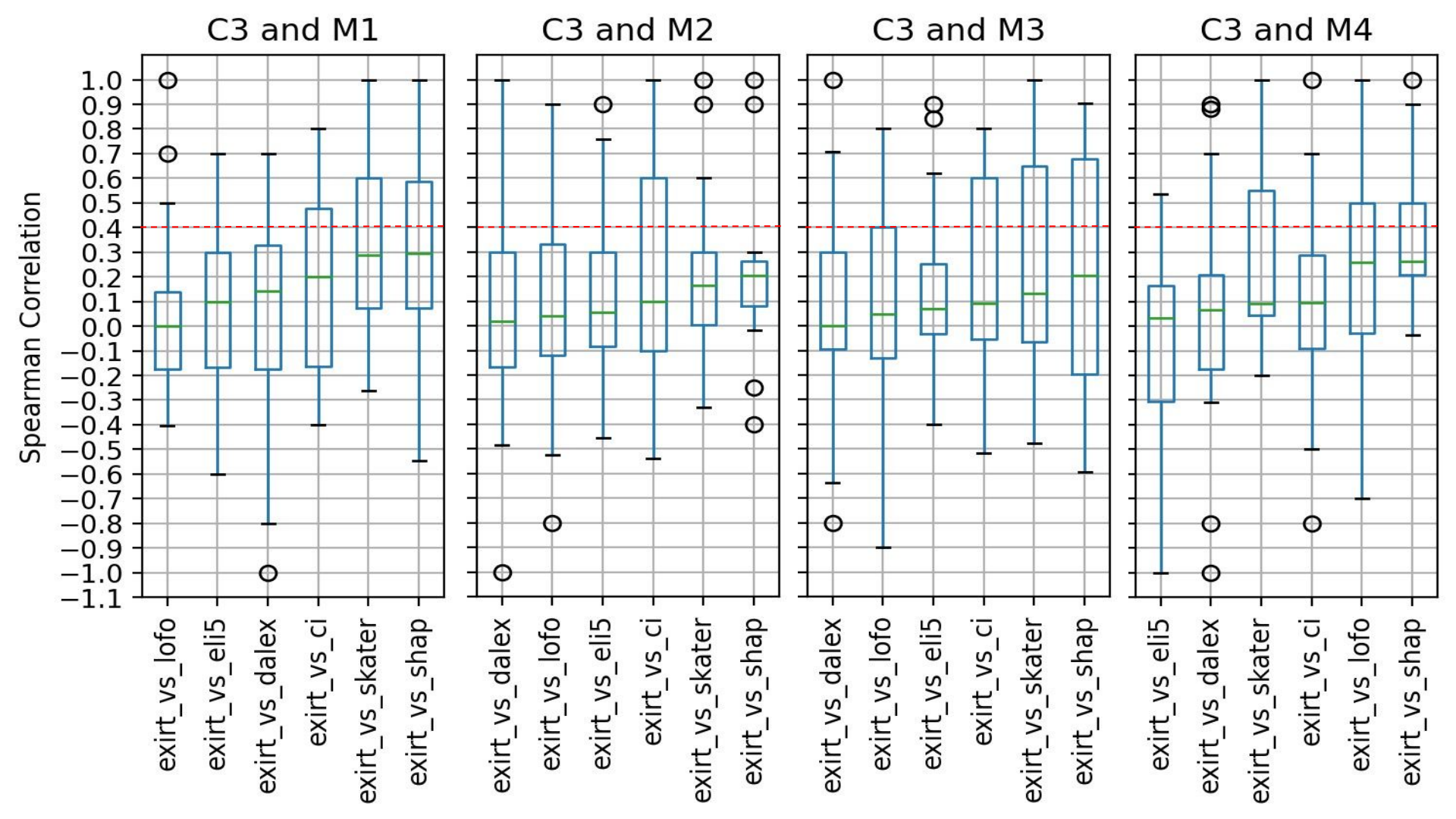

C3 and M1
p-value
skater_vs_shap
exirt_vs_shap
exirt_vs_ci
dalex_vs_shap
exirt_vs_skater
exirt_vs_dalex
exirt_vs_eli5
lofo_vs_shap
lofo_vs_eli5
eli5_vs_shap
ci_vs_shap
skater_vs_ci
lofo_vs_skater
dalex_vs_eli5
skater_vs_dalex
ci_vs_eli5
lofo_vs_dalex
lofo_vs_ci
ci_vs_dalex
exirt_vs_lofo
skater_vs_eli5


C3 and M2
p-value
skater_vs_shap
exirt_vs_ci
skater_vs_ci
lofo_vs_dalex
ci_vs_shap
dalex_vs_shap
lofo_vs_shap
skater_vs_dalex
lofo_vs_skater
exirt_vs_lofo
lofo_vs_eli5
ci_vs_dalex
exirt_vs_shap
lofo_vs_ci
exirt_vs_dalex
exirt_vs_skater
eli5_vs_shap
ci_vs_eli5
exirt_vs_eli5
dalex_vs_eli5
skater_vs_eli5

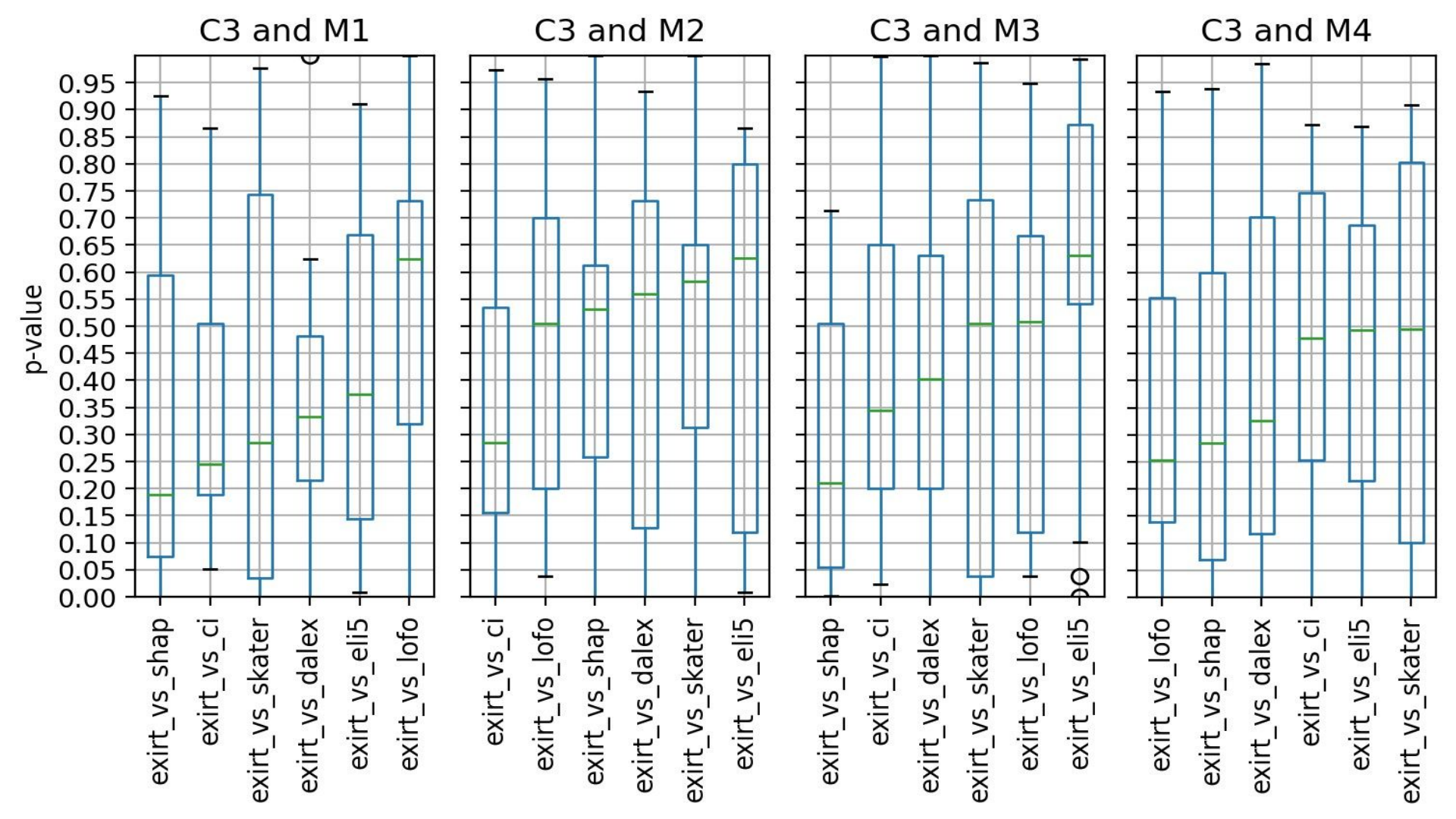

C3 and M3
p-value
skater_vs_shap
exirt_vs_shap
ci_vs_eli5
eli5_vs_shap
skater_vs_dalex
exirt_vs_ci
skater_vs_ci
ci_vs_shap
exirt_vs_dalex
dalex_vs_shap
lofo_vs_eli5
ci_vs_dalex
exirt_vs_skater
exirt_vs_lofo
lofo_vs_ci
dalex_vs_eli5
lofo_vs_skater
lofo_vs_shap
exirt_vs_eli5
lofo_vs_dalex
skater_vs_eli5


C3 and M4
p-value
dalex_vs_eli5
skater_vs_shap
dalex_vs_shap
skater_vs_dalex
exirt_vs_lofo
ci_vs_eli5
exirt_vs_shap
exirt_vs_dalex
ci_vs_dalex
eli5_vs_shap
lofo_vs_ci
ci_vs_shap
lofo_vs_eli5
exirt_vs_ci
exirt_vs_eli5
exirt_vs_skater
lofo_vs_skater
skater_vs_ci
lofo_vs_dalex
lofo_vs_shap
skater_vs_eli5

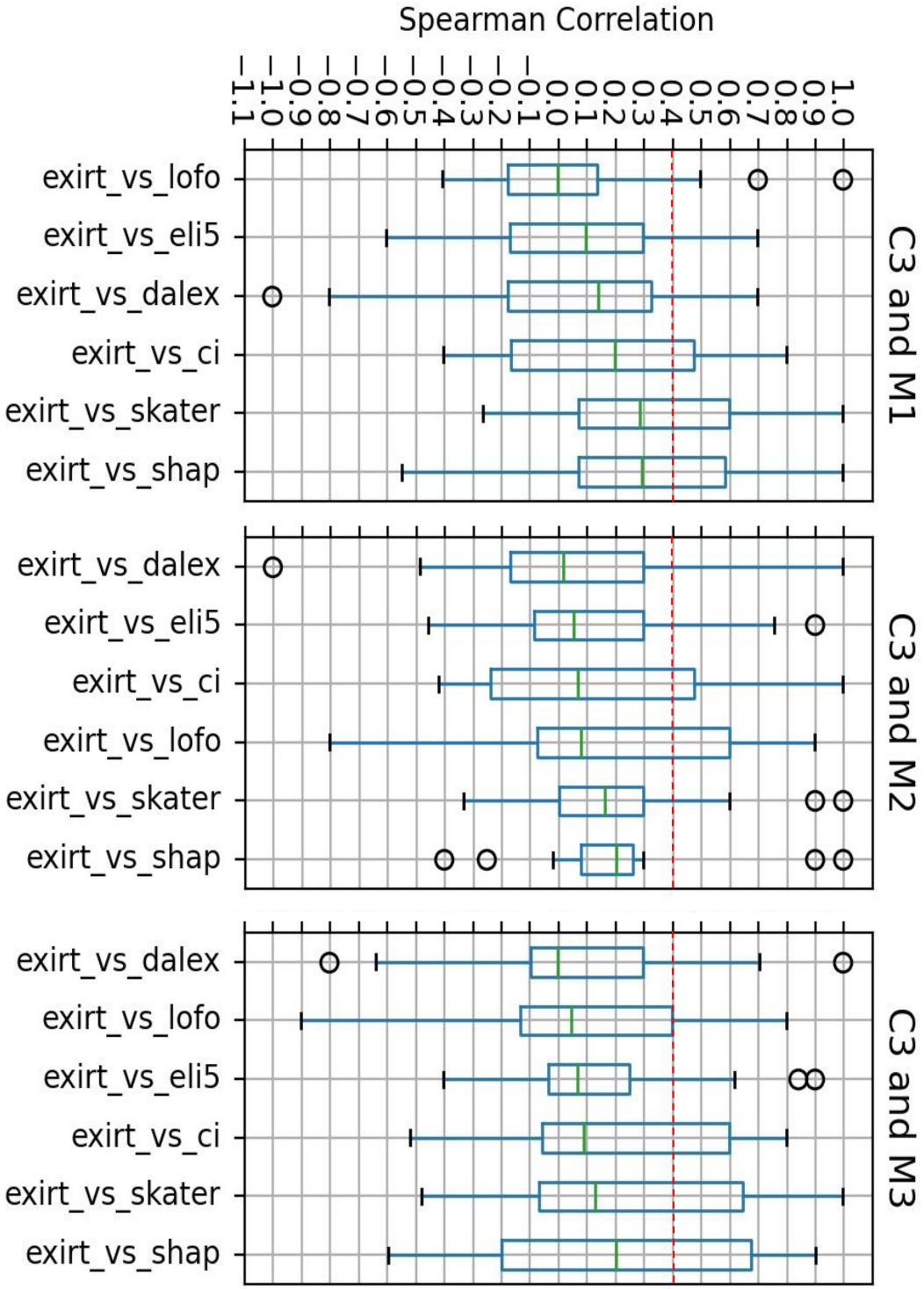
Only *eXirt* spearman comparisons of feature relevance ranks for models (M1 to M4) based on cluster 3

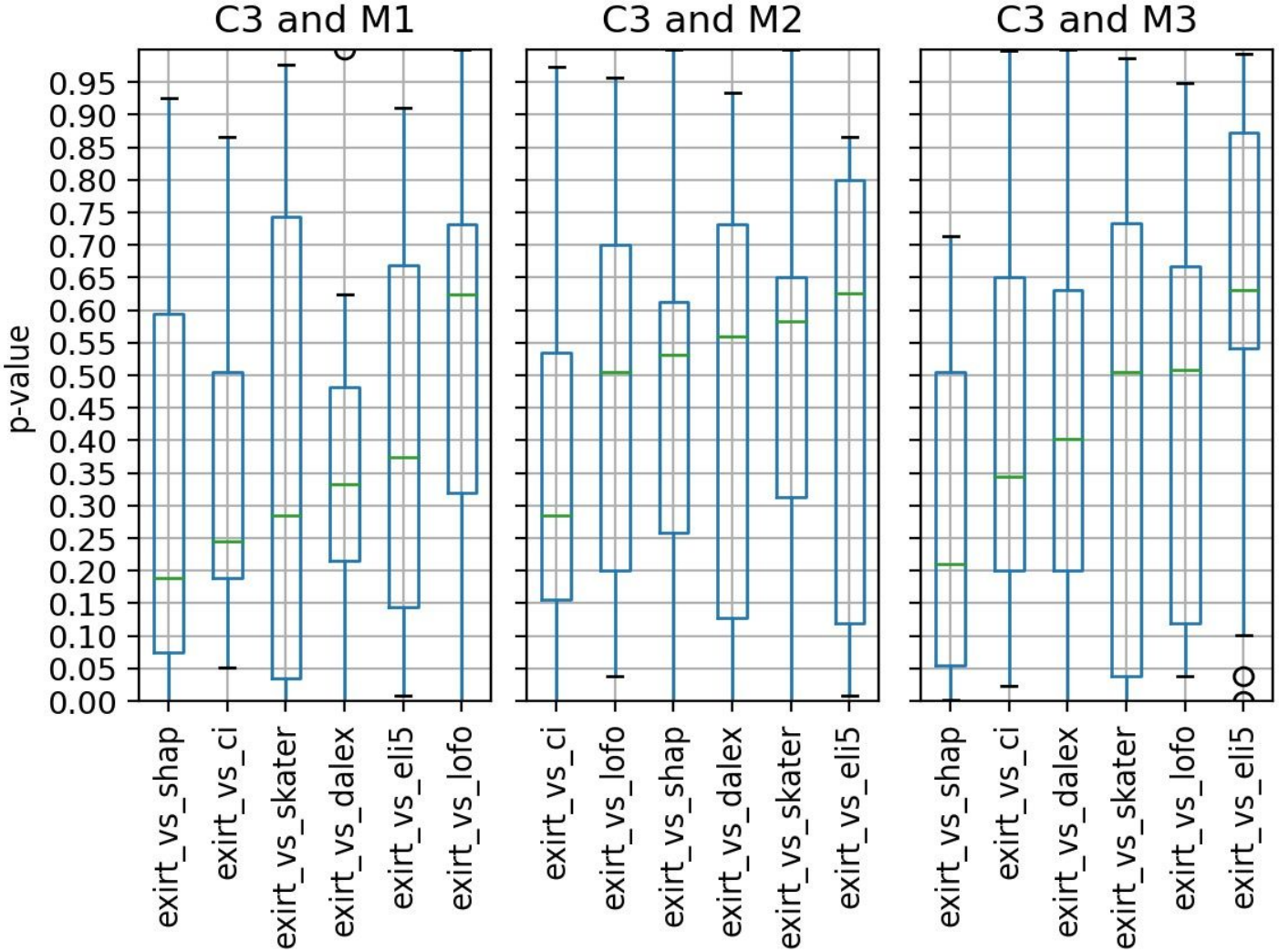

C3 and M1
C3 and M2
C3 and M3
C3 and M4
Spearman Correlation
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
−1.1
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_dalex
exirt_vs_ci
exirt_vs_skater
exirt_vs_shap
exirt_vs_dalex
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_ci
exirt_vs_skater
exirt_vs_shap
exirt_vs_dalex
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_ci
exirt_vs_skater
exirt_vs_shap
exirt_vs_eli5
exirt_vs_dalex
exirt_vs_skater
exirt_vs_ci
exirt_vs_lofo
exirt_vs_shap

Only *eXirt p-values* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 3

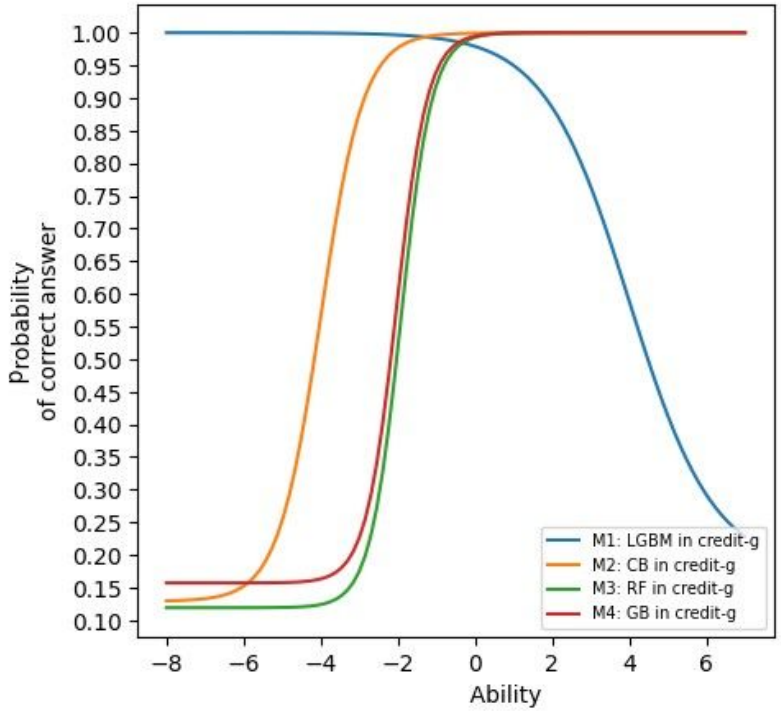

C3 and M1
C3 and M2
C3 and M3
C3 and M4
p-value
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
exirt_vs_shap
exirt_vs_ci
exirt_vs_skater
exirt_vs_dalex
exirt_vs_eli5
exirt_vs_lofo
exirt_vs_ci
exirt_vs_lofo
exirt_vs_shap
exirt_vs_dalex
exirt_vs_skater
exirt_vs_eli5
exirt_vs_shap
exirt_vs_ci
exirt_vs_dalex
exirt_vs_skater
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_lofo
exirt_vs_shap
exirt_vs_dalex
exirt_vs_ci
exirt_vs_eli5
exirt_vs_skater

Only e*Xirt* comparisons of feature relevance ranks for models (M1, M2, and M3) based on cluster 3

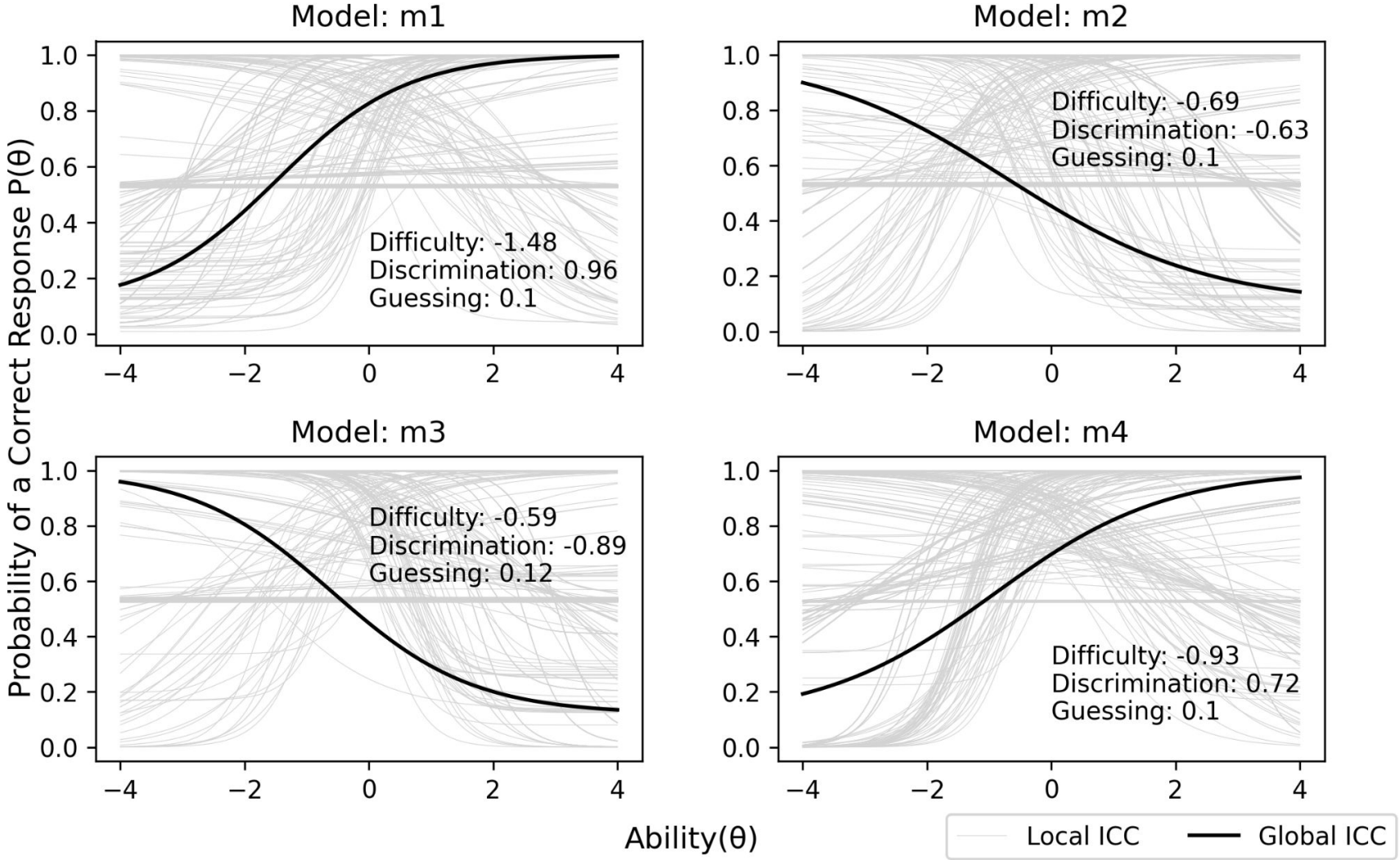

Spearman Correlation
−1.1
−1.0
−0.9
−0.8
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_dalex
exirt_vs_ci
exirt_vs_skater
exirt_vs_shap
C3 and M1
exirt_vs_dalex
exirt_vs_eli5
exirt_vs_ci
exirt_vs_lofo
exirt_vs_skater
exirt_vs_shap
C3 and M2
exirt_vs_dalex
exirt_vs_lofo
exirt_vs_eli5
exirt_vs_ci
exirt_vs_skater
exirt_vs_shap
C3 and M3


C3 and M1
C3 and M2
C3 and M3
p-value
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
exirt_vs_shap
exirt_vs_ci
exirt_vs_skater
exirt_vs_dalex
exirt_vs_eli5
exirt_vs_lofo
exirt_vs_ci
exirt_vs_lofo
exirt_vs_shap
exirt_vs_dalex
exirt_vs_skater
exirt_vs_eli5
exirt_vs_shap
exirt_vs_ci
exirt_vs_dalex
exirt_vs_skater
exirt_vs_lofo
exirt_vs_eli5

Item Characteristic Curve of the credit-g dataset (test) for models M1 to M4


1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
Probability
of correct answer
−8
−6
−4
−2
0
2
4
6
Ability
M1: LGBM in credit-g
M2: CB in credit-g
M3: RF in credit-g
M4: GB in credit-g

Mean Item Characteristic Curve of the credit-g dataset (test) for models M1 to M4


Model: m1
Difficulty: -1.48
Discrimination: 0.96
Guessing: 0.1
Model: m2
Difficulty: -0.69
Discrimination: -0.63
Guessing: 0.1
Model: m3
Difficulty: -0.59
Discrimination: -0.89
Guessing: 0.12
Model: m4
Difficulty: -0.93
Discrimination: 0.72
Guessing: 0.1
Probability of a Correct Response P(θ)
Ability(θ)
Local ICC
Global ICC