Supplementary Material: Faster Convergence with Lexicase Selection in Tree-based Automated Machine Learning

Nicholas Matsumoto, Anil Kumar Saini, Pedro Ribeiro, Hyunjun Choi, Alena Orlenko, Leo-Pekka Lyytikäinen, Jari O Laurikka, Terho Lehtimäki, Sandra Batista, and Jason H. Moore

Cedars-Sinai Medical Center, Tampere University, and Sydänsairaala Hospital

1 Number of Operators at Convergence

Table 1. Average number of operators in the pipelines at convergence points for various selection methods on different datasets. The third column shows the p-values obtained by applying the Mann-Whitney-U test on the distribution of values from both conditions in the corresponding column.

Dataset	NSGA-II	Lexicase	p-value
ANGES	3.66	2.84	7.252E-05
Digen-2	2.95	2.70	1.077E-01
Digen-4	3.18	2.85	1.055E-01
Digen-7	3.28	2.53	3.585E-04
Digen-14	3.05	2.73	1.466E-01
Digen-23	3.38	2.93	4.501E-02
Digen-24	2.63	2.13	1.515E-03
Digen-25	3.00	2.58	1.700E-02
Digen-27	2.33	2.18	3.438E-01
Digen-28	3.05	2.70	7.957E-02
Digen-30	3.30	2.80	3.293E-03
Digen-32	2.30	2.35	8.815E-01
Digen-35	2.35	2.43	3.465E-01
Digen-40	2.48	2.20	5.837E-02

2 Exploration Trie Metrics

Table 2. Metrics for NGSA-II and lexicase exploration tries on DIGEN and ANGES datasets at the final generation: (a) Mean nodal global efficiency with standard deviation (b) Mean total number of trie nodes with standard deviation (c) Mean leaf-to-node ratio with standard deviation.

Dataset	NSGA-II (a)	Lexicase (a)	NSGA-II (b)	Lexicase (b)	NSGA-II (c)	Lexicase (c)
ANGES	0.26 ± 0.02	0.29 ± 0.04	2791.9 ± 615.7	1584.3 ± 708.2	0.47 ± 0.05	0.58 ± 0.07
DIGEN-2	0.37 ± 0.02	0.41 ± 0.03	419.9 ± 65.0	292.4 ± 87.2	0.57 ± 0.04	0.67 ± 0.06
DIGEN-4	0.37 ± 0.02	0.4 ± 0.04	405.7 ± 54.4	293.5 ± 87.9	0.58 ± 0.04	0.66 ± 0.06
DIGEN-7	0.37 ± 0.02	0.41 ± 0.03	411.3 ± 66.3	263.2 ± 77.9	0.57 ± 0.04	0.68 ± 0.05
DIGEN-14	0.37 ± 0.02	0.41 ± 0.03	406.0 ± 66.3	276.8 ± 72.7	0.58 ± 0.03	0.67 ± 0.04
DIGEN-23	0.37 ± 0.02	0.41 ± 0.04	414.1 ± 59.9	275.1 ± 80.0	0.57 ± 0.04	0.67 ± 0.05
DIGEN-24	0.37 ± 0.02	0.42 ± 0.04	407.0 ± 60.3	250.6 ± 100.0	0.58 ± 0.03	0.67 ± 0.05
DIGEN-25	0.37 ± 0.02	0.41 ± 0.04	405.8 ± 51.1	269.5 ± 93.3	0.57 ± 0.03	0.68 ± 0.06
DIGEN-27	0.38 ± 0.02	0.43 ± 0.03	383.4 ± 57.9	224.2 ± 56.4	0.59 ± 0.04	0.69 ± 0.04
DIGEN-28	0.37 ± 0.02	0.42 ± 0.03	405.3 ± 67.2	260.0 ± 74.4	0.58 ± 0.04	0.68 ± 0.05
DIGEN-30	0.38 ± 0.02	0.41 ± 0.03	397.5 ± 63.1	284.4 ± 78.7	0.58 ± 0.03	0.67 ± 0.05
DIGEN-32	0.38 ± 0.02	0.42 ± 0.03	401.8 ± 55.5	254.2 ± 60.5	0.59 ± 0.03	0.68 ± 0.05
DIGEN-35	0.38 ± 0.01	0.42 ± 0.04	393.0 ± 48.9	254.3 ± 79.6	0.59 ± 0.02	0.68 ± 0.05
DIGEN-40	0.37 ± 0.02	0.43 ± 0.03	410.2 ± 57.0	228.4 ± 72.2	0.58 ± 0.03	0.69 ± 0.05

3 Feature Importance on ANGES Dataset

Permutation feature importance (PFI) analysis was conducted to compare the features that the best models for lexicase and NSGA-II found to affect accuracy most on the ANGES data. While PFI can change across GP runs and different initialization conditions, it provides useful comparison for the the impact of features on model accuracy. PFI rank showed that lexicase selected as the three most informative features known clinical risk factors for many cardiovascular diseases: myocardiac infarction signs, age and sex. Age, however ranked outside the top 10 informative features for the NGSA-II model. Administration of nitrate and statin medications appeared among top 10 informative features for NGSA-II and lexicase and glucose is a top informative feature for lexicase. All three features are known cardiovascular risk factors. Various lipid compounds that are potent predictors of cardiovascular disease were found among the top informative features of both models: LDL.TG in the best model for NGSA-II and VLDL.D and total fatty acid estimate in the best model for lexicase. Since the mechanistic relationships between clinical biomarkers and complex health endpoints, such as cardiovascular disease, are typically unknown, and since PFI only gives relative importance of features within a model, we cannot ascertain the most accurate features for cardiovascular disease However, both models selected features in

their models that are known well-supported clinical risk factors or predictors for cardiovascular disease. The top 10 features of the best models for each selection method and their affect on the model accuracy are given in Supplemental Figure 3 .

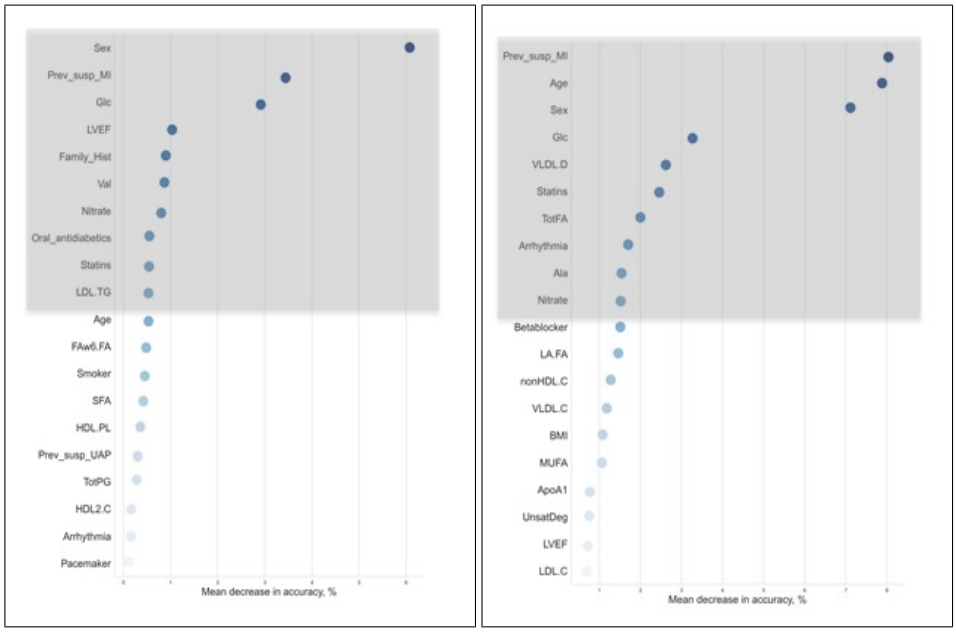

Fig. 1. Top 10 features in the best models for NSGA-II (left) and lexicase (right).

4 Viewing the Exploration Trie

The exploration trie can be viewed through the networkx library to display the data structure as a graph. There is a root node labeled 'NA' as a grey node. This is where all pipelines start. The exploration trie will branch through the space as the evolutionary algorithm provides new pipelines. The node label indicates the which machine learning algorithm is used, and the color indicates the performance of all pipelines that traverse through the node. Red is poor performance, and green is good performance.

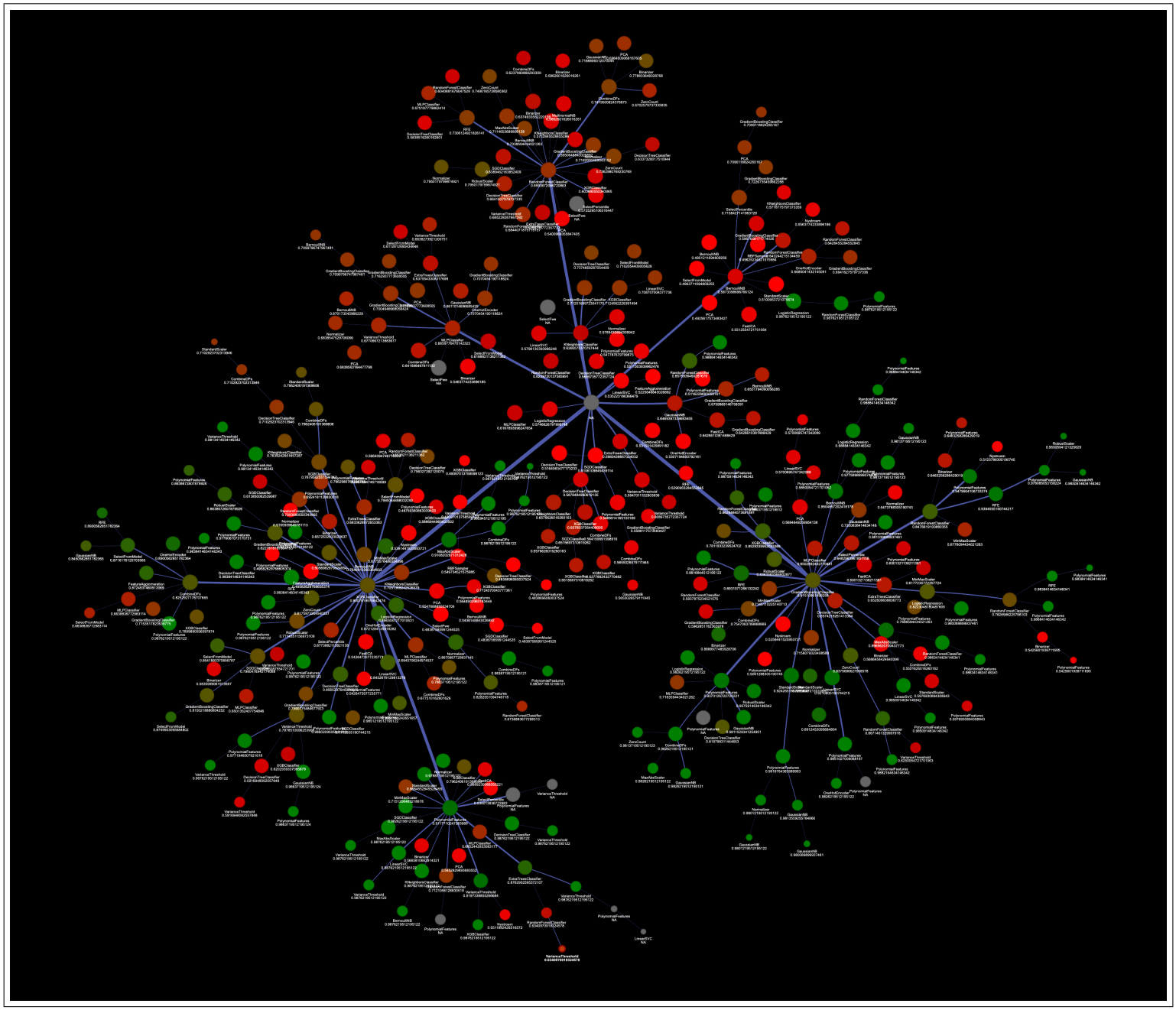

Fig. 2. Exploration Trie for DIGEN-24 using NSGA-II. The gray node in the middle is the root node. Nodal global efficiency: 0.396. Total number of trie nodes: 352. Leaf-to-node ratio: 0.599.

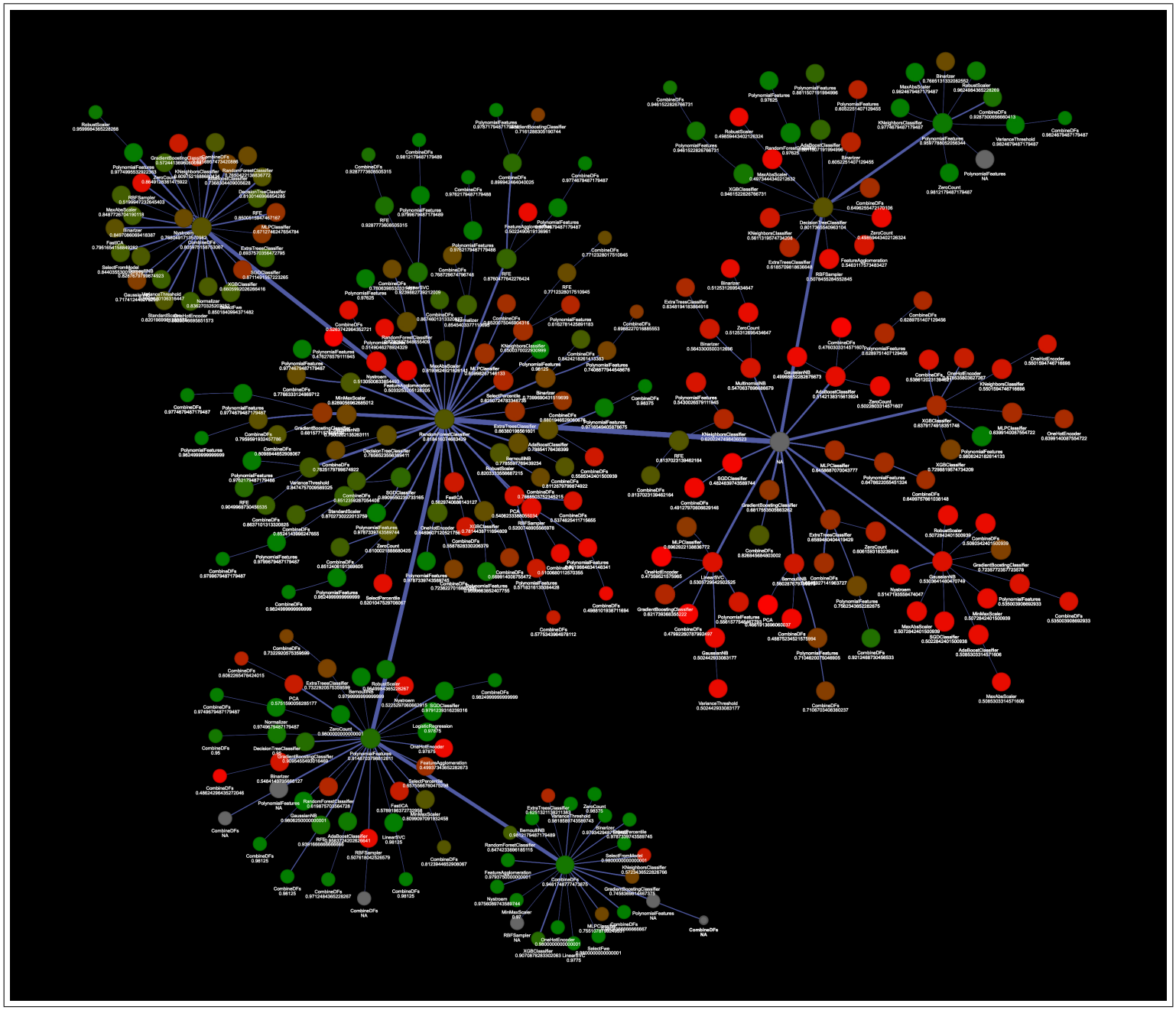

Fig. 3. Exploration Trie for DIGEN-24 using Lexicase. The gray node towards the center right is the root node. Nodal global efficiency: 0.399. Total number of trie nodes: 271. Leaf-to-node ratio: 0.657.