Online Appendix for "Leveraging Model-based Trees as Interpretable Surrogate Models for Model Distillation"

Julia Herbinger*[1,2][0000−0003−0430−8523], Susanne Dandl*[1,2][0000−0003−4324−4163], Fiona K. Ewald[1,2][0009−0002−6372−3401], Sofia Loibl[1], and Giuseppe Casalicchio ✉[1,2][0000−0001−5324−5966]

[1] Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany
[2] Munich Center for Machine Learning (MCML), Munich, Germany
Giuseppe.Casalicchio@stat.uni-muenchen.de

A Details on Quantifying the Rand index

The detailed steps for the quantification of the RI in the simulations in Section 4.2 can be described as follows:

1. Simulate evaluation data (50000 observations) from the DGP
2. **For each** simulation run in $1:100$ runs:
   (a) Simulate data ($n = 1500$) and perform train/test split ($\frac{2}{3}/\frac{1}{3}$)
   (b) Train MBT on the training data, calculate fidelity measures on the train and test set, and extract the number of leaf nodes
   (c) save the partitioning of the evaluation data defined through the trained MBT
3. **For each** of the $(100(100-1)/2 = 4950$ MBT pairs
   (a) Sample 1000 observations from the evaluation data sets
   (b) **If** both trees have the same number of leaf nodes, calculate the RI for the two partitions of the sampled evaluation data subset

B More Results and Details on Experiments

Here, we provide more details on the experiments described in Section 4.2. Therefore, details on model configurations are described as well as further results and detailed analyses are provided.

B.1 Hyperparameter Configurations

In the experiments of Section 4.2, an XGBoost algorithm was used as a black box model (besides a correctly specified lm or GAM) with correctly specified interaction terms. In Table 1 the XGBoost hyperparameter configurations, which were used for the simulations of Section 4.2, are defined.

* These authors contributed equally to this work.

Table 1. XGBoost hyperparameter configurations for the three scenarios linear smooth, linear categorical, and linear mixed of the experiments in Section 4.2.

	linear smooth	linear categorical	linear mixed
max_depth	5	3	5
eta	0.5	0.5	0.5
alpha	1	0.5	2
gamma	2	1	3.5
nrounds	400	350	500

B.2 Experimental Results: Linear Smooth

This section provides more detailed results on the scenario linear smooth described in Section 4.2. A similar mean number of leaves for all MBTs are given by $\gamma = \alpha = 0.05$. Thus, the analysis of stability for the lm (Figure 1) is based on the results for these configurations to enable comparability between the different algorithms.

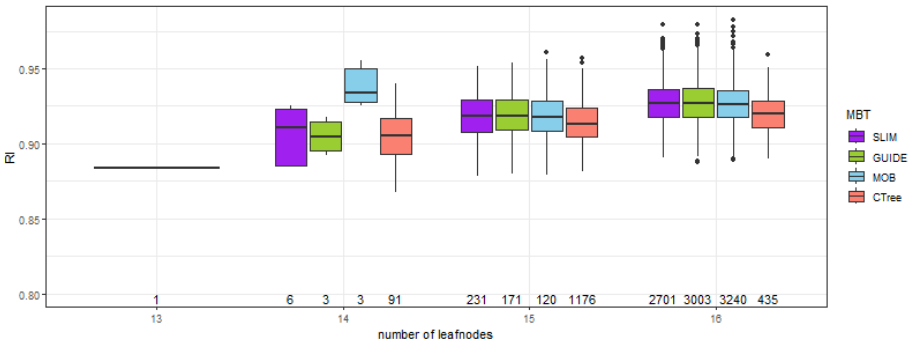

Fig. 1. RI for the four MBT algorithms when used as a surrogate model for an lm model for the scenario linear smooth with $\alpha = \gamma = 0.05$. The numbers below the boxplots indicate the number of tree pairs, for which both trees have the respective number of leaf nodes.

B.3 Experimental Results: Linear Categorical

This section provides more detailed results on the scenario linear categorical described in Section 4.2. Aggregated results on interpretability and fidelity are provided in Table 2. The analysis of fidelity for the lm (Figure 1) is based on the configuration $\gamma = \alpha = 0.05$. The results coincide with the results of the XGBoost model (Figure 1). Hence, MOB and CTree lead to a worse fidelity and due to many more leaf nodes to a worse interpretability than SLIM and GUIDE for this scenario.

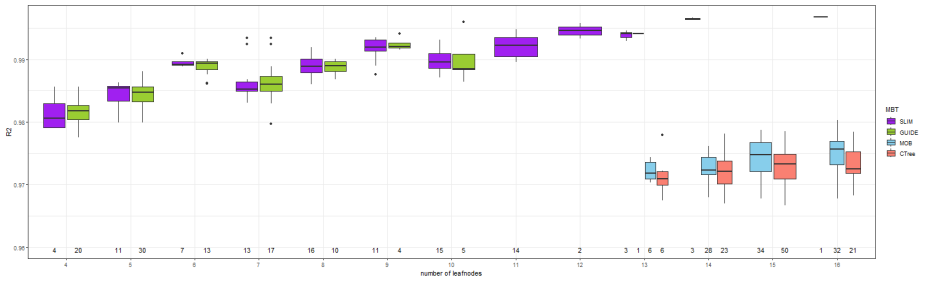

Fig. 2. Test accuracy R^2 vs. number of leaf nodes for the four MBT algorithms as surrogate models for lm for scenario linear categorical with $n_{train} = 1000$ and $n_{test} = 500$, $\alpha = \gamma = 0.05$. The numbers below the boxplots indicate the number of trees (from 100 simulation runs) which have the respective number of leaf nodes for the regarded algorithm.

B.4 Experimental Results: Linear Mixed

This section provides more detailed results on the scenario linear mixed described in Section 4.2. Aggregated results on interpretability and fidelity are provided for all four MBT algorithms in Table 3.

For $\alpha = \gamma = 0.05$ the mean number of leaf nodes is similar for all four MBT algorithms. Thus, these configurations are chosen for the analyses in Section 4.2 to enable comparability between the different MBT algorithms. Table 3 shows a slightly higher fidelity for SLIM and GUIDE compared to MOB and CTree for this configuration.

B.5 Non-linear effects

While the scenario definitions so far were only based on linear main effects, the following scenario will investigate how the choice of different objective functions affects the interpretability and fidelity of SLIM MBTs when non-linear main effects are included in the DGP. Since [2] provide flexible options to efficiently estimate non-linear main effects in the leaf nodes, only the SLIM algorithm is used.

Scenario definition Let $X_1, \ldots, X_5 \sim \mathcal{U}(-1, 1)$, $\mathbf{x}_6 \sim Bern(0.5)$ and the DGP based on n realizations is defined by $y = f(\mathbf{x}) + \epsilon$ with $f(\mathbf{x}) = \mathbf{x}_1 + 2\mathbf{x}_2^2 + \mathbf{x}_3 log(abs(\mathbf{x}_3)) + \mathbf{x}_4\mathbf{x}_5 + \mathbf{x}_1\mathbf{x}_4\mathbb{1}_{(\mathbf{x}_6=0)}$ and $\epsilon \sim \mathcal{N}(0, 0.01 \cdot \sigma^2(f(x)))$.

SLIM is fitted with the following four different model types in the leaf nodes:

1. Linear regression model (lm)
2. Polynomial regression of degree 2 with lasso penalization (penalized poly)
3. Linear regression with unpenalized linear B-spline transformations of the features (B-splines)
4. GAMs with integrated smoothness estimation [5]

Table 2. Simulation results on 100 simulation runs for all four MBTs on scenario linear categorical with sample sizes $n_{train} = 1000$ and $n_{test} = 500$ for different values of γ and α. The mean (standard deviation) fidelity on the training data for the GAM is 0.9702 (0.0018) and for the XGBoost is 0.9876 (0.0015). On the test data set the respective fidelity values for the GAM are 0.9694 (0.0029) and for the XGBoost are 0.9778 (0.0031).

Black box	MBT	γ/α	Number of Leaves			R^2_{train}		R^2_{test}		Share
			mean	min	max	mean	sd	mean	sd	x_2
xgboost	SLIM	0.15	2.00	2	2	0.8321	0.8321	0.8323	0.8323	0.0000
xgboost	GUIDE	0.15	2.00	2	2	0.8321	0.8321	0.8323	0.8323	0.0000
xgboost	MOB	0.001	13.45	11	16	0.9793	0.9793	0.9729	0.9729	0.8865
xgboost	CTree	0.001	11.96	10	14	0.9602	0.9602	0.9545	0.9545	0.9914
xgboost	SLIM	0.10	4.00	4	4	0.9923	0.9923	0.9870	0.9870	0.0000
xgboost	GUIDE	0.10	4.00	4	4	0.9923	0.9923	0.9870	0.9870	0.0000
xgboost	MOB	0.010	14.38	13	16	0.9831	0.9831	0.9765	0.9765	0.8656
xgboost	CTree	0.010	12.76	10	15	0.9612	0.9612	0.9550	0.9550	0.9897
xgboost	SLIM	0.05	4.00	4	4	0.9923	0.9923	0.9870	0.9870	0.0000
xgboost	GUIDE	0.05	4.00	4	4	0.9923	0.9923	0.9870	0.9870	0.0000
xgboost	MOB	0.050	14.63	13	16	0.9837	0.9837	0.9771	0.9771	0.8614
xgboost	CTree	0.050	13.46	10	16	0.9623	0.9623	0.9558	0.9558	0.9838
gam	SLIM	0.15	2.00	2	2	0.8528	0.8528	0.8513	0.8513	0.0000
gam	GUIDE	0.15	2.00	2	2	0.8528	0.8528	0.8513	0.8513	0.0000
gam	MOB	0.001	13.53	11	15	0.9773	0.9773	0.9718	0.9718	0.9824
gam	CTree	0.001	13.89	11	16	0.9773	0.9773	0.9720	0.9720	0.9973
gam	SLIM	0.10	2.64	2	4	0.8972	0.8972	0.8937	0.8937	0.0000
gam	GUIDE	0.10	2.64	2	4	0.8972	0.8972	0.8937	0.8937	0.0000
gam	MOB	0.010	14.28	13	16	0.9784	0.9784	0.9728	0.9728	0.9721
gam	CTree	0.010	14.47	12	16	0.9779	0.9779	0.9725	0.9725	0.9949
gam	SLIM	0.05	8.56	4	16	0.9910	0.9910	0.9893	0.9893	0.0017
gam	GUIDE	0.05	6.06	4	13	0.9875	0.9875	0.9859	0.9859	0.0084
gam	MOB	0.050	14.92	13	16	0.9797	0.9797	0.9740	0.9740	0.9572
gam	CTree	0.050	14.86	13	16	0.9783	0.9783	0.9729	0.9729	0.9925

Since the models fitted in the leaf nodes vary in complexity and thus interpretability, it is not sufficient to consider solely the number of leaf nodes as a measure of interpretability. In addition, the following criteria are analyzed:

- Effective degrees of freedoms of the leaf node models (lm and penalized poly), as sparse models are easier to interpret
- Proportion of splits for which features are chosen that are involved in feature interactions (in the DGP) vs. proportion of splits for which features are selected that only contain non-linear main effects.
- Interpretable formulas vs. only visual interpretability

In order to enable a comparison of the interpretability, the R^2 is used as an early stopping parameter in this scenario in addition to the early stopping procedure with γ. As soon as the R^2 exceeds a certain value in a node, no further

Table 3. Simulation results on 100 simulation runs for all four MBTs on scenario linear mixed with sample sizes $n_{train} = 1000$ and $n_{test} = 500$ for different values of γ and α. The mean (standard deviation) fidelity on the training data for the lm is 0.9902 (0.0006) and for the XGBoost is 0.9859 (0.0014). On the test data set the respective fidelity values for the lm are 0.9898 (0.0008) and for the XGBoost are 0.9682 (0.0042). The column "Share" defines the relative number of partitioning steps which used the numeric features $\mathbf{x}_1$ or $\mathbf{x}_2$ for splitting.

Black box	MBT	γ/α	Number of Leaves			R^2_{train}		R^2_{test}		Share
			mean	min	max	mean	sd	mean	sd	x_1 or x_2
xgboost	SLIM	0.15	4.47	2	13	0.9067	0.0336	0.9013	0.0339	0.9486
xgboost	GUIDE	0.15	4.37	2	13	0.9059	0.0335	0.9005	0.0339	0.9453
xgboost	MOB	0.001	14.82	13	17	0.9853	0.0018	0.9745	0.0047	0.9735
xgboost	CTree	0.001	15.03	13	17	0.9850	0.0017	0.9743	0.0042	0.9949
xgboost	SLIM	0.10	12.80	7	16	0.9832	0.0089	0.9724	0.0103	0.9044
xgboost	GUIDE	0.10	12.48	6	16	0.9822	0.0098	0.9715	0.0112	0.8737
xgboost	MOB	0.010	14.94	13	17	0.9854	0.0017	0.9746	0.0046	0.9727
xgboost	CTree	0.010	15.07	13	17	0.9850	0.0017	0.9743	0.0042	0.9947
xgboost	SLIM	0.05	14.80	12	17	0.9870	0.0018	0.9764	0.0044	0.9068
xgboost	GUIDE	0.05	14.47	12	17	0.9863	0.0022	0.9758	0.0047	0.8683
xgboost	MOB	0.050	14.94	13	17	0.9854	0.0017	0.9746	0.0046	0.9727
xgboost	CTree	0.050	15.07	13	17	0.9850	0.0017	0.9743	0.0042	0.9947
lm	SLIM	0.15	3.20	2	13	0.8879	0.0309	0.8806	0.0331	0.9705
lm	GUIDE	0.15	3.17	2	13	0.8872	0.0308	0.8799	0.0329	0.9707
lm	MOB	0.001	14.99	13	17	0.9882	0.0016	0.9838	0.0021	0.9637
lm	CTree	0.001	15.05	13	17	0.9880	0.0016	0.9841	0.0019	0.9994
lm	SLIM	0.10	13.07	5	16	0.9875	0.0098	0.9843	0.0108	0.8766
lm	GUIDE	0.10	12.66	7	16	0.9866	0.0095	0.9834	0.0106	0.8676
lm	MOB	0.010	14.99	13	17	0.9882	0.0016	0.9838	0.0021	0.9637
lm	CTree	0.010	15.05	13	17	0.9880	0.0016	0.9841	0.0019	0.9994
lm	SLIM	0.05	14.78	12	16	0.9913	0.0020	0.9885	0.0028	0.8723
lm	GUIDE	0.05	14.38	12	16	0.9905	0.0022	0.9876	0.0029	0.8611
lm	MOB	0.050	14.99	13	17	0.9882	0.0016	0.9838	0.0021	0.9637
lm	CTree	0.050	15.05	13	17	0.9880	0.0016	0.9841	0.0019	0.9994

split for this node is performed. We choose $\gamma = 0.05$ and an R^2 value of 0.9 for as early stopping configurations for the described simulation setting.

We simulate $n = 3000$ observations (2000 for training and 1000 testing). The SLIM MBTs are fitted as a surrogate model on the predictions of an XGBoost model. The hyperparameter configurations of the XGBoost model are defined in Table 4.

Results Table 5 provides an overview of the simulation results regarding the interpretability measures for the different model types used within the SLIM MBTs. For SLIM with linear regression models in the nodes, this results in large trees which provide low interpretability due to the high number of leaf nodes. In addition, on average 61% of all partitioning steps do not use a feature involved

Table 4. Tuned hyperparameter configurations for the XGBoost algorithm with correctly specified interaction effects for the scenario non-linear.

	non-linear
max_depth	4
eta	0.825
alpha	0.75
gamma	1
nrounds	700

in an interaction for splitting, but split based on an insufficiently modeled main effect. Moreover, a high number of features is used for splitting which further reduces interpretability. The advantage is that the models fitted in the leaf nodes provide an interpretable formula and not only a visual component to interpret the final results.

Using the second model type, meaning penalized polynomial regression, in the leave nodes when applying SLIM, leads to a comparable fidelity as using an lm but with fewer splits. Thus, interpretability increases in the sense that the number of leaf nodes drops. Also, the number of different split features is reduced, which again increases interpretability. While the proportion of partitioning steps that use main effect features for splitting is smaller, there is still more than one-third of splits that are not performed due to feature interactions.

Both, SLIM with linear B-Spline transformed features and GAMs require on average only two leaf nodes, i.e. one split, to achieve an R^2 of 0.9. Thus, with regard to the number of leaf nodes the interpretability highly increases. However, the models in the leaf nodes can only be interpreted visually, and no interpretable formula is provided. Since the number of models (2) is very small, the degree of interpretability is comparatively high. Moreover, these models actually only split by interactions, as the non-linear main effects are already modeled sufficiently well.

If an interpretable formula is not explicitly required and a visual interpretation is sufficient, it is recommended to use flexible models such as splines in the leaf nodes. GAMs are preferable to unpenalized B-Splines in terms of their generalization error, however, it needs to be considered that it is computationally more expensive as shown in Table 6.

B.6 Linear smooth with noise features

Here, we examine how noise features that have no influence on the target y affect the MBTs. Therefore, we use the scenario linear smooth and we add six noise features to the underlying data set. In addition to the four MBT algorithms used so far, SLIM models are fitted with lasso regularization [4,1]. Lasso models allow to fit sparse models, i.e., a feature selection within the models fitted in each node is automatically included. The strength of the feature selection depends strongly on the penalization parameter. For all regularized SLIM models, the penalization parameter is selected using the BIC criterion [3]. However, in the case of $df = 3$

Table 5. Simulation results with regard to interpretability for SLIM when different model types in the leaf nodes are fitted for the scenario non-linear. The different SLIM variants are applied as surrogate models on the XGBoost model. The interpretability measures are the number of leaf nodes, the number of split features used within the fitted trees, the share of splits which are based on main effect features, and the effective degrees of freedom.

DGP/ black box	Model	No. of leaves			No. of split feat			Share	Df	
		mean	min	max	mean	min	max	main	mean	sd
XGBoost	lm	19.82	2	33	4.76	1	6	0.6125	6.8671	0.1438
XGBoost	penalized poly	3.58	2	7	2.08	1	4	0.3880	8.5499	1.1147
XGBoost	B-Splines	1.86	1	3	0.84	0	1	0.0000		
XGBoost	GAM	1.90	1	4	0.88	0	2	0.0000		

Table 6. Simulation results with regard to fidelity for SLIM when different model types in the leaf nodes are fitted for the scenario non-linear. The different SLIM variants are applied as surrogate models on the XGBoost model. Besides the R^2 values of the model fits, also the average required time in seconds to fit one tree is provided in the last column.

DGP/ black box	Model	R^2_{train}		R^2_{test}		time in sec
XGBoost	lm	0.9200	0.0228	0.9023	0.0190	55.7771
XGBoost	penalized poly	0.9213	0.0079	0.9150	0.0087	35.4093
XGBoost	B-Splines	0.9382	0.0118	0.9296	0.0120	11.4670
XGBoost	GAM	0.9348	0.0118	0.9289	0.0117	384.7403
XGBoost	XGBoost	0.9386	0.0295	0.9199	0.0362	3.1163

or $df = 2$, the additional restriction is defined so that the effective degrees of freedom (df) must not exceed this value. This enforces especially sparse models.

Scenario definition Let $X_1, \ldots, X_{10} \sim \mathcal{U}(-1, 1)$ where the DGP based on n realizations is defined by $y = f(\mathbf{x}) + \epsilon$ with $f(\mathbf{x}) = \mathbf{x}_1 + 4\mathbf{x}_2 + 3\mathbf{x}_2\mathbf{x}_3$ and $\epsilon \sim \mathcal{N}(0, 0.01 \cdot \sigma^2(f(\mathbf{x})))$. The MBTs are fitted as surrogates on lm predictions on a data set with sample size $n = 3000$ (2000 training, 1000 test observations) using the early stopping parameter configurations $\alpha = 0.001$ and $\gamma = 0.1$. The simulation is repeated 250 times.

Results The aim of the simulation is to investigate whether the noise features are incorrectly chosen as splitting features. Table 7 shows an overview of the results of the described scenario. Noise features are not used as splitting features if the MBTs are used as surrogates for an lm model. However, when we applied the approaches directly on the data from the DGP (not on the model predictions) as a standalone model, SLIM and GUIDE use noise features for splitting. MOB and CTree always split with respect to non-noise features. GUIDE shows the highest share of selecting noise features for splitting. While the number of leaf

nodes for SLIM with the lasso df 2 regularization decreases compared to using SLIM without regularization, the performance values remain on a comparable level and thus, the performance vs. interpretability trade-off improves when the respective regularization is applied in this scenario.

Table 7. Simulation results on 250 simulation runs for an lm surrogate model and as a standalone (DGP) for all four MBTs on scenario linear mixed with noise features for different values of γ and α. The mean (standard deviation) fidelity on the training data for the lm is 0.9901 (0.0004). On the test data set the respective fidelity values for the lm are 0.9901 (0.0006). The column "Share" defines the proportion of trees in which at least one of the noise features is used for splitting. For comparison, we also show the results if the methods are directly applied to the DGP as a standalone model.

Black box	MBT	Share	number of leaves			R^2_{train}		R^2_{test}	
		$\mathbf{x}_{noise}$	mean	min	max	mean	sd	mean	sd
lm	SLIM	0.000	14.036	8	16	0.9987	0.0018	0.9984	0.0019
lm	SLIM lasso	0.000	13.736	8	16	0.9985	0.0023	0.9982	0.0027
lm	SLIM lasso df 3	0.000	14.008	8	16	0.9985	0.0023	0.9983	0.0026
lm	SLIM lasso df 2	0.000	11.160	5	14	0.9979	0.0018	0.9977	0.0020
lm	GUIDE	0.000	14.096	8	16	0.9988	0.0018	0.9984	0.0019
lm	MOB	0.000	15.960	15	16	0.9995	0.0000	0.9993	0.0001
lm	CTree	0.000	15.564	13	16	0.9994	0.0001	0.9992	0.0001
DGP	SLIM	0.072	11.988	5	17	0.9880	0.0048	0.9854	0.0049
DGP	SLIM lasso	0.092	11.048	5	16	0.9871	0.0049	0.9852	0.0051
DGP	SLIM lasso df 3	0.028	9.732	4	14	0.9863	0.0046	0.9848	0.0050
DGP	SLIM lasso df 2	0.028	9.648	4	15	0.9864	0.0044	0.9852	0.0047
DGP	GUIDE	0.104	11.788	5	16	0.9880	0.0047	0.9854	0.0048
DGP	MOB	0.000	11.096	8	14	0.9901	0.0005	0.9878	0.0007
DGP	CTree	0.000	13.140	10	16	0.9904	0.0004	0.9882	0.0007

References

[1] Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**(1), 1 (2010). https://doi.org/10.18637/jss.v033.i01

[2] Hu, L., Chen, J., Nair, V.N., Sudjianto, A.: Surrogate locally-interpretable models with supervised machine learning algorithms. arXiv preprint arXiv:2007.14528 (2020). https://doi.org/10.48550/arXiv.2007.14528

[3] Sabourin, J.A., Valdar, W., Nobel, A.B.: A permutation approach for selecting the penalty parameter in penalized model selection. Biometrics **71**(4), 1185–1194 (2015). https://doi.org/10.1111/biom.12359

[4] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[5] Wood, S.N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society Series B: Statistical Methodology **73**(1), 3–36 (2011). https://doi.org/10.1111/j.1467-9868.2010.00749.x