A Machine Learning Model Selection considering Tradeoffs between Accuracy and Interpretability

Zhumakhan Nazir, Dinmukhamed Kaldykhanov, Kozy-Korpesh Tolep and Jurn-Gyu Park
School of Engineering and Digital Sciences, Computer Science
Nazarbayev University, Nur-Sultan, Kazakhstan
zhumakhan.nazir, dinmukhamed.kaldykhanov, kozykorpesh.tolep, jurn.park@nu.edu.kz

***Abstract*—Using black box machine learning models (e.g., Deep Neural Networks) in high-stakes domains such as healthcare, criminal justice and real-time systems can cause serious problems due to their complexity and poor interpretability. Moreover, model selection with interpretability in addition to accuracy is one of emerging research areas with lack of model agnostic and quantitative interpretability metrics. In this work, we adopt a quantitative interpretability metric, and then, introduce a trade-offs methodology between accuracy and interpretability, which can be demonstrated by increasing interpretability of ML models while allowing accuracy to drop up to given thresholds. In our experimental results, interpretability in terms of simulatability operation count (SOC) is improved up to 76.2% with minimal 2.3% accuracy drop in a SVR estimator of the Auto MPG dataset (up to 64.3% with minimal 1.9% accuracy drop in the Forest Fire dataset of an MLP estimator).**

***Index Terms*—Interpretable Machine Learning (IML); Explainable ML; Model Selection techniques;**

I. Introduction

Machine learning (ML) models are actively being adopted in various fields including high-stakes fields such as public health and criminology. However, most of them can be described as 'black box' models with lack of transparency and accountability [1]. For example, a CNN model misleads its prediction due to incorrect learning (lack of accountability) when the model has learned to detect a metal token that radiology technicians place on the patient in the corner of the image field of view at the time they capture the image [2]; this is because it is challenging to detect such behaviour due to the intrinsic black-box model properties (no transparency).

According to interpretable machine learning (IML) by Molnar et al. [3], there is increasing interest to consider interpretability in addition to accuracy in model selection. However, due to lack of quantitative evaluation methods, assessing interpretability is a challenging task. Moreover, even if we can develop or adopt a quantitative interpretability metric, another challenging issue is how can we improve accuracy and interpretability together? (Or if it is extremely challenging, with trade-offs concepts, alternatively how can we improve interpretability with minimal accuracy drop?). Generally, accuracy and interpretability are inversely related to each other [18] [19]. Therefore, one realistic alternative can be the practicality of such trade-offs, in other words can we build simpler (high interpretability) and accurate 'enough' (acceptable accuracy drop) models?

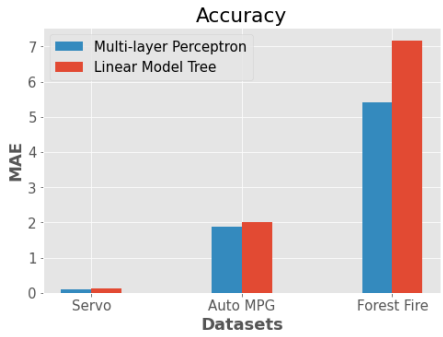
To answer these questions, we adopt a simple yet effective quantitative interpretability metric, simulatability operation count (SOC) [4], following the main contributions of this paper.

- Evaluate best accuracy and interpretability of popular regression models: linear model tree (LMT), multi-layer perceptron regression (MLP) and support vector regression (SVR), using mean absolute errors (MAE) and the number of SOC (#SOC) [4].
- Apply a trade-offs methodology between accuracy and interpretability to improve interpretability of the estimators, by allowing accuracy to decrease up to certain thresholds.

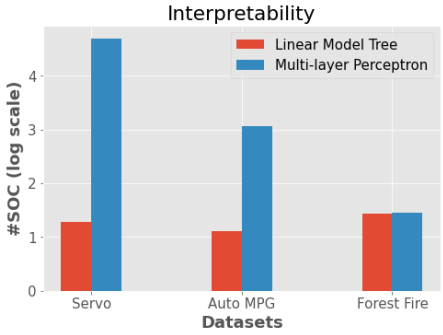
II. Motivation and Related Work

A. Motivation

Although superiority of black-box models lead to their widespread use, simple but accurate models like tree-based models (e.g., linear model tree) are able to perform competitively with high interpretability.



(a) Accuracy (MAE)



(b) Interpretability (#SOC)

Fig. 1: Motivating Example: Accuracy and interpretability of LMT and MLP Algorithms

As shown in Figure 1a of the motivating example, the linear model tree (LMT) compared to multi-layer perceptron regression (MLP), LMT has similar accuracy (MAE) on the two simple datasets (Servo and AutoMPG), and a lower accuracy (higher MAE) on a complex dataset (Forest Fire). Whereas, the interpretability of LMT in terms of #SOC is significantly better (lower SOC) than MLP in the Servo and AutoMPG datasets (Figure 1b).

Therefore, for relatively simple datasets (Servo and AutoMPG), we select the LMT model to improve interpretability as well as accuracy, within negligible accuracy drop. However, if the accuracy drop can not be negligible for complicated datasets like the Forest Fire, we can build models by cross-validating hyper-parameters of the model to make it more interpretable (e.g. by reducing number of hidden layers or neurons in MLP) with minimal accuracy drop.

B. Related Work

There are a number of research works devoted to interpretability/explainability of ML models. Rudin [1] argued that using interpretable models is preferred and that they are able to replace complex 'back box' models, in terms of transparency and accountability. Farquad et al. [5] extracted if-then based rules from Support Vector Machine with hybrid method: first fit data to SVM and get a reduced set of training data represented by support vectors and train another explainable model. Doshi-Velez and Kim [6] introduced that human-grounded proxy metrics can be built by analyzing human evaluation of interpretability of the model itself or a post-hoc interpretation of a black-box model. Slack et al. [7] performed a study of simulatability and 'what if' local explainability for decision tree, logistic regression and neural network. As a result, the run time operation count was proposed as a global interpretability metric. The authors [4] derived the SOC formula for several regression models and evaluated their interpretability. These SOC formulas will be adopted to compare interpretability of built models for our experiment.

III. Methodology

Figure 2 demonstrates an overview of our methodology. Phase I contains traditional model

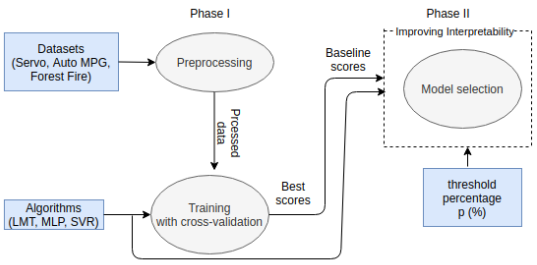


Fig. 2: An Overview of the Trade-offs Methodology

building steps, which consist of two stages: Data Preprocessing and Model Training. In the preprocessing stage for regression, we 1) convert categorical columns of datasets into numerical using OrdinalEncoder [9]; 2) scale using StandardScaler [9] to equate initial effect of each feature on regression coefficients; 3) detect and remove outlier data instances with z-score (greater than 3 [13]); 4) drop highly correlated columns to reduce serious multicollinearity (i.e. if Variance Inflation Factor (VIF) is greater than 10) [14]. In addition, in the model training stage, algorithms are trained using GridSearchCV from scikit-learn [9] which ensures that we can select the best model among all possible combinations of hyperparameters.

Phase II is the main part of our methodology for model selection. First, the #SOC of the selected models from the training stage will be calculated. Then, we try to reduce #SOC values by allowing the Mean Absolute Error (MAE) to increase by up to a threshold percentage ($p\%$) from the baseline accuracy scores obtained in the training stage (trade-offs between accuracy and interpretability). It can be achieved by reducing parameters that affect #SOC from Table I below.

Esti.	$K_t or A_t$	SOC formula
LMT	N/A	$2D + 2P + 1$
MLP	A_t	$\sum_{h=1}^{H}(2 \times N_h + A_t) \times N_{h+1} + 2 \times N_{h+1}$
	Relu	$A_t = 1$
	Sigmoid	$A_t = 4$
	Tanh	$A_t = 9$
SVR	K_t	$SV \times (K_t + 2)$
	Linear	$K_t = (2P - 1)$
	Polynomial	$K_t = (2P + 1 + d)$
	Sigmoid	$K_t = (2P + 10)$
	RBF	$K_t = (3P + 1)$

TABLE I: SOC formula of Estimators [4]

Since the number of features (P) will be fixed after feature selection, the depth parameter (D) can be reduced to minimize #SOC in LMT. For the MLP, the number of neurons (N), number of hidden layers (H) and type of activation functions (A_t) can be changed. Lastly, for SVR, reducing the number of support vectors (SV) (using NuSVR [9]) and choosing a simpler type of kernel function (K_t, e.g. Linear) reduces #SOC.

IV. Experimental Results

A. Experimental Setup

Datasets: To conduct the experiment, three benchmark datasets (from simple to complex) are used: Servo (simple) [10], Auto MPG (medium) [11] and Forest Fire (complex) [12]. The size of the chosen datasets are small because Phase II of our methodology would take enormous time otherwise.

1) Servo: The dataset has 5 features and 167 instances including a bold-highlighted target variable (class). All values of the dataset are discrete; and the target variable is continuous in $[0.13, 7.1]$ and means a raise time of the servo-mechanism. Feature are given in TABLE II

Feature	Description
motor	A,B,C,D,E
screw	A,B,C,D,E
pgain	3,4,5,6
vgain	1,2,3,4,5
class	**0.13 to 7.10**

TABLE II: Servo Features

2) Auto MPG: This dataset has 9 features and 398 instances, predicting attribute 'mpg' (miles per gallon) using 3 multi-valued discrete and 5 continuous attributes. Detailed feature descriptions are given in TABLE III

Feature	Description
mpg	**miles per gallon, continuous output variable**
model year	version of a car
cylinders	power unit of engine
displacement	measure of the cylinder volume
horsepower	power of engine produces
weight	weight of car
acceleration	amount of time taken for car to reach a velocity of 60 miles per hour
origin	multi-valued discrete
name	name of the car

TABLE III: Auto MPG features

3) Forest Fire: This dataset can be represented as a complex dataset having 13 features and 517 instances with a target variable 'area' described in TABLE IV.

Algorithms: LMT uses an estimator [21] based on Quinlan's M5 design [8]; and MLP Regressor and SVR are taken from scikit-learn library [9]. Classifier estimators can also be used, however the idea and steps will be identical.

Feature	Description
area	**in ha, 0 means less than** 1ha/100 (=100m^2)
X,Y	coordinates of place of fire
month, date	categorical value from jan. to dec. and mon. to sun. correspondingly
temp, wind, rain	meteorological data
RH	relative humidity
FFMC,DMC,DC,ISI	components of Fire Weather Index (FWI) of the Canadian system

TABLE IV: Forest Fire features

	Servo	Auto MPG	Forest Fire
LMT	0.133	1.889	6.847
MLP	**0.096**	1.890	5.376
SVR	0.183	**1.830**	**5.212**
Lin. Reg.	0.863	2.304	6.723
Other Ref.	0.220 [15]	2.020 [16]	6.334 [17]

TABLE V: Accuracy Performance (MAE) of Trained Models with references. (Lowest error values in bold).

B. Results and Analysis

Preprocessing: After the preprocessing stage, Servo is reduced to 152 data instances; and Auto MPG is reduced to 367 data instances with 6 features ('displacement' and 'horsepower' are dropped due to collinearity issue, i.e. VIF >10; and 'name' is not used); and Forest Fire is reduced to 468 data samples.

Model Training: As mentioned previously, models are trained using GridSearchCV which allows testing a wide range of hyper-parameter combinations. As illustrated in Figure 3, the lowest

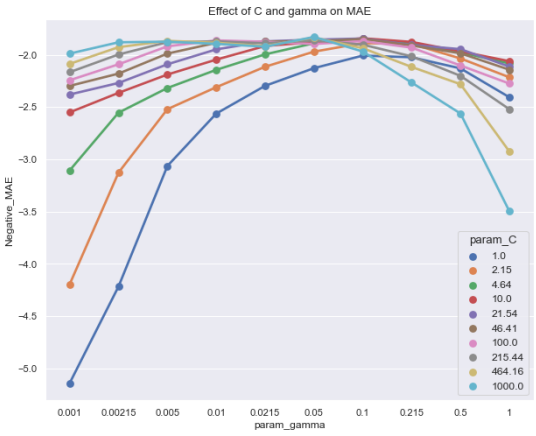


Fig. 3: Training SVR on Auto MPG dataset.

error corresponds to $gamma = 0.05$ and $C = 1000$ and concave down shape ensures that we select the suboptimal points.

Overall results using GridSearchCV in the model training stage are provided in Table V. MLP outperforms other estimators on Servo dataset, SVR on Auto MPG and Forest Fires datasets. For comparison purposes, results of the sklearn's default Linear Regression and other references are provided.

For Model Selection: Two approaches to improve interpretability can be available: 1) can a model be replaced by another simpler model? (described in the motivating example) and 2) can the same model be simplified using the threshold percentage? (described in Phase II of Fig 2).

Results of the first approach are shown in Figure 4. The scale on the axis means that an actual value is multiplied by a scaled value to fit on the graph. For example, the last entry in AutoMPG & SVR cell $(1062, 2.006)$ in Table VI corresponds to the utmost point (labeled as (9.4, 83.1) = (MAE, #SOC :2.066, 7696) in the scaled graph of SVR on Figure 4 $(7696 \approx 1062 \times 7.25, 2.066 \approx 2.006 \times 1.03)$. The scaled values are found by taking one of the graphs (e.g. the graph of MLP in Figure 4) as a base and scaling the other graphs' starting points to match the base graph's starting point. Thus, the x-axis scale value of SVR (7.25) is found by dividing 45393 (MLP's highest #SOC on Auto MPG) to 6264 (SVR's highest #SOC on Auto MPG). In the same way, the y-axis scale value of SVR, 1.03 is $1.890/1.830$.

The idea behind Figure 4 is to show how different models behave when they are optimized for interpretability using the trade-offs methodology. MLP and SVR behave similarly - #SOC reduced significantly $(97.4\%$ and 76.2% respectively) for small increases in error rate $(2.6\%$ and $2.3\%)$. However, accuracy of LMT is not reduced because the most accurate model has depth of 1 (which is also the most interpretable model). From the results (Figure 4 and Table VI) of the Auto MPG dataset, SVR is the most accurate and least interpretable, while LMT is most interpretable with moderate accuracy. For interpretability concerning tasks, LMT can be a suitable candidate.

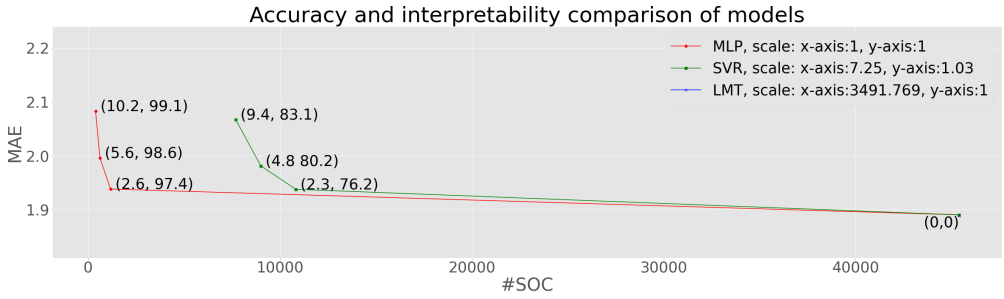


Fig. 4: Comparison of Models in Accuracy and Interpretability in Auto MPG dataset. (Each point is labeled in (a, b) format, where a is MAE increase and b is SOC decrease in percentage).

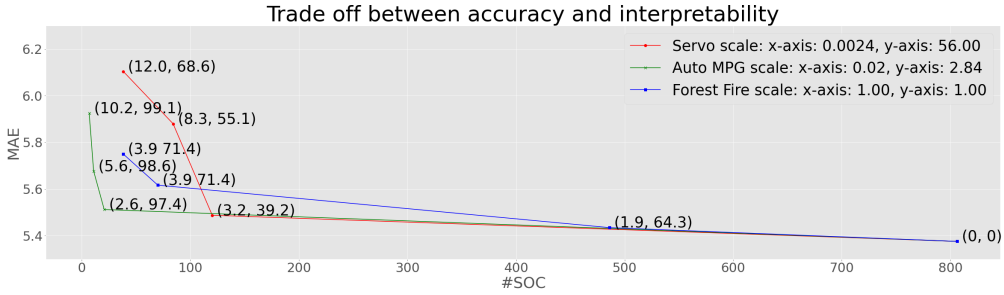


Fig. 5: Trade-off between Accuracy and Interpretability in MLP estimator. (Each point is labeled in (a, b) format, where a is MAE increase and b is SOC decrease (in percentage)).

Table VI is an extension of Figure 4 by adding the results of Servo and Forest Fire datasets. Similar to the figure, significant improvement in interpretability can be achieved with small reduction in accuracy.

	LMT	MLP	SVR
Servo	0.133, 19	0.096, 339371	0.183, 2280
	0.134, 17	0.098, 50583	0.186, 1905
	0.181, 15	0.105, 35403	0.191, 1740
	0.184, 13	0.109, 16103	0.205, 1545
Auto MPG	**1.889, 13**	1.890, 45393	**1.830, 6264**
	- -	1.938, 1163	1.876, 1494
	- -	1.996, 613	1.918, 1242
	- -	2.083, 383	2.006, 1062
Forest F.	6.847, 27	5.376, 806	5.212, 10550
	6.965, 27	5.435, 486	5.347, 7425
	7.144, 27	5.617, 70	5.479, 6850
	7.180, 27	5.751, 38	5.725, 5700

TABLE VI: Comparison of Models in terms of Accuracy and Interpretability. (Each entry in the table contains data in (MAE, #SOC) format).

Results of the second approach are shown in Figure 5 using MLP on the three datasets. The estimator behaves similarly on all the datasets - #SOC can be decreased significantly with small reduction in error rate. The elbow points can be suitable candidates for effective trade-offs between accuracy and interpretability, since after that point (when moving from right to left) a slope of the graph sharply increases in absolute magnitude. For example, for the Forest Fire dataset (blue line) interpretability can be improved (reduction in #SOC) by 64.3% with only 1.9% drop (increase in MAE) in accuracy. The same concept applies to other datasets also.

V. Conclusion

In this paper, we proposed a trade-offs methodology between accuracy and interpretability by adopting the model agnostic quantitative metric, SOC. For the LMT algorithm, due to its simple yet sophisticated structure (combination of decision tree and linear regression models), it is the most interpretable model among the evaluated estimators, having almost the same accuracy with MLP in relatively simple-medium datasets such

as Servo and Auto MPG. In addition, for all types of datasets, interpretability can be improved more by allowing accuracy to drop/sacrifice to a certain threshold, in terms of algorithms (Figure 4) and datasets (Figure 5). From the experiments, interpretability can be improved significantly with minimal accuracy reduction (up to 76.2% interpretability improvement with minimal 2.3% accuracy drop). Finally, although this work was evaluated with regression datasets, our trade-offs methodology can be also applicable for classification problems. (This will be an extended future work).

ACKNOWLEDGEMENT

This work was partly supported by the Nazarbayev University (NU), Kazakhstan, under FDCRGP grant 021220FD0851.

REFERENCES

[1] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

[2] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11), e1002683.

[3] Molnar, C., Casalicchio, G., Bischl, B. (2020). Interpretable MachineLearning–A Brief History, State-of-the-Art and Challenges. arXiv preprint arXiv:2010.09337

[4] Park, J.-G., Dutt, N., and Lim, S.-S. An Interpretable Machine Learning Model Enhanced Integrated CPU-GPU DVFS Governor. ACM Trans. Embed. Comput. Syst. (TECS), 20(6): 108:1-108:28 (2021).

[5] Farquad, M. A. H., Ravi, V., Raju, S. B. (2010). Support vector regressionbased hybrid rule extraction methods for forecasting. Expert Systems with Applications, 37(8), 5577-5589

[6] Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608

[7] Slack, D., Friedler, S. A., Scheidegger, C., Roy, C. D. (2019). Assessingthe Local Interpretability of Machine Learning Models. arXiv preprint arXiv:1902.03501

[8] Quinlan, J. R. (1992). "Learning with continuous classes." Proc., 5thAustralian Joint Conf. on Artificial Intelligence, Adams Sterling,eds., World Scientific, Singapore, 343–348

[9] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[10] Servo dataset (1993). Link: https://archive.ics.uci.edu/ml/datasets/Servo

[11] Auto Mpg dataset. Link: http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data-original

[12] Forest Fire dataset. Link: https://archive.ics.uci.edu/ml/datasets/forest+fires

[13] Engineering Statistics Handbook. Nist Sematech. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

[14] O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality quantity, 41(5), 673-690.

[15] Cortez, P., Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1-17. page 17.

[16] Birnbaum, L. A. (Ed.). (2014). Machine Learning Proceedings 1993: Proceedings of the Tenth International Conference on Machine Learning, University of Massachusetts, Amherst, June 27-29, 1993. Morgan Kaufmann. page 240.

[17] Stanford-Moore, A., Moore, B. Wildfire Burn Area Prediction. page 5.

[18] Johansson U, Sönströd C, Norinder U, Boström H. Trade-off between accuracy and interpretability for predictive in silico modeling. Future Med Chem. 2011 Apr;3(6):647-63. doi: 10.4155/fmc.11.23. PMID: 21554073.

[19] Mori, T., Uchihira, N. Balancing the trade-off between accuracy and interpretability in software defect prediction. Empir Software Eng 24, 779–825 (2019). https://doi.org/10.1007/s10664-018-9638-1

[20] Uzair, M., Jamil, N. (2020, November). Effects of Hidden Layers on the Efficiency of Neural networks. In 2020 IEEE 23rd International Multitopic Conference (INMIC) (pp. 1-6). IEEE.

[21] Dillard, L. lmt.py (2017). Link: https://gist.github.com/logandillard/lmt.py.