Should We Compute Feature Effects on Training or Validation Data?

Timo Heiß[1]

Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany
t.heiss@campus.lmu.de

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: Explainable AI · Feature effects · Partial dependence plot · Accumulated local effects

1 Introduction

Most Machine Learning (ML) models can be considered black boxes — opaque systems that intrinsically do not allow insight into their internal reasoning, making it often impossible to explain their decisions. However, this can be problematic in many domains and applications, such as the healthcare, legal, or finance sectors, where decisions must be transparent and accountable [1].

Interpretability is crucial to enhance trust [23, 24], address potential biases [10], fairness and ethical concerns [14], and ensure compliance with regulations such as the EU's General Data Protection Regulation (GDPR) [4] and AI Act [5]. To address these challenges, the field of Explainable AI / Interpretable ML has emerged [1]. Although there are many different methods[1], we will focus on feature effect methods like *Partial Dependence Plots (PDP)* [7] and *Accumulated Local Effects (ALE)* [2].

Due to the severity of many applications, it is crucial to utilize these explainability methods correctly. In general, there are many pitfalls to be aware of [22], including whether to compute explanations in-sample, i.e. on training data, or out-of-sample, which we refer to as validation data in the following. In loss-based methods such as *Permutation Feature Importance (PFI)* [3, 6], this choice is crucial and has already been studied (e.g., in [22]). Nevertheless, other explainability methods, such as the *Mean Decrease in Impurity (MDI)* (or *Gini Importance*) of Random Forests, or *SHAP* values [16, 17], have also been found to exhibit biases when computed on training data.

However, to the best of our knowledge, there exist no similar studies for feature effect methods like PDP and ALE. Existing works, including the original papers of PDP and ALE, rely on training data without further justification (e.g., [2, 7, 20]). In contrast, practitioners often advocate for using unseen

[1] For an overview of Explainable AI methods, see e.g. [1, 13, 20].

test/validation data[2], or base their choice on practical constraints like dataset size[3]. While the training set is usually larger and might thus lead to less variance in the feature effect estimates, a too large dataset can increase computation times substantially, particularly for the PDP [7]. On the other hand, although feature effects are not based on generalization error like PFI, it is not clear how much they are affected by overfitting or distribution shifts between training and validation data.

In this paper, we aim to answer this largely unaddressed, fundamental methodological question of whether to use training or validation data to compute feature effects. We perform an empirical simulation study, comparing feature effect error and uncertainty for PDP and ALE across training data, validation data, and cross-validation scenarios, considering various data scenarios and model types. Our main contributions in this paper are as follows.

1. We shed light on the question of whether to compute feature effects on training data, validation data, or in a cross-validated manner, grounded through our comprehensive simulation study, considering feature effect error, bias, and variance across different models and data scenarios.
2. We define methods to quantify feature effect error and uncertainty in a theoretical framework, building upon previous work in this area.
3. We provide an overview of commonly used test functions for simulation studies in Interpretable ML, including applications and purposes.

These contributions have several important implications: Our empirically grounded recommendations enable practitioners and researchers to make informed decisions about which dataset to select for feature effect computation, helping to understand potential implications of their choices. Additionally, our framework for evaluating feature effects and our systematic collection of test functions provide a foundation for future research in Interpretable ML, with the latter specifically facilitating test function choice for simulation studies.

The remainder of this paper is structured as follows. In Section 2, we introduce the considered feature effect methods PDP and ALE, as well as related works on feature effect error and uncertainty, and give an overview of common test functions for simulation studies. In Section ??, we extend existing works and give our definitions of feature effect errors and uncertainty. We then describe the methodology and set-up of our simulation studies in Section 3, present the results in Section 4, and discuss their implications and limitations in Section 5. In Section 6, we briefly conclude our work.

[2] see, e.g. https://github.com/SauceCat/PDPbox/issues/68 (10/27/2024)
[3] see, e.g. https://forums.fast.ai/t/partial-dependence-plot/98465 (10/27/2024)

2 Background & Related Work

2.1 Feature Effects

The *Partial Dependence Plot (PDP)* by Friedman [7] describes the marginal effect of one or two features on the prediction of a model $\hat{f}$. For a feature set X_S (with $S \subseteq \{1, \ldots, p\}$, $|S| = 1$ or $|S| = 2$), the PDP is defined as

$$PDP_{\hat{f},S} = \mathbb{E}_{X_C}[\hat{f}(x_S, X_C)] = \int f(x_S, x_C) d\mathbb{P}(x_C), \tag{1}$$

where X_C is the complement feature subset. $PDP_{\hat{f},S}$ is a function of x_S and can be estimated by Monte Carlo integration:

$$\widehat{PDP}_{\hat{f},S}(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)}). \tag{2}$$

Here, $x_C^{(i)}$ are the actual complement feature values from the dataset of n instances. To plot this function, a grid of G grid points $\{(x_S^{(g)}, \widehat{PD}_{\hat{f},S}(x_S^{(g)}))\}_{g=1}^{G}$ can be used [21]. Molnar et al. [22] recommend using quantile-based over equidistant grids.

The PDP assumes that the features in S are independent of the features in C. If this is violated, the perturbations may produce unrealistic data points outside the underlying joint distribution of the data. This extrapolation issue can cause misleading interpretations [20, 22].

The *Accumulated Local Effects (ALE)* plot is an alternative to the PDP that solves the extrapolation issue [2]. Using the notation above, for $|S| = 1$, the ALE plot is defined as

$$ALE_{\hat{f},S}(x_S) = \int_{x_{\min,s}}^{x_S} E_{X_C|X_S}\left[\hat{f}^S(X_S, X_C)|X_S = z_S\right] dz_S - \text{constant} \tag{3}$$

$$= \int_{x_{\min,s}}^{x_S} \int_{x_C} \hat{f}^S(z_S, x_C) \mathbb{P}(x_C|z_S) dx_C dz_S - \text{constant}, \tag{4}$$

where $\hat{f}^S(x_S, x_C) = \frac{\partial \hat{f}(x_S, x_C)}{\partial x_S}$. The constant is chosen so that $\widehat{ALE}_{\hat{f},S}(X_S)$ is centered with a mean of 0 w.r.t. the marginal distribution of X_S. The uncentered ALE can be estimated by

$$\widehat{\widetilde{ALE}}_{\hat{f},S}(x) = \sum_{k=1}^{k_S(x)} \frac{1}{n_S(k)} \sum_{i:x_S^{(i)} \in N_S(k)} \left[\hat{f}(z_{k,S}, x_C^{(i)}) - \hat{f}(z_{k-1,S}, x_C^{(i)})\right]. \tag{5}$$

Here, $\{N_S(k) = (z_{k-1,S}, z_{k,S}]\}_{k=1}^{K}$ partitions the samples $\{x_S^{(i)}\}_{i=1}^{n}$ into K intervals or neighborhoods $N_S(k)$. $n_S(k)$ denotes the number of observations in

the kth interval $N_S(k)$, $k_S(x)$ represents the index of the interval to which a particular value x of feature x_S belongs. The uncentered ALE is centered by

$$\widehat{ALE}_{\hat{f},S}(x) = \widehat{\widetilde{ALE}}_{\hat{f},S}(x) - \frac{1}{n}\sum_{i=1}^{n} \widehat{\widetilde{ALE}}_{\hat{f},S}(x_S^{(i)}) \tag{6}$$

to have a mean effect of 0. For the grid that defines the intervals, the quantiles of the empirical distribution of $\{x_S^{(i)}\}_{i=1}^n$ can be used [2].

Other feature effect methods include the *M-Plot (Marginal Plot)*, which, however, suffers from the omitted variable bias [2, 7], or *functional ANOVA (fANOVA)*, which decomposes feature effects into main and interaction effects [12].

Furthermore, methods exist that extend previous ones, such as *Robust and Heterogeneity-aware ALE (RHALE)* [8], or *Accumulated Total Derivative Effect (ATDEV)* plots, which can be decomposed into ALE and *Accumulated Cross Effects (ACE)* [15].

In addition to these global feature effect methods, there are regional effect plots such as *REPID* [11], as well as local methods, including *ICE (Individual Conditional Expectation)* curves [9] or SHAP dependence plots (e.g., [20]) based on SHAP values [18].

For the remainder of this paper, we will focus on PDP and ALE as the most popular global feature effect methods and refer to them when speaking of feature effects.

2.2 Feature Effect Error Decomposition

To quantify the error of a computed feature effect, a "ground truth" needs to be defined first. We follow the approach of Molnar et al. [21] and define ground truth versions of PDP and ALE directly on the data generating process (DGP) by applying PDP and ALE to the underlying ground truth function f (instead of model $\hat{f}$).

For PDP, we can directly use the definition of Molnar et al. [21]:

Definition 1 (Definition 1 from [21]). *The PDP ground truth is the PDP applied to function $f : \mathcal{X} \to \mathcal{Y}$ of the data generating process.*

$$PDP_{f,S}(x_S) = \mathbb{E}_{X_C}[f(x_S, X_C)] \tag{7}$$

As stated by Molnar et al. [21], their results also apply to conditional variants of the PDP such as ALE. We now make this definition explicit:

Definition 2. *The ALE ground truth is the ALE applied to function $f : \mathcal{X} \to \mathcal{Y}$ of the data generating process.*

$$ALE_{f,S}(x_S) = \int_{x_{min,s}}^{x_S} E_{X_C|X_S}\left[f^S(X_S, X_C)|X_S = z_S\right] dz_S - constant \tag{8}$$

where $f^S(x_S, x_C) = \frac{\partial f(x_S, x_C)}{\partial x_S}$ *and constant chosen such that the effect has a mean of* 0 *w.r.t. the marginal distribution of* X_S.

Note that different ground truth effects may also be defined, and our choices come with certain implications and limitations, such as omitting the *aggregation bias*[4] [19].

With more complex ground truth functions f, it may become increasingly difficult to derive the ground truth feature effects analytically, especially for ALE. In these cases, we therefore propose to also estimate those effects by Monte Carlo integration, yielding $\widehat{PDP}_{f,S}(x_S)$ and $\widehat{ALE}_{f,S}(x_S)$ (obtained by plugging in f instead of $\hat{f}$ into the estimators in equations (2) and (5) / (6)).

Summarizing, we have now defined four quantities per feature effect: $PDP_{f,S}$ $\widehat{PDP}_{f,S}$, $PDP_{\hat{f},S}$, and $\widehat{PDP}_{\hat{f},S}$ (analogue for ALE). We can now define different errors between each of these quantities. In this paper, we focus on the MSE, as it can be decomposed into bias and variance (see e.g. [?]). Taking, for example, $PDP_{f,S}$ as ground truth, we can define the MSE of $PDP_{\hat{f},S}$ at a point x_S as follows [21]:

$$\text{MSE}(x_S; PDP_{f,S}, PDP_{\hat{f},S}) = \mathbb{E}_F[(PDP_{f,S}(x_S) - \widehat{PDP}_{\hat{f},S}(x_S))^2] \tag{9}$$

$$= \underbrace{(PDP_f(x) - \mathbb{E}_F[PDP_{\hat{f}}(x)])^2}_{Bias^2} + \underbrace{\text{Var}_F[PDP_{\hat{f}}(x)]}_{Variance} \tag{10}$$

Here F denotes the distribution of trained models.

$\mathbb{E}_F$ and Var_F could be estimated by averaging over multiple models of the same inducer fitted to M different training data sets sampled independently from the DGP. In the simulation scenario, this yields:

$$\widehat{\text{MSE}}(x_S; PDP_{f,S}, PDP_{\hat{f},S}) = \frac{1}{M} \sum_{m=1}^{M} (PDP_{f,S}(x_S) - \widehat{PDP}_{\hat{f}^{(m)},S}(x_S))^2 \tag{11}$$

3 Methodology & Experimental Set-Up

4 Results

5 Discussion

6 Conclusion

Acknowledgments. A bold run-in heading in small font size at the end of the paper is used for general acknowledgments, for example: This study was funded by X (grant number Y).

[4] for details, see [11]

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Apley, D.W., Zhu, J.: Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. Journal of the Royal Statistical Society Series B: Statistical Methodology **82**(4), 1059–1086 (Sep 2020)
3. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (Oct 2001)
4. European Parliament and Council: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union **L 119**(2016/679), 1–88 (5 2016)
5. European Parliament and Council: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union (2024/1689) (7 2024), PE/24/2024/REV/1
6. Fisher, A., Rudin, C., Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research : JMLR **20**(177), 1–81 (2019)
7. Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics **29**(5), 1189–1232 (2001)
8. Gkolemis, V., Dalamagas, T., Ntoutsi, E., Diou, C.: RHALE: Robust and Heterogeneity-aware Accumulated Local Effects (Sep 2023), arXiv:2309.11193v1 [cs.LG]
9. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics **24**(1), 44–65 (Jan 2015)
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys **51**(5), 1–42 (Sep 2019)
11. Herbinger, J., Bischl, B., Casalicchio, G.: REPID: Regional Effect Plots with implicit Interaction Detection. In: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. pp. 10209–10233. PMLR (May 2022)
12. Hooker, G.: Discovering additive structure in black box functions. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 575–580. KDD '04, Association for Computing Machinery, New York, NY, USA (Aug 2004)
13. Kamath, U., Liu, J.: Introduction to Interpretability and Explainability. In: Kamath, U., Liu, J. (eds.) Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning, pp. 1–26. Springer International Publishing, Cham (2021)

14. Lipton, Z.C.: The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (Jun 2018)
15. Liu, X., Chen, J., Vaughan, J., Nair, V., Sudjianto, A.: Model Interpretation: A Unified Derivative-based Framework for Nonparametric Regression and Supervised Machine Learning (Sep 2018), arXiv:1808.07216v2 [cs, stat]
16. Loecher, M.: Debiasing MDI Feature Importance and SHAP Values in Tree Ensembles. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) Machine Learning and Knowledge Extraction. pp. 114–129. Springer International Publishing, Cham (2022)
17. Loecher, M.: Debiasing SHAP scores in random forests. AStA Advances in Statistical Analysis **108**(2), 427–440 (Jun 2024)
18. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. **54**(6) (Jul 2021)
20. Molnar, C.: Interpretable machine learning: a guide for making black box models explainable. Christoph Molnar, Munich, Germany, second edition edn. (2022)
21. Molnar, C., Freiesleben, T., König, G., Herbinger, J., Reisinger, T., Casalicchio, G., Wright, M.N., Bischl, B.: Relating the partial dependence plot and permutation feature importance to the data generating process. In: Longo, L. (ed.) Explainable Artificial Intelligence. pp. 456–479. Springer Nature Switzerland, Cham (2023)
22. Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B.: General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R., Samek, W. (eds.) xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers, pp. 39–68. Springer International Publishing, Cham (2022)
23. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (Aug 2016)
24. Teach, R.L., Shortliffe, E.H.: An analysis of physician attitudes regarding computer-based clinical consultation systems. Computers and Biomedical Research, an International Journal **14**(6), 542–558 (Dec 1981)