Pitfalls to Avoid when Interpreting Machine Learning Models

Christoph Molnar [1] **Gunnar König** [1 2] **Julia Herbinger** [1] **Timo Freiesleben** [3 4] **Susanne Dandl** [1]
Christian A. Scholbeck [1] **Giuseppe Casalicchio** [1] **Moritz Grosse-Wentrup** [2 5 6] **Bernd Bischl** [1]

Abstract

Modern requirements for machine learning (ML) models include both high predictive performance and model interpretability. A growing number of techniques provide model interpretations, but can lead to wrong conclusions if applied incorrectly. We illustrate pitfalls of ML model interpretation such as bad model generalization, dependent features, feature interactions or unjustified causal interpretations. Our paper addresses ML practitioners by raising awareness of pitfalls and pointing out solutions for correct model interpretation, as well as ML researchers by discussing open issues for further research.

1. Introduction

Traditionally, researchers have used parametric models, e.g., linear models, to conduct inference. However, a noticeable shift has happened over the last years towards more non-parametric and non-linear ML models. Models such as random forests, boosting or neural networks often outperform interpretable models on many prediction tasks, as most ML models handle feature interactions and non-linear effects automatically[1] (Fernández-Delgado et al., 2014).
Many disciplines benefit from the predictive performance of ML models and answer scientific questions using ML interpretation techniques. Examples of such efforts include modeling pre-evacuation decision making (Zhao et al., 2020), mapping canopy covers in savannas (Anchang et al., 2020), understanding wildlife diseases (Fountain-Jones et al., 2019), forecasting crop yield (Shahhosseini et al., 2020; Zhang et al., 2019), inferring behavior from smartphone usage (Stachl et al., 2019), and analyzing risk for teacher burnout (Posada-Quintero et al., 2020).
Practitioners are usually interested in the global effect that features have on the outcome and their importance for correct predictions. For certain model classes, e.g., linear models or decision trees, feature effects or importance scores can be inferred from the learned parameters and model structure. In contrast, complex non-linear models that, e.g., do not have intelligible parameters, make it more difficult to extract such knowledge. Therefore, interpretation methods necessarily simplify the relationships between features and the target, e.g., by marginalizing over other features. Prominent techniques for global feature effects include the partial dependence plot (PDP) (Friedman et al., 1991), accumulated local effects (ALE) (Apley & Zhu, 2016) and individual conditional expectation (ICE) (Goldstein et al., 2015). A common feature importance technique is the permutation feature importance (PFI) (Breiman, 2001; Fisher et al., 2019; Casalicchio et al., 2019). This paper will mainly focus on pitfalls of global interpretation techniques when the full functional relationship underlying the data is to be analyzed. Out of scope is the discussion of "local" interpretation methods such as LIME (Ribeiro et al., 2016) or counterfactual explanations (Wachter et al., 2017; Dandl et al., 2020), where individual predictions are to be explained – usually to explain decisions to individuals.
The shift towards ML modeling entails numerous pitfalls for model interpretations. ML models usually contain non-linear effects and higher-order interactions. Therefore, lower-dimensional or linear approximations can be inappropriate and misleading masking effects can occur. As interpretations are based on simplifying assumptions, the associated conclusions are only valid if we have checked that the assumptions underlying our simplifications are not substantially violated. In classical statistics this process is called "model diagnostics" (Fahrmeir et al., 2013) and we believe that a similar process is necessary for interpretable machine learning (IML) based techniques.
Contributions: We review pitfalls of global model-

[1]Department of Statistics, LMU Munich, Munich, Germany [2]Research Group Neuroinformatics, Faculty for Computer Science, University of Vienna [3]Munich Center for Mathematical Philosophy, LMU Munich [4]Graduate School of Systemic Neurosciences, LMU Munich [5]Research Platform Data Science @ Uni Vienna [6]Vienna Cognitive Science Hub. Correspondence to: Christoph Molnar <christoph.molnar@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[1]While the inclusion of non-linear and interactions effects in classical statistical models is possible, it comes with the increased cost of going more or less manually over many possible modelling options.

agnostic[2] interpretation techniques. Each section describes the pitfall, reviews (partial) solutions for practitioners and discusses open issues that require further research.

Related Work: A general warning about using and explaining ML models for high stakes decisions has been brought forward by Rudin (2019). She strictly argues against model-agnostic techniques in favour of inherently interpretable models. Krishnan (2019) criticizes the general conceptual foundation of interpretability, but does not dispute the usefulness of available methods. Likewise, Lipton (2018) criticizes interpretable ML (IML) for its lack of causal conclusions, trust and insights, but the author does not discuss any pitfalls in detail. Specific pitfalls due to dependent features are discussed by Hooker (2007) for partial dependence and functional ANOVA and by Hooker & Mentch (2019) for feature importance computations. Hall (2018) discusses recommendations for the application of particular IML methods, but does not address general pitfalls.

2. Bad Model Generalization

Pitfall: Under- or overfitting models will result in misleading interpretations regarding true feature effects and importance scores, as the model does not match the underlying data generating process well (Good & Hardin, 2012). In-sample evaluation (i.e., on training data) should not be used for ML models due to the danger of overfitting. We have to resort to out-of-sample validation such as cross-validation procedures. These resampling procedures are readily available in software and well-studied in theory and practice (Arlot & Celisse, 2010), although rigorous analysis of cross-validation is still considered an open problem (Shalev-Shwartz & Ben-David, 2014).

Formally, IML methods are designed to interpret the model instead of drawing inferences about the data generating process. In practice, however, the latter is the goal of the analysis, not the former. If a model approximates the data generating process well enough, its interpretation should reveal insights into the underlying process.

Solution: An interpretation can only be as good as its underlying model. It is crucial to properly evaluate models using training and test splits, ideally using a resampling scheme like (repeated) cross-validation for smaller sample sizes and nested setups, when computational model selection and hyperparameter tuning are involved (Bischl et al., 2012; Simon, 2007). Flexible models should be part of the model selection process so that the true data generating function is more likely to be discovered (Claeskens et al., 2008). This is important, as the Bayes error for most practical situations is unknown, and we cannot make absolute statements about whether a model already fits the data optimally.

[2]Model-agnostic methods can be applied to any ML model

3. Unnecessary Use of Complex Models

Pitfall: A common mistake is to use an opaque, complex ML model when an interpretable model would have been sufficient, i.e., when the performance of interpretable models is only negligibly worse – or maybe the same or even better – than the ML model. Although there are many model-agnostic methods to interpret complex ML models, it is usually preferable to use an interpretable model (Rudin, 2019). There are also some examples where complex ML models such as neural networks were not able to beat interpretable models (Makridakis et al., 2018; Baesens et al., 2003; Kuhle et al., 2018; Wu et al., 2010).

Solution: We recommend to start with simple, interpretable models such as (generalized) linear models, LASSO, generalized additive models, decision trees or decision rules and gradually increase complexity in a controlled, step-wise manner, where predictive performance is carefully measured and compared. Complex models should only be analyzed if the additional performance gain is both significant and relevant – a judgment call that the practitioner must ultimately make. Starting with simple models is considered best practice in data science, independent of the question of interpretability (Claeskens et al., 2008). The comparison of predictive performance between model classes of different complexity can add further insights for interpretation.

Open Issues: Measures of model complexity allow to quantify the trade-off between complexity and performance and to automatically optimize for multiple objectives beyond performance. Some steps have been made towards quantifying model complexity like Molnar et al. (2019) and Philipp et al. (2018). However, further research is required as there is no single perfect definition of interpretability but rather multiple, depending on the context (Doshi-Velez & Kim, 2017; Rudin, 2019).

4. Ignoring Feature Dependence

4.1. Interpretation with Extrapolation

Pitfall: When features are dependent, perturbation-based IML methods such as the PFI and PDP extrapolate in areas where the model was trained with little or no training data, which can cause misleading interpretations (Hooker & Mentch, 2019). Perturbations produce artificial data points that are used for model predictions, which in turn are aggregated to produce global interpretations (Scholbeck et al., 2020). Feature values can be perturbed by replacing original values with values from an equidistant grid of that feature, with permuted or with randomly subsampled values (Casalicchio et al., 2019), or with quantiles. We highlight two major issues. First, if features are dependent, all three perturbation approaches produce unrealistic data points, i.e., the new data points are located outside of the multivariate joint distribution of the data (see Figure 1). Second, even if

features are independent, using an equidistant grid can produce unrealistic values for the feature of interest. Consider a feature that follows a skewed distribution with outliers. An equidistant grid would generate a lot of values in between outliers and non-outliers. In contrast to the grid-based approach, the other two approaches maintain the marginal distribution of the feature of interest.

Both issues can result in misleading interpretations (illustrative examples given in Hooker & Mentch (2019); Molnar et al. (2020)) since the model is evaluated in areas of the feature space with few or no observed data points, where model uncertainty can be expected to be very high. This issue is aggravated if global interpretation methods integrate over such points with the same weight and confidence as for much more realistic samples with high model confidence.

Solution: Before applying interpretation methods, practitioners should check for dependencies between features in the data, e.g., via descriptive statistics or measures of dependence (see Section 4.2). When it is unavoidable to include dependent features in the model, which is usually the case in ML scenarios, additional information regarding the strength and shape of the dependence structure should be provided. Sometimes alternative interpretation methods can be used as a workaround or to provide additional information. ALE (Apley & Zhu, 2016) plots are preferable to the PDP when visualizing feature effects of dependent features. For other methods such as the PFI, conditional variants exist (Molnar et al., 2020; Candes et al., 2018; Strobl et al., 2008). Note, however, that conditional interpretations are often different and should not be used as a substitute for unconditional interpretations (see Section 4.3). Furthermore, dependent features should not be interpreted separately but rather jointly. This can be achieved by visualizing, e.g., a 2-dimensional ALE plot of two dependent features, which, admittedly, only works for very low-dimensional combinations. We recommend using quantiles or randomly subsampled values over equidistant grids. By default, many implementations of interpretability methods use an equidistant grid to perturb feature values (Greenwell, 2017; Molnar et al., 2018; Pedregosa et al., 2011), although some also allow to use user-defined values.

Open Issues: A comprehensive comparison of strategies addressing extrapolation, and how they affect an interpretation method, is currently missing. This also includes studying interpretation methods and their conditional variants when they are applied to data with different dependence structures.

4.2. Confusing Correlation with Dependence

Pitfall: Features with a Pearson correlation coefficient (PCC) close to zero can still be dependent and cause misleading model interpretations (see Figure 2). While independence between two features implies that the PCC is zero, the converse is generally false. The PCC, which is often used to analyze dependence, only tracks linear correlations and has other shortcomings such as sensitivity to outliers (Tjstheim et al., 2018). Any type of dependence between features can have a strong impact on the interpretation of the results of IML methods (see Section 4.1). Thus, knowledge about the (possibly non-linear) dependencies between features is crucial for an informed use of IML methods.

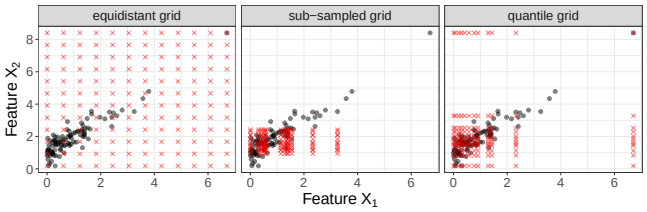

Figure 1. Illustration of artificial data points generated by three different perturbation approaches. The black dots refer to observed data points and the red crosses to the artificial data points.

Solution: Low-dimensional data can be visualized to detect dependence (e.g., scatter plots) (Matejka & Fitzmaurice, 2017). For high-dimensional data, several other measures of dependence in addition to PCC can be used. If dependence is monotonic, Spearman's rank correlation coefficient (Liebetrau, 1983) can be a simple, robust alternative to PCC. For categorical or mixed features, separate dependence measures have been proposed, such as Kendall's tau for ordinal features, or the phi coefficient and Goodman & Kruskals lambda for nominal features. (Khamis, 2008)

Studying non-linear dependencies is more difficult since a vast variety of possible associations have to be checked. Nevertheless, several non-linear association measures with sound statistical properties exist. Kernel-based measures such as kernel canonical correlation analysis (KCCA) (Bach & Jordan, 2002) or the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005) are commonly used. They have a solid theoretical foundation, are computationally feasible and robust (Tjstheim et al., 2018). In addition, there are information-theoretical measures such as (conditional) mutual information (Cover & Thomas, 2012) or the maximal information coefficient (MIC) (Reshef et al., 2011), that can however be difficult to estimate (Walters-Williams & Li, 2009; Belghazi et al., 2018). Other important measures are, e.g., the distance correlation (Székely et al., 2007), the randomized dependence coefficient (RDC) (Lopez-Paz et al., 2013), or the alternating conditional expectations (ACE) algorithm (Breiman & Friedman, 1985). In addition to using PCC we recommend using at least one measure that detects non-linear dependencies (e.g. HSIC).

4.3. Misunderstanding Conditional Interpretation

Pitfall: Conditional variants to estimate feature effects and importance scores require a different interpretation. While

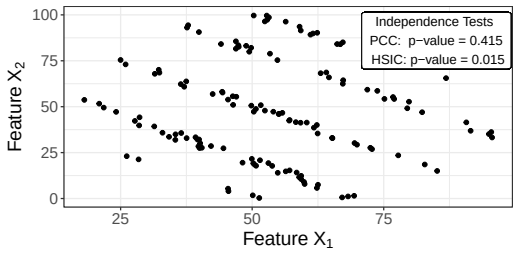

Figure 2. Highly dependent features X_1 and X_2 that have a correlation close to zero. A test (H_0: Features are independent) using Pearson correlation is not significant, but for HSIC the H_0-hypothesis gets rejected. Data from Matejka & Fitzmaurice (2017).

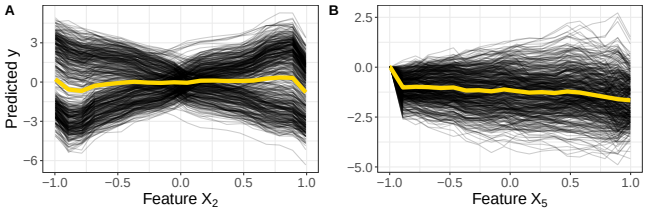

Figure 3. Simulation example with interactions: $y = 0.2 \cdot X_1 - 5 \cdot X_2 + 10 \cdot X_2 \mathbf{I}_{X_3>0} + 2 \cdot X_4 \cdot X_5 + \epsilon_i$ with $X_1, \ldots, X_5 \overset{i.i.d.}{\sim} U[-1, 1]$ and $\epsilon_i \overset{i.i.d.}{\sim} N(0, 1)$. **A:** PDP and ICE curves of X_2; **B:** PDP and centered ICE curves of X_5.

conditional variants for feature effects, e.g., the marginal plot (Apley & Zhu, 2016), feature importance scores (Candes et al., 2018; Watson & Wright, 2019; Molnar et al., 2020; Strobl et al., 2008), and conditional Shapley values (Lundberg et al., 2018) avoid model extrapolations, these methods answer a different question and have been argued to violate fundamental properties in the case of Shapley values (Janzing et al., 2019; Sundararajan & Najmi, 2019). Interpretation methods that perturb features independently of others also yield an unconditional interpretation, i.e., for feature effect methods such as the PDP, the effect can be interpreted as the isolated, average effect the feature has on the prediction. For the PFI, the importance can be interpreted as the drop in performance when the feature's information is "destroyed" (by perturbing it). Conditional variants do not replace values independently of other features, but in such a way that they conform to the conditional distribution. This changes the interpretation as the effects of all dependent features become entangled[3].

For dependent features, the PFI drops when using conditional variants since the conditional permutation answers the question: "How much does the model performance drop if we permute a feature, but given that we know the values of the other features?"[4].

To demonstrate how the interpretation can change, we trained a random forest to predict bike rentals (Fanaee-T & Gama, 2013), using the features "Temperature", "Apparent Temperature" and "Humidity". Temperature and the apparent temperature are highly linearly correlated, with a Pearson correlation coefficient of 0.992. The importance scores (measured as drop in mean absolute error) of the temperature (PFI 729; conditional PFI 285) and the apparent temperature (689; 266) drop considerably when using the conditional PFI instead of the marginal PFI. For the humidity, the importance scores of both variants are similar (578; 597).

Solution: The safest option would be to remove dependent features, but this is usually infeasible in practice. When features are highly dependent and conditional effects and importance scores are used, the practitioner has to be aware of the distinct interpretation. For feature effects, ALE plots (Apley & Zhu, 2016) provide an alternative with an unconditional interpretation. However, they only allow for an interval-wise interpretation.

Open Issues: Currently, no approach allows to simultaneously avoid model extrapolations and to allow a conditional interpretation of effects and importance scores for dependent features.

5. Misleading Effect due to Interactions

Pitfall: Global interpretation methods such as PDP or ALE plots can produce misleading interpretations when features interact. Figure 3 shows two examples where the global aggregated effects show almost no influence on the target, although an effect is clearly there by construction.

Solution: For the PDP, we recommend to additionally consider the corresponding ICE curves (Goldstein et al., 2015). While PDP and ALE average out interaction effects, ICE curves directly show the heterogeneity between individual predictions, as in Figure 3 A. Particularly for continuous interactions with ICE curves starting on different predictions, we recommend the use of derivative or centered ICE curves, which eliminate differences in intercepts and leave only differences due to interactions (Goldstein et al., 2015). As an example the diverging centered ICE curves of X_5 in Figure 3 B indicate that there must be an interaction with another feature. Other visualization techniques for discovering second-order interactions are 2-dimensional PDP or ALE plots and methods based on clustering ICE curves such as Visual Interaction Effects (VINE) (Britton, 2019).

Pitfall: Many interpretation methods cannot separate interactions from main effects. The PFI, for example, includes

[3]E.g., a feature that did not show an effect in the PDP might show an effect when using the marginal plot, when a dependent feature impacts the prediction.

[4]E.g., two highly dependent features might be individually important (based on the unconditional PFI), but have a very low conditional importance, since the information of one feature is contained in the other and vice versa.

both the importance of a feature and the importance of all its interactions with other features (Casalicchio et al., 2019).
Solution: Based on a PDP decomposition, the H-Statistic (Friedman & Popescu, 2008) quantifies the interaction strength between two features or between one feature and all others. Another similar interaction score based on partial dependencies is defined by Greenwell et al. (2018). Based on Shapley values Lundberg et al. (2018) proposed SHAP interaction values and Casalicchio et al. (2019) proposed a fair attribution of the importance of interactions to the individual features.
Open issues: Most methods that identify and visualize interactions are not able to identify higher-order interactions and interactions of dependent features. Instead of 2-dimensional PDPs, practitioners can use 2-dimensional ALE plots to visualize two-way interactions of dependent features. Furthermore, Hooker (2007) considers dependent features and decomposes the predictions in main and interaction effects. A way to identify higher-order interactions is shown in Hooker (2004). However, these issues are still a matter of further research. Furthermore, the presented solutions lack in automatic detection and ranking of all interactions of a model as well as specifying the type of modelled interaction.

6. Ignoring Estimation Uncertainty

Pitfall: Due to variance in the estimation process, interpretations of ML models can become misleading. Methods such as PDP and PFI use Monte Carlo sampling techniques to approximate expected values. These estimates vary, depending on the data used for the estimation. In particular, estimates may vary strongly for feature dependencies and interactions. Furthermore, the obtained ML model is also a random variable, as it is generated on randomly sampled data and the inducing algorithm might contain stochastic components as well. Hence, model variance has to be taken into account. The true effect of a feature may be flat, but purely by chance, especially on smaller data, an effect might algorithmically be detected. This effect could cancel out once averaged over multiple model fits. Figure 4 shows that a single PDP can be misleading because it does not show the variance due to PDP estimation and model fitting.
Solution: By repeatedly computing PDP and PFI with a given model, but with different permutations/bootstrap samples, the uncertainty of the estimate can be quantified, for example in the form of confidence intervals. For PFI, frameworks for confidence intervals and hypothesis tests exist (Watson & Wright, 2019; Altmann et al., 2010), but they assume a fixed model. If the practitioner wants to condition the analysis on the modeling process and capture the process' variance instead of conditioning on a fixed model, PDP and PFI should be computed on multiple model fits.
Open Issues: To the best of our knowledge, the uncertainty in feature effect methods such as ALE (Apley & Zhu, 2016) and PDP (Friedman et al., 1991) has not been studied in detail.

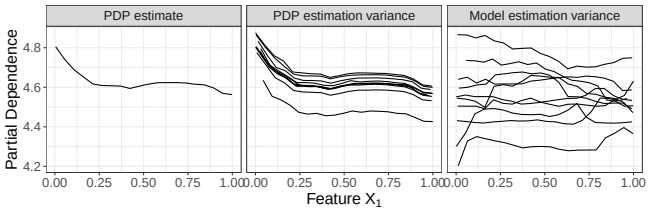

Figure 4. PDP for X_1 with $y = 0 \cdot X_1 + \sum_{j=2}^{10} X_j + \epsilon_i$ with $X_1, \ldots, X_{10} \sim U[0, 1]$ and $\epsilon_i \sim N(0, 0.9)$. **Left:** PDP for X_1 of a random forest trained on 100 data points. **Middle:** Multiple PDPs (10x) for the model from left plots, but with different samples (each n=100) for PDP estimation. **Right:** Repeated (10x) data samples of n=100 and newly fitted random forest.

7. Ignoring Multiple Comparisons

Pitfall: Simultaneously testing the importance of multiple features will result in false positive interpretations if the multiple comparisons problem (MCP) is ignored. The MCP is well known in significance tests for linear models and similarly exists in testing for feature importance in ML. For example, suppose we simultaneously test the importance of 50 features (with the H_0-hypothesis of zero importance) at the significance level 0.05. Even if all features are unimportant, the probability of observing that at least one feature is significantly important is $1 - \mathbb{P}(\text{'no feature important'}) = 1 - (1 - 0.05)^{50} \approx 0.923$. Multiple comparisons will even be more problematic, the higher dimensional our dataset is.
Solution: Methods such as Model-X knockoffs (Candes et al., 2018) directly control for the false discovery rate (FDR). For all other methods that provide p-values or confidence intervals, such as PIMP (Altmann et al., 2010), MCP is often ignored in practice to the best of our knowledge. Exceptions are, e.g., Stachl et al. (2019) and Watson & Wright (2019). One of the most popular MCP adjustment methods is the Bonferroni correction (Dunn, 1961), but it has the major disadvantage of increasing the probability of false negatives (Perneger, 1998). Since MCP is well known in statistics, we refer the practitioner to Dickhaus (2014) for an overview and discussion of alternative adjustment methods such as the Bonferroni-Holm method (Holm, 1979).

8. Unjustified Causal Interpretation

Pitfall: Practitioners are often interested in causal insights into the underlying data generating mechanisms, which IML methods in general do not provide. Common causal questions include the identification of causes and effects, predicting the effects of interventions, and answering counterfac-

tual questions (Pearl & Mackenzie, 2018). E.g., a medical researcher might want to identify risk factors or predict average and individual treatment effects (Knig & Grosse-Wentrup, 2019). In search for answers, a researcher can therefore be tempted to interpret the result of IML methods from a causal perspective.

However, a causal interpretation of predictive models is often not possible. Standard supervised ML models are not designed to model causal relationships but to merely exploit associations. A model may therefore rely on causes and effects of the target variable as well as on variables that help to reconstruct unobserved influences on Y, e.g., causes of effects (Weichwald et al., 2015). Consequently, the question whether a variable is relevant to a predictive model (indicated, e.g., by PFI > 0) does not directly indicate whether a variable is a cause, an effect or does not stand in any causal relation to the target variable.

Furthermore, even if a model would rely solely on direct causes for the prediction, the causal structure between features has to be taken into account. Intervening on a variable in the real world may affect not only Y but also other variables in the feature set. Without assumptions about the underlying causal structure IML methods cannot account for these adaptions and guide action (Karimi et al., 2020).

As an example, we constructed a dataset by sampling from a structural causal model (SCM), for which the corresponding causal graph is depicted in Figure 5. All relationships are linear Gaussian with variance 1 and coefficients 1. For a linear model fitted on the dataset all features were considered relevant based on the model coefficients ($\hat{y} = 0.329x_1 + 0.323x_2 - 0.327x_3 + 0.342x_4 + 0.334x_5$, $R^2 = 0.943$), although x_3, x_4 and x_5 do not cause Y.

Solution: The practitioner has to carefully assess whether sufficient assumptions can be made about the underlying data generating process, the learned model and the interpretation technique. If these assumptions are met, a causal interpretation may be possible. The PDP between a feature and the target can be interpreted as the respective average causal effect if the model performs well and the set of remaining variables is a valid adjustment set (Zhao & Hastie, 2019). When it is known whether a model is deployed in a causal or anti-causal setting, i.e., whether the models attempts to predict an effect from its causes or the other way round, a partial identification of the causal roles based on feature relevance is possible (under strong and non-testable assumptions) (Weichwald et al., 2015). Designated tools and approaches are available for causal discovery and inference (Peters et al., 2017).

Open issues: The challenge of causal discovery and inference remains an open key issue in the field of machine learning. Careful research is required to make explicit under which assumptions what insight about the underlying data generating mechanism can be gained by interpreting a machine learning model.

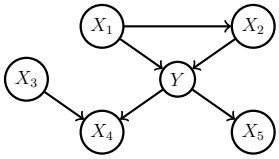

Figure 5. Causal graph

9. Discussion

In this paper, we have reviewed numerous pitfalls of global model-agnostic interpretation techniques, e.g., in the case of bad model generalization, dependent features, interactions between features, or causal interpretations. Although these pitfalls are far from complete, we believe that we cover common ones that pose a particularly high risk. We hope to encourage a more cautious approach when interpreting ML models in practice, to point practitioners to already (partially) available solutions and to stimulate further research on these issues. The stakes are high: ML algorithms are increasingly used for socially relevant decisions, and model interpretations play an important role in every empirical science. We therefore believe that users need concrete guidance on properties, dangers and problems of IML techniques – especially as the field is advancing at high speed. We need to strive towards a recommended, well-understood set of tools, which will require much more careful research. This especially concerns the meta-issues of comparisons of IML techniques, IML diagnostic tools to warn against misleading interpretations, and tools for analyzing multiple dependent or interacting features.

Acknowledgements

This work is funded by the Bavarian State Ministry of Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B) and supported by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A and the Graduate School of Systemic Neurosciences (GSN) Munich. The authors of this work take full responsibility for its content.

References

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010. doi: 10.1093/bioinformatics/btq134.

Anchang, J. Y., Prihodko, L., Ji, W., Kumar, S. S., Ross, C. W., Yu, Q., Lind, B., Sarr, M. A., Diouf, A. A., and Hanan, N. P. Toward operational mapping of woody canopy cover in tropical savannas using google earth engine. *Frontiers in Environmental Science*, 2020. doi: 10.3389/fenvs.2020.00004.

Apley, D. W. and Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*, 2016.

Arlot, S. and Celisse, A. A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79, 2010. doi: 10.1214/09-SS054. URL https://doi.org/10.1214/09-SS054.

Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(Jul):1–48, 2002. URL https://dl.acm.org/doi/abs/10.1162/153244303768966085.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003. doi: 10.1057/palgrave.jors.2601545.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International Conference on Machine Learning*, pp. 531–540, 2018.

Bischl, B., Mersmann, O., Trautmann, H., and Weihs, C. Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation*, 20(2):249–275, 2012.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001. doi: 10.1023/A:1010933404324.

Breiman, L. and Friedman, J. H. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985. doi: 10.1080/01621459.1985.10478157.

Britton, M. Vine: Visualizing statistical interactions in black box models. *arXiv preprint arXiv:1904.00561*, 2019.

Candes, E., Fan, Y., Janson, L., and Lv, J. Panning for gold:model-xknockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018. doi: 10.1111/rssb.12265.

Casalicchio, G., Molnar, C., and Bischl, B. Visualizing the feature importance for black box models. In Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., and Ifrim, G. (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 655–670, Cham, 2019. Springer International Publishing. doi: 10.1007/978-3-030-10925-7_40.

Claeskens, G., Hjort, N. L., et al. Model selection and model averaging. *Cambridge Books*, 2008. doi: 10.1017/CBO9780511790485.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2012.

Dandl, S., Molnar, C., Binder, M., and Bischl, B. Multi-objective counterfactual explanations. *arXiv preprint arXiv:2004.11165*, 2020.

Dickhaus, T. *Simultaneous Statistical Inference*. Springer-Verlag Berlin Heidelberg, 2014. doi: 10.1007/978-3-642-45182-9.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. *Regression: Models, Methods and Applications*. Springer-Verlag, Berlin, 2013. doi: 10.1007/978-3-642-34333-9.

Fanaee-T, H. and Gama, J. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pp. 1–15, 2013. doi: 10.1007/s13748-013-0040-3.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, 15(1):3133–3181, 2014. URL https://dl.acm.org/doi/10.5555/2627435.2697065.

Fisher, A., Rudin, C., and Dominici, F. All models are wrong, but many are useful: Learning a variables importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. URL http://jmlr.org/papers/v20/18-760.html.

Fountain-Jones, N. M., Machado, G., Carver, S., Packer, C., Recamonde-Mendoza, M., and Craft, M. E. How to make more from exposure data? an integrated machine learning pipeline to predict pathogen exposure. *Journal of Animal Ecology*, 88(10):1447–1461, 2019. doi: 10.1111/1365-2656.13076.

Friedman, J. H. and Popescu, B. E. Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2(3): 916–954, 09 2008. doi: 10.1214/07-AOAS148.

Friedman, J. H. et al. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991. doi: 10.1214/aos/1176347963.

Goldstein, A., Kapelner, A., Bleich, J., and 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, 2015. doi: 10.1080/10618600.2014.907095.

Good, P. I. and Hardin, J. W. *Common errors in statistics (and how to avoid them)*. John Wiley & Sons, 2012. doi: 10.1002/9781118360125.

Greenwell, B. M. pdp: An R package for constructing partial dependence plots. *The R Journal*, 9(1):421–436, 2017. doi: 10.32614/RJ-2017-016.

Greenwell, B. M., Boehmke, B. C., and McCarthy, A. J. A simple and effective model-based variable importance measure. *arXiv:1805.04755*, 2018.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, pp. 63–77. Springer, 2005. doi: 10.1007/11564089_7.

Hall, P. On the art and science of machine learning explanations. *arXiv preprint arXiv:1810.02909*, 2018.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL http://www.jstor.org/stable/4615733.

Hooker, G. Discovering additive structure in black box functions. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 04, pp. 575580, New York, NY, USA, 2004. Association for Computing Machinery. doi: 10.1145/1014052.1014122.

Hooker, G. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3): 709–732, 2007. doi: 10.1198/106186007X237892.

Hooker, G. and Mentch, L. Please stop permuting features: An explanation and alternatives. *arXiv preprint arXiv:1905.03151*, 2019.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causality problem. *arXiv preprint arXiv:1910.13413*, 2019.

Karimi, A.-H., Schlkopf, B., and Valera, I. Algorithmic Recourse: from Counterfactual Explanations to Interventions. *arXiv: 2002.06278*, 2020.

Khamis, H. Measures of association: how to choose? *Journal of Diagnostic Medical Sonography*, 24(3):155–162, 2008. doi: 10.1177/8756479308317006.

Krishnan, M. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 08 2019. doi: 10.1007/s13347-019-00372-9.

Kuhle, S., Maguire, B., Zhang, H., Hamilton, D., Allen, A. C., Joseph, K., and Allen, V. M. Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. *BMC Pregnancy and Childbirth*, 18(1):1–9, 2018. doi: 10.1186/s12884-018-1971-2.

Knig, G. and Grosse-Wentrup, M. A Causal Perspective on Challenges for AI in Precision Medicine, 2019. URL https://koenig.page/pdf/koenig2019_pmbc.pdf.

Liebetrau, A. *Measures of Association*. Number Bd. 32;Bd. 1983 in 07. SAGE Publications, 1983.

Lipton, Z. C. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018. URL https://dl.acm.org/doi/10.1145/3236386.3241340.

Lopez-Paz, D., Hennig, P., and Schölkopf, B. The randomized dependence coefficient. In *Advances in Neural Information Processing Systems*, pp. 1–9, 2013. URL https://dl.acm.org/doi/10.5555/2999611.2999612.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3), 2018. doi: 10.1371/journal.pone.0194889.

Matejka, J. and Fitzmaurice, G. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1290–1294, 2017. doi: 10.1145/3025453.3025912.

Molnar, C., Casalicchio, G., and Bischl, B. iml: An R package for interpretable machine learning. *Journal of Open Source Software*, 3(26):786, 2018. doi: 10.21105/joss.00786.

Molnar, C., Casalicchio, G., and Bischl, B. Quantifying model complexity via functional decomposition for better post-hoc interpretability. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 193–204. Springer, 2019. doi: 10.1007/978-3-030-43823-4_17.

Molnar, C., König, G., Bischl, B., and Casalicchio, G. Model-agnostic feature importance and effects with dependent features–a conditional subgroup approach. *arXiv preprint arXiv:2006.04628*, 2020.

Pearl, J. and Mackenzie, D. The ladder of causation. *The book of why: the new science of cause and effect. New York (NY): Basic Books*, pp. 23–52, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://dl.acm.org/doi/10.5555/1953048.2078195.

Perneger, T. V. What's wrong with bonferroni adjustments. *BMJ*, 316(7139):1236–1238, 1998. doi: 10.1136/bmj.316.7139.1236.

Peters, J., Janzing, D., and Scholkopf, B. *Elements of Causal Inference - Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Philipp, M., Rusch, T., Hornik, K., and Strobl, C. Measuring the stability of results from supervised statistical learning. *Journal of Computational and Graphical Statistics*, 27(4):685–700, 2018. doi: 10.1080/10618600.2018.1473779.

Posada-Quintero, H. F., Molano-Vergara, P. N., Parra-Hernández, R. M., and Posada-Quintero, J. I. Analysis of risk factors and symptoms of burnout syndrome in colombian school teachers under statutes 2277 and 1278 using machine learning interpretation. *Social Sciences*, 9(3):30, 2020. doi: 10.3390/socsci9030030.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011. doi: 10.1126/science.1205438.

Ribeiro, M. T., Singh, S., and 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. ACM, 2016. doi: 10.1145/2939672.2939778.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Scholbeck, C. A., Molnar, C., Heumann, C., Bischl, B., and Casalicchio, G. Sampling, intervention, prediction, aggregation: A generalized framework for model-agnostic interpretations. *Communications in Computer and Information Science*, pp. 205216, 2020. doi: 10.1007/978-3-030-43823-4_18.

Shahhosseini, M., Hu, G., and Archontoulis, S. V. Forecasting corn yield with machine learning ensembles. *arXiv preprint arXiv:2001.09055*, 2020.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Simon, R. Resampling strategies for model assessment and selection. In *Fundamentals of data mining in genomics and proteomics*, pp. 173–186. Springer, 2007.

Stachl, C., Au, Q., Schoedel, R., Buschek, D., Völkel, S., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., et al. Behavioral patterns in smartphone usage predict big five personality traits. 2019. doi: 10.31234/osf.io/ks4vd.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008. doi: 10.1186/1471-2105-9-307.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. *arXiv preprint arXiv:1908.08474*, 2019.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007. doi: 10.1214/009053607000000505.

Tjstheim, D., Otneim, H., and Stve, B. Statistical dependence: Beyond pearson's p. *arXiv preprint arXiv:1809.10455*, 2018.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Walters-Williams, J. and Li, Y. Estimation of mutual information: A survey. In *International Conference on Rough Sets and Knowledge Technology*, pp. 389–396. Springer, 2009.

Watson, D. S. and Wright, M. N. Testing Conditional Independence in Supervised Learning Algorithms. *arXiv preprint arXiv:1901.09917*, 2019.

Weichwald, S., Meyer, T., zdenizci, O., Schlkopf, B., Ball, T., and Grosse-Wentrup, M. Causal interpretation rules

for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59, 2015. doi: 10.1016/j.neuroimage.2015.01.036.

Wu, J., Roy, J., and Stewart, W. F. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical Care*, pp. S106–S113, 2010. doi: 10.1097/MLR.0b013e3181de9e17.

Zhang, Z., Jin, Y., Chen, B., and Brown, P. California almond yield prediction at the orchard level with a machine learning approach. *Frontiers in Plant Science*, 10:809, 2019. doi: 10.3389/fpls.2019.00809.

Zhao, Q. and Hastie, T. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, pp. 1–10, 2019. doi: 10.1080/07350015.2019.1624293.

Zhao, X., Lovreglio, R., and Nilsson, D. Modelling and interpreting pre-evacuation decision-making using machine learning. *Automation in Construction*, 113:103140, 2020. doi: 10.1016/j.autcon.2020.103140.