Evaluation of post-hoc XAI approaches through synthetic tabular data

Julian Tritscher[1], Markus Ring[2], Daniel Schlör[1], Lena Hettinger[1], and Andreas Hotho[1]

[1] University of Würzburg, Germany
{tritscher, schloer, hettinger, hotho}@informatik.uni-wuerzburg.de
[2] University of Coburg, Germany
markus.ring@hs-coburg.de

Abstract. Evaluating the explanations given by post-hoc XAI approaches on tabular data is a challenging prospect, since the subjective judgement of explanations of tabular relations is non trivial in contrast to e.g. the judgement of image heatmap explanations. In order to quantify XAI performance on categorical tabular data, where feature relationships can often be described by Boolean functions, we propose an evaluation setting through generation of synthetic datasets. To create gold standard explanations, we present a definition of feature relevance in Boolean functions. In the proposed setting we evaluate eight state-of-the-art XAI approaches and gain novel insights into XAI performance on categorical tabular data. We find that the investigated approaches often fail to faithfully explain even basic relationships within categorical data.

Keywords: Explainable AI · Evaluation · Synthetic data

1 Introduction

Black box classifiers such as deep neural networks (DNNs) have been established as state-of-the-art in many machine learning areas. Even though they give strong predictions, their models contain lots of non-linear dependencies, causing their decisions to become untraceable. As a result, a branch of research in explainable artificial intelligence (XAI) has developed, aiming to give local explanations for single predictions of trained black box models in a post-hoc fashion [6].

Problem. While there are many approaches to acquire such explanations, no unified evaluation method has been proposed so far. While easily comprehensible domains like image and text classification use simple presentation of explanations [3] and user studies [6], XAI behavior on tabular data is largely unexplored.

Objective. Post-hoc XAI explanations are inherently approximations, giving simplified but not necessarily faithful insights into highly complex models [11]. Therefore, assessing the limits of these approaches is a central point of research interest. We take a first step to investigate post-hoc XAI performance on DNNs trained on tabular data, by developing a test setting for categorical tabular data, where feature relationships can often be expressed via Boolean functions.

Approach and Contribution. In this paper, we design synthetic datasets reflecting typical relationships of real-world data such as logical AND, OR and XOR connections between categorical attributes. We propose a definition of feature importance in Boolean functions in the context of XAI to generate gold-standard explanations for our datasets. Using this data, we present an evaluation setting for XAI approaches, that allows for evaluation of data with underlying complex feature relationships. This setting is used to analyze and compare eight state-of-the-art XAI approaches. Evaluation on an expert-annotated real dataset suggests that our results translate well to real data. We publish our datsets to facilitate comparison of XAI approaches in a standardized evaluation setting.

2 Related Work

Aside from subjective image explanations [3] and resource intensive user studies [6], several approaches have been used to evaluate XAI performance.

In [5,9] model faithfulness of their approaches is evaluated by explaining predictions of inherently explainable linear models and comparing obtained explanations directly to the model. Performance of XAI approaches when explaining non-linear classifiers, however, can not be assessed in this evaluation setting.

Additionally, [9] also evaluate the faithfulness of their local approximation model by measuring if it corresponds to changed inputs in the same way as the classifier it approximates. This evaluation can, however, only be performed on XAI approaches that train a simplified classifier as a local approximation.

Perturbation-based evaluations, e.g. as used in [12], iteratively remove features with the highest relevance from data and re-classify. Explanations are rated higher, the faster the classification error increases. While perturbation-based evaluation can be applied to classification tasks where the removal of single features is expected to gradually impair the performance of the classifier, this assumption does not hold for categorical tabular data in general.

3 Investigated XAI Approaches

Perturbation-based explanation approaches mask or remove input features from data samples to observe the change in classifier output. While they pose no architectural constraints on the classifier, they are computationally intensive. LIME (Local Interpretable Model-agnostic Explanations) [9] uses perturbations to explore the classifier outputs locally around a given input. It trains a local linear classifier and uses the weights as scores of input feature relevance. Shapley Value sampling [4] is based on the Shapley value from cooperative game theory, a unique solution for distributing an achieved score onto cooperating players, under a list of desirable criteria. Shapley value sampling approximates this NP-complete problem with a sampling approach, evaluating the output of possible feature value combinations through perturbation. (Kernel) SHAP (Kernel SHapley Additive exPlanation) [6] uses a custom kernel for the LIME XAI approach, in order to adapt LIME to approximate Shapley values.

Gradient-based XAI approaches use the gradient of a gradient descent based classifier to approximate explanations through few backpropagations, considerably saving runtime in comparison to perturbation-based approaches. For our evaluation, we use the implementations of [1]. Saliency maps [14] highlight the most influential pixels using a first order approximation of the absolute gradient of the predicted output with respect to the input for a specific data sample. Gradient×Input [13] builds on the Saliency approach, multiplying the signed result of Saliency with the corresponding input feature. Integrated Gradients [15] computes the average output gradients with respect to different inputs. Gradients are computed for values on a linear path between the data sample and an uninformative baseline input. ϵ-LRP (ϵ-Layerwise Relevance Propagation) [3] defines the relevance of a neuron as all influence it has on the neurons of the next layer, multiplied by these neurons' activations for a specific data sample. In this work we use the reformulated implementation by [1]. DeepLIFT [12], like LRP, computes the relevance of a neuron by measuring the influence on neurons of the next layer, additionally subtracting the influence of an uninformative baseline.

4 Data Generation Approach

Since categorical attributes can be binarized (e.g. one-hot encoding) to Boolean features, we will focus on feature relationships modeled as Boolean functions.

In [7] a feature is considered influential in a sample if changing its value would also change the function output. While this is intuitive on some inputs, on others it assigns no influence to any feature. For example, consider the Boolean function $y = 0 \wedge 0 = 0$, where no single feature can be changed to change y. In the context of XAI, this would not allow to differentiate between the 0-inputs involved in the function and irrelevant features. To address this, we adapt the influence definition on basic Boolean operations (AND, OR, XOR) to assign influence to both features, if no single feature can be changed to change the function output. For more complex functions, we proceed as follows:

Definition 1. *Let the Boolean function y be represented by a Boolean binary expression tree [8]. For each data sample, we consider a child node c relevant to the explanation of its parent operation-node o, iff the value of the subtree formed by c, evaluated with respect to the sample, is relevant to the operation-node o.*

We thereby distribute the relevance of a complex function by decomposing it into its basic operations. We calculate their intermediate results for each data sample, propagating the relevance of the entire function through all basic operations down to its input features. The resulting explanations contain the input features that most determine the output of complex Boolean functions.

When assessing XAI performance, we have to take into account that non-matching explanations might be correct explanations of a weak classifier, instead of a poorly performing XAI approach. For this, we train our classifiers in a 5-fold stratified cross-validation setting, using only classifiers that reliably achieve 100% accuracy on training- and test-sets. Further, we generate synthetic datasets

including every permutation of categorical attributes exactly once. This guarantees that the test-sets contain permutations not seen during training, ensuring that the classifier learned to perfectly generalize to the unseen test-data without sensitivity to irrelevant inputs.

Following these restrictions, we expect XAI approaches to give the highest scores to the relevant features. Thus, we consider a data sample to be correctly explained, if the top scoring features given by an XAI approach match the relevant features of the ground truth explanation.

5 Experiments

The following setup is used throughout all of our experiments.

Datasets are generated following the criteria of Sect. 4. We set a fixed dataset size of n binary features, for which we include every permutation once in the dataset, giving 2^n data samples. We then generate the label for each data sample with a Boolean function and generate the explanation of every data sample according to Sect. 4. Features that were not used in the generation of the label hereby act as noise that XAI approaches may falsely consider relevant. All following experiments use 12 binary features and $2^{12} = 4096$ data samples.

Classifier & XAI setup also follow Sect. 4. We encode our Boolean input data with the values 1 for True and -1 for False, and train a feed forward neural network with 5 layers, 20 neurons per layer, and ReLU activations in a 5-fold stratified cross-validation setting. The classifiers reliably achieve 100% accuracy on training- and test-sets for all evaluated datasets. For each cross-validation fold, we compute the explanations of the test-data. We repeat the evaluation 10 times per dataset, reporting the average over the results.

Baseline Some XAI approaches replace classifier input features with uninformative values to observe classifier behavior with missing information. We let LIME and SHAP extract their own baseline from the cross-validation training data, using k-means clustering with $k = 20$ for SHAP. For the gradient-based approaches, we use 0 as baseline value, as discussed in [2].

5.1 Evaluation of basic Boolean operations

We initially evaluate the behavior of XAI approaches on datasets where two features are linked by common Boolean operations AND ($\wedge$), OR ($\vee$) or XOR ($\otimes$) in Table 1. We find that most XAI approaches fail to fully explain even the linear Boolean AND and OR operations, with only LIME and SHAP finding the most relevant features for each sample. Results on the XOR dataset show that LIME, due to training a local linear model around the sample, fails to give good explanations when the underlying local function is non-linear. SHAP appears to improve on LIMEs behavior, correctly matching the gold standard explanations with its kernel-based Shapley value adaptation of LIME.

Detailed analysis We take a closer look at the input permutations causing problems to the XAI approaches. We find that falsely explained samples for all

Table 1: XAI performance in percent correctly explained samples after Def. 1.

approach	$\wedge$	$\vee$	$\otimes$	$(\otimes) \wedge (\otimes)$	synthetic	real
LIME	**100.00**	**100.00**	43.76	10.12	**100.00**	84.78
Shapley sampling	99.83	99.92	69.10	56.34	79.79	77.15
SHAP	**100.00**	**100.00**	**100.00**	**95.01**	98.12	**93.34**
Saliency	95.81	95.69	85.99	55.55	59.86	59.12
Gradient×input	97.66	97.06	75.81	57.31	69.47	69.49
Integrated gradients	99.64	99.56	73.02	57.73	76.46	76.27
ϵ-LRP	97.66	97.06	75.84	57.34	69.48	69.71
Deeplift	99.67	99.59	73.28	58.83	75.48	76.77

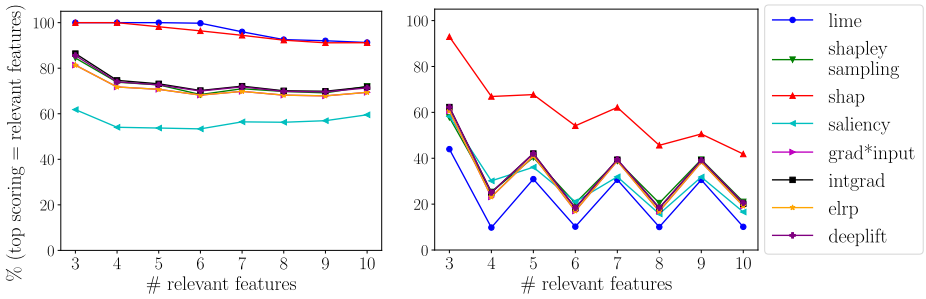
gradient-based approaches and Shapley sampling on the AND and OR datasets are caused by issues with $y = 0 \wedge 0$ and $y = 1 \vee 1$. In this case, the mentioned approaches consider one of the two features as irrelevant, even though both features are equally important to the label. Additionally, the Saliency approach shows issues on unequal inputs, where one feature speaks against the prediction outcome. Since the gradient of the output with respect to this feature is negative, the Saliency's absolute gradient causes this negatively influencial feature to overshadow the relevant feature. On the non-linearly separable XOR dataset, all approaches show a similar amount of errors for each input permutation.

5.2 Evaluation of Boolean functions with multiple variables

Using our relevance definition (Def. 1), we investigate XAI performance on complex Boolean functions. Results of the function $y = (f_1 \otimes f_2) \wedge (f_3 \otimes f_4)$, shown as $(\otimes) \wedge (\otimes)$ in Table 1, indicate that XAI performance deteriorates with an increased number of relevant features involved. To test this, we create similar datasets with increasing numbers of variables that may impact the output label.

Linearly separable Boolean functions. Since XAI performance on basic operations suggests different XAI behavior on linear and non-linear Boolean functions, we first investigate XAI performance with increasing function complexity on linear functions. For this, we generate eight datasets using the function $y = ((((f_1 \wedge f_2) \vee f_3) \wedge f_4) \vee ...)$, appending 3 to 10 relevant features as label. To validate that the used datasets are linearly separable, we ensure that a linear Support Vector Machine can perfectly separate each dataset.

The average results on each dataset are shown in Fig. 1a. We find LIME to be able to fully explain all samples of datasets with up to 6 relevant features. Both LIME and SHAP are capable of explaining a large amount of samples in all tested datasets. Shapley sampling and all gradient-based methods show difficulties with explaining functions with more then 2 variables involved, with performance declining further with more than 3 variables. The small inclines in explanation score with increased function complexity may be explained by all datasets consisting of a total of 12 variables. This means that when 10 variables



(a) Linear AND & OR operators (b) Non-linear XOR & AND operators

Fig. 1: XAI performance on datasets with multiple relevant features. Performance in percent of correctly explained samples according to Def. 1.

are involved in the function, randomly assigning the 2 non relevant variables the lowest scores may occur more often then with 6 relevant and irrelevant variables.

Non-linearly separable Boolean functions. We also evaluate performance on eight non-linearly separable datasets generated using the function $y = ((((f_1 \otimes f_2) \wedge f_3) \otimes f_4) \wedge ...)$ with 3 to 10 relevant features used for label generation. As seen in Fig. 1b, all approaches show lower performance compared to the non-linear XOR dataset (see Table 1). While SHAP still maintains stronger performance than other approaches on non-linear datasets throughout the experiment, its performance deteriorates when more than 3 features influence the label. All other approaches show poor performance on all datasets. The fluctuation between scores with increasing complexity may be caused by the alternating label-generation: If the last operator in the outermost brackets of the function is an AND, then for all samples that evaluate to $x \wedge 0$ with the entire previous term $x = 1$, the only relevant variable for this sample is the last 0. Therefore XAI approaches only have to find the last variable 0 as explanation, simplifying the problem down to a basic AND operation for several input permutations.

5.3 Application scenario

Next, we choose a setting from the intrusion detection domain, to show that our findings can be applied to realistic settings. A method to synthetically create flow-based network traffic is proposed in [10], where a flow describes a network connection between two hosts and contains attributes like transport protocol and TCP-flags. In this setting, only flows which represent TCP traffic are allowed to set any TCP-flags, creating the task of validating whether samples resemble valid network traffic. While transport protocol and TCP-flags can be represented as binary attributes with a set of predefined rules, the complexity of the problem is low enough to create expert-annotated labels. In this experiment, we use N-WGAN-GP from [10] to create 2048 correct and 2048 incorrect flows.

Each flow is represented by six categorical (*TCP flags*) and two categorical, one-hot-encoded features (*weekday, protocol*), six numerical features (*bytes, packets, duration, time, source- and destination-port*) and two values encoded with multiple numeric features (*IP-addresses*). The six *TCP flags* and the *protocol* are considered as relevant. To create a comparable synthetic setting, we generate a dataset using the function $y = (f_1 \vee -(f_2 \vee f_3 \vee f_4 \vee f_5 \vee f_6 \vee f_7))$. We then evaluate the XAI approaches on both datasets.

The results, marked as "synthetic" and "real" in Table 1, indicate a similar ranking of the XAI approaches with respect to their performance on both datasets. We observe that, while all perturbation-based approaches perform worse on real data, LIME achieves considerably better results on synthetic data in comparison to the real setting. This may be due to its local linear approximation that benefits more from the equal distribution of different sample permutations in the synthetic data. This experiment suggests that XAI performance on our synthetic datasets closely resembles real world application scenarios.

6 Discussion

Evaluation of eight post-hoc XAI approaches shows that many approaches fail to give satisfactory explanations even on basic categorical tabular data. The gradient-based approaches used in this work all show weaker performance than perturbation-based methods. While the approaches LIME and SHAP are both capable of well explaining basic linearly separable Boolean functions, only SHAP is capable of explaining non-linearly separable functions with up to 3 variables. Overall, we find that investigated approaches struggle to explain more complex, as well as non-linear Boolean functions. The datasets generated for these experiments may be used to gain first insights into XAI performance on categorical tabular data and will therefore be made available as benchmark datasets.[3]

7 Conclusion

In this paper, we investigated XAI performance on categorical tabular data, proposing a setting in which XAI approaches can be evaluated independently of classifier performance using synthetic datasets with gold standard explanations. We generated benchmark datasets containing typical relationships between binary attributes such as AND, OR and XOR, as well as explanations according to a novel definition of relevance of features in Boolean functions.

Using these datasets, we empirically evaluated eight state-of-the-art XAI approaches. We found that many approaches fail to capture simple feature relationships such as non-linear XOR connections, with performance decreasing with increasing relationship complexity. Overall, we found the tested gradient-based approaches to yield worse results than the perturbation-based methods. By evaluating an expert-annotated dataset from the intrusion detection domain

[3] http://www.dmir.uni-wuerzburg.de/projects/deepscan/xai-eval-data/

and comparing the results to explanations from synthetic data, we showed that the findings on our synthetic datasets can be applied to realistic data.

Acknowledgement

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany as part of the DeepScan project (01IS18045A).

References

1. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: A unified view of gradient-based attribution methods for deep neural networks. In: NIPS 2017 - Workshop on Interpreting, Explaining and Visualizing Deep Learning. ETH Zurich (2017)
2. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Gradient-based attribution methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 169–191. Springer (2019)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
4. Castro, J., Gómez, D., Tejada, J.: Polynomial calculation of the shapley value based on sampling. Computers & Operations Research **36**(5), 1726–1730 (2009)
5. Kindermans, P.J., Schütt, K.T., Alber, M., Müller, K.R., Erhan, D., Kim, B., Dähne, S.: Learning how to explain neural networks: Patternnet and patternattribution. In: Int. Conf. on Learning Representations (2018)
6. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. pp. 4765–4774 (2017)
7. O'Donnell, R.: Analysis of boolean functions. Cambridge University Press (2014)
8. Preiss, B.: Data Structures and Algorithms with Object-Oriented Design Patterns in Java (1999)
9. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data mining. pp. 1135–1144. ACM (2016)
10. Ring, M., Schlör, D., Landes, D., Hotho, A.: Flow-based Network Traffic Generation using Generative Adversarial Networks. Computer & Security **82**, 156–172 (2019)
11. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell **1**(5), 206 (2019)
12. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: 34th Int. Conf. on Machine Learning - Volume 70. pp. 3145–3153. JMLR. org (2017)
13. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 (2016)
14. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Bengio, Y., LeCun, Y. (eds.) ICLR (Workshop Poster) (2014)
15. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: 34th Int. Conf. on Machine Learning-Volume 70. pp. 3319–3328. JMLR. org (2017)