Explanation Groves – Controlling the Trade-off between the Degree of Explanation vs. its Complexity

Gero Szepannek[1,*,†]

[1]Stralsund University of Applied Sciences, Zur Schwedenschanze 15, 18435 Stralsund, Germany

Abstract

Regulatory requirements, such as the recently published EU AI Act, emphasize the need to explain machine learning models. Nonetheless, the limits of any given explanation have to be taken into account. From Psychology research it is known, that the human working memory capacity is limited. For this reason, any explanation must not be too complex. In this work, explanation groves are presented as a model agnostic tool to control the complexity of an explanation while simultanously maximizing the obtained degree of explanation. Explanation groves do result, if the degree of explanation is maximized over the search space of all sets of if-then rules of prespecified size. A user-friendly implementation of explanation groves is given in the R package `xgrove`. Its use is demonstrated for a random forest model trained on the Boston housing data. Explanation groves not only provide an easily understandable explanation but can be further used to analyze the trade-off between the obtained degree of explanation and its corresponding complexity.

Keywords

Model Agnostic XAI, Rule-based Explanations, Surrogate Models, Working Memory Capacity, Gradient Boosting

1. Introduction

Regulatory requirements, such as the recently published EU AI Act [1], emphasize the need to explain of machine learning models. A general process (TAX4CS) that links requirements of the different model stakeholders to existing methdology from the field of XAI in order to ensure transparent and auditable machine learning models in industry is proposed by [2]. Nonetheless, the limits of any given explanation have to be taken into account. From Psychology research it is known, that the human working memory capacity is limited (cf. e.g. [3, 4]). For this reason, any explanation must not be too complex. In this work, explanation groves are presented as a model agnostic tool to control the complexity of an explanation while simultanously maximizing the obtained degree of explanation.

In [5] explainable boosting machines are proposed which provide interpretable glassbox models based on the idea of generalized additive models. In contrast to this, the approach presented in this paper is dedicated to black box explanation and an explanation based on a set of additive weighted interpretable rules is derived. For random forest models, [6] and [7] aim to find the most representative tree (MRT) in a forest providing a set of interpretable rules. Further, in [8] the appropriateness of the resulting explanation is analyzed and it is worked out, that explanations can be improved by considering not only one single tree but a small number of representative trees, called groves, to explain forests. Note that for decision trees, a model simplification is obtained by pruning. Different strategies of pruning are investigated in [9]. Unlike MRTs, explanation groves proposed in this work, are model agnostic. It is shown that explanation groves do result, if the degree of explanation is maximized over the search space of all sets of if-then rules of prespecified size. For this purpose, a measure for the degree of explanation is required which is described in section 2 and proposed in [10]. A user-friendly implementation of explanation groves is given in the R package `xgrove` [11]. In section 3, its use is demonstrated: The method is applied to a random forest model trained on the Boston housing data. Explanation groves

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3d World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

gero.szepannek@hochschule-stralsund.de (G. Szepannek)
https://www.hochschule-stralsund.de/ws/personal-an-der-fakultaet-fuer-wirtschaft/szepannek-gero/ (G. Szepannek)
0000-0001-8456-1283 (G. Szepannek)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

not only provide an easily understandable explanation but can be further used to analyze the trade-off between the obtained degree of explanation and its corresponding complexity.

2. Explanation Groves

2.1. Measuring Explainability

In order to find the best explanation it has to be defined what characterizes a good explanation is. In literature, several concepts are proposed to analyze this (cf. e.g. [12]). For the purpose of this work the metric proposed in [13] is used: An explanation $XAI(x)$ is appropriate if it is close to the model of interest $\hat{f}(x)$ for any value of x. According to [13] this can be summarized by the expected squared difference:

$$ESD(XAI) = \int (\hat{f}(X) - XAI(X))^2 \, dP(X) \tag{1}$$

and a measure to quantify the appropriateness of an explanation is given by the degree of explanation:

$$\Upsilon = 1 - \frac{ESD(XAI)}{ESD_0} \tag{2}$$

where ESD_0 is the ESD based on $XAI(x) = c, \forall x$ being the constant average prediction $c := E(\hat{f}(X))$. By construction, Υ is similar to the R^2 coefficient of determination for regression problems and can be thus interpreted in a similar way: For a good explanation $ESD(XAI)$ will be close to 0 and thus the closer Υ is to the value of one the better the explanation.

2.2. Finding the Best Explanation

Based on the previous quantification of the appropriateness of an explanation, one can try to find the best explanation of by stagewise maximization of Υ. For this purpose, an iterative approach can be used. Let

$$XAI^{(m)}(x)$$

be the explanation after the m^{th} iteration, $m \in \mathbb{N}$. The ESD from the previous section measures the squared loss between the model's predictions and its explanation. As in gradient boosting theory [14] a greedy approach to minimizing the loss function is obtained by iteratively updating $XAI^{(m-1)}(x)$ into the direction of the steepest descent:

$$\begin{aligned}
& -\frac{\partial L(\hat{f}(x_i), XAI(x_i))}{\partial XAI(x_i)}\bigg|_{XAI(x_i)=XAI^{(m-1)}(x_i)} \\
&= -\frac{\partial(\hat{f}(x_i) - XAI(x_i))^2}{\partial XAI(x_i)}\bigg|_{XAI(x_i)=XAI^{(m-1)}(x_i)} \\
&= 2(\hat{f}(x_i) - XAI^{(m-1)}(x_i)) \\
&=: \tilde{y}_i
\end{aligned} \tag{3}$$

i.e. by iteratively fitting the pseudo-residuals $\tilde{y}_i$ between model and current explanation after stage $(m-1)$ to the data, where $i = 1, ..., N$ denotes the observation index. The optimal model-agnostic rule-based explanation is then given by the sum over a set of weighted rules, which can be written as:

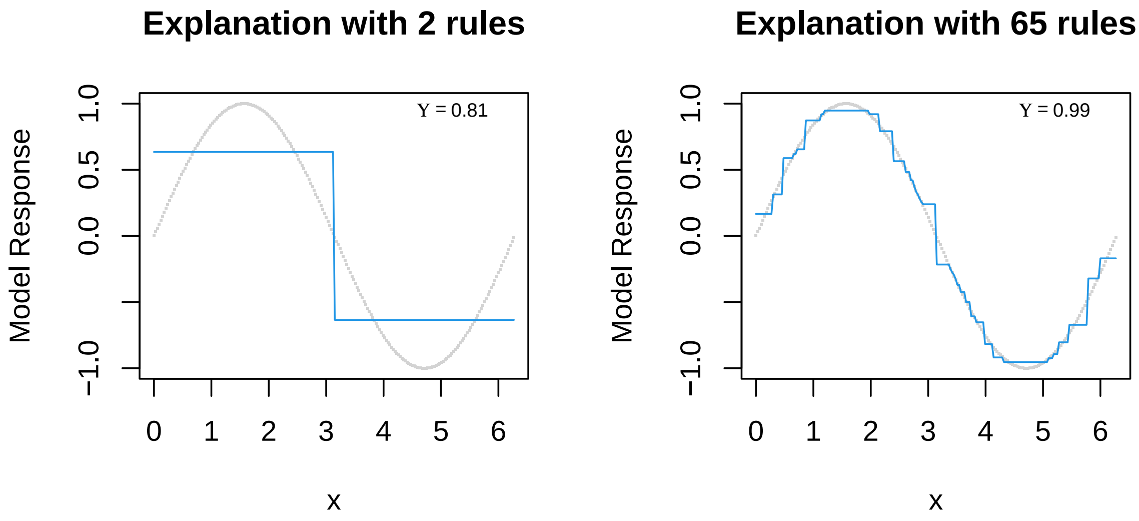

Figure 1: Simple example of explaining an artificial model f(x) = sin(x) by a set of 2 (left) and 65 (right) rules.

$$\begin{aligned} XAI^{(m)}(x) &= XAI^{(m-1)}(x) \\ &\quad + \gamma_{m+}\ \mathbf{1}_{(x \in R^{(m)})} \\ &\quad + \gamma_{m-}\ \mathbf{1}_{(x \notin R^{(m)})}. \end{aligned} \tag{4}$$

Here, $\mathbf{1}_{(x \in R^{(m)})}$ denotes the indicator function that returns 1 if rule $R^{(m)}$ holds for x and 0 otherwise. $R^{(m)}$ is a rule of the form $X_j \leq a$ in variable X_j for numeric variables or $X_j \in \mathcal{A}$ with $\mathcal{A} \subset |X_j|$ for categorical variables. γ_m is a weight that describes how the explanation changes, if rule $R^{(m)}$ holds. For this purpose, $R^{(m)}$, γ_{m+} and γ_{m-} can be computed simultaneously by fitting a gradient boosting model using squared loss and decision trees of depth one (stumps) to the predictions $\hat{f}(x_i)$ of the model of interest [15].

Note that the resulting optimal explanation $XAI(x)$ consists of a set of rules and corresponding weights $\{(R^{(m)}, \gamma_{m+}, \gamma_{m-})\}$ and thus represents a rule-based explanation as opposed to example-based explanations. For a comparison of both approaches for model explanation cf. e.g. [16]. The complexity of the resulting explanation is given by the number of rules and can be controlled by the number of iterations m.

2.3. Illustration

Figure 1 illustrates the aforementioned trade-off between complexity and appropriateness for explanation groves of different size. For an artificial (unknown) model $\hat{f}(x) = \sin(x)$ (grey dots) two explanation groves of different size are computed: The explanation on the left graph consists of only two rules (with a split at π) and is easy to understand. Although it captures the information that there are positive values below π and negative ones above, it is not close to the predictions of the model and thus not appropriate but a too simple explanation. In contrast, in the right plot, a grove of 65 rules, approximates the original model quite well (which is reflected by an $\Upsilon = 0.99$) but the corresponding large set of rules will be difficult to understand for humans.

Although this example is artificial, oversimplified and takes into account for only one single variable and not for the usual multivariate setting, it nicely illustrates the trade-off between adequacy and complexity of different explanations. In practice, several groves with different numbers of rules m can

Table 1
Coloumns of the Boston housing data.

Name	Description
cmedv	Corrected median value of owner-occupied homes in USD 1000’s.
lon	Longitude of census tract.
lat	Latitude of census tract.
crim	Per capita crime rate by town.
zn	Proportion of residential land zoned for lots over 25,000 sq.ft.
indus	Proportion of non-retail business acres per town.
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
nox	Nitric oxides concentration (parts per 10 million).
rm	Average number of rooms per dwelling.
age	Proportion of owner-occupied units built prior to 1940.
dis	Weighted distances to five Boston employment centers.
rad	Index of accessibility to radial highways.
tax	Full-value property-tax rate per USD 10,000.
ptratio	Pupil-teacher ratio by town.
b	$1000(B - 0.63)^2$ where B is the proportion of coloured people by town.
lstat	Percentage of lower status of the population.

be computed in order to analyze whether there is an explanation that is both easy to understand and appropriate.

Note that an explanation grove only denotes a surrogate model [cf. 17]. This means, it is only an approximation which mimics the model under investigation but there is no guarantee that the identified rules correctly describe the original model. In the example, a simple and understandable explanation could be obtained if it would be known that the underlying function is of trigononetric type. Anyway, in practice, the type of the underlying function is usually not known and a surrogate model can neverthless help understanding a model’s behaviour. In that sense, one may refer to the famous quote of George Box that ‘’all models are wrong but some are useful” [18].

3. Demonstration of the R Package xgrove

As it has been worked out before, the proposed methodology not only allows to find a set of rules of fixed size that maximizes the appropriateness of the resulting explanation but, furthermore, by comparing groves of different size, allows to analyze the trade-off between appropriateness and complexity. Explanation groves are implemented in the `R` package `xgrove` which is available on CRAN [11]. Its use is demonstrated to find an explanation for a random forest model that has been trained on the Boston housing data [19]. The data can be accessed via the UCI machine learning benchmark repository [20]. The data consist of median housing values (variable cmedv) from 506 census tracts in the suburbs of Boston and the goal is to predict the housing prices based on 15 explanatory variables such as the crime rate (crim), the average number of rooms (rm), the percentage of persons of lower status in the population (lstat) or the weighted distances to five Boston employment centers (dis).

Initially, a random forest model is trained using the `ranger` implementation [21]. A random forest has been chosen as an example here, as random forests turned out to perform good in many data situations [cf. e.g. 22]. In addition, random forests are comparatively insensitive to the choice of the hyperparameters [23, 24, 25]. For this reason, the default hyperparameters are used.

Note that explanation groves are model agnostic and the same code can be run for arbitrary models. As a default, it is presumed that the call `predict(model, data)` returns the desired predictions $\hat{f}(x)$ of the `model` to be analyzed (here: `rf`). It is possibile to define user-specific predict functions as it is done here by the function `pf`.

The total number splits over all 500 trees in the forest sums up to 80331 which is, of course, far too high to be interpretable. Instead, from Psychology research it is known that humans’ working memory

```r
# load data
library(pdp)
data(boston)

# train model
library(ranger)
set.seed(42)
rf    <- ranger(cmedv ~ ., data = boston)

# define predict function, if necessary
pf     <- function(model, data) {
             return(predict(model, data)$predictions)
             }

# include library
library(xgrove)

# specify desired grove sizes
ntrees <- c(4, 8, 16, 32, 64, 128)

# remove target variable from data
data   <-  boston[, colnames(boston) != "cmedv"]

# compute groves of different size
xg     <- xgrove(rf, data, ntrees, pfun = pf)

# visualize achieved degree of explanation vs. complexity
plot(xg)

# print rules of the grove with at maximum eight rules
xg$rules[["8"]]
```

Figure 2: R Code Demo of the xgrove package.

capacity is limited and restricted to a small number of items [3, 4].

Finally, explanation groves are computed using the function `xgrove()`, which requires three arguments: the `model`, the `data` as well as the desired number of rules (`ntrees`). For this example, six groves of different size are computed where the number of rules is successively doubled from four to 128. The target variable should not be used for the explanation. For this reason it is removed from the data here.[1] The additional `pfun` argument allows to define arbitrary predict functions and only needs to be specified if `predict(model, data)` does not directly return the desired predictions (cf. above).

The resulting S3 object (`xg`) summarizes the achieved degree of explaination Υ as well as the corresponding number of rules for the different groves (cf. figure 3). In `xg$groves`, all groves are of different size are stored as specified by the `ntrees` argument in the call. A similar, but more convenient output is given by `xg$rules` where identical rules with no data points inbetween both splits are aggregated. Thus, the resulting number of rules is smaller or equal than pre-specified. An example of a resulting grove is given in table 2.

Figure 3 compares the appropriateness of the explanations given by groves of different size. This figure can be created by calling the `plot()` method on the output object of the `xgrove()` call. It can be easily seen that the degree of explanation gets better with an increasing number of rules. A value of $\Upsilon \sim 0.9$ is already obtained for less than 20 rules in this case. On the other hand, if a degree of the

[1] This is done automatically if the `model` contains a `terms` component if the `remove.target` argument is specified as `TRUE` (default). Alternatively, this can be done manually as it is done here.

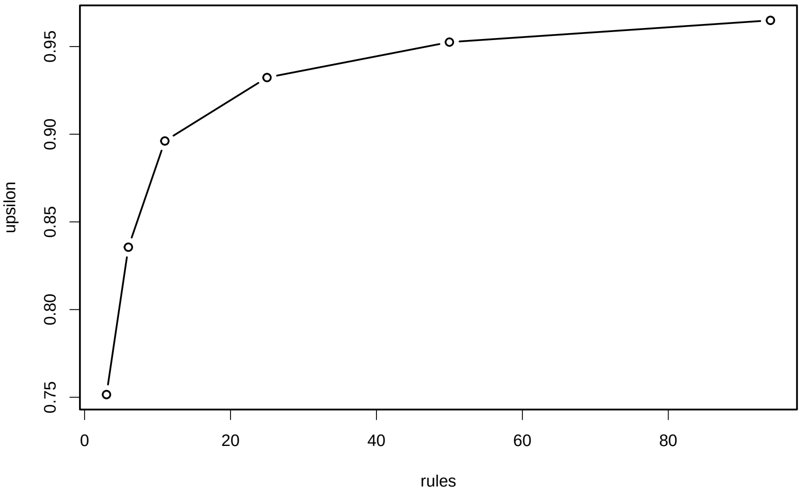

Figure 3: Complexity (abscissa) vs. degree of explanation Υ (ordinate): Groves of different size for a ranger model on the Boston housing data.

Table 2
Explanation grove with six rules for a random forest model on the Boston housing data.

Variable	Upper bound left	Δ left	Δ pright
Intercept		22.527	22.527
crim	9.33	0.474	-3.333
lon	-71.04	0.791	-1.445
lstat	4.55	4.951	-0.472
lstat	14.44	2.656	-4.937
rm	6.84	-1.407	6.777
rm	7.44	-0.374	5.030

explanation of at least $\Upsilon = 0.95$ is required, more than 80 rules are needed. This, in turn, will be hard to interpret.

The resulting grove is given in table 2 below for a grove of six rules. It can be easily seen, that the predicted house prices decrease from 22.53 by 3.33 if the crime rate in a census tract is above 9.33 percent and the model predicts slightly higher prices for more eastern census tracts (longitude > -71.04). A comparatively strong increase of house prices is assigned to census tracts with small percentages of persons with lower status (below 14.44 percent and even more if it is also below 4.55 percent). Finally, also a strong effect can be seen for census tracts with a high average room number above 6.84 or even above 7.44. Nonetheless, the degree of explanation given by these rules is only $\Upsilon \sim 0.836$. Ideally, there should exist a grove of few rules and a high degree of explanation (i.e. in the top left corner of the previous graph). It is up to the user to decide whether this can be considered as sufficient here, but at least, there should be awareness about the magnitude of the gap between the degree of explanation and the true model's responses.

4. Summary

Explanation groves are introduced as a model-agnostic tool to extract a set of understandable rules in order to explain arbitrary machine learning models. An algorithm is proposed that allows to find the best explanation by a prespecified number of rules.

The proposed method is available in the R package `xgrove` on CRAN. It is demonstrated how groves of different size can be easily computed in order to explain arbitratry machine learning models. The results consist in an set of understandable if-then rules. By increasing the number of rules, and thus the complexity of the explanation, the appropriateness of the resulting explanation will improve but it is well-known that human's working memory capacity is limited. In consequence, by creating groves of different size, explanation groves allow to analyze the trade-off between the appropriateness and the complexity of an explanation. The observed trade-off between the degree of explanation and its complexity should be taken into account whenever explainable machine learning is applied in practice.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] European Commission, EU artificial intelligence act, https://artificialintelligenceact.eu/the-act/, 2024.

[2] M. Bücker, G. Szepannek, A. Gosiewska, P. Biecek, TAX4CS – Transparency, auditability and explainability of machine learning models in credit scoring, Journal of the Operational Research Society (2021) 1–21. doi:10.1080/01605682.2021.1922098.

[3] G. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information, Psychological Review 63 (1956) 81–97. doi:10.1037/h0043158.

[4] N. Cowan, The magical mystery four: How is working memory capacity limited, and why?, Curr Dir Psychol Sci 19 (2010) 51–57. doi:10.1177/0963721409359277.

[5] Y. Lou, R. Caruana, J. Gehrke, G. Hooker, Accurate intelligible models with pairwise interactions, in: Proc. 19th ACM SIGKDD Int. Conf. on KDD, ACM, New York, NY, USA, 2013, p. 623–631. doi:10.1145/2487575.2487579.

[6] M. Banerjee, Y. Ding, A.-M. Noone, Identifying representative trees from ensembles, Stat Med 31 (2012) 1601–16. doi:10.1002/sim.4492.

[7] B.-H. Laabs, L. L. Kronziel, I. R. König, S. Szymczak, Construction of artificial most representative trees by minimizing tree-based distance measures, in: L. Longo, S. Lapuschkin, C. Seifert (Eds.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2024, pp. 290–310.

[8] G. Szepannek, B.-H. Laabs, Can't see the forest for the trees, Behaviormetrika 51 (2024) 411–423. doi:10.1007/s41237-023-00205-2.

[9] F. Esposito, D. Malerba, G. Semeraro, J. Kay, A comparative analysis of methods for pruning decision trees, IEEE Trans. on Pattern Analysis and Machine Intelligence 19 (1997) 476–491. doi:10.1109/34.589207.

[10] G. Szepannek, K. Lübke, Explaining artificial intelligence with care, KI - Künstliche Intelligenz (2022). doi:10.1007/s13218-022-00764-8.

[11] G. Szepannek, xgrove: Explanation groves – R package version 0.1-15, 2025. URL: https://CRAN.R-project.org/package=xgrove.

[12] D. Alvarez-Melis, T. S. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Curran Associates Inc., Red Hook, NY, USA, 2018, p. 7786–7795.

[13] G. Szepannek, K. Lübke, How much do we see? on the explainability of partial dependence plots for credit risk scoring, Argumenta Oeconomica 50 (2023). doi:10.15611/aoe.2023.1.07.

[14] J. Friedman, Greedy function approximation: A gradient boosting machine, Annals of Statistics 29 (2001) 1189–1232.

[15] G. Ridgeway, Generalized boosted models: A guide to the gbm package, 2024. URL: https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.

[16] J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating xai: A comparison of rule-

based and example-based explanations, Artificial Intelligence 291 (2021). doi:10.1016/j.artint.2020.103404.

[17] C. Molnar, Interpretable Machine Learning, 2 ed., 2022. URL: https://christophm.github.io/interpretable-ml-book.

[18] G. Box, Robustness in the strategy of scientific model building, Academic Press, 1979, pp. 201–246. doi:10.1016/B978-0-12-438150-6.50018-2.

[19] D. Harrison, D. Rubinfeld, Hedonic prices and the demand for clean air, J. of Environmental Economics and Managemen 5 (1978) 81–102.

[20] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: "http://archive.ics.uci.edu/ml".

[21] M. N. Wright, A. Ziegler, ranger: A fast implementation of random forests for high dimensional data in C++ and R, Journal of Statistical Software 77 (2017) 1–17. doi:10.18637/jss.v077.i01.

[22] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems?, J. Mach. Learn. Res. 15 (2014) 3133–3181.

[23] P. Probst, A.-L. Boulesteix, B. Bischl, Tunability: Importance of hyperparameters of machine learning algorithms, J. Mach. Learn. Res. 20 (2021) 1934–1965.

[24] G. Szepannek, On the practical relevance of modern machine learning algorithms for credit scoring applications, WIAS Report Series 29 (2017) 88–96. doi:10.20347/wias.report.29.

[25] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A.-L. Boulesteix, D. Deng, M. Lindauer, Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges, WIREs Data. Mining. Knowl. Discov. 13 (2023). doi:10.1002/widm.1484.

A. Online Resources

The corresponding R package `xgrove` is avaliable on CRAN.