Experimental Investigation and Evaluation of Model-based Hyperparameter Optimization

Working Paper

Eva Bartz
Bartz & Bartz GmbH
Goebenstr. 10
51643 Gummersbach
eva.bartz@bartzundbartz.de

Martin Zaefferer
Bartz & Bartz GmbH
Goebenstr. 10
51643 Gummersbach
martin.zaefferer@bartzundbartz.de

Olaf Mersmann
Bartz & Bartz GmbH
Goebenstr. 10
51643 Gummersbach
olaf.mersmann@bartzundbartz.de

Thomas Bartz-Beielstein
IDE+A, TH Köln
Steinmüllerallee 1
51643 Gummersbach
thomas.bartz-beielstein@th-koeln.de

July 19, 2021

ABSTRACT

Machine learning algorithms such as random forests or xgboost are gaining more importance and are increasingly incorporated into production processes in order to enable comprehensive digitization and, if possible, automation of processes. Hyperparameters of these algorithms used have to be set appropriately, which can be referred to as *hyperparameter tuning* or optimization. Based on the concept of *tunability*, this article presents an overview of theoretical and practical results for popular machine learning algorithms. This overview is accompanied by an experimental analysis of 30 hyperparameters from six relevant machine learning algorithms. In particular, it provides (i) a survey of important hyperparameters, (ii) two parameter tuning studies, and (iii) one extensive global parameter tuning study, as well as (iv) a new way, based on consensus ranking, to analyze results from multiple algorithms. The R package mlr is used as a uniform interface to the machine learning models. The R package SPOT is used to perform the actual tuning (optimization). All additional code is provided together with this paper.

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) methods are gaining more importance and are increasingly incorporated into production processes in order to enable comprehensive digitization and, if possible, automation of processes. As a rule, several hyperparameters of the methods used have to be set appropriately. Previous projects carried out produced inconsistent results in this regard. For example, with Support Vector Machines (SVMs) it could be observed that the tuning of the hyperparameters is critical to success with the same data material, with random forests the results do not differ too much from one another despite different selected hyperparameter values. While some methods have only one or a few hyperparameters, others provide a large number. In the latter case, optimization using a (more or less) fine grid (grid search) quickly becomes very time-consuming and can therefore no longer be implemented. In addition, the question of how the optimality of a selection can be measured in a statistically valid way (test problem: training / validation / test data and resampling methods) arises for both many and a few hyperparameters.

This article deals with the hyperparameter tuning of Machine Learning (ML) algorithms. In particular, it provides

- a survey of important model parameters,
- two parameter tuning studies and

- one extensive global parameter tuning study, and
- a new way, based on consensus ranking, to analyze results from multiple algorithms.

More than 30 hyperparameters from six relevant ML models are analyzed. This paper is structured as follows: Sec. 2 presents models (algorithms) and hyperparameters. Hyperparameter Tuning (HPT) methods are introduced in Sec. 3. Three case studies are presented next. These hyperparameter tuning studies are using the Census-Income (KDD) Data Set (CID), which will be described in Sec. 4.2. The first case study, which describes HPT for decision trees (`rpart`), is presented in Sec. 4. This case study might serve as a starting point for the interested reader, because `rpart` can be evaluated very quickly. The second case study analyses eXtreme Gradient Boosting and is presented in Sec. 5. A global case study, which analyses *tunability*, is presented in Sec. 6. Results are discussed in Sec. 7.

2 Models and Hyperparameters

2.1 Overview

2.1.1 Methodology

In the following, we provide a survey and description of relevant parameters of six different ML models. We emphasize that this is not a complete list of their parameters, but covers parameters that are set quite frequently according to the literature.

Since the specific names and meaning of hyperparameters may depend on the actual implementation used, we have chosen a reference implementation for each model. The implementations chosen are all packages from the statistical programming language R. Thus, we provide a description that is consistent with what users experience, so that they can identify the relevant parameters when tuning ML algorithms in practice.

In particular, we cover the following models, with the respective R packages:

- k-Nearest-Neighbor: `kknn`
- Elastic Net: `glmnet`
- Decision Tree: `rpart`
- Random Forest: `ranger`
- eXtreme Gradient Boosting: `xgboost`
- Support Vector Machine: `e1071`

Table 1 presents an overview of these algorithms and their associated hyperparameters. After a short, general description of the specific hyperparameter, the following features will be described for every hyperparameter:

Type: Describes the type (e.g., integer) and complexity (e.g., scalar). These data types are described in Sec. 2.1.2.

Sensitivity / robustness: Describes how much the parameter is affected by changes, i.e, how much its setting depends on the considered problem and data. There is a close relationship between sensitivity and tunability as defined by [Probst et al., 2019a], because tunability is the potential for improvement of the parameter in the vicinity of a reference value.

Heuristics for determination: Describes ways to find good hyperparameter settings.

Plausible range of values: Describes feasible values, i.e., lower and upper bounds, constraints, etc.

Further constraints: Additional constraints, specific for certain settings or algorithms.

Interactions: Describes interactions between the parameters.

Each algorithm description concludes with a brief survey of examples from the literature, that gives hints how the algorithm was tuned.

2.1.2 Data Types

Our implementation is done in the R programming language, where data and functions are represented as objects. Each object has a data type. The basic (or *atomic*) data types are shown in Table 2.

In addition to these data types, R uses an *internal storage mode* which can be queried using `typeof()`. Thus there are two storage modes for the `numeric` data type:

Table 1: Overview. Models and parameters analyzed in this study.

Model	Hyperparameter	Comment
knn	`k`	number of neighbors
	`p`	p norm
Elastic net	`alpha`	weight term of the loss-function
	`lambda`	trade-off between model quality and complexity
	`thresh`	threshold for model convergence, i.e., convergence of the internal coordinate descent
Decision trees	`minsplit`	minimum number of observations required for a split
	`minbucket`	minimum number of observations in an end node (leaf)
	`cp`	complexity parameter
	`maxdepth`	maximum depth of a leaf in the decision tree
Random forest	`num.trees`	number of trees that are combined in the overall ensemble model
	`mtry`	number of randomly chosen features are considered for each split
	`sample.fraction`	number of observations that are randomly drawn for training a specific tree
	`replace`	replacement of randomly drawn samples
	`respect.-`	
	`unordered.factors`	handling of splits of categorical variables
xgBoost	`eta`	learning rate, also called "shrinkage" parameter.
	`nrounds`	number of boosting steps
	`lambda`	regularization of the model
	`alpha`	parameter for the L1 regularization of the weights
	`subsample`	portion of the observations that is randomly selected in each iteration
	`colsample_bytree`	number of features that chosen for the splits of a tree
	`gamma`	number of splits of a tree by assuming a minimal improvement for each split
	`max_depth`	maximum depth of a leaf in the decision trees
	`min_child_weight`	restriction of the number of splits of each tree
Support vector machines	`degree`	degree of the polynomial (parameter of the polynomial kernel function)
	`gamma`	parameter of the polynomial, radial basis, and sigmoid kernel functions
	`coef0`	parameter of the polynomial and sigmoid kernel functions
	`cost`	regularization parameter, weights constraint violations of the model
	`epsilon`	regularization parameter, defines ribbon around predictions

Table 2: Atomic data types of the programming language R

Data type	Description	Examples
NULL	empty set	NULL
logical	boolean values	TRUE, FALSE
numeric	integer and real values	1, 0.5
complex	complex values	1+1i
character	characters and strings of characters	"123", "test"

- `integer` for integers and

- `double` for real values.

The corresponding variables are referred to as *numeric*. *Factors* are used in R to represent nominal (qualitative) features. Ordinal features can also be represented by factors in R (see argument `ordered` of the function `factor()`). However, this case is not considered here. Factors are generated with the generating function `factor()`. Factors are not atomic data types. Internally in R, factors are stored by numbers (integers), externally the name of the factor is used. We call the corresponding variables *categorical*.

2.1.3 Terminology

This article deals with the tuning of the hyperparameters of methods from the field of supervised learning (both regression and classification). In the following, the scope of application is explained, including a terminology of important terms.

The *data points* x come from an input data space $\mathcal{X}$ ($x \in \mathcal{X}$) and can have different scale levels (nominal: no order; ordinal: order, no distance; cardinal: order, distances). Nominal and ordinal data are mostly discrete, cardinal data are continuous. Usually, the data points are n-dimensional vectors. The vector elements are also called features or dependent variables. The number of data points in a data set is m.

In addition, we consider output data (dependent variables) $y \in \mathcal{Y}$. These can also have different scale levels (nominal, ordinal, cardinal). Output data are usually scalar. Variable identifiers like x and y represent scalar or vector quantities, the meaning results from the context.

We study ML algorithms, also referred to as models or methods. *Supervised learning* means that for each observed data point x_i with $i = 1, 2, ..., m$ of the input space an associated dependent output value y_i is known. Supervised learning models learn the relationship between input and output. The goal is to predict the output value for a new/unknown data point x^*: $y^* = f(x^*)$, with $f : \mathcal{X} \mapsto \mathcal{Y}$.

Regression deals with a real output space ($y_i \in \mathbb{R}$). Integer output spaces are also included in this domain. Example question: how does relative humidity depend on temperature?

Classification deals with an output space of discrete classes (categories), $y_i \in \{a_1, a_2, ..., a_c\}$. Where $c \in \mathbb{N}$ is the natural number of classes. Example: How does the default on a loan (yes, no) depend on the income of the borrower?

The *model quality* describes the performance of a model. This can be expressed in different ways, e.g. as a measure of the fit of the model to the observed data values.

Hyperparameters are settings or configurations of the methods (models), which are freely selectable within a certain range and influence model quality Example of hyperparameters: The number of units or layers of a neural network. The number of hyperparameters of a model is d. Hyperparameters are distinct from model parameters, among others. Model parameters are chosen during the learning process by the model itself. Example of model parameters: The weights of each connection in a neural network.

The determination of the best possible hyperparameters is called *tuning* (HPT). For the tuning procedure (or tuner for short), we use a model-based search. The corresponding model is called a surrogate model.

The term *tunability* is used according to the definition presented in Probst et al. [2019a]. Tunability describes a measure for modeling algorithms as well as for individual hyperparameters. It is the difference between the model quality for default values (or reference values) and the model quality for optimized values (after tuning is completed). Or in the words of Probst et al. [2019a]: "measures for quantifying the tunability of the whole algorithm and specific hyperparameters based on the differences between the performance of default hyperparameters and the performance of the hyperparameters when this hyperparameter is set to an optimal value". Tunability of individual hyperparameters can also be used as a measure of their *relevance*, *importance*, or *sensitivity*. Parameters with high tunability are accordingly of greater importance for the model. The model reacts strongly to (i.e., is sensitive to) changes in these hyperparameters.

The term *complexity* or model complexity generally describes, how many functions of different difficulty can be represented by a model. Example 1: for linear models, this can be influenced by the number of model coefficients. Example 2: for support vector machines, this can be influenced by the parameter `cost`.

2.2   Model: k-Nearest-Neighbor

The k-Nearest-Neighbor (k-NN) model determines for each x' the n neighbors with the least distance to x', e.g., based on the Minkowski distance:

$$\mathrm{d}(x, x') = \left(\sum_{i=1}^{n} |x_i - x'_i|^{\mathtt{p}} \right)^{1/\mathtt{p}} = ||x - x'||_{\mathtt{p}}. \tag{1}$$

For regression, the mean of the neighbors is used [James et al., 2017]. For classification, the prediction of the model is the most frequent class observed in the neighborhood. Two relevant hyperparameters (k, p) result from this. Additionally, one categorical hyperparameter could be considered: the choice of evaluation algorithm (e.g., choosing between brute force or KD-Tree) [Friedman et al., 1977]. However, this mainly influences computational efficiency, rather than actual performance.

We consider the implementation from the R package kknn[1] [Schliep et al., 2016].

2.2.1   k-Nearest-Neighbor Parameter: k

The parameter k determines the number of neighbors that are considered by the model. In case of regression, it affects how smooth the predicted function of the model is. Similarly, it influences the smoothness of the decision boundary in case of classification.

Small values of k lead to fairly non-linear predictors (or decision boundaries), while larger values tend towards more linear shapes [James et al., 2017]. The error of the model at any training data sample is zero if k =1 but this does not determine anything about the generalization error [James et al., 2017]. Larger values of k may help to deal with rather noisy data. Moreover, larger values of k increase the runtime of the model.

Type: integer, scalar.

Sensitivity / robustness: Determining the size of the neighborhood via k is a fairly sensitive decision. James et al. [2017] describe this as a drastic effect. However, this is only true as long as the individual classes are hard to separate (in case of classification). If there is a large margin between classes, the shape of the decision boundary becomes less relevant (see [Domingos, 2012, Fig. 3]). Thus, the sensitivity of the parameter depends on the considered problem and data. Probst et al. [2019a] also identify k as a sensitive (or *tunable*) parameter. There is a close relationship between sensitivity and tunability as defined by Probst et al, since tunability is the potential for improvement of the parameter in the vicinity of a reference value.

Heuristics for determination: As mentioned above, the choice of k may depend on properties of the data. Hence, no general rule can be provided. In individual cases, determining the distance between and within classes may help to find an approximate value: k =1 is better than k $>$ 1, if the distance within classes is larger than the distance between classes [Cover and Hart, 1967]. Another empirical suggestion from the literature is k $= \sqrt{m}$ [Lall and Sharma, 1996, Probst et al., 2019a].

Plausible range of values: k ≥ 1, k $\ll m$, where m is the number of data samples. Only integer values are valid.

Further constraints: none.

Interactions: We are not aware of any interactions between the parameters. However, both k and p change the perceived neighborhood of samples and thus the shape of the decision boundaries. Hence, an interaction between these parameters is likely.

2.2.2   k-Nearest-Neighbor Parameter: p

The parameter p affects the distance measure that is used to determine the nearest neighbors in k-NN. Frequently, this is the Minkowski distance: $\mathrm{d}(x, x') = (\sum_{i=1}^{n} |x_i - x'_i|^{\mathtt{p}})^{1/\mathtt{p}}$. Moreover, it has to be considered that other distances could be chosen for non-numerical features of the data set (i.e., Hamming distance for categorical features). The implementation used in this article (kknn) transforms categorical variables into numerical variables via dummy-coding, then using the Minkowski distance on the resulting data. Similarly to k, p changes the observed neighborhood. While p does not change the number of neighbors, it still affects the choice of neighbors.

Type: double, scalar.

---

[1] https://cran.r-project.org/package=kknn

Sensitivity / robustness: It has to be expected, that the model is less sensitive to changes in p than to changes in k, since fairly extreme changes are required to change the neighborhood set of a specific data sample. This explains, why many publications do not consider p during tuning, see Table 3. However, the detailed investigation of Alfeilat et al. [2019] showed, that changes of the distance measure can have a significant effect on the model accuracy. Alfeilat et al. only tested special cases of the Minkowski distance: Manhattan distance (p = 1), Euclidean distance (p = 2) and Chebyshev distance (p = ∞). They give no indication whether other values may be of interest as well.

Heuristics for determination: The choice of distance measure (and hence p) depends on the data, a general recommendation or rule-of-thumb is hard to derive [Alfeilat et al., 2019].

Plausible range of values: Often, the interval $1 \leq p \leq 2$ is considered. The lower boundary is $p > 0$. Note: The Minkowski distance is not a metric if $p < 1$ [Alfeilat et al., 2019]. Theoretically, a value of $p = \infty$ is possible (resulting into Chebyshev distance), but this is not possible in the kknn implementation.

Further constraints: none.

Interactions: We are not aware of any known interactions between parameters. However, both k and p change what is perceived as the neighborhood of samples, and hence the shape of decision boundaries. Hence, an interaction between those parameters is likely.

In conclusion, Table 3 provides a brief survey of examples from the literature, where k-NN was tuned.

Table 3: Survey of examples from the literature, for tuning of k-NN. Asterisk (*) denotes results that depend on data set (multiple data sets).

Hyperparameter	Lower bound	Upper bound	Result	Notes
[Schratz et al., 2019], weighted k-NN variant, spacial data, 1 data set				
k	10	400	NA	
p	1	100	NA	integer
[Khan et al., 2020], detection of bugs, 5 data sets				
k	1	17	NA	
p	0,5	5	NA	
[Osman et al., 2017], detection of bugs, 5 data sets				
k	1	5	2 or 5 *	
[Probst et al., 2019a], various applications, 38 data sets				
k	1	30	2 to 30 *	
[Doan et al., 2020], impact damage on reinforced concrete, 1 data set				
k	7	51	9	
p	1	11	3	

2.3 Model: Regularized regression—Elastic Net

Elastic Net (EN) is a regularized regression method [Zou and Hastie, 2005]. Regularized regression can be employed to fit regression models with a reduced number of model coefficients. Special cases of EN are Lasso and Ridge regression.

Regularization is useful, when data sets are high-dimensional (especially but not exclusively if $n > m$), or when variables in the data sets are heavily correlated with each other [Zou and Hastie, 2005]. Less complex models (i.e., with fewer coefficients) help to reduce overfitting[2]. The resulting models are also easier to understand for humans, due to their reduced complexity.

During training, non-regularized regression reduces only the model error (e.g., via the least squares method). EN also considers a penalty term, which grows with the number of coefficients included in the model (i.e., are non-zero).

As a reference implementation, we use the R package glmnet[3] [Friedman et al., 2020, Simon et al., 2011].

---

[2]Overfitting means, that the model is extremely well adapted to the training data, but generalizes poorly as a result (i.e., predicts poorly for unseen data).

[3]https://cran.r-project.org/package=glmnet

2.3.1 Elastic Net Parameter: `alpha`

The parameter `alpha` (α) weights the two elements of the penalty term of the loss-function in the EN model [Friedman et al., 2010]:

$$\min_{\beta_0,\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^\mathrm{T}\beta)^2 + \lambda P(\alpha, \beta). \tag{2}$$

The penalty term $P(\alpha, \beta)$ is [Friedman et al., 2010]:

$$(1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1, \tag{3}$$

with the vector of p model coefficients $\beta \in \mathbb{R}^p$ and the 'intercept' coefficient $\beta_0 \in \mathbb{R}$. The value `alpha` =0 corresponds to the special case of Ridge regression, `alpha` =1 corresponds to Lasso regression [Friedman et al., 2010].

The parameter `alpha` allows to find a compromise or trade-off between Lasso and Ridge regression. This can be advantageous, since both variants have different consequences. Ridge regression affects that coefficients of strongly correlated variables match to each other (extreme case: identical variables receive identical coefficients) [Friedman et al., 2010]. In contrast, Lasso regression tends to lead to a single coefficient in such a case (the other coefficients being zero) [Friedman et al., 2010].

Type: double, scalar.

Sensitivity / robustness: Empirical results from Friedman et al. show that the EN model can be rather sensitive to changes in `alpha` [Friedman et al., 2010].

Heuristics for determination: We are not aware of any heuristics to set this parameter. As described by Friedman et al. [2010], `alpha` can be set to a value of close to 1, if a model with few coefficients without risk of degeneration is desired.

Plausible range of values: `alpha` $\in [0, 1]$.

Further constraints: none.

Interactions: `lambda` interacts with `alpha`, see Sec. 2.3.2.

2.3.2 Elastic Net Parameter: `lambda`

The hyperparameter `lambda` influences the impact of the penalty term $P(\alpha, \beta)$ in equation 2. Very large `lambda` lead to many model coefficients (β) being set to zero. Correspondingly, only few model coefficients become zero if `lambda` is small (close to zero). Thus, `lambda` is often treated differently than other hyperparameters: In many cases, several values of `lambda` are of interest, rather than a single value [Simon et al., 2011]. There is no singular, optimal solution for `lambda`, as it controls the trade-off between model quality and complexity (number of coefficients that are not zero). Hence, a whole set of `lambda` values will often be suggested to users, who are then able to choose a resulting model that provides a specific trade-off of their liking.

Type: double, scalar.

Sensitivity / robustness: EN is necessarily sensitive to `lambda`, since extreme values lead to completely different models (i.e., all coefficients are zero or none are zero). This is also shown in figure 1 by Friedman et al. [2010].

Heuristics for determination: Often, `lambda` gets determined by a type of grid-search, where a sequence of decreasing `lambda` is tested [Friedman et al., 2010, Simon et al., 2011]. The sequence starts with a sufficiently large value of `lambda`, such that $\beta = 0$. The sequence ends, if the resulting model starts to approximate the unregularized model [Simon et al., 2011].

Plausible range of values: `lambda` $\in (0, \infty)$ (Note: `lambda` =0 is possible, but leads to a simple unregularized model). Using a logarithmic scale seems reasonable, as used in the study by Probst et al. [2019a], to cover a broad spectrum of very small and very large values.

Further constraints: none.

Interactions: `lambda` interacts with `alpha`. Both are central for determining the coefficients β (see also [Friedman et al., 2010, Fig. 1]).

2.3.3 Elastic Net Parameter: `thresh`

The parameter `thresh` is a threshold for model convergence (i.e., convergence of the internal coordinate descent). Model training ends, when the change after an update of the coefficients drops below this value [Friedman et al., 2020]. Unlike parameters like `lambda`, `thresh` is not a regularization parameter, hence there is clear connection between `thresh` and the number of model coefficients.

As a stopping criterion, `thresh` influences the duration of model training (larger values of `thresh` result into faster training), and the quality of the model (larger values of `thresh` may decrease quality).

Type: double, scalar.

Sensitivity / robustness: As long as `thresh` is in a reasonable range of values, the model will not be sensitive to changes. Extremely large values can lead to fairly poor models, extremely small values may result into significantly larger training times.

Heuristics for determination: none are known.

Plausible range of values: $\text{thresh} \approx 0$, $\text{thresh} > 0$. It seems reasonable to set `thresh` on a log-scale with fairly coarse granularity, since `thresh` has a low sensitivity for the most part. Example: $\text{thresh} = 10^{-20}, 10^{-18}, ..., 10^{-4}$.

Further constraints: none.

Interactions: none are known.

In conclusion, Table 4 provides a brief survey of examples from the literature, where EN was tuned.

Table 4: Survey of examples from the literature, for tuning of EN.
*: Results depend on data set (multiple data sets).
**: the integrated, automatic tuning procedure from `glmnet` was used.

Hyperparameter	Lower bound	Upper bound	Result	Notes
[Probst et al., 2019a], various applications, 38 data sets				
alpha	0	1	0,003 to 0,981 *	
lambda	2^{-10}	2^{10}	0,001 to 0,223 *	
[Wong et al., 2019], medical data, 1 data set				
alpha	1	1	1	not tuned, constant
lambda	**	**	0.001	

2.4 Model: Decision Trees

Decision and regression trees are models that divide the data space into individual segments with successive decisions (called splits).

Basically, the procedure of a decision tree is as follows: Starting from a root node (which contains all observations) a first split is carried out. Each split affects a variable (or a feature). This variable is compared with a threshold value. All observations that are less than the threshold are assigned to a new node. All other observations are assigned to another new node. This procedure is then repeated for each node until a termination criterion is reached or until there is only one observation in each end node.

End nodes are also called leaves (following the tree analogy).

A detailed description of tree-based models is given by James et al. [2014]. An overview of decision tree implementations and algorithms is given by Zharmagambetov et al. [2019]. Gomes Mantovani et al. [2018] describe the tuning of hyperparameters of several implementations. As a reference implementation, we refer to the R package `rpart` [Therneau and Atkinson, 2019, Therneau et al., 2019].

2.4.1 Decision Tree Parameter: `minsplit`

If there are fewer than `minsplit` observations in a node of the tree, no further split is carried out at this node. Thus, `minsplit` limits the complexity (number of nodes) of the tree. With large `minsplit` values, fewer splits are made. A suitable choice of `minsplit` can thus avoid overfitting. In addition, the parameter influences the duration of the training of an decision tree [Hastie et al., 2017].

Type: integer, scalar.

Sensitivity / robustness: Trees can react very sensitively to parameters that influence their complexity. Together with `minbucket`, `cp` and `maxdepth`, `minsplit` is one of the most important hyperparameters [Gomes Mantovani et al., 2018].

Heuristics for determination: `minsplit` is set to three times `minbucket` in certain implementations, if this parameter is available [Therneau and Atkinson, 2019].

Plausible range of values: `minsplit` $\in [1, m]$, where `minsplit` $<< m$ is recommended, since otherwise trees with extremely few nodes will arise. Only integer values are valid.

Further constraints: `minsplit` $>$ `minbucket`. This is a soft constraint, i.e., valid models are created even if violated, but `minsplit` would no longer have any effect.

Interactions: The parameters `minsplit`, `minbucket`, `cp` and `maxdepth` all influence the complexity of the tree. Interactions between these parameters are therefore likely. In addition, `minsplit` has no effect for certain values of `minbucket` (see Constraints). Similar relationships (depending on the data) are also conceivable for the other parameter combinations.

2.4.2 Decision Tree Parameter: `minbucket`

`minbucket` specifies the minimum number of data points in an end node (leaf) of the tree. The meaning in practice is similar to that of `minsplit`. With larger values, `minbucket` also increasingly limits the number of splits and thus the complexity of the tree.

Type: integer, scalar.

Sensitivity / robustness: see `minsplit`.

Heuristics for determination: `minbucket` is set to a third of `minsplit` in the reference implementations, if this parameter is available Therneau et al. [2019].

Plausible range of values: `minbucket` $\in [1, m]$, where `minbucket` $<< m$ is recommended, as otherwise trees with extremely few nodes will arise. Only integer values are valid.

Further constraints: `minsplit` $>$ `minbucket` (this is a soft constraint, i.e., valid models are created even if violated, but `minsplit` would no longer have any effect).

Interactions: see `minsplit`. Due to the similarity of `minsplit` and `minbucket`, it can make sense to only tune one of the two parameters.

2.4.3 Decision Tree Parameter: `cp`

The threshold complexity `cp` controls the complexity of the model in that split decisions are linked to a minimal improvement. This means that if a split does not improve the tree-based model by at least the factor `cp`, this split will not be carried out. With larger values, `cp` increasingly limits the number of splits and thus the complexity of the tree.

Therneau and Atkinson [2019] describe the `cp` parameter as follows:

> The complexity parameter `cp` is, like `minsplit`, an advisory parameter, but is considerably more useful. It is specified according to the formula
>
> $$R_{\mathtt{cp}}(T) \equiv R(T) + \mathtt{cp} \times |T| \times R(T_1) \tag{4}$$
>
> where T_1 is the tree with no splits, $|T|$ is the number of splits for a tree, and R is the risk. This scaled version is much more user-friendly than the original CART formula since it is unit less. A value of `cp` $= 1$ will always result in a tree with no splits. For regression models the scaled `cp` has a very direct interpretation: if any split does not increase the overall R^2 of the model by at least `cp` (where R^2 is the usual linear-models definition) then that split is decreed to be, a priori, not worth pursuing. The program does not split said branch any further, and saves considerable computational effort. The default value of .01 has been reasonably successful at 'pre-pruning' trees so that the cross-validation step need only remove 1 or 2 layers, but it sometimes over prunes, particularly for large data sets.

Type: double, scalar.

Sensitivity / robustness: see `minsplit`.

Table 5: Decision Tree: survey of examples from the literature. Tree-based tuning example configurations. $*$ denotes that results depend on data sets.

Hyperparameter	Lower bound	Upper bound	Result
[Probst et al., 2019a], various applications, 38 data sets			
`minsplit`	1	60	6.7 to 49.15 $*$
`maxdepth`	1	30	9 to 28 $*$
`cp`	0	1	0 to 0.528 $*$
`minbucket`	1	60	1 to 44.1 $*$
[Wong et al., 2019], medical data, 1 data set			
`cp`	10^{-6}	10^{-1}	10^{-2}
[Khan et al., 2020], software bug detection, 5 data sets			
`minbucket`	1	50	NA
[Gomes Mantovani et al., 2018], various data sets, 94 data sets			
`minsplit`	1	50	
`minbucket`	1	50	
`cp`	0.0001	0.1	
`maxdepth`	1	50	
This study, see Sec. 4, CID			
`minsplit`	1	300	16 (not relevant)
`minbucket`	0.1	0.5	0.17 (not relevant)
`cp`	10^{-10}	1	10^{-3} (most relevant hyperparameter)
`maxdepth`	1	30	> 10

Heuristics for determination: none known.

Plausible range of values: paramcp $\in [0, 1[$.

Further constraints: none.

Interactions: see `minsplit`. Since `cp` expresses a relative factor for the improvement of the model, an interaction with the corresponding quality measure is also possible (split parameter).

2.4.4 Decision Tree Parameter: `maxdepth`

The parameter `maxdepth` limits the maximum depth of a leaf in the decision tree. The depth of a leaf is the number of nodes that lie on the path between the root and the leaf. The root node itself is not counted [Therneau and Atkinson, 2019].

The meaning in practice is similar to that of `minsplit`. Both `minsplit` and `maxdepth` can be used to limit the complexity of the tree. However, smaller values of `maxdepth` lead to a lower complexity of the tree. With `minsplit` it is the other way round (larger values lead to less complexity).

Type: integer, scalar.

Sensitivity / robustness: see `minsplit`

Heuristics for determination: none known.

Plausible range of values: `maxdepth` $\in [0, m]$. Only integer values are valid.

Further constraints: none.

Interactions: see `minsplit`.

Table 5 shows examples from the literature.

2.5 Model: Random Forest

The model quality of decision trees can often be improved with ensemble methods. Here, many individual models (i.e., many individual trees) are merged into one overall model (the ensemble).

An example is the Random Forest (RF) model. In this process, many decision trees are created at the same time, and their prediction is then usually made using the mean (in case of regression) or by majority vote (in case of classification). The variant of RF described by Breiman [2001] uses two important steps to reduce generalization error: First,

when creating individual trees, only a random subset of the features is considered for each split. Second, each tree is given a randomly drawn subset of the observations to train. Typically, the approach of bootstrap aggregating or bagging [James et al., 2017] is used. A comprehensive discussion of random forest models is provided by Louppe [2015] and a detailed discussion of hyperparameters is provided by Louppe [2015]. Theoretical results on hyperparameters of RF models are summarized by Scornet [2017]. Often, tuning of RF also takes into account parameters for the decision trees themselves see Sec. 2.4. Our reference implementation studied in this report is from the R package `ranger`[4] [Wright, 2020, Wright and Ziegler, 2017].

2.5.1 Random Forest Parameter: `num.trees`

`num.trees` determines the number of trees that are combined in the overall ensemble model. In practice, this influences the quality of the model (more trees lead to a better quality) and the runtime of the model (more trees lead to longer runtimes for training and prediction).

Type: integer, scalar.

Sensitivity / robustness: According to Breiman [2001], the generalization error of the model converges with increasing number of trees towards a lower bound. This means, that the model will become less sensitive to changes of `num.trees` with increasing values of `num.trees`. This is also shown in the benchmarks of Louppe [2015]. Only with relatively small values (`num.trees` < 50) the model is rather sensitive to changes in that parameter. The empirical results of Probst et al. [2019a] also show that the *tunability* of `num.trees` is estimated to be rather low.

Heuristics for determination: There are theoretical results about the convergence of the model in relation to `num.trees` [Breiman, 2001, Scornet, 2017]. This however does not result into a clear heuristic approach to setting this parameter. One common recommendation is to choose `num.trees` sufficiently high [Probst et al., 2019b] (since more trees are usually better), while making sure that the runtime of the model does not become too large.

Plausible range of values: `num.trees` $\in [1, \infty)$. Several hundred or thousands of trees are commonly used, see also Table 6.

Further constraints: none.

Interactions: none are known.

2.5.2 Random Forest Parameter: `mtry`

`mtry` determines, how many randomly chosen features are considered for each split. Thus, it controls an important aspect, the randomization of individual trees. Values of `mtry` $\ll n$ imply that differences between trees will be larger (more randomness). This increases the potential error of individual trees, but the overall ensemble benefits [Breiman, 2001, Louppe, 2015]. One useful side effect is, that `mtry` $\ll n$ may also reduce the runtime considerably [Louppe, 2015]. Nevertheless, findings about this parameter largely depend on heuristics and empirical results. According to Scornet, no theoretical results about the randomization of split features are available [Scornet, 2017].

Type: integer, scalar.

Sensitivity / robustness: According to Breiman, RF is relatively insensitive to changes of `mtry`: "*But the procedure is not overly sensitive to the value of F. The average absolute difference between the error rate using F=1 and the higher value of F is less than 1.*" [Breiman, 2001] (here: F corresponds to `mtry`).

This seems to be at odds with the benchmarks by Louppe [2015], which determine that `mtry` may indeed have a considerable impact, especially for low values of `mtry`. The investigation of tunability by Probst et al. [2019a] also identifies `mtry` as an important (i.e., tunable) parameter. This is not necessarily a contradiction to Breiman's observation, since Probst et al. [2019a] determine RF as the least tunable model in their experimental investigation. So while `mtry` might have some impact (compared to other parameters), it may be less sensitive when compared relative to hyperparameters of other models.

Heuristics for determination: Breiman [2001] propose the following heuristic:

$$\texttt{mtry} = \text{floor}(\log_2(n) + 1).$$

Should categorical features be present, Breiman suggests doubling or tripling that value. No theoretical motivation is given.

---

[4]`https://cran.r-project.org/package=ranger`

Another frequent suggestion is $\mathtt{mtry} = \sqrt{n}$ (or $\mathtt{mtry} = \mathrm{floor}(\sqrt{n})$). While these are used in various implementations of RF, there is no clear theoretical motivation given. For $n < 20$ both heuristics provide very similar values.

Some implementations distinguish between classification ($\mathtt{mtry} = \sqrt{n}$) and regression ($\mathtt{mtry} = n/3$). Empirical results with these heuristics are described by Probst et al. [2019b].

Plausible range of values: $\mathtt{mtry} \in [1, n]$.

Further constraints: none.

Interactions: none are known.

2.5.3 Random Forest Parameter: `sample.fraction`

The parameter `sample.fraction` determines, how many observations are randomly drawn for training a specific tree.

Probst et al. [2019b] write, that `sample.fraction` a similar effect as `mtry`. That means, it influences the properties of the trees: With small `sample.fraction` (corresponding to small `mtry`) individual trees are weaker (in terms of predictive quality), yet the diversity of trees is increased. This improves the ensemble model quality. Smaller values of `sample.fraction` reduce the runtime [Probst et al., 2019b] (if all other parameters are equal).

Type: double, scalar.

Sensitivity / robustness: `sample.fraction` can have a relevant impact on model quality. Scornet reports: "*However, according to empirical results, there is no justification for default values in random forests for sub-sampling or tree depth, since optimizing either leads to better performance.*"

Heuristics for determination: none known.

Plausible range of values: `sample.fraction` $\in (0, 1]$.

Further constraints: none.

Interactions: Potentially, `sample.fraction` interacts with parameters that influence training individual trees Decision Tree (DT) (e.g., `maxdepth`, `minsplit`, `cp`). Scornet: "*According to the theoretical analysis of median forests, we know that there is no need to optimize both the subsample size and the tree depth: optimizing only one of these two parameters leads to the same performance as optimizing both of them*" [Scornet, 2017]. However, this theoretical observation is only valid for the respective *median trees* and not necessarily for the classical RF model we consider

2.5.4 Random Forest Parameter: `replace`

The parameter `replace` specifies, whether randomly drawn samples are replaced, i.e., whether individual samples can be drawn multiple times for training of a tree (`replace =TRUE`) or not (`replace =FALSE`). If `replace =TRUE`, the probability that two trees receive the same data sample is reduced. This may further decorrelate trees and improve quality.

Type: logical, scalar.

Sensitivity / robustness: The sensitivity of `replace` is often rather small. Yet, the survey of Probst et al. [2019b] notes a potentially detrimental bias for `replace =TRUE`, if categorical variables with a variable number of levels are present.

Heuristics for determination: Due to the aforementioned bias, the choice could be made depending on the variance of the cardinality in the data features. However, a quantifiable recommendation is not available.

Plausible range of values: `replace` $\in \{\mathrm{TRUE}, \mathrm{FALSE}\}$.

Further constraints: none.

Interactions: One obvious interaction occurs with `sample.fraction`. Both parameters control the random choice of training data for each tree. The setting (`replace = TRUE` $\wedge$ `sample.fraction = 1`) as well as the setting (`replace = FALSE` $\wedge$ `sample.fraction < 1`) implies that individual trees will not see the whole data set.

2.5.5 Random Forest Parameter: `respect.unordered.factors`

This parameter decides how splits of categorical variables are handled. A detailed discussion is given by Wright and König [2019]. A standard that is also used by [Breiman, 2001], is `respect.unordered.factors =partition`. In

that case, all potential splits of a nominal, categorical variable are considered. This leads to a good model, but the large number of considered splits can lead to an unfavorable runtime.

A naive alternative is `respect.unordered.factors =ignore`. Here, the categorical nature of a variable will be ignored. Instead, it is assumed that the variables is ordinal, and splits are chosen just as with numerical variables. This reduces runtime but can decrease model quality.

A better choice should be `respect.unordered.factors =order`. Here, each categorical variable first is sorted, depending on the frequency of each level in the first of two classes (in case of classification or depending on the average dependent variable value (regression). After this sorting, the variable is considered to be numerical. This allows for a runtime similar to that with `respect.unordered.factors =ignore` but with potentially better model quality. This may not be feasible for classification with more than two classes, due to lacking a clear sorting criterion [Wright and König, 2019, Wright, 2020].

In specific cases, `respect.unordered.factors =ignore` may work well in practice. This could be the case, when the variable is actually nominal (unknown to the analyst).

Type: character, scalar.

Sensitivity / robustness: unknown.

Heuristics for determination: none.

Plausible range of values: `respect.unordered.factors` $\in$ {`ignore`, `order`, `partition`}. The parameter `respect.unordered.factors` can also be understood as a binary value. Then `TRUE` corresponds to `order` and `FALSE` to `ignore` [Wright, 2020].

Further constraints: none.

Interactions: none are known.

In conclusion, Table 6 provides a brief survey of examples from the literature, where RF was tuned.

Table 6: RF: survey of examples from the literature for tuning of random forest. *: Results depend on data set (multiple data sets).

Hyperparameter	Lower bound	Upper bound	Result	Notes
[Probst et al., 2019a], various applications, 38 data sets				
`num.trees`	1	2000	187,85 to 1908,25 *	
`replace`			FALSE	binary
`sample.fraction`	0.1	1	0,257 to 0,974 *	
`mtry`	0	1	0,035 to 0,954 *	transformed: `mtry` $\times m$
`respect.unordered.factors`			FALSE oder TRUE	binary
`min.node.size`	0	1	0,007 to 0,513 *	transformed: $n^{\text{min.node.size}}$
[Schratz et al., 2019], spatial data, 1 data set				
`num.trees`	10	10000	NA	
`mtry`	1	11	NA	
[Wong et al., 2019], medical data, 1 data set				
`num.trees`	10	2000	1000	
`mtry`	10	200	50	

2.6 Model: eXtreme Gradient Boosting

Boosting is an ensemble process. In contrast to random forests, the individual models (here: decision trees) are not created and evaluated at the same time, but rather sequentially. The basic idea is that each subsequent model tries to compensate for the weaknesses of the previous models.

For this purpose, a model is created repeatedly. The model is trained with weighted data. At the beginning these weights are identically distributed. Data that are poorly predicted or recognized by the model are given larger weights in the next step and thus have a greater influence on the next model. All models generated in this way are combined as a linear combination to form an overall model [Freund and Schapire, 1997, Drucker and Cortes, 1995].

An intuitive description of this approach is *slow learning*, as the attempt is not made to understand the entire database in a single step, but to improve the understanding step by step [James et al., 2014]. Gradient Boosting (GB) is a

variant of this approach, with one crucial difference: instead of changing the weighting of the data, models are created sequentially that follow the gradient of a loss function. In the case of regression, the models learn with residuals of the sum of all previous models. Each individual model tries to reduce the weaknesses (here: residuals) of the ensemble [Friedman, 2001].

In the following we consider the hyperparameters of one version of GB: Extreme Gradient Boosting (XGBoost) [Chen and Guestrin, 2016]. In principle, any models can be connected in ensembles via boosting. We apply XGBoost to decision trees. As a reference implementation, we refer to the R package xgboost [Chen et al., 2020]. Brownlee [2018] describes some empirical hyperparameter values for tuning XGBoost.

2.6.1   eXtreme Gradient Boosting Parameter: `eta`

The parameter `eta` is a learning rate and is also called "shrinkage" parameter. The parameter controls the lowering of the weights in each boosting step [Chen and Guestrin, 2016, Friedman, 2002]. It has the following practical meaning: lowering the weights helps to reduce the influence of individual trees on the ensemble. This can also avoid overfitting [Chen and Guestrin, 2016].

Type: double, scalar.

Sensitivity / robustness: Empirical results show that XGBoost is more sensitive to `eta` when `eta` is large [Friedman, 2001]. Generally speaking, smaller values are better. In an empirical study, Probst and Boulesteix [2018] describe `eta` as a parameter with comparatively high tunability.

Heuristics for determination: A heuristic is difficult to formulate due to the dependence on other parameters and the data situation, but Hastie et al. [2017] recommend

> ... the best strategy appears to be to set `eta` to be very small ($eta < 0.1$) and then choose `nrounds` by early stopping.

This may lead to correspondingly longer runtimes due to large `nrounds`. Brownlee [2018] mentions a heuristic, which describes a search range depending on `nrounds`.

Plausible range of values: $eta \in [0, 1]$. Using a logarithmic scale seems reasonable, e.g., $2^{-10}, \ldots, 2^0$), as used in the studies by Probst and Boulesteix [2018] or Sigrist [2020], because values close to zero often show good results.

Further constraints: none.

Interactions: There is a connection between `eta` and `nrounds`: If one of the two parameters increases, the other should be decreased if the error remains the same [Friedman, 2001], [Probst et al., 2019a]. This is also demonstrated by Hastie et al. [2017]:

> Smaller values of `eta` lead to larger values of `nrounds` for the same training risk, so that there is a trade-off between them.

In addition, Hastie et al. [2017] also point to correlations with the `subsample` parameter: In an empirical study, `subsample` = 1 and `eta` = 1 show significantly worse results than `subsample` = 0.5 and `eta` = 0.1. If `subsample` = 0.5 and `eta` = 1, the results are even worse than for `eta` = 1 and `subsample` = 1. In the best case (`subsample` = 0.5 and `eta` = 0.1), however, larger values of `nrounds` are required to achieve optimal results.

2.6.2   eXtreme Gradient Boosting Parameter: `nrounds`

The parameter `nrounds` specifies the number of boosting steps. Since a tree is created in each individual boosting step, `nrounds` also controls the number of trees that are integrated into the ensemble as a whole. Its practical meaning can be described as follows: larger values of `nrounds` mean a more complex and possibly more precise model, but also cause a longer running time. The practical meaning is therefore very similar to that of num.trees in random forests. In contrast to num.trees, overfitting is a risk with very large values, depending on other parameters such as `eta`, `lambda`, `alpha`. For example, the empirical results of Friedman [2001] show that with a low `eta`, even a high value of `nrounds` does not lead to overfitting.

Type: integer, scalar.

Sensitivity / robustness Similar to the random forests parameter num.trees, `nrounds` also has a higher sensitivity, especially with low values [Friedman, 2001].

Heuristics for determination: Heuristics cannot be derived from the literature. Often values of several hundred to several thousand trees are set as the upper limit [Brownlee, 2018].

Plausible range of values: $\in [1, \infty[$. Only integer values are valid.

Further constraints: none.

Interactions: There is a connection between the hyperparameters beta, rounds, and `subsample`.

2.6.3 eXtreme Gradient Boosting Parameter: `lambda`

The parameter `lambda` is used for the regularization of the model. This parameter influences the complexity of the model [Chen and Guestrin, 2016, Chen et al., 2020] (similar to the parameter of the same name in elastic net). Its practical significance can be described as follows: as a regularization parameter, `lambda` helps to prevent overfitting [Chen and Guestrin, 2016]. With larger values, smoother or simpler models are to be expected.

Type: double, scalar.

Sensitivity / robustness: not known.

Heuristics for determination: none known.

Plausible range of values: `lambda` $\in [0, \infty[$. A logarithmic scale seems to be useful, e.g., $2^{-10}, \ldots, 2^{10}$, as used in the study by Probst et al. [2019a] to cover a wide range of very small and very large values.

Further constraints: none.

Interactions: Because both `lambda` and `alpha` control the regularization of the model, an interaction is likely.

2.6.4 eXtreme Gradient Boosting Parameter: `alpha`

Chen and Guestrin [2016] did not mention this parameter. The documentation of the reference implementation does not provide any detailed information on `alpha` either. Due to the description as a parameter for the L1 regularization of the weights [Chen et al., 2020], a highly similar use as for the parameter of the same name in elastic net is to be assumed. Its practical meaning can be described as follows: similar to `lambda`, `alpha` also functions as a regularization parameter.

Type: double, scalar.

Sensitivity / robustness: unknown.

Heuristics for determination: No heuristics are known.

Plausible range of values: `alpha` $\in [0, \infty[$. A logarithmic scale seems to be useful, e.g., $2^{-10}, \ldots, 2^{10}$, as used in the study by Probst et al. [2019a] to cover a wide range of very small and very large values.

Further constraints: none.

Interactions: Since both `lambda` and `alpha` control the regularization of the model, an interaction is likely.

2.6.5 eXtreme Gradient Boosting Parameter: `subsample`

In each boosting step, the new tree to be created is usually only trained on a subset of the entire data set, similar to random forest [Friedman, 2002]. The `subsample` parameter specifies the portion of the data approach that is randomly selected in each iteration. Its practical significance can be described as follows: an obvious effect of small `subsample` values is a shorter running time for the training of individual trees, which is proportional to the `subsample` [Hastie et al., 2017].

Type: double, scalar.

Sensitivity / robustness: The study by Friedman [2002] shows a high sensitivity for very small or large values of `subsample`. In a relatively large range of values from `subsample` (around 0.3 to 0.6), however, hardly any differences in model quality are observed.

Determination heuristics: Hastie et al. [2017] suggest `subsample` = 0.5 as a good starting value, but point out that this value can be reduced if `nrounds` increases. With many trees (nround is large) it is sufficient if each individual tree sees a smaller part of the data, since the unseen data is more likely to be taken into account in other trees.

Plausible range of values: `subsample` $\in]0, 1]$. Based on the empirical results Friedman [2002], Hastie et al. [2017], a logarithmic scale is not recommended.

Further constraints: none.

Interactions: There is a connection between the `eta`, `nrounds`, and `subsample`.

2.6.6   eXtreme Gradient Boosting Parameter: `colsample_bytree`

The parameter `colsample_bytree` has similarities to the mtry parameter in random forests. Here, too, a random number of features is chosen for the splits of a tree. In XGBoost, however, this choice is made only once for each tree that is created, instead of for each split [xgboost developers, 2020]. Here `colsample_bytree` is a relative factor. The number of selected features is therefore `colsample_bytree` $\times n$. Its practical meaning is similar to mtry: `colsample_bytree` enables the trees of the ensemble to have a greater diversity. The runtime is also reduced, since a smaller number of splits have to be checked each time (if `colsample_bytree` < 1).

Type: double, scalar.

Sensitivity / robustness: The empirical study by Probst et al. [2019a] shows that the model is particularly sensitive to changes in the area of `colsample_bytree` = 1. However, this sensitivity decreases in the vicinity of more suitable values.

Heuristics for determination: none known.

Plausible range of values: `colsample_bytree` $\in]0, 1]$.   Brownlee [2018] mentions search ranges such as `colsample_bytree` $= 0.4, 0.6, 0.8, 1$, but mostly works with `colsample_bytree` $= 0.1, 0.2, \ldots, 1$.

Further constraints: none.

Interactions: none known.

2.6.7   eXtreme Gradient Boosting Parameter: `gamma`

This parameter of a single decision tree is very similar to the parameter `cp`: Like `cp`, `gamma` controls the number of splits of a tree by assuming a minimal improvement for each split. According to the documentation [Chen et al., 2020]:

> Minimum loss reduction required to make a further partition on a leaf node of the tree. the larger, the more conservative the algorithm will be.

The main difference between `cp` and `gamma` is the definition of `cp` as a relative factor, while `gamma` is defined as an absolute value. This also means that the value range differs.

Plausible range of values: `gamma` $\in [0, \infty[$. A logarithmic scale seems to make sense, e.g., $2^{-10}, \ldots, 2^{10}$, as, e.g., in the study by Thomas et al. [2018] to cover a wide range of very small and very large values.

2.6.8   eXtreme Gradient Boosting Parameter: `maxdepth`

This parameter of a single decision tree is already known as `maxdepth`.

Sensitivity / heuristics: Hastie et al. [2017] state:

> Although in many applications $J = 2$ will be insufficient, it is unlikely that $J > 10$ will be required. Experience so far indicates that $4 \leq J \leq 8$ works well in the context of boosting, with results being fairly insensitive to particular choices in this range.[5]

2.6.9   eXtreme Gradient Boosting Parameter: `min_child_weight`

Like `gamma` and `maxdepth`, `min_child_weight` restricts the number of splits of each tree. In the case of `min_child_weight`, this restriction is determined using the Hessian matrix of the loss function (summed over all observations in each new terminal node) [Chen et al., 2020, Sigrist, 2020]. In experiments by Sigrist [2020], this parameter turns out to be comparatively difficult to tune: The results show that tuning with `min_child_weight` gives worse results than tuning with a similar parameter (limitation of the number of samples per sheet) [Sigrist, 2020].

Type: double, scalar.

Sensitivity / robustness: unknown.

Heuristics for determination: none known.

Plausible range of values: `min_child_weight` $\in$ $[0, \infty[$.   A logarithmic scale seems to make sense, e.g., $2^{-10}, \ldots, 2^{10}$, as used in the study by Probst et al. [2019a] to cover a wide range of very small and very large values.

---

[5] J is the number of nodes in a tree that is strongly influenced by `maxdepth`.

Table 7: Survey: examples from literature about XGBoost tuning. * denotes that results depend on the data (several data sets).

Hyperparameter	Lower bound	Upper bound	Result	Comments
[Probst et al., 2019a], several applications, 38 data sets				
nrounds	1	5000	920.7 to 4847.15 *	
eta	2^{-10}	2^0	0.002 to 0.445 *	
subsample	0.1	1	0.545 to 0.964 *	
max_depth	1	15	2.6 to 14 *	
min_child_weight	2^0	2^7	1.061 to 7.502 *	
colsample_bytree	0	1	0.334 to 0.922 *	
lambda	2^{-10}	2^{10}	0.004 to 29.755 *	
alpha	2^{-10}	2^{10}	0.002 to 6.105 *	
[Thomas et al., 2018], several applications, 16 data sets				
eta	0.01	0.2		
gamma	2^{-7}	2^6		
subsample	0.5	1		
max_depth	3	20		
colsample_bytree	0.5	1		
lambda	2^{-10}	2^{10}		
alpha	2^{-10}	2^{10}		
[Wang and Ni, 2019], Risk Classification, 1 data set				
eta	0.005	0.2		
subsample	0.8	1		
max_depth	5	30		
min_child_weight	0	10		
colsample_bytree	0.8	1		
gamma	0	0.02		
[Zhou et al., 2020], Tunnel construction, 1 data set				
nrounds	1	150	103	
eta	0.00001	1	0.152	
max_depth	1	15	15	
lambda	1	15	13	
alpha	1	15	1	
This study, see Sec. 5, CID				
nrounds	0	32	256	
eta	2^{-10}	0	0.125	

Further constraints: none.

Interactions: Interactions with parameters such as gamma and maxdepth are probable, since all three parameters influence the complexity of the individual trees in the ensemble.

Table 7 shows example parameter settings from the literature.

2.7   Model: Support Vector Machine

The Support Vector Machine (SVM) is a kernel-based model. A kernel is a real-valued, symmetrical function $k(x, x')$ (usually positive-definite), which often expresses some form of similarity between two observations x, x'. The usefulness of kernels can be explained by the Kernel-Trick. The Kernel-Trick describes the ability of kernels, to transfer data into a higher-dimensional feature space. This allows, e.g., classification with linear decision boundaries (hyperplanes) even in cases where the data in the original features space are not linearly separable [Schölkopf and Smola, 2001].

As reference implementation, we use the R package e1071[6] [Meyer et al., 2020]. Based on libsvm [Chang and Lin, 2011].

---

[6]https://cran.r-project.org/package=e1071

2.7.1   Support Vector Machine Parameter: `kernel`

The parameter `kernel` is central for the SVM model. It describes the choice of the function $k(x, x')$. In practice, $k(x, x')$ can often be understood to be a measure of similarity. That is, the kernel function describes how similar to observations are to each other, depending on their feature values. In that context, the kernel function can express important properties about the relation between observations, such as differentiability.

Type: character, scalar.

Sensitivity / robustness: The empirical investigation of Probst et al. [2019a] shows: "*In svm the biggest gain in performance can be achieved by tuning the kernel, `gamma` or `degree`, while the cost parameter does not seem to be very tunable.*" This does not necessarily mean that `cost` should not be tuned, as the tunability investigated by Probst et al. [2019a] always considers a reference value (e.g., the default).

Heuristics for determination: Informally, it is often recommended to use `kernel = radial basis`. This also matches well to results and observations from the literature [Probst et al., 2019a, Guenther and Schonlau, 2016]. With very large numbers of observations and/or features Hsu et al. [2016] suggest to use `kernel = linear`. These are infallible rules, other kernels may perform better depending on the data set. This stresses the necessity of using hyperparameter tuning to choose kernels.

Plausible range of values: • linear: $k(x, x') = x^\mathrm{T} x'$
  • polynomial: $k(x, x') = (\texttt{gamma } x^\mathrm{T} x' + \texttt{coef0})^{\texttt{degree}}$
  • radial basis: $k(x, x') = \exp(-\texttt{gamma } ||x - x'||^2)$
  • sigmoid: $k(x, x') = \tanh(\texttt{gamma } x^\mathrm{T} x' + \texttt{coef0})$

Further constraints: none.

Interactions: The kernel functions themselves have parameters (`degree`, `gamma`, and `coef0`), whose values only matter if the respective function is chosen.

2.7.2   Support Vector Machine Parameter: `degree`

The parameter `degree` influences the kernel function:

  • polynomial: $k(x, x') = (\texttt{gamma } x^\mathrm{T} x' + \texttt{coef0})^{\texttt{degree}}$

Integer values of `degree` determine the degree of the polynomial. Non-integer values are possible, even though not leading to a polynomial in the classical sense. If `degree` has a value close to one, the polynomial kernel approximates the linear kernel. Else, the kernel becomes correspondingly non-linear.

Type: double, scalar.

Sensitivity / robustness: The empirical investigation of Probst et al. [2019a] shows: "*In svm the biggest gain in performance can be achieved by tuning the kernel, `gamma` or `degree`, while the cost parameter does not seem to be very tunable.*"

Heuristics for determination: none are known.

Plausible range of values: $\texttt{degree} \in (0, \infty)$.

Further constraints: none.

Interactions: The parameter only has an impact if `kernel` =polynomial.

2.7.3   Support Vector Machine Parameter: `gamma`

The parameter `gamma` influences three kernel functions:

• polynomial: $k(x, x') = (\texttt{gamma } x^\mathrm{T} x' + \texttt{coef0})^{\texttt{degree}}$
• radial basis: $k(x, x') = \exp(-\texttt{gamma } ||x - x'||^2)$
• sigmoid: $k(x, x') = \tanh(\texttt{gamma } x^\mathrm{T} x' + \texttt{coef0})$

In case of polynomial and sigmoid, `gamma` acts as a multiplier for the scalar product of two feature vectors. For radial basis, `gamma` acts as a multiplier for the distance of two feature vectors.

In practice `gamma` scales how far the impact of a single data sample reaches in terms of influencing the model. With small `gamma`, an individual observation my potentially influence the prediction in a larger vicinity, since with increasing distance between x and x', their similarity will decrease more slowly (esp. with `kernel = radial basis`).

Type: double, scalar.

Sensitivity / robustness: The empirical investigation of van Rijn and Hutter [2018] shows that `gamma` is rather sensitive.

Heuristics for determination: The reference implementation uses a simple heuristic, to determine `gamma`: `gamma` $= 1/n$ [Meyer et al., 2020]. Another implementation (the `sigest` function in `kernlab`[7]) first scales all input data, so that each feature has zero mean and unit variance. Afterwards, a good interval for `gamma` is determined, by using the 10% and 90% quantile of the distances between the scaled data samples. By default, 50% randomly chosen samples from the input data are used.

Plausible range of values: `gamma` $\in [0, \infty)$. Using a logarithmic scale seems reasonable (e.g.: $2^{-10}, ..., 2^{10}$ as used by Probst et al. [2019a]), to cover a broad spectrum of very small and very large values.

Further constraints: none.

Interactions: This parameter has no effect when `kernel` = linear. In addition, empirical results show a clear interaction with `cost` [van Rijn and Hutter, 2018].

2.7.4   Support Vector Machine Parameter: `coef0`

The parameter `coef0` influences two kernel functions:

- polynomial: $k(x, x') = (\text{gamma } x^T x' + \text{coef0})^{\text{degree}}$
- sigmoid: $k(x, x') = \tanh(\text{gamma } x^T x' + \text{coef0})$

In both cases, `coef0` is added to the scalar product of two feature vectors.

Type: double, scalar.

Sensitivity / robustness: Empirical results of Zhou et al. [2011] show that `coef0` has a strong impact in case of the polynomial kernel (but only for `degree` =2).

Heuristics for determination: Guenther and Schonlau [2016] suggest to leave this parameter at `coef0` =0.

Plausible range of values: `coef0` $\in \mathbb{R}$.

Further constraints: none.

Interactions: This parameter is only active if `kernel` =polynomial or `kernel` =sigmoid.

2.7.5   Support Vector Machine Parameter: `cost`

The parameter `cost` (often written as C) is a constant that weights constraint violations of the model. C is a typical regularization parameter, which controls the complexity of the model [Cherkassky and Ma, 2004], and may help to avoid overfitting or dealing with noisy data.[8]

Type: double, scalar.

Sensitivity / robustness: The empirical results of van Rijn and Hutter [2018] show that `cost` has a strong impact on the model, while the investigation of Probst et al. [2019a] determines only a minor tunability. This disagreement may be explained, since `cost` may have a huge impact in extreme cases, yet good parameter values are found close to the default values.

Heuristics for determination: Cherkassky and Ma [2004] suggest the following: `cost` $= \max(|\bar{y}+3\sigma_y|, |\bar{y}-3\sigma_y|)$. Here, $\bar{y}$ is the mean of the observed y values in the training data, and σ_y is the standard deviation. They justify this heuristic, by pointing out a connection between `cost` and the predicted y: As a constraint, `cost` limits the output values of the SVM model (regression) and should hence be set in a similar order of magnitude as the observed y [Cherkassky and Ma, 2004].

Plausible range of values: `cost` $\in [0, \infty)$. Using a logarithmic scale seems reasonable (e.g.: $2^{-10}, ..., 2^{10}$ as used by Probst et al. [2019a]), to cover a broad spectrum of very small and very large values.

Further constraints: none.

Interactions: Empirical results show a clear interaction with `gamma` [van Rijn and Hutter, 2018].

---

[7] https://cran.r-project.org/package=kernlab

[8] Here, complexity does not mean the number of model coefficients (as in linear models) or splits (decision trees), but the potential to generate more active / rugged functions. In that context, C influences the number of support vectors in the model. A high model complexity (many support vectors) can create functions with many peaks. This may lead to overfitting.

2.7.6   Support Vector Machine Parameter: `epsilon`

The parameter `epsilon` defines a ribbon around predictions. Residuals within that ribbon are tolerated by the model, i.e., are not penalized [Schölkopf and Smola, 2001]. The parameter is only used for regression with SVM, not for classification. In the experiments of this investigation (see Sec. 6) `epsilon` is only considered when SVM is used for regression.

Similarly to `cost`, `epsilon` is a regularization parameter With larger values, `epsilon` allows for larger errors / residuals. This reduces the number of support vectors (and incidentally, also the run time). The model becomes more smooth (cf. [Schölkopf and Smola, 2001, Fig. 9.4]). This can be useful, e.g., to deal with noisy data and avoid overfitting. However, the model quality may be decreased.

Type: double, scalar.

Sensitivity / robustness: As described above, `epsilon` has a significant impact on the model.

Heuristics for determination: For SVM regression, Cherkassky and Ma [2004] suggest based on simplified assumptions and empirical results: `epsilon` $= 3\sigma\sqrt{\frac{\ln(n)}{n}}$. Here, σ^2 is the noise variance, which has to be estimated from the data, see, e.g., equations (22), (23), and (24) in Cherkassky and Ma [2004]. The noise variance is the remaining variance of the observations y, which can not be explained by an ideal model. This ideal model has to be approximated, e.g., with the nearest neighbor model [Cherkassky and Ma, 2004]), resulting into additional computational effort.

Plausible range of values: `epsilon` $\in (0, \infty)$.

Further constraints: none.

Interactions: none are known.

In conclusion, Table 8 provides a brief survey of examples from the literature, where SVM was tuned.

Table 8: Survey of examples from the literature, for tuning of SVM ("*" denotes that results depend on data set (multiple data sets)).

Hyperparameter	Lower bound	Upper bound	Result	Notes
[Probst et al., 2019a], various applications, 38 data sets				
`kernel`			radial basis	
`cost`	2^{-10}	2^{10}	0,002 to 963,81 *	
`gamma`	2^{-10}	2^{10}	0,003 to 276,02 *	
`degree`	2	5	2 to 4 *	
[Mantovani et al., 2015], various applications, 70 data sets				
`cost`	2^{-2}	2^{15}		
`gamma`	2^{-15}	2^{3}		
[van Rijn and Hutter, 2018], various applications, 100 data sets				
`cost`	2^{-5}	2^{15}		
`gamma`	2^{-15}	2^{3}		
`coef0`	-1	1		only sigmoid
`tolerance`	10^{-5}	10^{-1}		
[Sudheer et al., 2013], flow rate prediction (hydrology), 1 data set				
`cost`	10^{-5}	10^{5}	1,12 to 1,93 *	
`epsilon`	0	10	0,023 to 0,983 *	
`gamma`	0	10	0,59 to 0,87 *	

2.8   Concluding Overview

On the basis of our literature survey, we recommend tuning the hyperparameters. Reasonable bounds are summarized in Table 9.

3   Tuning Procedures

This section gives an overview of the various relevant tuning procedures, which could be applied to the problem of ML algorithm tuning.

Table 9: Overview of hyperparameters in the experiments. For data type, we employ the signifiers used in R. For categorical parameters, we list categories instead of providing bounds.

Model	Hyperparameter	Data type	Lower bound	Upper bound	Lower bound (trans.)	Upper bound (trans.)	Transformation
k-NN	`k`	integer	1	30	1	30	–
	`p`	numeric	0.1	100	-1	2	10^p
EN	`alpha`	numeric	0	1	0	1	–
	`thresh`	numeric	10^{-8}	10^{-1}	-8	-1	$10^{\texttt{thresh}}$
DT	`minsplit`	integer	1	300	1	300	–
	`minbucket`	integer	1	150	0.1	0.5	$\mathrm{round}(\max(\texttt{minsplit}\times$ $\texttt{minbucket}, 1))$
	`cp`	numeric	10^{-10}	10^0	-10	0	$10^{\texttt{cp}}$
	`maxdepth`	integer	1	30	1	30	–
RF	`num.trees`	integer	1	2048	0	11	$\mathrm{round}(2^{\texttt{num.trees}})$
	`mtry`	integer	1	n	1	n	–
	`sample.fraction`	numeric	0.1	1	0.1	1	–
	`replace`	factor	TRUE, FALSE				–
	`respect.-` `unordered.factors`	factor	ignore, order				–
xgBoost	`eta`	numeric	2^{-10}	2^0	-10	0	$2^{\texttt{eta}}$
	`nrounds`	integer	1	2048	0	11	$2^{\texttt{nrounds}}$
	`lambda`	numeric	2^{-10}	2^{10}	-10	10	$2^{\texttt{lambda}}$
	`alpha`	numeric	2^{-10}	2^{10}	-10	10	$2^{\texttt{alpha}}$
	`subsample`	numeric	0.1	1	0.1	1	–
	`colsample_bytree`	numeric	$1/n$	1	$1/n$	1	–
	`gamma`	numeric	2^{-10}	2^{10}	-10	10	$2^{\texttt{gamma}}$
	`max_depth`	integer	1	15	1	15	–
	`min_child_weight`	numeric	1	128	0	7	$2^{\texttt{min\_child\_weight}}$
SVM	`kernel`	factor	radial, sigmoid				–
	`gamma`	numeric	2^{-10}	2^{10}	-10	10	$2^{\texttt{gamma}}$
	`coef0`	numeric	-1	1	-1	1	–
	`cost`	numeric	2^{-10}	2^{10}	-10	10	$2^{\texttt{coef0}}$
	`epsilon`	numeric	10^{-8}	10^0	-8	0	$10^{\texttt{epsilon}}$

3.1 Model-free Search

3.1.1 Manual Search

A frequently applied approach is that ML algorithms are configured manually [Bergstra and Bengio, 2012]. Users apply their own experience and trial-and-error to find reasonable parameter values.

In individual cases, this approach may indeed yield good results: when expert knowledge about data, algorithms and parameters is available. At the same time, this approach has major weaknesses, e.g., it may require significant amount of work time by the users, bias may be introduced due to wrong assumptions, limited options for parallel computation, and extremely limited reproducibility. An automated approach is hence of interest.

3.1.2 Undirected Search

Undirected search algorithms determine new hyperparameter values independently of any results of their evaluation. Two important examples are Grid Search and Random search.

Grid Search covers the search space with a regular grid. Each grid point is evaluated. Random Search selects new values at random (usually independently, uniform distributed) in the search space.

Grid Search is a frequently used approach, as it is easy to understand and implement (including parallelization). As discussed by Bergstra and Bengio [2012], Random Search shares the advantages of Grid Search. However, they show that Random Search may be preferable to Grid Search, especially in high-dimensional spaces or when the importance of individual parameters is fairly heterogeneous. They hence suggest to use Random Search instead Grid Search if

such simple procedures are required. Probst et al. [2019a] also use a Random Search variant to determine the tunability of models and hyperparameters. For these reasons, we employ Random Search as a baseline for the comparison in our experimental investigation.

Next to Grid Search and Random Search, there are other undirected search methods. Hyperband is an extension of Random Search, which controls the use of certain resources (e.g., iterations, training time) [Li et al., 2018]. Another relevant set of methods are Design of Experiments methods, such as Latin Hypercube Designs [Leary et al., 2003].

3.1.3 Directed Search

One obvious disadvantage of undirected search is, that a large amount of the computational effort may be spent on evaluating solutions that cover the whole search space. Hence, only a comparatively small amount of the computational budget will be spent on potentially optimal or at least promising regions of the search space.

Directed search on the other hand may provide a more purposeful approach. Basically any gradient-free, global optimization algorithm could be employed. Prominent examples are Iterative Local Search [Hutter et al., 2009] and Iterative Racing [López-Ibáñez et al., 2016]. Metaheuristics like Evolutionary Algorithms or Swarm Optimization are also applicable [Yang and Shami, 2020]. In comparison to undirected search procedures, directed search has two frequent drawbacks: an increased complexity that makes implementation a larger issue, and being more complicated to parallelize.

In our own investigation, we employ a *model-based* directed search procedure. This is described in the following Sec. 3.2.

3.2 Model-based Search

A disadvantage of model-free, directed search procedures is that they may require a relatively large number of evaluations (i.e., long runtimes) to approximate the values of optimal hyperparameters.

When tuning ML algorithms, this can easily become problematic: if complex ML algorithms are tuned on large data sets, runtimes of a single hyperparameter evaluation may easily go up into the range of hours.

Model-based search is one approach to resolve this issue. These search procedures use information gathered during the search to learn the relationship between hyperparameter values and quality measures (e.g., misclassification error). The model that encodes this learned relationship is called the surrogate model.

The advantage of this approach is that a considerable part of the evaluation burden (i.e., the computational effort) can be shifted from 'real' evaluations to evaluations of the surrogate model, which should be faster to evaluate.

One variant of model-based search is Sequential Parameter Optimization Toolbox (SPOT) [Bartz-Beielstein et al., 2005]. Details of SPOT its application in practice is given by Bartz-Beielstein et al. [2021c]. SPOT was originally developed for tuning of optimization algorithms. The requirements and challenges of algorithm tuning in optimization broadly reflect those of tuning machine learning models. SPOT uses the following approach. A detailed, algorithmic description is given by Preuss and Bartz-Beielstein [2007].

1. Design: In a first step, several candidate solutions (here: different combinations of hyperparameter values).
2. Evaluation: All new candidate solutions are evaluated (here: training the respective ML model with the specified hyperparameter values and measuring the quality / performance).
3. Termination: Check whether a termination criterion has been reached (e.g., number of iterations, evaluations, runtime).
4. Training: If not terminated, the surrogate model will be trained with all data derived from the evaluated candidate solutions, thus learning how hyperparameters affect model quality.
5. Surrogate search: The trained model is used to perform a search for new, promising candidate solutions
6. Go to 2.

Note, that it can be useful to allow for user interaction with the tuner after the evaluation step. Thus, the user may affect changes of the search space (stretch or shrink bounds on parameters, eliminate parameters). However, we will consider a purely automatic search in our experiments. Figure 1 summarizes the approach of SPOT.

We use the R implementation of SPOT, as provided by the R package SPOT [Bartz-Beielstein et al., 2021b,c].

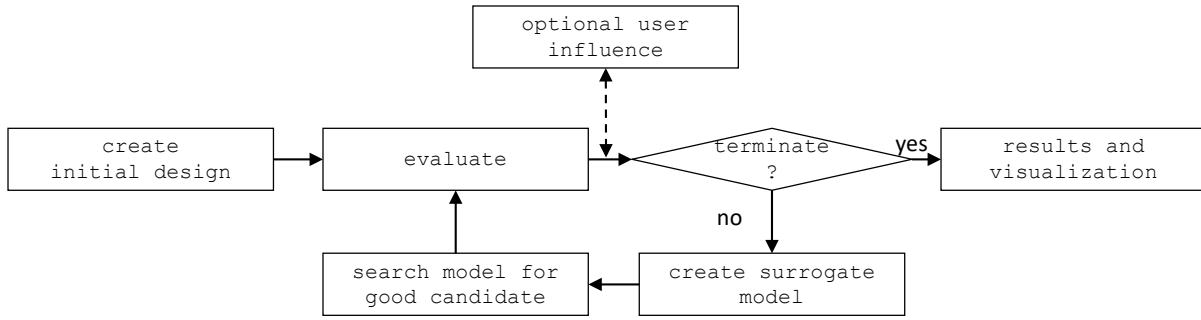Table 10 (taken from Bartz-Beielstein et al. [2021a]) compares SPOT with tuning procedures.

Figure 1: Visual representation of model-based search with SPOT.

Table 10: Overview of common features of tuning procedures. (✓) means that a feature is included. (◯) means that an implementation of that feature is possible with minimal effort, according to the respective authors. Table taken from Bartz-Beielstein et al. [2021a].

Features \Software	IRACE	SMAC	ParamILS	GGA	SPOT	Manual
Supports Numerical Variables	✓	✓	✓	✓	✓	✓
Supports Categorical Variables	✓	✓	✓	✓	✓	✓
Interactive Execution					✓	✓
Multi Instance Problems		✓				
Parallelization	✓		◯	✓	◯	
Model-Based		✓		✓	✓	
Visualization	✓				✓	
Noise Handling	✓				✓	
Graphical User Interface					✓	

4   Case Study I: Tuning Decision Trees (rpart)

This case study considers the Decision Tree algorithm which is detailed in Sec. 2.4. The hyperparameter tuning study applies Decision Tree to the CID, which will be described in Sec. 4.2. This case study considers a classification task.

We will explain how to find suitable parameter values and bounds, and perform experiments w.r.t. the following four Decision Tree hyperparameters:

1. `minsplit`
2. `minbucket`
3. `cp`
4. `maxdepth`

The R package Machine Learning in R (mlr) is used to set up a learning task. The corresponding parameters of this mlr task can be accessed and modified via SPOT. Section 4.1 illustrates how the R packages can be combined.

4.1   rpart: Project Setup

We start with the general setup. Since we are considering classification (income class, i.e, high income versus low income), the mlr task.type is set to `classif`. Regression trees are specified via `rpart`.

```
task.type <- "classif"
model <- "rpart"
data.seed <- 1
tuner.seed <- 1
timebudget <- 60  # budget for tuning in seconds
timeout <- NA  # do not use timeout (avoid overhead for quick tests)
```

4.2   rpart: The Data Set

For the investigation we choose the CID, which is made available, for example, via the UCI Machine Learning Repository. [9] This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. It includes 41 demographic and employment related variables. For our investigation, we will access the version of the data set that is available via the platform openml.org under the data record ID 45353 [Vanschoren et al., 2013]. This data set is an excerpt from the current population surveys for 1994 and 1995, compiled by the U.S. Census Bureau. It contains $m = 299,285$ observations with 41 features on demography and employment. The data set is comparatively large, has many categorical features with many levels and fits well with the field of application of official statistics.

The CID can be downloaded as follows:

```
dirName = "oml.cache"
if (!(dir.exists(dirName))) dir.create(dirName)
setOMLConfig(apikey = "c1994bdb7ecb3c6f3c8f3b35f4b47f1f", cachedir = dirName)

## OpenML configuration:
##    server          : http://www.openml.org/api/v1
##    cachedir        : oml.cache
##    verbosity       : 1
##    arff.reader     : farff
##    confirm.upload  : TRUE
##    apikey          : **************************47f1f

dataOML <- getOMLDataSet(4535)$data
```

Its 42 columns contain the following information:

```
str(dataOML)

## 'data.frame': 299285 obs. of  42 variables:
##  $ V1 : num  73 58 18 9 10 48 42 28 47 34 ...
##  $ V2 : Factor w/ 9 levels " Federal government",..: 4 7 4 4 4 5 5 5 2 5 ...
##  $ V3 : num  0 4 0 0 0 40 34 4 43 4 ...
##  $ V4 : num  0 34 0 0 0 10 3 40 26 37 ...
##  $ V5 : Factor w/ 17 levels " 10th grade",..: 13 17 1 11 11 17 10 13 17 17 ...
##  $ V6 : num  0 0 0 0 1200 0 0 876 0 ...
##  $ V7 : Factor w/ 3 levels " College or university",..: 3 3 2 3 3 3 3 3 3 3 ...
##  $ V8 : Factor w/ 7 levels " Divorced"," Married-A F spouse present",..: 7 1 5 5 5 3 3 5 3 3 ...
##  $ V9 : Factor w/ 24 levels " Agriculture",..: 15 5 15 15 15 7 8 5 6 5 ...
##  $ V10: Factor w/ 15 levels " Adm support including clerical",..: 7 9 7 7 7 11 3 5 1 6 ...
##  $ V11: Factor w/ 5 levels " Amer Indian Aleut or Eskimo",..: 5 5 2 5 5 1 5 5 5 5 ...
##  $ V12: Factor w/ 10 levels " All other"," Central or South American",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ V13: Factor w/ 2 levels " Female"," Male": 1 2 1 1 1 1 2 1 1 2 ...
##  $ V14: Factor w/ 3 levels " No"," Not in universe",..: 2 2 2 2 2 1 2 2 1 2 ...
##  $ V15: Factor w/ 6 levels " Job leaver",..: 4 4 4 4 4 4 4 2 4 4 ...
##  $ V16: Factor w/ 8 levels " Children or Armed Forces",..: 3 1 3 1 1 2 1 7 2 1 ...
##  $ V17: num  0 0 0 0 0 ...
##  $ V18: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ V19: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ V20: Factor w/ 6 levels " Head of household",..: 5 1 5 5 5 3 3 6 3 3 ...
##  $ V21: Factor w/ 6 levels " Abroad"," Midwest",..: 4 5 4 4 4 4 4 4 4 4 ...
##  $ V22: Factor w/ 51 levels " ?"," Abroad",..: 37 6 37 37 37 37 37 37 37 37 ...
##  $ V23: Factor w/ 38 levels " Child <18 ever marr not in subfamily",..: 30 21 8 3 3 37 21 36 37 21 ...
##  $ V24: Factor w/ 8 levels " Child 18 or older",..: 7 5 1 3 3 8 5 6 8 5 ...
##  $ V25: num  1700 1054 992 1758 1069 ...
##  $ V26: Factor w/ 10 levels " ?"," Abroad to MSA",..: 1 4 1 6 6 1 6 1 1 6 ...
##  $ V27: Factor w/ 9 levels " ?"," Abroad",..: 1 9 1 7 7 1 7 1 1 7 ...
##  $ V28: Factor w/ 10 levels " ?"," Abroad",..: 1 10 1 8 8 1 8 1 1 8 ...
##  $ V29: Factor w/ 3 levels " No"," Not in universe under 1 year old",..: 2 1 2 3 3 2 3 2 2 3 ...
##  $ V30: Factor w/ 4 levels " ?"," No"," Not in universe",..: 1 4 1 3 3 1 3 1 1 3 ...
```

---

[9] https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)

```
##  $ V31: num  0 1 0 0 0 1 6 4 5 6 ...
##  $ V32: Factor w/ 5 levels " Both parents present",..: 5 5 5 1 1 5 5 5 5 5 ...
##  $ V33: Factor w/ 43 levels " ?"," Cambodia",..: 41 41 42 41 41 32 41 41 41 41 ...
##  $ V34: Factor w/ 43 levels " ?"," Cambodia",..: 41 41 42 41 41 41 41 41 41 41 ...
##  $ V35: Factor w/ 43 levels " ?"," Cambodia",..: 41 41 42 41 41 41 41 41 41 41 ...
##  $ V36: Factor w/ 5 levels " Foreign born- Not a citizen of U S ",..: 5 5 1 5 5 5 5 5 5 5 ...
##  $ V37: num  0 0 0 0 0 2 0 0 0 0 ...
##  $ V38: Factor w/ 3 levels " No"," Not in universe",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ V39: num  2 2 2 0 0 2 2 2 2 2 ...
##  $ V40: num  0 52 0 0 0 52 52 30 52 52 ...
##  $ V41: num  95 94 95 94 94 95 94 95 95 94 ...
##  $ V42: Factor w/ 2 levels " - 50000."," 50000+.": 1 1 1 1 1 1 1 1 1 1 ...
```

4.3   rpart: Create an mlr Task

The target variable of the data set is the income class (income below or above US $ 50,000). It is stored in column 42.

```
task <- makeClassifTask(data = dataOML, target = "V42")
```

Using mlr, a resampling strategy can be defined as follows:

```
rsmpl <- makeResampleDesc("Holdout", split = 0.6)
```

4.4   rpart: Define the Learner (Algorithm)

The we define the mlr learner and print the learner's parameters. Please note, that the parameter bounds given by mlr are not always correct. If they are used, they may lead to invalid configurations (the tuning procedure will try to work around this via a "penalty", but this may be an avoidable waste of run time).

```
learner <- paste(task.type, model, sep = ".")
print(getParamSet(learner))

##                      Type len  Def    Constr Req Tunable Trafo
## minsplit          integer   -   20 1 to Inf   -    TRUE     -
## minbucket         integer   -    - 1 to Inf   -    TRUE     -
## cp                numeric   - 0.01   0 to 1   -    TRUE     -
## maxcompete        integer   -    4 0 to Inf   -    TRUE     -
## maxsurrogate      integer   -    5 0 to Inf   -    TRUE     -
## usesurrogate     discrete   -    2    0,1,2   -    TRUE     -
## surrogatestyle   discrete   -    0      0,1   -    TRUE     -
## maxdepth          integer   -   30  1 to 30   -    TRUE     -
## xval              integer   -   10 0 to Inf   -   FALSE     -
## parms             untyped   -    -        -   -    TRUE     -
```

4.5   rpart: Define the Experiment Configuration

The experiment configuration contains information about the mlr learner to be tuned, the task, the resampling strategy, parameter names as specified for the mlr learner, lower and upper bounds (see also transformations), data type of each parameter, and further information. Fixed parameters can be specified, in our case, there are none, so the list is empty. The same holds for the factor levels for categorical parameters. There are none of them in our configuration.

Important note: minbucket is set relative to minsplit, i.e., we are using numerical values for minbucket that represent *percentages* relative to minsplit. If minbucket = 1.0, then minbucket = minsplit (minsplit values should be greater equal minbucket values).

```
cfg <- list(
  learner = learner,
  task = task,
  resample = rsmpl,
  tunepars = c("minsplit", "minbucket", "cp", "maxdepth"),
```

```
  lower = c(1,
            0.1,-10,
            1),
  upper = c(300,
            0.5,
            0,
            30),
  type =  c("integer", "numeric", "numeric", "integer"),
  fixpars = list(),
  factorlevels = list(),
  transformations = c(trans_id, #identity transformation
                      trans_id,
                      trans_10pow, # x^10 transformation
                      trans_id),
  dummy = FALSE,
  # set parameters relative to other parameters. here: minbucket relative to minsplit
  relpars = list(minbucket = expression(round(
    max(minsplit * minbucket, 1)
  )))
  # if no relative parameters needed, use empty list for relpars
)
```

The, the objective function can be defined:

```
objf <- get_objf(config = cfg, timeout = timeout)
```

4.6   rpart: Run Tuning with SPOT

Now, everything is prepared and the SPOT run can be started. Note, that using `timebudget/60` converts the timebudget to minutes.

```
result <- spot(fun = objf, lower = cfg$lower, upper = cfg$upper, control = list(types = cfg$type,
    maxTime = timebudget/60, plots = FALSE, progress = FALSE, model = buildKriging,
    optimizer = optimDE, noise = TRUE, seedFun = 123, seedSPOT = tuner.seed, designControl = list(size = 5 *
        length(cfg$lower)), funEvals = Inf, modelControl = list(target = "y", useLambda = TRUE,
        reinterpolate = TRUE), optimizerControl = list(funEvals = 100 * length(cfg$lower))))

save(result, file = "rpartResult.RData")
```

```
load(file = "R.d/case1Result.RData")
```

4.7   rpart: Visualize the Results

First, we visualize the progress of the SPOT tuning procedure.

```
size <- 5 * length(cfg$lower)
plot(result$y, type = "l")
abline(v = size, col = "red")
y0 <- min(result$y[1:size])
abline(h = y0, col = "blue")
```

The best hyperparameter setting can be printed as follows:

```
(paraNames <- cfg$tunepars)

## [1] "minsplit"  "minbucket" "cp"        "maxdepth"

print(result$xbest)

##      [,1]      [,2]       [,3] [,4]
## [1,]   16 0.1770681 -3.213333   28
```
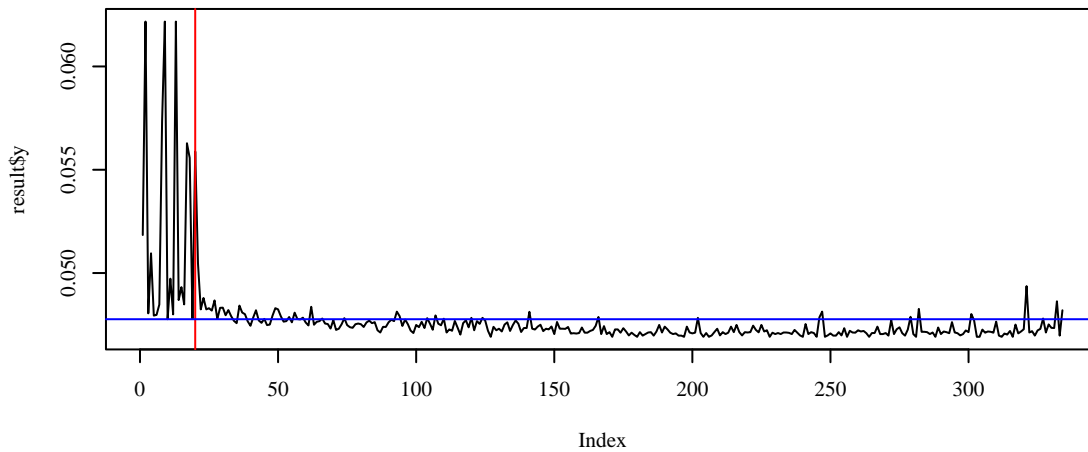
Figure 2: Case Study I: SPOT search. The vertical red line denotes the end of the initial design phase. The best result from the (random) initial phase is shown in blue.
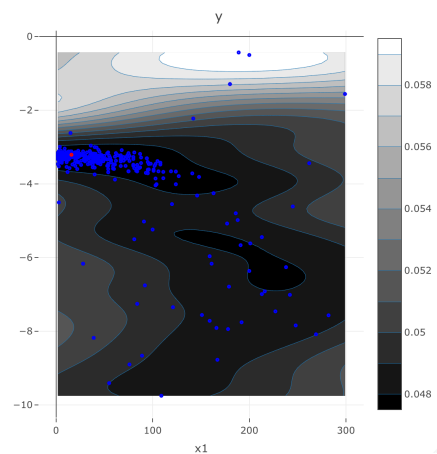
Figure 3: Surface plot: mce as a funtion of minsplit and cp. Note that the cp values (on the vertical axis) use the 10^x transformation.

The error (mce) can be shown as a function of minsplit and cp (based on surrogate) as shown in Fig. 3.

```
plot_surface(result, yvar = 1, which = c(1, 3))
```

Next, mce vs all parameters (based on evaluations) is shown in the parallel plot in Fig. 4.

```
plot_parallel(result, yvar = 1)
```

In addition, a sensitivity plot can be shown.

```
plot_sensitivity(result, type = "agg")
```

Finally, we fit a simple regression tree to analyze the most important hyperparameters.

```
library("rpart")
library("rpart.plot")
fitTree <- buildTreeModel(x = result$x, y = result$y, control = list(xnames = paraNames))  # paste0('x', 1:n)))
rpart.plot(fitTree$fit)
```
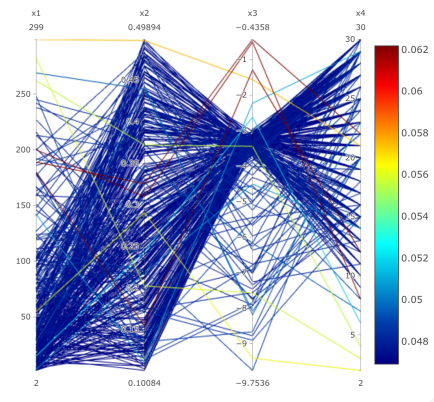
Figure 4: Parallel plot. Red lines denote poor configurations, e.g., large x_3 (cp) values are not recommended.
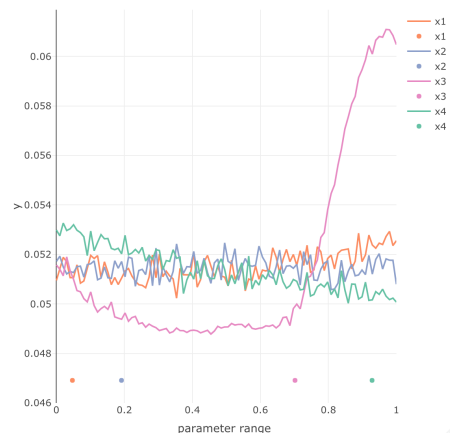
Figure 5: Sensitivity plot. Effect of normalized hyperparamter values. cp (x_3) appears to be the most important parameter.
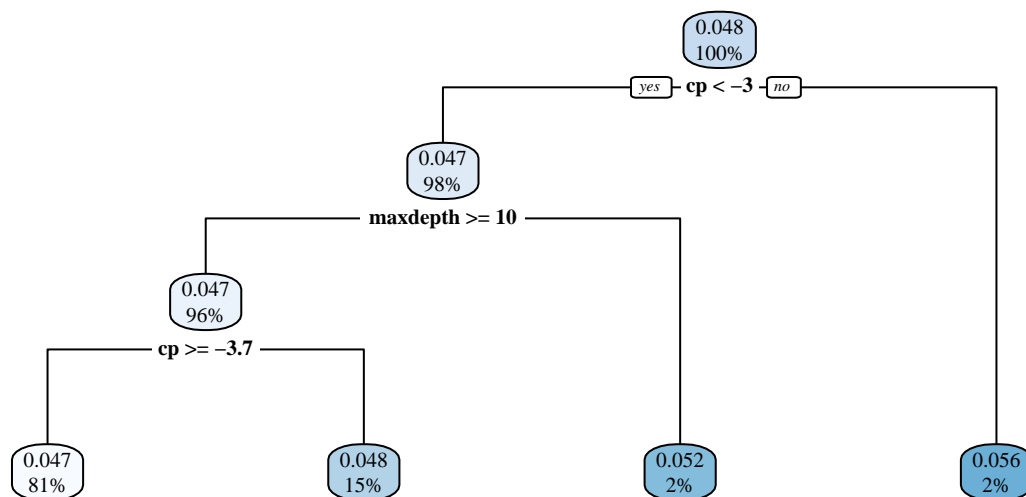
Our visualizations are based on the transformed values, i.e., the power transformation was applied to the cp values. Natural hyperparameter values can be obtained as follows:

```
result$x[, 3] <- trans_10pow(result$x[, 3])
```

4.8   rpart: Summary

Data visualizations shown in Sec. 4.7 indicate that hyperparameter `cp` has the greatest effect on the algorithm's performance. The recommended value is $cp \approx 10^{-3} = 0.001$.

5   Case Study II: Tuning XGBoost

Consider the eXtreme Gradient Boosting algorithm which is detailed in Sec. 2.6. We will explain how to find suitable parameter values and bounds, and perform experiments w.r.t. the following eXtreme Gradient Boosting parameters:

1. `nrounds`
2. `eta`
3. `lambda`
4. `alpha`
5. `subsample`
6. `colsample_bytree`
7. `gamma`
8. `max_depth`
9. `min_child_weight`

5.1   The Census-Income (KDD) Data Set

The first step is identical to the step in the `rpart` example in Sec. 4.

```
dirName = "oml.cache"
if (!(dir.exists(dirName))) dir.create(dirName)
setOMLConfig(apikey = "c1994bdb7ecb3c6f3c8f3b35f4b47f1f", cachedir = dirName)

## OpenML configuration:
##    server           : http://www.openml.org/api/v1
##    cachedir         : oml.cache
##    verbosity        : 1
##    arff.reader      : farff
##    confirm.upload   : TRUE
##    apikey           : ***************************47f1f

dataOML <- getOMLDataSet(4535)$data
task <- makeClassifTask(data = dataOML, target = "V42")
rsmpl <- makeResampleDesc("Holdout", split = 0.6)
```

5.2   eXtreme Gradient Boosting: Project Setup

The project setup is also similar to the setup described in Sec. 4. Because the XGBoost algorithm is more complex than the Decision Tree (DT), the computational budget is increased. The budget for tuning in seconds is set to $15 \times 3,600$ seconds or 15 hours.

```
task.type <- "classif"
data.seed <- 1
tuner.seed <- 1
timebudget <- 15 * 3600
timeout <- timebudget/20   # use 1/20 times the budget before tuning is stopped
```

5.3 eXtreme Gradient Boosting: Learner (Algorithm) Definition

The mode name is "xgboost".

```r
model <- "xgboost"
learner <- paste(task.type, model, sep = ".")
```

The complete list of XGBoost hyperparameters can be shown using the mlr function `getParamSet`. Since this is a long list, it is not shown here.

```r
print(getParamSet(learner))
```

5.4 eXtreme Gradient Boosting: Experiment Configuration

Now follows the most important part of the HPT task: the specification of a reasonable hyperparameter setting. We will use the examples from literature shown in Table 7 in Sec. 2.6 as a guideline.

```r
tunepars <- c("nrounds","eta","lambda","alpha","subsample","colsample_bytree","gamma","max_depth","min_child_we
lower <- c(0,
            -10,
            -10,
            -10,
            0.1,
            1/getTaskNFeats(task),
            ## slight deviation from Prob19a.
            ## reason: 0 makes no sense. At least one feature should be chosen via colsample.
            -10, ## smaller than Thom18a
            1,
            0)
upper <- c(5,   ## set similar as random forest (which is less than Prob19a used: 5000)
            0,
            10,
            10,
            1,
            1,
            10,  ## larger than Thom18a
            15,
            7)
type <-   c("numeric","numeric","numeric","numeric","numeric","numeric","numeric","integer","numeric")
if(task.type=="classif"){
  fixpars <- list(eval_metric="logloss",# suppress warning given when default metric is used.
              nthread=1) #one thread, not parallel
}else{
  fixpars <- list(eval_metric="rmse",# suppress warning given when default metric is used.
              nthread=1) #one thread, not parallel
}
factorlevels <- list()
transformations <- c(trans_2pow_round, ## differs from Prob19a
                     trans_2pow,
                     trans_2pow,
                     trans_2pow,
                     trans_id,
                     trans_id,
                     trans_2pow, ## Thom18a
                     trans_id,
                     trans_2pow)
dummy=TRUE
relpars <- list()
```

The imputation of missing values can be implemented as follows.

```r
task <- impute(task, classes = list(factor = imputeMode(), integer = imputeMedian(),
    numeric = imputeMean()))$task
```

All factor features will be replaces with their dummy variables. Internally model.matrix is used. Non-factor features will be left untouched and passed to the result.

```
if (dummy) {
    task <- createDummyFeatures(task)
}
str(task)

## List of 6
##  $ type       : chr "classif"
##  $ env        :<environment: 0x7fe0f65be088>
##  $ weights    : NULL
##  $ blocking   : NULL
##  $ coordinates: NULL
##  $ task.desc  :List of 13
##   ..$ id                : chr "dataOML"
##   ..$ type              : chr "classif"
##   ..$ target            : chr "V42"
##   ..$ size              : int 299285
##   ..$ n.feat            : Named int [1:4] 409 0 0 0
##   .. ..- attr(*, "names")= chr [1:4] "numerics" "factors" "ordered" "functionals"
##   ..$ has.missings      : logi FALSE
##   ..$ has.weights       : logi FALSE
##   ..$ has.blocking      : logi FALSE
##   ..$ has.coordinates   : logi FALSE
##   ..$ class.levels      : chr [1:2] " - 50000." " 50000+."
##   ..$ positive          : chr " - 50000."
##   ..$ negative          : chr " 50000+."
##   ..$ class.distribution: 'table' int [1:2(1d)] 280717 18568
##   .. ..- attr(*, "dimnames")=List of 1
##   .. .. ..$ : chr [1:2] " - 50000." " 50000+."
##   ..- attr(*, "class")= chr [1:3] "ClassifTaskDesc" "SupervisedTaskDesc" "TaskDesc"
##  - attr(*, "class")= chr [1:3] "ClassifTask" "SupervisedTask" "Task"
```

We can set the seed to improve reproducibility:

```
set.seed(data.seed)
```

The information is compiled to a list.

```
cfg <- list(learner = learner, tunepars = tunepars, lower = lower, upper = upper,
    type = type, fixpars = fixpars, factorlevels = factorlevels, transformations = transformations,
    dummy = dummy, relpars = relpars, task = task, resample = rsmpl)
```

The objective function is defined: this function receives a configuration for a tuning experiment, and returns an objective function to be tuned via SPOT.

```
objf <- get_objf(config = cfg, timeout = timeout)
```

5.5   eXtreme Gradient Boosting: Tuning Run with SPOT

The SPOT tuning run can be started as follows.

```
resultXGB <- spot(fun = objf,
              lower=cfg$lower,
              upper=cfg$upper,
              control = list(types=cfg$type,
                             maxTime = timebudget/60, #convert to minutes
                             plots=TRUE,
                             progress = TRUE,
                             model=buildKriging,
                             optimizer=optimDE,
                             noise=TRUE,
```
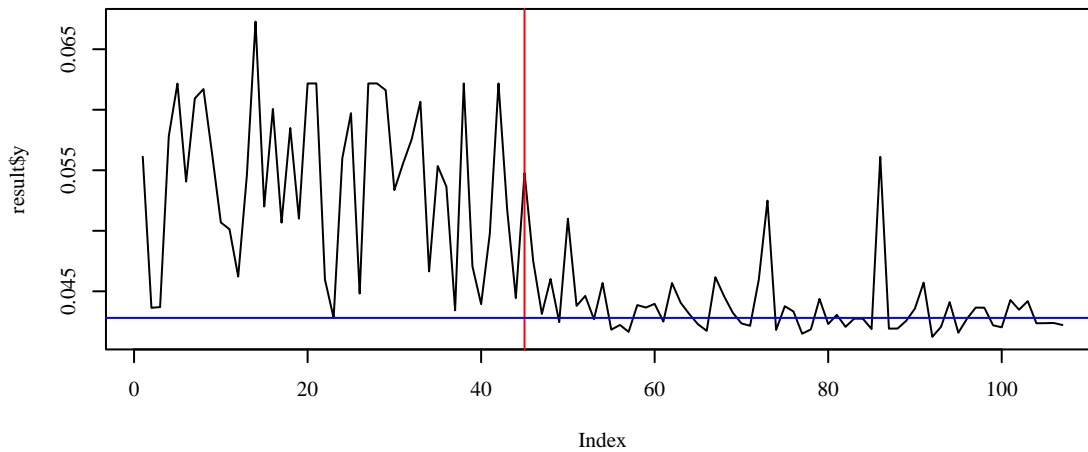
Figure 6: Case Study II: SPOT search. The vertical red line denotes the end of the initial design phase. The best result from the (random) initial phase is shown in blue.

```
                              seedFun=123,
                              seedSPOT=tuner.seed,
                              designControl=list(size=5*length(cfg$lower)),
                              funEvals=Inf,
                              modelControl=list(target="y",
                                                useLambda=TRUE,
                                                reinterpolate=TRUE),
                              optimizerControl=list(funEvals=100*length(cfg$lower))
              )
)
save(resultXGB, file="xgbResult.RData")
```

```
load(file = "xgbResult.RData")
result <- resultXGB
```

5.6   eXtreme Gradient Boosting: Visualization of the Results

First, we visualize the progress of the SPOT tuning procedure.

```
size <- 5 * length(cfg$lower)
plot(result$y, type = "l")
abline(v = size, col = "red")
y0 <- min(result$y[1:size])
abline(h = y0, col = "blue")
```

The best hyperparameter setting can be printed as follows:

```
(paraNames <- cfg$tunepars)

## [1] "nrounds"          "eta"              "lambda"          "alpha"
## [5] "subsample"        "colsample_bytree" "gamma"           "max_depth"
## [9] "min_child_weight"

print(result$xbest)

##         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7] [,8]
```
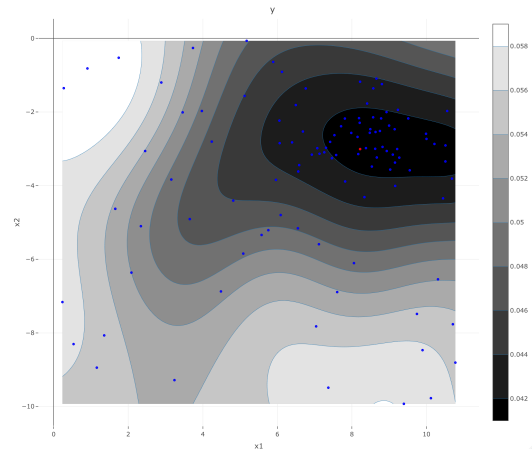
Figure 7: mce versus `nrounds` and `eta`. Values of x_1 (nrounds) are in the range from 2^0 to 2^{10} and values of x_2 (eta) in the range from 2^{-10} to 2^0.
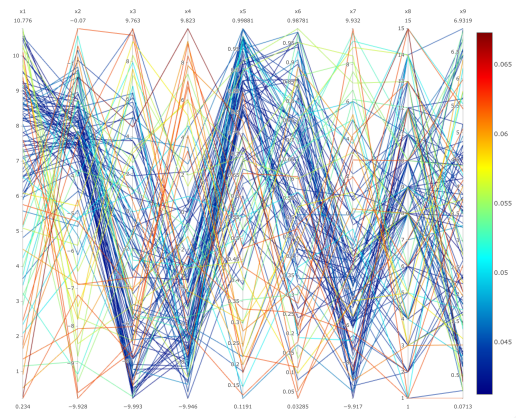
Figure 8: mce: parallel plot. Red lines denote poor configurations, blue lines better configurations. This plot indicates that x_1 and x_5 should be large whereas x_7 values should be low.

```
## [1,] 8.22029 -3.009606 3.767863 -0.5625179 0.9684651 0.6746676 -2.636762    9
##          [,9]
## [1,] 2.756446
```

First plot, based on surrogate:

```
plot_surface(result, yvar = 1)
```

Next, mce vs all parameters (based on evaluations) is shown in Fig. 8.

```
plot_parallel(resultXGB, yvar = 1)
```

In addition, a sensitivity plot can be shown.

```
plot_sensitivity(result, type = "agg")
```

Finally, we fit a simple regression tree to analyze the most important hyperparameters.

```
library("rpart")
library("rpart.plot")
fitTree <- buildTreeModel(x = result$x, y = result$y, control = list(xnames = paraNames))  # paste0('x', 1:n)))
rpart.plot(fitTree$fit)
```
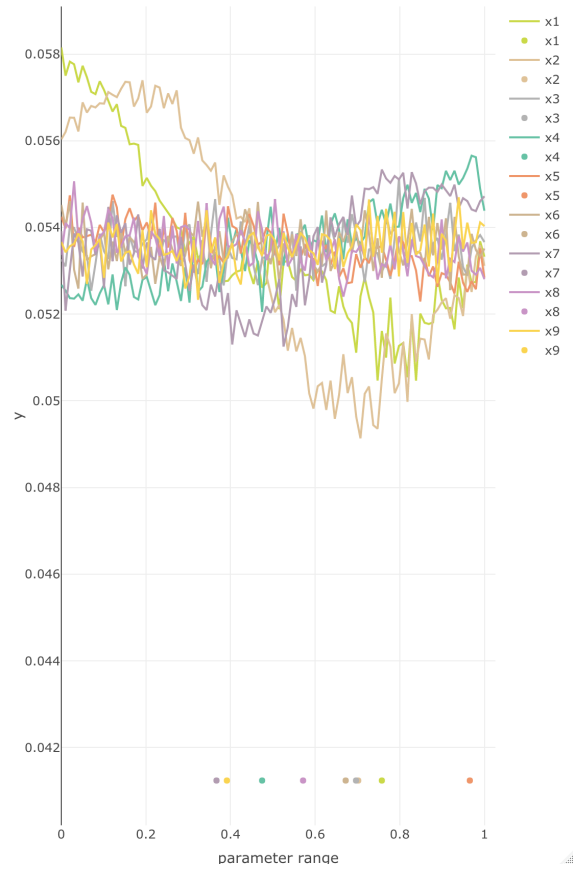
Figure 9: Sensitivity plot. Effect of normalized hyperparamter values. x_1 and x_2 appear to be the most important parameters.
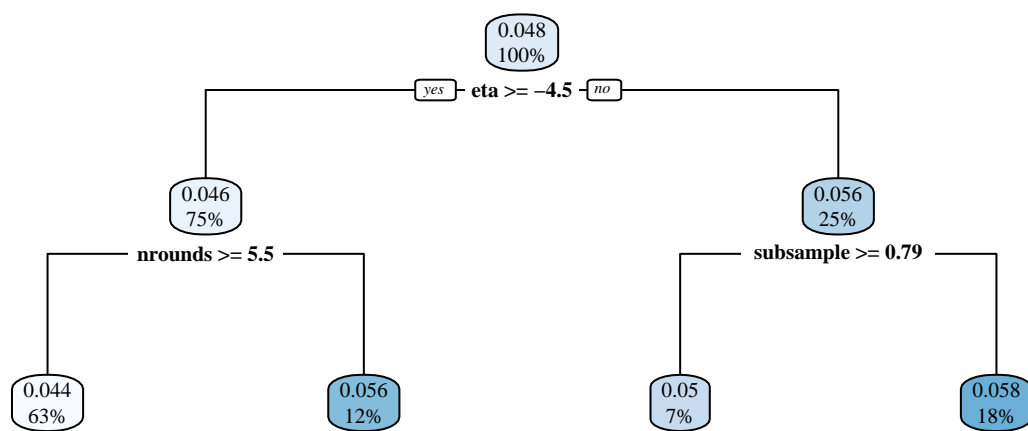
Our visualizations are based on the transformed values.

5.7 eXtreme Gradient Boosting: Summary

Data visualizations shown in Sec. 4.7 indicate that the hyperparameters `eta` and `nrounds` have the greatest effect on the algorithm's performance. The recommended values are `nrounds` $\approx 2^8 = 256$ and `eta` $\approx 2^{-3} = 0.125$.

6 Case Study III: Global Experimental Investigation

6.1 Research Questions

As described in Sec. 3, SPOT offers a robust approach for the tuning of ML algorithms, especially if the training and/or evaluation runtime become large.

In practice, the learning process of models and hence the choice of their hyperparameters is influenced by a plethora of other factors. On the one hand, this complex situation further motivates the use of tuning procedures, since the ML algorithms have to be adapted to new data or situations. On the other hand, this raises the question of how such factors influence the tuning process itself. We want to investigate this in a structured manner.

In detail, the following factors are objects of our investigation, testing their influence on the tuning procedures.

- Number of numerical features in the data (`nnumericals`).
- Number of categorical features in the data (`nfactors`).
- Cardinality of the categorical features, i.e., the maximal number of levels (`cardinality`).
- Number of observations in the data (m).
- Task type, classification or regression.
- Choice of model.

We want to investigate the following hypotheses.

(H-1) Tuning is necessary to find good parameter values (compared to defaults).

(H-2) Data: Properties of the data influence (the difficulty of) the tuning process.
- Information content: If the data has little information content, models are easier to tune, since more parameter configurations achieve near-optimal quality. In general, changing parameters has less impact on model quality in this case.
- Number of features: A larger number of features leads to longer runtimes, which affects how many evaluations can be made during tuning.
- Type of features: The number of numerical and / or categorical features as well as their cardinality influences how much information is available to the model, hence may affect the difficulty of tuning.
- Number of observations m: With increasing m, the average runtime of model evaluations will increase.

(H-3) Target variable: There is no fundamental difference between tuning regression or classification models.

(H-4) Model: The choice of model (e.g., EN or SVM) affects the difficulty of the tuning task, but not necessarily the choice of tuning procedure.

(H-5) Benchmark: The aptitude of the employed tuners can be measured in a statistically sound manner.

6.2 Setup

6.2.1 Generation of Data Sets

For this investigation we choose a slighly modified version of the CID data set, which was introduced in Sec. 4.2. This data set suits well to our research questions, as it is comparatively large, has many categorical features, several of the categorical features have a lot of levels, and it can be easily used to generate different classification as well as regression problems.

To investigate our research questions, the data set needs to be pre-processed accordingly. This pre-processing is documented in the included source code, and is briefly explained in the following.

- Feature 24 (instance weight MARSUPWT) is removed. This feature describes the number of persons in the population who are represented by the respective observation. This is relevant for data understanding, but should not be an input to the machine learning models.

- Several features are encoded as numerical(integer) variables, but are in fact categorical. Example: Feature 3 (industry code ADTIND) is encoded as an integer. Since the respective integers represent discrete codes for different sectors of industry, they have no inherent order and should be encoded as categorical features.

- The data set sometimes contains `NA` values (missing data). These `NA` values are replaced before modeling. For categorical features, the most frequently observed category is imputed (mode). For integer features, the median is imputed, and for real-valued features the mean.

- As the only model in the investigation, `xgboost` is not able to work directly with categorical features. In that case (only for `xgboost`) the categorical data features are transferred into a dummy-coding. For each category of the categorical feature, a new binary feature is created, which specifies whether an observation is of the respective category or not.

- Finally, we split the data randomly into test data (40% of the observations) and training data (60%).

In addition to these general pre-processing steps, we change properties of the data set for individual experiments, to cover our various hypotheses.

Arguably, we could have done this by using completely different data sets where each covers different objects of investigation (i.e., different numbers of features or different m). We decided to vary a single data set instead, generating new data sets with different properties, because this allows us to reasonably compare results between the individual variations. This way, we generate multiple data sets that cover different aspects and problems in detail. While they all derive from the same data set (CID), they all have different characteristics: Number of categorical features, number of numerical features, cardinality, number of observations, and target variable. These characteristics can be quantified with respect to difficulty as discussed in Sec. 6.4.4.

In detail, we vary:

- Target variable: The original target variable of the data set is the income class (below / above 50 000 USD). We choose *age* as the target variable instead. For classification experiments, age will be discretized, into two classes: age < 40 and age $>= 40$. For regression, age remains unchanged. This choice intends to establish comparability between both experiment groups (classification, regression).

- Number of categories (`cardinality`): To create variants of the data set with different cardinality of categorical features, we merge categories into new, larger categories. For instance, for feature 35 (country of birth self PENATVTY) the country of origin is first merged by combining all countries from a specific continent. This reduces the cardinality, with 6 remaining categories (medium cardinality). For a further reduction (low cardinality) to three categories, the data is merged into the categories unknown, US, and abroad. Similar changes to other features are documented in the source code. For our experiments, this pre-processing step results into data sets with the levels of cardinality: low (up to 15 categories), medium (up to 24 categories) and high (up to 52 categories).

- Number of numerical features (`nnumericals`): To change the number of features, individual features are included or removed from the data set. This is done separately for categorical and numerical features and results into four levels for `nnumericals` (low: 0, medium: 4, high: 6, complete: 7).

- Number of categorical features (`nfactors`): Correspondingly, we receive four levels for `nfactors` (low: 0, medium: 8, high: 16, complete: 33). Note, that these numbers become somewhat reduced, if cardinality is low (low: 0, medium: 7, high: 13, complete: 27). The reason is, that some features might become redundant when merging categories.

- Number of observations (m): To vary m, observations are randomly sampled from the data set. We test five levels on a logarithmic scale from 10^4 to 10^5: 10 000, 17 783, 31 623, 56 234, and 100 000. In addition, we made a separate test with the complete data set, i.e., 299 285 observations.

For classification, we performed for each tuner and each model the experiments summarized in Table 11. A reduced number of experiments were performed for regression, see Table 12.

To judge the impact of the creating random subsets of the data (i.e., to reduce m), and to consider the test / train split, three data sets are generated for each configuration. All experiments are repeated for each of those three configurations.

6.2.2 Model Configuration

The experiments include the models k-NN, DT, EN, RF, eXtreme Gradient Boosting (XGBoost), and SVM. Their respective parameters are listed in Table 13).

Table 11: Experiments for classification: Investigated combinations of the number of categorical features (nfactors), numerical features(nnumericals), cardinality, and m. An empty field for cardinality occurs for low nfactors. Here, no categorical features are present, so the number of categories is irrelevant.

nfactors	nnumericals	cardinality	m
high	low	low	$10^4, 10^{4.25}, ..., 10^5$
medium	medium	low	$10^4, 10^{4.25}, ..., 10^5$
low	high		$10^4, 10^{4.25}, ..., 10^5$
high	high	low	$10^4, 10^{4.25}, ..., 10^5$
high	low	medium	$10^4, 10^{4.25}, ..., 10^5$
medium	medium	medium	$10^4, 10^{4.25}, ..., 10^5$
high	high	medium	$10^4, 10^{4.25}, ..., 10^5$
high	low	high	$10^4, 10^{4.25}, ..., 10^5$
medium	medium	high	$10^4, 10^{4.25}, ..., 10^5$
high	high	high	$10^4, 10^{4.25}, ..., 10^5$
complete data set			299285

Table 12: Experiments for regression: Investigated combinations of the number of categorical features (nfactors), numerical features(nnumericals), cardinality, and m. An empty field for cardinality occurs for low nfactors. Here, no categorical features are present, so the number of categories is irrelevant.

nfactors	nnumericals	cardinality	m
low	high		$10^4, 10^{4.25}, ..., 10^5$
high	high	low	$10^4, 10^{4.25}, ..., 10^5$
high	high	high	$10^4, 10^{4.25}, ..., 10^5$

For EN, `lambda` is not optimized by the tuner. Rather, the `glmnet` implementation itself tunes that parameter. Here, a sequence of different `lambda` values is tested [Hastie and Qian, 2016, Friedman et al., 2020].

For SVM, the choice of the kernel (`kernel`) is limited to `radial` and `sigmoid`, since we experienced surprisingly large runtimes for `linear` and `polynomial` in a set of preliminary experiments. Hence, `degree` is also excluded, as it is only relevant for the `polynomial` kernel. Due to experiment runtime, we also did not perform experiments with SVM and k-NN on data sets with $m \geq 10^5$ observations. These would require using model variants that are able to deal with huge data sets (e.g., some sparse SVM type).

Table 9 lists hyperparameters that are actually tuned in the experiments, including data type, bounds and employed transformations. Here, we mostly follow the bounds and transformations as used by Probst et al. [2019a]. Fundamentally, these are not general suggestions. Rather, reasonable bounds will usually require some considerations with respect to data understanding, modeling / analysis, and computational resources. Bounds on values which affect run time should be chosen so that experiments are still possible within a reasonable time frame. Similar consideration can apply to memory requirements Where increasing / decreasing parameters may lead to increasing / decreasing sensitivity of the model, a suitable transformation (e.g., log-transformation) should be applied.

Most other configurations of the investigated models remain at default values. The only exceptions are:

- `ranger`: For the sake of comparability with other models, model training and evaluation are performed in a single thread, without parallelization (`num.threads = 1`).

- `xgboost`: Similarly to`ranger`,we set `nthread=1`. For regression, the evaluation metric is set to the root-mean-square error (`eval_metric="rmse"`). For classification, log-loss is chosen (`eval_metric="logloss"`).

6.2.3 Further Configurations

Overall, the implementation of our experiments is as follows: The R package `mlr` is used as a uniform interface to the machine learning models. The R package `SPOT` is used to perform the actual tuning (optimization). All additional code is provided together with this paper. We also provide examples for creating visualizations of the tuning results.

The following configurations are of importance:

- General configuration:

- Model quality measures: The evaluation of hyperparameter values requires a measure of quality, which determines how well the resulting models perform.
  * Regression: For the regression experiments, we use Root Mean Squared Error (RMSE): RMSE= $\sqrt{(1/m)\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$. Here, $\hat{y}_i$ is the predicted value of the corresponding model for the i-th observation in the data set, and y_i is the actually observed value.
  * Classification: For the classification experiments, we use mean Mis-Classification Error (MCE): MCE= $(1/m)\sum_{i=1}^{m}\mathbf{I}(y_i \neq \hat{y}_i)$.
  The tuner uses these two measures to determine better hyperparameter values.

- Further measures: In addition to RMSE and MCE, we also record runtime (Overall runtime of a model evaluation, runtime for prediction, runtime for training). Additional performance measure from the `mlr` package could be easily integrated when necessary[10].

- Runtime budget: To mirror a realistic use case, we specify a fixed runtime budget for the tuner. This limits how long the tuner may take for finding potentially optimal hyperparameter values. We set a budget of 5 hours for SVM, k-NN, RF, and XGBoost. Since EN and DT are much faster to evaluate and less complex, they receive a considerably lower budget (EN: 1 hour, DT: 5 minutes).

- Timeout: For a majority of the models, the runtime of a single evaluation (training + prediction) is hard to predict and may easily become excessive if parameters are chosen poorly. In extreme cases, the runtime of a single evaluation may become so large that it takes up the majority of tuner runtime, or even more. In such a case, there would be insufficient time to test different hyperparameter values. To prevent this, we specify a limit for the runtime of a single evaluation, which we call timeout. If the timeout is exceeded by the model, the evaluation will be aborted. During the experiments we set the timeout to a twentieth of the tuner's overall runtime budget. Exceptions are the experiments with DT (`rpart`): Since `rpart` evaluates extremely quickly, (in our experiments: usually much less than a second) and has a correspondingly reduced runtime budget (5 minutes), the timeout is not required. In fact, using the timeout would add considerable overhead to the evaluation time in this case.

- Returned quality values in case of errors: If the evaluation gets aborted (e.g., due to timeout or in case of some numerical instability), we still require a quality value to be returned to the tuner, so that the search can continue. This return value should be chosen, so that, e.g., additional evaluations with high runtimes are avoided. At the same time, the value should be on a similar scale as the actual quality measure, to avoid a deterioration of the underlying surrogate model. To achieve this, we return the following values when an evaluation aborts.
  * Regression: Model quality for simply predicting the mean of the training data.
  * Classification: Model quality for simply predicting the mode of the training data.

- Benchmark of runtime: The experiments are not run on entirely identical hardware, but on somewhat diverse nodes of a cluster. Hence, we have to consider that single nodes are faster or slower than others (i.e., a tuning run on one node may perform more model evaluations than on another node, simply because of having more computational power). To preserve some sort of comparability, we compute a corrective factor that will be multiplied with the runtime budget. Before each tuning run, we compute a short performance benchmark. The time measured for that benchmark will be divided by the respective value measured at a reference node, to determine the runtime multiplicator. We use the `benchmark_std` function from the `benchmarkme` R package [Gillespie, 2021].

- Configuration of SPO Toolbox (SPOT):

  - Design: The initial design is created by Latin Hypercube Sampling [Leary et al., 2003]. The size of that design (number of sampled configurations of hyperparameters) corresponds to $5d$. Here, d is the number of hyperparameters. An exponential dependence such as d^5 would be possible in principle, but might lead to an excessive size of the initial design (e.g., just $d = 4$ parameters yield $4^5 = 1024$ initial evaluations).

  - Surrogate model: SPOT can use arbitrary regression models as surrogates, e.g., random forest or Gaussian process models (Kriging). We chose a Gaussian process model, since it performs well in our experience, and the Gaussian process implementation can work well with discrete as well as continuous hyperparameters. Random forest is less suited for continuous parameters, as it has to approximate them in a step-wise constant manner.

    The Gaussian process model is fitted via Maximum Likelihood Estimation. That is, using numerical optimization, the model parameters are determined, so that the observed data are judged to have maximum likelihood according to the model. We use Differential Evolution [Storn and Price, 1997] for this purpose.

---

[10]A complete list of measures in `mlr` can be found at `https://mlr.mlr-org.com/articles/tutorial/measures.html`

In one central issue, we deviate from more common configurations of surrogate model based optimization algorithms: For the determination of the next candidate solution to be evaluated, we directly use the predicted value of the Gaussian process model, instead of the so-called expected improvement. Our reason is, that the expected improvement may yield worse results if the number of evaluations is low, or the dimensionality rather high [Rehbach et al., 2020, Wessing and Preuss, 2017]. With the strictly limited runtime budget, our experience is that the predicted value is the better choice. A similar observation is made by De Ath et al. [2019].

– Optimizing the predicted value: To find the next candidate solution, the predicted value of the surrogate model is optimized via Differential Evolution [Storn and Price, 1997][11]. Thus, SPOT searches for the hyperparameter configuration that is predicted to result into the best possible model quality.. We allow for $200d$ surrogate evaluations in each iteration of SPOT.

– All other parameters of SPOT remain at default values.

• Configuration of Random Search (RS):

– With RS, hyperparameter values will be sample uniformly from the search space.

– All related configurations (timeout, runtime budget, etc.) correspond to those of SPOT.

• Defaults:

– As a comparison basis, we perform an additional experiment for each model, where all hyperparameter values remain at the models default settings.

– However, in those cases we do *not* set a timeout for evaluation. Since no search takes place, the overall runtime for default values is anyways considerable lower than the runtime of SPOT or RS.

– All other settings correspond to those of SPOT and RS.

• Replications: To roughly estimate the variance of results, we repeat all experiments (each tuner for each model on each data set) three times.

6.3 Results

In this section, we provide an exploratory overview of the results of the experiments. A detailed discussion of the results in terms of the research questions defined in Sec. 6.1 follows in Sec. 6.4.

To get an impression of the overall results, we show exemplary boxplots[12]. Since different results are achieved depending on the data set and optimized model, a preprocessing step is performed first: For each drawn data set and each model the mean value of all observed results (model goodness of the best solutions found) is determined. This mean is then subtracted from each individual observed value of the corresponding group. Subsequently, these subtracted individual values are examined. This allows a better visualization of the difference between the tuners without compromising interpretability. The resulting values are no longer on the original scale of the model quality, but the units remain unchanged. Thus, differences on this scale can still be interpreted well.

For the classification experiments, Figure 10 shows the results for a case with many features and observations (nfactors, nnumericals and cardinality are all set to high and $m = 10^5$). The figure first shows, that both tuners (RS, SPOT) achieve a significant improvement over default values. The value of this improvement is in the range of about 1% MCE. In almost all cases, the tuners show a smaller dispersion of the quality values than a model with default values. Except for the tuning of `rpart`, the results obtained by SPOT are better than those of RS.

Likewise for classification, Figure 11 shows the case with $m = 10^4$, without categorical features (nfactors=low) and with maximum number of numerical features (nnumericals=high). Here the data set contains much less information, since a large part of the features is missing and only few observations are available. Still, both tuners are significantly better than using default values. In this case, it is mostly not clear which of the two tuners provides better results. For `rpart` (DT) and SVM, RS seems to work better.

In the same form as for classification, figures 12 and 13 show results for regression models. Unlike for classification, the results here are somewhat more diverse. In particular, `glmnet` shows a small difference between the tuners and default values. There are also differences of several orders of magnitude between the individual models (e.g. RF and SVM). For example, for RF the differences between tuners and default values are about 0.02 years (the target size is the age in years). As shown in figures 12 and 13, the interquartile range is about 0.01 years.

For XGBoost, on the other hand, there is a difference of about 20 years between tuners and default values. Here, the default values seem to be particularly poorly suited.

---

[11]We could use other global optimization algorithms as well. Even random search would be a feasible strategy.

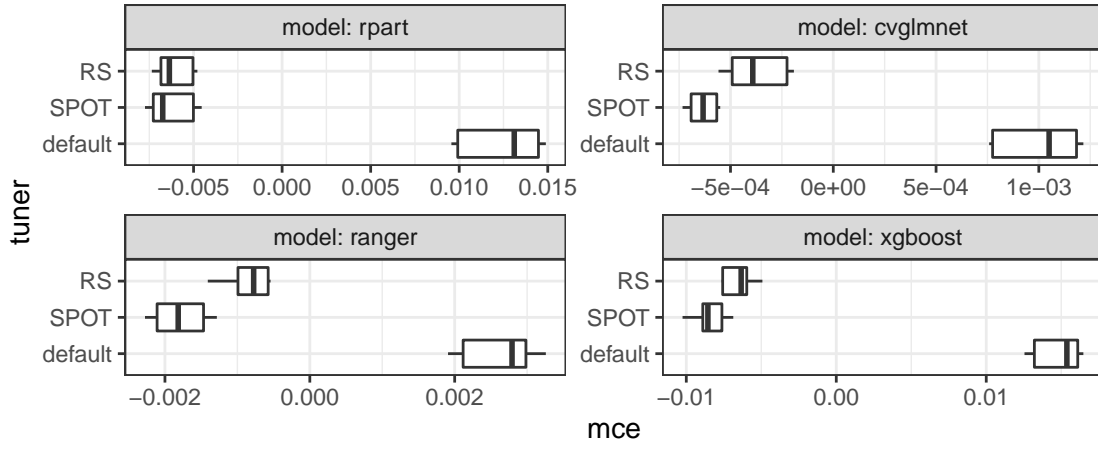[12]Corresponding boxplots for all experiments can be found in the appendix of this document.

Figure 10: Boxplot comparing tuners for different classification models, with $m = 10^5$, nfactors=high, nnumericals=high, cardinality=high. For this value of m, no experiments with k-NN and SVM were performed. The presented quality measure (mce) shows values after preprocessing, where the mean value for each problem instance is subtracted.
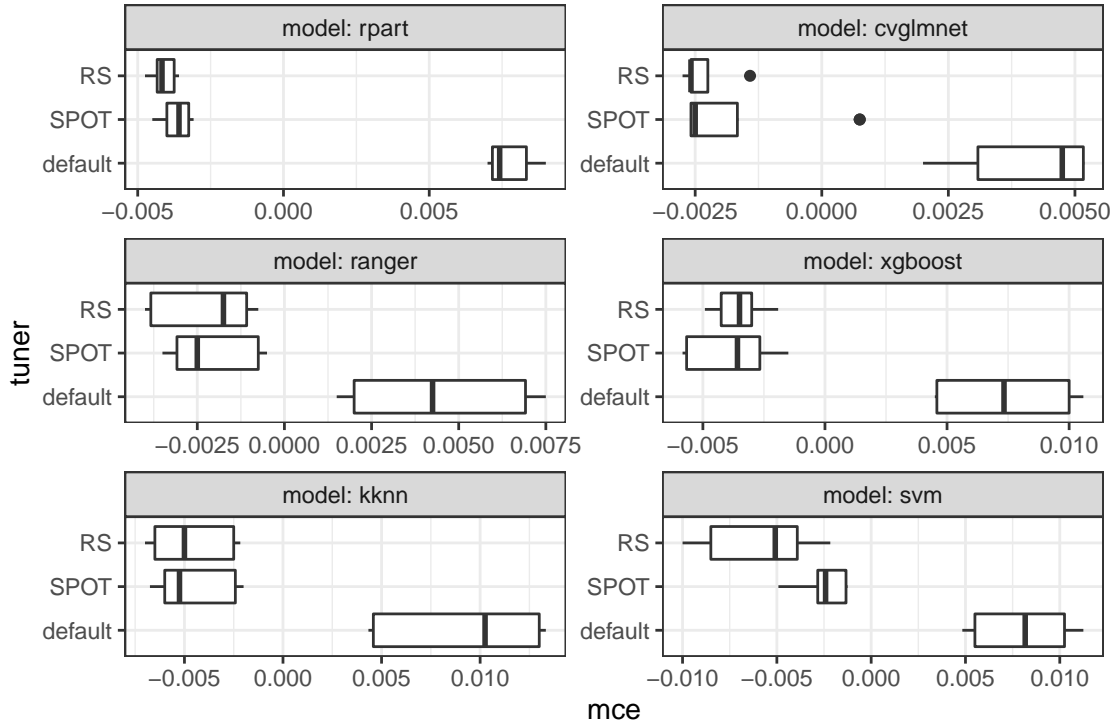

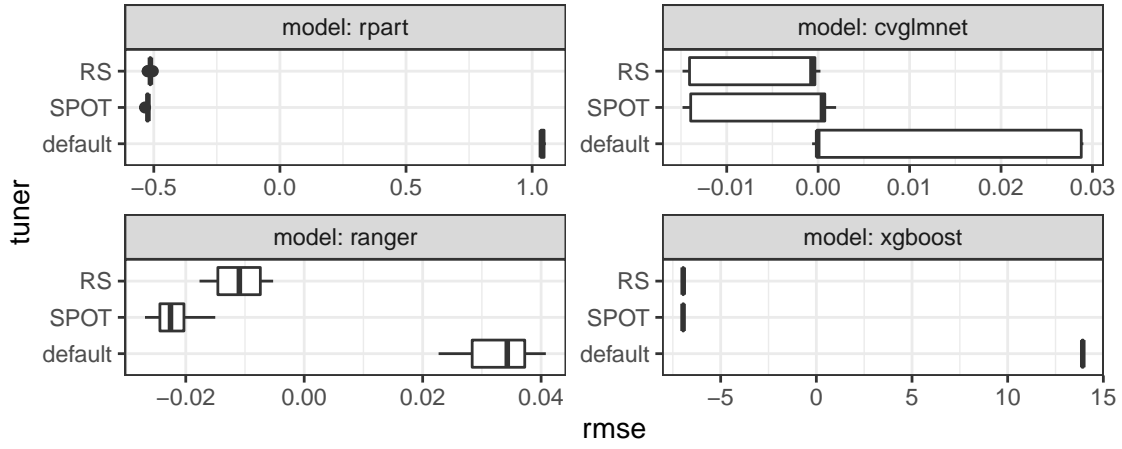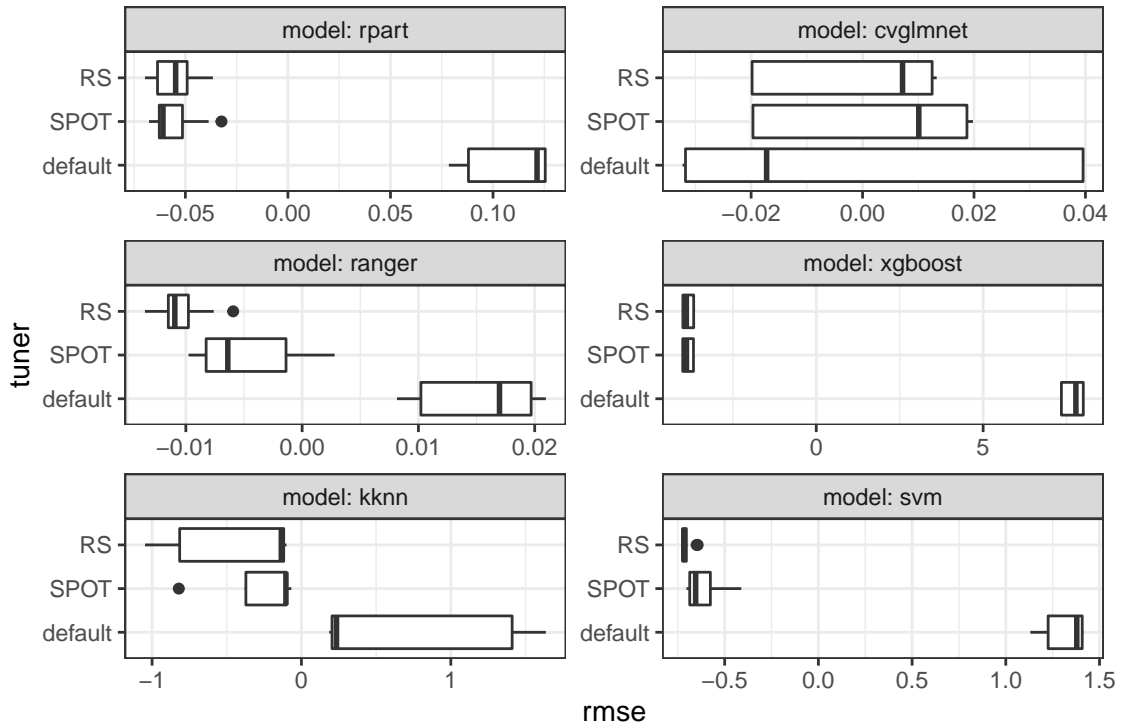**Figure 11:** Boxplot comparing tuners for different classification models, with $m = 10^4$, nfactors=low, nnumericals=high. The presented quality measure (mce) shows values after preprocessing, where the mean value for each problem instance is subtracted.

Figure 12: Boxplot comparing tuners for different regression models, with $m = 10^5$, nfactors=high, nnumericals=high, cardinality=high. For this value of m no experiments with k-NN and SVM were performed. The presented quality measure (mce) shows values after preprocessing, where the mean value for each problem instance is subtracted.

Figure 13: Boxplot comparing tuners for different regression models, with $m = 10^4$, nfactors=low, nnumericals=high. The presented quality measure (mce) shows values after preprocessing, where the mean value for each problem instance is subtracted.

6.4 Discussion

6.4.1 Rank-Analysis

We analyze the results of the experiments using rankings of the result values instead of the raw results. Rankings are scale invariant, so aggregating the results from different problem instances (resulting from data set and model choice) with different scales for the result values is possible. This is important because for the analysis of individual study items we have to aggregate over very diverse problem instances.

To aggregate rankings (also called consensus rankings), we follow the so-called "optimal ranking" approach of Kemeny [1959]. Here, the consensus ranking is determined such that the mean distance between the consensus ranking and observed ranking is minimal. The distance measure used for this purpose is Kendall's tau [Kendall, 1938], which counts the number of pairwise comparisons in which two rankings contradict each other. The ranking determined using this approach can be interpreted as the "median" of the individual rankings.

This procedure has the following advantages [Hornik and Meyer, 2007, Mersmann et al., 2010]:

- Scale invariant
- Invariant to irrelevant alternatives.
- Aggregation of large sets of comparisons, over arbitrary factors.
- Easy/intuitive to interpret results & visualizable
- Generates relevant information: selection of the best one
- Additional weights for preferences can be inserted
- Fast evaluation
- Non-parametric method, no distribution assumptions
- Visualization of clusters over distance is possible, identification of problem classes with similar algorithm behavior.

However, in addition to these advantages, there are also disadvantages:

- Estimating uncertainty in the ranking is difficult
- The ranking does not have to induce a strict ordering, ties are possible

We generate the consensus ranking by respectively combining rankings of tuners (SPOT, RS, default) of individual experiments. We always summarize the rankings of 9 experiments (3 repetitions of the tuner runs on each of 3 randomly drawn data sets). Then, to aggregate across different variables related to the study subjects (e. B. m, nfactors), we count how often the tuners reach a certain consensus rank.

This count is divided by the total number of experiments (or number of rankings) on the corresponding problem instances. Thus, for each tuner, the number of times a particular Consensus Rank is achieved is obtained.

Simplified example:

- For case 2 (i.e., for a fixed choice of nnumericals, nfactors, cardinality, model, m, target variable), the comparison of SPOT, RS, and default methods resulted in the ranks

$$\{1, 3, 2\} \quad \{1, 2, 3\} \quad \{2, 1, 3\}.$$

  The consensus ranking for this case is $\{1, 2, 3\}$.

- For case 2 (i.e., for *another* fixed choice of nnumericals, nfactors, cardinality, model, m, target variable), the comparison of SPOT, RS, and default methods resulted in the ranks

$$\{3, 2, 1\} \quad \{1, 2, 3\} \quad \{2, 1, 3\}.$$

  The consensus ranking for this case is $\{2, 1, 3\}$.

- When both experiments are combined for an analysis, the frequencies for the obtained rankings are as follows:
    - SPOT: rank 1 with 50%, rank 2 with 50%, rank 3 with 0%.
    - RS: rank 1 with 50%, rank 2 with 50%, rank 3 with 0%.
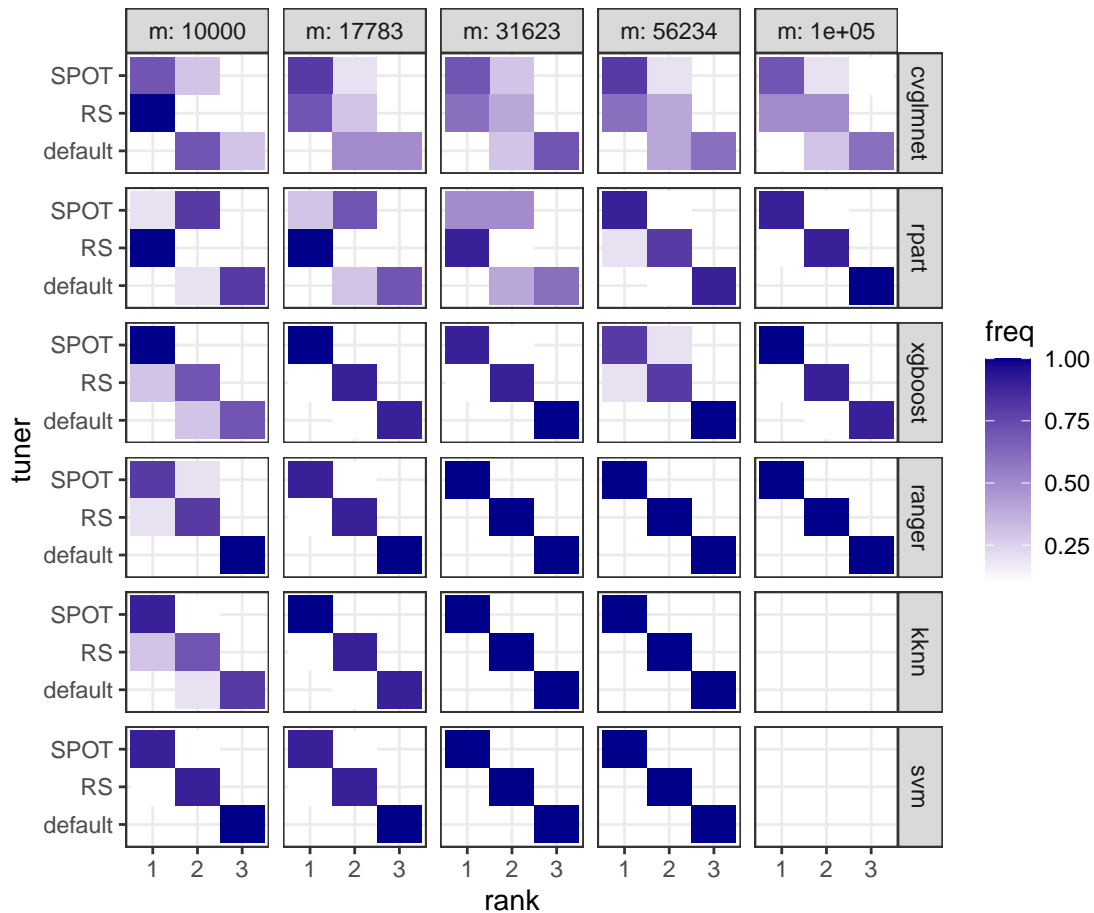   - Default: rank 1 with 0%, rank 2 with 0%, rank 3 with 100%.

Figure 14: Rank of tuners depending on number of observations (m) and model, for classification.

6.4.2 Rank-Analysis: Classification

Based on the analysis method described in Sec. 6.4.1, figure 14 shows the relationship between the tuners, the number of observations m, and the optimized models. It shows that mostly SPOT and RS beat the default setting and SPOT also usually performs better than RS. However, some cases deviate from this. Especially, glmnet (EN) and rpart (DT) seem to profit less from tuning: here the distinction between the ranks of tuners is more difficult. In addition, when the number of observations is small, there tends to be a greater uncertainty in the results. These results can be partly explained by the required running time of the models. With a smaller number of observations, the runtime of the individual models decreases, and glmnet and rpart are the models with the lowest runtime (in the range of a few seconds, or below one second in the case of rpart). If the evaluation of the models itself takes hardly any time, it is advantageous for RS that the runtime overhead is low. SPOT, on the other hand, requires a larger overhead (for the surrogate model and the corresponding search for new parameter configurations).

Figure 15 shows the corresponding summary of the results depending on the number of categorical features (nfactors). Again, SPOT usually performs best, followed by RS. There is a tendency for the greatest uncertainty to be found in case of a few categorical features. Two explanations are possible: On the one hand, the reduction of features also means a reduction of the required running time. On the other hand, the difficulty of the modeling increases, since with fewer features less information is available to separate the classes.

The corresponding results for the number of numerical features can be found in Figure 16. There are hardly any differences, the number of numerical features seems to have little influence. It should be noted that the data set contains fewer numerical than categorical features anyway.
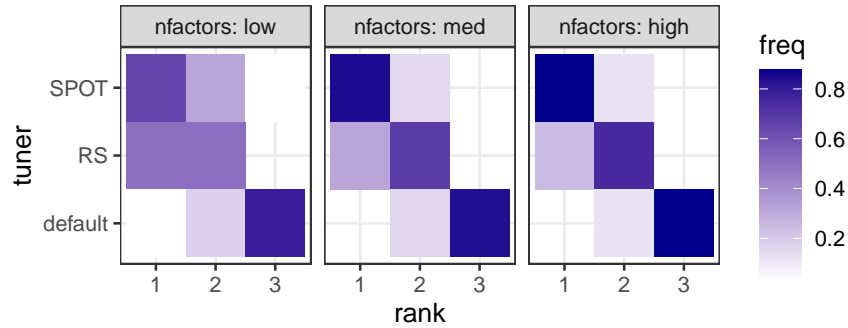
Figure 15: Rank of tuners as a function of number of categorical features (nfactors) and model, for classification.
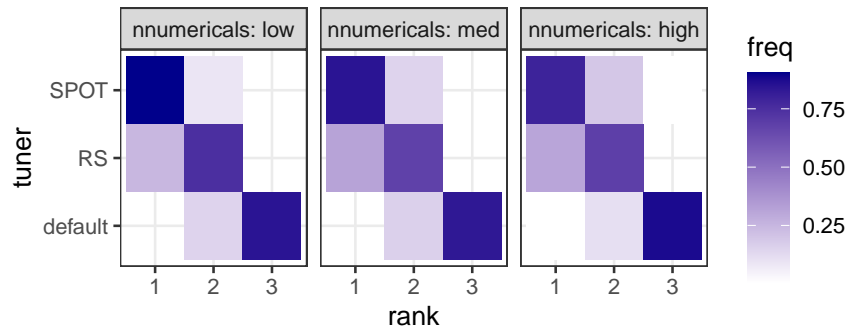

Figure 16: Rank of tuners depending on number of numerical features (nnumericals) and model, for classification.

The cardinality of the categorical features also has little influence, see Figure 17. However, there is a slight tendency: at higher cardinality, the distinction between the first rank (SPOT) and second rank (RS) is clearer. This can be explained (similarly to nfactors) by the higher information content of the data set.

Finally, Figure 18 shows the result of the tuners on the unmodified, complete data set. For each case, SPOT gets rank 1 and RS rank 2. This result is in line with the trends described above, since the complete data set contains the most information and also leads to the largest runtimes (for model evaluations).
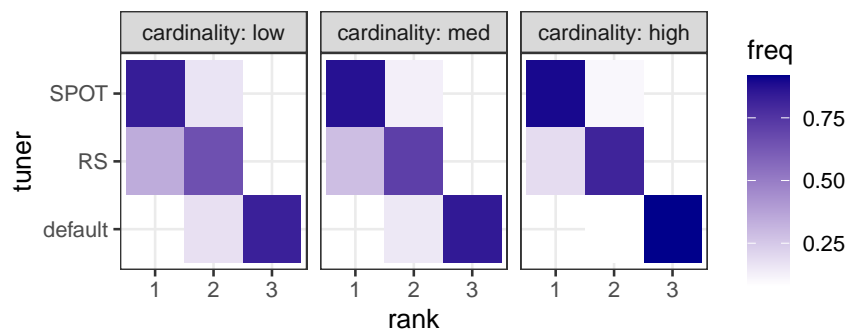
Figure 17: Rank of tuners as a function of cardinality of categorical features and model for classification. Note: This figure does not include the cases where the data set no longer contains categorical features (cardinality cannot be determined).
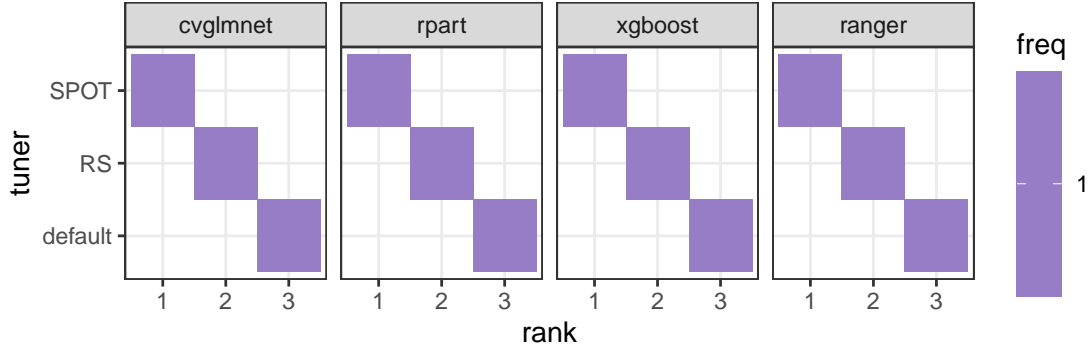
Figure 18: Rank of tuners depending on the model for classification on complete data set. Due to the increased runtime, `kknn` (k-NN) and `e1071` (SVM) are not included.

6.4.3 Rank-Analysis: Regression

In addition to the classification experiments, a smaller set of experiments with regression as an objective were also conducted. Figure 19 shows the results for this case separately for each optimized model. Here, too, SPOT is usually ranked 1. Unlike classification, however, there is more uncertainty.

For `glmnet` (EN), default values occasionally even achieve rank 1. It seems that linear modeling with `glmnet` is unsuitable for regression with the present data set, so that the tuners can hardly achieve differences to the default values. This behavior was already indicated in Figure 13.

In the case of SVM, RS is more often ranked 1 than SPOT, the reason for this is not clear. Possible cause is a lower influence of the hyperparameters on the model goodness of fit of SVM for regression (compared to classification). However, it should also be taken into account that for regression a smaller number of experiments is used for the evaluation.

The dependence on the number of observations m is shown in Figure 20. The results for regression with respect to m are again largely consistent with those for classification. With increasing number of observations, SPOT is more clearly ahead of RS.

The correlation with categorical features (number, cardinality) is also consistent with the classification results, see Figure 21. With larger number of features and larger cardinality, a clearer separation between the tuners is observed.

6.4.4 Problem Analysis: Difficulty

In the context of this study, an interesting question arose as to how the difficulty of the modeling problem is related to the results of the tuning procedures. We investigate this on the basis of the data obtained from the experiments.

In general, there are many measures of the difficulty of modeling problems in the literature. An overview is given by Lorena et al. [2019]. Of these measures, the overlap volume (F2, see [Lorena et al., 2019]) is of interest, since it is easily interpretable and not specific to a particular model. It is first determined for each feature individually how far
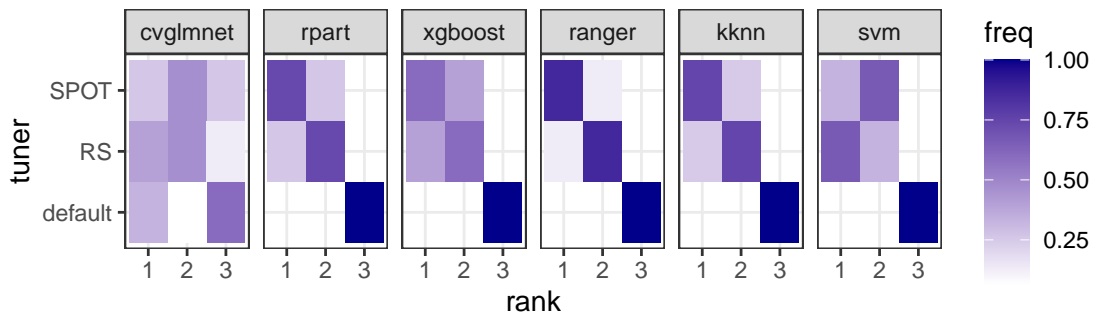

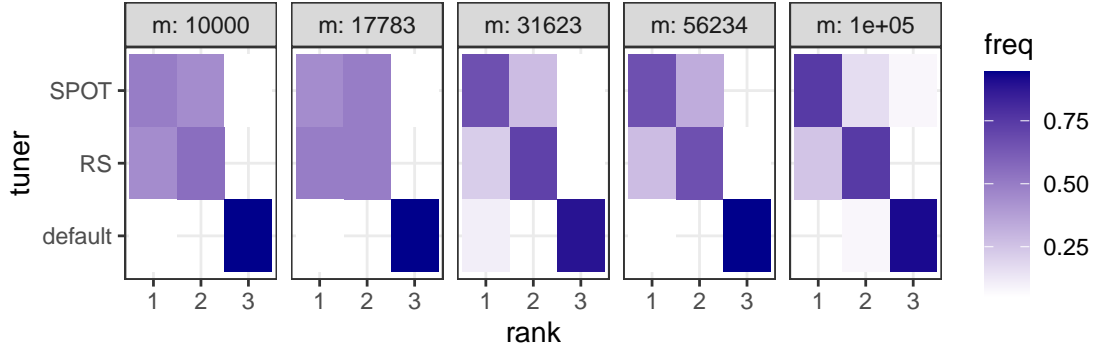**Figure 19:** Rank of tuners depending on model for regression.

Figure 20: Rank of tuners depending on number of observations (m) for regression.
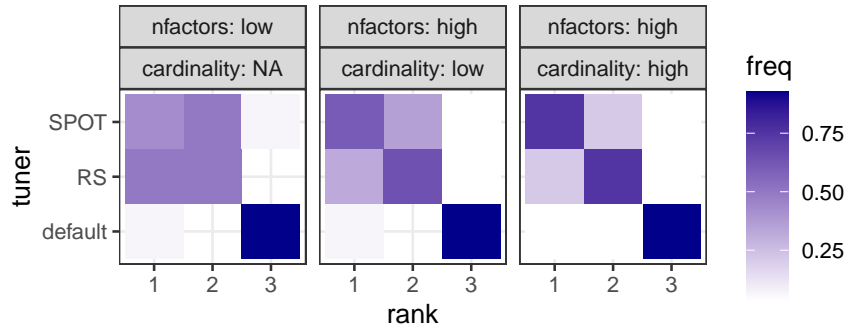

Figure 21: Rank of tuners depending on cardinality and number of categorical features (nfactors) for regression.

the value ranges of two classes overlap. This is called the *overlap* of the feature. The *overlap volume* of the data set is then the product of the individual overlap values. However, this measure is unsuitable for categorical features (even after dummy coding). Furthermore, outliers are very problematic for this measure.

We therefore use a slight modification, by calculating the proportion of sample values for each feature, which could occur in both classes (i.e. for which a swap of the classes based on the feature value is possible). As an example, this is illustrated for a numeric and a categorical feature in Figure 22 for a numerical and a categorical feature. For the overall data set, the individual overlap values of each feature are multiplied. Subsequently, we refer to this measure as *sample overlap*.

Figure 23 shows the dependence of the sample overlap on our data properties (m, nfactors, nnumericals, cardinality). Our data sets can be grouped into 4 difficulty levels based on these values of sample overlap:

1. Sample overlap ≈ 0.39: nfactors=high and cardinality=high

2. Sample overlap ≈ 0.54: nfactors=high and cardinality=med

3. Sample overlap ≈ 0.76: all others

4. Sample overlap ≈ 1.00: nfactors=low

Here, 4 corresponds to the highest level of difficulty. For the range relevant in the experiments, there is almost no change depending on m or nnumericals. For nfactors and cardinality a strong correlation can be seen.

Based on the 4 difficulty levels, the ranks already determined in previous sections can be re-ranked. The result is summarized in Figure 24. It turns out that as the difficulty of the problem increases, the rank differences are less robust. That is, with larger sample overlap it is harder to estimate which tuner works best. This is plausible with respect to the theoretical extreme case: A maximally difficult data set cannot be learned by any model, since the features no longer contain any information about the classes. In this case, the hyperparameters no longer have any influence and the tuners cannot improve.
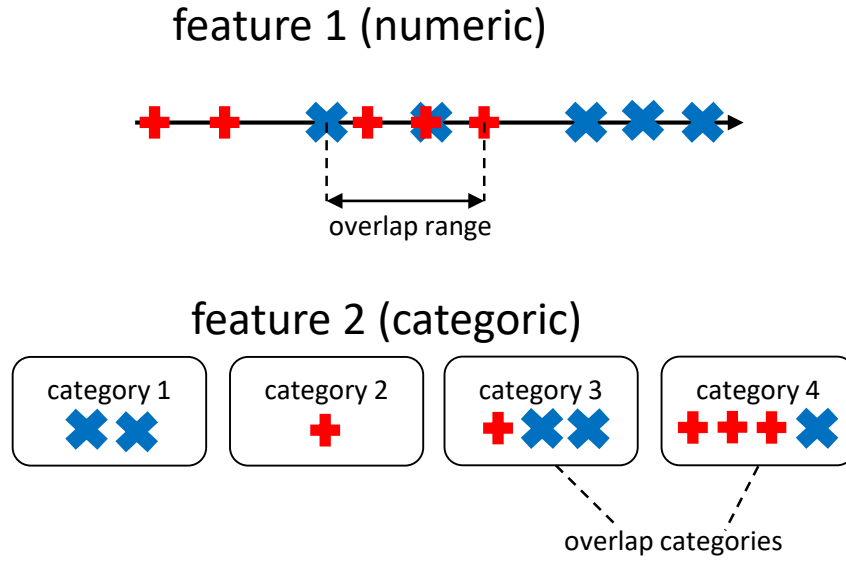
Figure 22: Example of sample overlap for two features of a classification problem with two classes. The crosses show samples from the data set. All samples with class 1 are red (+). All samples with class 2 are blue (x). For feature 1, the overlap is 50% (number of samples in the overlap area divided by the total number of samples). For feature 2, the overlap is 70% (number of samples in the overlap categories divided by total number of samples). For both characteristics together, the sample overlap is $0.5 \times 0.7 = 0.35$.
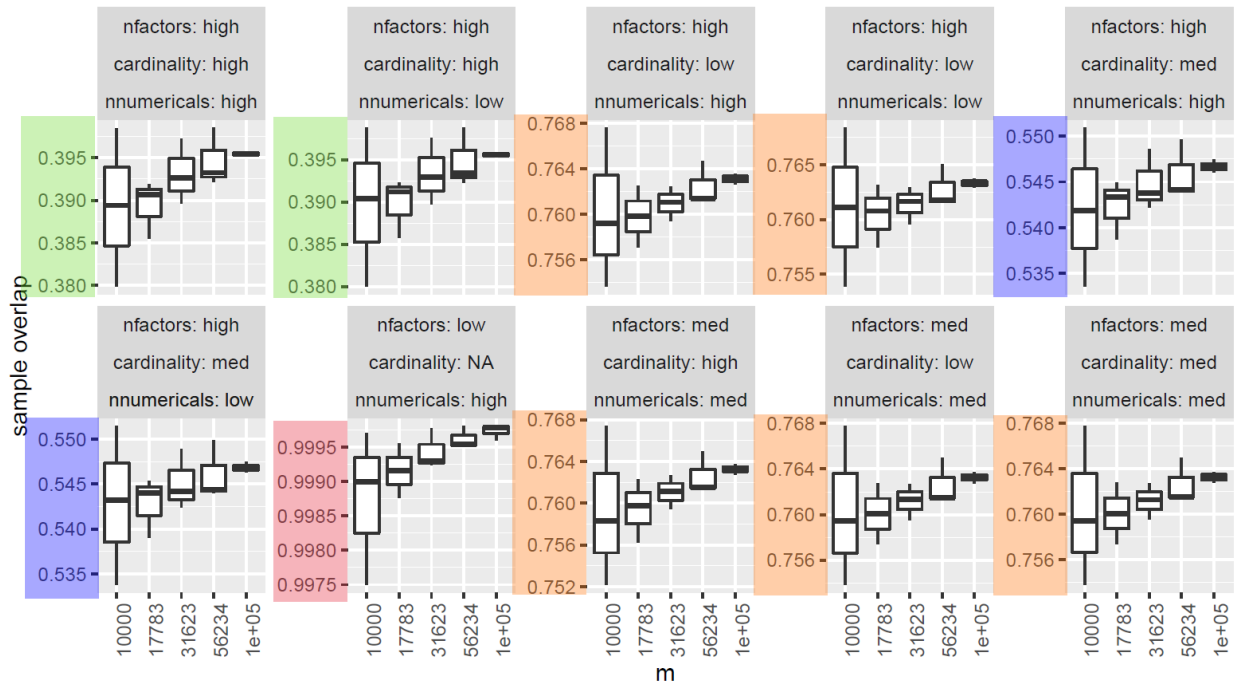

Figure 23: Sample overlap depending on the properties of the data set varied in the experiments. The sample overlap is used as a measure of problem problem difficulty and leads to the definition of four difficulty levels (marked in color).
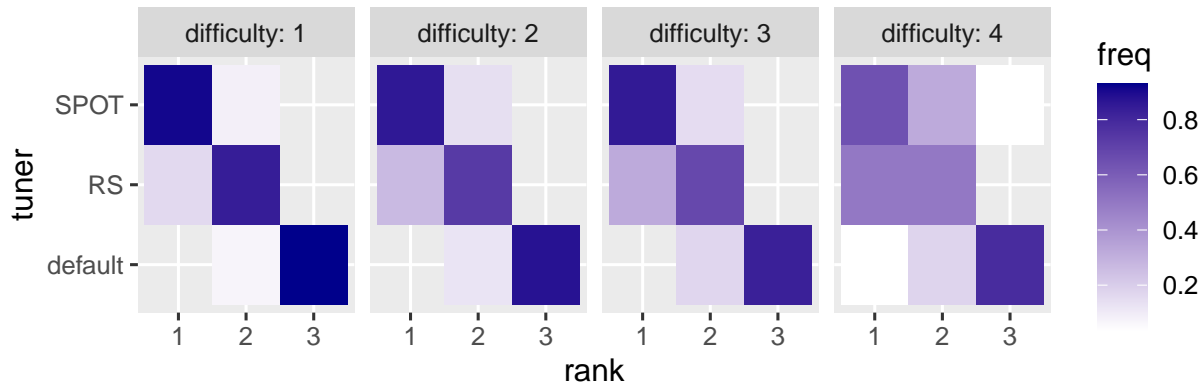
Figure 24: Frequency of the achieved ranks of the tuners depending on the difficulty level.

7   Conclusion

7.1   Summary

Finally, we discuss the hypotheses from Sec. 6.1.

(H-1) *Tuning is necessary:* In almost all cases, significantly worse results are obtained with default settings. In addition, the variance of the results decreases when tuners are used (see figures 10 - 13).

(H-2) *Data:* Differences between tuners become apparent for data sets with a high information content and for larger data volumes. If the number of levels or features (categorical and numerical) decreases, it can be observed that the differences between the tuners become smaller.

It has been confirmed that as the number of features and observations increases, the running time increases. As the mean runtime of the models increases (i.e., for more complex models or larger data sets), SPOT performs increasingly better than RS, as the ratio of overhead to evaluation time decreases more for SPOT.

(H-3) *Target variable:* The execution of the tuning is not affected by the change of the target variable. A peculiarity that requires further investigation occurred when tuning SVM for regression: RS seems to perform better than SPOT. We recommend using a larger database for this case.

(H-4) *Model:* The choice of tuning method is not fundamentally influenced by the models. Models that can be evaluated very quickly (e.g. `rpart`) benefit from the larger number of evaluations. This is due to the fact that time was chosen as a termination criterion.

(H-5) *Benchmark:* As described in Sec. 6.4.1, analysis methods based on consensus ranking can be used to evaluate the suitability of tuners in a simple and statistically valid way.

Overall, the result of the experimental investigation underlines, that tuning of machine learning models is essential to produce good configurations of the hyperparameters and thus models of high quality.

It is especially important to use tuners at all, since the biggest differences are usually observed in comparison to default settings. However, there are also differences between tuners, which in our investigation were mostly for the use of SPOT, i.e. a model-based method. For the prospective focus on tuning with relatively large data sets (large m and n), this is all the more evident, since the resulting high model runtime favors the use of model-based methods.

Since the number of hyperparameters is manageable ($<< 100$), the addition of a few more parameters does not significantly increase the complexity of the tuning. For both Random Search[Bergstra and Bengio, 2012] and SPOT, the addition of a few parameters is harmless, even if they have only a small effect on the model quality.

Rather, it results in a benefit:

- Possible software bugs can be detected,
- seemingly unimportant parameters are detected as important,
- unknown interactions can be uncovered.

The surrogate model allows to learn the influence of the hyperparameters on the model values. Unimportant parameters are thus automatically weighted less in the search.

Finally, we recommend for the selection of the tuner:

- Random Search can be used when either very little time (order of magnitude: time is sufficient for a single-digit number of sequential evaluations) but a lot of parallel computing capacity is available or when models can be evaluated extremely fast (seconds range).

- If time-consuming computations with complex data and models need to be performed in more time (with relatively less parallel computing capacity), then we recommend model-based tuning with SPO.

- In exceptional cases, which deviate from the objective considered here for complex data and models (extremely large amount of computing time available, average evaluation times for the models), the use of surrogate model-free tuning methods can also be considered.

7.2 Special Case: Monotonous Parameters

Table 13: Global hyperparameter overview. The column "Quality" shows all parameter, where a monotonous relationship between parameter values and model quality is to be expected. (↑↑: quality increases if parameter value increases, ↑↓: quality decreases if parameter value increases). Correspondingly, the column "run time" shows the same information for the relationship of parameter values and run time.

Model	Hyperparameter	Quality	Run time
k-NN	k		↑↑
	p		
EN	alpha		
	lambda		
	thresh	↑↓	↑↓
DT	minsplit		↑↓
	minbucket		↑↓
	cp		↑↓
	maxdepth		↑↑
RF	num.trees	↑↑	↑↑
	mtry		↑↑
	sample.fraction		↑↑
	replace		
	respect.unordered.factors		
xgBoost	eta		
	nrounds		↑↑
	lambda		
	alpha		
	subsample		↑↑
	colsample_bytree		↑↑
	gamma		↑↓
	max_depth		↑↑
	min_child_weight		↑↓
SVM	kernel		
	degree		
	gamma		
	coef0		
	cost		
	epsilon		

A special case are hyperparameters with monotonous effect on the quality and runtime (and/or memory requirements) of the tuned model. In our survey (see Table 13), two examples are included: `num.trees` (RF), `thresh` (EN). Due to the monotonicity properties, treating these parameters differently is an likely consideration. In the following, we focus the discussion on `num.trees` as an example, since this parameter is frequently discussed in literature and online communities [Probst and Boulesteix, 2018].

It is known from the literature that larger values of `num.trees` generally lead to better models. As the size increases, a saturation sets in, leading to progressively lower quality gains. It should be noted that this is not necessarily true for

every quality measure. Probst and Boulesteix [2018], for example, show that this relation holds for log-loss and Brier score, but not for Area Under the receiver operating characteristic Curve (AUC).

Because of this relationship, Probst and Boulesteix recommend, that `num.trees` should not be optimized. Instead, they recommend setting the parameter to a "computationally feasible large number" [Probst and Boulesteix, 2018]. For certain applications, especially for relatively small or medium sized data sets, we support this assessment. However, at least in perspective, the present report considers tuning hyperparameters for very large data sets (many observations and/or many features). For this use case, we do not share the recommendation of Probst and Boulesteix, because the required runtime of the model plays an increasingly important role and is not explicitly considered in the recommendation. In this case, a 'computationally feasible large number' is not trivial to determine.

In total, we consider five alternatives:

1. Set manually: The parameter is set to the largest possible value that is still feasible with the available computing resources. Risks:
   (a) Single evaluations during tuning waste time unnecessarily.
   (b) Interactions with parameters (e.g. `mtry`) are not considered.
   (c) The value may be unnecessarily large (from a model quality point of view).
   (d) The determination of this value can be difficult, it requires detailed knowledge regarding: size of the data set, efficiency of the model implementation, available resources (memory / computer cores / time).

2. Manual adjustment of the tuning: After a preliminary examination (as represented e.g. by the inital design step of SPOT) a user intervention takes place. Based on the preliminary investigation, a value that seems reasonable is chosen by the user and is not changed in the further course of the tuning. Risks:
   (a) The preliminary investigation itself takes too much time.
   (b) The decision after the preliminary investigation requires intervention by the user (problematic for automation). While this is feasible for individual cases, it is not practical for numerous experiments with different data (as in the experiments of the study in Sec. 6). Moreover, this reduces the reproducibility of the results.
   (c) Depending on the scope and approach of the preliminary study, interactions with other parameters may not be adequately accounted for.

3. No distinction: parameters like `num.trees` are optimized by the tuning procedure just like all other hyperparameters. Risks:
   (a) The upper bound for the parameter is set too low, so potentially good models are not explored by the tuning procedure. (Note: bounds set too tight for the search space are a general risk that can affect all other hyperparameters as well).
   (b) The upper bound is set too high, causing individual evaluations to use unnecessary amounts of time during tuning.
   (c) The best found value may become unnecessarily large (from a model quality point of view).

4. Multi-objective: runtime and model quality can be optimized simultaneously in the context of multi-objective optimization. Risks:
   (a) Again, manual evaluation is necessary (selection of a sector of the Pareto front) to avoid that from a practical point of view irrelevant (but possibly Pareto-optimal) solutions are investigated.
   (b) This manual evaluation also reduces reproducibility.

5. Regularization via weighted sum: The number of trees (or similar parameters) can be incorporated into the objective function. In this case, the objective function becomes a weighted sum of model quality and number of trees (or runtime), with a weighting factor θ.
   (a) The new parameter of the tuning procedure, θ, has to be determined.
   (b) Moreover, the optimization of a weighted sum cannot find certain Pareto-optimal solutions if the Pareto-front is non-convex.

In our experimental investigation, we use solution 3. That is, the corresponding parameters are tuned but do not undergo any special treatment during tuning. Due to the large number of experiments, user interventions would not be possible and would also complicate the reproducibility of the results. In principle, we recommend this solution for use in practice.

In individual cases, or if a good understanding of algorithms and data is available, solution 2 can also be used. For this, SPOT can be interrupted after the first evaluation step, in order to set the corresponding parameters to a certain value or to adjust the bounds if necessary (e.g., if `num.trees` was examined with too low an upper bound).

Appendix

Abbreviations

AUC Area Under the receiver operating characteristic Curve

DT Decision Tree

EN Elastic Net

k-NN k-Nearest-Neighbor

ML Machine Learning

MCE mean Mis-Classification Error

RF Random Forest

RMSE Root Mean Squared Error

RS Random Search

SPOT SPO Toolbox

SVM Support Vector Machine

XGBoost eXtreme Gradient Boosting

References

Haneen Arafat Abu Alfeilat, Ahmad B.A. Hassanat, Omar Lasassmeh, Ahmad S. Tarawneh, Mahmoud Bashir Alhasanat, Hamzeh S. Eyal Salman, and V.B. Surya Prasath. Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big Data*, 7(4):221–248, dec 2019. doi: 10.1089/big.2018.0175.

Thomas Bartz-Beielstein, Christian Lasarczyk, and Mike Preuß. Sequential parameter optimization. In B. McKay et al., editors, *Proceedings of the Congress on Evolutionary Computation (CEC 2005)*, pages 773–780, New York, NY, USA, 2005. IEEE.

Thomas Bartz-Beielstein, Frederik Rehbach, and Margarita Rebolledo. Tuning algorithms for black-box optimization: State of the art and future perspectives. In Panos Pardalos, Varvara Rasskazova, and Michael Vrahatis, editors, *Black Box Optimization, Machine Learning and No-Free Lunch Theorems*, number 170 in Springer Optimization and Its Applications. Springer, 2021a. doi: https://doi.org/10.1007/978-3-030-66515-9_3. Accepted, in print.

Thomas Bartz-Beielstein, Joerg Stork, Martin Zaefferer, Margarita Rebolledo, Christian Lasarczyk, and Frederik Rehbach. Spot: Sequential parameter optimization toolbox v2.3.0. online, 2021b. Available: `https://cran.r-project.org/package=SPOT`, accessed: 2021-03-15.

Thomas Bartz-Beielstein, Martin Zaefferer, and Frederik Rehbach. In a nutshell – the sequential parameter optimization toolbox. *arXiv*, 2021c. arXiv:1712.04076v2.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012. URL `http://jmlr.org/papers/v13/bergstra12a.html`.

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

Jason Brownlee. *XGBoost With Python – Gradient Boosted Trees with XGBoost and scikit-learn*. Machine Learning Mastery, v1.10 edition, 2018.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM. *ACM Transactions on Intelligent Systems and Technology*, 2(3): 1–27, apr 2011. doi: 10.1145/1961189.1961199. Updated version from November 29, 2019.

Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. doi: 10.1145/2939672.2939785.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, and Yutian Li. Package xgboost (reference manual, v1.2.0.1). online, retrieved: 2020-11-22, 2020. `https://cran.r-project.org/web/packages/xgboost/xgboost.pdf`.

Vladimir Cherkassky and Yunqian Ma. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, 17(1):113–126, jan 2004. doi: 10.1016/s0893-6080(03)00169-2.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, jan 1967. doi: 10.1109/tit.1967.1053964.

George De Ath, Richard M. Everson, Alma A. M. Rahat, and Jonathan E. Fieldsend. Greed is good: Exploration and exploitation trade-offs in bayesian optimisation. *ArXiv e-prints*, 2019.

Quoc Hoan Doan, Duc-Kien Thai, and Ngoc Long Tran. A hybrid model for predicting missile impact damages based on k-nearest neighbors and Bayesian optimization. *Journal of Science and Technology in Civil Engineering (STCE) - NUCE*, 14(3):1–14, aug 2020. doi: 10.31814/stce.nuce2020-14(3)-01.

Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, oct 2012. doi: 10.1145/2347736.2347755.

Harris Drucker and Corinna Cortes. Boosting decision trees. In *NIPS 95: Proceedings of the 8th International Conference on Neural Information Processing Systems*, pages 479–485, November 1995.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, aug 1997. doi: 10.1006/jcss.1997.1504.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010. doi: 10.18637/jss.v033.i01.

Jerome Friedman, Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Kenneth Tay, Noah Simon, and Junyang Qian. Package glmnet (reference manual, v4.0-2). online, retrieved: 2020-11-22, 2020. `https://cran.r-project.org/web/packages/glmnet/glmnet.pdf`.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5): 1189–1232, 2001.

Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, feb 2002. doi: 10.1016/s0167-9473(01)00065-2.

Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw.*, 3(3):209–226, September 1977. ISSN 0098-3500. doi: 10.1145/355744.355745. URL `https://doi.org/10.1145/355744.355745`.

Colin Gillespie. Package benchmarkme (reference manual, v1.0.5). online, retrieved: 2021-02-09, 2021. `https://cran.r-project.org/web/packages/ranger/ranger.pdf`.

Rafael Gomes Mantovani, Tomáš Horváth, Ricardo Cerri, Sylvio Barbon Junior, Joaquin Vanschoren, and André Carlos Ponce de Leon Ferreira de Carvalho. An empirical study on hyperparameter tuning of decision trees. *arXiv*, 2018. arXiv:1812.02207, v2.

Nick Guenther and Matthias Schonlau. Support vector machines. *The Stata Journal: Promoting communications on statistics and Stata*, 16(4):917–937, dec 2016. doi: 10.1177/1536867x1601600407.

Trevor Hastie and Junyang Qian. An introduction to glmnet. online, retrieved: 2020-11-22, 2016. `https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet.pdf`.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2017. doi: 10.1007/978-0-387-84858-7. 12th printing.

Kurt Hornik and David Meyer. Deriving consensus rankings from benchmarking experiments. In Reinhold Decker and Hans J. Lenz, editors, *Advances in Data Analysis*, pages 163–170, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classication. online, retrieved: 2020-11-29, 2016. `https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

Frank Hutter, Holger H. Hoos, Kevin Leyton-Brown, and Thomas Stützle. Paramils: An automatic algorithm configuration framework. *Journal of Artificial Intelligence Research*, 36(1):267—306, September 2009. ISSN 1076-9757.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 7th edition, 2014.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2017. doi: 10.1007/978-1-4614-7138-7.

John G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959. URL `http://www.jstor.org/stable/20026529`.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938.

Faiza Khan, Summrina Kanwal, Sultan Alamri, and Bushra Mumtaz. Hyper-parameter optimization of classifiers, using an artificial immune network and its application to software bug prediction. *IEEE Access*, 8:20954–20964, 2020. doi: 10.1109/access.2020.2968362.

Upmanu Lall and Ashish Sharma. A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, 32(3):679–693, mar 1996. doi: 10.1029/95wr02966.

Stephen Leary, Atul Bhaskar, and Andy Keane. Optimal orthogonal-array-based latin hypercubes. *Journal of Applied Statistics*, 30(5):585–598, jun 2003. doi: 10.1080/0266476032000053691.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 6389–6399. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.

Manuel López-Ibáñez, Jérémie Dubois-Lacoste, Leslie Pérez Cáceres, Mauro Birattari, and Thomas Stützle. The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, 3:43–58, 2016. doi: 10.1016/j.orp.2016.09.002.

Ana C. Lorena, Luís P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, and Tin Kam Ho. How complex is your classification problem?: A survey on measuring classification complexity. *ACM Computing Surveys*, 52(5):1–34, oct 2019. doi: 10.1145/3347711.

Gilles Louppe. *Understanding random forests – from theory to practice*. PhD thesis, University of Liege, 2015.

Rafael G. Mantovani, Andre L. D. Rossi, Joaquin Vanschoren, Bernd Bischl, and Andre C. P. L. F. de Carvalho. Effectiveness of random search in SVM hyper-parameter tuning. In *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, jul 2015. doi: 10.1109/ijcnn.2015.7280664.

Olaf Mersmann, Heike Trautmann, Boris Naujoks, and Claus Weihs. Benchmarking evolutionary multiobjective optimization algorithms. In *IEEE Congress on Evolutionary Computation*. IEEE, jul 2010. doi: 10.1109/cec.2010.5586241.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. Package e1071 (reference manual, v1.7-4). online, retrieved: 2020-11-29, 2020. https://cran.r-project.org/web/packages/e1071/e1071.pdf.

Haidar Osman, Mohammad Ghafari, and Oscar Nierstrasz. Hyperparameter optimization to improve bug prediction accuracy. In *2017 IEEE Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE)*. IEEE, feb 2017. doi: 10.1109/maltesque.2017.7882014.

Mike Preuss and Thomas Bartz-Beielstein. Sequential Parameter Optimization Applied to Self-Adaptation for Binary-Coded Evolutionary Algorithms. In Fernando Lobo, Claudio Lima, and Zbigniew Michalewicz, editors, *Parameter Setting in Evolutionary Algorithms*, pages 91–120. Springer, Berlin, Heidelberg, New York, 2007.

Philipp Probst and Anne-Laure Boulesteix. To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18(181):1–18, 2018. URL http://jmlr.org/papers/v18/17-269.html.

Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20(53):1–32, 2019a.

Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), jan 2019b. doi: 10.1002/widm.1301.

Frederik Rehbach, Martin Zaefferer, Boris Naujoks, and Thomas Bartz-Beielstein. Expected improvement versus predicted value in surrogate-based optimization. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pages 868–876, Cancun, Mexico, July 2020. ACM. doi: 10.1145/3377930.3389816.

Klaus Schliep, Klaus Hechenbichler, and Antoine Lizee. Package kknn (reference manual, v1.3.1). online, retrieved: 2020-11-29, 2016. https://cran.r-project.org/web/packages/kknn/kknn.pdf.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, 2001.

Patrick Schratz, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, and Alexander Brenning. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406:109–120, aug 2019. doi: 10.1016/j.ecolmodel.2019.06.002.

Erwan Scornet. Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, 60:144–162, 2017. doi: 10.1051/proc/201760144.

Fabio Sigrist. Gradient and newton boosting for classication and regression. *arXiv*, 2020. arXiv:1808.03064, v7.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for coxs proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5), 2011. doi: 10.18637/jss.v039.i05.

Rainer Storn and Kenneth Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.

Ch. Sudheer, R. Maheswaran, B. K. Panigrahi, and Shashi Mathur. A hybrid SVM-PSO model for forecasting monthly streamflow. *Neural Computing and Applications*, 24(6):1381–1389, feb 2013. doi: 10.1007/s00521-013-1341-y.

Terry M. Therneau and Elizabeth J. Atkinson. An introduction to recursive partitioning using the rpart routines. online, retrieved: 2020-11-22, 2019. `https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf`.

Terry M. Therneau, Elizabeth J. Atkinson, and Brian Ripley. Package rpart (reference manual, v4.1-15). online, retrieved: 2020-11-22, 2019. `https://cran.r-project.org/web/packages/rpart/rpart.pdf`.

Janek Thomas, Stefan Coors, and Bernd Bischl. Automatic gradient boosting. In *International Workshop on Automatic Machine Learning at ICML*, 2018.

Jan N. van Rijn and Frank Hutter. Hyperparameter importance across datasets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, jul 2018. doi: 10.1145/3219819.3220058.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL `http://doi.acm.org/10.1145/2641190.2641198`.

Yan Wang and Xuelei Sherry Ni. A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. *International Journal of Database Management Systems*, 11(01):01–17, feb 2019. doi: 10.5121/ijdms.2019.11101.

Simon Wessing and Mike Preuss. The true destination of EGO is multi-local optimization. In *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*. IEEE, nov 2017. doi: 10.1109/la-cci.2017.8285677.

Jenna Wong, Travis Manderson, Michal Abrahamowicz, David L Buckeridge, and Robyn Tamblyn. Can hyperparameter tuning improve the performance of a super learner? *Epidemiology*, 30(4):521–531, jul 2019. doi: 10.1097/ede.0000000000001027.

Marvin N. Wright. Package ranger (reference manual, v0.12.1). online, retrieved: 2020-11-22, 2020. `https://cran.r-project.org/web/packages/ranger/ranger.pdf`.

Marvin N. Wright and Inke R. König. Splitting on categorical predictors in random forests. *PeerJ*, 7:e6339, feb 2019. doi: 10.7717/peerj.6339.

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

xgboost developers. xgboost release 1.2.1 (documentation). online, retrieved: 2020-11-28, 2020. `https://xgboost.readthedocs.io/_/downloads/en/stable/pdf/`.

Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, nov 2020. doi: 10.1016/j.neucom.2020.07.061.

Arman Zharmagambetov, Suryabhan Singh Hada, Miguel A. Carreira-Perpinan, and Magzhan Gabidolla. An experimental comparison of old and new decision tree algorithms. *arXiv*, 2019. arXiv:1911.03054, v2.

Jian Zhou, Yingui Qiu, Shuangli Zhu, Danial Jahed Armaghani, Manoj Khandelwal, and Edy Tonnizam Mohamad. Estimation of the TBM advance rate under hard rock conditions using XGBoost and bayesian optimization. *Underground Space*, jul 2020. doi: 10.1016/j.undsp.2020.05.008. corrected proof.

Junyi Zhou, Jing Shi, and Gong Li. Fine tuning support vector machines for short-term wind speed forecasting. *Energy Conversion and Management*, 52(4):1990–1998, apr 2011. doi: 10.1016/j.enconman.2010.11.007.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 13697412, 14679868. URL `http://www.jstor.org/stable/3647580`.