




# Imbalanced regression and extreme value prediction

Rita P. Ribeiro<sup>1,2</sup> · Nuno Moniz<sup>1,2</sup> 

Received: 15 January 2020 / Revised: 31 July 2020 / Accepted: 11 August 2020 /  
Published online: 4 September 2020

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2020

## Abstract

Research in imbalanced domain learning has almost exclusively focused on solving classification tasks for accurate prediction of cases labelled with a rare class. Approaches for addressing such problems in regression tasks are still scarce due to two main factors. First, standard regression tasks assume each domain value as equally important. Second, standard evaluation metrics focus on assessing the performance of models on the most common values of data distributions. In this paper, we present an approach to tackle imbalanced regression tasks where the objective is to predict extreme (rare) values. We propose an approach to formalise such tasks and to optimise/evaluate predictive models, overcoming the factors mentioned and issues in related work. We present an automatic and non-parametric method to obtain relevance functions, building on the concept of relevance as the mapping of target values into non-uniform domain preferences. Then, we propose *SERA*, a new evaluation metric capable of assessing the effectiveness and of optimising models towards the prediction of extreme values while penalising severe model bias. An experimental study demonstrates how *SERA* provides valid and useful insights into the performance of models in imbalanced regression tasks.

**Keywords** Supervised learning · Imbalanced domain learning · Imbalanced regression · Extreme value prediction

## 1 Introduction

The primary assumption of standard supervised learning tasks is that each value of the domain is equally important. However, this is not always true. In domains such as finance, meteorology or environmental sciences, the goal is often the prediction of uncommon events, also known as rare/extreme cases. Imbalanced domain learning tasks

---

Editors: Ira Assent, Carlotta Domeniconi, Aristides Gionis, Eyke Hüllermeier.

---

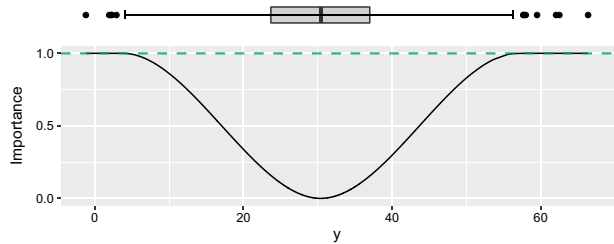
✉ Nuno Moniz  
nmmoniz@inesctec.pt

Rita P. Ribeiro  
rpribeiro@fc.up.pt

<sup>1</sup> INESC TEC, Porto, Portugal

<sup>2</sup> Department of Computer Science, Faculty of Sciences, University of Porto, Porto, Portugal

**Fig. 1** Illustration of the importance of values for a target variable distribution in a regression task: the assumption of uniform domain preferences (dashed green), as in standard regression tasks, and our objective—non-uniform domain preferences biased to extreme values (black) (Color figure online)



formalise such predictive modelling scenarios. These have two characteristics (Branco et al. 2016): (i) skewed distribution of target variables and (ii) domain preference for underrepresented cases.

Research concerning imbalanced domain learning spans over 20 years, addressing various aspects (Fernández et al. 2018; Branco et al. 2016; López et al. 2013; Krawczyk 2016; He and Ma 2013). These include (i) the formalisation of the task, (ii) shortcomings of standard learning algorithms, (iii) strategies to overcome such limitations, and (iv) the search for appropriate evaluation metrics. The problem of imbalanced classification and, especially, binary classification, has been the main focus of research in this topic. In comparison, the volume of research concerning imbalanced regression tasks is negligible.

Imbalanced regression faces two significant challenges. First, to provide a principled approach capable of describing non-uniform preferences over continuous domains. Figure 1 illustrates the difference between the standard assumption of uniform and that of non-uniform preferences. Although trivial for classification tasks (e.g. deciding the positive class), ad-hoc solutions would require a specification of preferences over a potentially infinite domain. As such, we may require automatic methods for specifying those preferences. Nonetheless, we cannot base such methods on static assumptions concerning the shape of the distribution, such as the assumption of normality. Such assumption has long-standing reports of its theoretical and practical inconsistencies (Hald 1998; Wilcox 1990) especially when assumed in continuous distributions (Wilcox 2005). Nonetheless, it is a common assumption in machine learning and earlier work in the related topic of utility-based regression. The second challenge has to do with finding appropriate evaluation and optimisation criteria capable of improving the predictive ability of models towards extreme values without severe model biasing. As in classification tasks, standard metrics for numerical prediction are not appropriate for these endeavours, given their focus on the normal behaviour of models. Also, concerning alternative metrics, these either consider all target values as equally important (Crone et al. 2005), i.e. uniform preferences, or allow for extraordinary model biasing towards extreme values, resulting in low generalisation capability. In this work, we tackle these challenges.

The main objective of this paper is to provide a new basis for the formalisation of imbalanced regression tasks and metrics for model evaluation and optimisation in this context. Such implies the identification of issues in related work and the proposal of new methods. The contributions of this paper are as follows:

- a review of earlier work on utility-based regression, and non-standard evaluation metrics for numerical prediction;
- an automatic and non-parametric method to infer non-uniform domain preferences biased to extreme values, tackling the assumption of an underlying normal distribution in earlier work (Torgo and Ribeiro 2007);

- a new evaluation metric *SERA* (Squared Error-Relevance Area) allowing the optimisation and evaluation of models as to their ability to predict extreme values, while robust to severe model biasing;
- an extensive experimental study which shows that (i) the evaluation metric *SERA* provides a robust tool for selection and optimisation procedures, (ii) as well as the analysis of prediction models' performance and, (iii) its impact and predictive trade-offs;
- an open-source package containing all developed methods.<sup>1</sup>

The remainder of this paper is organised as follows. The task of imbalanced regression is described in Sect. 2, along with a motivating example. Section 3 formalises the concept of relevance and proposes a new non-parametric method to generate relevance functions automatically. Section 4 reviews evaluation metrics for imbalanced regression tasks and proposes a new metric for assessing model performance. Section 5 presents an extensive experimental study to support initial claims, followed by a discussion of results in Sect. 6, and conclusions in Sect. 7.

## 2 Imbalanced regression

Let  $D$  be a training set defined by  $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ , where  $\mathbf{x}_i$  is a feature vector from the feature space  $\mathcal{X}$  composed by predictor (independent) variables and  $y_i$  is an instance of the target (dependent) variable  $Y$  with domain  $\mathcal{Y}$ . For supervised learning tasks, the objective is to learn the best approximation of an unknown function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ . Depending on the domain of the target variable, we may have a classification problem (if  $\mathcal{Y}$  is discrete) or a regression problem (if  $\mathcal{Y}$  is continuous). The approximation  $h$  is a model obtained by optimising a preference criterion  $\mathcal{L}$  on the training set. We base such optimisation on the search over the parameter space of the algorithm that builds the model.

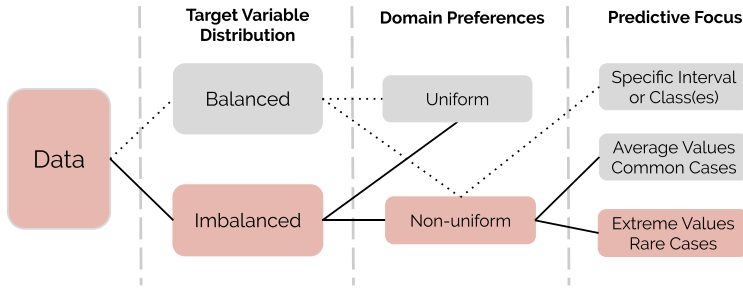
In regression, overall error estimates, such as mean absolute error or mean squared error, are used as standard preference criteria. They focus on minimising the average error across the domain of the target variable. Given the predominance of cases with target values within/near the central tendency of the distributions, reducing the error of predictions for cases with common values will have a considerable impact on model performance. Naturally, model performance is mainly related to the accurate prediction of such values in detriment of extreme values.

However, as previously mentioned, for many real-world domains, not all values of the target variable are equally important. In many occasions, the prediction of extreme values is pertinent to be particularly accurate. Such raises the problem of imbalanced regression, described by the factors in the following taxonomy.

### 2.1 Taxonomy for imbalanced domain learning

Three factors must be considered when analysing imbalanced learning tasks: (i) the distribution of the target variable, (ii) domain preferences concerning accurate predictions, and (iii) their predictive focus. A schematic definition of such type of tasks is presented

<sup>1</sup> <https://github.com/nunompmniz/IRon>.



**Fig. 2** Taxonomy of imbalanced domain learning tasks. The distribution of target variables can be balanced or imbalanced. Each of these may have uniform preferences or non-uniform. Balanced distributions with non-uniform preferences have a predictive focus on specific intervals or classes. As for imbalanced distributions, their predictive focus is either on average values/common cases or extreme values/rare cases. The characteristic conditions for imbalanced domain learning are marked red

in Fig. 2. The following sections describe each of the factors, focusing on the problem of imbalanced regression.

### 2.1.1 Target variable distribution

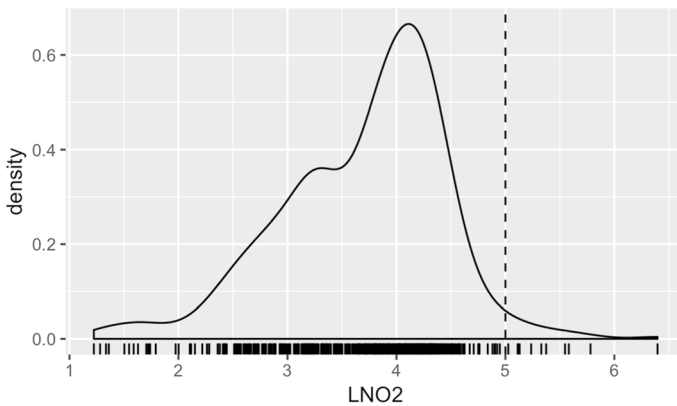
The probability distribution of a target variable  $y$  may be balanced or imbalanced, depending on the frequency of all possible values, and if they present similar probabilities or not. In classification tasks, assuming a binary class scenario, the target variable distribution is considered imbalanced if the discrepancy in class cardinality is large, i.e. given a data set  $\mathcal{D}$ , subsets labelled with common ( $\mathcal{D}_N$ ) or rare ( $\mathcal{D}_R$ ) classes imply  $|\mathcal{D}_N| \gg |\mathcal{D}_R|$ . For regression tasks, we consider that the target variable distribution is imbalanced when extreme values (or outliers) are present. Commonly—but not necessarily—it translates to a skewed distribution.

### 2.1.2 Domain preferences

Given a particular prediction problem end-users have domain preferences. These describe the importance that each value of the domain has in terms of obtaining a precise prediction. This factor can be uniform or non-uniform. The former considers that each value of the target variable domain is equally important. The latter stipulates specific ranges of target values as more or less critical. For example, in a season of wildfire danger, consider the task of predicting the temperature at noon for the following day: a predicted value of 30 °C for true values of 20 °C or 40 °C incurs an error of equal magnitude. Nevertheless, the latter presents a more hazardous situation.

### 2.1.3 Predictive focus

Finally, the third factor—predictive focus—entails the definition of which intervals have more or less importance w.r.t precise predictions. In case the underlying distribution is considered balanced, under non-uniform domain preferences, the predictive focus will concern specific intervals of values or a given class (or set of classes). As for the case where the distribution of values is considered imbalanced, possibilities mainly concern focusing on



**Fig. 3** The pdf of the log-transformed NO<sub>2</sub> hourly concentration values (LNO2). The vertical line indicates the limit threshold established by Directive 2008/50/EC

average/common values or extreme/rare values. Consider the previous example of anticipating the temperature at noon for the following day. Assuming moderate climate conditions, the goal of the first scenario (true value 20 °C) may be the prediction of values within the central tendency of the distribution (disregarding extreme values). The latter scenario (40 °C) describes an objective focused on accurately predicting values outside of such central tendency, e.g. extreme temperatures. Thus, predictive focus and the importance of values is related to probability distributions, either in a proportional (focus on average values) or inversely proportional manner (focus on extreme values).

According to the presented taxonomy, learning tasks are considered imbalanced regression tasks when, given a particular distribution of continuous values, (i) such distribution shows the presence of outliers, (ii) domain preferences are not uniform, and (iii) predictive focus is on extreme values. The following section describes an example of an imbalanced regression task, motivating our contributions.

## 2.2 Application: prediction of outdoor air pollution

Road traffic, industries and forest fires are among the main factors that affect air quality. Due to increasing percentages of unhealthy substances, the World Health Organization (WHO) determined the establishment of concentration limits for some toxic compounds (Organization 2005)—AQGs (“Air Quality Guideline”).

Real data from a study (Aldrin and Haff 2005) that relates air pollution on a road with traffic volume and meteorology is used. The data<sup>2</sup> consists of 500 observations collected by the Norwegian Public Roads Administration. The target variable (LNO2) is the log-transformed concentration values of NO<sub>2</sub> measured in µg/m<sup>3</sup> for each hour, at Alnabru (Oslo, Norway), between October 2001 and August 2003. Figure 3 depicts the probability density function (pdf) of the LNO2 variable.

According to the ruling Directive 2008/50/EC, concentration values above 150 µg/m<sup>3</sup> ( $\ln(150 \mu\text{g}/\text{m}^3) \approx 5.0$ ) are dangerous—they should be avoided. Anticipating them is highly

<sup>2</sup> Available on the StatLib Datasets Archive: <http://lib.stat.cmu.edu/datasets/>.

important. Thus, the goal is to obtain a model particularly accurate in the prediction of alarming levels of  $NO_2$  emissions.

The question that may arise is why not simplify the predictive task by casting it as a classification task, aggregating in a class all the extreme and dangerous values? In addition to multiple issues extensively raised in previous research (Royston et al. 2006), there are two key issues. First, it does not solve the imbalance issue, i.e. the most important class would be much less frequent than any other. Second, even though the number of bins and the respective cut-off points can be defined based on domain knowledge, this process creates crisp and artificial divisions between values of the target variable. This causes the relationship between the response and the predictors to be flat within intervals, thus dismissing the notion of numeric precision to a great extent. To illustrate, take the mentioned air pollution study. Given the limit concentration value of  $150 \mu\text{g}/\text{m}^3$ , let  $\hat{y}_1 = 149$  and  $\hat{y}_2 = 200$  be predictions of models  $M_1$  and  $M_2$ , respectively, for a true value  $y = 151$ . If transformed into a classification problem, model  $M_2$  would be more precise, although  $M_1$  presents the best approximation (smallest numeric deviation). Also, consider a prediction from a third model  $M_3$  where  $\hat{y}_3 = 151$ . As a classification problem, predictions of models  $M_2$  and  $M_3$  would be considered equally correct, despite  $M_3$  presenting a precise prediction of the true value.

So far, we have provided a formalisation of the learning task and a description of its characteristics. Nonetheless, progress in research concerning imbalanced regression faces two challenges. First, to describe non-uniform preferences over continuous (and infinite) domains; second, to properly optimise and evaluate predictive models in such settings. We address such challenges in the following sections.

### 3 Relevance functions in imbalanced regression

There are many real-world imbalanced problems for which the prediction of a continuous target value is essential, and where a classification approach may not be appropriate. Examples include domains such as climate/weather (Freemeteo 2017), electricity (Koprinska et al. 2011) and water demand (Herrera et al. 2010) or financial markets (Akbulgic et al. 2014). In such domains, there are multiple tasks focused on models' ability to anticipate extreme values and where magnitude is a factor. Examples include the prediction of extremely high or low stock market returns, abnormally high-temperature levels, electricity load or water consumption or harmful outdoor air pollution levels. Nevertheless, it is difficult to map continuous target variables and domain preferences. In this section, we review and formalise an approach to obtain such mapping through relevance functions, and propose a non-parametric approach for its automatic definition.

#### 3.1 Definition

To our knowledge, Torgo and Ribeiro (2007) are the first to mention the concept of relevance in a similar context. Introduced as a domain-dependent concept for utility-based regression tasks, it translates a target value into a scale of relevance, describing the importance of obtaining an accurate prediction. Ribeiro (2011) proposes a first approach for obtaining relevance functions, used in early work concerning utility-based regression and forecasting tasks (e.g. Torgo et al. 2013; Branco et al. 2016; Moniz et al. 2017b, a; Branco et al. 2019). It is formally defined as follows.

**Definition 1** (*Relevance function*) A relevance function  $\phi(Y) : \mathcal{Y} \rightarrow [0, 1]$  is a continuous function that expresses the application-specific bias concerning the target variable domain  $\mathcal{Y}$  by mapping it into a  $[0, 1]$  scale of relevance, where 0 and 1 represent the minimum and maximum relevance, respectively.

In imbalanced classification, specifying the relevance of a target variable is, in most cases, choosing the positive (minority) class. For imbalanced regression, given the potentially infinite nature of the target variable domain, specifying the relevance of all values is virtually impossible, requiring an approximation. Two essential components are necessary: a set of data points where relevance is known, i.e. control points, and a decision on which interpolation method to use.

### 3.1.1 Control points

In order to obtain a relevance function,  $\phi : \mathcal{Y} \rightarrow [0, 1]$ , a set of control points  $S = \{(y_k, \phi(y_k), \phi'(y_k))\}_{k=1}^s$  must be given as input to an interpolation algorithm. This set must contain information on (i) the target value  $y_k$ , (ii) its respective relevance value  $\phi(y_k)$  and on (iii) the intended derivative of the relevance function at that point  $\phi'(y_k)$ . By default, control points are assumed as local minimum or maximum of relevance and, thus, all have derivative  $\phi'(y_k)$  equal to zero. But, other values may be provided, adjusting them so that monotonicity is preserved.

Ideally, control points should be introduced based on domain knowledge. However, it is common for such knowledge to be unavailable or nonexistent even. To tackle such scenarios, we propose an approach to automatically obtain control points based on the distribution of target variables, described in Sect. 3.3.

### 3.1.2 Interpolation

Interpolation concerns the estimation of values within a range given by a set of data points (Phillips 2003). We aim at estimating the relevance of target values given a set of data points for which such relevance is known, i.e. control points. There are two main types of interpolation methods: statistical and spline smoothing. The former includes methods such as loess smoothing (Cleveland et al. 1992). The latter includes methods such as nearest-neighbour, bilinear, bicubic and shape-preserving interpolation (Basu et al. 2015).

The interpolation required within the scope of our objective has several aspects that help decide which methods are most appropriate. For example, the set of control points will typically have a very low cardinality. Such restriction hampers the application of statistical methods. These methods focus on minimising the residual sum of distances between two continuous functions and require reasonably large and densely sampled sets of data points (Basu et al. 2015). Concerning spline smoothing methods, two aspects help focus our decision. First, interpolation methods should have certain properties as to guarantee that the relevance of data points is distinguishable. Such includes properties of continuity, convexity and monotonicity. Second, the interpolation method must guarantee that the estimated relevance values match the relevance values at control points. Given this, shape-preserving interpolation methods provide the best option. The main reasons are that these are not prone to unrealistic overshoots of estimated values (as in cubic splines Barker and McDougall 2020) and that they are bounded by the data points provided.

We propose, as in early work by Ribeiro (2011), the use of Piecewise Cubic Hermite Interpolating Polynomials (Dougherty et al. 1989) (*pchip*), using a set of  $n$  known relevance values, i.e. control points. We base this option on the ability of *pchip* in guaranteeing the smoothness of interpolation and allowing user control on the generated function’s shape—it requires the derivative at each control point. By restricting the first derivative at control points, the method preserves local positivity, monotonicity and convexity of the data. As mentioned, these are convenient properties in the context of our target applications as we want to induce a continuous function that reproduces, as closely as possible, the relevance values of the supplied control points. Algorithm 1 shows how *pchip* is performed over a set of control points  $S$ .

---

**Algorithm 1** *pchip*( $S$ ) Piecewise Cubic Hermite Interpolating Polynomial.

---

```

Input:  $S = \{(y_k, \varphi(y_k), \varphi'(y_k))\}_{k=1}^s$ , set of control points with  $y_1 < y_2 < \dots < y_s$ , their respective relevance values  $\varphi(y_k)$  and preliminary derivatives  $\varphi'(y_k)$ .
Output:  $\phi(y)$ : a Piecewise Cubic Hermite Interpolating Polynomial.
1: for  $k \leftarrow 1$  to  $s - 1$  do
2:    $h_k \leftarrow y_{k+1} - y_k$ 
3:    $\delta_k \leftarrow (\varphi(y_{k+1}) - \varphi(y_k))/h_k$ 
4:    $a_k \leftarrow \varphi(y_k)$ 
5: end for
6:  $\{b_k\}_{k=1}^{s-1} \leftarrow \text{check\_slopes}(\{\varphi'(y_k)\}_{k=1}^{s-1}, \{\delta_k\}_{k=1}^{s-1})$  // Algorithm 2 [26]
7: for  $k \leftarrow 1$  to  $s - 1$  do
8:    $c_k \leftarrow (3\delta_k - 2b_k + b_{k+1})/h_k$ 
9:    $d_k \leftarrow (b_k - 2\delta_k + b_{k+1})/h_k^2$ 
10: end for
11: return  $\phi(y) = a_k + b_k(y - y_k) + c_k(y - y_k)^2 + d_k(y - y_k)^3, y \in [y_k, y_{k+1}[$ 

```

---

A key feature of this algorithm is finding the right slopes at given points. This guarantees that the interpolant is *piecewise monotone*, i.e. its derivative does not change the sign in any interval defined by control points. This task is ensured by the *check\_slopes* method proposed by Fritsch and Carlson (1980) and presented in Algorithm 2. The method estimates reasonable derivatives at each control point. Once the first derivative values are known, the four coefficients returned by *pchip* are calculated for each interval of the interpolant. If a control point is a local maximum or minimum, the method ensures a zero derivative.



**Table 1** Control points of LNO2 concentration thresholds according to Directive 2008/50/EC

$y_k$ : LNO2 concentration values		$\varphi(y_k)$	$\varphi'(y_k)$
Low concentration:	$\ln(3 \mu\text{g}/\text{m}^3) \approx 1.1$	0.0	0.0
Annual mean guideline:	$\ln(40 \mu\text{g}/\text{m}^3) \approx 3.7$	0.0	0.0
Limit threshold:	$\ln(150 \mu\text{g}/\text{m}^3) \approx 5.0$	1.0	0.0

**Algorithm 2** check\_slopes( $\Phi, \Delta$ ) Fritsch and Carlson (1980) method [26].

```

Input:  $\Phi = \{\varphi'(y_k)\}_{k=1}^{s-1}$ ,  $\Delta = \{\delta_k\}_{k=1}^{s-1}$ 
1: for  $k \leftarrow 1$  to  $s - 1$  do
2:   if  $\delta_k = 0$  then
3:      $\varphi'(y_k) \leftarrow \varphi'(y_{k+1}) \leftarrow 0$ 
4:   else
5:      $\alpha \leftarrow \varphi'(y_k) / \delta_k$ 
6:      $\beta \leftarrow \varphi'(y_{k+1}) / \delta_k$ 
7:     if  $\varphi'(y_k) \neq 0 \wedge \alpha < 0$  then
8:        $\varphi'(y_k) \leftarrow -\varphi'(y_k)$ 
9:        $\alpha \leftarrow \varphi'(y_k) / \delta_k$ 
10:    end if
11:    if  $\varphi'(y_{k+1}) \neq 0 \wedge \beta < 0$  then
12:       $\varphi'(y_{k+1}) \leftarrow -\varphi'(y_{k+1})$ 
13:       $\beta \leftarrow \varphi'(y_{k+1}) / \delta_k$ 
14:    end if
15:     $\tau_1 \leftarrow 2\alpha + \beta - 3$ 
16:     $\tau_2 \leftarrow \alpha + 2\beta - 3$ 
17:    if  $\tau_1 > 0 \wedge \tau_2 > 0 \wedge \alpha(\tau_1 + \tau_2) < \tau_1\tau_2$  then
18:       $\tau \leftarrow 3\delta_k / \sqrt{\alpha^2 + \beta^2}$ 
19:       $\varphi'(y_k) \leftarrow \alpha\tau$ 
20:       $\varphi'(y_{k+1}) \leftarrow \beta\tau$ 
21:    end if
22:  end if
23: end for
24: return  $\Phi = \{\varphi'(y_k)\}_{k=1}^{s-1}$ 

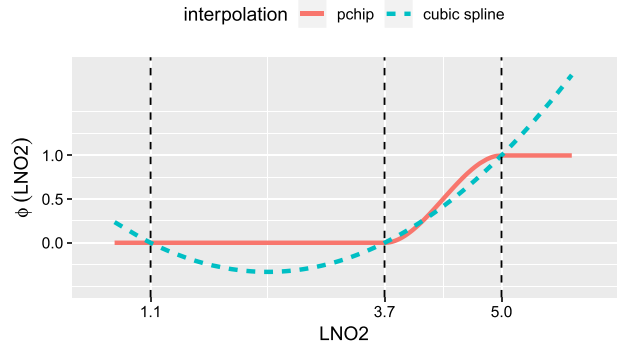
```

At the end of the interpolation process, evaluating the relevance of value  $y \in \mathcal{Y}$ , i.e.  $\phi(y)$ , corresponds to estimating the interpolant  $\phi(y)$  such that  $[y_k, y_{k+1}[$  is the interval of control points interpolation to which  $y$  belongs. Beyond the maximum and the minimum values supplied in the set of control points, the relevance function is guaranteed to be constant, by linear extrapolation.

**3.2 Illustrative example**

Considering the NO<sub>2</sub> Emissions prediction problem described in Sect. 2.2, the Directive 2008/50/EC contains information on the relevance of certain data points. In particular, the goal to maintain the LNO2 hourly concentration values below a limit equal to  $\ln(150 \mu\text{g}/\text{m}^3) \approx 5.0$ —a value with maximum relevance, and the annual mean guideline of  $\ln(40 \mu\text{g}/\text{m}^3) \approx 3.7$ —minimum relevance. To complete our comprehension of the domain (Table 1), the lowest concentration value of LNO2 is attributed minimum relevance.

**Fig. 4** Relevance functions for the prediction of LNO2 concentration values obtained by two different interpolation methods: piecewise cubic hermite interpolating polynomial (*pchip*) versus cubic splines. The data points used are based on domain knowledge from Directive 2008/50/EC



Based on this information of control points, Fig. 4 presents two possible relevance functions  $\phi()$ : one obtained by Algorithm 1 (*pchip*) and another obtained by `splinefun`, a standard cubic interpolation algorithm available on R software (R Core Team 2017).

Results show that the relevance function produced by *pchip* suits best the application goals. Also, cubic spline interpolation does not allow much control over the function. It does not confine the relevance function to the stipulated  $[0, 1]$  interval scale. Such is addressed by the *pchip* method using appropriate derivatives at control points, guaranteeing the properties of positivity, monotonicity and convexity. These are crucial for a principled mapping of domain preferences.

The main caveat for this approach rises when domain knowledge is unavailable or non-existent. The following section proposes a non-parametric method to obtain relevance functions based on a target variable distribution automatically, assuming extreme values as the most important to anticipate.

### 3.3 Automatic and non-parametric relevance functions

Lack of sufficient domain knowledge to define precise control points is a common situation. To obtain relevance functions in such conditions, we require an automatic approach that determines which target values have minimum and maximum relevance. Given that the most extreme values of the distribution are considered the most important to predict accurately, these should have maximum relevance. On the contrary, the most common and well-represented values of the distribution should have minimum relevance.

Determining which values of a distribution should be considered extreme is a long-standing topic in the literature, with an emphasis in outlier/anomaly detection and extreme value analysis. From an unsupervised perspective, there are several approaches to tackle this issue (Chandola et al. 2009). Such includes statistical-based, proximity-based, using notions of distance or density, and clustering-based detection techniques. Given that our problem is univariate, the most direct approach to reaching our goal is to resort to distribution-based analysis, i.e. statistical techniques. These are broadly divided into parametric and non-parametric approaches. The former include approaches based on Gaussian or a mixture of parametric distributions. The latter is commonly based on solutions using histograms and kernel functions. For a thorough analysis and discussion of outlier/extreme value detection methods, we point to several contributions such as the work of Aggarwal (2013), Chandola et al. (2009) and Hodge and Austin (2004).

Tukey's boxplot rule (1970) is one of the most frequently used statistical graphic methods (Wickham and Stryjewski 2012) for depicting data from continuous distributions. This method illustrates information concerning the location, spread, skewness and tails of the distribution. It uses a box to represent the interquartile range (IQR) and two whiskers that define the fences of the boxplot, based on the IQR. This interval frames average values; when outside such interval, values are considered probable outliers—the target cases in imbalanced regression. Nonetheless, the standard rule proposed by Tukey assumes a normal distribution of data points, i.e. distribution symmetry. When learning with skewed and asymmetrical distributions, this rule is prone to erroneously classifying specific data points as being outliers (Hoaglin et al. 1983). One can apply a transformation, e.g. logarithm, to make the data distribution symmetrical and detect outliers afterwards. However, finding a robust transformation to symmetry with application to all distributions is far from trivial. An adaptation to Tukey's original proposal (Tukey 1970) is the adjusted boxplot, proposed by Hubert and Vandervieren (2008). The objective is to correct the symmetry issue using a robust measure of skewness when determining the fences of the boxplot, i.e. limits for values considered normal.

The interval proposed by Tukey (1970) for determining outliers cutoff values is  $[Q_1 - 1.5 IQR, Q_3 + 1.5 IQR]$  where  $Q_1$  and  $Q_3$  are the first and third quartile, respectively, and  $IQR = Q_3 - Q_1$  is the interquartile range. To make this interval less prone to bias, Hubert and Vandervieren (2008) propose to incorporate the medcouple, introduced by Brys et al. (2004), into the definition of the whiskers. The medcouple is a robust alternative to the classical skewness coefficient, based on variance and skewness of data. It is location and scale-invariant, and defined as

$$MC = \underset{x_i \leq Q_2 \leq x_j}{med} h(x_i, x_j), \quad (1)$$

where  $Q_2$  is the second quartile (median) and for all  $x_i \neq x_j$  the kernel function  $h$  is given by

$$h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i} \quad (2)$$

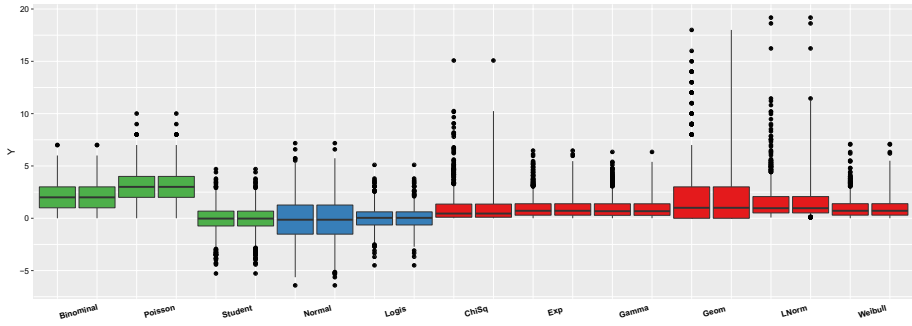
According to the value of  $MC$ , the following intervals will mark points outside them as potential outliers:

- if  $MC \geq 0$ , then the interval is  $[Q_1 - 1.5 e^{-4MC} IQR, Q_3 + 1.5 e^{3MC} IQR]$ ;
- if  $MC < 0$ , then the interval is  $[Q_1 - 1.5 e^{-3MC} IQR, Q_3 + 1.5 e^{4MC} IQR]$ .

According to the study carried by Hubert and Vandervieren (2008), the use of such exponential functions, allows the boxplot to be more skewness-adjusted as the fences might be asymmetric around the box.

Given the context of imbalanced regression tasks, the adjusted boxplot method presents a better alternative for two main reasons. First, it is non-parametric, therefore more flexible to underlying distributions. Second, by using a robust measure of skewness, the method is better suited to avoid missing real cases of extreme values (outliers).

To illustrate the difference between the two types of boxplots, in Fig. 5 we depict the Tukey's boxplot and the adjusted boxplot for a set of 1000 artificial generated values from different theoretical distributions. As it is possible to observe, for *Binomial*, *Logistic* and *Poisson* distributions there is almost no difference (green). Then, we observe that symmetric distributions such as *Normal* and *Student t* distributions present a slight difference



**Fig. 5** A comparison of the standard Tukey’s boxplot (left) and the adjusted boxplot (right) for a set of artificial samples from symmetric and skewed theoretical distributions, and their degree of dissimilarity: none (green), residual (blue) and considerable (red) (Color figure online)

**Table 2** Control points for high LNO2 extreme values inferred by the automatic and non-parametric approach to relevance functions based on the adjusted boxplot

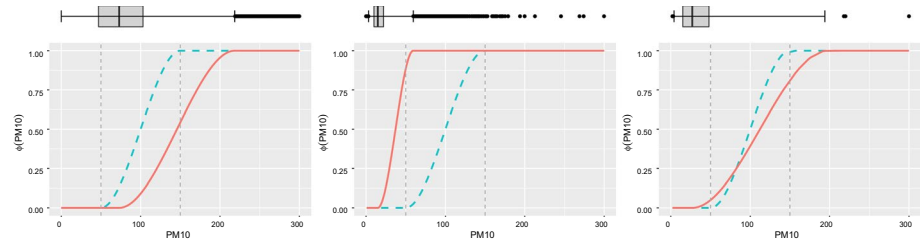
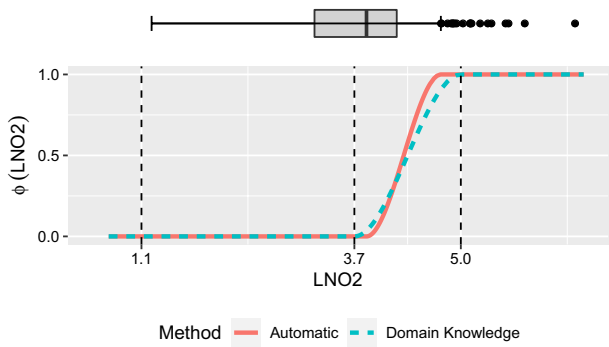
$y_k$ : LNO2 concentration values		$\varphi(y_k)$	$\varphi'(y_k)$
Lower adjacent value:	$\ln(3.4 \mu\text{g}/\text{m}^3) \approx 1.2$	0.0	0.0
Median:	$\ln(46.9 \mu\text{g}/\text{m}^3) \approx 3.8$	0.0	0.0
Upper adjacent value:	$\ln(116.3 \mu\text{g}/\text{m}^3) \approx 4.8$	1.0	0.0

between the two boxplots (blue). Finally, we have heavily skewed distributions (red), namely  $\chi^2$ , Exponential, Gamma, Geometric, LogNormal and Weibull distributions. For these distributions, the difference between Tukey’s boxplot and adjusted boxplot becomes more evident: the number of extreme values (outliers) identified by the latter is much smaller in comparison to the former. Such confirms that the adjusted boxplot rule is more appropriate for automatic outlier detection when dismissing any assumption concerning the distribution of the data.

We propose the use of adjusted boxplot to automatically supply the control points, based on the methodology presented by Ribeiro (2011)—initially designed to handle Tukey’s boxplot supplied control points. This method fulfils the objective of obtaining a continuous relevance function that maps the domain of the target variable  $Y$  to the relevance interval  $[0, 1]$  so that the extreme values of  $Y$  are assigned maximum relevance. As such, the upper and lower adjacent values are considered threshold values for extremes. Also, the median value of  $Y$  is considered as a centrality value of irrelevance. Three points compose the set of control points: the median value of  $Y$  with relevance value equal to zero and the upper and lower adjacent values with relevance value equal to one.<sup>3</sup> All these control points are assumed to have a derivative of zero so that they represent local maximum and minimum of the relevance function. The *pchip* interpolation method (cf. Algorithm 1) receives this set of control points and derives an extreme-based relevance function  $\phi()$ .

<sup>3</sup> By default we assume that both high and low extreme values are relevant. In case only one type is relevant, only the corresponding adjacent value has relevance value equal to one.

**Fig. 6** Relevance functions when **a** automatically obtained for the prediction of high extreme of LNO2 concentration values (red), and **b** guided by domain knowledge—Directive 2008/50/EC values (blue) (Color figure online)



**Fig. 7** Relevance functions using domain-guided value-relevance pairs according to standard international guidelines (blue, dashed) and the non-parametric automatic method proposed (red, solid), with respective adjusted boxplots, for data concerning  $PM_{10}$  concentration levels in Beijing (China) (Zheng et al. 2013), rural background stations in Germany (Pebesma 2012) and a station in Alnabru, Oslo (Norway) (Aldrin 2006). Left and right vertical dashed lines represent official values considered normal ( $50 \mu\text{g}/\text{m}^3$ ) and dangerous ( $150 \mu\text{g}/\text{m}^3$ ) for 24-h averages (Color figure online)

Resorting to the air pollution scenario, one can use the referred method to define the set of control points automatically,<sup>4</sup> as shown in Table 2. Using such set of control points and the *pchip* interpolation method, the obtained relevance function  $\phi()$  is depicted (red) in Fig. 6. It is interesting to notice that, regarding the most critical values in this particular data set, the proposed automatic method (red) obtains a relevance function similar to the relevance function obtained by the established guidelines based on Directive 2008/50/EC (blue).

### 3.4 Discussion on relevance functions

In the previous section, we presented an automatic and non-parametric method to obtain the set of control points to build the relevance function, based on the adjusted boxplot. The question one may ask is whether these automatically induced relevance functions meet real-world domain preferences.

We resort to the domain of air pollution and the indicator  $PM_{10}$ , using three publicly available data sets. First, concerning hourly averages of concentration levels in Beijing

<sup>4</sup> In this particular example, the focus is only on high extreme values, as low extreme values have no harmful impact on human health.

**Table 3** Artificial scenario with the predictions of two models  $M_1$  and  $M_2$ , for the same set of true  $LNO_2$  values and their respective loss values

True $LNO_2$	2.71	3.35	3.36	3.63	4.08	4.16	4.31	5.55	5.78	6.40
$M_1$	2.68	3.30	3.43	3.72	3.96	4.29	4.55	5.91	7.03	4.72
$M_1$ Loss	<b>0.03</b>	<b>0.05</b>	<b>0.07</b>	<b>0.09</b>	<b>0.12</b>	<b>0.13</b>	<b>0.24</b>	<b>0.37</b>	<b>1.25</b>	<b>1.67</b>
$M_2$	1.04	4.61	3.73	3.87	4.21	4.04	4.41	5.62	5.73	6.37
$M_2$ Loss	<b>1.67</b>	<b>1.25</b>	<b>0.37</b>	<b>0.24</b>	<b>0.13</b>	<b>0.12</b>	<b>0.09</b>	<b>0.07</b>	<b>0.05</b>	<b>0.03</b>

(China) from August 2012 to March 2013 (Zheng et al. 2013). Second, daily averages for rural background stations in Germany from 1998 to 2009 (Pebesma 2012). Third, hourly values for a station in Alnabru, Oslo (Norway), between October 2001 and August 2003 (Aldrin 2006).

We base the set of control points used on official recommendations per the Organization (2005) for denoting 24-h averages as normal or dangerous:  $\phi(50\mu\text{g}/\text{m}) = 0$  and  $\phi(150\mu\text{g}/\text{m}) = 1$ , respectively. The baseline (blue, dashed) and the automatic and non-parametric (red, solid) relevance functions for each data set are presented in Fig. 7. We also include an illustration of the adjusted boxplot.<sup>5</sup>

Our proposal to automatically obtain relevance functions on information is based on control points relayed by the adjusted boxplot. Therefore, its derivation is solely dependent on the underlying distribution of the data sample. By comparison, results show that without the introduction of domain knowledge, automatic relevance functions cannot approximate such knowledge naturally. This might only occur when the control points derived by the sample distribution approximate those of domain knowledge.

In effect, the ideal scenario is to have access to domain knowledge that would allow the relevance function to express as closely as possible the reality. As that is often hard to obtain, we present a non-parametric alternative to automatically induce a relevance function based on the target variable sample distribution. We assume that extreme values are the most relevant ones. The similarity between the relevance functions obtained by these two methods is dependent on the representativeness of our data sample concerning the domain information. If that is the case, then the boxplot-induced and the domain provided control points would be identical, and both relevance functions will be similar; we observe this in the example provided in Sect. 3.3 (Fig. 6). Otherwise, the relevance function will be specific to the data sample distribution. Using the case of air pollution, this means that concentration values considered relevant by the automatic method rely solely on sample values rather than WHO reference values. They represent a contextual (sample-based) notion of relevance.

## 4 Evaluation metrics

In regression tasks, researchers commonly resort to standard metrics such as the Mean Squared Error ( $MSE$ ). These metrics assume uniform domain preferences, focusing solely on the magnitude of the prediction error. As such, they raise several issues when evaluating

<sup>5</sup> We excluded values  $PM_{10} > 300$  from the figure on the left for scaling and comparison purposes.

imbalanced domain learning tasks (Torgo et al. 2013). To demonstrate one of such potential problems, Table 3 describes a predictive modelling scenario using synthetically generated data for  $NO_2$  emissions. It contains the predictions of two models ( $M_1$  and  $M_2$ ) for the same set of true values.<sup>6</sup>

Results show that model  $M_1$  is more accurate concerning lower values of the data, and model  $M_2$  is more precise at higher values. However, standard metrics such as Mean Squared Error and Mean Absolute Deviation ( $MSE$  and  $MAD$ , respectively), report no difference between these two models: a score 0.460 for  $MSE$  and 0.402 for  $MAD$ . Such occurs because the overall magnitude of errors is equal, and such metrics consider all domain values equally relevant.

#### 4.1 Alternative evaluation metrics to standard error metrics

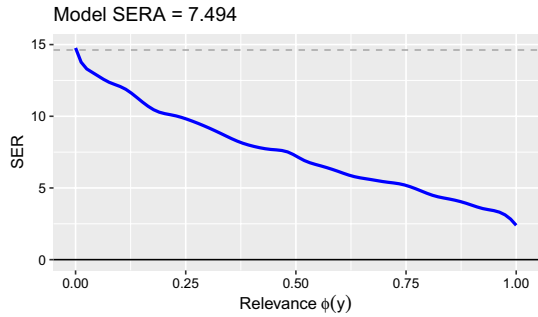
Several authors have proposed alternative evaluation metrics to account for non-uniform domain preferences. In finance, (Christoffersen and Diebold 1996) proposed the *LIN-LIN* error metric aiming at distinguishing prediction errors depending on them being under- or over-predictions. By introducing a different penalisation (weight) to these two cases, this metric includes an asymmetric notion of error. Based on such notion, other metrics were proposed (Zellner 1986; Cain and Janssen 1995; Christoffersen and Diebold 1996, 1997; Granger 1999; Crone et al. 2005; Lee 2007) combining linear, quadratic or exponential costs. Also, the proposal of ROC space for regression (RROC) by Hernández-Orallo (2013) plots the total over-estimation and under-estimation error of models in  $X$ -axis and  $Y$ -axis, respectively. As in the context of classification, this graphical metric enables the inference of dominance when analysing multiple models, providing an important tool of performance analysis. However, distinguishing under- and over-predictions is not sufficient to adequately evaluate imbalanced regression tasks. Regardless of their relevance, errors of the same magnitude are considered equal—all cases are assumed equally relevant.

The Regression Error Characteristic (REC) curves, proposed by Bi and Bennett (2003), depict the cumulative distribution of models' prediction errors. The authors use the notion of error tolerance ( $X$ -axis) and accuracy ( $Y$ -axis), translating to the percentage of cases with a prediction error smaller (or equal to) a given tolerance  $\epsilon$ . However, REC curves do not account for non-uniform domain preferences, allowing for errors with the same magnitude—but different relevance—to be equally considered. Torgo (2005) proposed an extension to REC curves: the Regression Error Characteristic Surfaces (RECS). This graphical metric includes an additional dimension to plot the cumulative distribution of the target variable. As such, it depicts the prediction errors across the domain of the target variable. The importance of this proposal lies in allowing the study of models' ability in predicting ranges of extreme target values. Nonetheless, a proper illustration of the link between the magnitude of the error and the relevance of cases is not clear.

Finally, Ribeiro (2011) describes a utility-based precision/recall evaluation framework. This proposal focuses on the ability of models in accurately predicting cases with high relevance. It analyses the usefulness of predictions as a function of numeric prediction error and the relevance of both predicted and true values. However, the proposal has two caveats. First, it requires an ad-hoc relevance threshold. Second, it does not account for the predictive ability of models in cases where both the relevance of the predicted and true values are

<sup>6</sup> Example based on previous work by Ribeiro (2011).

**Fig. 8** An example of the squared error-relevance area (SERA) metric for an artificial model, based on the integration of Squared Error-Relevance ( $SER_t$ ) for cutoff relevance  $\phi(\cdot)$  values  $t$ . The grey dashed line depicts the sum of squared errors for all cases (Color figure online)



below the mentioned threshold. This issue is equivalent to that of F-Measure (Rijsbergen 1979) in classification tasks. The combination of such caveats allows the optimisation of models that may neglect the impact of prediction cases with low relevance—i.e. majority of cases. In turn, this could lead to naive extreme models with no discernible capability of generalisation.

Based on the review of previous work, we propose a new evaluation metric to overcome the challenges posed to the evaluation of imbalanced regression tasks. Such a metric must encompass critical characteristics, such as:

1. focus on minimising prediction errors in cases with extreme target values, i.e. high relevance, by countering the dominance of low relevance cases;
2. ability to prevent over-fitting of models, biased to predicting extreme (or near extreme) target values and disregarding all other cases;
3. allow for an asymmetric notion of loss, i.e. errors of equal magnitude have different impacts depending on their relevance;
4. allowing model discrimination, comparison and dominance analysis.

### 4.2 Squared error-relevance area (SERA)

Consider a data set  $\mathcal{D} = \{\langle x_i, y_i \rangle\}_{i=1}^N$  and a relevance function  $\phi : \mathcal{Y} \rightarrow \{0, 1\}$  defined for the target variable  $Y$ . We define the subset  $\mathcal{D}^t \subseteq \mathcal{D}$  formed by the cases for which the relevance of the target value is above or equal a cutoff  $t$ , i.e.  $\mathcal{D}^t = \{\langle x_i, y_i \rangle \in \mathcal{D} \mid \phi(y_i) \geq t\}$ . We can obtain an estimate of the Squared Error-Relevance of a model with respect to a cutoff  $t$  ( $SER_t$ ), as follows,

$$SER_t = \sum_{i \in \mathcal{D}^t} (\hat{y}_i - y_i)^2 \tag{3}$$

where  $\hat{y}_i$  and  $y_i$  are the predicted and true values for case  $i$ , respectively. For this estimate, only the subset of predictions composed by the cases  $i \in \mathcal{D}^t$ , for which the relevance of the true target value is above a specific cutoff  $t$ , are considered.

Given the bounds of relevance values— $\phi(y) \in [0, 1]$ , we may represent a curve, where each point represents the value of  $SER_t$  for a possible relevance cutoff  $t$ . This curve has interesting properties. The highest and the lowest value of  $SER_t$  are attained when all ( $t = 0$ ) or only the most relevant cases ( $t = 1$ ) are included, respectively. Additionally, for any  $\delta \in \mathbb{R}^+$  such that  $t + \delta \leq 1$ , we have that  $SER_t \geq SER_{t+\delta}$ , given that  $SER_{t+\delta}$  considers



a subset (or all) of the cases included in  $SER_t$ . These properties ensure that the curve is decreasing and monotonic.

In this paper, we propose the Squared Error-Relevance Area ( $SERA$ ). It represents the area below the  $SER_t$  curve, obtained via integration<sup>7</sup> (cf. Eq. 4), illustrated in Fig. 8.

$$SERA = \int_0^1 SER_t dt = \int_0^1 \sum_{i \in \mathcal{D}'} (\hat{y}_i - y_i)^2 dt \quad (4)$$

It is important to note that the  $SER_t$  curve provides an overview of the magnitude of the prediction errors in the domain, along with different relevance cutoff values. Thus, the smaller is the area under this curve ( $SERA$ ), the better is the model. Also, we should note that, when uniform preferences are assumed,  $\phi(\mathcal{Y}) = 1$ ,  $SERA$  is equivalent to the sum of squared errors.

**Optimisation of  $SERA$ .** To optimise the squared error, we must find the constant that minimises it. Given the target variable domain  $\mathcal{Y}$ , we know that the squared loss function is differentiable for every predicted value in that domain. Likewise,  $SER_t$  is also differentiable w.r.t. the predicted value: the constant  $m_t$  that minimises  $SER_t$  is the average of true target values, concerning only the target values whose relevance values are equal or above the specified cutoff  $t$ . This is shown in Eq. 5 (cf. proof of Theorem 1 in “Appendix”).

$$m_t = \frac{\sum_{i \in \mathcal{D}'} y_i}{|\mathcal{D}'|} \quad (5)$$

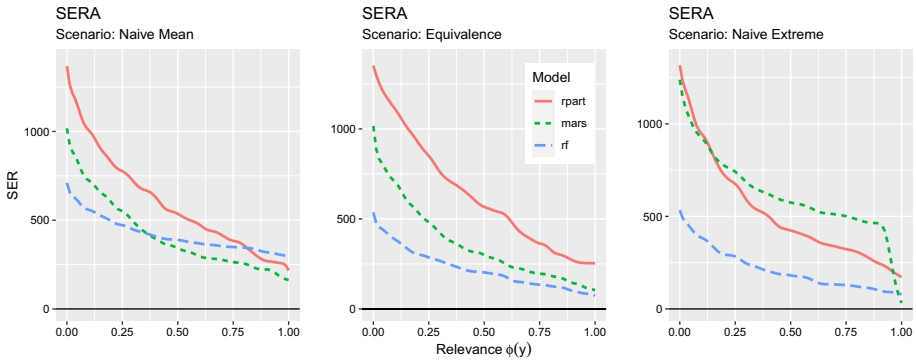
Additionally, we should note that it is also possible to find the constant that minimises  $SERA$ .  $SERA$  corresponds to the integration of  $SER_t$  over the  $[0, 1]$  relevance interval. Given that  $SER_t$  is differentiable, by applying the Fundamental Theorem of Calculus,  $SERA$  is also differentiable. The constant  $m$  that minimises  $SERA$  is given by Eq. 6 (cf. proof of Theorem 2 in “Appendix”).

$$m = \frac{\int_0^1 \sum_{i \in \mathcal{D}'} y_i dt}{\int_0^1 |\mathcal{D}'| dt} \quad (6)$$

#### 4.2.1 Analysis

Although our goal is mainly to estimate the effectiveness of models in predicting extreme values,  $SERA$  does not entirely discard the impact of models' performance in cases with average target values. Prediction errors made in lower relevance cases have much less impact than those in highly relevant cases. The errors of the latter are counted more times along the relevance cutoff values in the overall sum that constitutes each point the curve. Therefore,  $SERA$  encompasses previously mentioned characteristics for an appropriate evaluation metric in imbalanced regression tasks (characteristics 1–3, Sect. 4.1): the reduction of prediction errors in extreme target values via model optimisation with an asymmetric notion of loss while preventing over-fitting of models towards highly relevant cases.

<sup>7</sup> Riemann sums with the trapezoidal rule is used, with a default delta of 0.001.



**Fig. 9** Visual comparison of three prediction models from different learning algorithms in three distinct and common evaluation scenarios in imbalanced regression: (i) naive mean, (ii) equivalence between the *MSE* and *SERA* evaluation metrics, and (iii) naive extreme

**Table 4** *MSE* and *SERA* estimates for the evaluation scenarios depicted in Fig. 9. The best scores in each scenario, and for each evaluation metric, are denoted in bold

Model	Naive Mean		Equivalence		Naive Extreme	
	<i>MSE</i>	<i>SERA</i>	<i>MSE</i>	<i>SERA</i>	<i>MSE</i>	<i>SERA</i>
rpart	1392.8	596.8	1373.1	630.1	1335.6	515.4
mars	1033.2	<b>414.1</b>	1026.9	357.9	1252.7	616.8
rf	<b>722.8</b>	416.3	<b>548.9</b>	<b>220.2</b>	<b>540.4</b>	<b>210.9</b>

Building on the graphical dimension of this metric, *SERA* is capable of allowing model discrimination, comparison and dominance analysis (characteristic 4, Sect. 4.1), as illustrated in Fig. 9. We use different learning algorithms available in **R** packages: random forests (**rf**) (Wright and Ziegler 2017), CART regression tree (**rpart**) (Therneau and Atkinson 2018) and multiple adaptive regression splines (**mars**) (Milborrow 2019), all with the default parameters. The figure depicts a comparison between three models from these learning algorithms, using a standard evaluation metric (*MSE*) and *SERA*, in three distinct and common scenarios in imbalanced regression: (i) naive mean, (ii) equivalence, and (iii) naive extreme. The first and the third represent configurations where the best model according to *MSE* is biased towards predicting cases with an average value (naive mean), or a model is biased towards predicting all cases with extreme (or near-extreme) values (naive extreme). The second (equivalence) depicts a scenario where conclusions of *MSE* and *SERA* agree, w.r.t the rank of models. Individual model scores are described in Table 4.

For the naive mean scenario (left), the **rf** model obtains the best *MSE* score but a worse *SERA* score when compared to the **mars** model. Based on the graphical capabilities of *SERA* for dominance analysis, although the **rf** model presents a lower overall prediction error (value at relevance 0), it shows low ability to correctly model cases with extreme values, i.e. the majority of prediction errors for this model. Such is illustrated further by its low slope across the domain w.r.t  $\phi(\cdot)$ . The **mars** model shows a predictive advantage for cases with relevance greater than 0.3. Given our objective to accurately predict target values of highly relevant cases, this shows how *MSE* scores may be misleading in such context.

Concerning naive extreme scenarios (right), results show that the **rf** model is the best model for both *MSE* and *SERA* metrics. Nevertheless, results concerning **rpart** and **mars** models are contradictory: although the former presents a worse *MSE* score, it shows an advantage concerning *SERA*. By analysing *SERA*, we observe that although the **mars** model demonstrates lower prediction errors concerning the most extreme values, it also achieves considerably higher levels of prediction error for the remainder of the domain, showing a considerable bias towards accurately predicting extreme values at the cost of low representation of the average behaviour of data. As such, optimisation with the *SERA* metric provides the ability to prevent over-fitting of models biased to predict extreme (or near extreme) target values.

Regarding the scenario of equivalence, this illustrates an agreement between the *MSE* and *SERA* metrics as to the rank of the **rf**, **mars** and **rpart** models.

Concluding, we show that *SERA* allows a thorough analysis of prediction models w.r.t prediction errors in varying levels of relevance, as well as dominance analysis. This combination of characteristics results in a significant contribution to the evaluation of imbalanced regression tasks, allowing an assessment focused on the predictive ability of models towards extreme values.

## 5 Experimental study

We present a broad experimental study focusing on the problem of imbalanced regression. First, an experimental evaluation is performed over a diverse list of data sets, using multiple learning algorithms. The objective is to compare the selection of models using a standard evaluation metric and the proposed metric *SERA*. Second, we provide an analysis of results concerning the impact of using *SERA* as a preference criterion for the optimisation of models in imbalanced regression tasks. Given such objectives, our goal is to answer the following research questions.

- RQ1** What is the impact on models performance when using standard evaluation metrics for model selection, in comparison to using the proposed metric *SERA*?
- RQ2** Which are the predictive trade-offs associated with models performance estimation using *SERA* as a criterion?
- RQ3** Is the *SERA* metric appropriate for model optimisation processes when our goal is to improve the prediction of extreme values?

### 5.1 Data

In our experimental study, we used a diverse group of regression data sets from different domains.<sup>8</sup> First, we collected a considerable number of data sets from public repositories. Then, we applied the adjusted box-plot method (previously described) to assess if the distribution of target values in each data set demonstrated the existence of extreme values. The data sets used in this experimental study are those that demonstrated the existence of such extreme values. Table 5 describes their characteristics. For each data set, the description includes information on the number of instances and the number of nominal and

<sup>8</sup> Data sets are made available in <https://github.com/nunompmniz/IRon>.

**Table 5** Datasets ( $\mathcal{D}$ ) used in the experimental study with no. of instances ( $|\mathcal{D}|$ ); nr. of nominal ( $Nom$ ) and numerical ( $Num$ ) variables; nr. of instances with maximum ( $|\mathcal{D}^1|$ ) and non-maximum ( $|\mathcal{D}\setminus\mathcal{D}^1|$ ) values of relevance; Imbalance Ratio ( $IR$ ); and *Type* of extremes according to the adjusted boxplot: upper (U), lower (L) or both (B)

#	Datasets ( $\mathcal{D}$ )	$ \mathcal{D} $	$Nom$	$Num$	$ \mathcal{D}\setminus\mathcal{D}^1 $	$ \mathcal{D}^1 $	$IR$	<i>Type</i>
1	diabetes	43	0	2	39	4	9.75	U
2	triazines	186	0	60	181	5	36.20	B
3	a7	198	3	8	187	11	17.00	U
4	elecLen1	495	0	2	489	6	81.50	U
5	housingBoston	506	0	13	455	51	8.92	B
6	forestFires	517	0	12	508	9	56.44	U
7	strikes	625	0	6	620	5	124.00	U
8	mortgage	1049	0	15	971	78	12.45	L
9	treasury	1049	0	15	953	96	9.93	L
10	musicorigin	1059	0	117	1043	16	65.19	B
11	airfoild	1503	0	5	1490	13	114.62	U
12	acceleration	1732	3	11	1702	30	56.73	B
13	fuelConsumption	1764	12	25	1725	39	44.23	B
14	availablePower	1802	7	8	1712	90	19.02	B
15	maxTorque	1802	13	19	1747	55	31.76	B
16	debutenizer	2394	0	7	2278	116	19.64	U
17	space_ga	3107	0	6	3077	30	102.57	B
18	pollen	3848	0	4	3811	37	103.00	B
19	abalone	4177	1	7	3693	484	7.63	B
20	wine	6497	0	11	5220	1277	4.09	U
21	deltaAilerons	7129	0	5	6459	670	9.64	B
22	heat	7400	3	8	7351	49	150.02	B
23	cpuAct	8192	0	21	7898	294	26.86	L
24	kinematics8fh	8192	0	8	8125	67	121.27	B
25	kinematics32fh	8192	0	32	8121	71	114.38	B
26	pumaRobot	8192	0	32	8070	122	66.15	B
27	deltaElevation	9517	0	6	7265	2252	3.23	U
28	sulfur1	10,081	0	5	9518	563	16.91	B
29	sulfur2	10,081	0	5	9340	741	12.60	B
30	ailerons	13,750	0	40	13,515	235	57.51	B
31	elevators	16,599	0	17	14,589	2010	7.26	B
32	calHousing	20,640	0	8	20,613	27	763.44	L
33	house8H	22,784	0	8	22,387	397	56.39	B
34	house16H	22,784	0	16	22,387	397	56.39	B

numeric variables. Also, we provide information regarding the number of instances with maximum,  $|\mathcal{D}^1|$  ( $\phi(y) = 1$ ), and non-maximum relevance,  $|\mathcal{D}\setminus\mathcal{D}^1|$  ( $\phi(y) < 1$ ). Such information is obtained using the proposed approach for non-parametric auto-generated relevance functions (see Sect. 2). The imbalance ratio ( $IR$ ) is calculated as  $\frac{|\mathcal{D}\setminus\mathcal{D}^1|}{|\mathcal{D}^1|}$ .

**Table 6** Models parameters considered for grid search in the experimental study

Algorithm	Parameters	R Package
rpart	minsplit $\in$ {10, 20, 30} cp $\in$ {0.001, 0.005, 0.01}	<i>rpart</i> (Therneau and Atkinson 2018)
mars	thresh $\in$ {0.001, 0.005, 0.01}	<i>earth</i> (Milborrow 2019)
svm	kernel $\in$ {linear, polynomial, radial, sigmoid} epsilon $\in$ {0.1, 0.05, 0.01}	<i>e1071</i> (Meyer et al. 2019)
rf	ntrees $\in$ {100, 250, 500}	<i>ranger</i> (Wright and Ziegler 2017)
bagging	nbag $\in$ {25, 50, 100}	<i>ipred</i> (Peters and Hothorn 2018)

## 5.2 Methods

In this section, we describe the learning and evaluation methods used in the experimental study. We also present information concerning parametrisation for reproducibility purposes.

The following learning algorithms are used: CART regression trees (**rpart**), multivariate adaptive regression splines (**mars**), support vector machines (**svm**), random forests (**rf**) and **bagging**, based on implementations in **R** packages.

Parametrisation is carried out using a grid search approach, using the values in Table 6. Each model is identified by a number corresponding to the respective combination of parameters. For example, **rpart**<sub>1</sub> corresponds to parameters *minsplit* and *cp* with respective values 10 and 0.001, and **rpart**<sub>2</sub> with values 10 and 0.005. Experimental results are presented using evaluation metrics *MSE* and *SERA*, estimated using a 2 × 5-fold cross validation evaluation methodology.

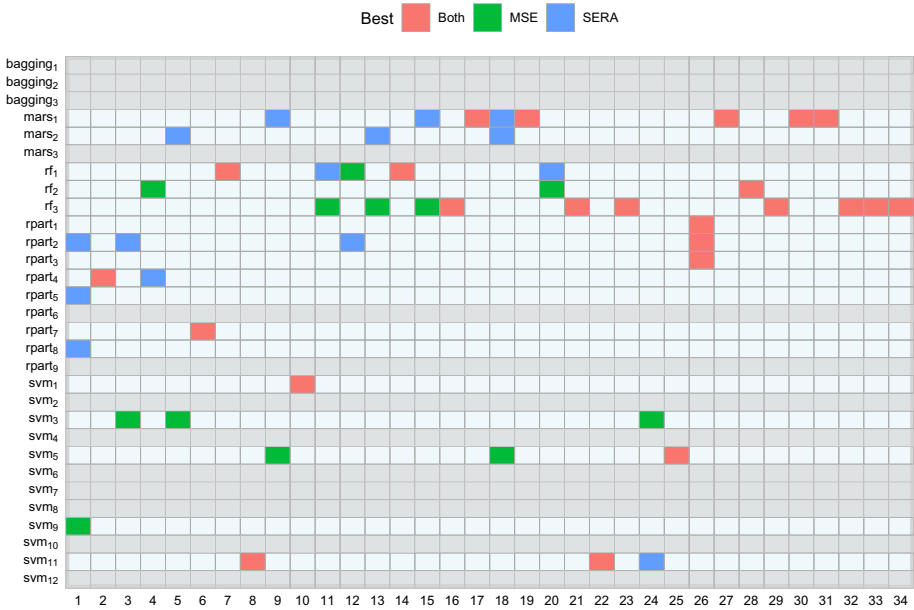
## 5.3 Results: model selection

In Sect. 4.2, the *SERA* evaluation metric is proposed, and the shortcomings of standard metrics, such as *MSE*, in imbalanced regression tasks are described. In this section, an experimental evaluation is carried out to understand the impact of assessing the performance of models when using either of these two metrics.

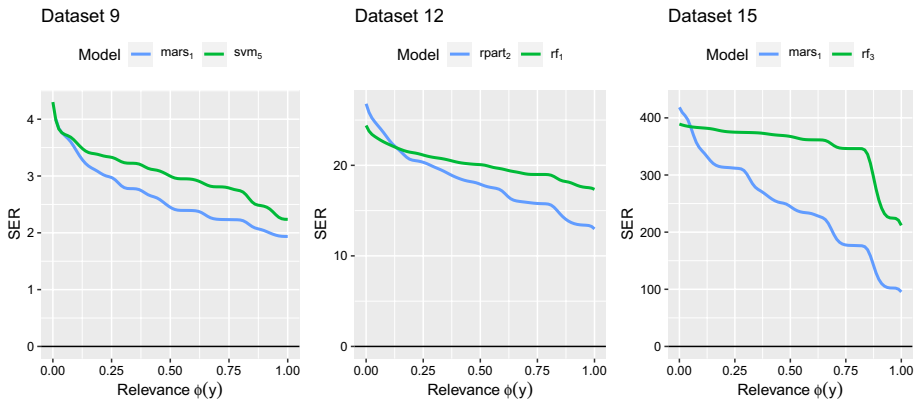
We employ the following methodology:

1. for each data set (Table 5), we split the data into train and test sets using a 70%/30% random partition of cases;
2. each combination of learning algorithm/parameter configuration (Table 6) is used to create a model—30 models for each data set;
3. models are evaluated with the test set using *MSE* and *SERA* metrics;
4. for each data set, we select the models that provide the best approximation according to each evaluation metric used; these are denoted as “oracles”.

Our objective is to analyse how the different evaluation criteria impact the selection process of learning algorithms and their respective parameterisation. In Fig. 10, we present the results obtained from applying the methodology described, which indicates for each

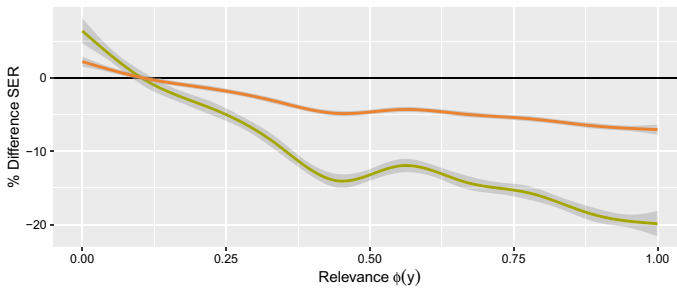


**Fig. 10** Best model (oracle) for all 34 data sets of the experimental evaluation, according to *MSE* and/or *SERA* metrics, using the grid search described in Table 6



**Fig. 11** *SER* curves of the best model according to the *MSE* metric (green) and the best model according to the *SERA* metric (blue) in datasets 9, 12 and 15 (Color figure online)

data set (column), the best models according to the evaluation metrics *SERA* (blue) and *MSE* (green). If both metrics select the same model, these are denoted in red. From the 30 models created for each dataset, 17 were selected by either evaluation metric in at least one dataset. Models that were never selected are marked with a grey background. The subscript of the models represents the different parameter configurations of each learning algorithm, as described in Sect. 5.2 (Table 6).



**Fig. 12** Average percentage difference between the best models (oracles) according to either *MSE* or *SERA* metrics for: (i) all data sets (orange) (ii) data sets where different models were selected by the two metrics (green) (Color figure online)

Results show that the evaluation metrics *MSE* and *SERA* selected different models in 12 of the 34 data sets. In 9 of such cases (26% overall), the best prediction models according to the mentioned metrics belong to different learning algorithms; in 3 cases (9%), the models belong to the same algorithm while using different parameter settings. Based on these results, we observe the discrepancy of outcomes when assessing model performance with these two evaluation metrics. Nonetheless, such analysis does not adequately illustrate the impact in predictive performance of those models, in the context of imbalanced regression tasks. With such aim, Fig. 11 provides a depiction of the *SER* curves for data sets 9, 12 and 15 (Table 5), where *MSE* (green) and *SERA* (blue) metrics select models from different learning algorithms.

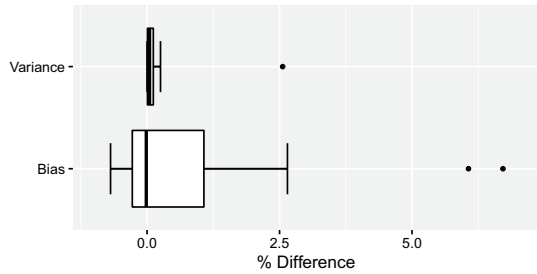
Results show that models selected by the *MSE* metric tend to demonstrate a lower sum of squared errors for the entire domain (observed when relevance is 0). However, such an advantage is mostly due to the high density of cases in the central tendency of the target variable distribution—cases with lower relevance. Once we progressively focus on those with higher relevance, we rapidly observe a trade-off point where models selected by the *SERA* metric provide a considerably better predictive performance. To provide further evidence of this trade-off, Fig. 12 illustrates a smoothed conditional mean of the percentage difference of the *SER* score concerning the oracle (best) model according to *MSE* and the oracle model according to *SERA*. Furthermore, such illustration distinguishes two settings. First, using all the available data sets (orange); second, using the data sets in which the models selected according to *MSE* and *SERA* metrics are different (green). We carry out the computation of the percentage difference of  $SER_t$  at each relevance value  $t$  as follows:

$$\frac{SER_t^s - SER_t^m}{SER_t^m} * 100 \quad (7)$$

where  $SER_t^s$  and  $SER_t^m$  represent the *SER* score of the oracle according to *SERA* and the oracle according to *MSE*, respectively.

Based on this illustration, we are capable of further understanding the impact of employing the *SERA* metric in the context of imbalanced regression tasks. In comparison to those selected by the *MSE* metric, the best models, according to *SERA*, are considerably more able to reduce prediction loss in cases with higher relevance. Also, we observe that such ability does not come at the cost of significant bias for extreme values: we observe a favourable trade-off towards the best models according to *SERA* for cases with relevance above 0.1, approximately. Also, we observe that such conclusion is valid both when focusing on

**Fig. 13** Distribution of the percentage difference, regarding bias and variance error decomposition, between the best model according to *MSE* and to *SERA* metrics for all 14 data sets where the selected models are different



data sets where the *MSE* and *SERA* metrics select different models, but also when accounting for all data sets—although with different magnitude. However, this comes at a cost for cases in the centre of the distribution—slight decrease in predictive accuracy. Nonetheless, results show two important conclusions. First, it is possible to improve predictive accuracy towards extreme target values by selecting models using *SERA*. Second, this does not require a significant bias of prediction models towards the naive mean or extreme scenarios (**RQ1**). For clarity, we should highlight the magnitude of the trade-off between the best models according to *MSE* and *SERA*. As such, we analyse the mean squared error decomposition of the best models according to *MSE* and *SERA*, using the bias-variance framework described by Geman et al. (1992). We present the results of such study in Fig. 13 for all data sets where the selected models are different.

Results provide interesting insights concerning the normal behaviour of the models. As previously stated, the process of selecting a model that minimises *SERA* will likely increase the *MSE* score—although presenting a far better ability at predicting target values that diverge from the mean of the distribution. Figure 13 shows that this is mostly due to an increase in variance—models that perform better in imbalanced regression tasks are more sensitive towards extreme values. However, we observe a considerable fluctuation of bias: despite a slight increase on average, results show that bias is reduced in 57.1% of the data sets analysed (8 data sets), in contrast with the variance—1 data set (7.1% of cases). Overall, results confirm that the predictive ability of the best models according to *SERA* present a slight decrease in performance for low relevance cases. In contrast, they present a significant increase in the ability to anticipate target values for cases with higher relevance (**RQ2**)—the goal in imbalanced regression tasks.

## 5.4 Results: model optimisation

In the previous section, we compared the predictive performance of the best prediction models (oracles) when selected by either *MSE* or the proposed *SERA* metric. In this section, we assess the ability of the *SERA* metric when used for model optimisation. Specifically, we address the problem of optimising the parametrisation of learning algorithms, to minimise *SERA*.

The most common approach for optimising a learning algorithm's parameters is using a parameter grid search with a *k*-fold cross-validation methodology. The data set is divided into *k* partitions. An iterative process is applied where we use each partition as a validation set, and the remainder partitions are grouped, forming a train set. In each pair of train and validation sets, we use the train set to learn a prediction model

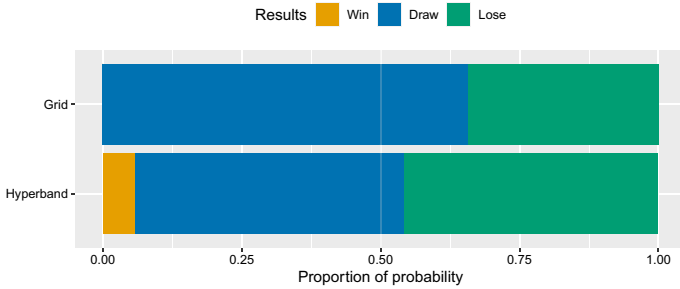


for every combination of parameters in the user-defined grid. We use the validation set to obtain an estimate of the prediction error for each model. The expectation is that the average metric score across the  $k$  validation sets will be a good proxy of the real predictive error—test set. Although it is the most common approach, this is a greedy methodology and therefore, very time consuming and computationally expensive. This limitation is one of the core motivations for automated machine learning (Brazdil et al. 2008) (AutoML), where the goal is to generate automatic recommendations or rankings of predictive solutions.

Research in AutoML has provided multiple search procedures and optimisation algorithms that are capable of fulfilling such goal with reasonable time constraints (Pinto et al. 2017). Examples include the Data Mining Advisor (Giraud-Carrier 2005), a meta-learning approach that relates the characteristics of data sets and the performance of learning algorithms and their respective parametrisation. Bayesian optimisation methods are also well-known for their ability to optimise the parametrisation of learning algorithms efficiently. For example, SMAC (Hutter et al. 2011), a Bayesian optimisation method, maps the relationship between the performance of learning algorithms and their parametrisation. Additionally, we should mention the Hyperband method (Li et al. 2017), a pure-exploration algorithm for multi-armed bandits. Such method approaches the automatic model selection problem as an automatic model evaluation problem. Most importantly, Hyperband has provided evidence of its ability to improve over the results of bayesian optimisation methods. For a thorough review of the state-of-art in AutoML, we refer to the work of He et al. (2019).

In this experimental evaluation, we employ a similar methodology to the one used in Sect. 5.3, as follows.

1. We split each data set (Table 5) into train and test sets using a 70%/30% random partition of cases;
2. In the training data set, we apply two optimisation methods: (i) grid search, as described in Table 6, and (ii) Hyperband. We aim at comparing a greedy approach—grid search—with a method that optimises the parametrisation of algorithms based on the direct minimisation of the *SERA* metric—Hyperband.
  - (a) We apply the grid search method with a 2x5-fold cross-validation methodology. We average the results to obtain an estimation of the best learning algorithm and its parametrisation;
  - (b) We also apply the Hyperband method with a 2x5-fold cross-validation methodology for each learning algorithm. We average the results to obtain an estimation of which learning algorithm is best;
3. We use the outcome of the grid search method (the best combination of learning algorithm-parametrisation) to learn a prediction model. The Hyperband method is applied to the entire train set using the best learning algorithm according to the cross-validation results;
4. Models optimised with grid search and Hyperband methods are used to predict the test set, retrieving an estimation of out-of-sample prediction performance.



**Fig. 14** Proportion of probability concerning wins, draws and losses of models optimised through grid search or Hyperband methods, in comparison to the oracle models, according to the Bayes Sign test applied to the *SERA* scores of the models. ROPE is defined as the interval  $[-1\%, 1\%]$ , concerning the percentage difference between *SERA* scores

We should stress that the test set data is not used in the optimisation process, and is only used for the final prediction of the optimised models and to obtain the *SERA* metric estimates.

After the collection of experimental results using the described methodology, results are analysed using the Bayes Sign Test (Benavoli et al. 2014, 2017). In order to provide such analysis, we require the use of standardised values over the multiple data sets used in this experimental evaluation. As such, we use the oracle models from the previous experimental evaluation as a baseline. These are the best models (oracles) in an out-of-sample evaluation according to either *MSE* or *SERA* metrics (Sect. 5.3). We use such baselines to obtain the percentage difference between the out-of-sample *SERA* score of the models obtained with the grid search and Hyperband methods, and the *SERA* score of the oracle models. We carry out the percentage difference as follows:

$$\frac{SERA_a - SERA_o}{SERA_o} * 100 \tag{8}$$

where  $SERA_a$  and  $SERA_o$  represent the *SERA* score of the model under comparison and the score of the oracle model, respectively.

Given this, we can define the *region of practical equivalence* (ROPE) (Kruschke and Liddell 2015). In the context of Bayesian analysis, practical equivalence means that the probability of the difference of values being inside a specific range can be considered as having virtually no effect. Therefore, the main idea of using ROPE is to define an area around the null value (no difference between predictive solutions) encapsulating values considered equivalent to such null value for practical purposes Kruschke (2015). In classification tasks, the interval  $[-1\%, 1\%]$  for the average difference between accuracy scores is considered a reasonable value for the ROPE (Benavoli et al. 2017), Kruschke (2015). In regression tasks, (Kruschke and Liddell 2017) suggest that this value can be by default set to a range of  $-0.1$  to  $0.1$  of a certain standardised parameter. For thoroughness, we define ROPE to be the interval  $[-1\%, 1\%]$ , given that it is more restrictive. Therefore, based on Eq. 8, we consider that: (i) if the percentage difference of *SERA* scores between model *a* and the oracle model is less than  $-1\%$ , the former outperforms the latter (win); (ii) if the percentage difference is within the interval  $[-1\%, 1\%]$ , they are of practical equivalence (draw); and (iii) if the percentage difference is greater than  $1\%$ , the oracle outperforms model *a* (lose). Figure 14 shows the proportion of probability for

wins, draws and losses against the oracle models, considering all data sets used in the experimental evaluation.

We observe that the optimised models are either of practical equivalence to or outperform the oracle models with more than 50% of probability. In the case of models optimised with grid search, this value is 66%; for models optimised with Hyperband, the value is 54% (6% of wins). Such outcome provides evidence of an essential aspect of the *SERA* metric: its usefulness as a metric for the optimisation of learning algorithms and their parametrisation in the context of imbalanced regression tasks (**RQ3**).

## 6 Discussion

In this section, we discuss our contribution in light of recent work and related topics. Also, we discuss the relationship between metrics *MSE*, *SERA* and the utility-based metric  $F_u$  used in the evaluation of utility-based regression tasks.

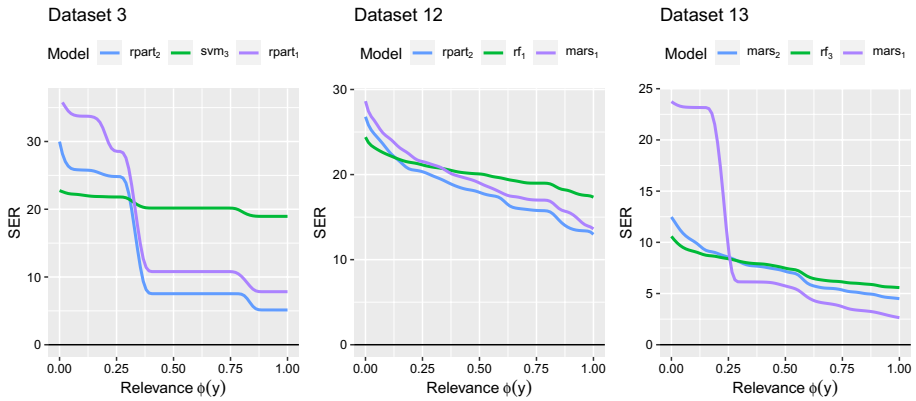
### 6.1 Recent work

Recent contributions address the topic of extreme values, such as the work of Siffer et al. (2017), Ding et al. (2019) and Wang et al. (2019). Based on the formalisation or use of tools from extreme value theory (the first two references) or combinations of pairwise preference classification and ordinal ranking, they illustrate well the issues tackled in this paper. The mentioned works are dependent on the definition of thresholds to distinguish target and non-target cases. As argued, discretisation disregards the magnitude of prediction errors, raising extensively discussed issues (Royston et al. 2006). Also, the contributions mentioned do not tackle the learning assumption of uniform domain preferences, and the evaluation metrics used are debatable. Siffer et al. (2017) use an error rate based on the quantiles of the distribution and a ROC analysis on discretised values. Ding et al. (2019) use the Root Mean Squared Error and the F-Score with discretised values. Wang et al. (2019) use statistical distance metrics for distributions and ROC analysis on discretised values. Such is an illustration of the problem faced by imbalanced regression, as none of the mentioned metrics is appropriate for evaluating learning tasks where domain preferences are non-uniform, and the aim is to anticipate extreme values—an issue addressed by the evaluation metric proposed in our contribution, *SERA*.

### 6.2 Related topics

Relevant discussions are raised when analysing the use of support vector machine regression (Drucker et al. 1996) algorithms or extreme value theory (Goodwin and Wright 2010), providing interesting inputs concerning the impact of our work.

In support vector machine regression, algorithms such as SVR Drucker et al. (1996) are based on hinge loss. This metric assumes the dismissal of errors below a particular value  $\epsilon$ , allowing easier convergence. In this paper, the formalisation of the learning task and model performance assessment with the *SERA* metric allows for a non-uniform weighting of cases, via relevance functions, mimicking the logic of hinge loss in SVR. Additionally, our work does not assume distribution symmetry, as with hinge loss. Such is also a significant departure from previous work in utility-based regression, originating our proposal for non-parametric automatic generation of relevance functions.



**Fig. 15** SER curves of the best models according to the MSE (green), SERA (blue) and  $F_u$  (purple) metrics in datasets 3, 12 and 13 (Color figure online)

As for extreme value theory, methods attempt to avoid the bias introduced by assuming a normal distribution, concentrating the analysis on extremes. Often, methods require the definition of ad-hoc thresholds to distinguish extreme values (i.e. events), raising a known issue: lower thresholds increase the number of cases to estimate distribution parameters but include items closer to the central tendency of distributions; a higher threshold might lack sufficient data to estimate such parameters accurately. Our formalisation of imbalanced regression, and the use of the SERA evaluation metric, dismisses the common need of such thresholds. Such is also a significant difference to utility-based regression (Torgo and Ribeiro 2007), which requires user-defined thresholds for distinguishing normal and rare/extreme cases.

### 6.3 Utility-based metrics

Utility-based regression aims to address learning problems where a pre-specified range of highly relevant target values are associated with some sort of actionable decision—similar to activity monitoring (Fawcett and Provost 1999). Our formalisation of imbalanced regression tasks does not follow such perspective. The goal is to address problems by learning models that achieve good overall performance with an emphasis on the extreme (and higher relevance) values of the target variable.

In this paper, we have not made a direct comparison of SERA and metrics tailored for utility-based regression tasks (Ribeiro 2011; Moniz et al. 2018), such as the utility-based F-Score  $F_u$ —a combination of utility-based precision ( $Prec_u$ ) and recall ( $Rec_u$ ). There are several reasons for such a decision. First, the  $F_u$  metric evaluates the predictive ability of models solely in specific intervals of the domain; contrary to this, both MSE and SERA focus on evaluating predictive performance in the entire domain. Second,  $F_u$  requires the definition of an ad-hoc threshold, which ultimately dismisses the impact of cases where both true and predicted values have a relevance score lower than the ad-hoc threshold.

Regardless, we acknowledge the interest in clarifying the different impact that SERA and  $F_u$  metrics have in model selection/optimisation processes. Accordingly, we collected the results of the utility-based metric  $F_u$  in the experimental evaluation process described in Sect. 5.3, using an ad-hoc threshold of 1 (maximum). Figure 15 provides a depiction of

the *SER* curves for data sets 3, 12 and 13 (details in Table 5), where *MSE* (green), *SERA* (blue) and  $F_u$  (purple) metrics select different models. We remind that such experimental evaluation is based on the selection of the best performing model w.r.t. each metric (oracles).

By comparing the results of the metrics *MSE* and  $F_u$ , we observe that the best models according to the latter demonstrate the ability to reduce the prediction error in cases with high relevance. Nonetheless, they may demonstrate a considerable increase in the overall error across the domain (as observed when relevance is 0). As for the comparison between the best models according to *SERA* or  $F_u$ , results show that both reduce predictive error in higher relevance cases. However, results also show that models optimised with the *SERA* metric reduce the overall error more effectively—they do not focus solely on highly relevant cases. Additionally, the best models according to  $F_u$  show a consistent and noteworthy increase in the overall error across the domain, a trade-off that selection/optimisation processes using the *SERA* metric are capable of containing more effectively.

## 7 Conclusions

This paper addresses the problem of imbalanced regression and the prediction of extreme values. Although imbalanced domain learning has been a popular topic for over two decades, research concerning regression domains is, in comparison, negligible. This paper proposes methods that allow a robust process of formalisation and evaluation for imbalanced regression tasks. First, given the frequent absence of information concerning the domain, we propose an automatic and non-parametric approach to approximate domain preferences. Second, we propose a new evaluation metric—*SERA*, which is not only capable of considering non-uniform preferences in a domain but also provides a tool for dominance analysis. Results demonstrate the magnitude of our contribution, which we expect will foster future work in imbalanced regression. Further research directions can focus on two main aspects: the exploration of other methods to obtain relevance functions and the further study of *SERA*, regarding other evaluation metrics, e.g. rank-based metrics.

**Acknowledgements** This work is financed by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020. The authors would like to thank Vítor Cerqueira, Mariana Oliveira and Francesco Renna for their comments and suggestions. The authors would also like to thank Mia Hubert for her insights and discussion.

## Appendix: Minimization of *SERA*

Consider a data set  $\mathcal{D} = \{\langle x_i, y_i \rangle\}_{i=1}^N$ , a relevance function  $\phi : \mathcal{Y} \rightarrow \{0, 1\}$  defined for the target variable  $Y$ . We define the subset  $\mathcal{D}^t \subseteq \mathcal{D}$  formed by the cases for which the relevance of the target value is above or equal a threshold  $t$ , i.e.  $\mathcal{D}^t = \{\langle x_i, y_i \rangle \in \mathcal{D} \mid \phi(y_i) \geq t\}$ .

**Theorem 1** *The constant  $m_t$  that minimizes the Squared Error-Relevance  $SER_t = \sum_{i \in \mathcal{D}^t} (\hat{y}_i - y_i)^2$  is  $m_t = \frac{\sum_{i \in \mathcal{D}^t} y_i}{|\mathcal{D}^t|}$*

**Proof** To minimize the function  $SER_t$  with respect to  $m_t$ , we must have that  $\frac{\partial SER_t}{\partial m_t} = 0$ .

$$\begin{aligned} \frac{\partial SER_t}{\partial m_t} &= \frac{\partial}{\partial m_t} \sum_{i \in \mathcal{D}'} (m_t - y_i)^2 = \sum_{i \in \mathcal{D}'} \frac{\partial}{\partial m_t} (m_t - y_i)^2 \\ &= \sum_{i \in \mathcal{D}'} \left( \frac{\partial m_t^2}{\partial m_t} - \frac{\partial 2m_t y_i}{\partial m_t} + \frac{\partial y_i^2}{\partial m_t} \right) \\ &= \sum_{i \in \mathcal{D}'} (2m_t - 2y_i) = 2 \sum_{i \in \mathcal{D}'} (m_t - y_i) \\ \frac{\partial SER_t}{\partial m_t} = 0 &\Leftrightarrow 2 \sum_{i \in \mathcal{D}'} (m_t - y_i) = 0 \\ &\Leftrightarrow \sum_{i \in \mathcal{D}'} m_t - \sum_{i \in \mathcal{D}'} y_i = 0 \\ &\Leftrightarrow m_t \sum_{i \in \mathcal{D}'} 1 - \sum_{i \in \mathcal{D}'} y_i = 0 \\ &\Leftrightarrow m_t = \frac{\sum_{i \in \mathcal{D}'} y_i}{\sum_{i \in \mathcal{D}'} 1} \Leftrightarrow m_t = \frac{\sum_{i \in \mathcal{D}'} y_i}{|\mathcal{D}'|} \end{aligned}$$

**Theorem 2** The constant  $m$  that minimizes the Squared Error-Relevance Area  $SERA = \int_0^1 SER_t dt$  is  $m = \frac{\int_0^1 \sum_{i \in \mathcal{D}'} y_i dt}{\int_0^1 |\mathcal{D}'| dt}$

**Proof** To minimize the function  $SERA$  with respect to  $m$ , we must have that  $\frac{\partial SERA}{\partial m} = 0$ .

$$\begin{aligned} \frac{\partial SERA}{\partial m} = 0 &\Leftrightarrow \frac{\partial}{\partial m} \int_0^1 \sum_{i \in \mathcal{D}'} SER_t dt = 0 \Leftrightarrow \int_0^1 \frac{\partial}{\partial m} \sum_{i \in \mathcal{D}'} SER_t dt = 0 \\ &\Leftrightarrow \int_0^1 2 \sum_{i \in \mathcal{D}'} (m - y_i) dt = 0 \Leftrightarrow 2 \int_0^1 \sum_{i \in \mathcal{D}'} (m - y_i) dt = 0 \\ &\Leftrightarrow \int_0^1 \sum_{i \in \mathcal{D}'} m dt - \int_0^1 \sum_{i \in \mathcal{D}'} y_i dt = 0 \Leftrightarrow m \int_0^1 \sum_{i \in \mathcal{D}'} 1 dt = \int_0^1 \sum_{i \in \mathcal{D}'} y_i dt \\ &\Leftrightarrow m = \frac{\int_0^1 \sum_{i \in \mathcal{D}'} y_i dt}{\int_0^1 \sum_{i \in \mathcal{D}'} 1 dt} \Leftrightarrow m = \frac{\int_0^1 \sum_{i \in \mathcal{D}'} y_i dt}{\int_0^1 |\mathcal{D}'| dt} \end{aligned}$$

□

## References

Aggarwal, C. C. (2013). *Outlier analysis*. Berlin: Springer.

Akbulgic, O., Bozdogan, H., & Balaban, M. E. (2014). A novel hybrid RBF neural networks model as a fore-caster. *Statistics and Computing*, 24(3), 365–375.

Aldrin, M. (2006). Improved predictions penalizing both slope and curvature in additive models. *Computational Statistics and Data Analysis*, 50(2), 267–284.

Aldrin, M., & Haff, I. H. (2005). Generalised additive modelling of air pollution, traffic volume and meteorology. *Atmospheric Environment*, 39(11), 2145–2155.

Barker, P. M., & McDougall, T. J. (2020). Two interpolation methods using multiply-rotated piecewise cubic hermite interpolating polynomials. *Journal of Atmospheric and Oceanic Technology*, 37(4), 605–619. <https://doi.org/10.1175/JTECH-D-19-0211.1>.

- Basu, K., Mariani, M., Serpa, L., & Sinha, R. (2015). Evaluation of interpolants in their ability to fit seismometric time series. *Mathematics*, 3(3), 666–689. <https://doi.org/10.3390/math3030666>.
- Benavoli, A., Mangili, F., Corani, G., Zaffalon, M., & Ruggeri, F. (2014). A Bayesian Wilcoxon signed-rank test based on the Dirichlet process. In *Proceedings of the 31st international conference on international conference on machine learning, ICML'14* (Vol. 32, pp. II-1026–II-1034), JMLR.org.
- Benavoli, A., Corani, G., Demšar, J., & Zaffalon, M. (2017). Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *The Journal of Machine Learning Research*, 18(1), 2653–2688.
- Bi, J., & Bennett, K. P. (2003). Regression error characteristic curves. In *Proceedings of the 20th international conference on machine learning* (pp. 43–50).D
- Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2), 31:1–31:50.
- Branco, P., Torgo, L., & Ribeiro, R. P. (2019). Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, 343, 76–99.
- Brazdil, P., Giraud-Carrier, C., Soares, C., & Vilalta, R. (2008). *Metalearning: Applications to data mining* (1st ed.). Berlin: Springer.
- Brys, G., Hubert, M., & Struyf, A. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13(4), 996–1017. <https://doi.org/10.1198/106186004X12632>.
- Cain, M., & Janssen, C. (1995). Real estate price prediction under asymmetric loss. *Annals of the Institute of Statistical Mathematics*, 47(3), 401–414.
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1541882. <https://doi.org/10.1145/1541880.1541882>.
- Christoffersen, P. F., & Diebold, F. X. (1996). Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics*, 11(5), 561–571.
- Christoffersen, P. F., & Diebold, F. X. (1997). Optimal prediction under asymmetric loss. *Econometric Theory*, 13(06), 808–817.
- Cleveland, W. S., Grosse, E., & Shyu, W. M. (1992). *Local regression models* (Vol. 8). Belmont: Wadsworth & Brooks/Cole.
- Crone, S. F., Lessmann, S., & Stahlbock, R. (2005). Utility based data mining for time series analysis: Cost-sensitive learning for neural network predictors. In *Proceedings of the 1st international workshop on utility-based data mining*. (pp. 59–68). ACM.
- Ding, D., Zhang, M., Pan, X., Yang, M., & He, X. (2019). Modeling extreme events in time series prediction. In *Proceedings of the 25th ACM SIGKDD* (pp. 1114–1122). ACM.
- Dougherty, R. L., Edelman, A., & Hyman, J. M. (1989). Nonnegativity-, monotonicity-, or convexity-preserving cubic and quintic Hermite interpolation. *Mathematics of Computation*, 52(186), 471–494.
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. In *Proceedings of the 9th international conference on neural information processing systems, NIPS'96* (pp. 155–161) MIT Press.
- Fawcett, T., & Provost, F. (1999). Activity monitoring: Noticing interesting changes in behavior. In *Proceedings of the Fifth ACM SIGKDD, KDD'99* (pp. 53–62). ACM.
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Berlin: Springer. <https://doi.org/10.1007/978-3-319-98074-4>.
- Freemeteo. (2017). <http://freemeteo.com.pt/>. Accessed March 30, 2017.
- Fritsch, F. N., & Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17, 238–246.
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1), 1–58.
- Giraud-Carrier, C. (2005). The data mining advisor: Meta-learning at the service of practitioners. In *Proceedings of the fourth international conference on machine learning and applications, ICMLA'05* (pp. 113–119). IEEE Computer Society, USA. <https://doi.org/10.1109/ICMLA.2005.65>.
- Goodwin, P., & Wright, G. (2010). The limits of forecasting methods in anticipating rare events. *Technological Forecasting and Social Change*, 77(3), 355–368.
- Granger, C. W. (1999). Outline of forecast theory using generalized cost functions. *Spanish Economic Review*, 1(2), 161–173.
- Hald, A. A. (1998). *A history of mathematical statistics from 1750 to 1930*. New York: Wiley.
- He, X., Zhao, K., & Chu, X. (2019). Automl: A survey of the state-of-the-art. 1908.00709.
- He, H., & Ma, Y. (2013). *Imbalanced learning: foundations, algorithms, and applications* (1st ed.). New York: Wiley-IEEE Press.
- Hernández-Orallo, J. (2013). Roc curves for regression. *Pattern Recognition*, 46(12), 3395–3411.



- Herrera, M., Torgo, L., Izquierdo, J., & Pérez-García, R. (2010). Predictive models for forecasting hourly urban water demand. *Journal of Hydrology*, 387, 141–150.
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.
- Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2), 85–126. <https://doi.org/10.1023/B:AIRE.0000045502.10941.a9>.
- Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52(12), 5186–5201.
- Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In *Proceedings of the 5th international conference on learning and intelligent optimization, LION'05* (pp. 507–523). Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-642-25566-3\\_40](https://doi.org/10.1007/978-3-642-25566-3_40).
- Koprinska, I., & Rana, M., & Agelidis, V. (2011). Yearly and seasonal models for electricity load forecasting. In *Proceedings of IJCNN* (pp. 1474–1481).
- Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221–232.
- Kruschke, J. K. (Ed.). (2015). *Doing Bayesian data analysis* (2nd ed.). Boston: Academic Press.
- Kruschke, J. K., & Liddell, T. M. (2015). *The Bayesian new statistics: Two historical trends converge*. New York: SSRN eLibrary.
- Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 2017, 1–29.
- Lee, T. H. (2007). *Loss functions in time series forecasting*. Int Encyclopedia of the Social Sciences.
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(1), 6765–6816.
- López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250, 113–141.
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package v1.7-0.1.
- Milborrow, S. (2019). earth: Multivariate Adaptive Regression Splines. R package v4.7.0.
- Moniz, N., Ribeiro, R., Cerqueira, V., & Chawla, N. (2018). Smoteboost for regression: Improving the prediction of extreme values. In *2018 IEEE 5th international conference on data science and advanced analytics (DSAA)* (pp 150–159).
- Moniz, N., Branco, P., & Torgo, L. (2017a). Resampling strategies for imbalanced time series forecasting. *International Journal of Data Science and Analytics*, 3(3), 161–181.
- Moniz, N., Torgo, L., Eirinaki, M., & Branco, P. (2017b). A framework for recommendation of highly popular news lacking social feedback. *New Generation Computing*, 35(4), 417–450.
- Organization, W. H. (2005). Who air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide.
- Pebesma, E. (2012). spacetime: Spatio-temporal data in r. *Journal of Statistical Software, Articles*, 51(7), 1–30.
- Peters, A., & Hothorn, T. (2018). ipred: Improved Predictors. R package v0.9-8.
- Phillips, G. (2003). *Interpolation and approximation by polynomials*. CMS Books in Mathematics. Springer, <https://books.google.pt/books?id=87vciTxMcF8C>.
- Pinto, F., Cerqueira, V., Soares, C., & Mendes-Moreira, J. (2017). autobagging: Learning to rank bagging workflows with metalearning. 1706.09367.
- R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.
- Ribeiro, R. P. (2011). Utility-based regression. PhD thesis, Dep. Computer Science, Faculty of Sciences, University of Porto.
- Rijsbergen, C. J. V. (1979). *Information retrieval* (2nd ed.). Oxford: Butterworth-Heinemann.
- Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, 25(1), 127–141.
- Siffer, A., Fouque, P. A., Termier, A., & Largouet, C. (2017). Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD, KDD'17* (pp. 1067–1075). ACM.
- Therneau, T., & Atkinson, B. (2018). rpart: Recursive Partitioning and Regression Trees. R package v4.1-12.



- Torgo, L. (2005). Regression error characteristic surfaces. In *Proceedings of the eleventh ACM SIGKDD, KDD'05* (pp. 697–702). ACM.
- Torgo, L., & Ribeiro, R. (2007). Utility-based regression. In *Proceedings of 11th European conference on principles and practice of knowledge discovery in databases, PKDD* (pp. 597–604). Springer Berlin Heidelberg.
- Torgo, L., Branco, P., Ribeiro, R. P., & Pfahringer, B. (2013). Resampling strategies for regression. *Expert Systems*, 32(3), 465–476.
- Tukey, J. W. (1970). *Exploratory data analysis* (Prelim ed.). Reading: Addison-Wesley.
- Wang, X., Varol, O., & Eliassi-Rad, T. (2019). L2P: an algorithm for estimating heavy-tailed outcomes. CoRR abs/1908.04628.
- Wickham, H., & Stryjewski, L. (2012). 40 years of boxplots. Tech. rep., had.co.nz.
- Wilcoxon, R. R. (1990). Comparing the means of two independent groups. *Biometrical Journal*, 32(7), 771–780. <https://doi.org/10.1002/bimj.4710320702>.
- Wilcoxon, R. (2005). *Introduction to robust estimation and hypothesis testing. Statistical modeling and decision science*. Amsterdam: Elsevier Science.
- Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.
- Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394), 446–451.
- Zheng, Y., Liu, F., & Hsieh, H.P. (2013). U-Air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD* (pp. 1436–1444). ACM.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.