Data, Intelligence & Graphs (DIG) Seminar
Télécom Paris
29 March 2022

Imbalanced Regression and Extreme Value Prediction

Rita P. Ribeiro

INESC TEC
Department of Computer Science, Faculty of Sciences
University of Porto

Motivation

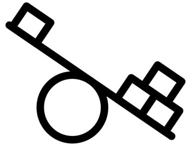
Standard Predictive Learning

- Uniform importance of values across the domain of the target variable Y
- The cases are equally relevant and thus the costs of the errors is the same

Motivation

Standard Predictive Learning

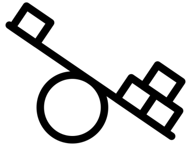
- Uniform importance of values across the domain of the target variable Y
- The cases are equally relevant and thus the costs of the errors is the same
- To achieve an overall good performance, the learning algorithm focus on the most frequent cases



Motivation

Standard Predictive Learning

- Uniform importance of values across the domain of the target variable Y
- The cases are equally relevant and thus the costs of the errors is the same
- To achieve an overall good performance, the learning algorithm focus on the most frequent cases



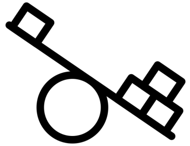
Predictive Learning in Imbalanced Domains

- Non-uniform importance of values across the domain of the target variable Y
- The cases that are more relevant are poorly represented in the training set.

Motivation

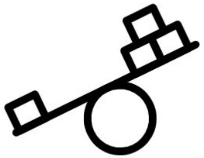
Standard Predictive Learning

- Uniform importance of values across the domain of the target variable Y
- The cases are equally relevant and thus the costs of the errors is the same
- To achieve an overall good performance, the learning algorithm focus on the most frequent cases



Predictive Learning in Imbalanced Domains

- Non-uniform importance of values across the domain of the target variable Y
- The cases that are more relevant are poorly represented in the training set.
- The costs of the errors is dependent on the relevance of the cases.



Motivation

Predictive Learning in Imbalanced Domains

- Classification Tasks
  - prediction of minority (positive) class(es);
  - e.g. fraud detection, rare disease diagnosis;

Motivation

Predictive Learning in Imbalanced Domains

- Classification Tasks
  - prediction of minority (positive) class(es);
  - e.g. fraud detection, rare disease diagnosis;

- Regression Tasks
  - several applications exist of numeric prediction tasks in imbalanced domains;
  - a specific range of values of the target variable, scarcely represented in the data set, maybe of the highest importance for the domain;
  - in most of the cases, the accurate prediction of extreme values is more critical;
  - e.g. extreme temperature values, high energy consumption demand.

Motivation

Imbalanced Domain Learning

- Far more research exists in classification than in regression.
- There are mainly two reasons for this gap.

1) How to define non-uniform preferences over continuous and possibly infinite domain of the target variable?

Motivation

Imbalanced Domain Learning

- Far more research exists in classification than in regression.
- There are mainly two reasons for this gap.

1) How to define non-uniform preferences over continuous and possibly infinite domain of the target variable?

First Contribution:

- Proposal of a method, based on previous work on utility-based regression, to obtain the so-called relevance function focused on extreme values, using an approach that is both automatic and non-parametric.

Motivation

Imbalanced Domain Learning

2) How to properly evaluate the models in an imbalanced regression setting to allow model selection and optimisation?

- in classification, it is known that standard metrics (e.g. accuracy) are not appropriate;
- in regression, the same happens with standard metrics (e.g. MSE);
- these metrics focus the model's performance on the cases with average target values.

Motivation

Imbalanced Domain Learning

2) How to properly evaluate the models in an imbalanced regression setting to allow model selection and optimisation?

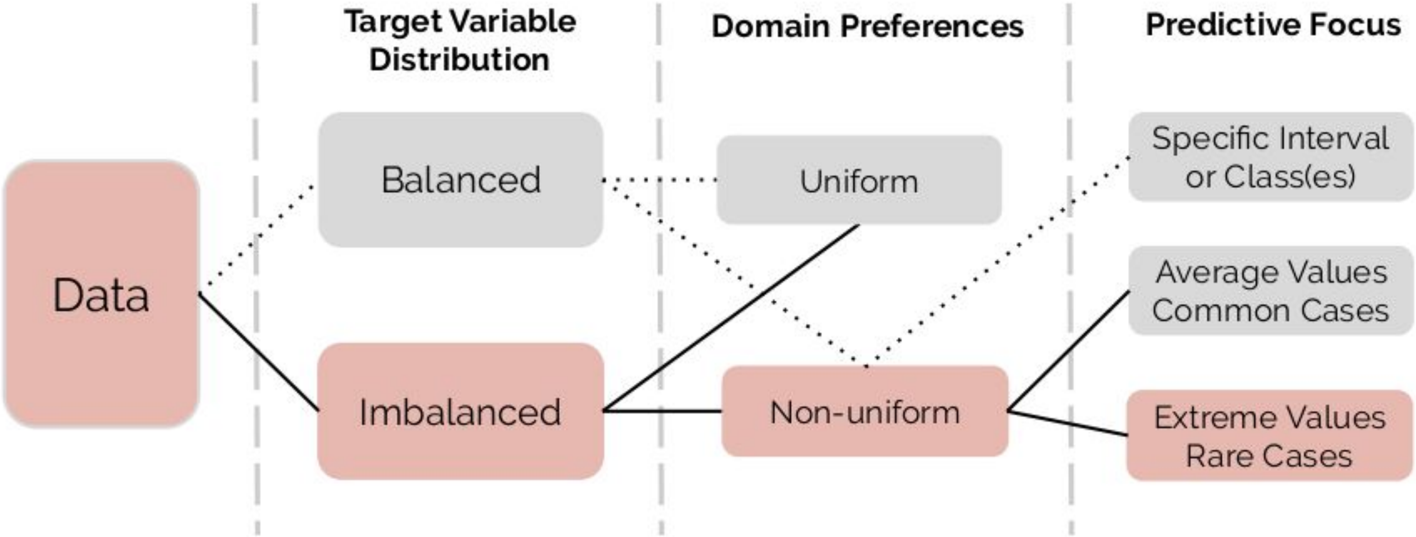
- in classification, it is known that standard metrics (e.g. accuracy) are not appropriate;
- in regression, the same happens with standard metrics (e.g. MSE);
- these metrics focus the model's performance on the cases with average target values.

Second Contribution

- Proposal of a new evaluation metric that allows evaluation of models as to their ability to predict extreme values, while robust to severe model bias.

Imbalanced Domain Learning

Problem Definition

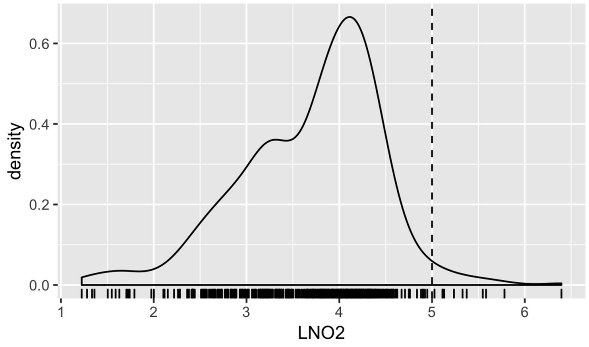


Imbalanced Regression

Air Pollution Example

- WHO Directive establishes that:

LNO2 concentration values	
low concentration	$ln(3\ \mu g/m^3) \approx 1.1$
annual mean guideline	$ln(40\ \mu g/m^3) \approx 3.7$
limit threshold	$ln(150\ \mu g/m^3) \approx 5.0$



LNO2: Log-transformed NO2 hourly concentration values in Oslo, during 2 years.

- LNO2 values above 5 are less frequent but are the most important ones for the model to be accurate on, as they are dangerous to human health.

Imbalanced Regression

Open Challenges

1) How to define non-uniform preferences over continuous and possibly infinite domain of the target variable?

2) How to properly evaluate the models in an imbalanced regression setting to allow model selection and optimisation?

Relevance Function

- Torgo and Ribeiro (2007) have proposed the concept of relevance function

$$\phi(Y) : \mathcal{Y} \rightarrow [0, 1]$$

  that maps the target variable domain to a scale of importance regarding the model predictions (1 is the maximum relevance).

- Given the infinite nature of the target variable, specifying the relevance values of all values is unfeasible.

- An approximation is necessary.

Relevance Function

- Torgo and Ribeiro (2007) have introduced the concept of relevance function

$$\phi(Y) : \mathcal{Y} \rightarrow [0,1]$$

  that maps the target variable domain to a scale of importance regarding the model predictions (1 is the maximum relevance).

- Given the infinite nature of the target variable, specifying the relevance values of all values is unfeasible.

- An approximation is necessary.

- Interpolation method by *Piecewise Cubic Hermite Splines* given a set of control points (e.g. the points established by WHO Directive)

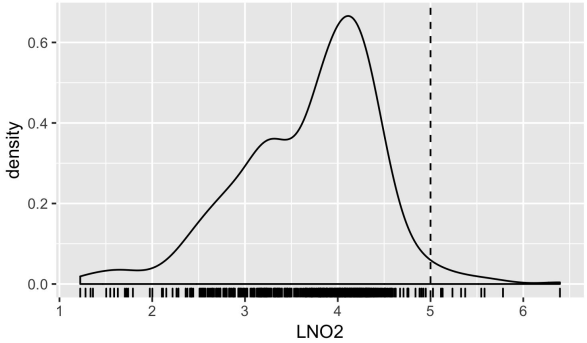
L. Torgo, R. P. Ribeiro : Utility-Based Regression. PKDD 2007: 597-604

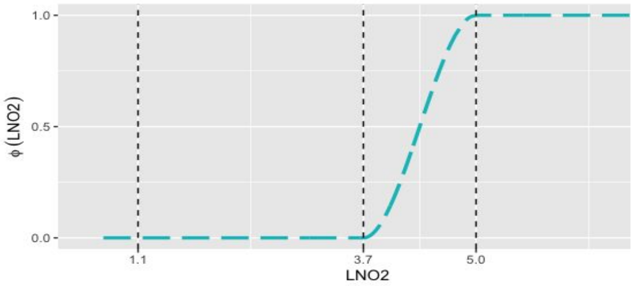
Imbalanced Regression

Air Pollution Example

- WHO Directive

LNO2 concentration values		ϕ(LNO2)
low concentration	$ln(3\ \mu g/m^3) \approx 1.1$	*0*
annual mean guideline	$ln(40\ \mu g/m^3) \approx 3.7$	*0*
limit threshold	$ln(150\ \mu g/m^3) \approx 5.0$	*1*



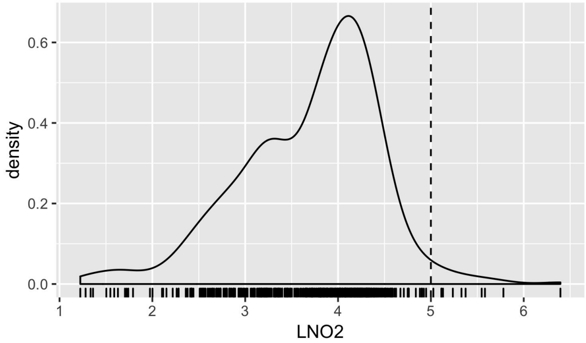


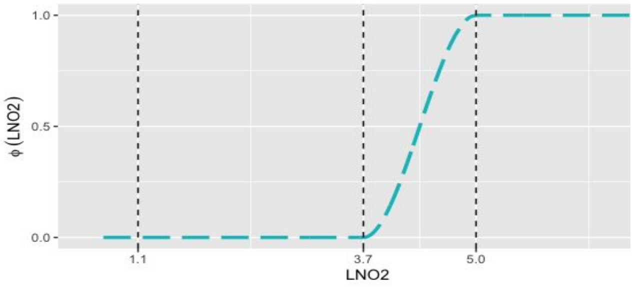
Imbalanced Regression

Air Pollution Example

- WHO Directive

LNO2 concentration values		ϕ(LNO2)
low concentration	$ln(3\ \mu g/m^3) \approx 1.1$	*0*
annual mean guideline	$ln(40\ \mu g/m^3) \approx 3.7$	*0*
limit threshold	$ln(150\ \mu g/m^3) \approx 5.0$	*1*



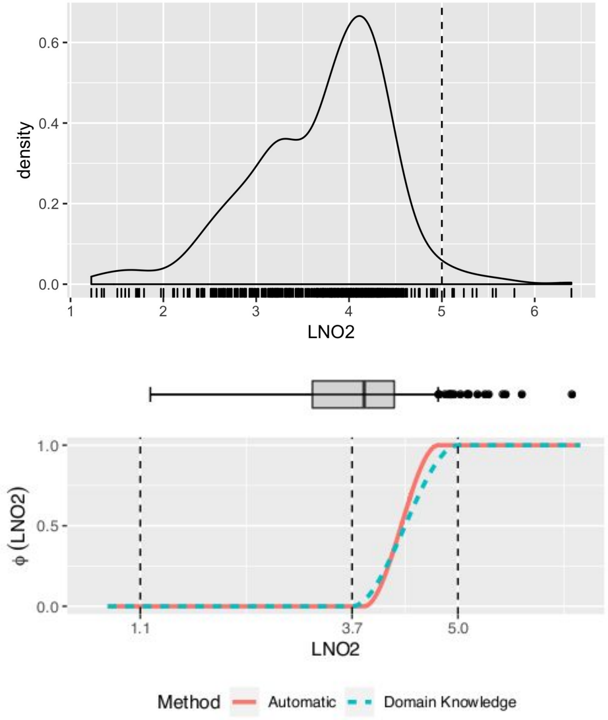


- Most of the times, there is no domain knowledge available to define the control points.

Relevance Functions

Air Pollution Example

- Assuming that the extreme values are the most important ones, a method (Ribeiro, 2011) exists based on the boxplot to supply the control points automatically.
- We propose the use of adjusted boxplot:
    - non-parametric and, thus, more flexible to underlying distributions of the data sample;
    - uses a robust measure of skewness;
    - avoids signalling “false” cases of extreme values.

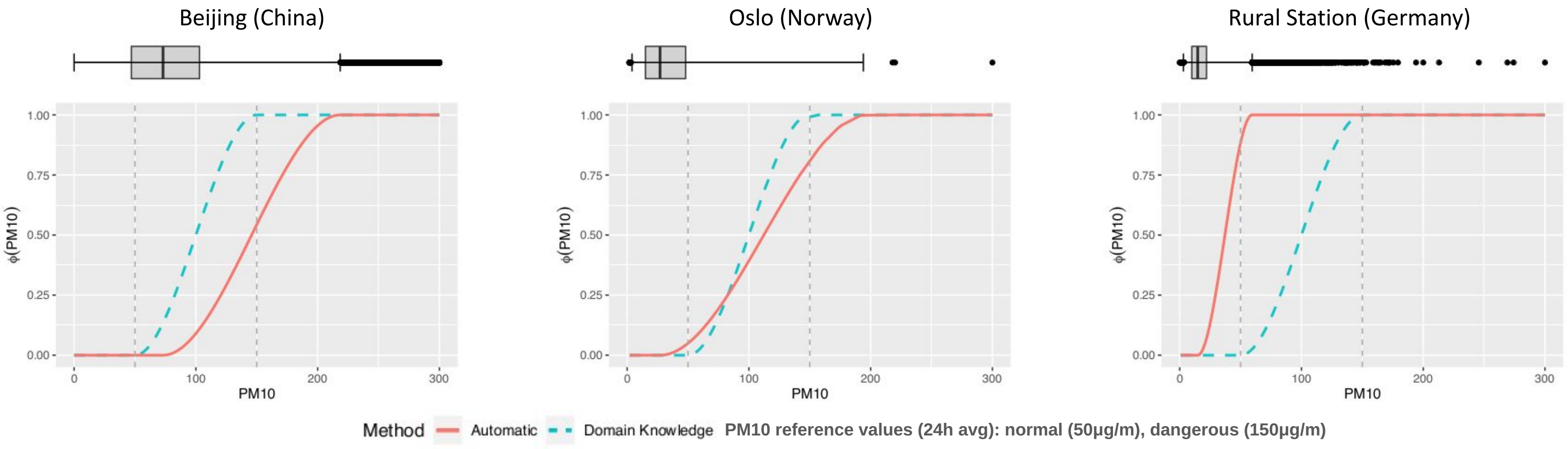


R. P. Ribeiro (2011) Utility-based Regression. PhD Thesis submitted to Faculty of Sciences of University of Porto

Relevance Functions

Air Pollution Example

- The similarity between the relevance functions based on domain knowledge and on boxplot depends on the representativeness of data sample concerning the domain.



Imbalanced Regression

Open Challenges

1) How to define non-uniform preferences over continuous and possibly infinite domain of the target variable?

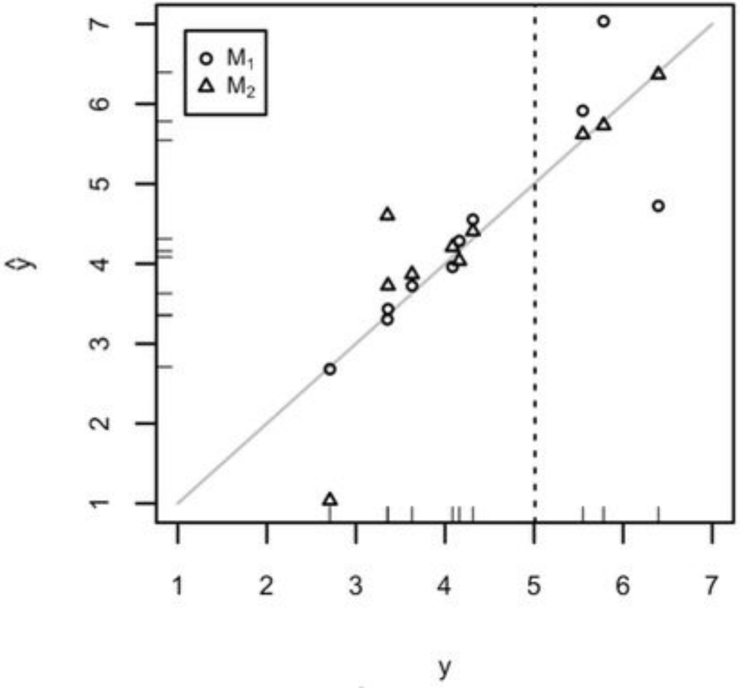
2) How to properly evaluate the models in an imbalanced regression setting to allow model selection and optimisation?

Evaluation Metrics: Issues

Air Pollution Example

Prediction of LNO2 Emissions

- M1 and M2 models achieve an MSE of 0.460
- Should they be considered equal from the domain perspective?

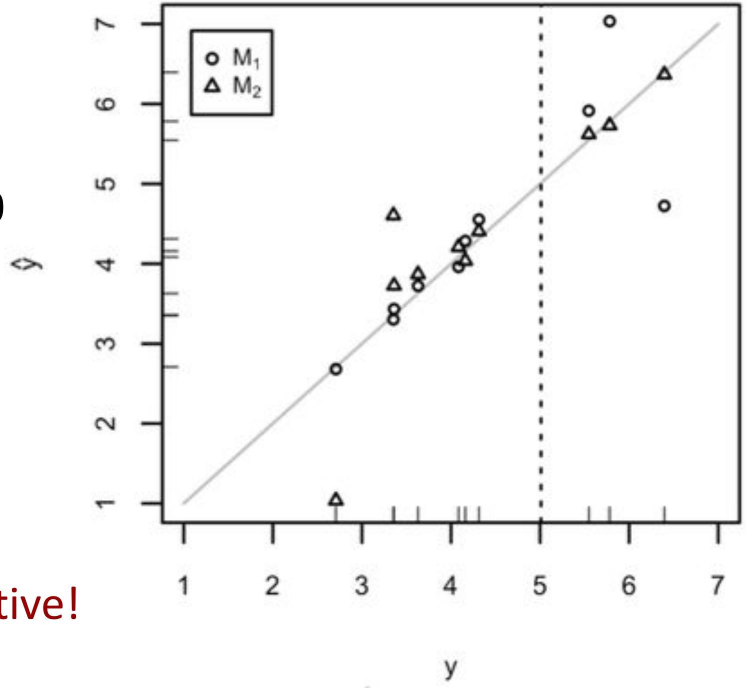


Evaluation Metrics: Issues

Air Pollution Example

Prediction of LNO2 Emissions

- M1 and M2 models achieve an MSE of 0.460
- Should they be considered equal from the domain perspective?
- M2 is more accurate at higher NO2 concentration values , the most important to predict accurately.
- M2 is more useful from the domain perspective!



Evaluation Metrics: Issues

- Standard performance metrics (e.g. MSE) are not suitable.
- Assume uniform domain preferences in the target variable.
- Focus solely on the magnitude of the errors and are heavily biased by the errors committed at the cases with the most frequent target values.

Evaluation Metrics: Issues

- Standard performance metrics (e.g. MSE) are not suitable.
- Assume uniform domain preferences in the target variable.
- Focus solely on the magnitude of the errors and are heavily biased by the errors committed at the cases with the most frequent target values.
- An evaluation metric is necessary to:
  - focus on minimising errors in cases with extreme target values;
  - prevent overfitting bias to the extreme values, disregarding all other cases;
  - allow for errors of equal magnitude have different impacts depending on the relevance values;
  - provide model discrimination, comparison and dominance analysis.

Squared Error-Relevance: SER

- Given a data set $\mathcal{D} = \{\langle \mathrm{x}_i , y_i \rangle\}_{i=1}^{N}$ and a relevance function $\phi(Y) : \mathcal{Y} \to [0, 1]$
- Let $\mathcal{D}^t = \{\langle \mathrm{x}_i , y_i \rangle \in \mathcal{D} \,|\, \phi(y_i) \geq t\} \subseteq \mathcal{D}$
- Squared Error-Relevance (SER) of a model w.r.t a cutoff t is

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

where $\hat{y}_i$ and y_i are the predicted and true value for case i in $\mathcal{D}^t$, respectively.

Squared Error-Relevance: SER

- Given a data set $\mathcal{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{N}$ and a relevance function $\phi(Y) : \mathcal{Y} \rightarrow [0, 1]$
- Let $\mathcal{D}^t = \{\langle \mathbf{x}_i, y_i \rangle \in \mathcal{D} \,|\, \phi(y_i) \geq t\} \subseteq \mathcal{D}$
- Squared Error-Relevance (SER) of a model w.r.t a cutoff t is

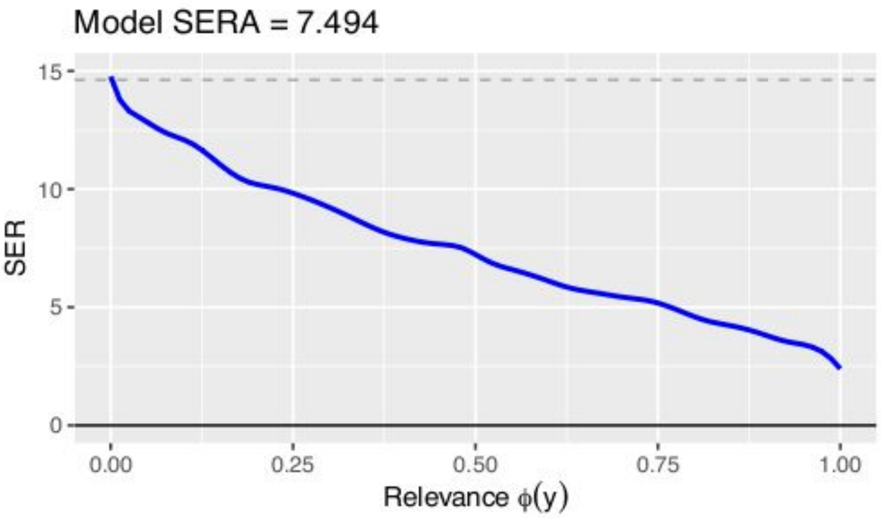
$$SER_t = \sum_{i \in \mathcal{D}^t} (\hat{y}_i - y_i)^2$$

where $\hat{y}_i$ and y_i are the predicted and true value for case i in $\mathcal{D}^t$, respectively.

- 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 is decreasing and monotonic.

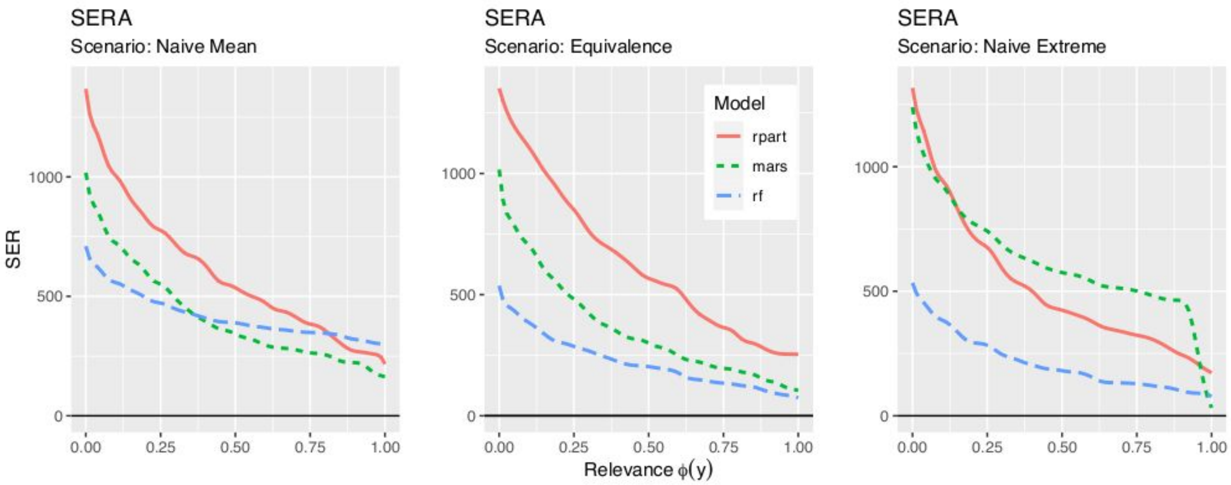
Squared Error-Relevance Area: SERA

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



Squared Error-Relevance Area: SERA

Model comparison and dominance analysis, robust to severe model biasing.



Experimental Study

- What is the impact of using standard evaluation metrics vs SERA in model selection?

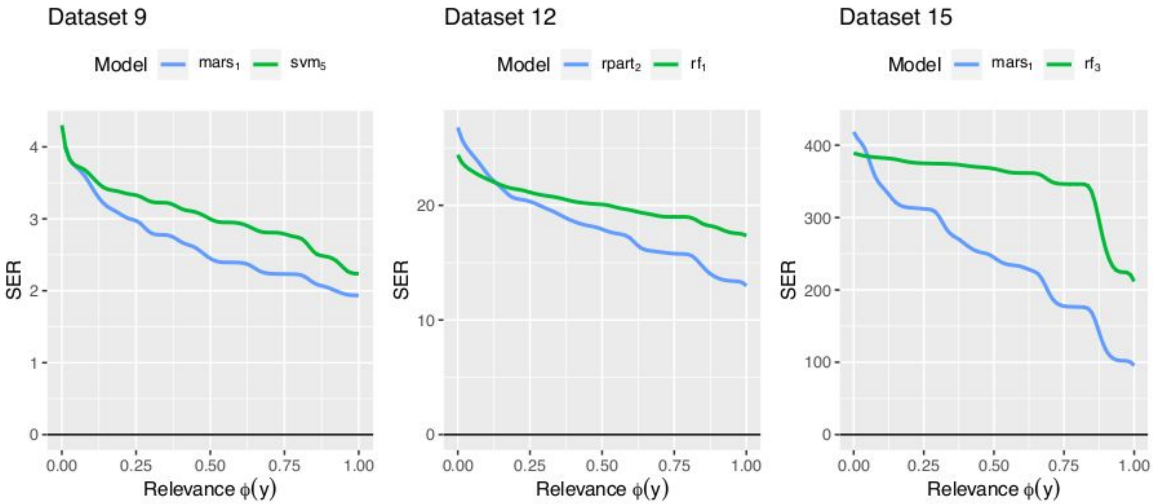
- Is SERA appropriate for model optimisation processes to improve the prediction of extreme values?

Experimental Setup:

- 34 data sets with extreme values according to the adjusted boxplot
- 5 algorithms: rpart, mars, svm, rf, bagging
- grid search of hyper-parameters
- 2x5-fold cross validation

Experimental Study: Model Selection

Data sets where different models were selected according to **MSE** and **SERA**.



Models selected by **MSE**

- lower sum of squared errors (SER) when considering all values with equal relevance.
- due to high density of cases in the central tendency, but with low relevance.

Models selected by **SERA**

- exhibit a better performance of SER when progressively focusing on cases with higher relevance.

Experimental Study: Model Optimization

Experimental methodology:

- train and test sets 70%/30% random partition;
- two optimization methods for SERA in training set (2x5-fold CV)
  - grid search optimization;
  - Hyperband (Li L et al, 2017);
- use the best grid search and Hyperband outcomes to learn a model using the entire training set;
- models selected in the optimization process are applied to test set, obtaining an estimation out-of-sample of prediction performance.

Experimental Study: Model Optimization

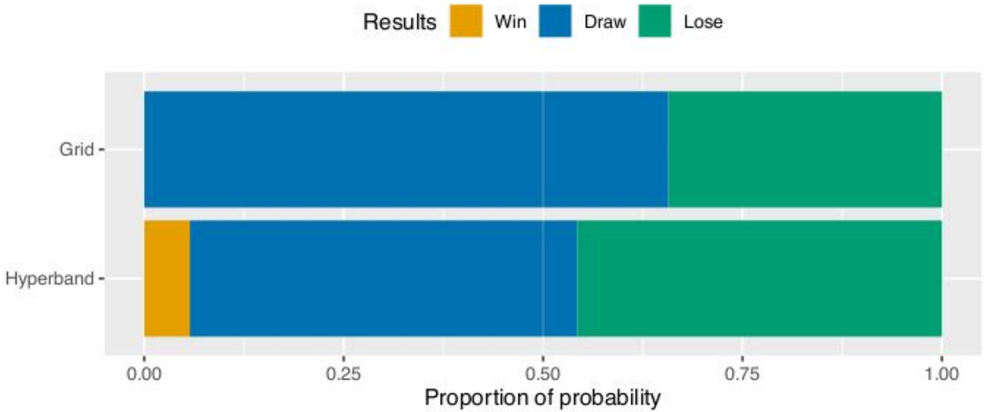
ROPE (Region of Practical Equivalence) with [-1%,1%] interval

By comparing SERA scores between optimised model and selected model:

Win: % diff SERA is less than -1%

Draw: % diff SERA is within [-1%,1%]

Lose: % diff SERA is greater than 1%



Optimised models are of practical equivalence or outperform the selected models with more than 50% probability → indicator of usefulness of SERA for optimization in learning algorithms

Conclusions

Tackle imbalanced regression tasks for the prediction of extreme values by means of:

- an automatic and non-parametric approach to approximate domain preferences, through the adjusted boxplot, for the definition of the relevance function for the target variable $\phi(Y)$;

- a new evaluation metric – SERA to assess the effectiveness of models towards the prediction of extreme values, penalizing severe model bias and low generalization capability.

- SERA is also a tool for dominance analysis.

References

- Ribeiro, R. P. and N. Moniz (2020).
  "Imbalanced regression and extreme value prediction".
  In: Machine Learning. DOI: [10.1007/s10994-020-05900-9](https://doi.org/10.1007/s10994-020-05900-9).

- All the proposed methods are available in **`IRon`** package in R.
  https://github.com/nunompmoniz/IRon

- Imbalanced Domain Learning with R
  Nuno Moniz and Rita P. Ribeiro
  Coming out late 2022!

CRC Press
Taylor & Francis Group

Data, Intelligence & Graphs (DIG) Seminar
Télécom Paris
29 March 2022

Thank you for your attention!

Questions?

Rita P. Ribeiro - rpribeiro@fc.up.pt