Imbalanced Regression: Definition, Impacts and Solutions

Denys Pommeret, Chaire ACTIONS, I2M (Aix-Marseille Université)

6th European
Congress of Actuaries
www.eca2026.org

Introduction

Data

Different types: structured (**tabular**), unstructured, time series, textual, geospatial, etc.

Learning...

Supervised (classification, **regression**) vs unsupervised (clustering, etc.) vs self-supervised

... from imbalanced data

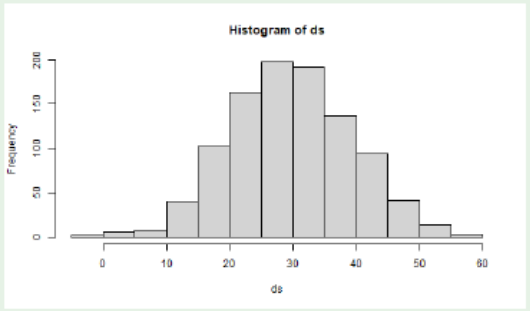
A dataset whose target variable distribution is "imbalanced"?
Examples: rare diseases, fraud, churn, catastrophes, etc.

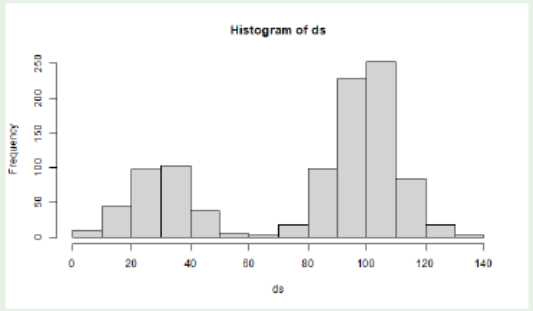
Statistical learning and imbalance bias

- Standard learning models:
  - Uniform importance of observations
  - Global and average metrics
- Syndrome: impact of imbalanced / non-representative samples
- Paradox: importance of rare values for the study

Notions of imbalance

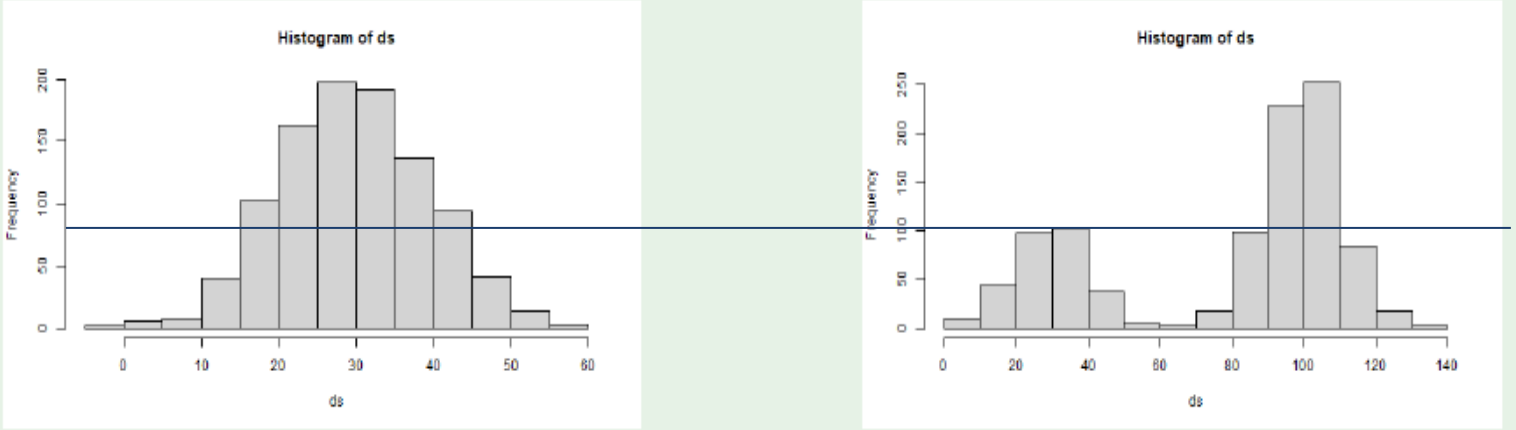
- Imbalance $\Longleftrightarrow$ Rarity? Extremes? Selection bias?
- Link with (statistical) extreme value theory?





Notions of imbalance

- Imbalance $\Longleftrightarrow$ Rarity? Extremes? Selection bias?
- Link with (statistical) extreme value theory?



Sources of imbalance?

- Intrinsic $\equiv$ natural asymmetry of the variable
- Extrinsic $\equiv$ external factor(s)
  e.g. selection bias $\Longrightarrow$ statistical survey theory

Notation

Notation

- Dataset $\mathbf{x} = (x_{ij})_{i=1,\cdot,n;j=1,\cdot,p} \in \mathcal{X}$
  n observations and p variables
- $y = (y_i)_{i=1,\cdot,n} \in \mathcal{Y}$ the associated target variable

In our example, *Telematics*

- $\mathbf{x}$ represents the features such as:
  - Driver information (age, gender, marital status, geographic area, etc.)
  - Vehicle information (age, usage, etc.)
  - Driving behaviour (speed, acceleration, braking, turns and distances, etc.)
- y represents the claims: claim amount, frequency, occurrence of claim(s)

Illustration

Experimental protocol

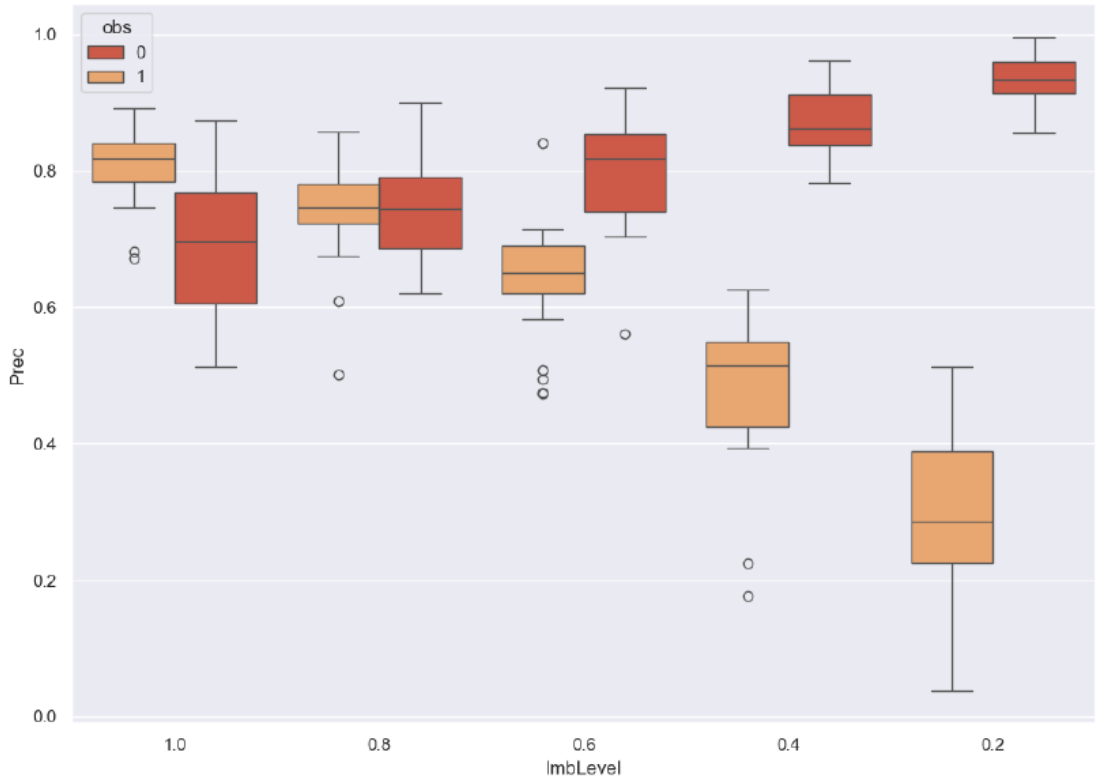
- Construction of a balanced test set (approximately uniform)
- Construction of a balanced training set
- Construction of several imbalanced training sets: from mildly to severely imbalanced, of various sizes
- k-fold approach (k=5)
- AutoML prediction (GLM, RF, Boosting, ANN, etc.)

Supervised learning

- Binary classification: claim vs no claim
- Regression: claim amount ($\in]0; 15000]$)

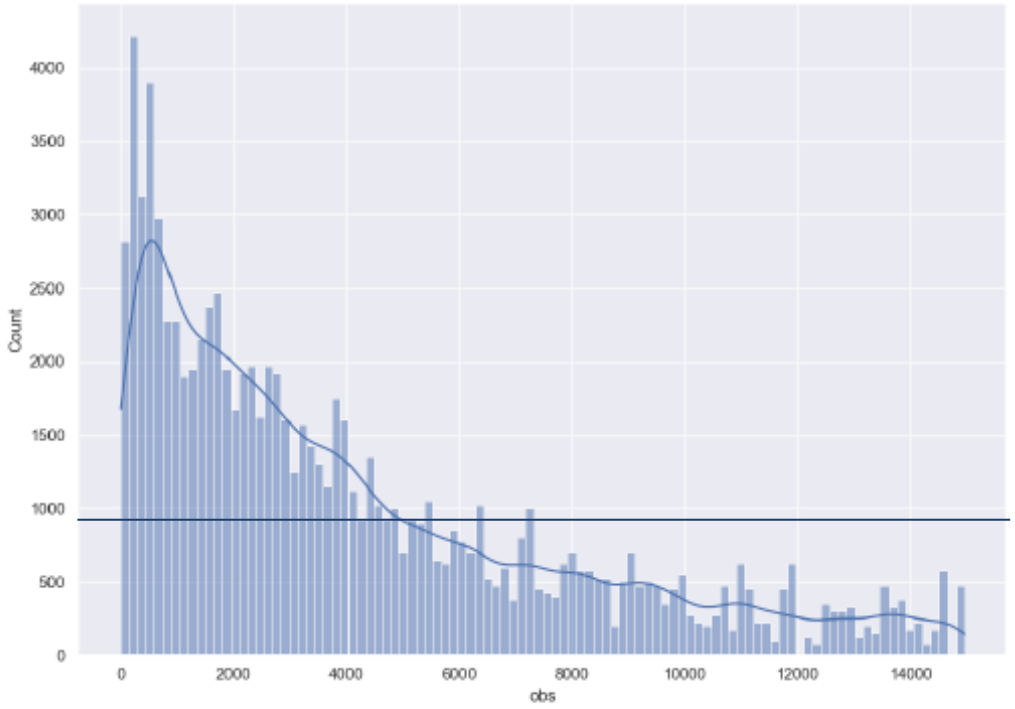
Imbalance between the proportion of pos. and neg.

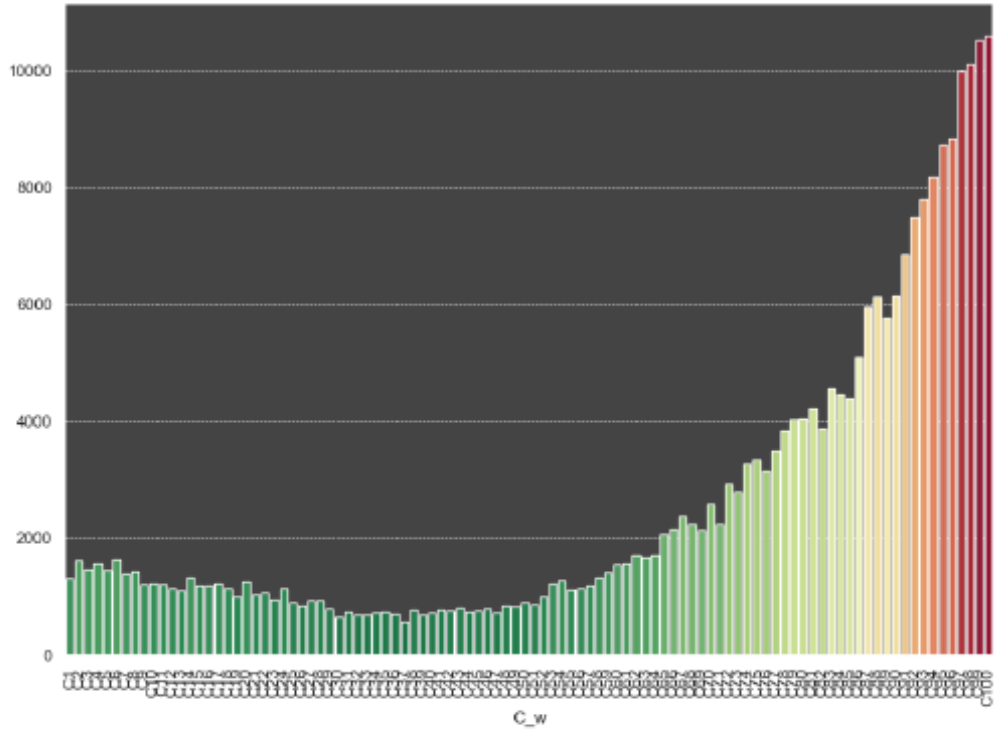
Classification



Imbalance continuously

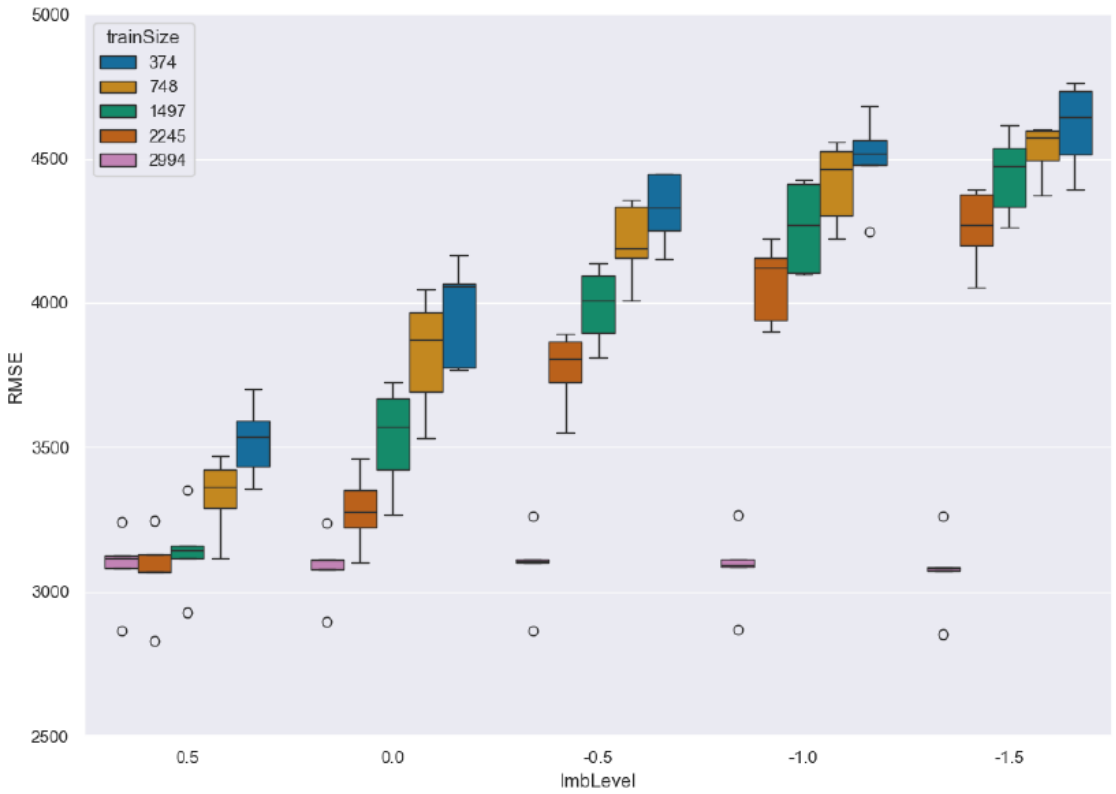
Regression





Effect on MSE

Regression



Definition

Imbalanced Regression

Supervised learning on data with $y = (y_i)_{i=1,\cdot,n} \in \mathcal{Y} \subset \mathbb{R}^n$ imbalanced

How to measure this imbalance? How to identify rare values?

First formal definition

[Rib11]: *Utility-Based Learning* approach: $\exists$ a utility function $\phi : \phi(y) : \mathcal{Y} \longrightarrow [0, 1]$ and $\equiv \phi(y)$ is not uniform on $\mathcal{Y}$

Identification of rare values

[TR07]: Binarisation of y values using a threshold t_r applied to $\phi(y)$:

$$D_R : \phi(y) > t_r \text{ and } D_N : \phi(y) \leq t_r$$

$\Longrightarrow$ Adaptation of *Imbalanced Classification* metrics and solutions

Definitions

Imbalanced Regression

Supervised learning on data with y in $\mathbb{R}$ imbalanced

How to measure this imbalance? How to identify rare values?

First formal definition

Utility-Based Learning approach: $\exists$ a utility function $\phi : \phi(y) : \mathcal{Y} \longrightarrow [0, 1]$

Identification of rare values

Binarisation of y values using a threshold t applied to $\phi(y)$:

$$y^* = 1 \text{ if } \phi(y) > t \text{ and } y^* = 0 \text{ if } \phi(y) \leq t$$

$\Longrightarrow$ Adaptation of *Imbalanced Classification* metrics and solutions

Contribution

Imbalanced $\Longleftrightarrow$ "rarity" of observations

But rarity with respect to ?

Let f be the **observed** density

Let f_0 be the **theoretical or expected** density, say the "target"

Rarity wrt f_0 means that f differs too much from f_0

Formal definition

Imbalanced (α, β)-imbalanced regression (wrt f_0) if $\exists A \subset \mathcal{Y}$ such that:

$$\int_A f_0(y)dy \geq \beta \quad \text{and} \quad \frac{\int_A f(y)dy}{\int_A f_0(y)dy} < \alpha.$$

Imbalanced Regression $\iff$ training sample significantly different from a reference distribution or under-represented (if $f_0 \sim \mathcal{U}$) for at least one significant subset of the support of Y

$\implies$ Definition encompassing both extremes and non-extreme rare values

Contributions: Generic rewriting of generators

Solutions to the *Imbalanced Learning* problem

Pre-processing vs in-processing vs post-processing
Pre-processing: under-/over- sampling ; synthetic data generation

Generalized Kernel density Estimate

$$g_{\boldsymbol{X}^*}(\boldsymbol{x}^*|\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \omega_i K_i(\boldsymbol{x}^*, \boldsymbol{x})$$

with $(K_i)_{i\in\mathcal{I}}$ a collection of kernels, $(\omega_i)_{i\in\mathcal{I}}$ a sequence of positive weights such that: $\sum_{i\in\mathcal{I}} \omega_i = 1$, and $\mathcal{I}$ denotes a subset of $\{1, 2, \cdots, n\}$.
Example : Smote, Gaussian Noise, etc.

Contribution

Imbalanced $\Longleftrightarrow$ "rarity" of observations

But rarity with respect to ?

Let f be the **observed** density

Let f_0 be the **theoretical or expected** density, say the "target"

Rarity wrt f_0 means that f differs too much from f_0

Solutions

Proposals for *Imbalanced Regression*

Sampling weights $\equiv$ inverse of the kde of the target variable Y
$\equiv$ the rarer the value, the higher its weight (non-parametric)

$$\omega_i := \frac{q_i}{\sum_j q_j}, \text{ with } q_i := \frac{1}{\widehat{f}_y(y_i)^\alpha},$$

with $\widehat{f}_y$ a kde and α a scaling hyperparameter

Generation of $Y^*|X^*$

Generation of y^* conditionally on $\boldsymbol{x}^*$: Wild-Bootstrap applied to the prediction error distribution of a random forest, adjusted with respect to $\boldsymbol{x}^*$

$$y_s^* := y_s + |\epsilon_k| v_s \times sign(\widehat{y_s} - \widehat{y_s^*})$$

Contributions

Proposals for *Imbalanced Regression*

Sampling weights $\equiv$ inverse of the kde of the target variable Y
$\equiv$ the rarer the value, the higher its weight (non-parametric)

$$\omega_i = \begin{cases} \frac{f_0(y_i)}{\widehat{f}(y_i)}, \\ \frac{1}{\widehat{f}_y(y_i)}, \\ \frac{1}{\widehat{f}_y(y_i)^\alpha}, \end{cases}$$

with $\widehat{f}_y$ a kde and $\alpha > 0$ a scaling hyperparameter

Contributions

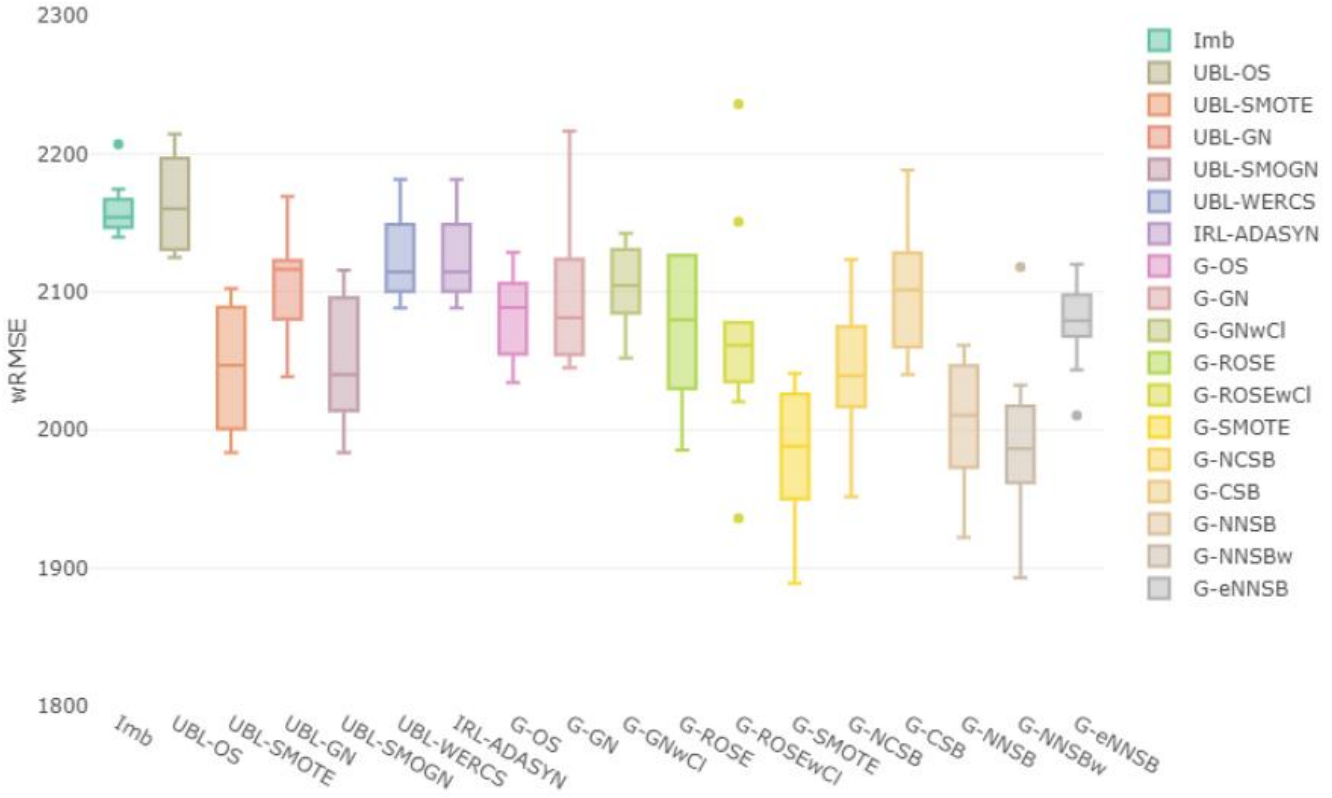
Generation of $Y^*|X^*$

Generation of y^* conditionally on $\boldsymbol{x}^*$: Wild-Bootstrap applied to the prediction error distribution of a random forest, adjusted with respect to $\boldsymbol{x}^*$

$$y_s^* := y_s + |\epsilon_s| v_s \times sign(\widehat{y_s} - \widehat{y_s^*}),$$

where ϵ are the residuals, v are suitable positive random variables.

Solutions



(a) Boxplots of wRMSE

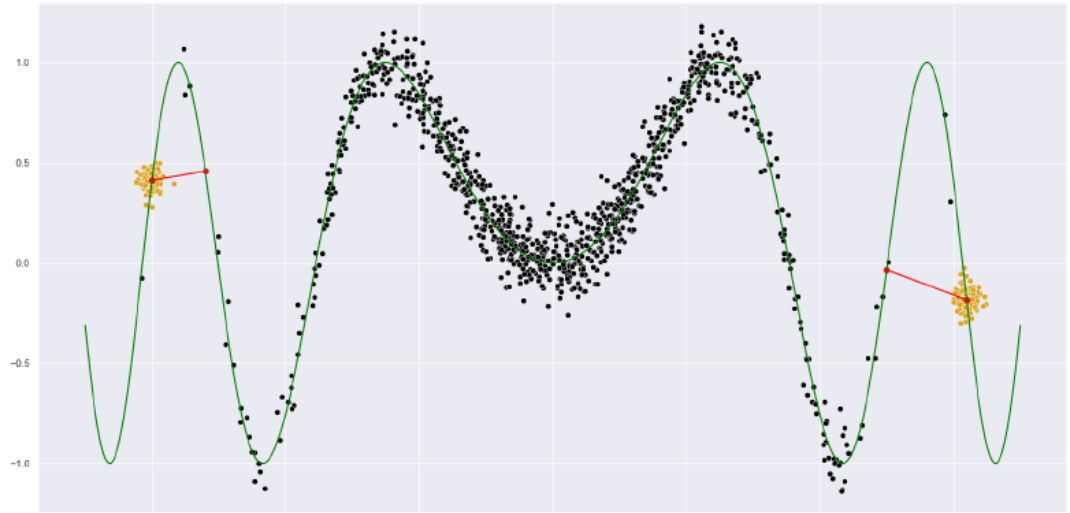
Dataset	Telematics
wRMSE gain	**GOLIATH**
G-OS	-3%
G-GN	-3%
G-GNwCl	-3%
G-ROSE	-5%
G-ROSEwCl	-4%
G-SMOTE	-8%
G-NCSB	-6%
G-CSB	-3%
G-NNSB	-7%
G-NNSBw	-8%
G-eNNSB	-4%

(b) Median wRMSE gains

Solutions

Limitations of GOLIATH

- Objectives
  - Synthetic data generation: rare values
  - Preserve correlations between data
- Limitations of GOLIATH
  - Data complexity: non-linear correlations
  - Mixed variables (categorical)



Solutions

Intuition

Changing the data representation space, more suitable for generation
$\Longrightarrow$ Latent space: correlations (latent variables); mixed data

Approaches

$\Longrightarrow$ Generate data from a latent space while capturing non-linear relationships $\Longrightarrow$ **Variational Autoencoder**, GAN, Diffusion Models, etc.

Solutions

Variational Autoencoder

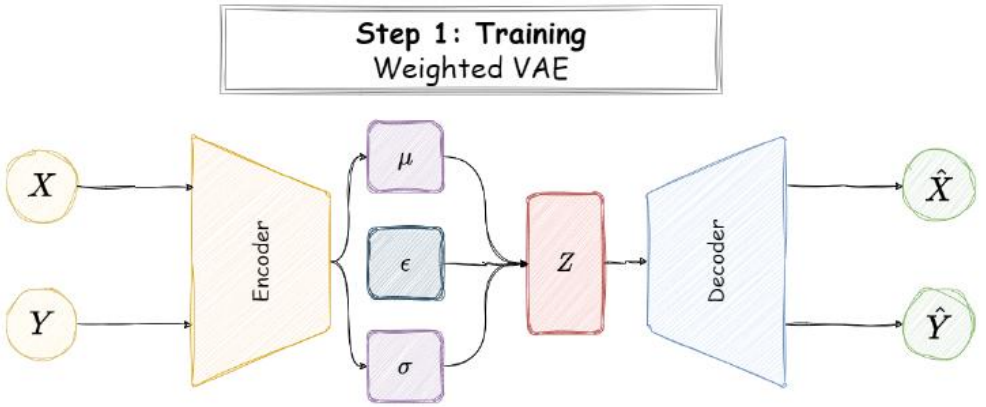
+ Regularity of the space, known latent distributions (Gaussian)
– Neglects rare values: not suited for IR, imposed latent distributions

Deep Smoothed Bootstrap for Imbalanced Regression

- Loss function specific to IR
- Generation from a Smoothed Bootstrap (GOLIATH) in the new representation space

Solutions

Architecture



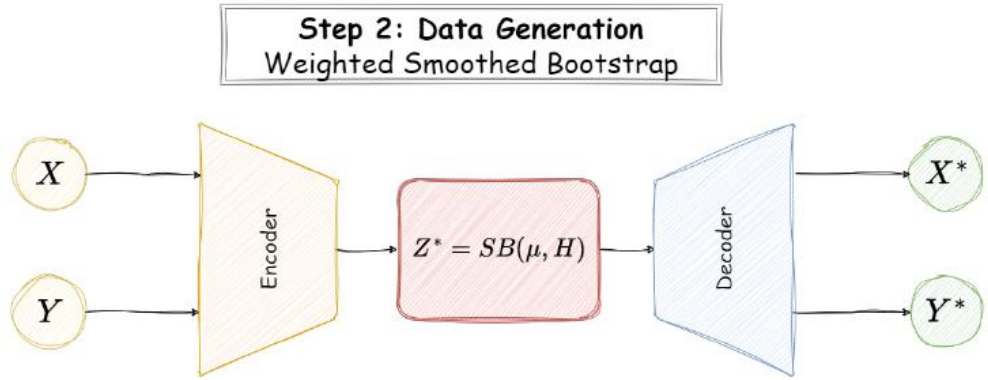


Figure: DAVID Algorithm

Solutions

Training

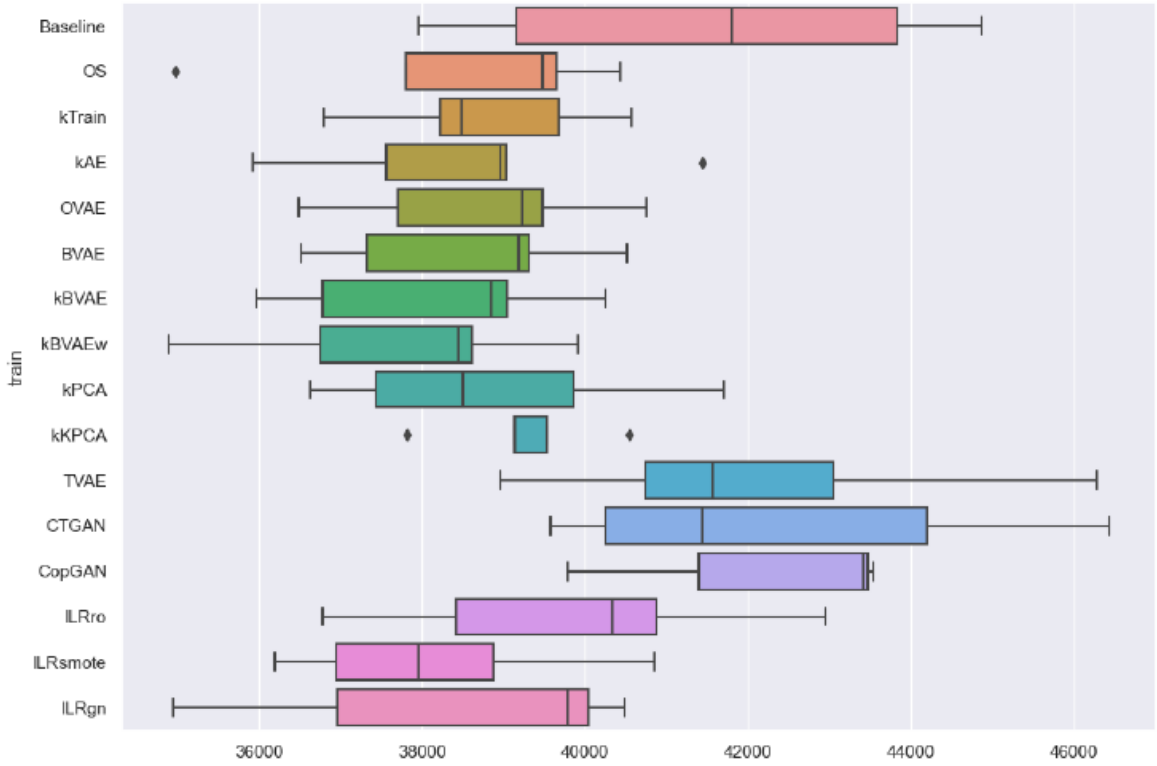
$$\mathcal{L}(\theta, \phi, \boldsymbol{x}, y) = \beta_x \mathbb{E}_q[log\ p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta_{KL} D_{KL}(q_\theta(z|\boldsymbol{x})||p_\theta(\boldsymbol{z})) + \frac{\beta_y}{\widehat{f}(Y)^\alpha} \mathbb{E}_q[log\ p_\theta(y|\boldsymbol{z})]$$

Generation

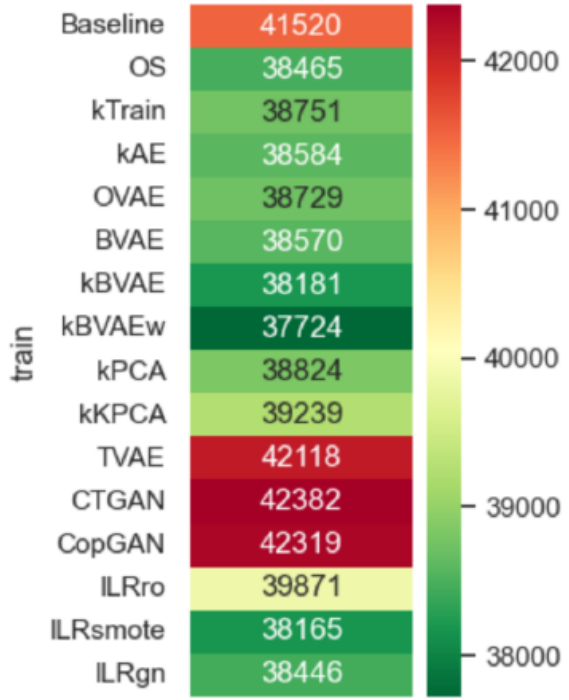
$$g_{z^*}(z^*|\mu) = \sum_{i=1}^{n} \omega_i K_{H_n}(z^* - \mu_i)$$

$$\omega_i := \frac{1}{\hat{f}_Y(y_i)^\alpha}$$

Solutions



(a) Boxplots of wMSE



(b) Mean wMSE

Conclusion

Summary

- Impact of imbalance on learning, not limited to the target variable
- Modelling of rare values (complementarity with EVT)
- Formal definitions
- *Imbalanced* Y: continuous weighting, imbalance measurement, consistent generation
- Natural extension to classification, unsupervised learning and synthetic data generation

Q&A

Thank you for your attention