FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

Evaluating the Stability of Model Explanations in Instance-dependent Cost-sensitive Credit Scoring

Matteo Ballegeer, Matthias Bogaert, Dries F. Benoit

August 28th, 2025

GHENT UNIVERSITY - RESEARCH GROUP DATA ANALYTICS

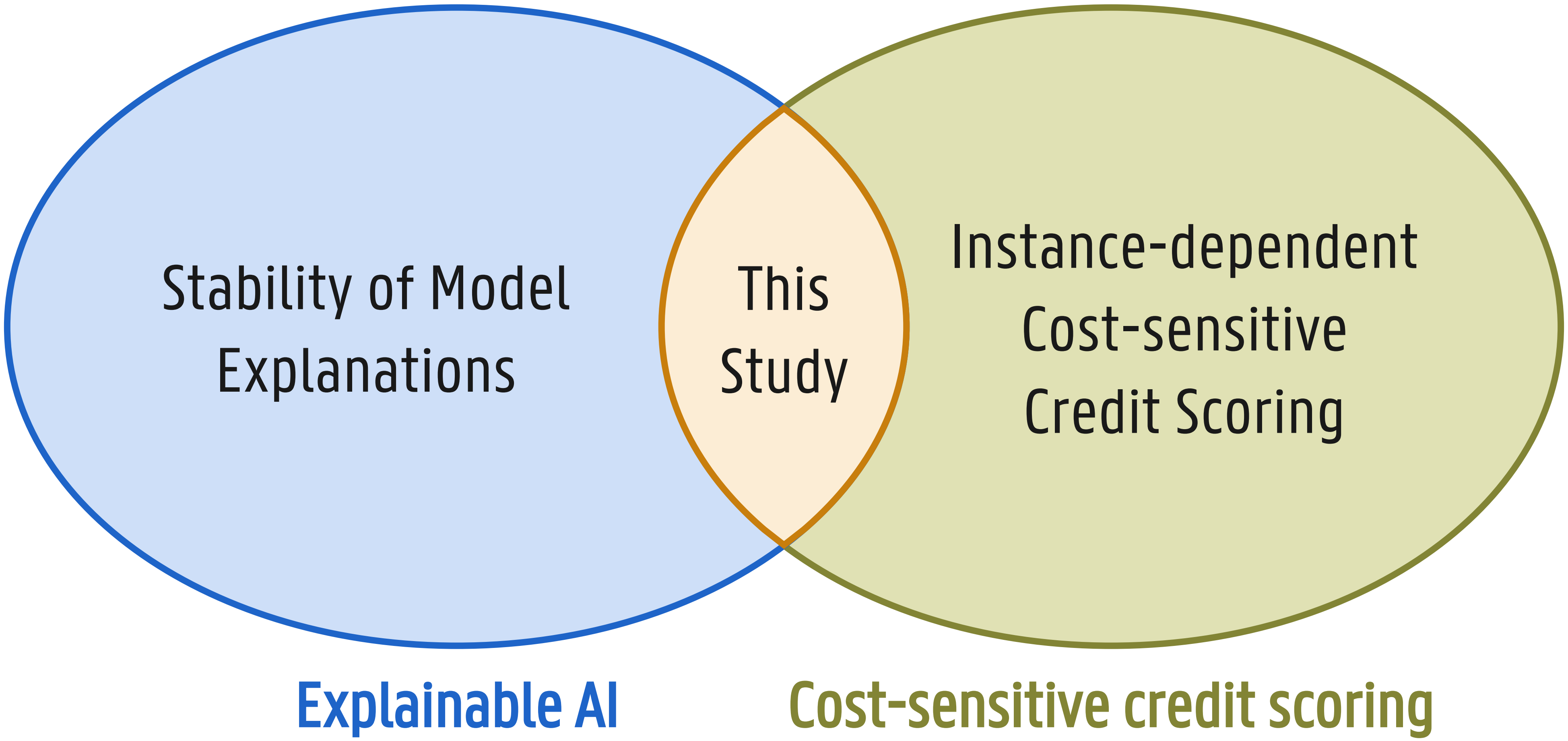
GHENT UNIVERSITY

Credit Research Centre

FLANDERS MAKE
DRIVING INNOVATION IN MANUFACTURING

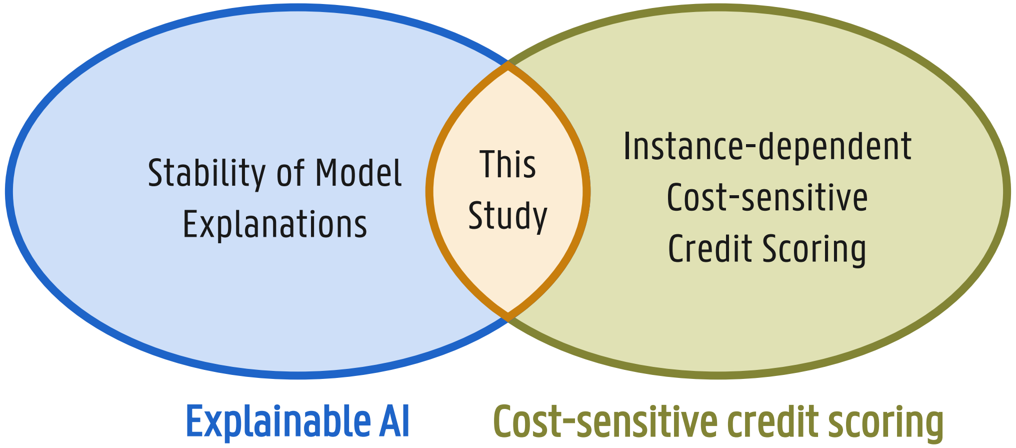
Research Positioning and Contributions

A bridge between model explainability and cost-sensitive learning



Research Positioning and Contributions

A bridge between model explainability and cost-sensitive learning



Main Contributions

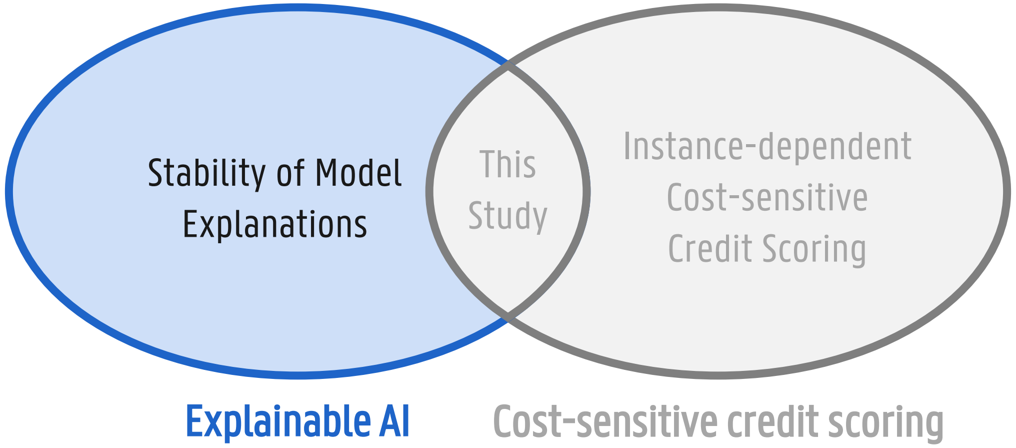
1. First credit scoring study relating explainable AI to cost-sensitive learning.
2. Demonstrate the impact of IDCS classifiers on the stability of post-hoc explainable AI techniques.
3. Show that a known negative impact of class imbalance on explanation stability[1] is amplified when using IDCS classifiers.
4. Introduce a novel, cross-dataset comparable evaluation metric for cost-sensitive learning.

GHENT UNIVERSITY

[1] Chen, Y., Calabrese, R., Martin-Barragan, B., 2024. Interpretable machine learning for imbalanced credit scoring datasets.

Research Positioning and Contributions

A bridge between model explainability and cost-sensitive learning



Main Contributions

1. First credit scoring study relating explainable AI to cost-sensitive learning.
2. Demonstrate the impact of IDCS classifiers on the stability of post-hoc explainable AI techniques.
3. Show that a known negative impact of class imbalance on explanation stability[1] is amplified when using IDCS classifiers.
4. Introduce a novel, cross-dataset comparable evaluation metric for cost-sensitive learning.

GHENT UNIVERSITY

[1] Chen, Y., Calabrese, R., Martin-Barragan, B., 2024. Interpretable machine learning for imbalanced credit scoring datasets.

Explainable AI

An increasing focus on regulating the use of black-box machine learning models


EBA DISCUSSION PAPER ON MACHINE LEARNING FOR IRB MODELS

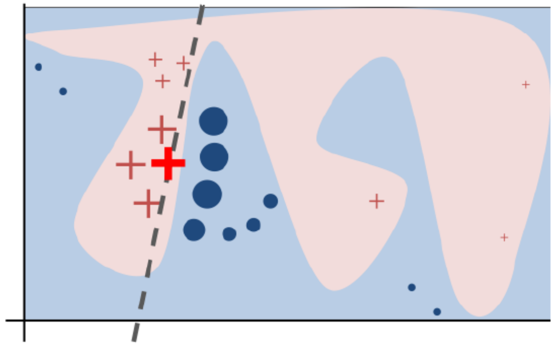
EBA/DP/2021/04

11 NOVEMBER 2021

Explainable AI

In response, XAI techniques to explain black-box models are rising in popularity

SHAP

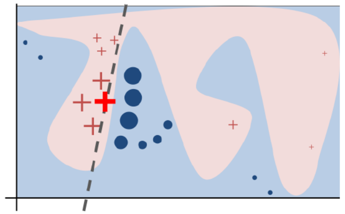


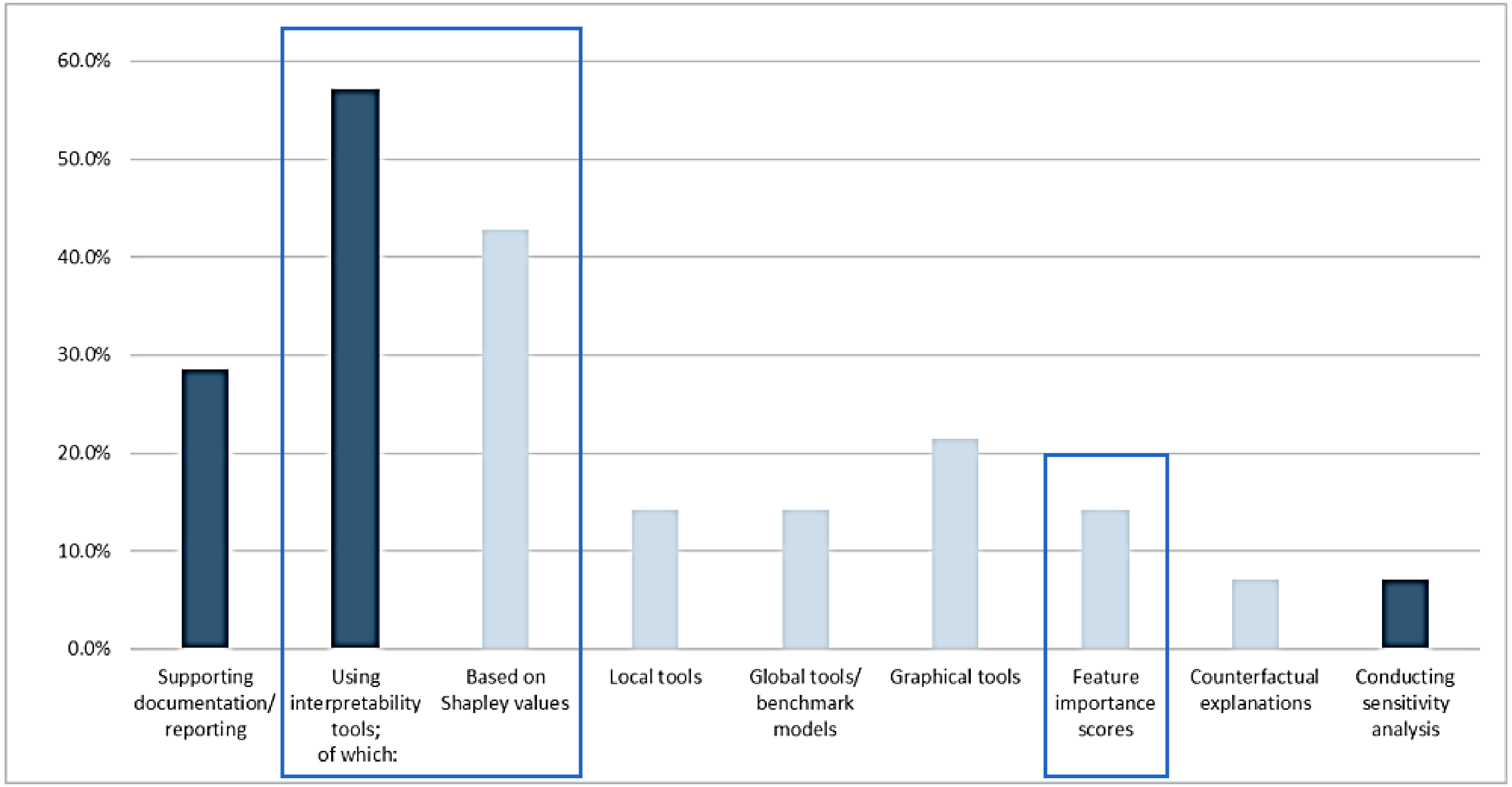
LIME

GHENT
UNIVERSITY

Explainable AI

In response, XAI techniques to explain black-box models are rising in popularity



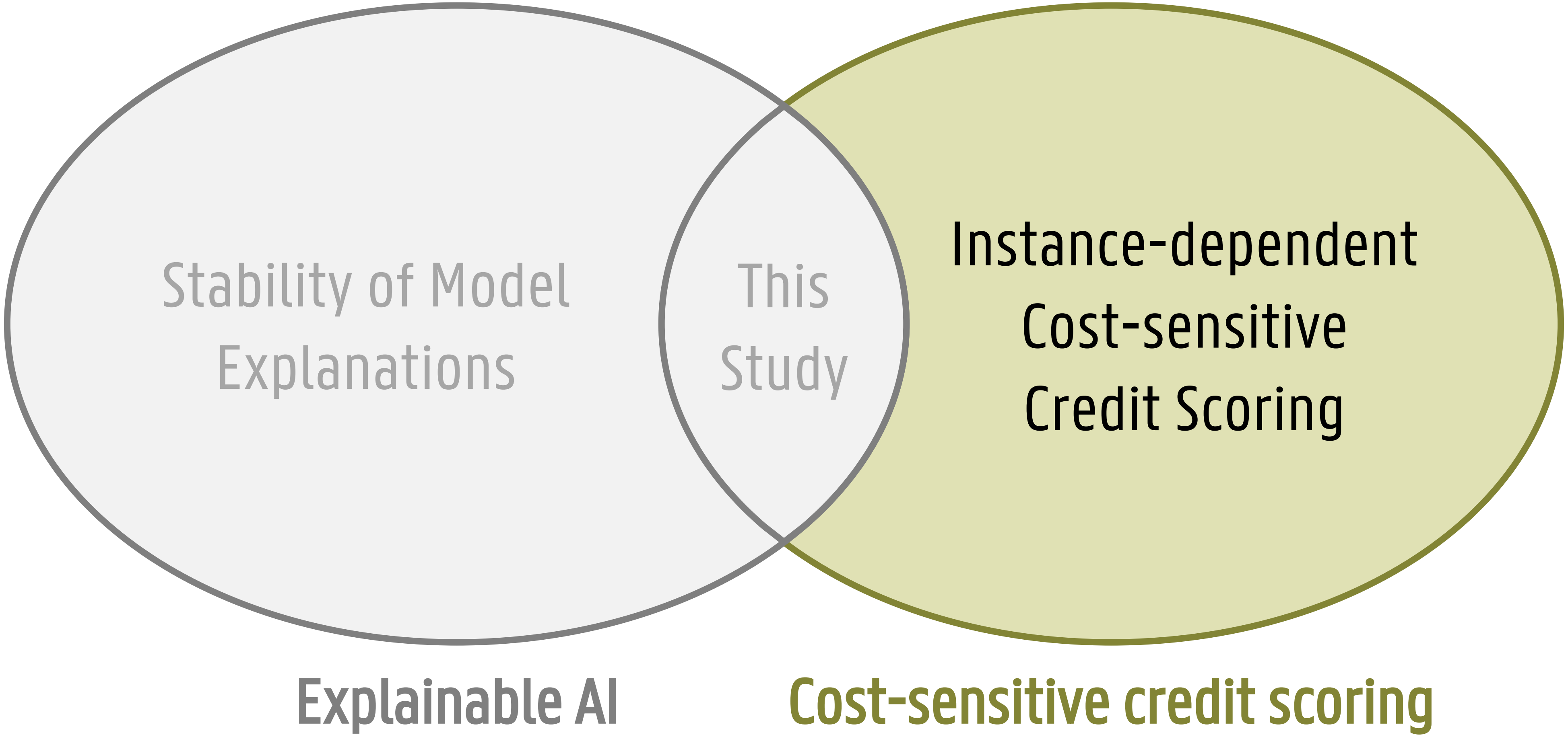


GHENT UNIVERSITY

European Banking Authority, 2023. Machine Learning for IRB Models: Follow-up Report from the Consultation on the Discussion Paper on Machine Learning for IRB Models.

Research Positioning and Contributions

A bridge between model explainability and cost-sensitive learning



Cost-sensitive Learning

Different misclassifications do not result in equal costs

Predicted \ Actual	No Default	Default
No Default	True negative	False negative
Default	False positive	True positive

Symmetrical misclassification costs

Predicted \ Actual	No Default	Default
No Default	Cost (0|0)	Cost (0|1)
Default	Cost (1|0)	Cost (1|1)

Asymmetrical at the class level

GHENT UNIVERSITY

Instance-dependent Cost-sensitive Learning

The requested loan amounts (and associated cost) between customers can vary a lot

		Actual	
		No Default	Default
Predicted	No Default	Cost (0|0)	Cost (0|1)
	Default	Cost (1|0)	Cost (1|1)

Asymmetrical at the class level

		Actual	
		No Default	Default
Predicted	No Default	$Cost_i(0\|0)$	$Cost_i(0\|1)$
	Default	$Cost_i(1\|0)$	$Cost_i(1\|1)$

Asymmetrical at the instance level

GHENT UNIVERSITY

Vanderschueren, T., Verdonck, T., Baesens, B., Verbeke, W., 2022a. Instance-dependent cost-sensitive learning: do we really need it?.

Instance-dependent Cost-sensitive Credit Scoring

The most popular credit scoring cost matrix from literature is used

		Actual	
		No Default	Default
Predicted	No Default	$Cost_i(0\|0)$	$Cost_i(0\|1)$
	Default	$Cost_i(1\|0)$	$Cost_i(1\|1)$

Asymmetrical at the instance level

		Actual	
		No Default	Default
Predicted	No Default	0	$Amount_i \cdot LGD$
	Default	$\boldsymbol{r_i} + Cost_{alt}$	0

$$Cost_{alt} = -\bar{r} \cdot \pi_0 + \overline{Amount} \cdot LGD \cdot \pi_1$$

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2014). Example-dependent cost-sensitive logistic regression for credit scoring.

GHENT UNIVERSITY

Instance-dependent Cost-sensitive Credit Scoring

Cost-efficient decision-making comes down to minimizing the AEC

		Actual	
		No Default	Default
Predicted	No Default	$Cost_i(0\|0)$	$Cost_i(0\|1)$
	Default	$Cost_i(1\|0)$	$Cost_i(1\|1)$

		Actual	
		No Default	Default
Predicted	No Default	0	$Amount_i \cdot LGD$
	Default	$r_i + Cost_{alt}$	0

$$Cost_{alt} = -\bar{r} \cdot \pi_0 + \overline{Amount} \cdot LGD \cdot \pi_1$$

$$\begin{aligned} Average\ Expected\ Cost\ (AEC)_i \ &= y_i \cdot [(1 - s(default)_i) \cdot (Amount_i \cdot LGD)] \\ &+ (1 - y_i) \cdot \left[s(default)_i \cdot \left(r_i + (-\bar{r} \cdot \pi_0 + \overline{Amount} \cdot LGD \cdot \pi_1)\right)\right] \end{aligned}$$

GHENT UNIVERSITY

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2014). Example-dependent cost-sensitive logistic regression for credit scoring.

Instance-dependent Cost-sensitive Credit Scoring

Relative AEC is introduced as dimensionless adaptation of AEC

$$\begin{aligned} \mathit{Average\ Expected\ Cost\ (AEC)}_i = {} & y_i \cdot [(1 - s(default)_i) \cdot (Amount_i \cdot LGD)] \\ & + (1 - y_i) \cdot \left[s(default)_i \cdot \left(r_i + (-\bar{r} \cdot \pi_0 + \overline{Amount} \cdot LGD \cdot \pi_1) \right) \right] \end{aligned}$$

$$\boldsymbol{Relative}\ \mathit{Average\ Expected\ Cost}\ (\boldsymbol{rel}AEC)_i = 1 - \frac{AEC(yi, s(default)_i, Amount_i)_i}{AEC(yi, \pi_1, Amount_i)}$$

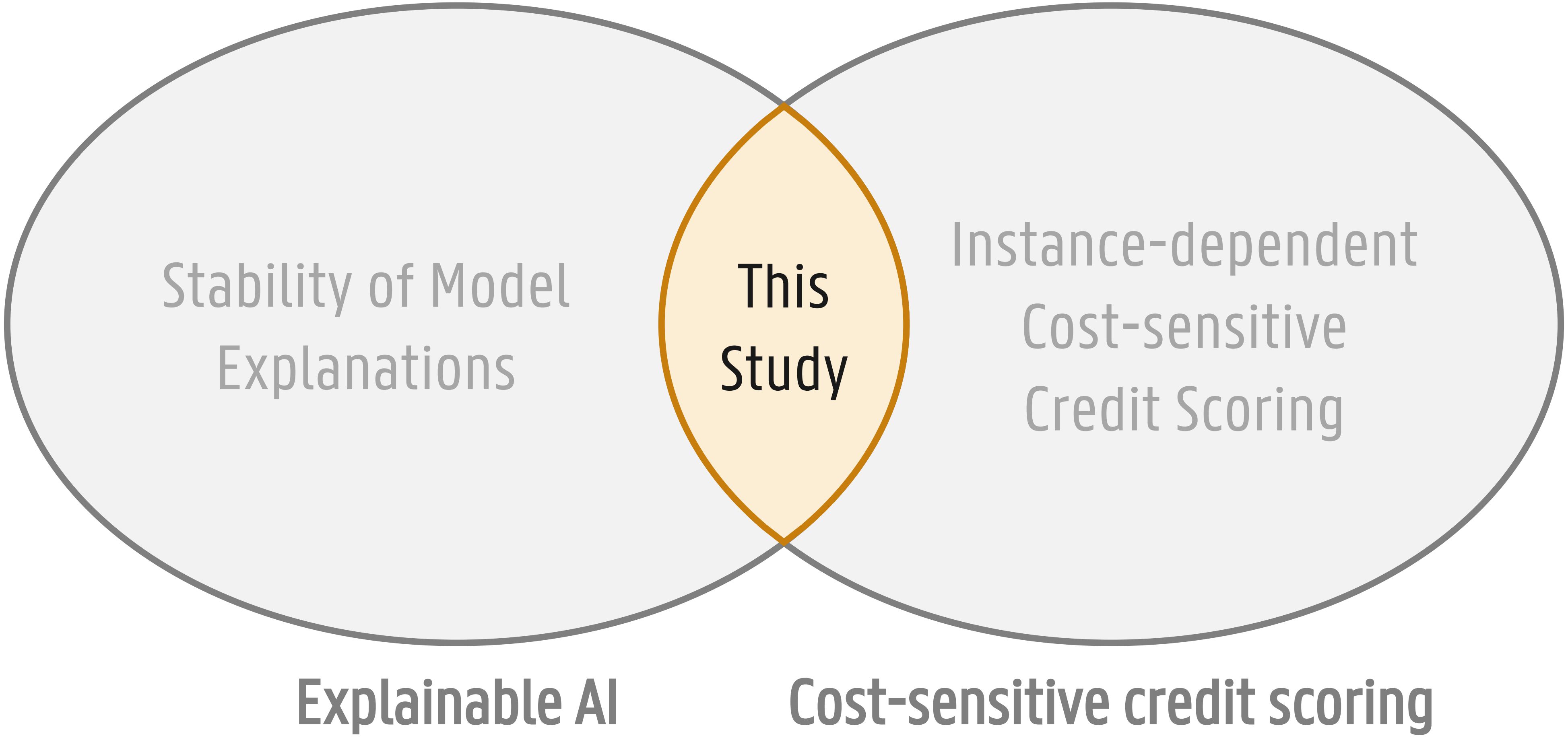
Worse than no model | Value

$-\infty$ 0 1

Similar interpretation as the savings metric!

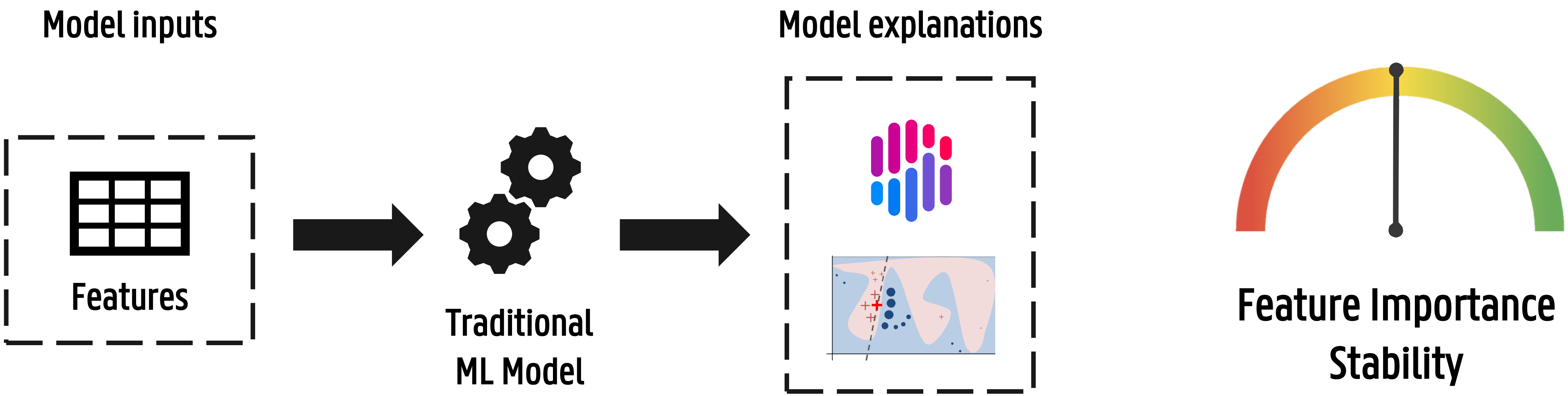
Research Positioning and Contributions

A bridge between model explainability and cost-sensitive learning



Explanation Stability in IDCS Credit Scoring

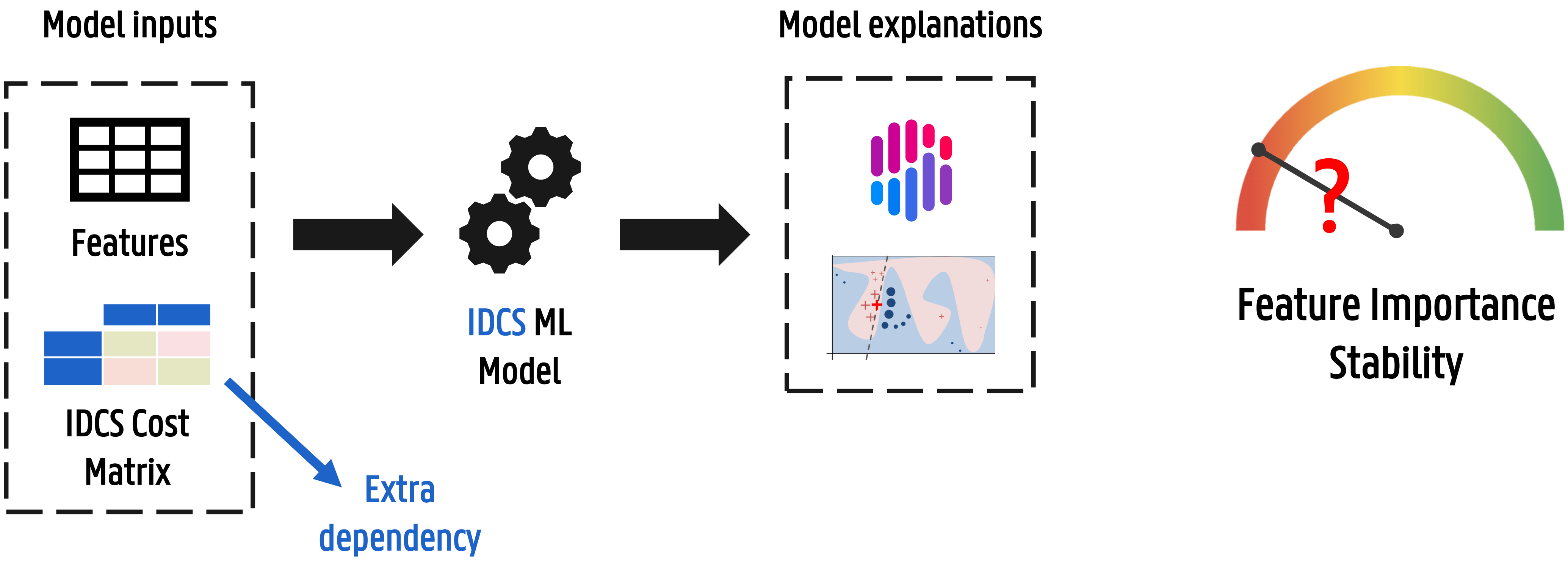
Does optimizing for an IDCS loss function impact the stability of model explanations?



GHENT
UNIVERSITY

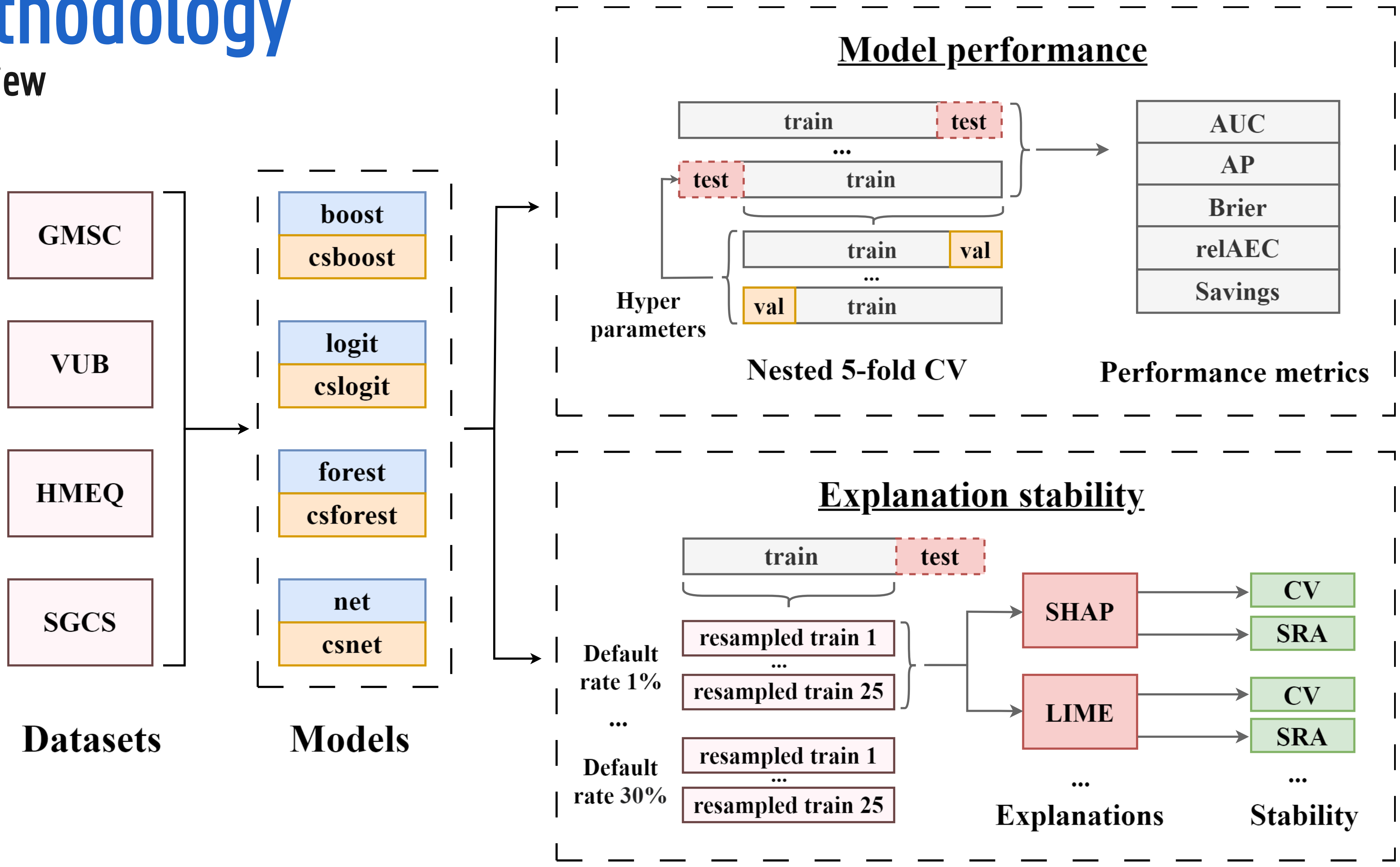
Explanation Stability in IDCS Credit Scoring

Does optimizing for an IDCS loss function impact the stability of model explanations?



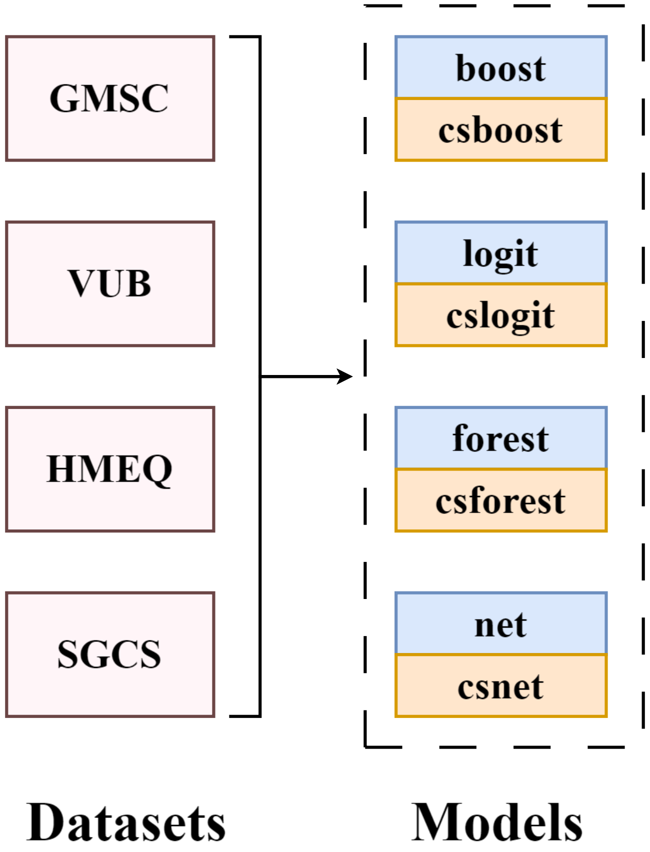
De Vos, S., en, T., Verdonck, T., Verbeke, W., 2023. Robust instance-dependent cost-sensitive classification

Methodology
Overview



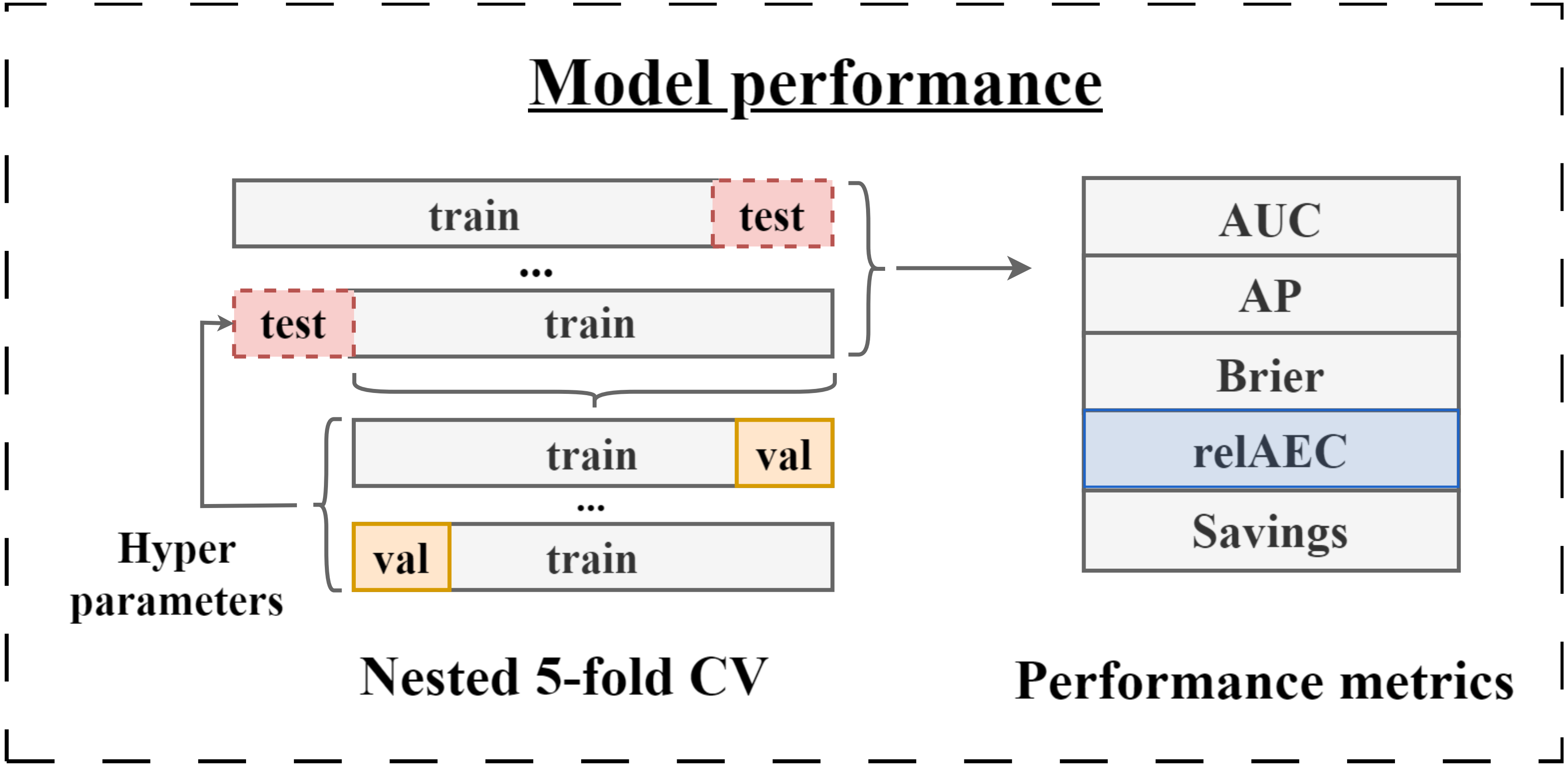
Methodology

Datasets and models



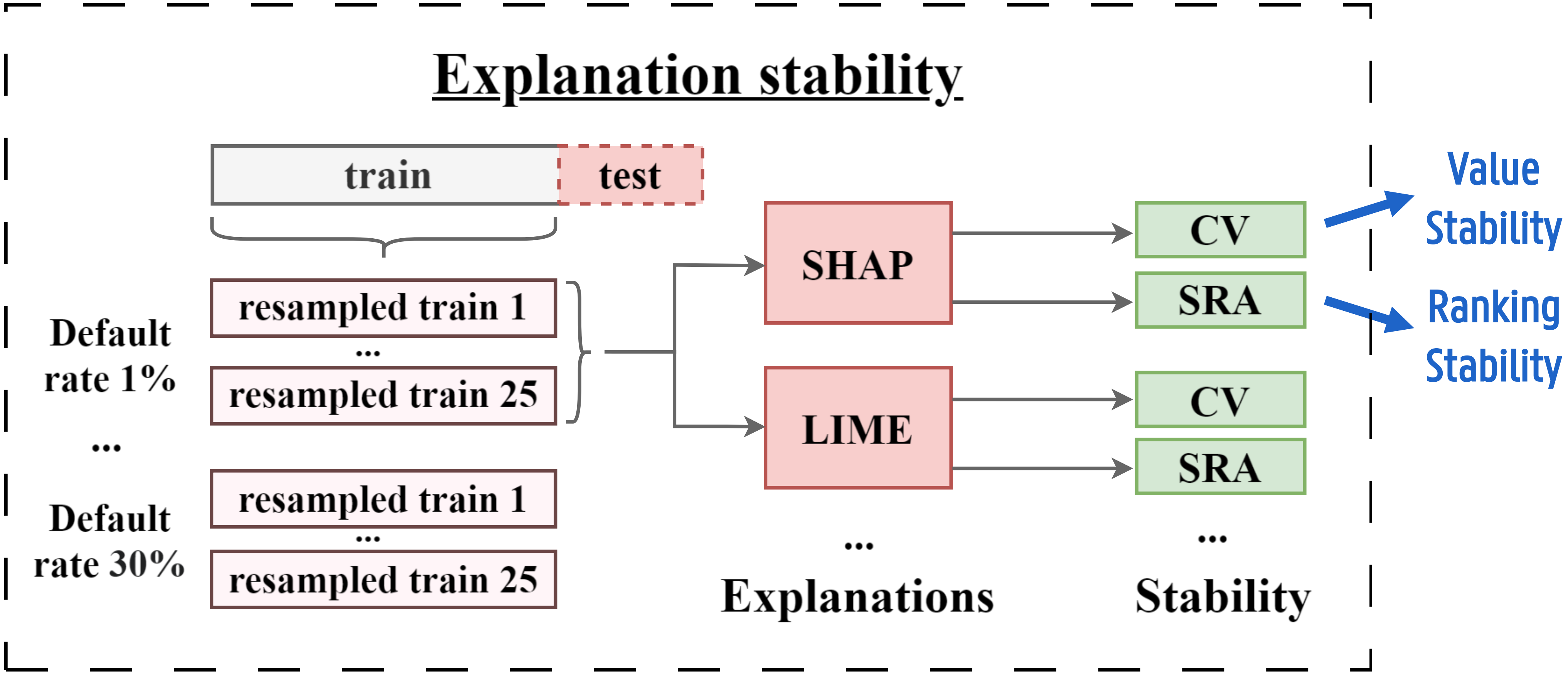
Methodology

Model performance



Methodology

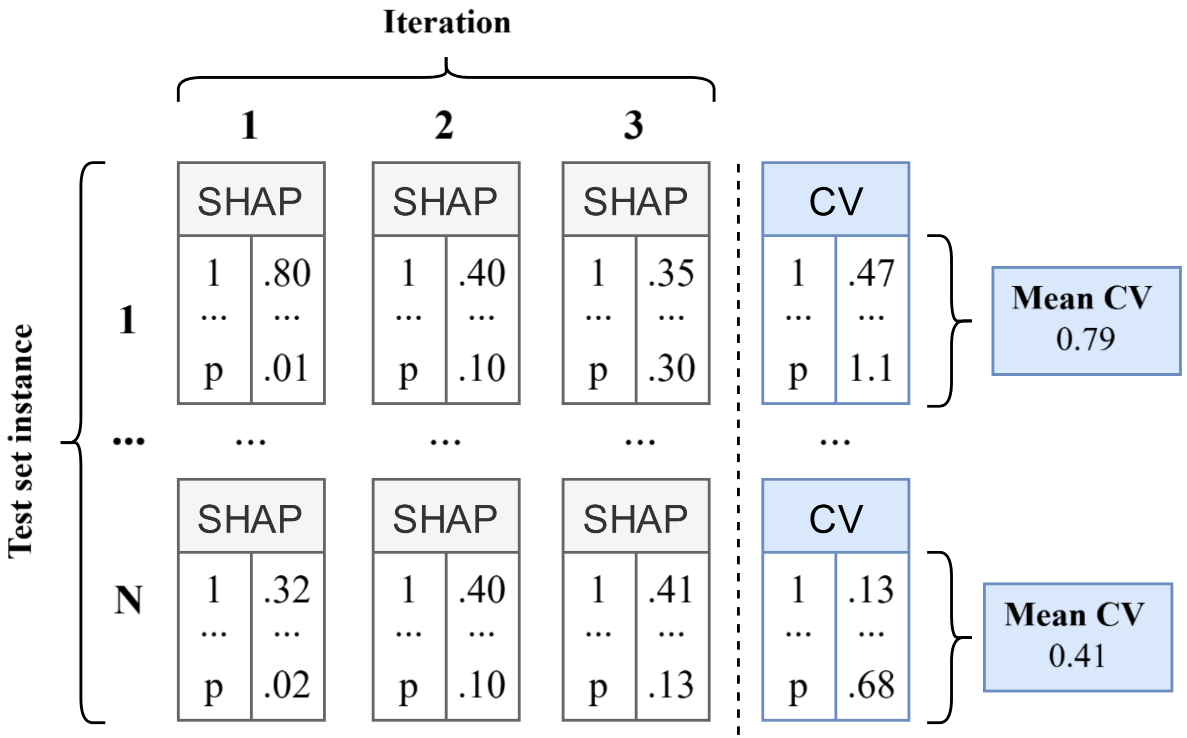
Explanation stability



Chen, Y., Calabrese, R., Martin-Barragan, B., 2024. Interpretable machine learning for imbalanced credit scoring datasets.

Methodology

Stability metrics



Xp	Φ_1	Φ_2	Φ_3
A	0.20	0.21	0.36
B	0.84	0.22	0.28
C	0.55	0.33	0.89
D	0.93	0.65	0.77
E	0.12	0.10	0.11

Value Stability

Coefficient of Variation (CV)

Ranking Stability

Sequential Rank Agreement (SRA)

GHENT UNIVERSITY

Ekstrøm, C.T., Gerds, T.A., Jensen, A.K., 2018. Sequential rank agreement methods for comparison of ranked lists. Biostatistics 20, 582598.

Methodology

Stability metrics

Xp	Φ_1	Φ_2	Φ_3
A	0.20	0.21	0.36
B	0.84	0.22	0.28
C	0.55	0.33	0.89
D	0.93	0.65	0.77
E	0.12	0.10	0.11

→

Xp	R_1	R_2	R_3	Var_L
A	4	4	3	0.33
B	2	3	4	1
C	3	2	1	1
D	1	1	2	0.33
E	5	5	5	0

Ranking Stability

Sequential Rank Agreement (SRA)

Ekstrøm, C.T., Gerds, T.A., Jensen, A.K., 2018. Sequential rank agreement methods for comparison of ranked lists. Biostatistics 20, 582598.

Methodology

Stability metrics

Xp	R_1	R_2	R_3	Var_L
A	4	4	3	0.33
B	2	3	4	1
C	3	2	1	1
D	1	1	2	0.33
E	5	5	5	0

Depth	S(d)	SRA
1	{C, D}	0.67
2	{B, C, D}	0.77
3	{A, B, C, D}	0.67
4	{A, B, C, D}	0.67
5	{A, B, C, D, E}	0.53

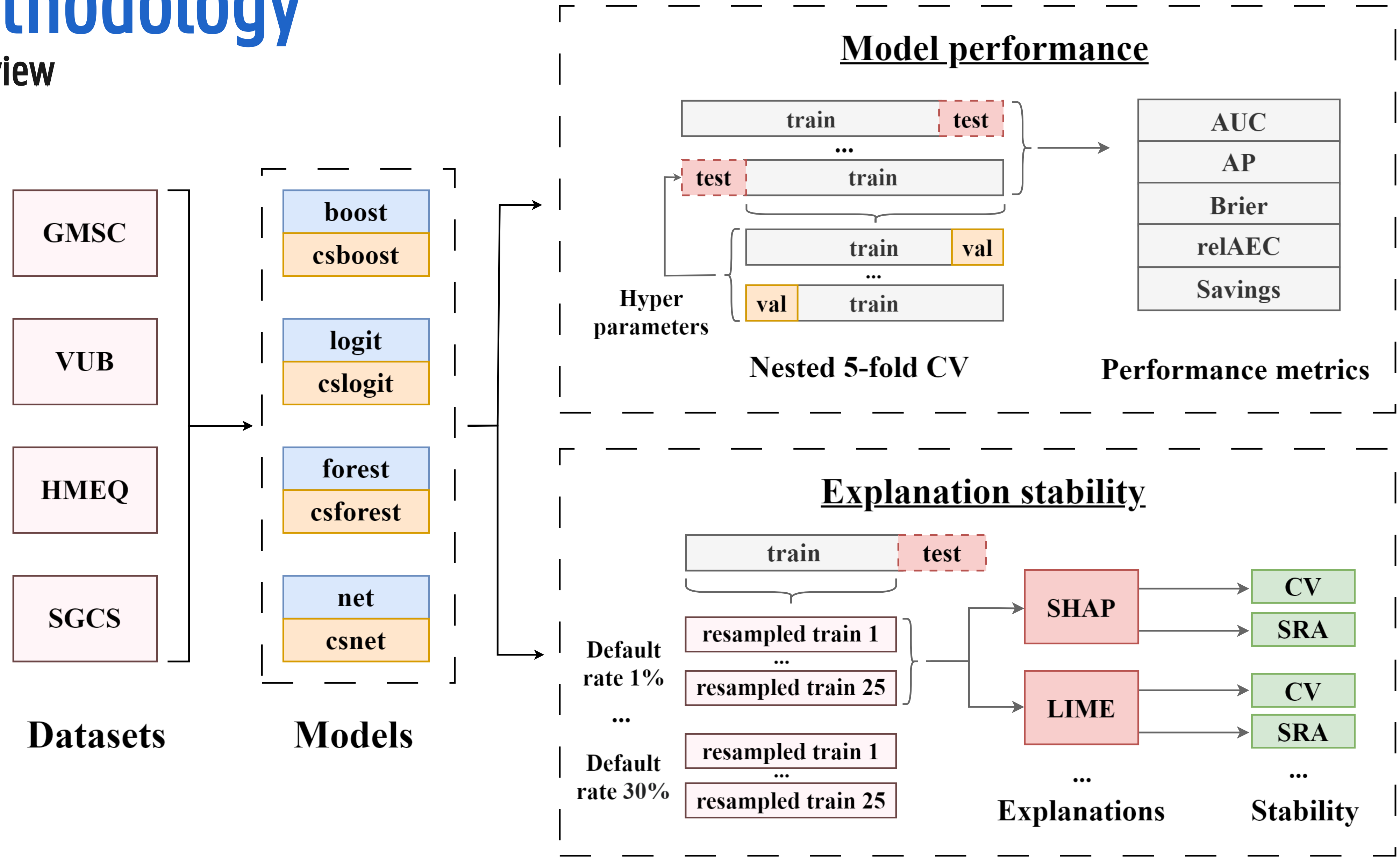
Ranking Stability

Sequential Rank Agreement (SRA)

Ekstrøm, C.T., Gerds, T.A., Jensen, A.K., 2018. Sequential rank agreement methods for comparison of ranked lists. Biostatistics 20, 582598.

Methodology

Overview



GHENT UNIVERSITY

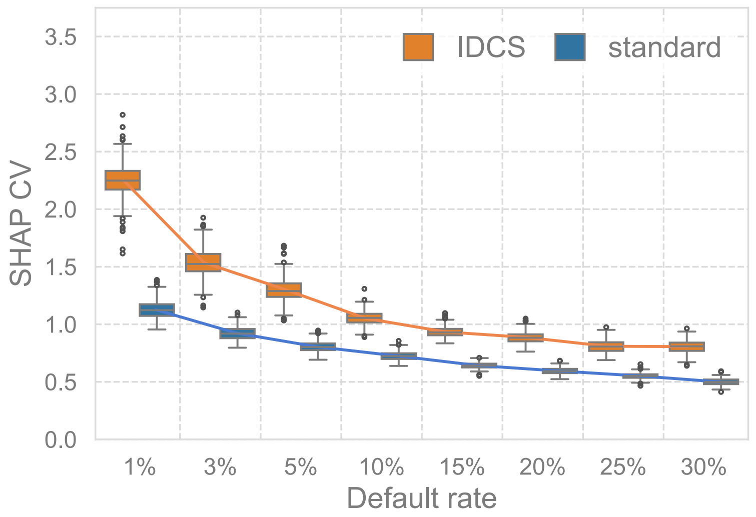
Explanation Stability Results

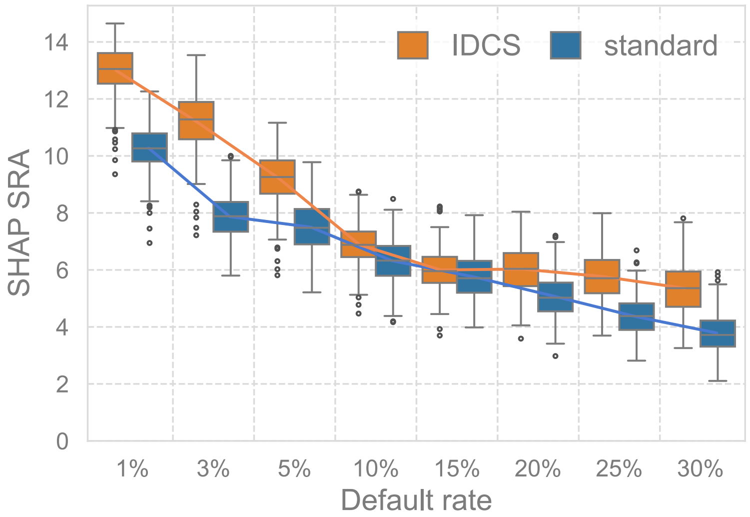
HMEQ dataset

CV

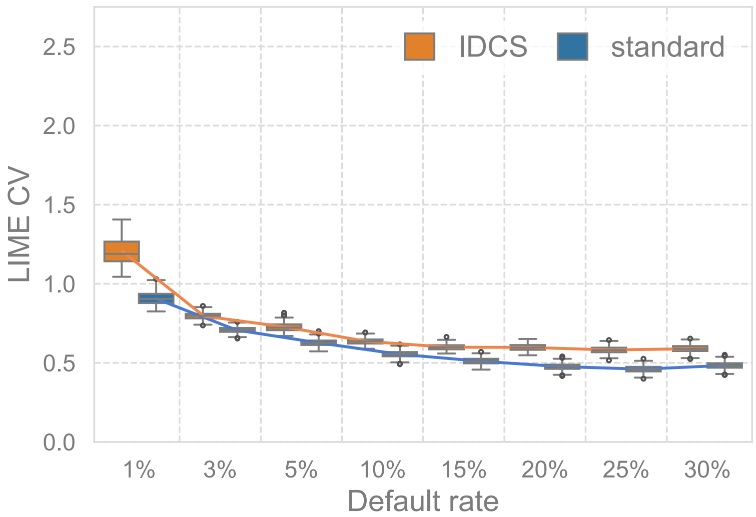
SRA

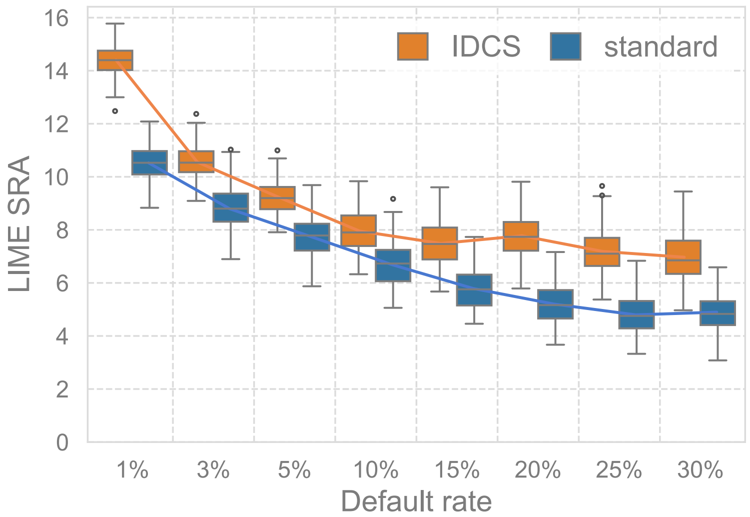
SHAP





LIME





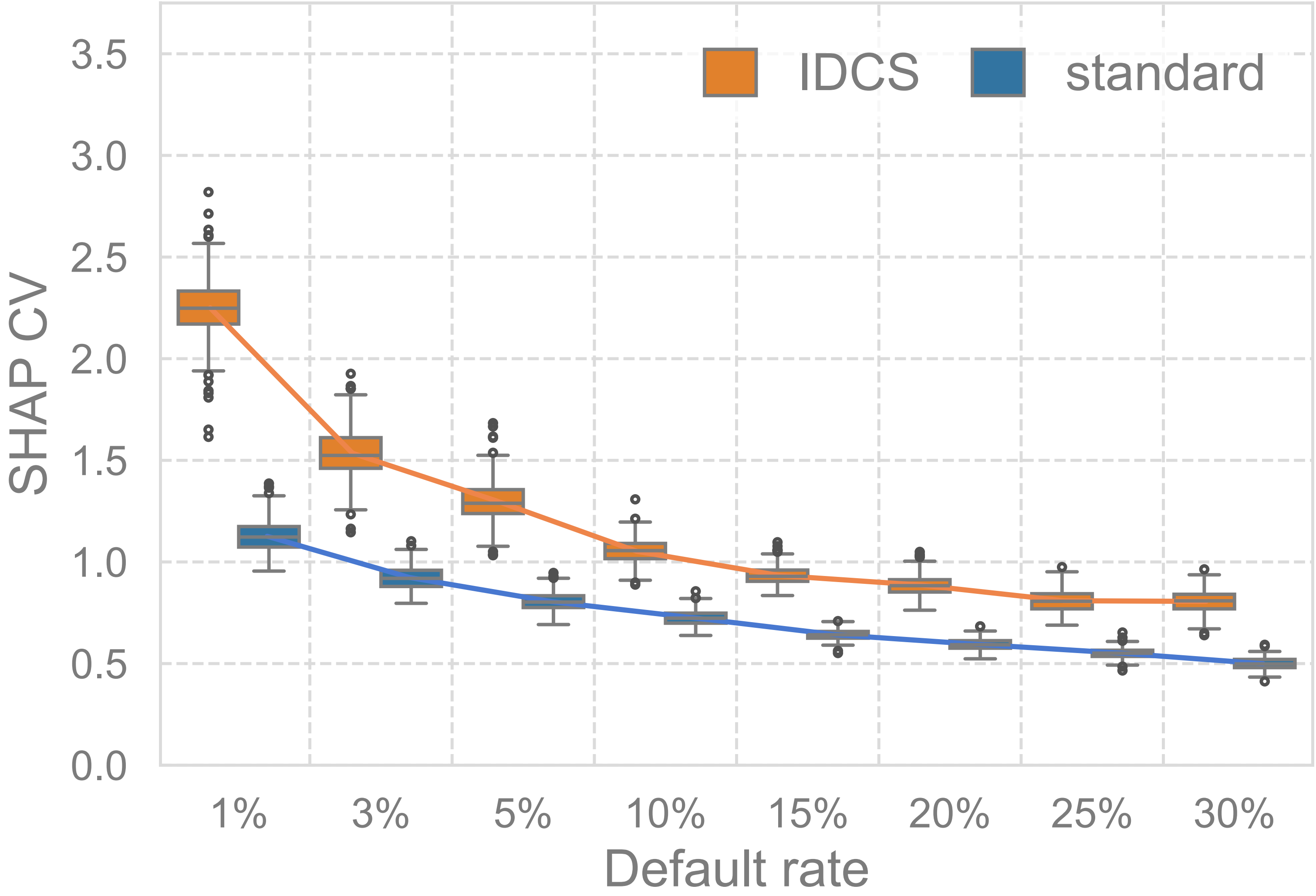
GHENT UNIVERSITY

Explanation Stability Results

HMEQ dataset

CV

SHAP



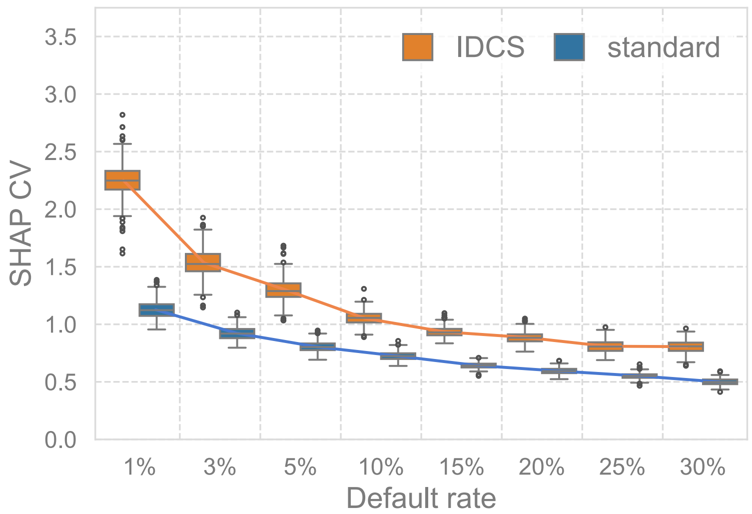
Explanation Stability Results

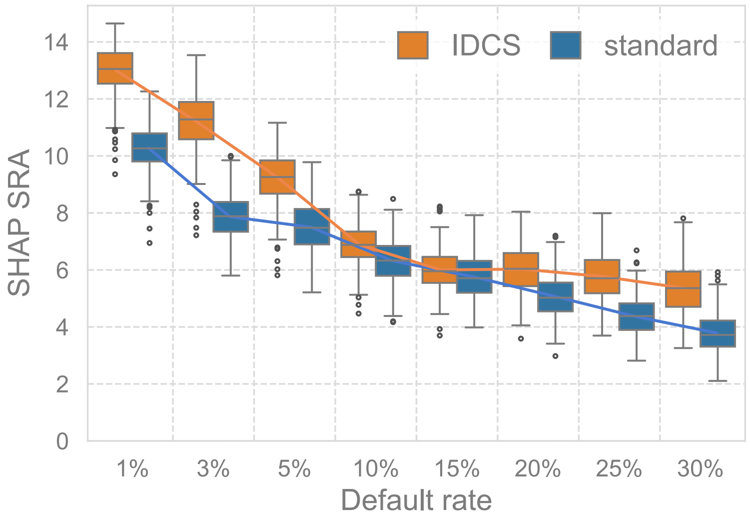
HMEQ dataset

CV

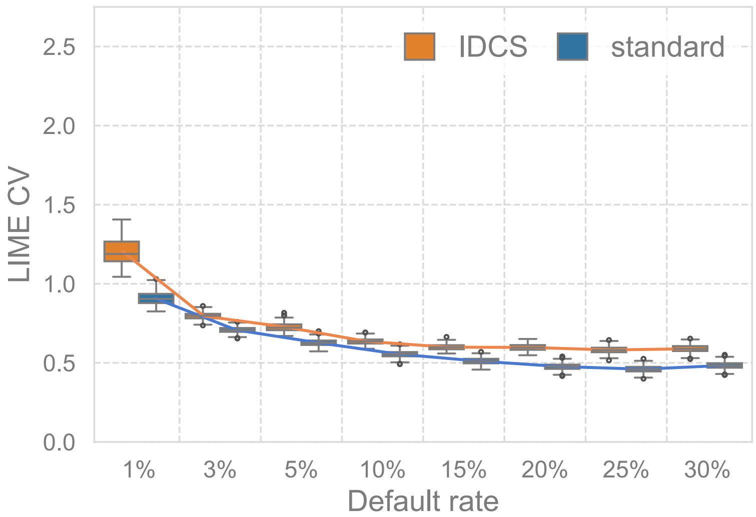
SRA

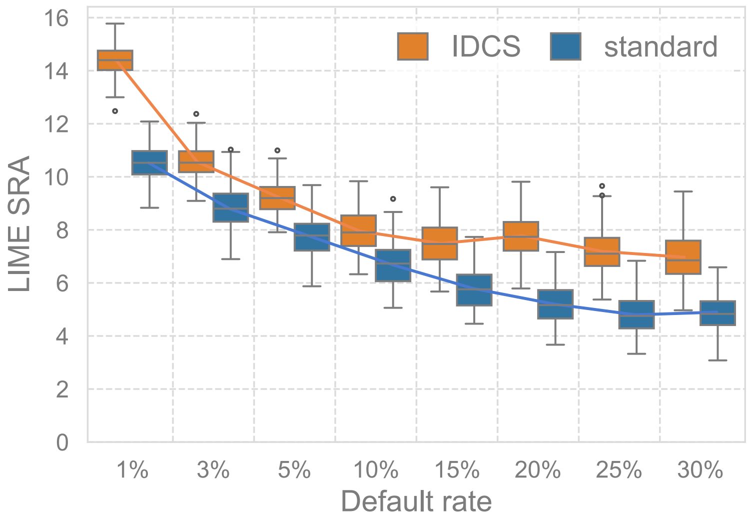
SHAP





LIME

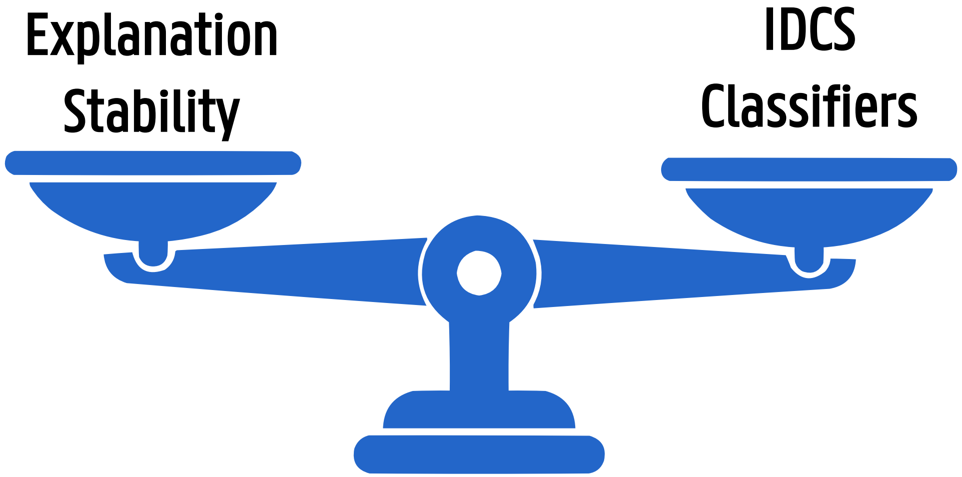




GHENT UNIVERSITY

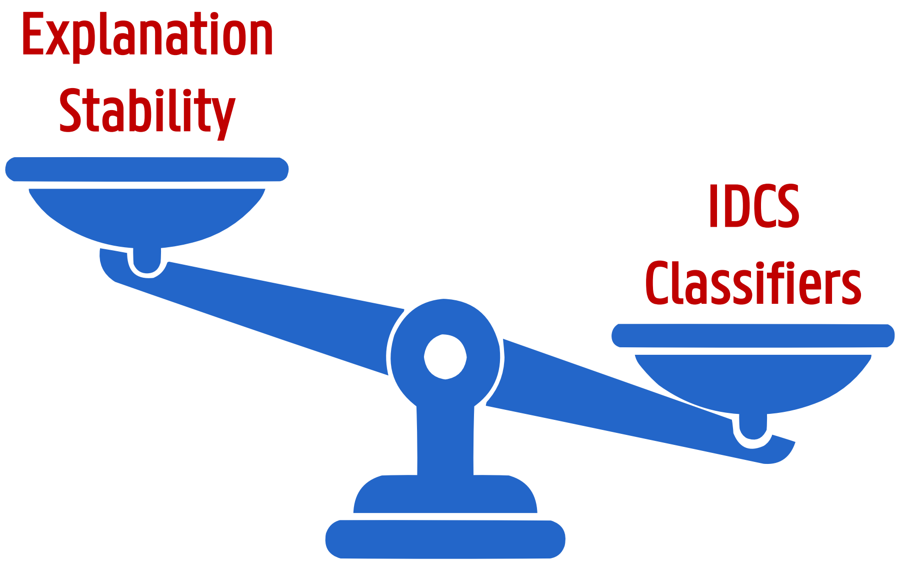
Conclusions

1. There is a trade-off between directly optimizing for costs and providing stable explanations.



Conclusions

1. There is a trade-off between directly optimizing for costs and providing stable explanations.
2. In their current state, IDCS classifiers will fail to meet the regulatory XAI standards to be used in practice.
3. Further research should explore safety measures or remediations to make IDCS model explanations more stable.



GHENT UNIVERSITY

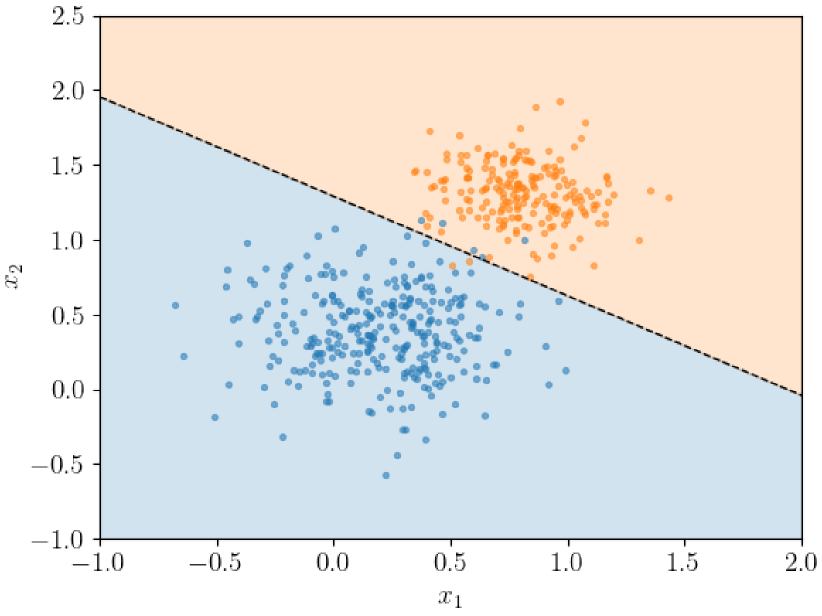
Discussion: the importance of a training sample

Traditional ML models: $s(\boldsymbol{x}) = \mathbb{E}\,(\mathrm{Y} \mid \boldsymbol{x})$

- Feature distribution
- Class imbalance (cfr Chen et al.)

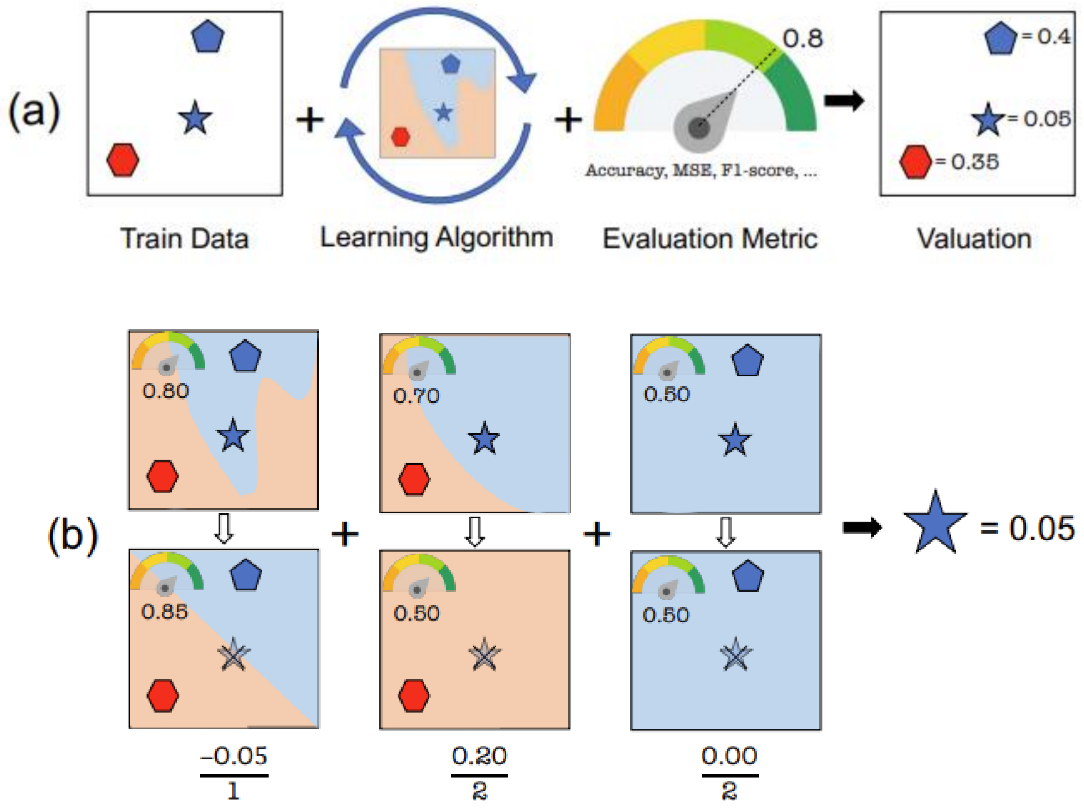
IDCS ML models: $\mathrm{AEC}(\mathrm{Y}, s(\boldsymbol{x}), \mathrm{C})$

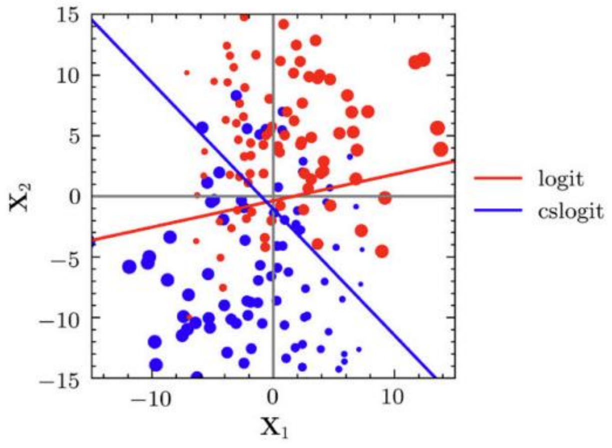
- Feature distribution
- Class imbalance
- **COST**



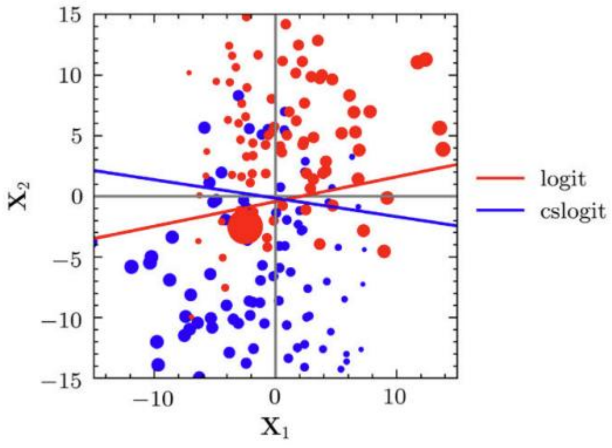
Chen, Y., Calabrese, R., Martin-Barragan, B., 2024. Interpretable machine learning for imbalanced credit scoring datasets.

Discussion: IDCSL and data valuation





(a) Example on synthetic data:
Optimal behavior of logit and cslogit



(b) Example on synthetic data:
Influence of outliers on cslogit

Ghorbani, A., & Zou, J. (2019). Data shapley: Equitable valuation of data for machine learning. In International conference on machine learning (pp. 2242-2251). PMLR.
De Vos, S., Vanderschueren, T., Verdonck, T., & Verbeke, W. (2023). Robust instance-dependent cost-sensitive classification. Advances in data analysis and classification, 17(4), 1057-1079.

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

Matteo Ballegeer

PhD Researcher

Research group Data Analytics
matteo.ballegeer@ugent.be

Read the study here

European Journal of Operational Research 326 (2025) 630–640

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Analytics, Computational Intelligence and Information Management

Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring

Matteo Ballegeer *, Matthias Bogaert, Dries F. Benoit

Ghent University, Research Group Data Analytics, Faculty of Economics and Business Administration, Tweekerkenstraat 2, 9000 Gent, Belgium
FlandersMake@UGent—Corelab CVAMO, Tweekerkenstraat 2, 9000 Gent, Belgium

GHENT UNIVERSITY

CRC Credit Research Centre

FLANDERS MAKE
DRIVING INNOVATION IN MANUFACTURING