

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

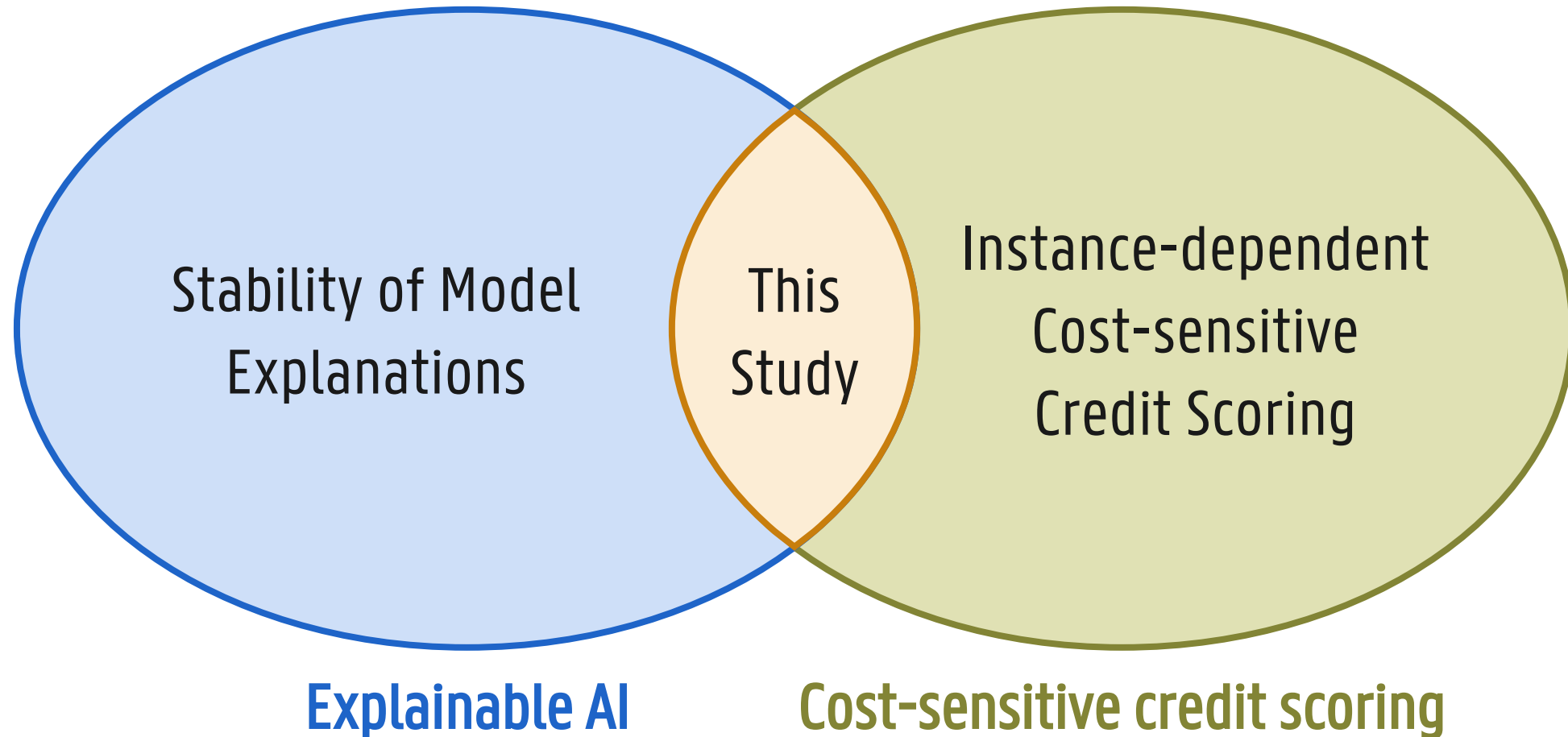
Matteo Ballegeer, Matthias Bogaert, Dries F. Benoit

August 28th, 2025

GHENT UNIVERSITY - RESEARCH GROUP DATA ANALYTICS

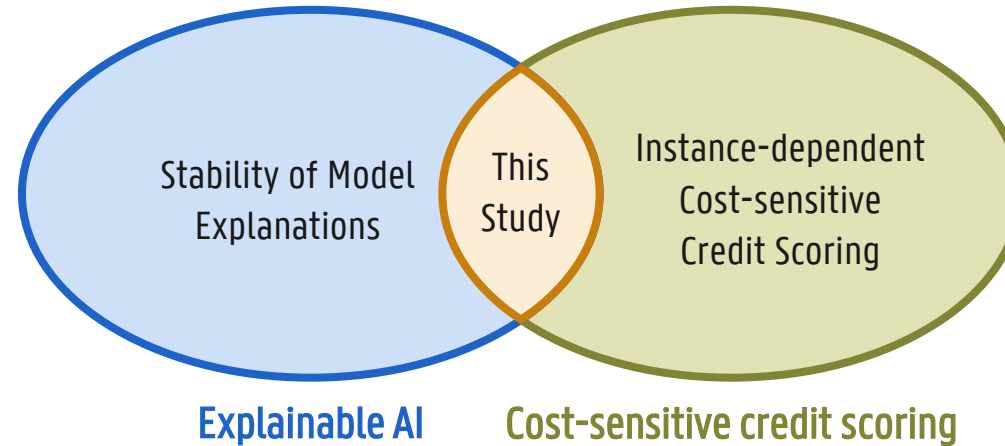
Research Positioning and Contributions

A bridge between model explainability and cost-sensitive learning



Research Positioning and Contributions

A bridge between model explainability and cost-sensitive learning

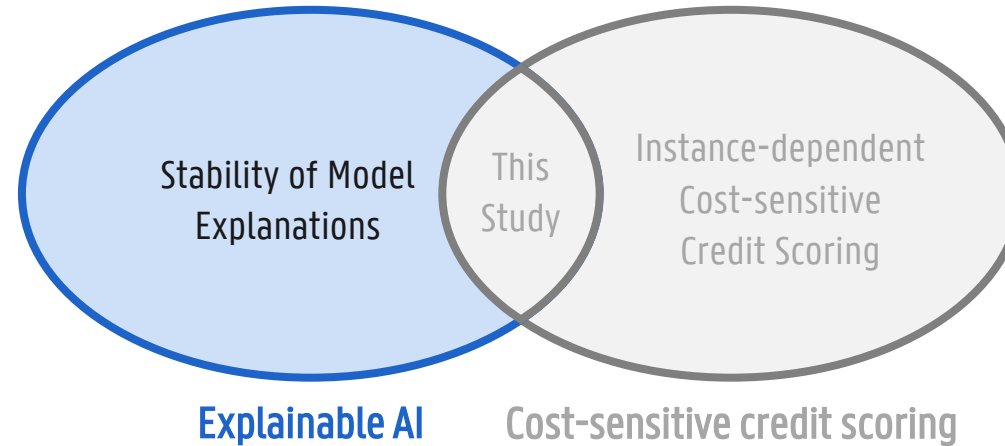


Main Contributions

- 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 stability¹ is amplified when using IDCS classifiers.
- 4 Introduce a novel, cross-dataset comparable evaluation metric for cost-sensitive learning.

Research Positioning and Contributions

A bridge between model explainability and cost-sensitive learning

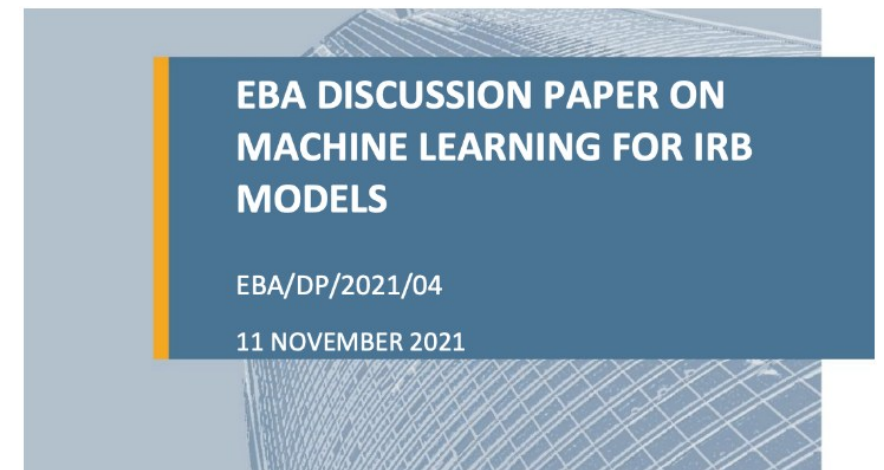
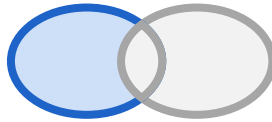


Main Contributions

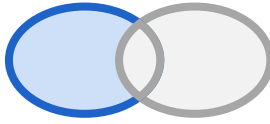
- 1 First credit scoring study relating explainable AI to cost-sensitive learning.
- 2 Demonstrate the impact of IDCS classifiers on the stability of post-hoc explainable AI techniques.
- 3 Show that a known negative impact of class imbalance on explanation stability¹ is amplified when using IDCS classifiers.
- 4 Introduce a novel, cross-dataset comparable evaluation metric for cost-sensitive learning.

Explainable AI

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



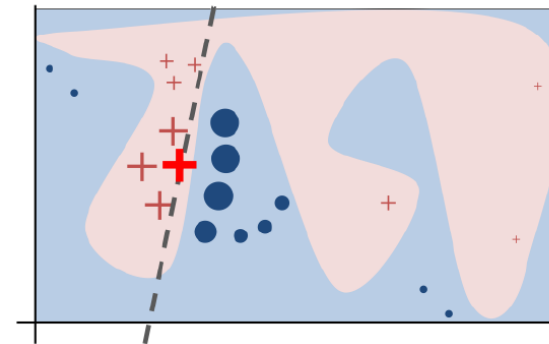
Explainable AI



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



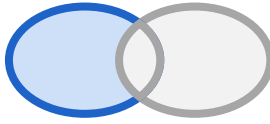
SHAP



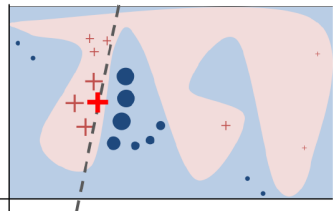
LIME

Explainable AI

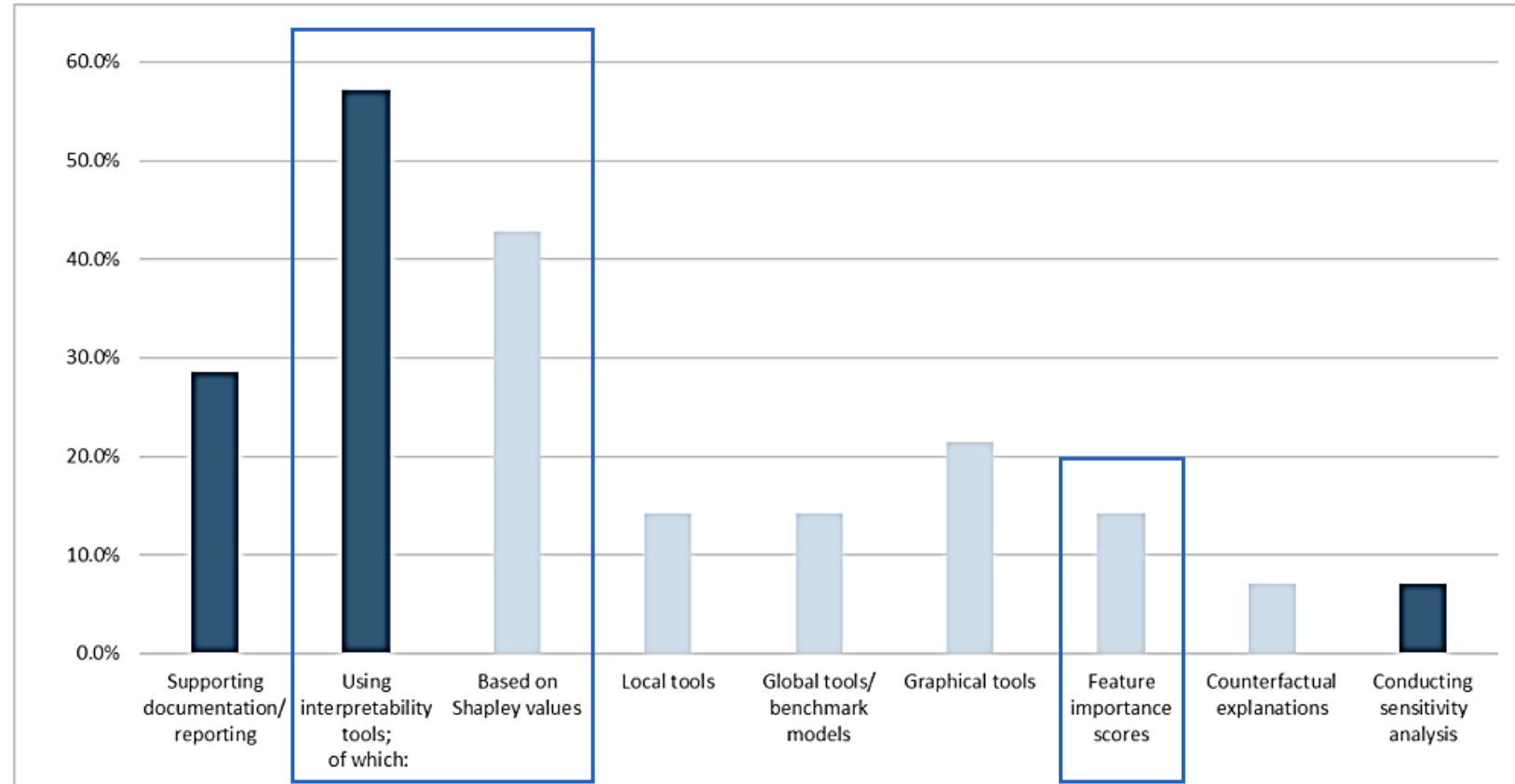
In response, XAI techniques to explain black-box models are rising in popularity



SHAP

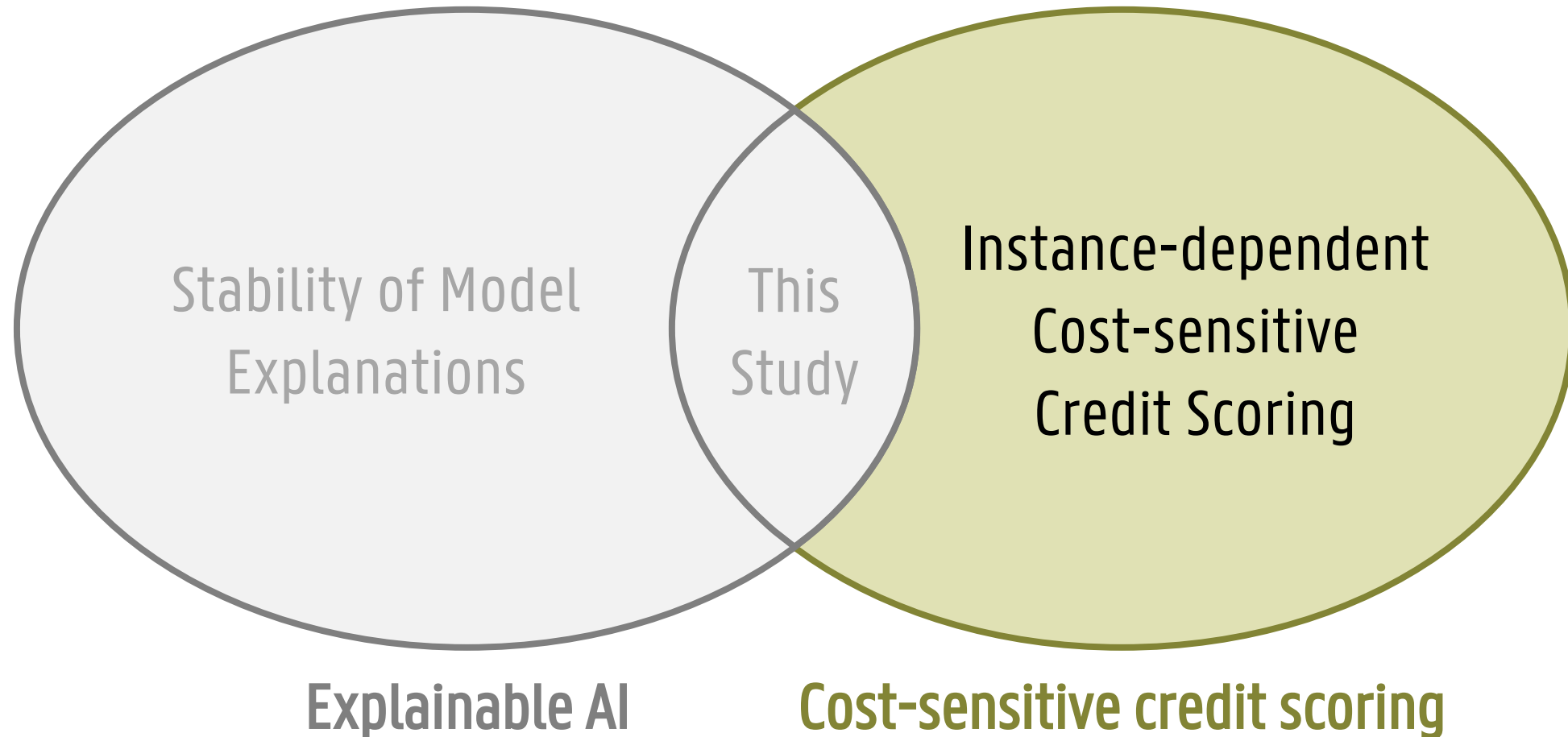


LIME



Research Positioning and Contributions

A bridge between model explainability and cost-sensitive learning



Cost-sensitive Learning

Different misclassifications do not result in equal costs



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

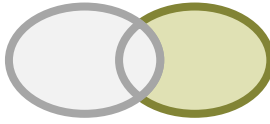
Symmetrical misclassification costs



		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**

Instance-dependent Cost-sensitive Learning



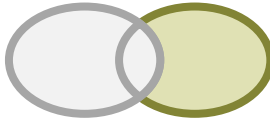
The requested loan amounts (and associated cost) between customers can vary a lot

		Actual		→			Actual	
		No Default	Default				No Default	Default
Predicted	No Default	Cost (0 0)	Cost (0 1)		Predicted	No Default	Cost _i (0 0)	Cost _i (0 1)
	Default	Cost (1 0)	Cost (1 1)			Default	Cost _i (1 0)	Cost _i (1 1)

Asymmetrical at the class level

Asymmetrical at the instance level

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	$Cost_i(1 0)$	$Cost_i(1 1)$

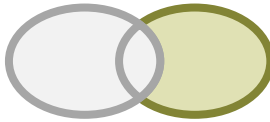


		Actual	
		No Default	Default
Predicted	No Default	0	$Amount_i \cdot LGD$
	Default	$r_i + Cost_{alt}$	0

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

Asymmetrical at the instance level

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	$Cost_i(1 0)$	$Cost_i(1 1)$

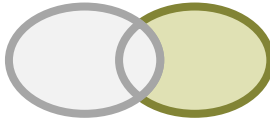


		Actual	
		No Default	Default
Predicted	No Default	0	$Amount_i \cdot LGD$
	Default	$r_i + Cost_{alt}$	0

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

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

Instance-dependent Cost-sensitive Credit Scoring



Relative AEC is introduced as dimensionless adaptation of AEC

$$\text{Average Expected Cost (AEC)}_i = y_i \cdot [(1 - s(\text{default})_i) \cdot (\text{Amount}_i \cdot \text{LGD})] \\ + (1 - y_i) \cdot [s(\text{default})_i \cdot (r_i + (-\bar{r} \cdot \pi_0 + \overline{\text{Amount}} \cdot \text{LGD} \cdot \pi_1))]$$



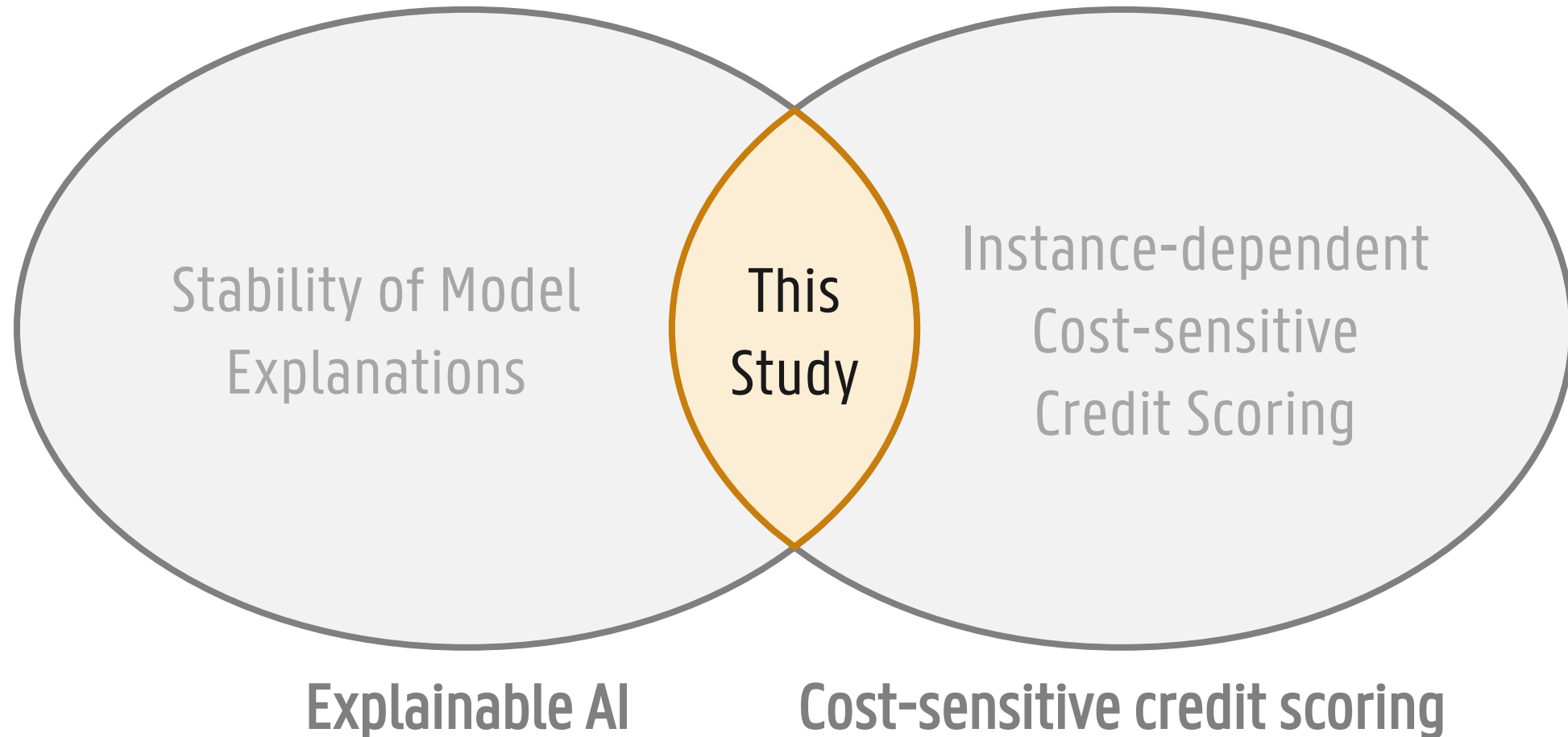
$$\text{Relative Average Expected Cost (relAEC)}_i = 1 - \frac{\text{AEC}(y_i, s(\text{default})_i, \text{Amount}_i)_i}{\text{AEC}(y_i, \pi_1, \text{Amount}_i)}$$

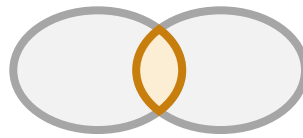


Similar interpretation as the **savings** metric!

Research Positioning and Contributions

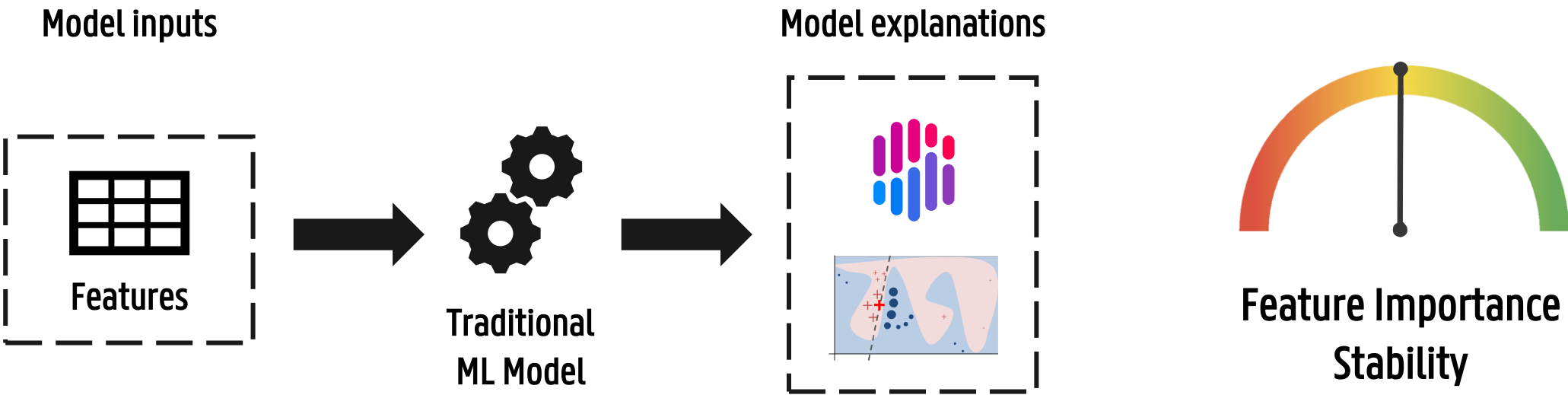
A bridge between model explainability and cost-sensitive learning

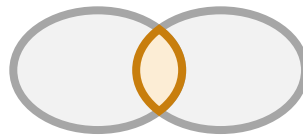




Explanation Stability in IDCS Credit Scoring

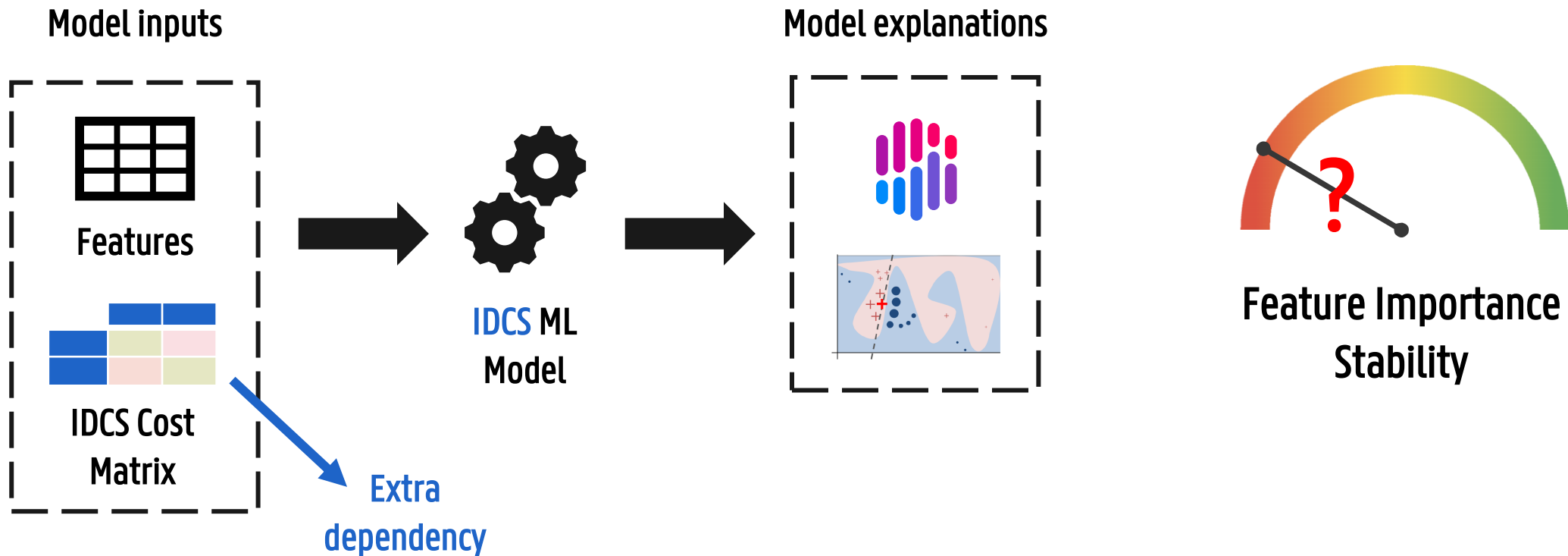
Does optimizing for an IDCS loss function impact the stability of model explanations?





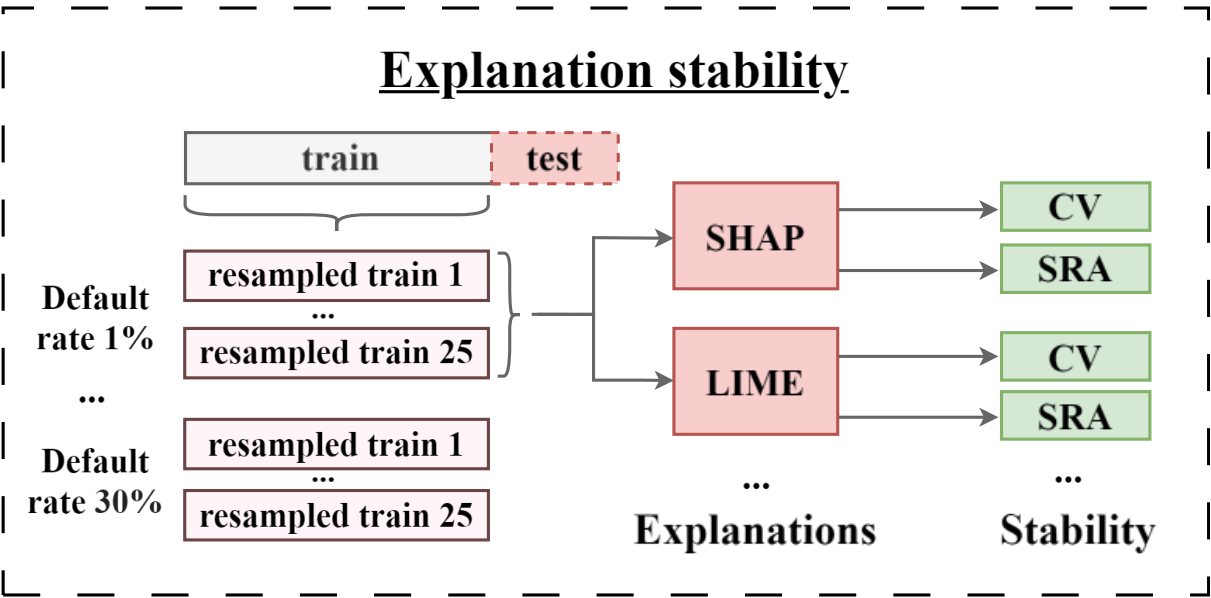
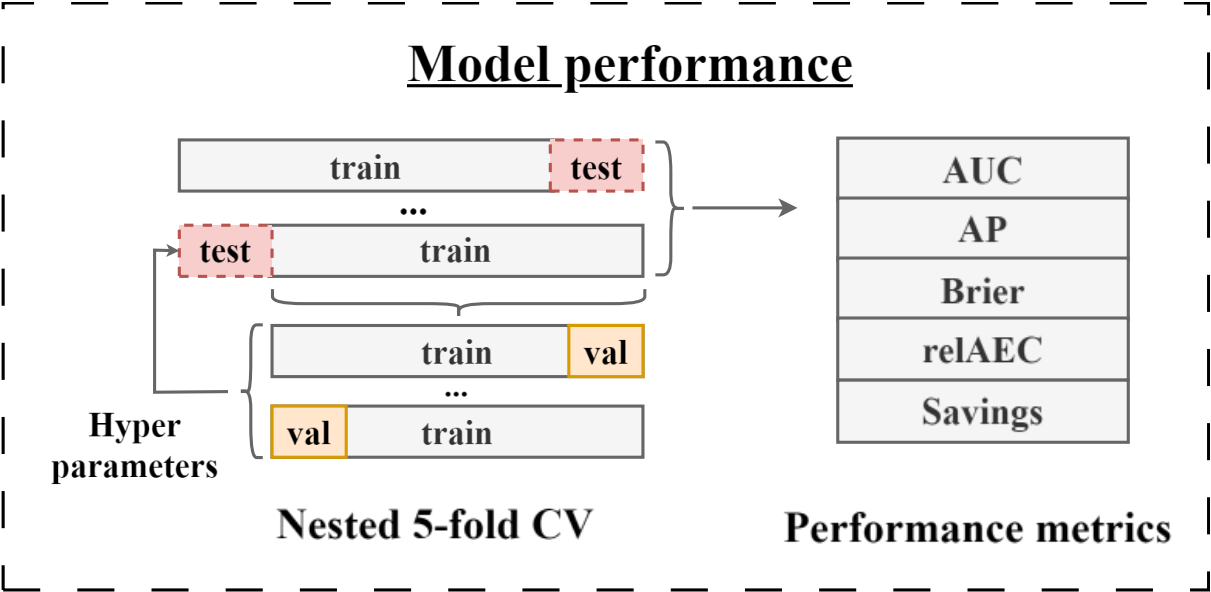
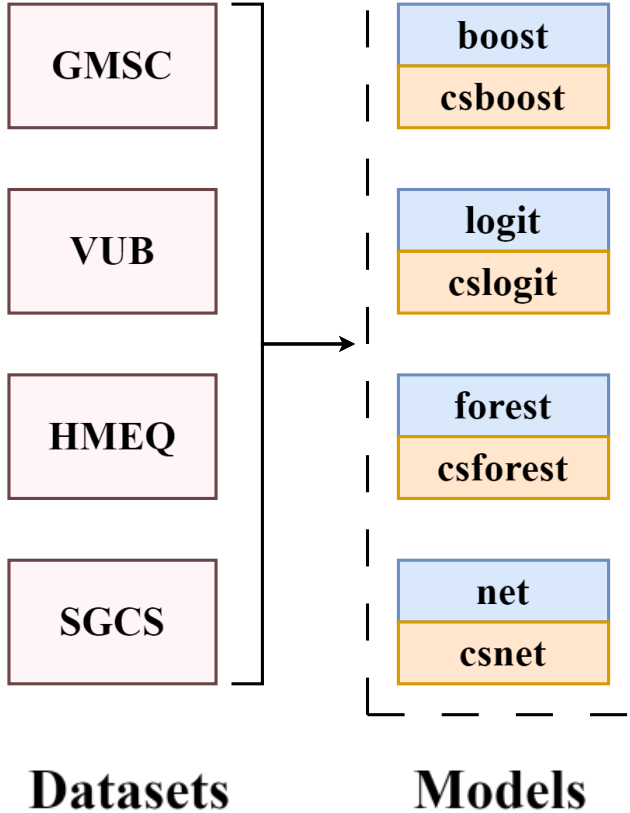
Explanation Stability in IDCS Credit Scoring

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



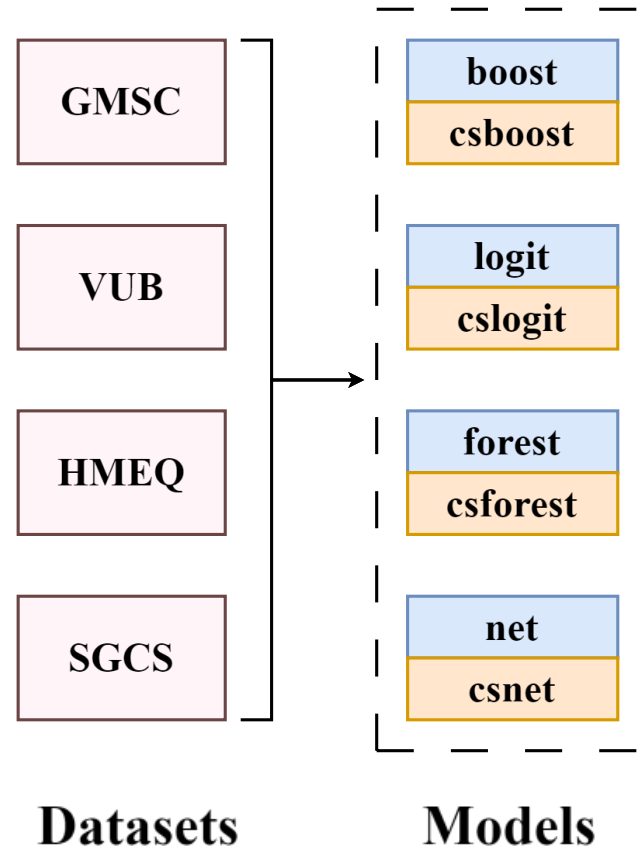
Methodology

Overview



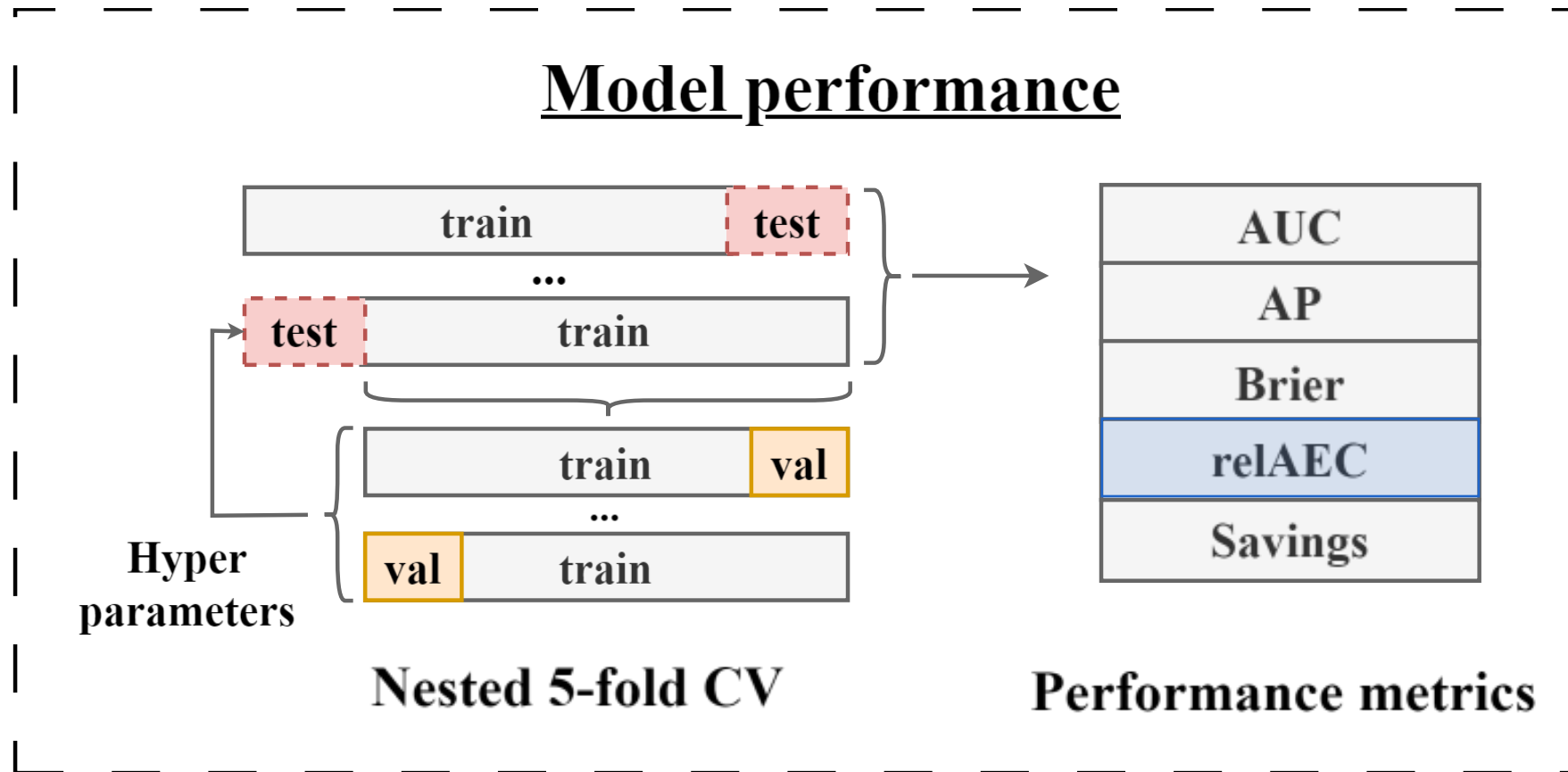
Methodology

Datasets and models



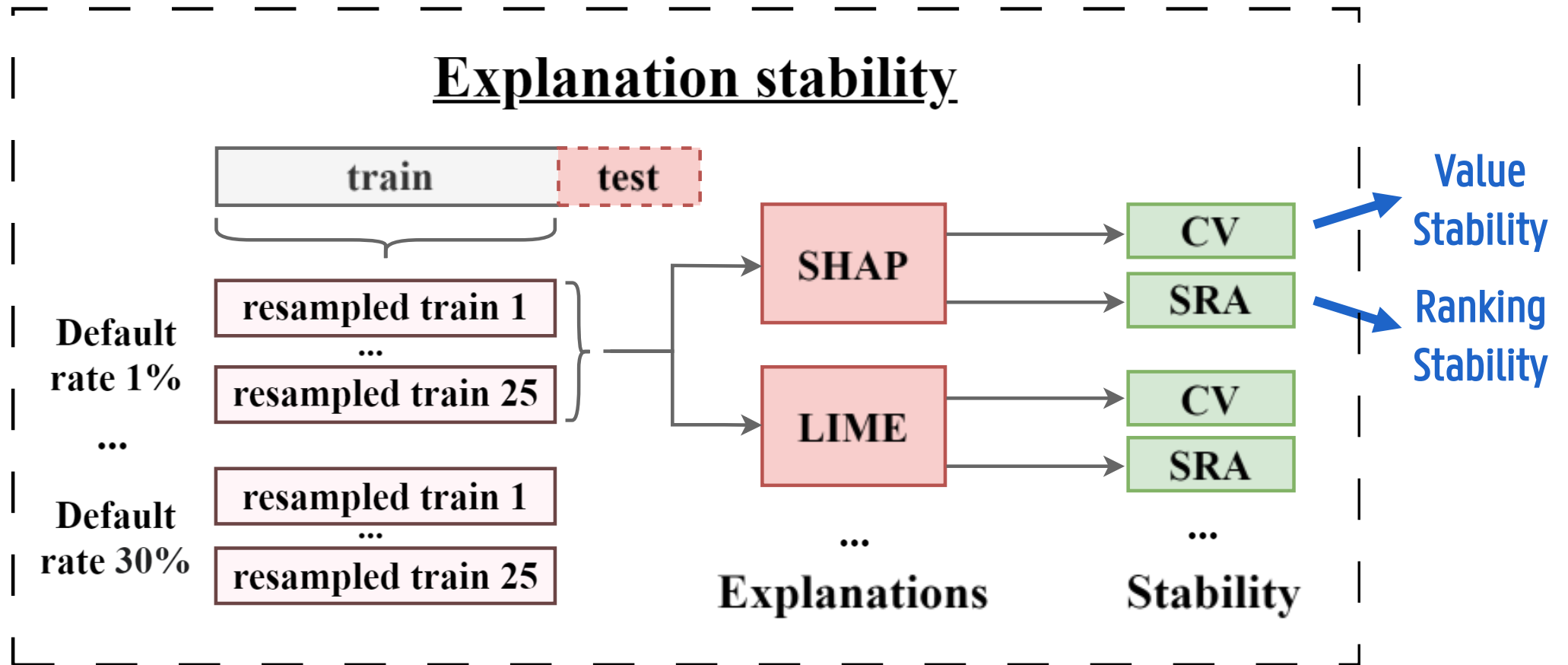
Methodology

Model performance



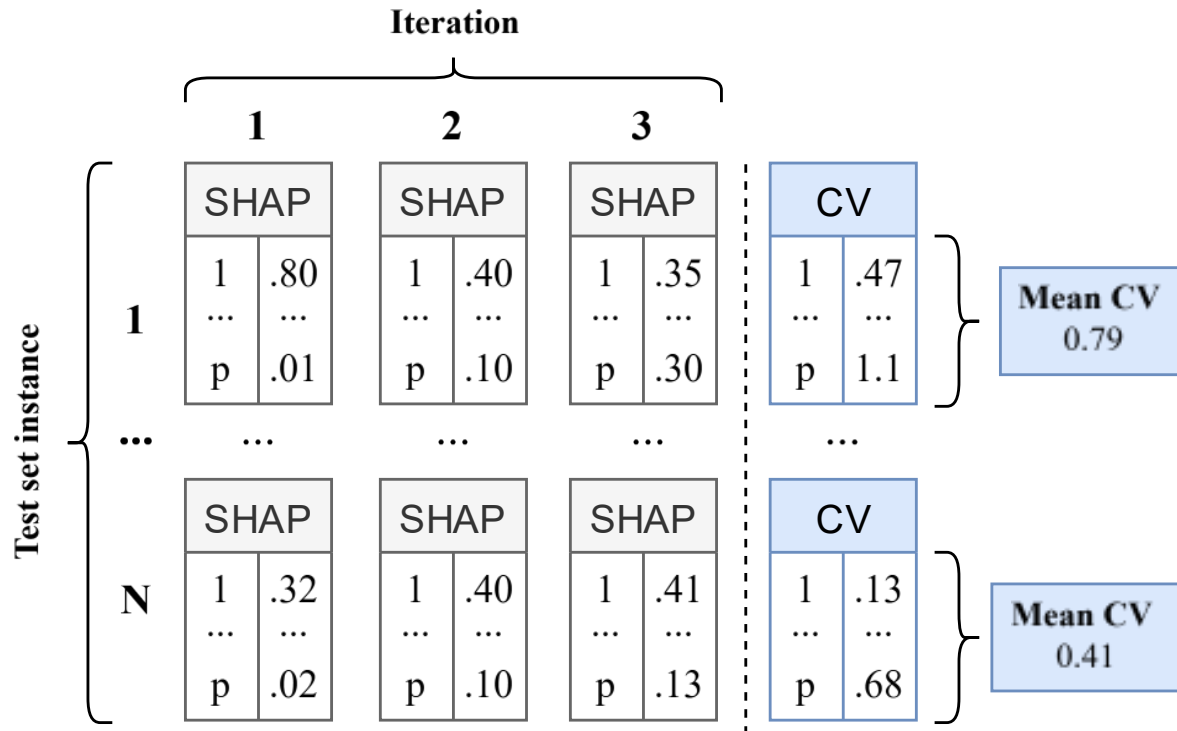
Methodology

Explanation stability



Methodology

Stability metrics



Value Stability

Coefficient of Variation (CV)

X_p	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

Ranking Stability

Sequential Rank Agreement (SRA)

Methodology

Stability metrics

X_p	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



X_p	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)

Methodology

Stability metrics

X_p	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



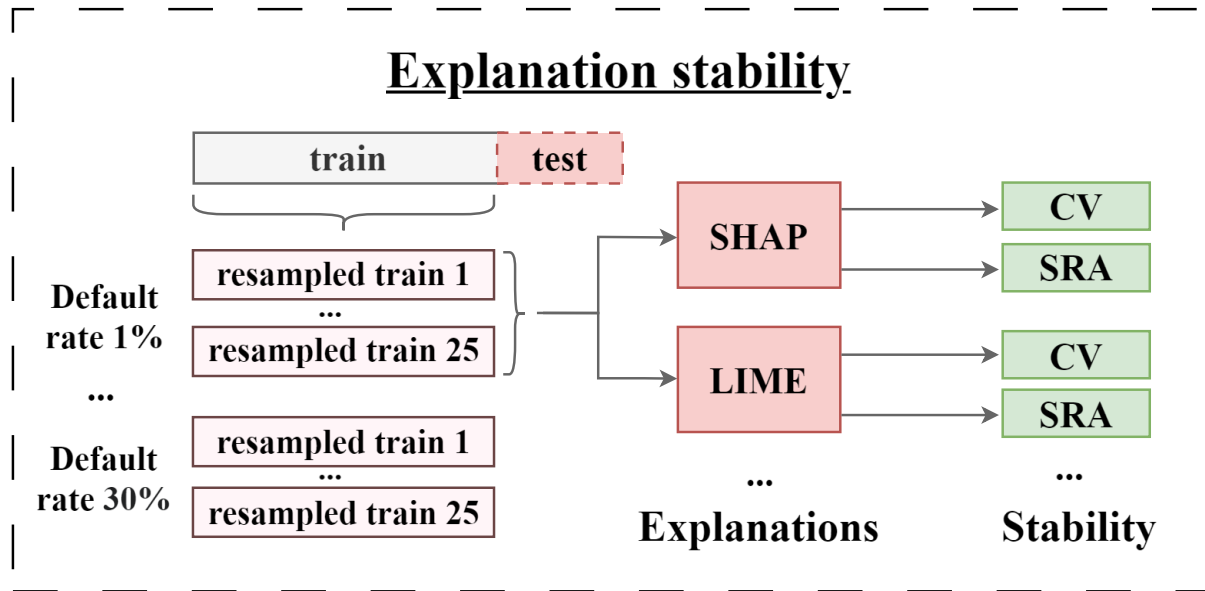
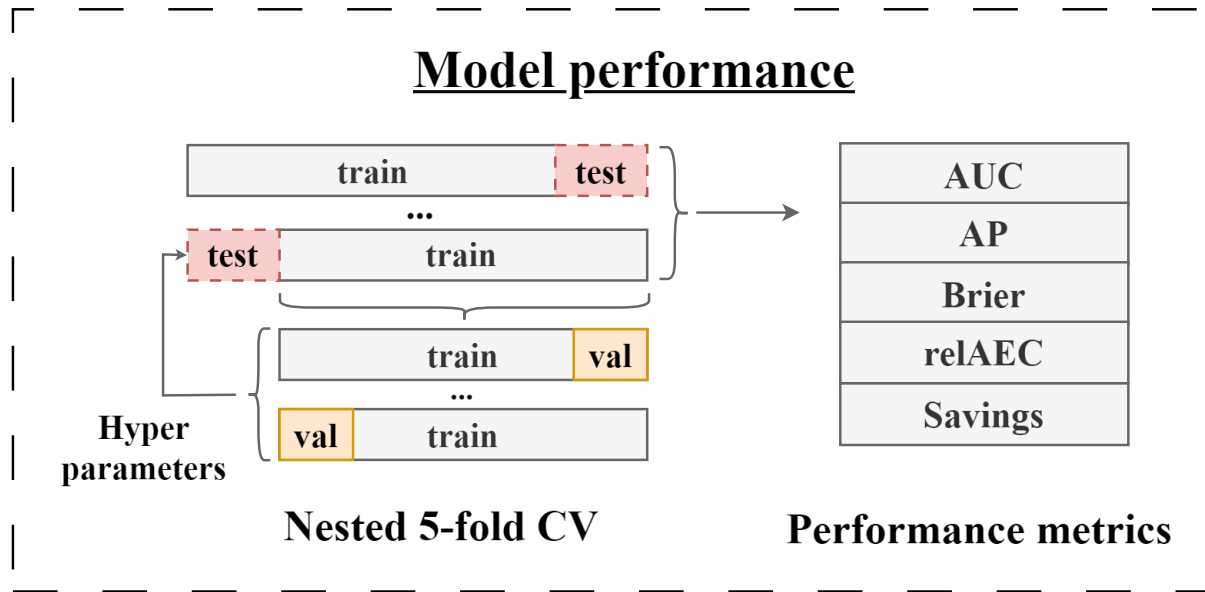
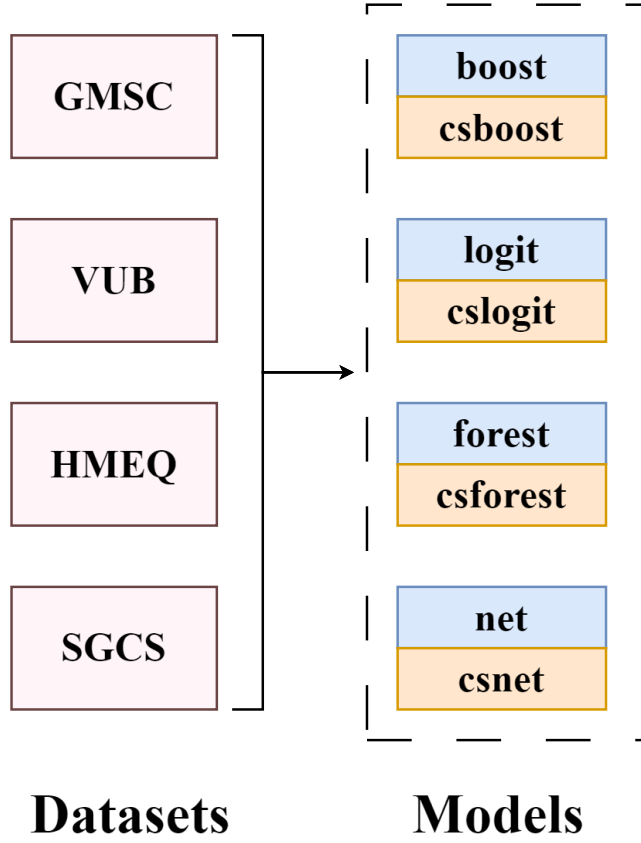
Depth	$S(d)$	SRA
1	{C, D}	0.67
2	{B, C, D}	0.77
3	{A, B, C, D}	0.67
4	{A, B, C, D}	0.67
5	{A, B, C, D, E}	0.53

Ranking Stability

Sequential Rank Agreement (SRA)

Methodology

Overview

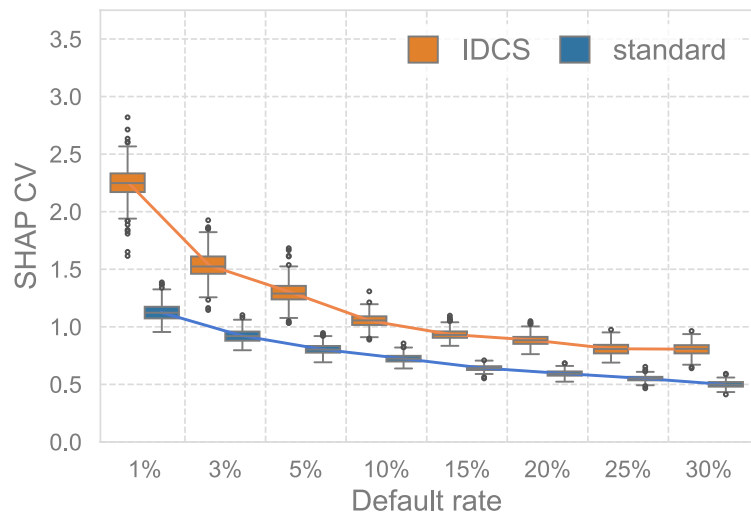


Explanation Stability Results

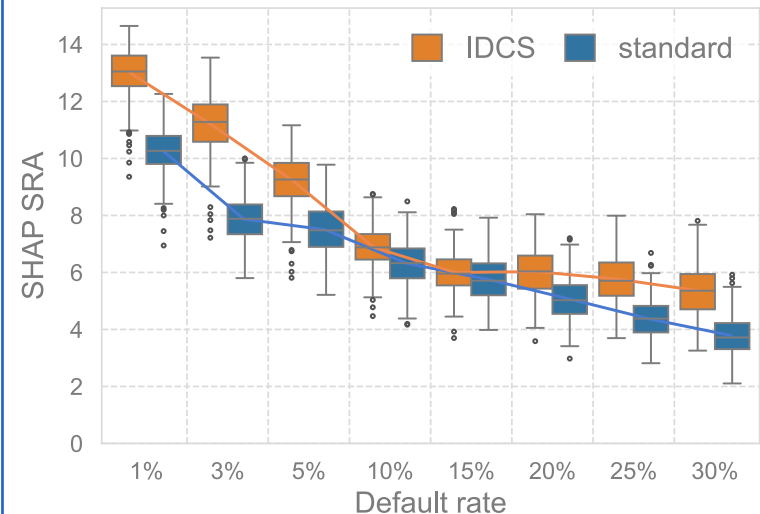
HMEQ dataset

SHAP

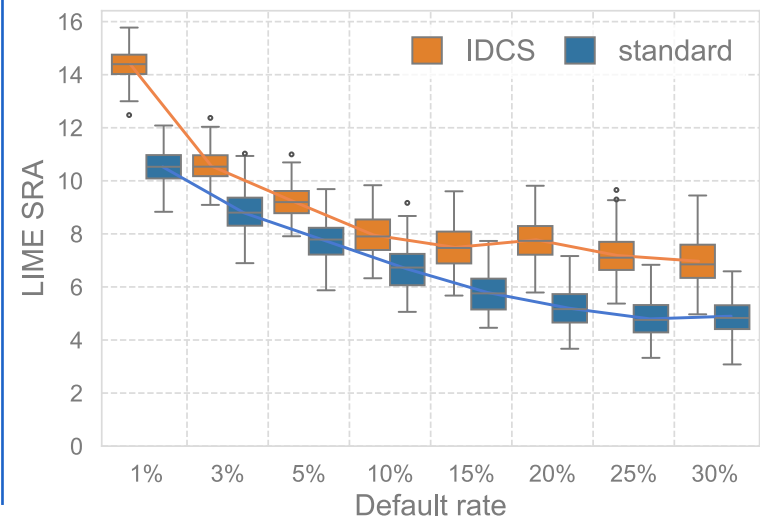
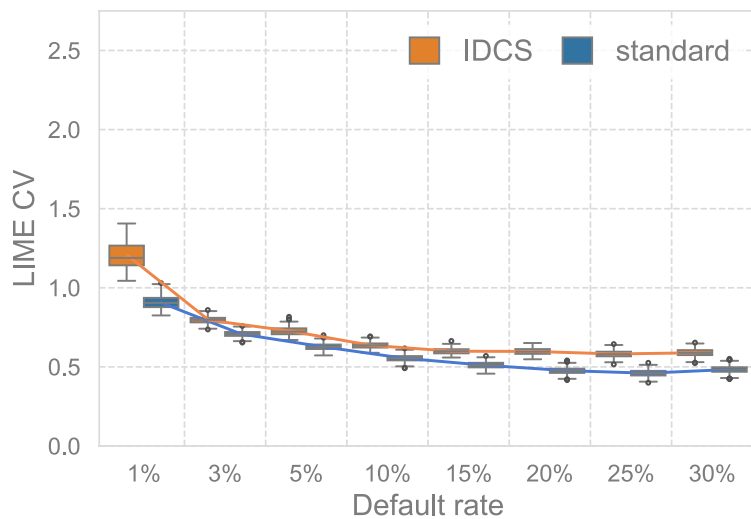
CV



SRA



LIME

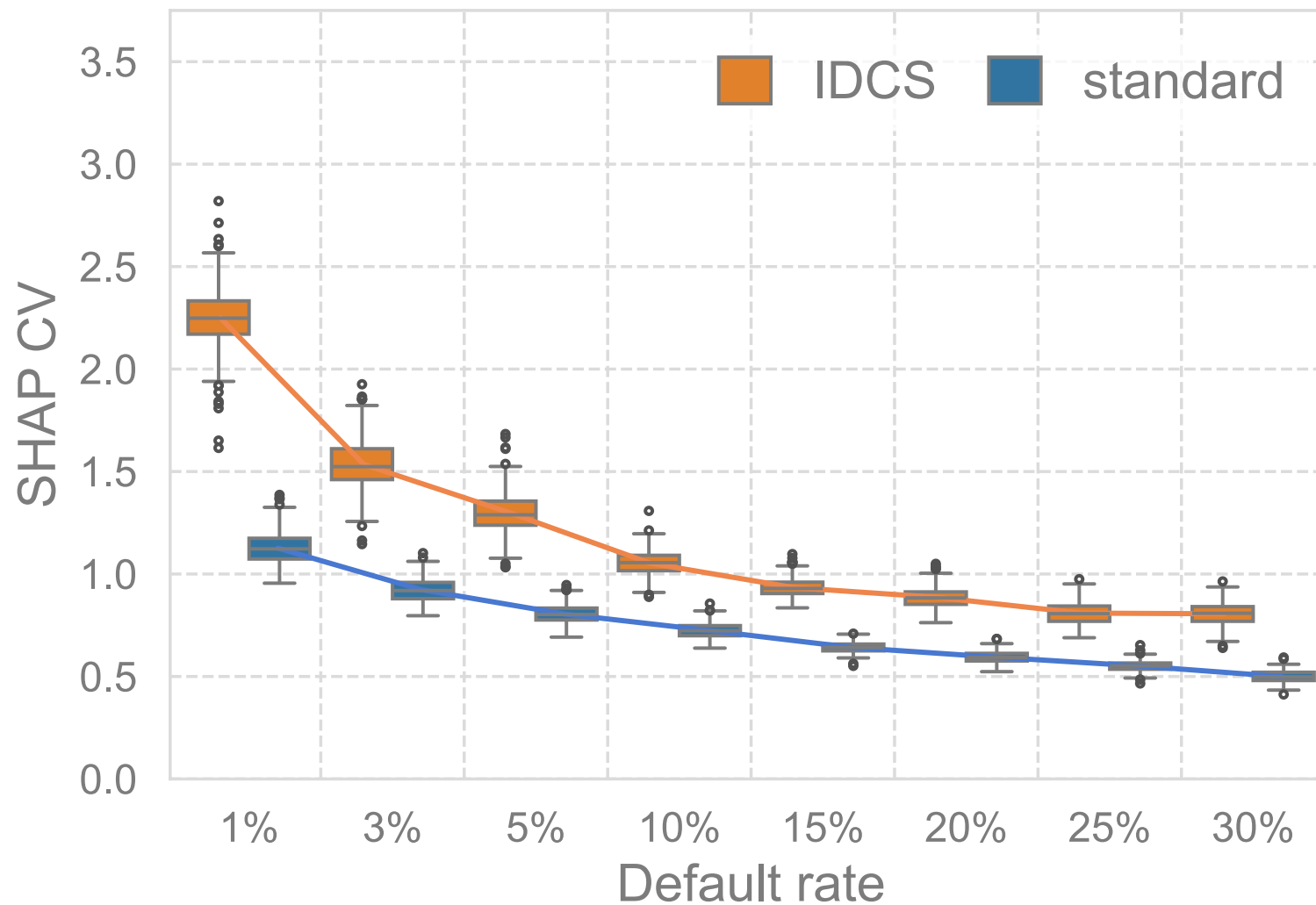


Explanation Stability Results

HMEQ dataset

SHAP

CV

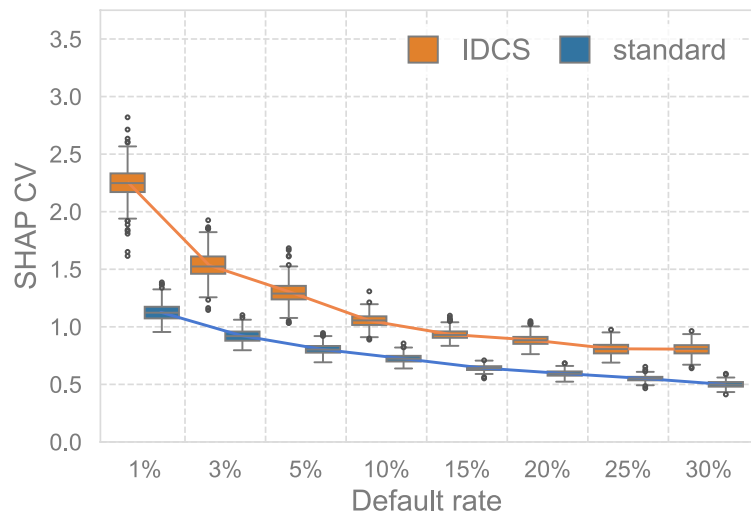


Explanation Stability Results

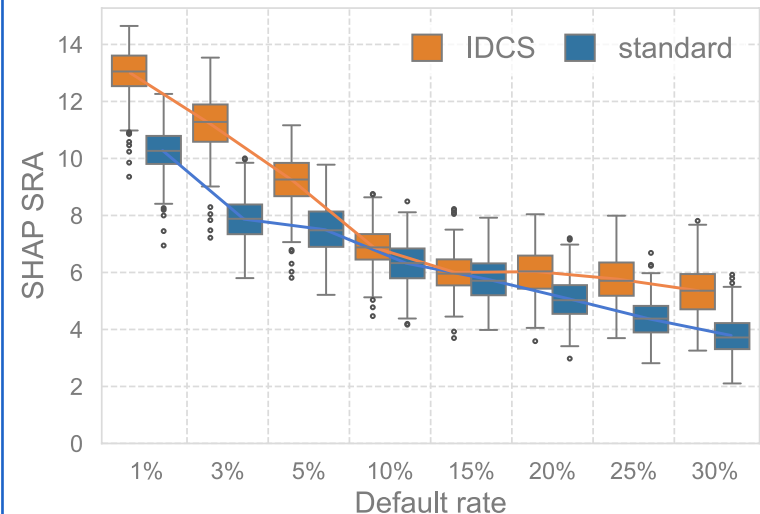
HMEQ dataset

SHAP

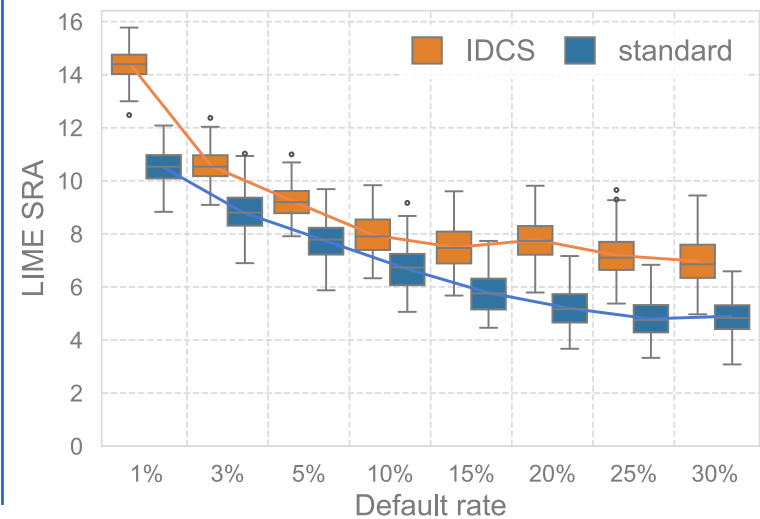
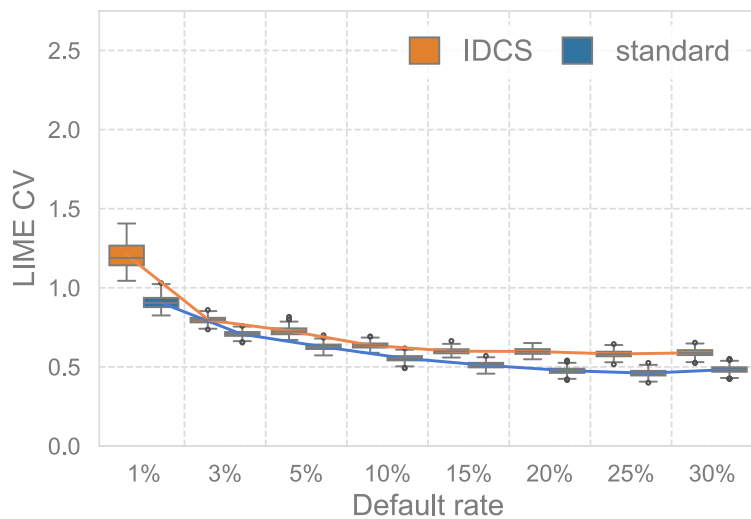
CV



SRA

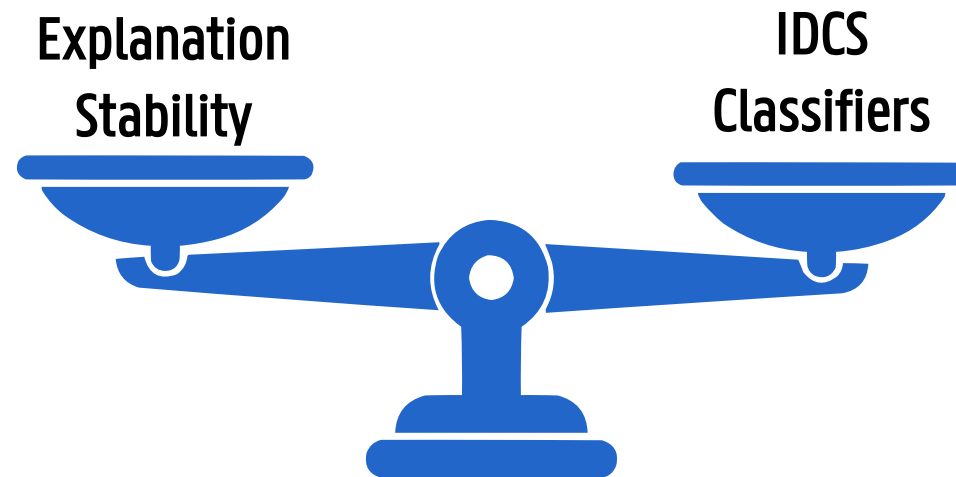


LIME



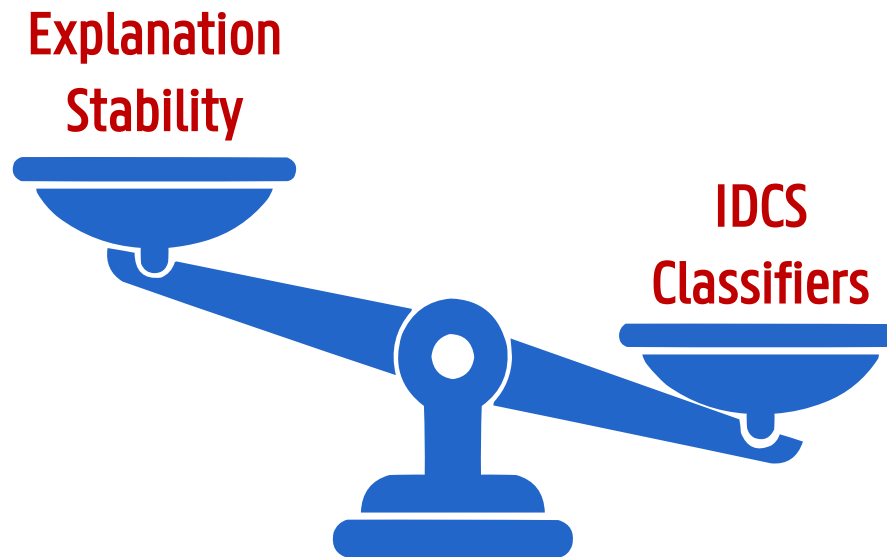
Conclusions

- 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.



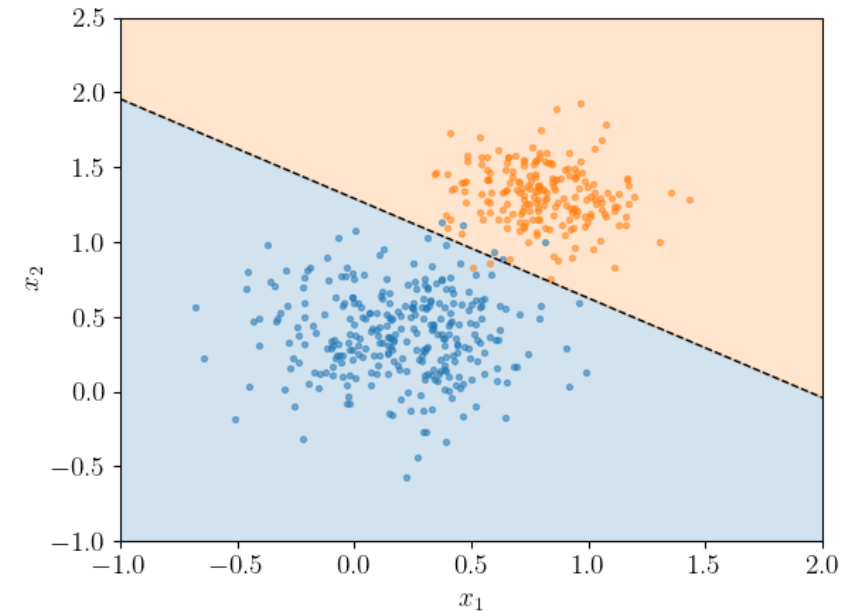
Discussion: the importance of a training sample

Traditional ML models: $s(\mathbf{x}) = \mathbb{E}(Y | \mathbf{x})$

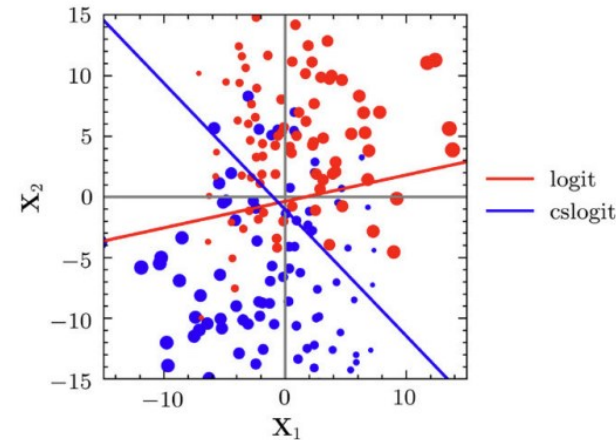
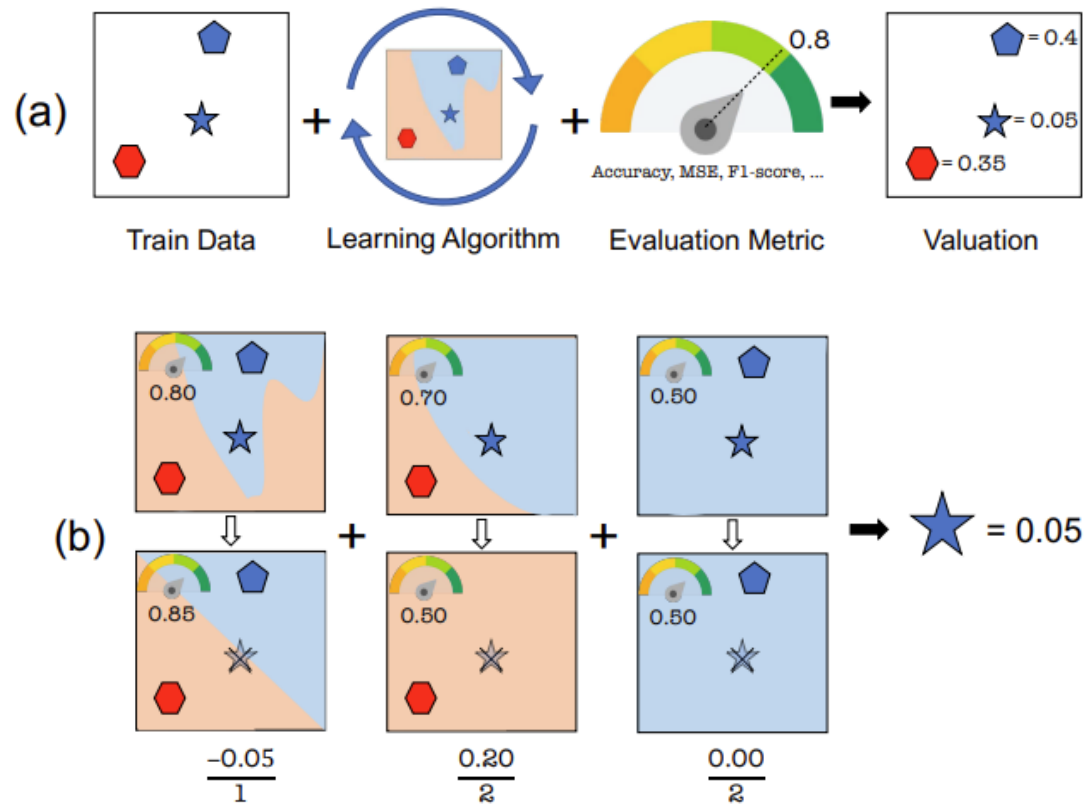
- Feature distribution
- Class imbalance (cfr Chen et al.)

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

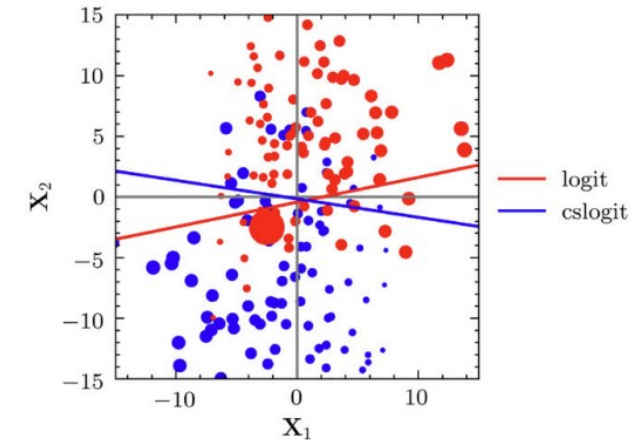
- Feature distribution
- Class imbalance
- **COST**



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



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

ELSEVIER

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, Tweeterkerstraat 2, 9000 Gent, Belgium
FlandersMake@UGent—Corelab CVAMO, Tweeterkerstraat 2, 9000 Gent, Belgium

