From Credit Risk to Explainable AI Research

FABRIZIO RUSSO

UCL - 20.05.22

Content

- Overview of modelling data from Credit Risk Agencies (CRAs)
- Credit Risk Modelling for Retail Application: from regression to machine learning
- Ongoing Research: Causal and Explainable Neural Networks
- Final Remarks

CRA data*

WHAT INFORMATION DO BANKS USE TO ASSESS YOUR CREDIT WORTHINESS?

*MATERIAL FROM RUSSO, 2019

Application Credit Checks Data

Credit Scores

Closed User Group Information (CUG)

- Number of accounts
- Outstand-ing balance
- Repay-ment behaviour
- Types of accounts
- Credit card utillisation

Credit Searches

- Recent credit activity

Public Data

- Electoral roll
- Court judgeme-nts
- Bankrupt-cies

Associates

- Financial associates

Postcode Level

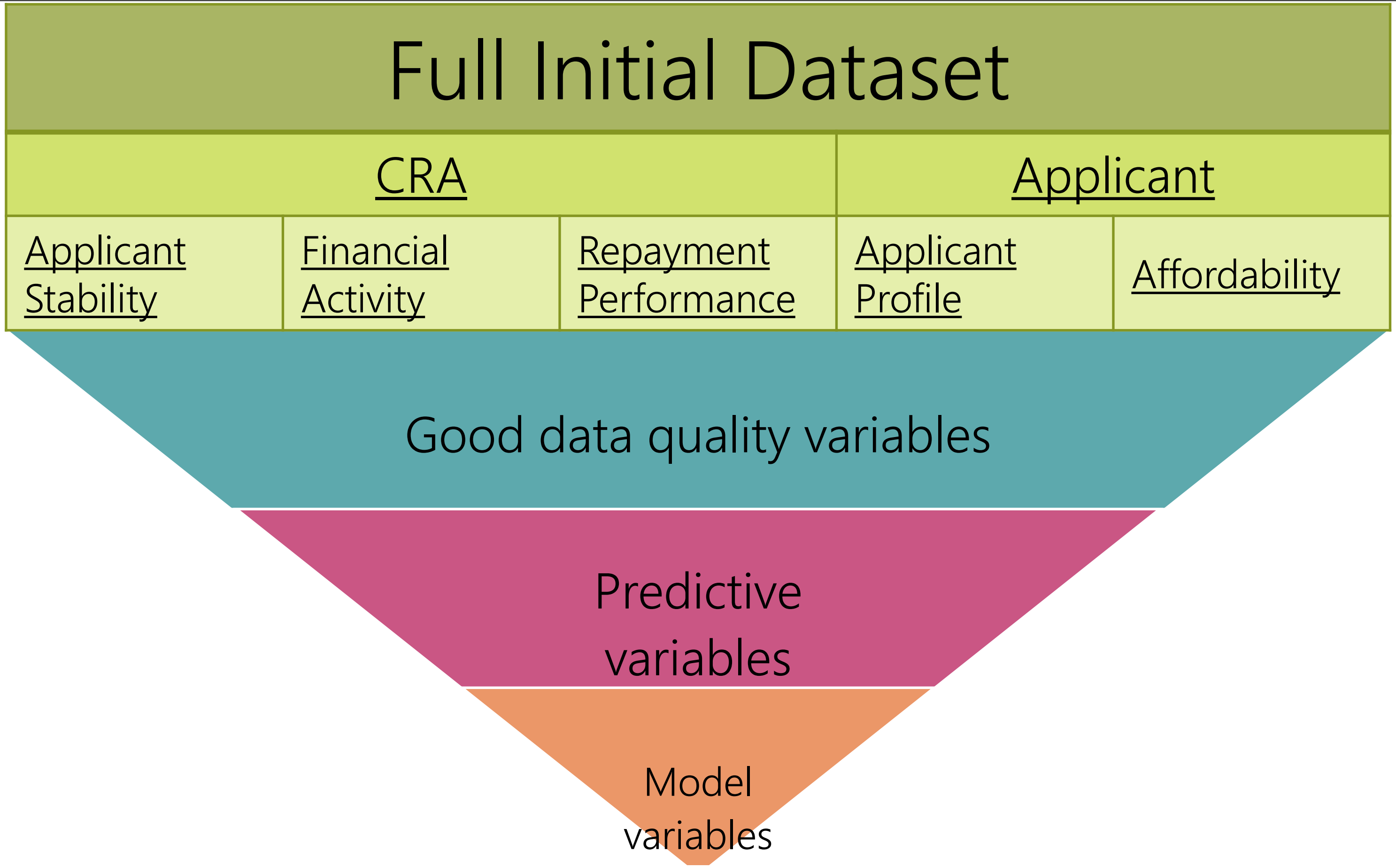
- Geo-demogra-phic profile

Transactional Data

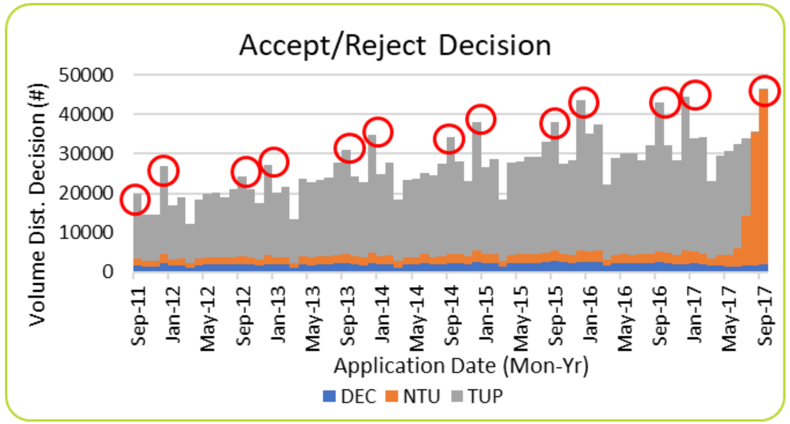
Tasks that Use Extensive CRA Data

- Origination Strategy
- Reject Inference
- Customer Management Strategy
- Regulatory Impacts
- New Products

Data Analysis – Quantity & Types

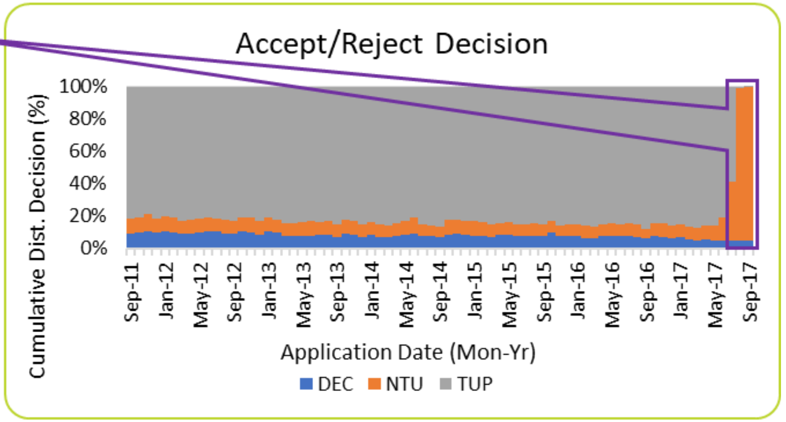


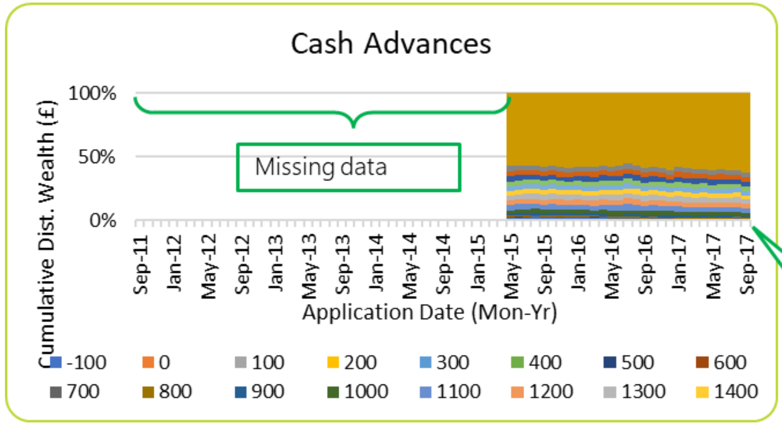
Data Analysis – Cleansing



Take-up can take up to 3 months

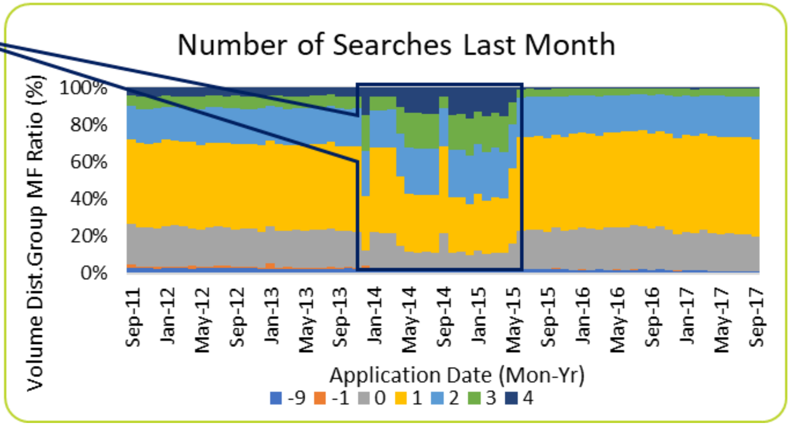
Application spikes in September and December





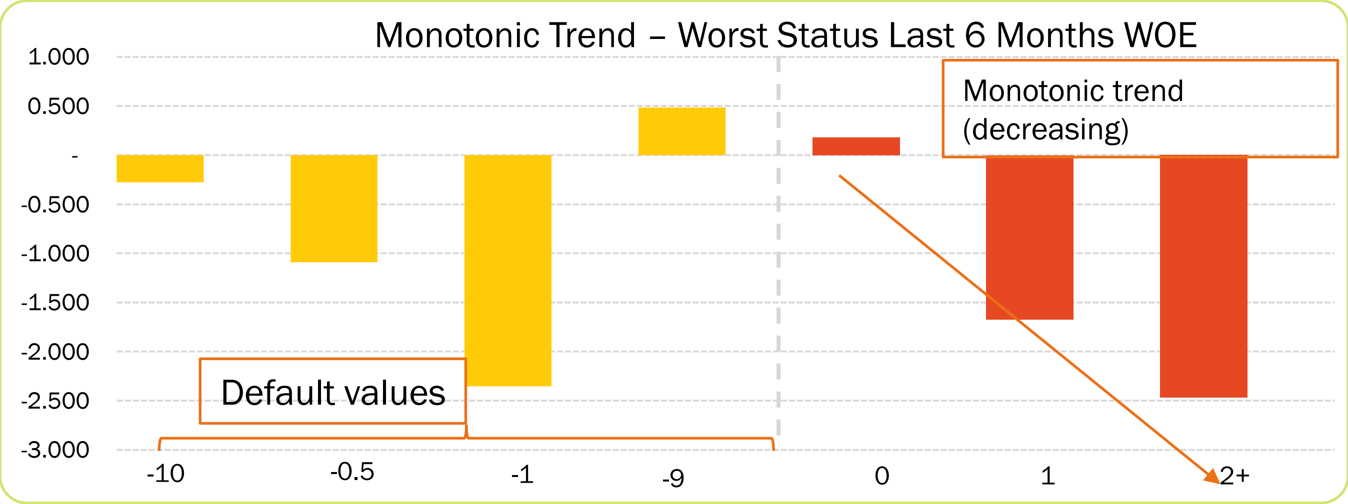
Unusual behaviour from December '13 to May '15

~200 cases per month with default value of -100



Data Analysis – Feature Engineering

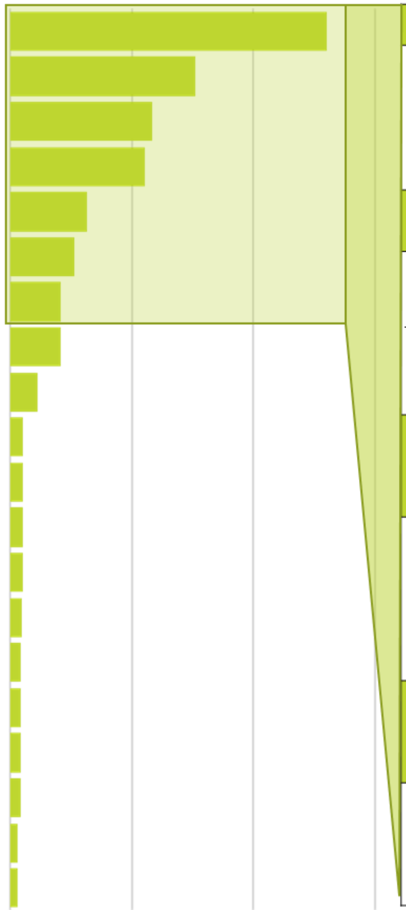
Variable	Overall IV	Data Group	Potential for Modelling
Worst Status Last 6 Months	1.22	CUG	y
Number of Delinquent Accounts	1.22	CUG	y
Value of Delinquent Accounts	1.22	CUG	maybe
Months Since Delinquency	1.19	CUG	y
Value of Unsecured Delinquent Debt	1.18	CUG	no
Number of Unsecured Delinquencies	1.18	CUG	Y
Time Since Most Recent Default	1.05	CUG	Y
Value of Defaults	1.03	CUG	no
Number of Defaults	1.03	CUG	Y
Months Since Mortgage Default	1.00	CUG	y
Value of Mortgage Default	0.99	CUG	maybe
Number of Mortgage Defaults	0.99	CUG	y
Confirmed at Address	0.31	ER	y
Number of Judgements	0.28	Public	y
Tine Since Judgement	0.28	Public	y
Time on ER at Current Address	0.27	ER	y
Number of All Public Judgement Records	0.26	Public	y
Time Since Bankruptcy	0.26	Public	y
Value of Bankruptcy	0.26	Public	y
Applicant Age	0.25	Internal	y
Confirmed at Current Address	0.18	ER	y
Worst Status of Active Accounts Last 12 Months	0.92	CUG	y
Credit Limit Utilisation	0.92	CUG	y
Worst Current Status	0.89	CUG	y
Worst Status Last 3 Motnhs	0.83	CUG	y
Months Since Most Recent Delinquency	0.78	CUG	y



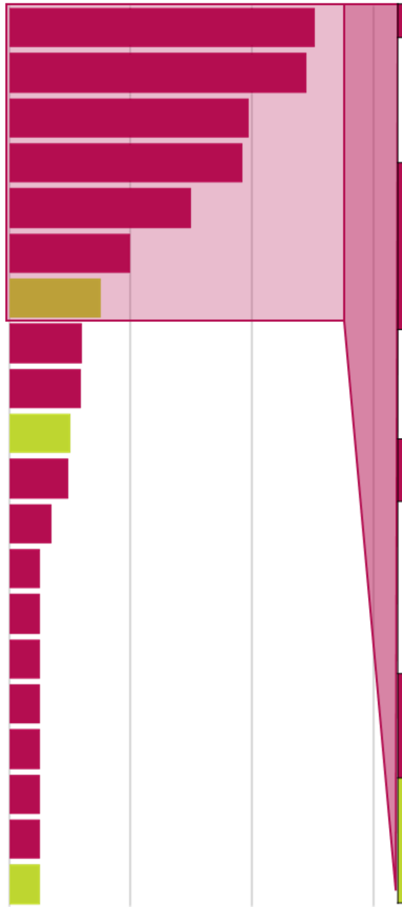
$$WoE = LnOdds(attribute) - LnOdds(population)$$

$$IV = Avg_{Good}(WoE) - Avg_{Bad}(WoE)$$

Traditional Scorecard – Internal & CRA Data



Variable	Values	Score
Loan to Value	Low to 25	24
	26 - 40	10
	41 - 50	5
	51 - 60	2
	61 - 69	-3
	70 - 79	-7
	80 to high	-11
Good Existing Customer	New Customer	0
	Yes	10
	No	-15
Time in Employment (MM)	No Info	-10
	Low to 36	-10
	37 - 66	-7
	67 - 91	-3
	92 - 120	-2
	121 - 143	2
	144 - 184	5
	185 to high	13
Residential Status	No Info	-1
	Public Tenant	-5
	Living with Parents	5
	Private Tenant	10
	Owner	15
Applicant Age (YY)	Low to 22	-6
	23 - 25	-5
	26 - 29	-3
	30 - 33	-2
	34 - 37	0
	38 - 42	2
	43 - 48	5
	49 - high	9
Declared Unsecured Debt to Income	No Info	-2
	Low to 15	4
	16 - 29	-2
	30 - 49	-6
	50 to high	-10
Employment Status	No info	2
	Unemployed	-10
	House Person	-3
	Contractor	6
	Part Time	8
	Full Time	12



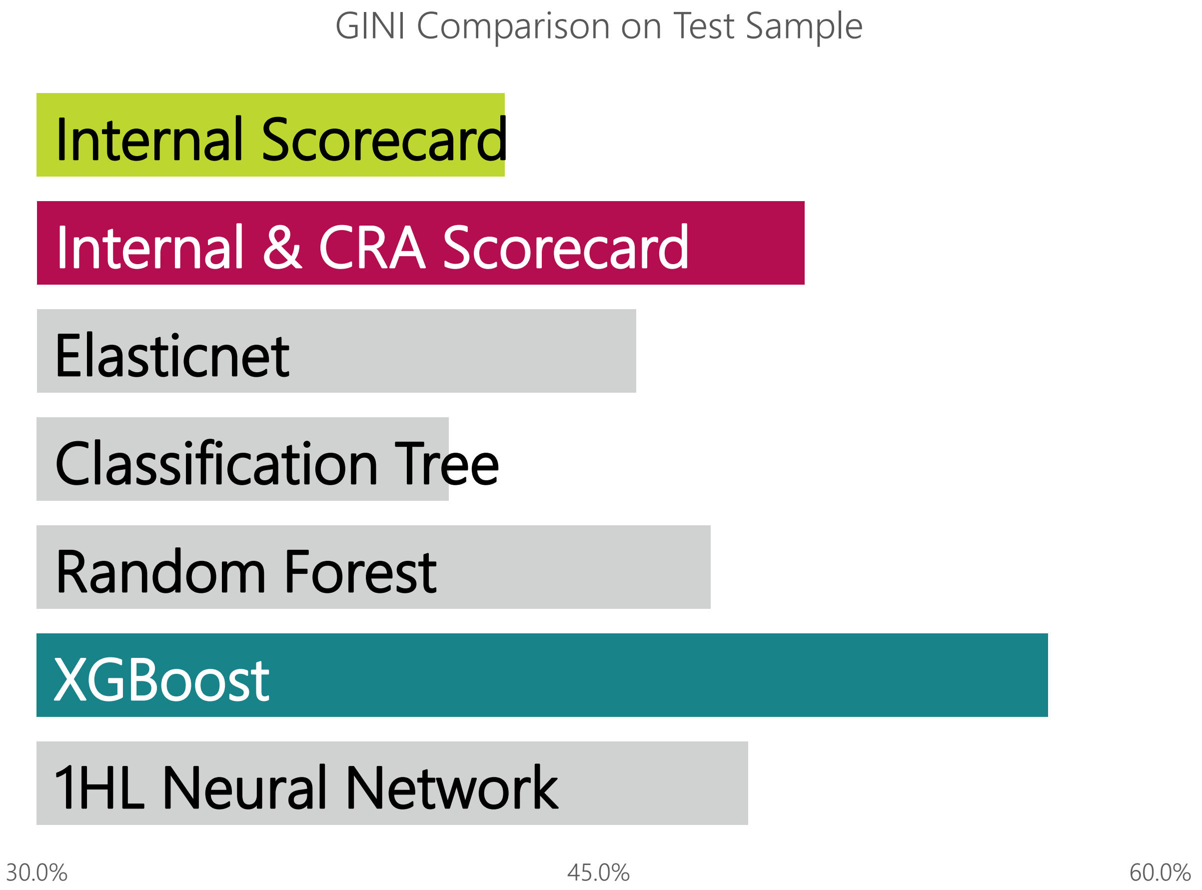
Variable	Values	Score
Worst Payment Status Last 6 Months	No Accounts	-11
	0	6
	1	-11
	2	-21
	3	-28
	4 to high	-30
Credit Limit Utilisation (%)	No Credit Card	-5
	0	-1
	Jan-16	28
	17 - 33	15
	34 - 48	6
	49 - 63	-2
	64 - 81	-13
	82 to high	-24
Time Since Last Delinquency (MM)	No Delinq Accounts	5
	Low to 12	-21
	13 - 36	-17
	37 - 50	-13
	51 to high	-9
Number of Defaulted Accounts	No Default	4
	01-Feb	-13
	3 to high	-16
Age of Oldest Active Account	No Active Account	-11
	Low to 52	-12
	53 - 90	-8
	91 - 113	-5
	114 - 132	-1
	133 - 152	2
	153 - 179	4
	180 to high	5
Time Since Opening Mortgage Account	No Mortgage	-8
	low to 23	19
	24 - 54	14
	55 - 91	10
	92 to high	7
Time Since Missed Payment (Existing Customers)	New Customer	0
	None	15
	Low to 6	-10
	7 – 12	-5
	13 – 36	-1
	37 to high	0

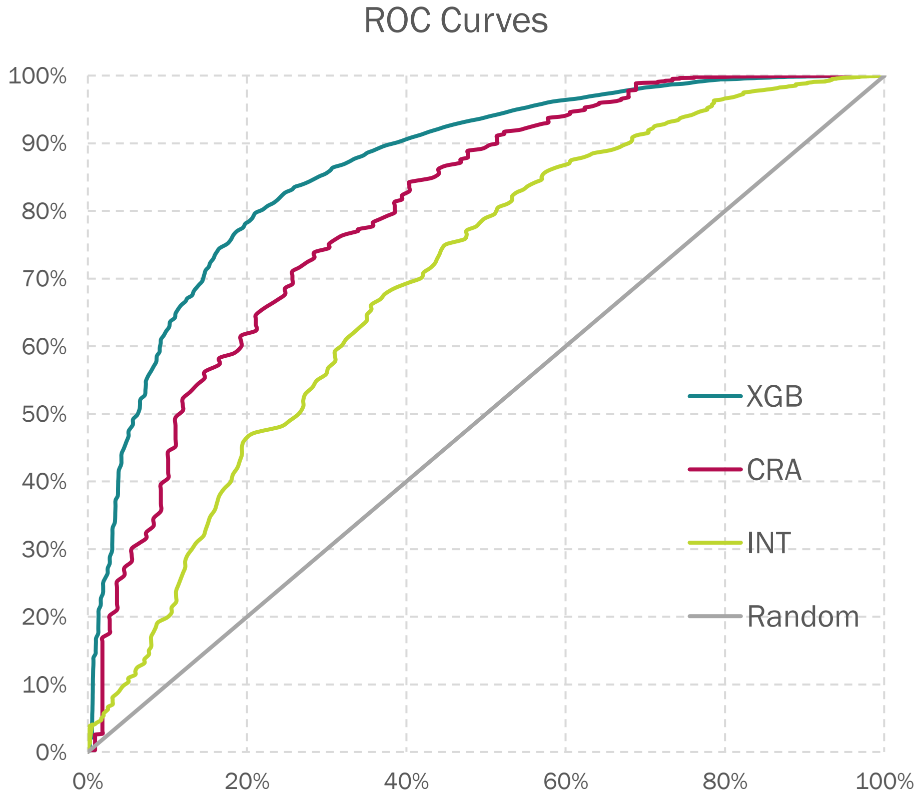
From Regression to Machine Learning*

CAN WE GET MORE OUT OF THE SAME DATA?

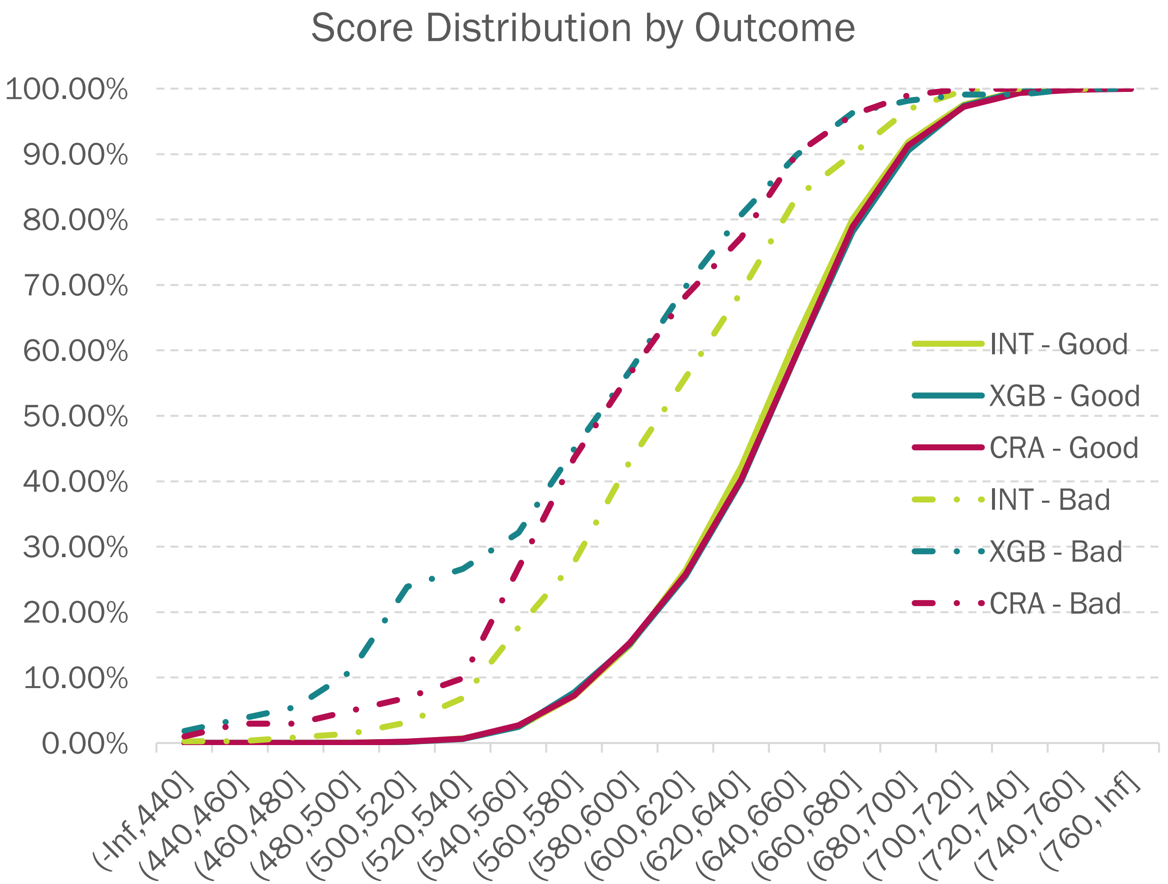
*MATERIAL FROM RUSSO ET AL, 2019. SEE FROM RISK SCORECARDS WITH MACHINE LEARNING (BRIZIORUSSO.GITHUB.IO)

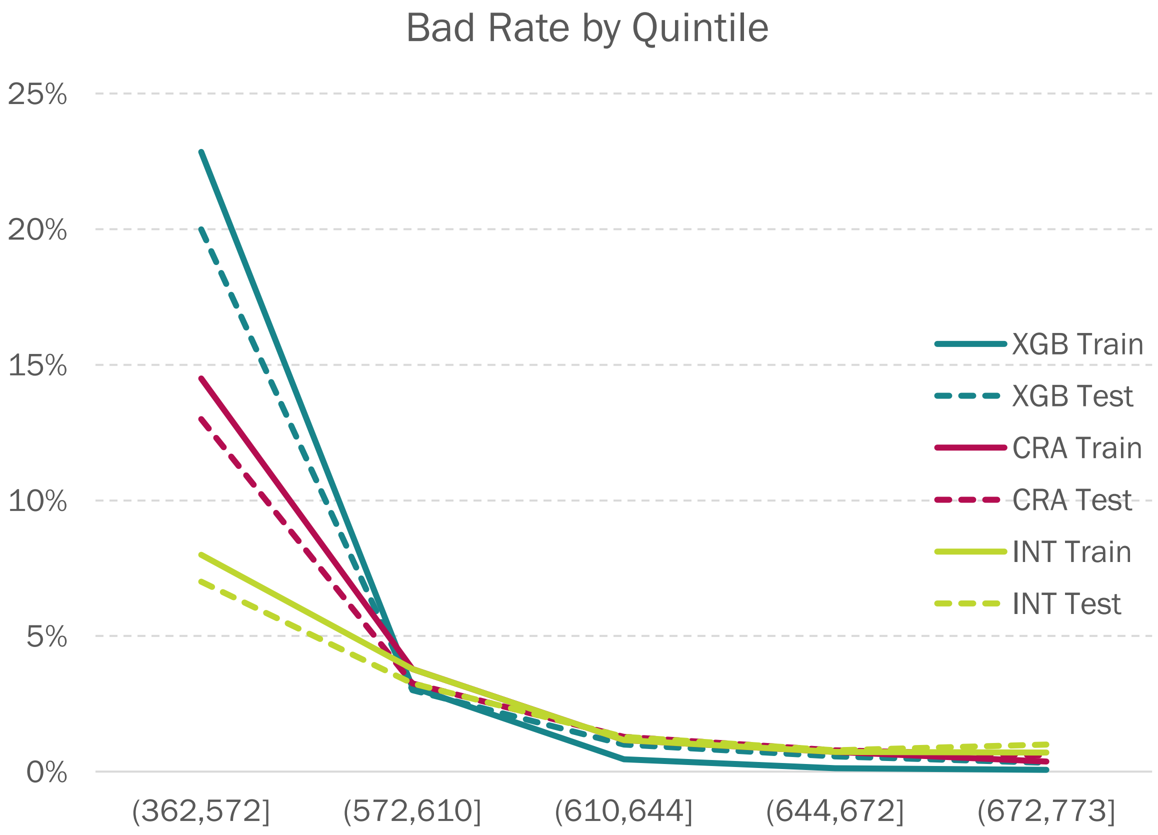
Algorithm Comparison



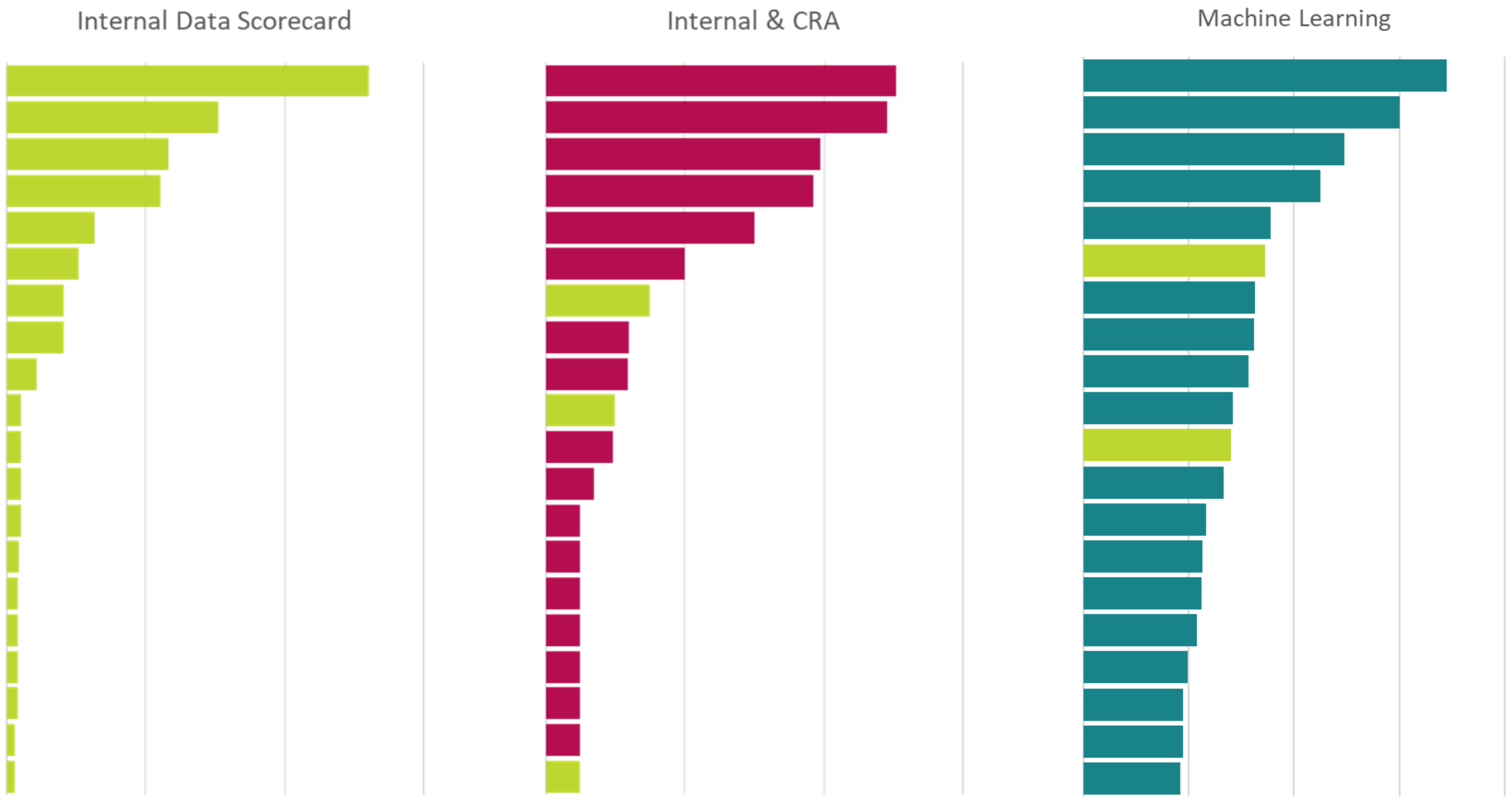


Performance Comparison





Variable Importance Comparison



Regulatory Considerations

Governance

Transparency

Consistent Decisions & Treat Customers Fairly

From Transparent Machine Learning to Causal XAI*

HOW DO WE GO FROM ASSESSING MODELS EX-POST TO MAKING SURE THEY LOOK AT THE RIGHT RELATIONSHIPS?

*MATERIAL FROM RUSSO & TONI, 2022.

Causal Discovery and Injection for Feed-Forward Neural Networks

- In finance many hard problems are tackled with models (e.g. fraud, pricing, credit scoring, trading, planning etc.)
- Practitioners often have a lot of domain (causal) knowledge
- Regulation is quite strict in requiring model stakeholders to understand and "own" their models
- Machine Learning models (e.g. Neural Networks) do not easily allow knowledge integration nor interpretation

Causal Injection into Neural Networks

- Introducing causality into neural networks not only makes them more robust and reliable, but it is also a step towards their interpretability

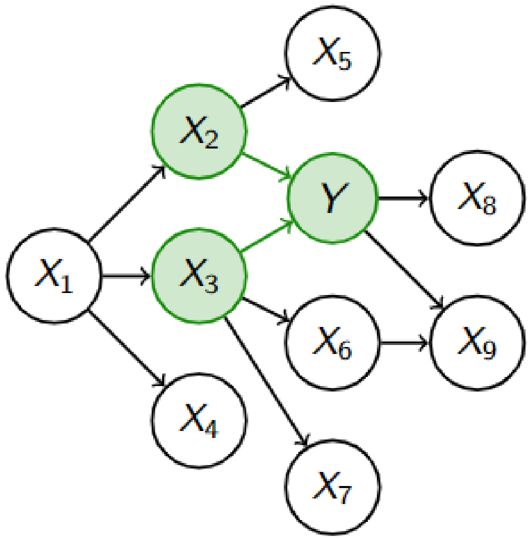
Formal Set-up

Supervised Learning setting

- $\mathbf{X} = [X_1, \dots, X_d] \in \mathcal{X} \subseteq \mathbb{R}^d$ (input features)
- $Y \in \mathcal{Y} \subseteq \mathbb{R}$ (target)
- $\mathcal{P}_{\boldsymbol{X},Y}$ joint distribution of input and target (DGP)
- $\mathcal{D} = \{(\boldsymbol{X}_i, Y_i), i \in \{1, \dots, N\}\}$
  - N i.i.d samples from $\mathcal{P}_{\boldsymbol{X},Y}$
- $f_Y : \mathcal{X} \to \mathcal{Y}$
- Goal: find $\hat{f}_Y$ in $\mathcal{H}$ (hypothesis space)
- $\mathcal{H}$ too complex → Regularize

Causal framework (Pearl, 2009)

- Causal Structure is a DAG G = ⟨V , E ⟩
  - $V = \{Y , X_1, . . . , X_{d+1} \}$ the set of vertices
  - $E \subseteq V \times V$ the set of edges
- $v_i = f_i\ (pa_i , u_i)$
  - v_i is a value for $V_i \in V$ with parents Pa_i having values pa_i
  - f_i any function
  - u_i representing the errors due to omitted factors



(a)

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.005	0.017	0.008	0.002	0.042	0.02	0.005	0.059	0.05
X_1	0.006	0.0	0.063	0.054	0.068	0.009	0.006	0.013	0.006	0.008
X_2	0.088	0.036	0.0	0.022	0.019	0.124	0.008	0.011	0.006	0.008
X_3	0.087	0.034	0.021	0.0	0.024	0.005	0.107	0.104	0.006	0.009
X_4	0.009	0.032	0.02	0.023	0.0	0.01	0.013	0.01	0.005	0.005
X_5	0.026	0.006	0.017	0.004	0.004	0.0	0.012	0.002	0.005	0.018
X_6	0.025	0.006	0.008	0.011	0.005	0.017	0.0	0.014	0.002	0.114
X_7	0.029	0.003	0.007	0.011	0.002	0.024	0.029	0.0	0.011	0.01
X_8	0.036	0.002	0.004	0.003	0.004	0.006	0.009	0.006	0.0	0.006
X_9	0.024	0.003	0.003	0.004	0.003	0.005	0.079	0.01	0.004	0.0

(b) $w_{ik} = \sqrt{\sum_{j=1}^{h} \left(\Theta_1^{i,j,k}\right)^2}$

Synthetic Data Example

(a) Example DAG from Kyono, Zhang and Schaar 2020.
(b) Adjacency Matrix produced by CASTLE (Kyono, Zhang and Schaar 2020) when fitted to the synthetic data produced following the DAG to the left.

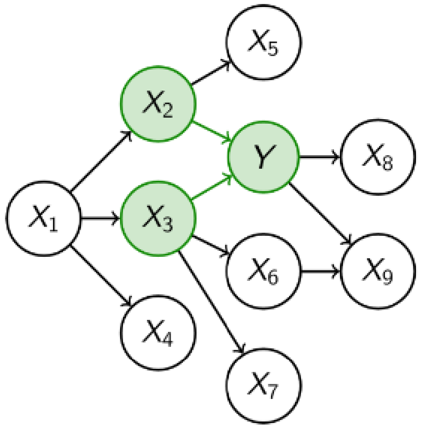
Causal Injection – The Intuition

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.005	0.017	0.008	0.002	0.042	0.02	0.005	0.059	0.05
X_1	0.006	0.0	0.063	0.054	0.068	0.009	0.006	0.013	0.006	0.008
X_2	0.088	0.036	0.0	0.022	0.019	0.124	0.008	0.011	0.006	0.008
X_3	0.087	0.034	0.021	0.0	0.024	0.005	0.107	0.104	0.006	0.009
X_4	0.009	0.032	0.02	0.023	0.0	0.01	0.013	0.01	0.005	0.005
X_5	0.026	0.006	0.017	0.004	0.004	0.0	0.012	0.002	0.005	0.018
X_6	0.025	0.006	0.008	0.011	0.005	0.017	0.0	0.014	0.002	0.114
X_7	0.029	0.003	0.007	0.011	0.002	0.024	0.029	0.0	0.011	0.01
X_8	0.036	0.002	0.004	0.003	0.004	0.006	0.009	0.006	0.0	0.006
X_9	0.024	0.003	0.003	0.004	0.003	0.005	0.079	0.01	0.004	0.0

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.059	0.05
X_1	0.0	0.0	0.063	0.054	0.068	0.0	0.0	0.0	0.0	0.0
X_2	0.088	0.0	0.0	0.0	0.0	0.124	0.0	0.0	0.0	0.0
X_3	0.087	0.0	0.0	0.0	0.0	0.0	0.107	0.104	0.0	0.0
X_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.114
X_7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Objective:

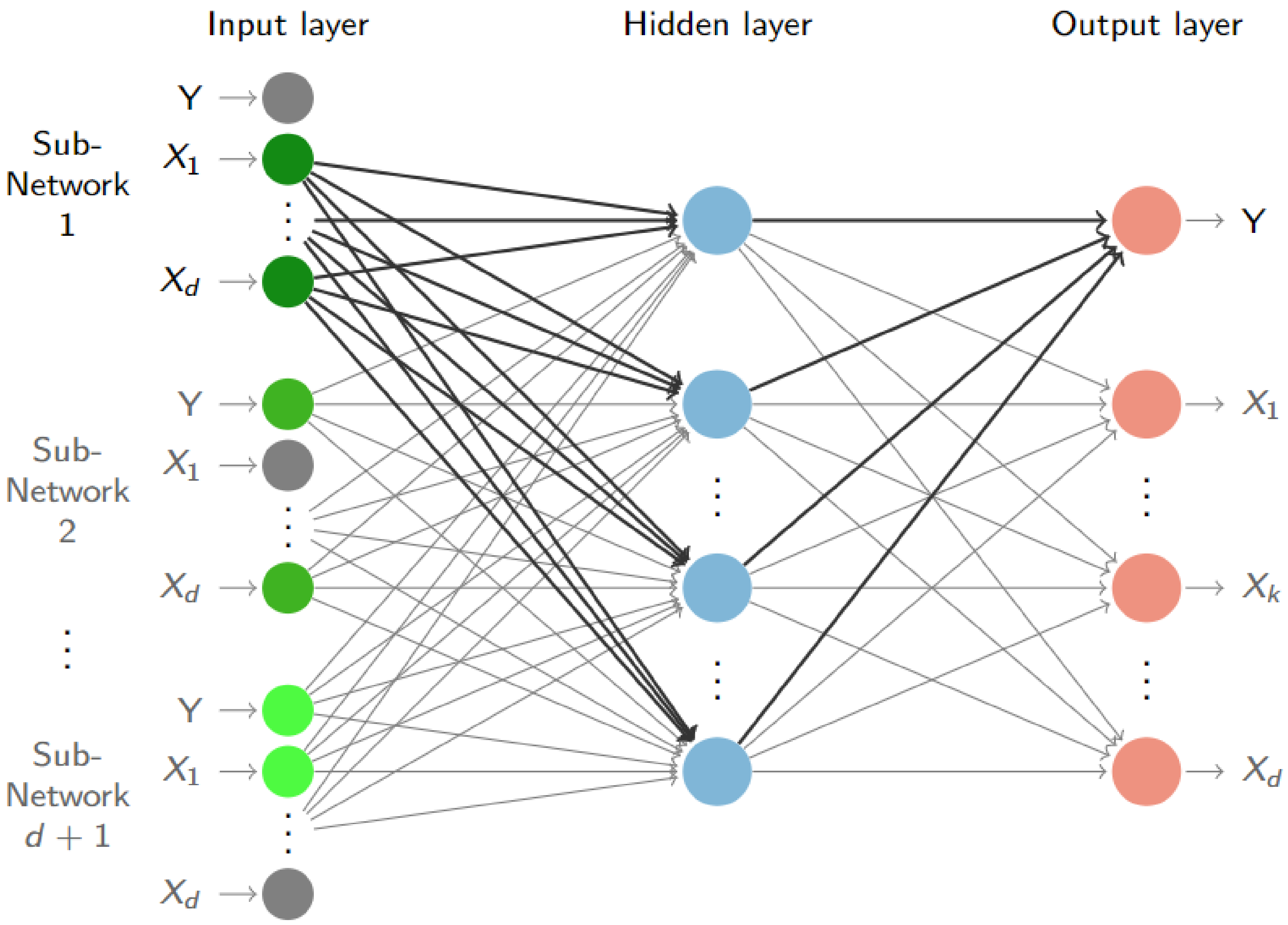
have the network use only the relationships contained in the DAG i.e. predict each of the features using only its parents.



	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.059	0.05
X_1	0.0	0.0	0.063	0.054	0.068	0.0	0.0	0.0	0.0	0.0
X_2	0.088	0.0	0.0	0.0	0.0	0.124	0.0	0.0	0.0	0.0
X_3	0.087	0.0	0.0	0.0	0.0	0.0	0.107	0.104	0.0	0.0
X_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.114
X_7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

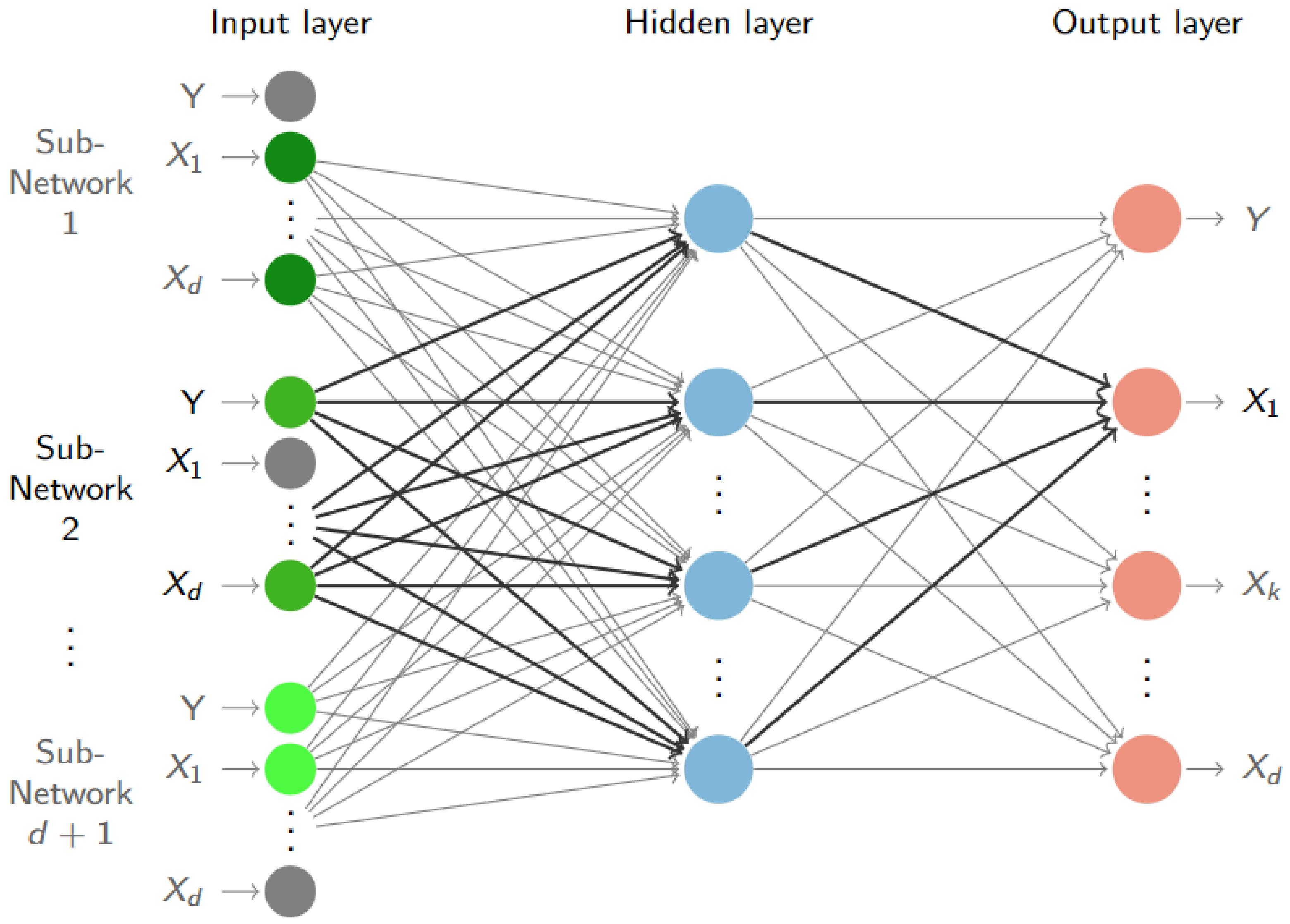
Joint Network

Predict the target while reconstructing all other input features



Joint Network

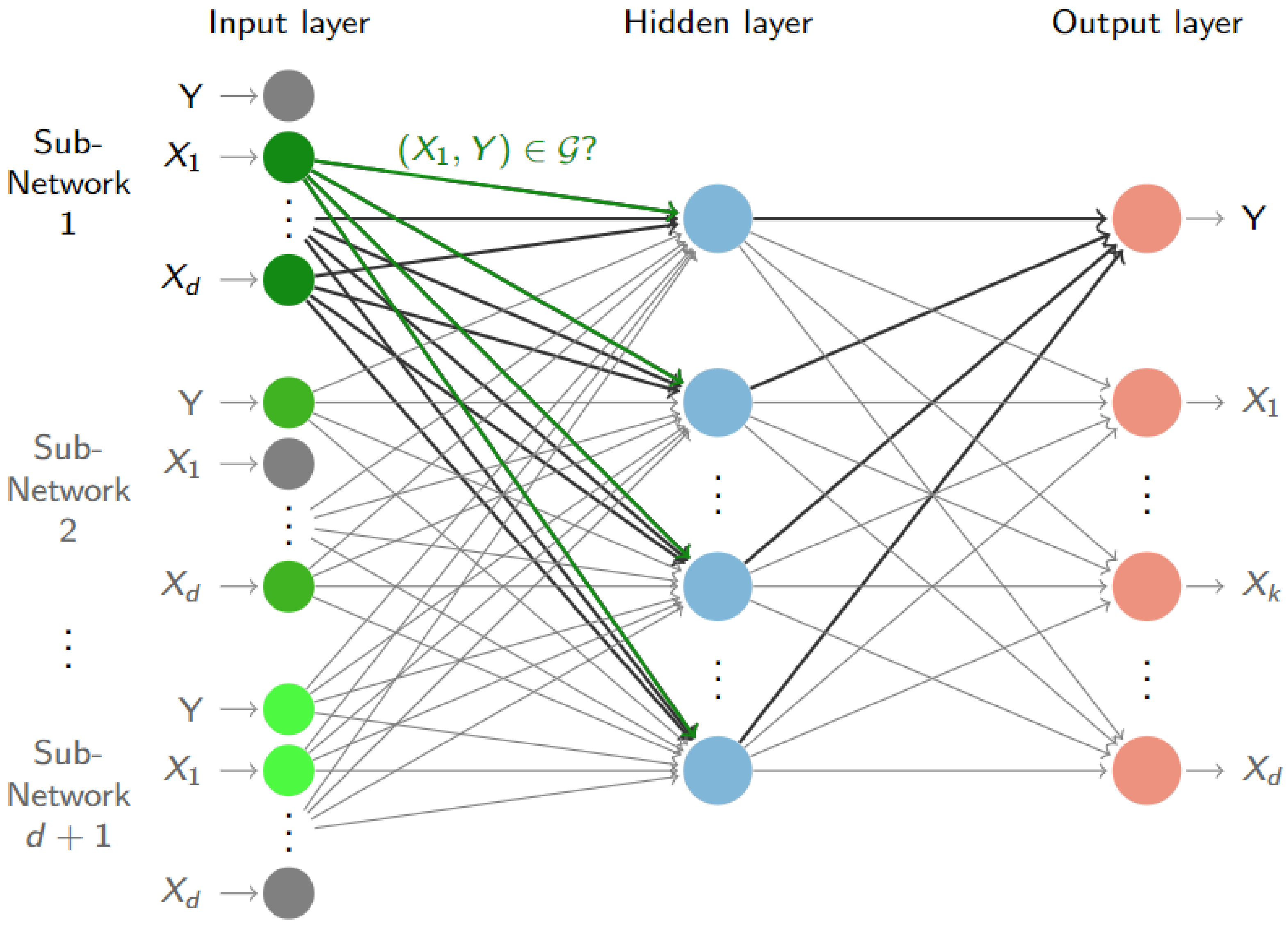
Predict the target while reconstructing all other input features



Joint Network

Predict the target while reconstructing all other input features

Is this input-output relationship contemplated in my causal DAG?

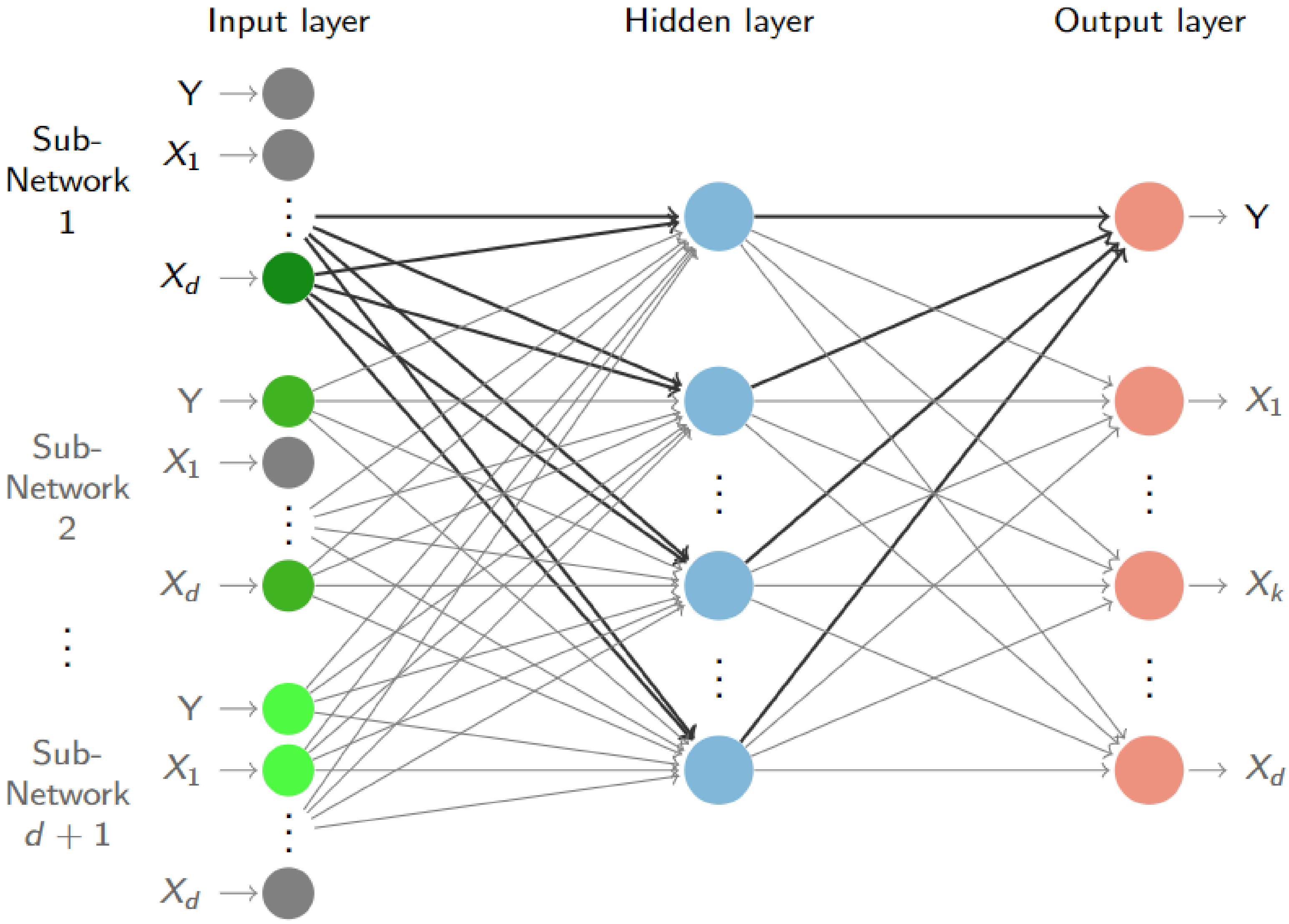


Joint Network

Predict the target while reconstructing all other input features

NO?

➢“Semantic” Regularization



Limitations of Proposed Algorithm

- It requires a complete DAG (covering all variables considered in the problem and the data)
- Full causal DAG is rare and often impractical to build
- We propose a second algorithm that involves Subject Matter Experts (SMEs) providing their input

Algorithm 2 – Human-AI Collaboration

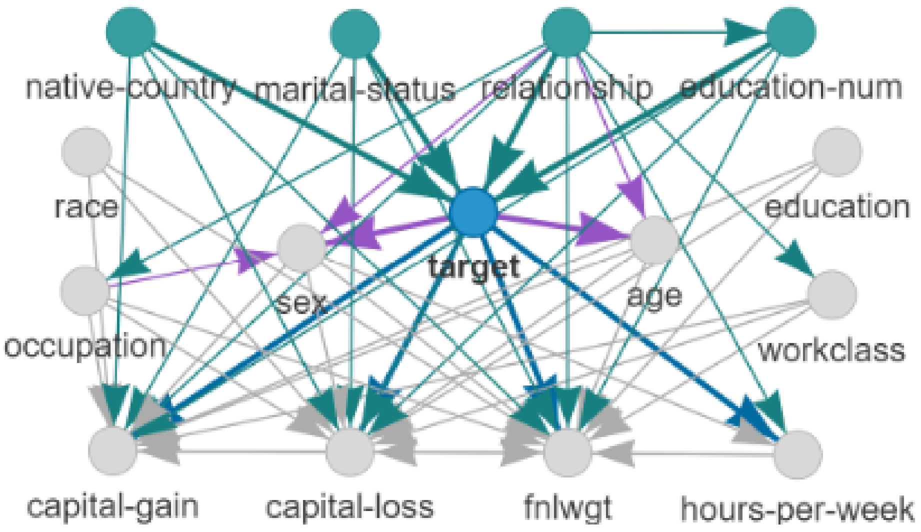


Figure 3: Example of computed DAG for Adult dataset (see Section 5.3.3). Cyan nodes at the top are computed causes for the target ("Income>50K"), edges coming out of the target are in blue while in purple are the edges into nodes that cannot be caused (as per basic assumptions in Section 5.3.2).

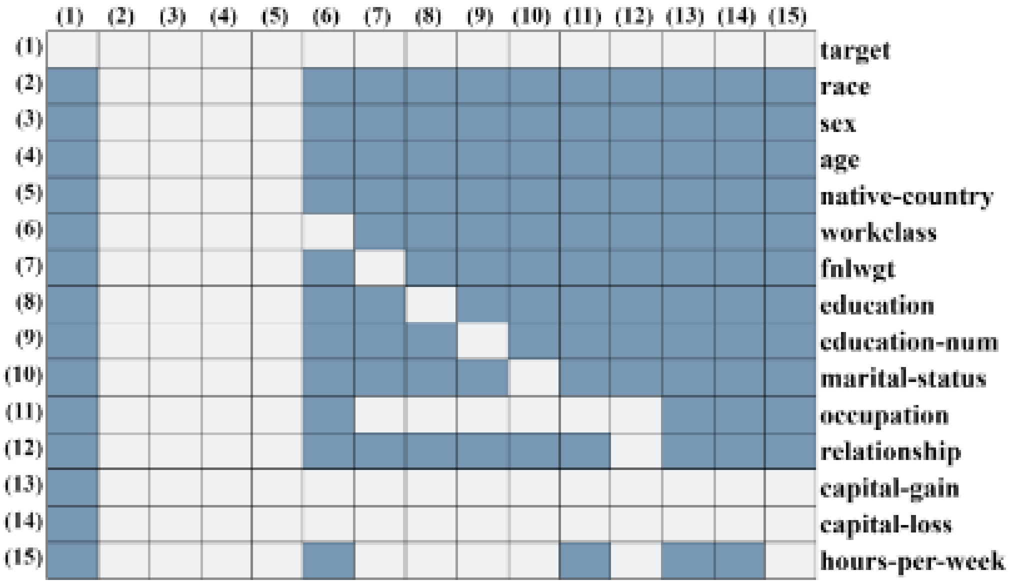


Figure 2: Input graph $\mathcal{G}_p$, as partial causal knowledge for the Adult dataset, in the form of an adjacency matrix W. Blue represents edges; missing edges in white (hard constraints).

HCI Causal Injection – Results

Table 1: Experiments with real data in the financial/economics sector. We report MSE (AUC) for regression (classification) across different sample sizes of the training data (best results in bold). We also detail, for each dataset, the number of features/nodes $|V|$ and the number of edges $|E|$ in the injected DAG (for our method) and in the (graph drawn from the) underlying adjacency matrix (for CASTLE). NA indicates a data size (N) bigger than the full dataset. CASTLE and *Injected* columns refer to Section 5.3.1, for *Partial* and *Refined* columns see Sections 5.3.2 and 5.3.3, respectively.

	REGRESSION (Metric: MSE)				CLASSIFICATION (Metric: AUC)					
	California ($\|V\| = 8$)		Boston ($\|V\| = 14$)		HELOC ($\|V\| = 23$)		Adult ($\|V\| = 14$)			
Data size (N)	CASTLE $\|E\| = 72$	*Injected* $\|E\| = 31$	CASTLE $\|E\| = 182$	*Injected* $\|E\| = 48$	CASTLE $\|E\| = 552$	*Injected* $\|E\| = 85$	CASTLE $\|E\| = 210$	*Injected* $\|E\| = 46$	*Partial* $\|E\| = 116$	*Refined* $\|E\| = 30$
100	7.05 (12.81)	**2.94 (2.63)**	112.04 (91.06)	**86.17 (13.75)**	**0.75 (0.02)**	0.74 (0.04)	0.67 (0.03)	**0.69 (0.04)**	0.66 (0.02)	**0.69 (0.04)**
500	2.33 (1.39)	**2.25 (1.07)**	21.95 (6.84)	**20.45 (5.12)**	**0.79 (0.01)**	0.78 (0.01)	0.72 (0.04)	**0.74 (0.02)**	0.71 (0.02)	**0.74 (0.02)**
1000	2.96 (4.12)	**1.68 (1.14)**	NA	NA	0.78 (0.01)	0.78 (0.01)	0.75 (0.03)	**0.76 (0.03)**	0.74 (0.03)	**0.76 (0.02)**
2000	3.86 (3.68)	**1.71 (0.57)**	NA	NA	**0.79 (0.01)**	0.78 (0.01)	0.74 (0.03)	**0.77 (0.01)**	**0.76 (0.03)**	**0.77 (0.02)**
5000	4.91 (7.41)	**1.51 (0.62)**	NA	NA	0.79 (0.01)	0.79 (0.01)	0.75 (0.03)	**0.79 (0.03)**	**0.76 (0.02)**	**0.79 (0.03)**
10000	1.74 (1.70)	**1.16 (0.31)**	NA	NA	**0.80 (0.01)**	0.79 (0.01)	0.75 (0.02)	**0.85 (0.01)**	**0.76 (0.02)**	**0.85 (0.01)**
20000	**0.66 (0.08)**	1.02 (0.35)	NA	NA	NA	NA	0.76 (0.02)	**0.86 (0.01)**	**0.77 (0.02)**	**0.86 (0.01)**

Conclusion

- CRA Data is what enables (more) accurate credit worthiness assessment in UK
- Logistic Regression is to this day the most used technique for its interpretability
- Other ML algorithms can achieve similar levels of transparency
- Statistical relationship is not the same as Causal Relationship
- High-stakes decision models should look at both statistical and causal relationships

Questions?

GET IN TOUCH

fabrizio@imperial.ac.uk

briziorusso.github.io

References

T. Kyono, Y. Zhang, and M. van der Schaar. 2020. CASTLE: Regularization via Auxiliary Causal Graph Discovery. In Proc. NeurIPS

J. Pearl. 2009. Causality (2 ed.). Cambridge University Press.

F. Russo. 2019. Credit Risk Modelling: Data and Techniques Used in the UK Banking Industry. THE USE OF CREDIT REGISTER DATA FOR FINANCIAL STABILITY PURPOSES AND CREDIT RISK ANALYSIS, Danmarks Nationalbank Conference.

F. Russo, T. Ringsjø, D. Smith, J. Woodcock, T. Pile, L. Koteva. 2019. Risk Scorecards with Machine Learning. Modelling with big data and machine learning: interpretability and model uncertainty, Joint Conference by the Bank of England and the Data Analytics for Finance and Macro Research Centre at King's College London,

F. Russo and F. Toni. 2022. Causal Discovery and Injection for Feed-Forward Neural Networks. arXiv:2205.09787. arXiv. http://arxiv.org/abs/2205.09787 arXiv:2205.09787

Acknowledgments

Credit risk material was adapted from presentations given while the author was working at 4most Europe, thanks to the team that helped with its preparation.