Model Multiplicity and Predictive Arbitrariness in Recidivism Risk Assessment

[Paper ID: 1303]

Ashwin S
TU Wien, Austria

Carlos Castillo
ICREA and Universitat Pompeu Fabra, Spain

Preprint

Definitions

Rashomon Set
Set of models with near-optimal loss, up to an additive ε.

For Hypothesis Class H, Dataset S, and Loss L,

$$R(H, \varepsilon, S) = \{h \in H : L_S(h) \leq L_S(h^*) + \varepsilon\}$$

where $h^* = \arg\min_{h \in H} L_S(h)$ and $\varepsilon \geq 0$.

Self-Consistency
How often do two models output the same prediction?

For a finite model set $P = \{h_1, \ldots, h_K\}$ and $x \in S$

$$SC_P(x) = \binom{K}{2}^{-1} \sum_{i<j} \mathbf{1}[h_i(x) = h_j(x)] \,.$$

DATA LABELING | MODEL TRAINING | MODEL DEPLOYMENT | RISK PREDICTION | MULTIPLICITY RESOLUTION

RisCanvi

Decision Support System, Operational in Catalan Prisons since 2009.

Manual Labeling
Experts apply legal rules to post-release records.

Logistic Regression
Learn exactly one model from manually labelled data.

One Model
No fairness intervention, or monotonicity constraints.

One Risk Estimate
Risk scores are split into low, medium and high.

No Mechanism

Proposed Pipeline

Automated Labeling — C1

Rule-Based Algorithm based on the Spanish Penal Code.

Iteratively Designed and Validated with Experts.

17x Larger Dataset

Linear Models — C2

Learn a linear integer model h*.

Initialize MILP with h* to build $P \subseteq R$ of Less Discriminatory Models (**LDMs**).

Generalize Search for LDMs to m>2 Groups

Finite Model Set $P \subseteq R$

Comparable **Accuracy** and **Error-Rate Disparity**.

Monotonicity: Rehabilitative progress lowers risk scores.

Model Multiplicity
Different Coefficients

Multiple Risk Scores

P { h_1 — Low ; h_K — High }

Predictive Arbitrariness
Different Predictions

Lowest-Risk Policy — C4

Final Risk Assessment
Low

Assign the lowest risk score output by any model $h \in P$.

Tight Lower Bound on Expected Self-Consistency — C3

Despite **Model Multiplicity**, **Predictive Agreement** can far exceed Worst-Case Guarantees.

Perfect Predictive Agreement on ~**70%** Test Instances.

	Observed	Lower Bound
$\mathbb{E}_{x \sim S}[SC_P(x)]$	0.91	0.53

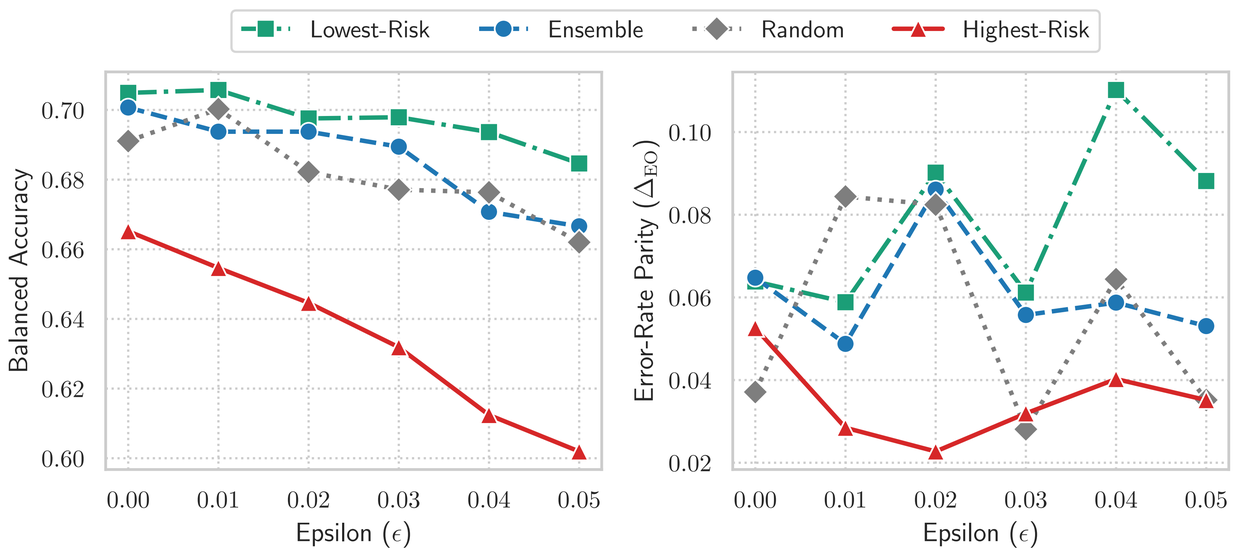
Predictive Performance and Algorithmic Fairness

Train Years: 2010-2014 [**complete**] + 2015-2018 [**recidivists only**]

Test Year: 2019 Target: Five-Year Penal Recidivism

	RisCanvi	Sparse Linear Integer Model [**SLIM**]	FairSLIM + Lowest-Risk Policy
Balanced Accuracy	55.3	69.4	70.6
False Positive Rate	5.2	44.4	27.8
False Negative Rate	84.3	17.9	31.1
Error Rate Parity	15.0	9.1	5.9

Policies for Resolving Predictive Multiplicity



Ask Me About

- Lowest Risk Policy for Punitive Interventions
- Mechanism Design for Addressing Predictive Arbitrariness
- Procedural, Existential and Worst-Case Predictive Multiplicity