

Data	Model	N	In-Sample Accuracy	Out-of-Sample Accuracy	Training Time (s)
Original	Sparse Logistic Reg.	500	81.2%	78.0%	3.9
		1000	82.1%	81.0%	15.9
		5000	78.1%	77.9%	22.9
		10000	%	%	
	Decision Tree	500	91.6%	70.2%	95.9
		1000	91.3%	76.5%	497.9
		5000	91.4%	75.7%	3255.3
		10000	%	%	
	Transformer	500	100%	86.4%	91.2
		1000	90.7%	76.0%	246.6
		5000	96.6%	86.4%	1167.6
		10000	87.6%	87.7%	1964.4
	Concept Bottleneck	500	98.4%	84.8%	110.9
		1000	99.2%	73.2%	211.6
		5000	90.4%	88.2%	892.8
		10000	92.5%	89.0%	1854.0
Featurized	Sparse Logistic Reg.	500	92.6%	87.8%	0.4
		1000	88.8%	85.4%	8.1
		5000	86.5%	85.3%	32.7
		10000	%	%	0.4
	Decision Tree	500	74.4%	69.6%	1.8
		1000	87.9%	70.6%	14.7
		5000	90.2%	75.1%	185.5
		10000	%	%	

Table 1: In and out-of-sample accuracies of models applied to the simulation study. For cross-validated models, training time was averaged across folds. Training time was computed on a 2022 MacBook Pro with 16GB RAM and an M2 GPU. Models differ in their inherent interpretability, manual feature curation effort, and generalization performance. Deep learning models can overfit the training sample while still generalizing well. Intrinsically interpretable models can be substantially improved through effective featurization.