Missingness-Avoiding Machine Learning

Prediction Models That Learn to Avoid Missing Values

Lena **Stempfle**, Anton **Matsson**, Newton **Mwai**, and Fredrik D. **Johansson**

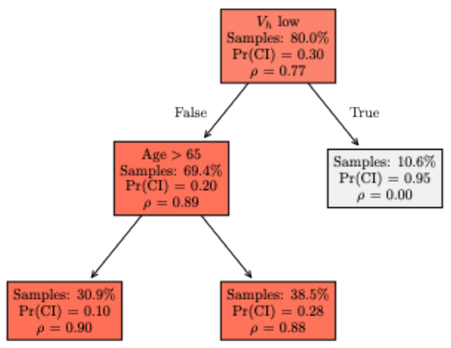
Chalmers University of Technology and University of Gothenburg

Overview

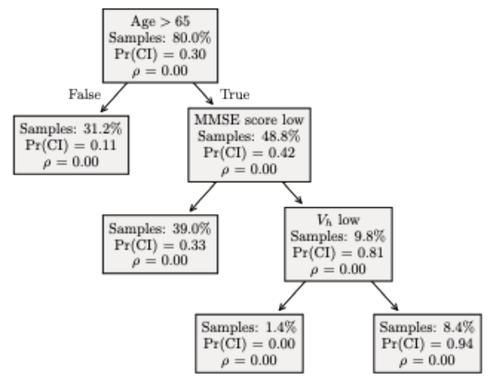
- **Challenge:** Common approaches to handling test-time missing values, such as imputation or adding missingness indicators, often reduce interpretability or introduce bias.
- **Insight:** If features are not necessary for predictions, models should not ask for them.
- **Approach:** We propose missingness-avoiding (MA) machine learning, a general framework for training models to minimize dependence on missing or imputed features.
- **Models:** We develop customized MA algorithms for decision trees, tree ensembles, and sparse linear models by regularizing their objectives to reduce reliance on missing values.
- **Results:** MA models match the predictive accuracy of baselines while significantly reducing reliance on missing data, enabling more interpretable and robust predictions.

A Motivating Example

- Missing values are rarely missing completely at random—they are missing for a reason!
- Example (cognitive impairment (CI) classification): cognitive test scores are available only for older patients, and MRI scans primarily for those with low cognitive test scores.
- Tree-based MA models exploit these patterns in the data-generating process to contextually avoid missing values.



DT | AUC = 0.74 | $\rho = 0.77$



MA-DT | AUC = 0.73 | $\rho = 0.00$

The MA Learning Framework

- **Setup:** Supervised learning with input $X = [X_1, \dots, X_d]^\intercal$ and label $Y \in \mathcal{Y}$; missing values in X are indicated by a mask $M \in \{0,1\}^d$.
- **Missingness reliance:** A model $h \in \mathcal{H}$ has missingness reliance $\rho(h, x) = 1$ for input x if computing $h(x)$ requires any missing feature in x; otherwise, $\rho(h, x) = 0$.
- **Goal:** Learn a model that balances predictive loss and missingness reliance under the (fixed) distribution $p(X, M, Y)$:

$$\min_{h \in \mathcal{H}} \mathbb{E}_p\big[L\big(Y, h(X)\big)\big] + \alpha \cdot \mathbb{E}_p[\rho(h, X)].$$

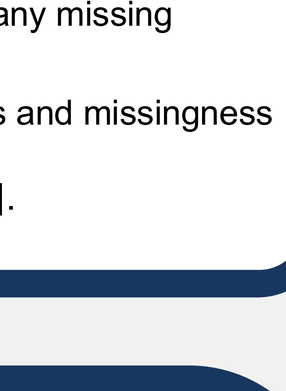
MA Models

MA trees and MA tree ensembles

- A tree h has missingness reliance for input x if there exists a node u along the decision path $\pi_h(x)$ where the split is based on a feature j_u that is missing in x: $\rho(h, x) \coloneqq \max_{u \in \pi_h(x)} \mathbf{1}\big[x_{j_u} = \mathbf{na}\big]$.
- For a tree ensemble e, we define $\rho(e, x) \coloneqq \max_{h \in e} \rho(h, x)$.
- We regularize the node splitting criteria C, splitting a leaf node ℓ assigned training indices S_ℓ by selecting the feature-threshold pair j, τ that solves $\min_{j,\tau} C(\ell, \mathcal{D}; j, \tau) + \alpha \cdot \sum_{i:S_\ell} \frac{\sigma_{ij}}{|S_\ell|} \mathbf{1}\big[x_{ij} = \mathbf{na}\big]$.
- The feature weight $\sigma_{ij} \in \{0,1\}$ prevents double penalization in gradient boosted trees when a feature is reused in later trees.

	A	B	C	D
x	1	0	**na**	0

	A	B	C	D
θ	5	1	**0**	3



MA sparse linear models

- For a linear model $h(x) = \theta^\intercal x$, we define

$$\rho(h, x) \coloneqq \max_j \mathbf{1}\big[|\theta|_j > 0\big]\, \mathbf{1}\big[x_j = \mathbf{na}\big].$$

- We solve $\min_\theta \frac{1}{n} \sum_{i=1}^n L(y_i, \theta^\intercal x_i) + \sum_{j=1}^d \big(\lambda + \alpha \bar{m}_j\big)|\theta_j|$.

Empirical Results

- We compare MA models to standard baselines on tabular datasets with varying sizes and levels of missingness (4 of 6 included here).
- MA models achieve comparable predictive performance while significantly reducing reliance on missing values.
- The trade-off parameter α is selected via cross-validation.

	ADNI[1]		FICO[2]		LIFE[3]		NHANES[4]	
Model	AUC	ρ	AUC	ρ	AUC	ρ	AUC	ρ
LR	72.0	64.1	79.1	75.4	98.7	63.4	85.1	100.0
MA-LR	68.4	11.8	75.8	5.7	97.7	21.0	84.5	0.4
DT	72.7	11.7	73.8	5.7	92.2	18.3	82.3	0.1
MA-DT	74.1	0.1	73.8	5.7	89.7	12.3	82.4	0.0
RF	75.0	66.1	77.3	65.4	97.8	82.8	81.6	100.0
MA-RF	78.0	3.2	76.3	5.8	97.2	24.2	84.0	0.2
GBT	79.0	60.9	78.4	69.3	98.2	56.1	84.8	95.7
MA-GBT	78.5	1.6	75.6	5.9	95.9	16.5	83.2	0.3

Conclusion

- MA learning enables models that rely less on (imputed) missing values without sacrificing predictive performance.
- Alternative definitions of $\rho(h, x)$ are possible; for example, the fraction of missing features used when computing $h(x)$.
- The trade-off parameter α can be selected based on application-specific needs.
- Open question: Can the MA framework be extended to other model classes, such as neural networks?

CHALMERS

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

FOOTNOTES.
[1]Alzheimer's Disease Neuroimaging Initiative: diagnosis change prediction (1,337 samples | 39 features).
[2]FICO Explainable ML Challenge: credit repayment prediction (10,549 | 23).
[3]World Health Organization: life expectancy classification (above/below dataset median) (2,864 | 18).
[4]National Health and Nutrition Examination Survey: hypertension prediction (10,000 | 42).