Hybrid Tree-based Interpretable Pricing

Panyi Dong

University of Illinois Urbana-Champaign

2025-08-01

Synopsis

- **Background & Motivation**
- **Methodology**
- **Empirical Experiments**
- **Conclusion**

Background

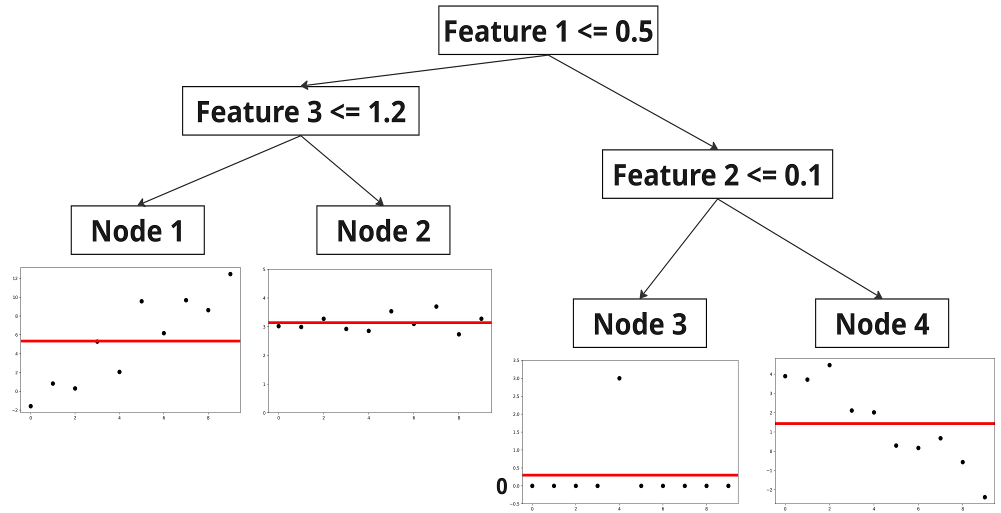

Fig. 1: Classification and Regression Tree (CART)

- CART (Breiman et al., 1984)
  - Intuitive data splits
    - Easy for **interpretation**
  - Address data heterogeneity
    - **Homogeneous** leaf nodes
    - **Mean** as predictions
- However
  - Insurance claims are
    - **Compound** frequency-severity
    - Classification + Regression

Background

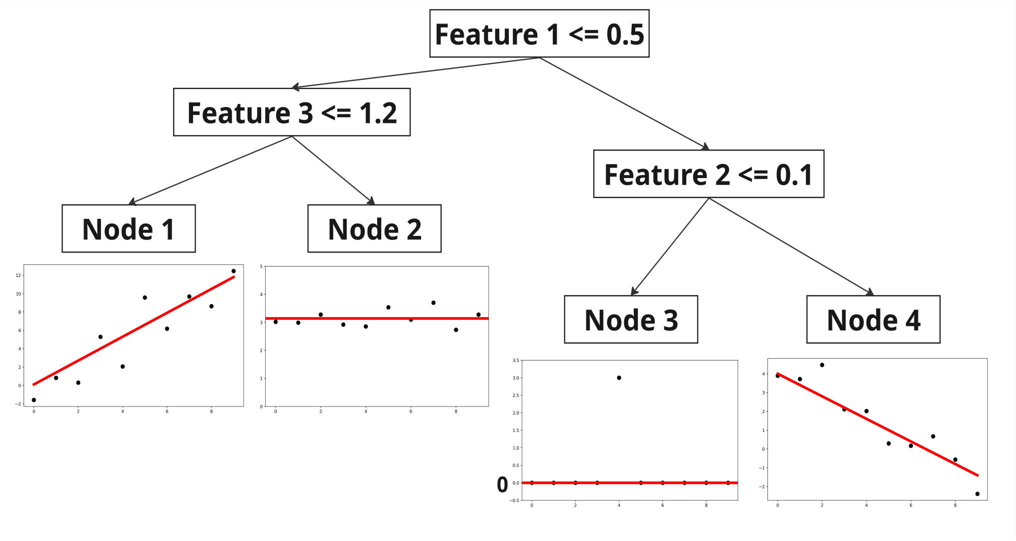

Fig. 2: HybridTree (HT)

- HybridTree (Quan et al., 2023)
  - Compound tree structure to capture insurance claims distribution
  - Classification tree: Frequency
    - Identification of risk
  - Regression leaf nodes: Severity
    - Quantification of reported claims
    - Zeroes for **excess Zeros**
    - Mean for **not data-sufficient** nodes
    - **Linear regression** for homogeneous nodes

Motivation

- Modification of HT
  - Previous HT
    - **Fixed** classification tree
    - Limited **growing/pruning** measures
  - **Solutions**
    - New implementation of HT from scratch
    - Introduces classification- and regression-based measures
    - Risk loading as post hoc modification

Motivation

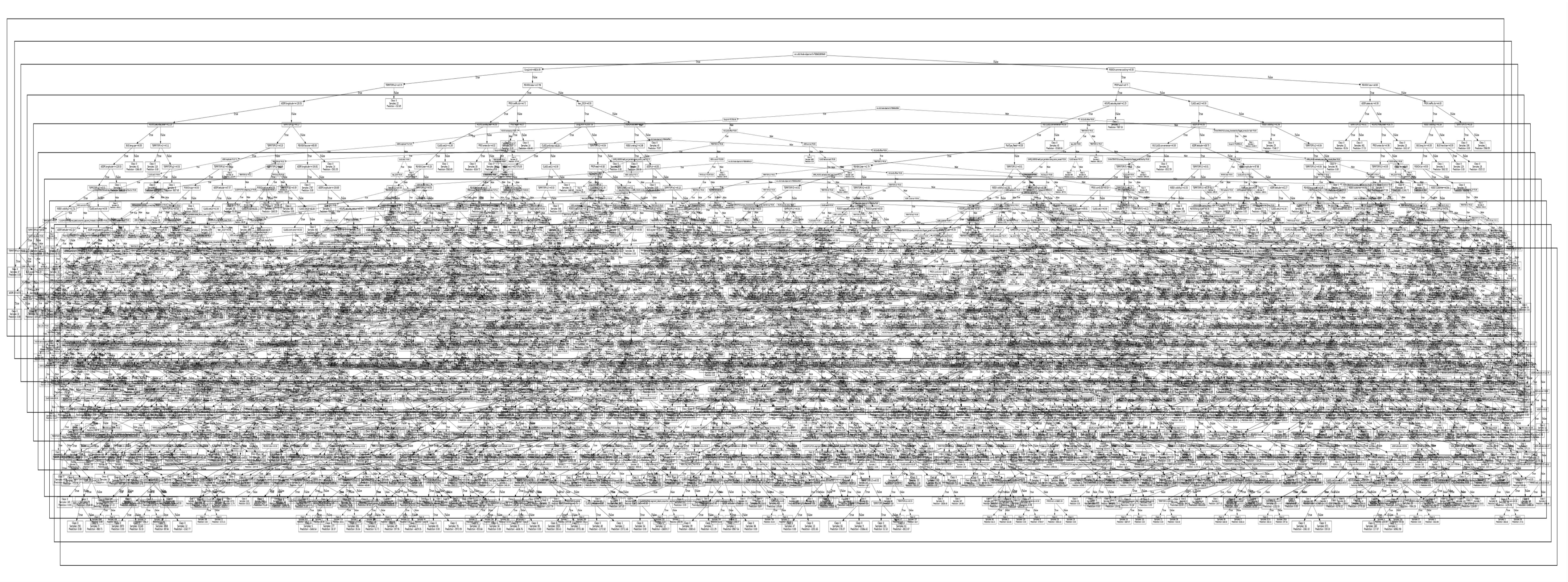

Fig. 3: Ten deep HTs from a HT ensemble

Motivation

- Interpretable HT pricing tool
  - Trees are supposed to be easily interpretable
    - Modern insurance datasets are much larger
    - **Deep** and **large ensemble** trees are almost impossible to interpret
  - To generate interpretable pricing tools for actuaries
    - **Extract** a few **critical nodes** from large ensembles
    - **Reconstruct** a simple pricing model with competitive predictive capability

Related Work

- Modification of CART
  - Weighted CART (wCART, Lopez et al, 2019)
    - Reweight observations with Kaplan–Meier (KM) weights
  - Novel splitting measure (Hwang et al, 2020)
    - Purity measure-inspired criteria with tunable hyperparameters
  - Imbalanced loss fucntions (Hu et al, 2022)
    - Modifies CART splitting criteria for imbalanced learning
  - Expectation-Boosting (EB, Hou et al, 2025)
    - Utilizes Gradient Boosting Decision Tree (GBDT) to estimate mixture models

Related Work

- Risk loading
  - An "ancient" idea to cover expenses or profits by adjusting risk premiums
    - Borch, K., 1960; Buhlmann, H., 1970; Benjamin, S., 1986.
- Rule extraction from tree-based models
  - Stable and Interpretable RUle Set (SIRUS, Benard et al, 2021)
    - Extract decision rules from tree models and reconstruct a simple linear model
  - Reformulate binary classification (Verwer and Zhang, 2019)
    - As rule-based linear programming optimization to increase modeling efficiency
  - Meta Rule (Li et al, 2023)
    - Existence of common decision paths in tree-based models

Methodology: Modified HT

- HT growing
  - Classification- and regression-based impurity
- HT pruning
  - **Retain** CART *minimal cost-complexity pruning*
  - More pruning cost functions
- Leaf node regression models
  - Generalized Linear Model (GLM) + *GLM Net* + *Probability-based GLM/GLM Net*

Methodology: Risk loading

- Risk loading post hoc modification
  - Introduces risk loading to leaf nodes to modify predictions

$$\hat{y}_s = f_i(\mathbf{x}_s) + r_i \sqrt{\frac{\sum_{m:m \in \mathcal{M}_i} (y_m - \bar{y}_i)^2}{|\mathcal{M}_i|}}$$

where r_i is some risk loading factors at leaf node i, $f_i(\mathbf{x}_s)$ is the original model predictions, squared root part is the standard deviation at leaf node i.

  - Risk loading factors can be difficult to quantify
    - Experts adjust these factors based on **experience**
    - **Data-based** optimization: Maximize Gini index while retaining Percentage Error (PE)

Methodology: Rule Extraction and Reconstruction

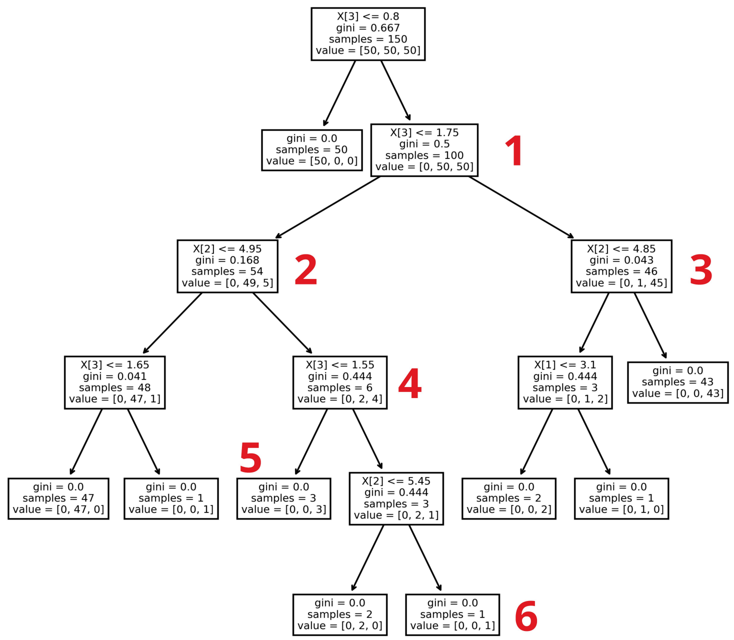

Fig. 3: Example of a decision tree

Definition 1 (Extended child node). A node T^{EC} is considered an extended child of node T if it resides within the subtree rooted at T, such that the removal of node T would also eliminate T^{EC} from the tree.

Example 1

- Node 2 is a (EC) child node of Node 1
- Node 5/6 are EC child nodes of Node 2
- Node 5/6 are NOT (EC) child node of Node 3

Methodology: Rule Extraction and Reconstruction

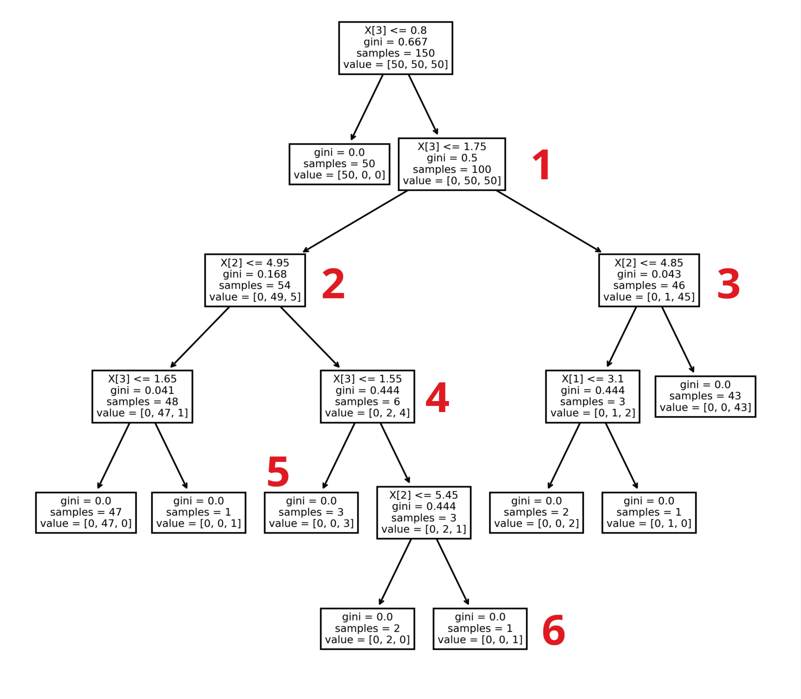

Fig. 3: Example of a decision tree

Definition 2 (Decision path). A series of tree nodes $\{T^{(1)}, T^{(2)}, \ldots, T^{(Q)}\}$ form a Q-layer decision path $h^{(Q)}$ in the hybrid tree if for every $q = 1, 2, \ldots, Q-1$, $T^{(q+1)}$ is a extended-child node of $T^{(q)}$. Furthermore, the decision path can be expressed as $h^{(Q)} = T^{(1)} \cap T^{(2)} \cap \ldots \cap T^{(Q)}$ where $T^{(q)}$ is the node at layer q.

Example 2

- Node 1-2-4 forms a decision path
- Node 2-4-6 forms a decision path
- Node 3-2-4 do NOT form decision path

Methodology: Rule Extraction and Reconstruction

Definition 3 (Pricing path). For S HTs, a pricing path h is a decision path that exists in at least $\lceil bS \rceil$ HTs, for some practical occurrence probability $b \in (0, 1)$.

- Commonly observed pricing paths represent
  - Critical data splitting rules
  - Crucial pricing decisions
- Extract these pricing paths and reconstruct a simplified pricing model
  - **Transparent** and **interpretable** insurance pricing

Methodology: Rule Extraction and Reconstruction

- Pricing path requires multiple trees
  - Bagging ensemble
  - Multiple HT -> **heterogeneity**
  - Each trained on subset -> **preserve critical decision nodes**
- Directly extracting pricing paths is **computational expensive**
  - Exhaustive search is almost impossible
  - First translate into extraction of *sharing node*

Definition 4 (Sharing node). A sharing node T_s is a non-terminal HT node that exists in at least $\lceil bS \rceil$ HTs, for some practical occurrence probability $b \in (0, 1)$.

Methodology: Rule Extraction and Reconstruction

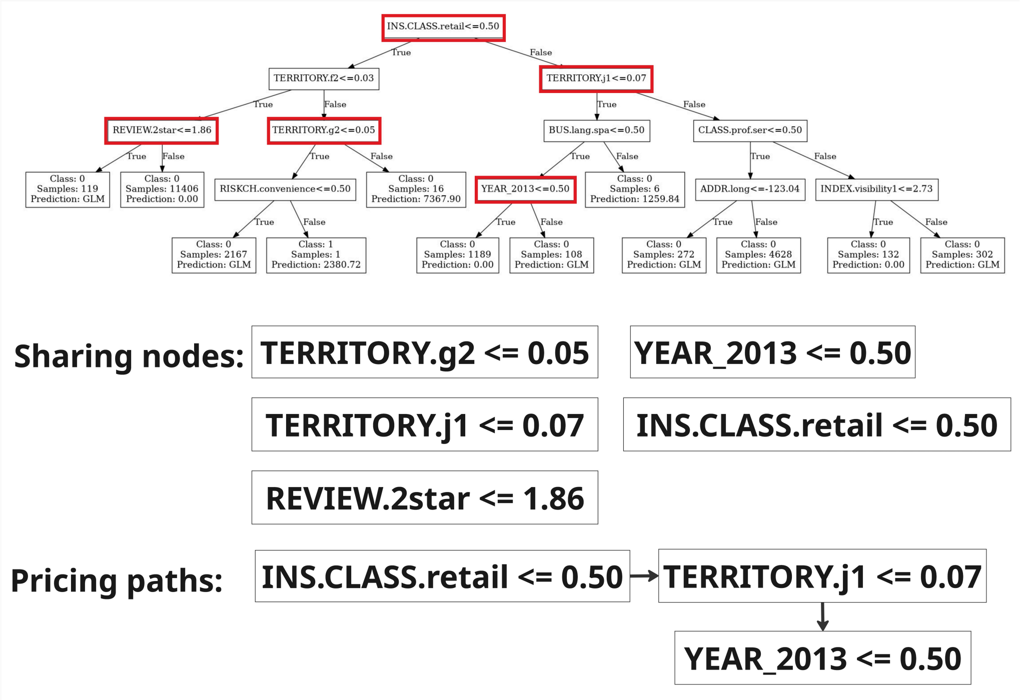

Fig. 4: Example of pricing path extraction

- To extract longest possible pricing paths
  - Start with L (*number of sharing nodes*)
  - Permutate sharing nodes to form candidate pricing paths
  - Validate the candidates
  - If
    - Found, return those pricing paths
    - Not, reduce length by 1 and repeat
    - Length is 1, return sharing nodes

Methodology: Rule Extraction and Reconstruction

- Benefits
  - Complexity of extraction $O(2^L)$
    - Number of sharing nodes L is usually small (~3-6)
  - **Order of nodes** is critical in pricing paths
    - Permutation ignores the order
    - Validation inherently encodes the hierarchical structure
  - Extracted pricing paths are usually **straightforward**
    - Easy to identify and categorize by actuaries

Methodology: Rule Extraction and Reconstruction

- With the extracted pricing paths
  - Reconstrucut a insurance pricing model with **competitive performance**
- As an intuitive solution
  - Replace the data splitting space using pricing paths
  - All possible splits -> A few feature + threshold pairs
  - Resulting model is **transparent** and **interpretable**
  - With risk loading, combine the reconstructed tree with risk loading

Empirical Experiments[1]: Real-life InsurTech Dataset

- InsurTech-enhanced Dataset
    - Introduced in Quan et al, (2025)
    - Collection of Business Owner's Policy (BOP) policies across 10-year time span
    - Identical pre-processing, data split is adopted
        - Selected business personal property (BP) coverage
        - 137,875 policies in the train set
        - 27,575 policies in the test set
        - 586 Insurance + InsurTech-enhanced features

[1] HT, Rule extraction and reconstruction: https://github.com/PanyiDong/HybridTree

Empirical Experiments: Real-life InsurTech Dataset

- Results

Model	Dataset	Gini	ME	MAE	Dataset	Gini	ME	MAE
Insurance in-house		0.59	-9.68	277.37		0.58	-15.08	270.75
Mean		-0.02	0.00	271.83		0.06	-5.92	265.74
Tweedie GLM		0.68	-0.03	262.64		0.36	-5.67	262.31
LightGBM		0.78	0.23	259.11		0.59	-7.17	262.78
HT	train	0.68	13.42	246.24	test	0.41	3.98	251.61
HT + Risk loading		0.69	11.57	245.62		0.54	3.49	249.38
HT ensemble		0.92	27.88	229.00		0.56	3.94	251.58
Rule reconstruction		0.54	8.93	258.96		0.42	1.61	255.30
Rule reconstruction + Risk loading		0.60	8.73	257.50		0.47	1.58	253.46

Empirical Experiments: Real-life InsurTech Dataset

- HT visualization

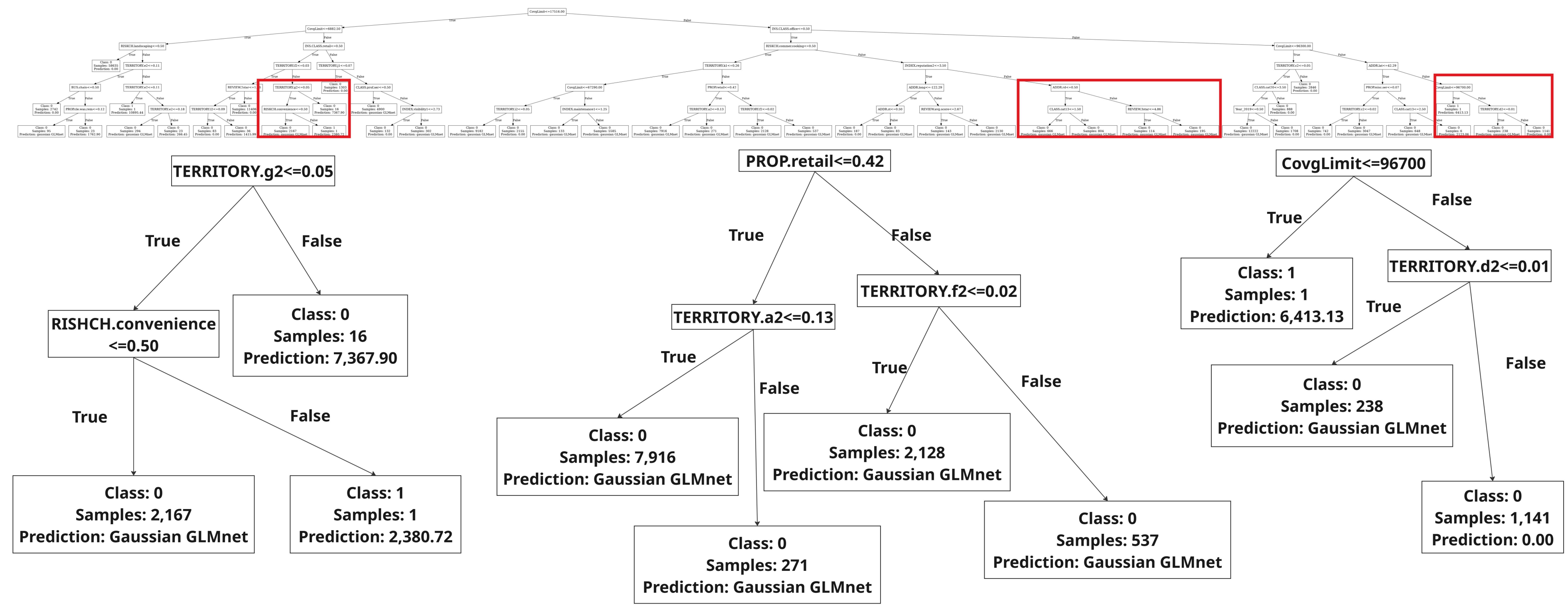

Fig. 5: HT trained on real-life data

Empirical Experiments: Real-life InsurTech Dataset

- HT visualization
  - Sharing nodes (**>80%** in **40** trees)
  - No *length*>=2 pricing paths found

Feature: *Year_2010*; Threshold: *0.50*
Feature: *Year_2011*; Threshold: *0.50*
Feature: *INS.CLASS.office*; Threshold: *0.50*
Feature: *TERRITORY.b2*; Threshold: *0.01*
Feature: *TERRITORY.c2*; Threshold: *0.02*

Empirical Experiments: Real-life InsurTech Dataset

- HT visualization

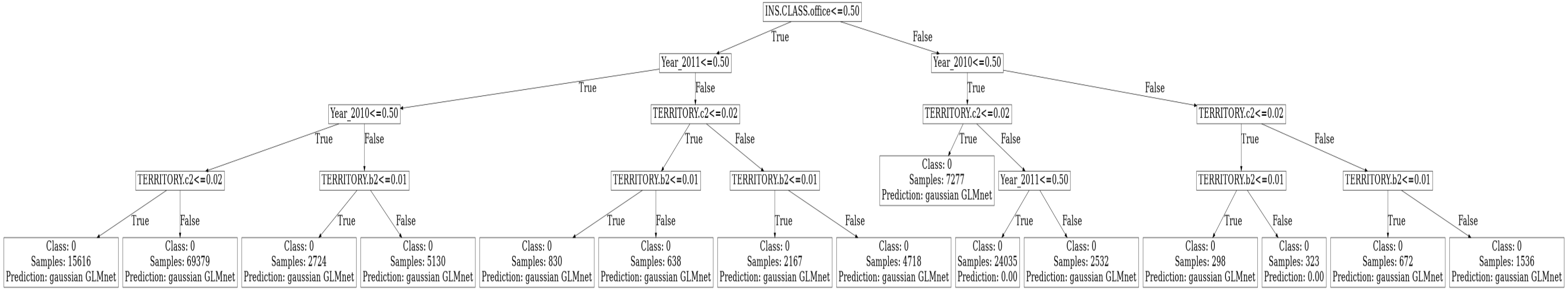

Fig. 6: Reconstructed HT on real-life data

Conclusion

- HybridTree
  - An alternative of CART to capture compound insurance frequency-severity
  - Modifications allows more flexible tree growing/pruning
  - Risk loading as post hoc modification to serve insurer's expectations
- Rule-based insurance pricing
  - Extract critical decision paths/nodes
  - Reconstruct a **transparent** and **interpretable** insurance pricing model

Selected Reference

- B´enard, C., Biau, G., da Veiga, S., and Scornet, E. (2021). Interpretable Random Forests via Rule Extraction. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 937–945. PMLR.
- Benjamin, S. (1986). Loadings for Insurance Premiums. *The Geneva Papers on Risk and Insurance*, 11(39):110–125.
- Borch, K. (1960). The safety loading of reinsurance premiums. *Scandinavian Actuarial Journal*, 1960(3-4):163–184.
- Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, New York.
- Buhlmann, H. (1970). *Mathematical Methods in Risk Theory*. Grundlehren Der Mathematischen Wissenschaft. Springer, Berlin, Heidelberg.

Selected Reference

- Frees, E. W. J., Meyers, G., and Cummings, A. D. (2014). Insurance ratemaking and a gini index. *Journal of Risk and Insurance*, 81(2):335–366.
- Hou, Y., Li, J., and Gao, G. (2025). Insurance loss modeling with gradient tree-boosted mixture models. *Insurance: Mathematics and Economics*, 121:45–62.
- Hu, C., Quan, Z., and Chong, W. F. (2022). Imbalanced learning for insurance using modified loss functions in tree-based models. *Insurance: Mathematics and Economics*, 106:13–32.
- Hwang, S., Yeo, H. G., and Hong, J.-S. (2020). A New Splitting Criterion for Better Interpretable Trees. *IEEE Access*, 8:62762–62774.
- Li, Q., Xie, C., Xu, X., Liu, X., Zhang, C., Li, B., He, B., and Song, D. (2023). Effective and Efficient Federated Tree Learning on Hybrid Data.

Selected Reference

- Lopez, O., Milhaud, X., and Th´erond, P.-E. (2019). A TREE-BASED ALGORITHM ADAPTED TO MICROLEVEL RESERVING AND LONG DEVELOPMENT CLAIMS. *ASTIN Bulletin: The Journal of the IAA*, 49(3):741–762.
- Quan, Z., Hu, C., Dong, P., and Valdez, E. A. (2025). Improving business insurance loss models by leveraging insurtech innovation. *North American Actuarial Journal*, 29(2):247–274.
- Quan, Z., Wang, Z., Gan, G., and Valdez, E. A. (2023). On hybrid tree-based methods for short-term insurance claims. *Probability in the Engineering and Informational Sciences*, 37(2):597–620.
- Verwer, S. and Zhang, Y. (2019). Learning Optimal Classification Trees Using a Binary Linear Program Formulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1625–1632.

Thank you! Q&A

Appendix: Methodology: CART

- Traditional CART
  - Best data split = Largest impurity decrease

$$Im(\mathbf{y}) - \frac{|\mathcal{M}_L|}{|\mathcal{M}|} Im(\mathbf{y}_L) - \frac{|\mathcal{M}_R|}{|\mathcal{M}|} Im(\mathbf{y}_R)$$

  - Growing impurity measures
    - Gini index: $Im_{gini}(\mathbf{y}) = 1 - \sum_{k=1}^{K} p_k^2$
    - Entropy: $Im_{entropy}(\mathbf{y}) = -\sum_{k=1}^{K} p_k \log(p_k)$

where $p_k = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} 1_{y_m = k}$ for $\mathcal{M}$ observations at the leaf node.

Appendix: Methodology: Modified HT

- Classification-based impurity
  - Mis-classifications rate: $Im_{mis}(\mathbf{y}) = \frac{\sum_{m=1}^{|\mathcal{M}|} 1_{y_m \neq \hat{y}}}{|\mathcal{M}|}$
  - Balanced mis-classifications rate: $Im_{bal\_mis}(\mathbf{y}) = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{m=1}^{|\mathcal{M}|} 1_{y_m \neq \hat{y}} 1_{y_m = k}}{\sum_{m=1}^{|\mathcal{M}|} 1_{y_m = k}}$
- Regression-based impurity[2]
  - Mean Absolute Error (MAE): $Im_{mae}(\mathbf{y}) = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} |y_m - \hat{y}|$
  - Mean Squared Error (MSE): $Im_{mse}(\mathbf{y}) = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} (y_m - \hat{y})^2$

[2] Regression-based impurity measures may be misaligned with the classification-oriented goal of identifying risk segments. However, they are retained to provide users with greater flexibility when applying HTs to regression tasks.

Appendix: Methodology: Modified HT

- Pruning
  - Minimal cost-complexity pruning

    $$CC(I) = C(I) + \alpha|I|$$

  where C is cost function, α denotes complexity parameter (cp) for a tree with I leaf nodes.
  - Pruning criteria: Mis-classification, MAE, and MSE
  - **Retain** CART pruning process

Appendix: Methodology: Modified HT

- Leaf node regression models
  - Generalized Linear Regression (GLM)
    - Gaussian family (simple linear regression) sufficient in most senarios
  - GLM net
    - High dimensional data
  - Probability-based GLM/GLM net
    - Two-step model
    - Probability of claims + Expected claims

Appendix: Methodology: Rule Extraction and Reconstruction

- Algorithm summary

Algorithm 1: Pricing path extraction

Input: S hybrid trees $E = \{E_1, E_2, ..., E_S\}$; Occurrence threshold b

Output: Extracted decision paths/nodes H

$H = \{\}$;
$T_S \leftarrow GetSharingNode(E)$;
/* Get sharing decision nodes that appears in at least $\lceil bS \rceil$ trees */
$L = length(T_S)$; /* Length of all sharing nodes */
$l = L$; **while** $l > 1$ **do**
    if $l = 1$ **then**
        return T_S;
    end
    for $h \in Comb(L, l)$ **do**
        /* Loop through all combinations of sharing nodes with length l */;
        if *h is a valid path in $\lceil bS \rceil$ trees* **then**
            $H \leftarrow h$;
        end
    end
    $l = l - 1$;
end
return H;

Appendix: Empirical Experiments: Simulated Tweedie Dataset

- Data generation $\mathcal{D} = (\mathbf{X}, \mathbf{y})$
  - Features: $\mathbf{X} = [\mathbf{X}_{cat}, \mathbf{X}_{con}]$ for *20* categorical variables $\mathbf{X}_{cat}$ and *20* continuous variables $\mathbf{X}_{con}$.
    - Categorical variables $\mathbf{X}_{cat}$: i.i.d. from $(-3, -2, 1, 4)$ with equal probability
    - Continuous variables $\mathbf{X}_{con}$: multi-variate normal with mean of $\mathbf{0}$ and identity covariance matrix
  - Response variable: $\mathbf{y} = (1 + 0.25|\delta|)\mathbf{y}_{true},\ \text{if}\ \mathbf{y}_{true} > 0; 0, \text{otherwise}.$
    - $\delta \sim \mathcal{N}(0, 1)$ is Gaussian noise

Appendix: Empirical Experiments: Simulated Tweedie Dataset

- Data generation $\mathcal{D} = (\mathbf{X}, \mathbf{y})$
  - True response variable: $\mathbf{y}_{true} \sim Gam(|Poi(\hat{\tau})|, \hat{\mu}^{0.5})$
    - Tweedie distribution with power of 1.5 and dispersion of 2
    - $\hat{\tau} = \frac{\tau}{\bar{\tau}}$ and $\hat{\mu} = 1000\frac{\mu}{\bar{\mu}}$
    - Poisson component: $\tau = e^{(-0.1+\mathbf{X}_{con}\beta_{Poi}+\mathbf{X}_{cat}\beta_{Poi})/2}$
    - Gamma component: $\mu = e^{6+\mathbf{X}_{con}\beta_{Gam}+\mathbf{X}_{cat}\beta_{Gam}}$
    - Coefficients of Poisson component: $\beta_{Poi,j} = -0.4 + 0.05j$
    - Coefficients of Gamma component: $\beta_{Gam,j} = -0.08 + 0.01j$

Appendix: Empirical Experiments: Simulated Tweedie Dataset

- Results

Model	Dataset	Gini	ME	MAE	Dataset	Gini	ME	MAE
Mean	train	-0.03	0.00	89.56	test	0.12	0.61	90.33
Tweedie GLM		0.91	-0.07	39.42		0.91	-1.13	41.28
HT		0.77	-10.25	62.86		0.77	-11.92	71.89
HT + Risk loading		0.79	-0.20	58.41		0.79	-1.10	65.95
HT ensemble		0.88	-6.22	57.90		0.86	-5.24	66.43
Rule reconstruction		0.52	7.69	57.56		0.62	9.13	59.57
Rule reconstruction + Risk loading		0.55	0.48	62.51		0.64	4.48	59.65

Appendix: Empirical Experiments: Simulated Tweedie Dataset

- HT visualization

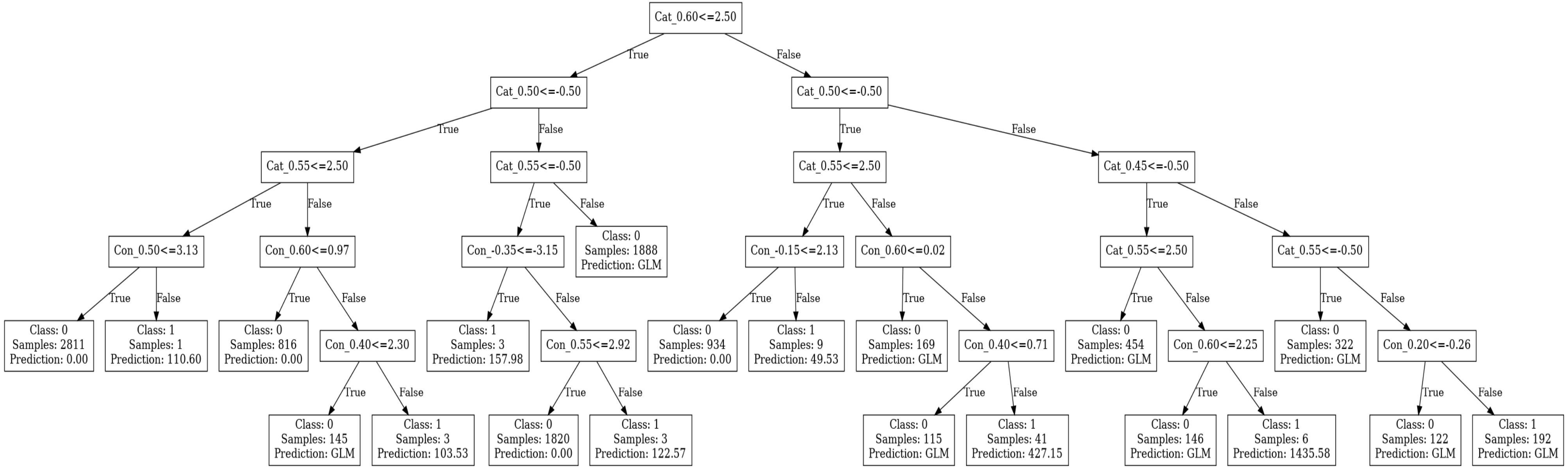

Fig. 7: HT trained on simulation data

Appendix: Empirical Experiments: Simulated Tweedie Dataset

- HT visualization

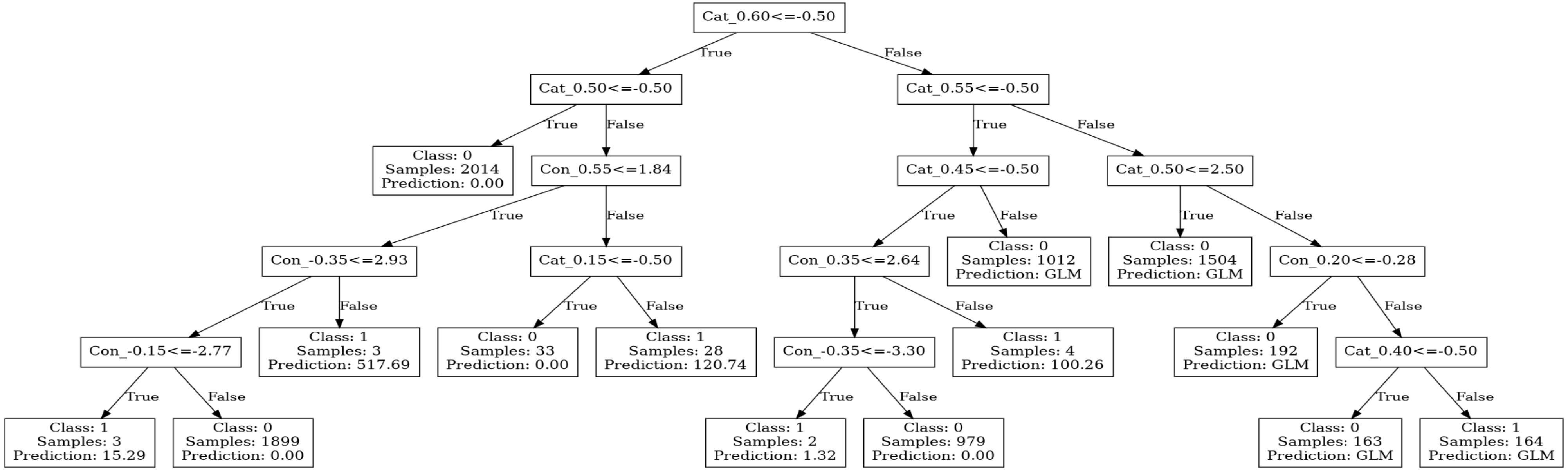

Fig. 8: First HT in the ensemble on simulation data

Appendix: Empirical Experiments: Simulated Tweedie Dataset

- HT visualization
  - Sharing nodes (**>=60%** in **100** HTs)

Feature: *Cat_0.40*; Threshold: *-0.5*
Feature: *Cat_0.45*; Threshold: *-0.5*
Feature: *Cat_0.45*; Threshold: *2.50*
Feature: *Cat_0.50*; Threshold: *-0.5*
Feature: *Cat_0.50*; Threshold: *2.50*
Feature: *Cat_0.55*; Threshold: *-0.5*
Feature: *Cat_0.55*; Threshold: *2.50*
Feature: *Cat_0.60*; Threshold: *2.50*

Appendix: Empirical Experiments: Simulated Tweedie Dataset

- HT visualization
  - Extracted pricing paths

Feature: *Cat_0.45*; Threshold: *-0.5* --- Feature: *Cat_0.60*; Threshold: *2.50*

Feature: *Cat_0.50*; Threshold: *-0.5* --- Feature: *Cat_0.60*; Threshold: *2.50*

Feature: *Cat_0.50*; Threshold: *2.50* --- Feature: *Cat_0.60*; Threshold: *2.50*

Feature: *Cat_0.55*; Threshold: *-0.5* --- Feature: *Cat_0.60*; Threshold: *2.50*

Feature: *Cat_0.55*; Threshold: *2.50* --- Feature: *Cat_0.60*; Threshold: *2.50*

Appendix: Empirical Experiments: Simulated Tweedie Dataset

- HT visualization

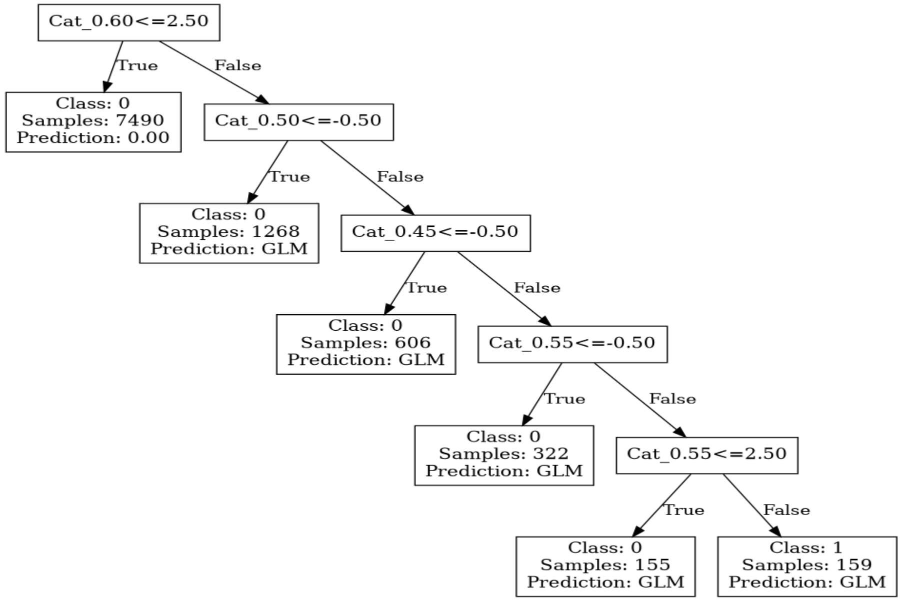

Fig. 9: Reconstructed HT on simulation data

Appendix: Notation

Notation	Description
Im	impurity measure
$(\mathbf{X}, \mathbf{y})$	pair of feature matrix and response vector
$\mathcal{M}$	partition index of leaf node
p_k	the probability for each class k
f	leaf node regression model
r	risk-loading factor
T	decision node
$h^{(Q)}$	Q-layer decision path
b	occurance probability
L	maximum length of decision paths
E	hybrid tree model

Appendix: Evaluation Metrics

$$Gini(y, \hat{y}) = 1 - \frac{2}{N-1}\left(N - \frac{\sum_{n=1}^{N} n y_{[n]}}{\sum_{n=1}^{N} y_{[n]}}\right)$$

$$ME(y, \hat{y}) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)$$

$$MAE(y, \hat{y}) = \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n|$$