Hybrid Tree-based Interpretable Pricing

Panyi Dong

University of Illinois Urbana-Champaign

2026-07-02

Synopsis

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

Background

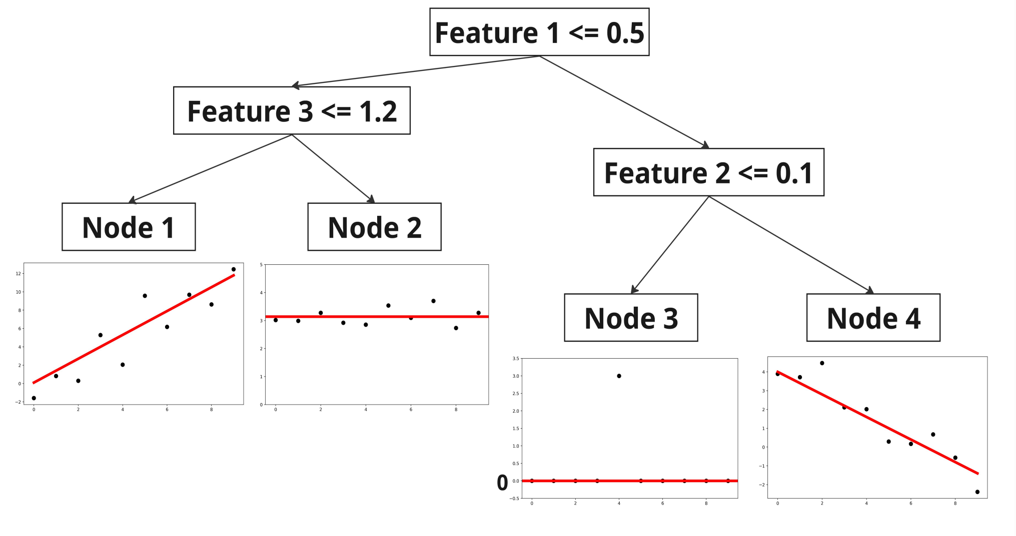

Fig. 1: HybridTree (HT)

- 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
        - **Conventional** classification tree
        - Limited **growing/pruning** measures
    - **Proposed framework**
        - New implementation of HT from scratch
        - Introduces classification- and regression-based measures
        - Risk adjustment as post hoc modification

Motivation

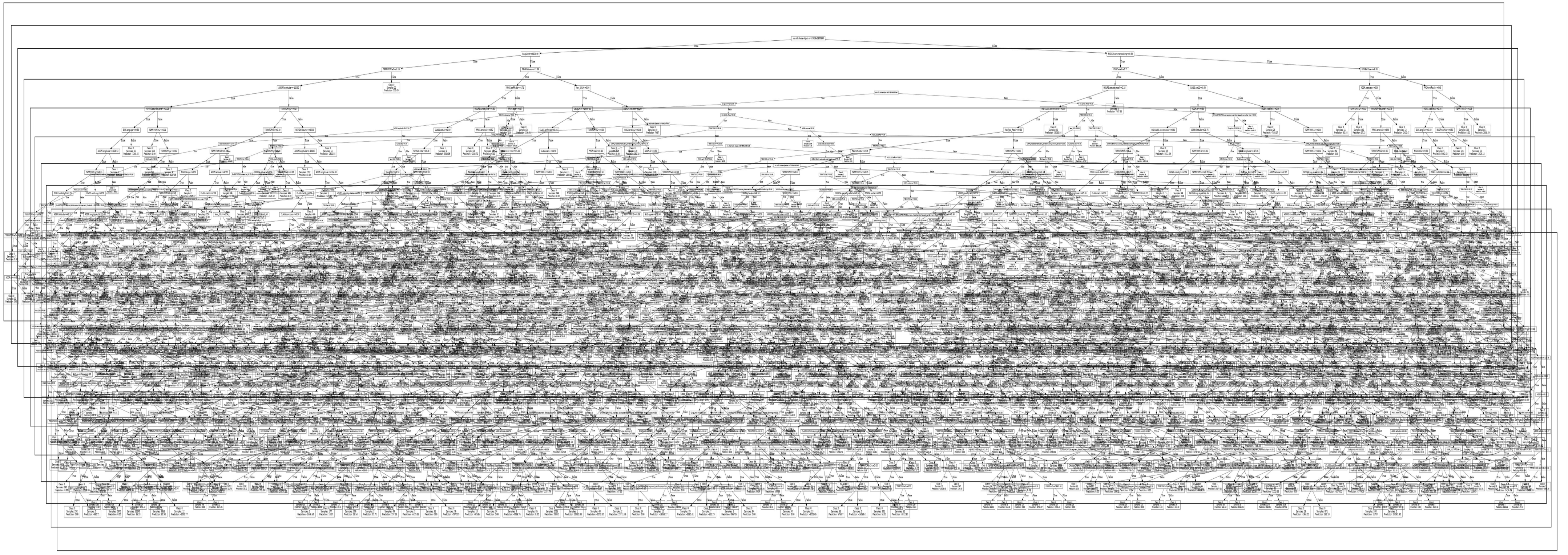

Fig. 2: Ten deep HTs from a HT ensemble

Motivation

- Interpretable HT pricing tool
  - Trees are supposed to be easily interpretable
    - Trees built on the 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

Methodology: Overview

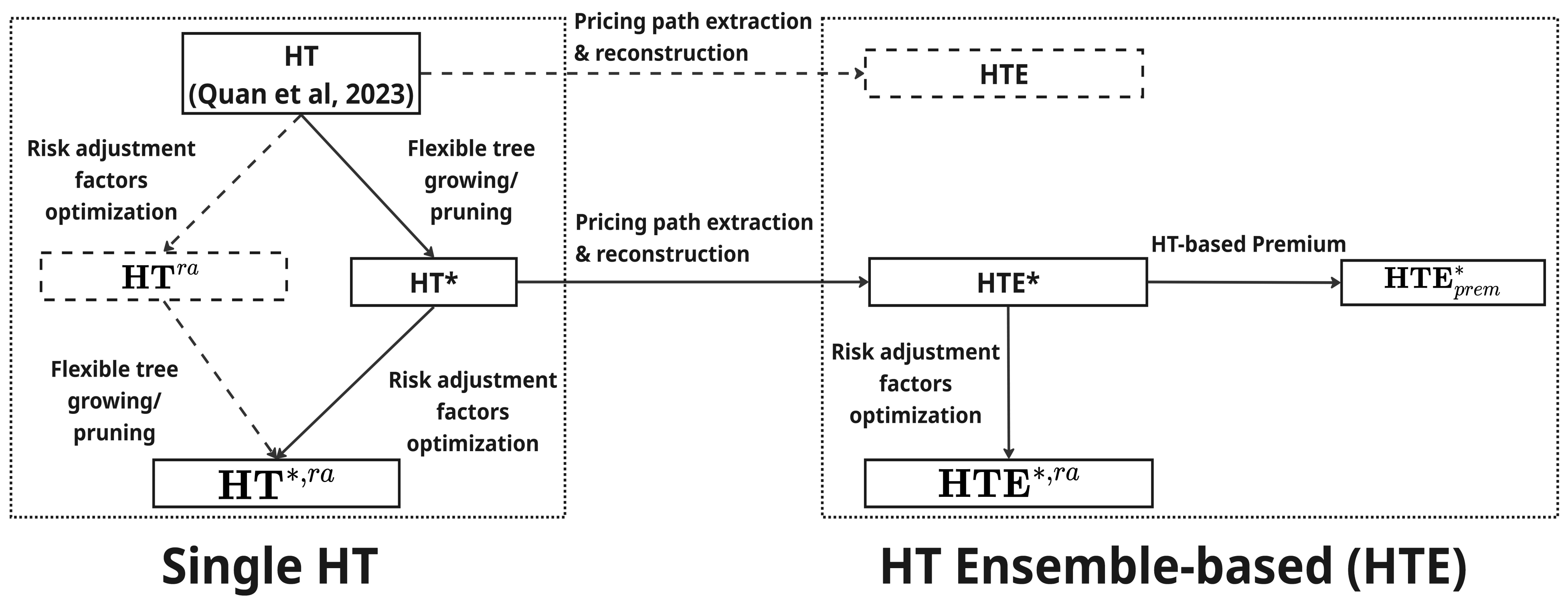

Fig. 3: HT workflow

Methodology: Modified HT (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 adjustment (RA)

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

$$\hat{y}_{j_s} = g_i(\mathrm{x}_j) + r_i \sqrt{\frac{\sum_{j:j\in \mathrm{M}_i} (y_{j_s} - \overline{y}_{i_s})^2}{|\mathrm{M}_i|}}$$

where r_i is the risk adjustment factors at leaf node i, $g_i(\mathrm{x}_j)$ is the original model predictions, squared root part is the standard deviation at leaf node i.

  - Risk adjustment factors can be determined by
    - Experts adjust these factors based on **experience**
    - **Data-driven** optimization: Maximize Gini index while retaining Percentage Error (PE)

Methodology: Extraction of Rule-based Pricing (HTE*)

- Pricing path
  - **Important** decision paths in HT
  - **Frequently observed** in decision-making
- 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*

Methodology: Extraction of Rule-based Pricing (HTE*)

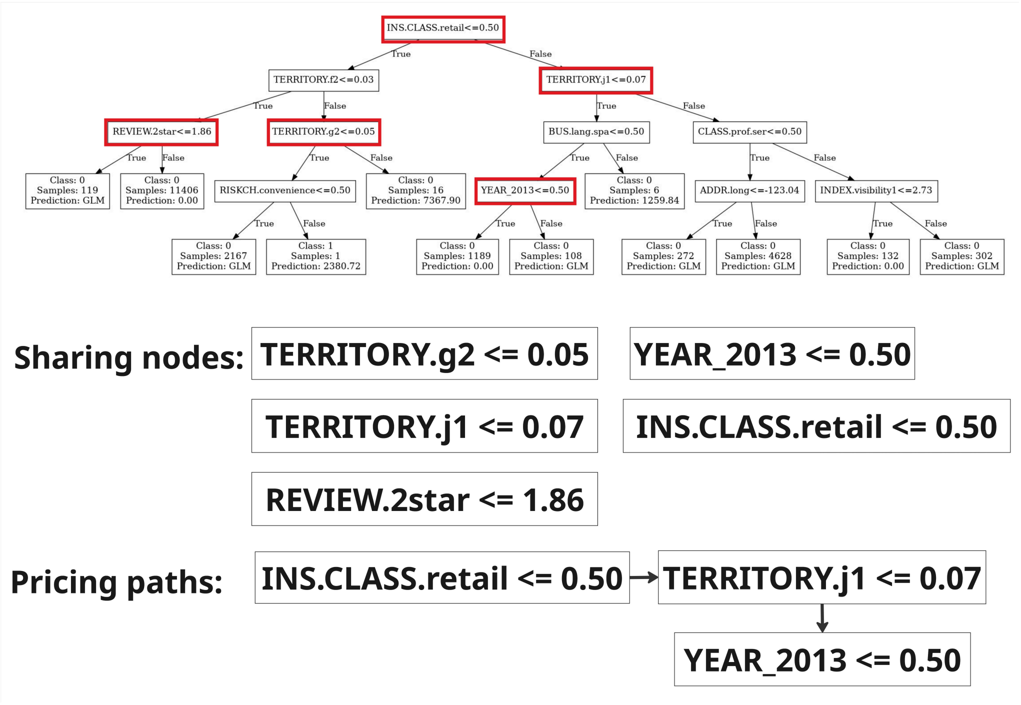

Fig. 4: Example of pricing path extraction

- 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: Extraction of Rule-based Pricing (HTE*)

- 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 adjustment, combine the reconstructed tree with risk adjustment

Methodology: HT-based Premium

- HT modeling (or general ML modeling)
  - Accurate loss predictions
  - Difficult for applications such as **premium pricing**
- HT-based premium
  - $P(\mathrm{x}_j) = \kappa(\mathrm{x}_j)\theta(\mathrm{x}_j)$ (frequency-severity decomposition)
  - Frequency: $\kappa(\mathrm{x}_j) = \frac{\sum_{j:j\in \mathrm{M}_i} 1_{y_{j_f}=1}}{|\mathrm{M}_i|}$ (node probability)
  - Severity:

$$\theta(\mathrm{x}_j) = \begin{cases} P^{st}, & \text{if} \quad \kappa(\mathrm{x}_j) < \eta \quad \text{and} \quad \hat{y}^r_{j_s} < \xi P^{st} \\ \hat{y}^r_{j_s}, & \text{o.w.} \end{cases}$$

where $\eta \in (0, 1]$ and $\xi \in (0, 1]$ are user-defined thresholds that control prediction confidence and severity magnitude, P^{st} is standard premium.

Empirical Experiments: 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

Empirical Experiments: Real-life InsurTech Dataset

- Results

Dataset	Model	GI	ME	MAE
	Insurance in-house model	0.59	-9.68	277.37
	Mean	-0.02	0.00	271.83
	Tweedie GLM	0.68	-0.03	262.64
	LightGBM	0.78	0.23	259.11
train	HT*	0.68	13.42	246.24
	HT^{*,ra}	0.69	11.57	245.62
	HTE*	0.54	8.93	258.96
	HTE^{*,ra}	0.60	8.73	257.50
	Insurance in-house model	0.58	-15.08	270.75
	Mean	0.06	-5.92	265.74
	Tweedie GLM	0.36	−5.67	262.31
	LightGBM	0.59	−7.17	262.78
test	HT*	0.41	3.98	251.61
	HT^{*,ra}	0.54	3.49	249.38
	HTE*	0.42	1.61	255.30
	HTE^{*,ra}	0.47	1.58	253.46

Tab. 1: Model performance on real-life InsurTech dataset

Empirical Experiments: Real-life InsurTech Dataset

- HT visualization

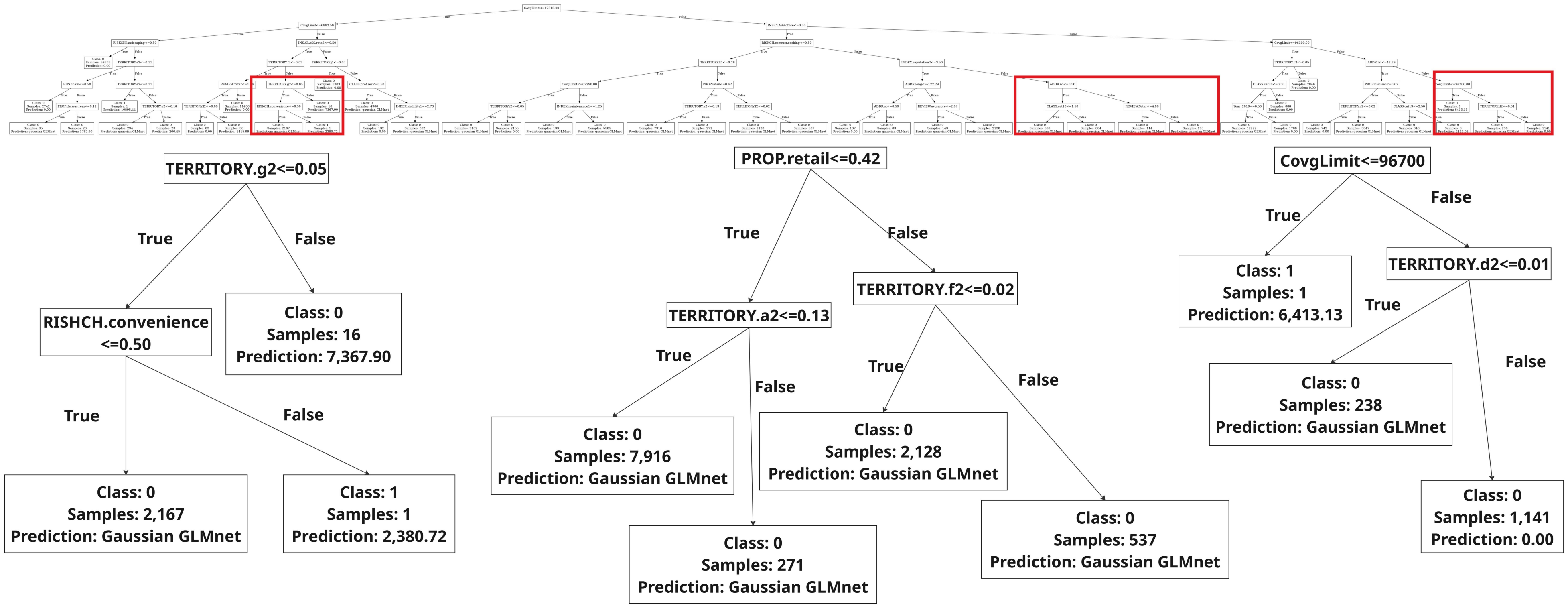

Fig. 5: HT trained on real-life data

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

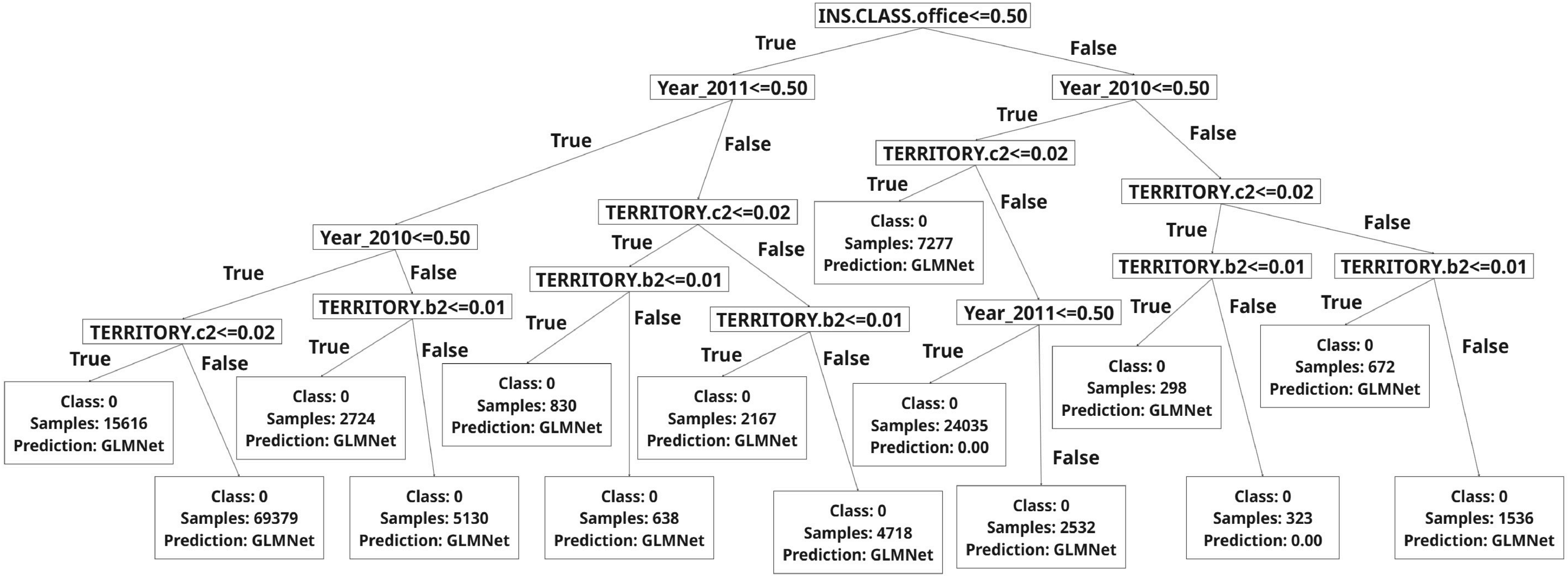

Fig. 6: Reconstructed HT on real-life data

Empirical Experiments: Real-life InsurTech Dataset

- HT-based Premium

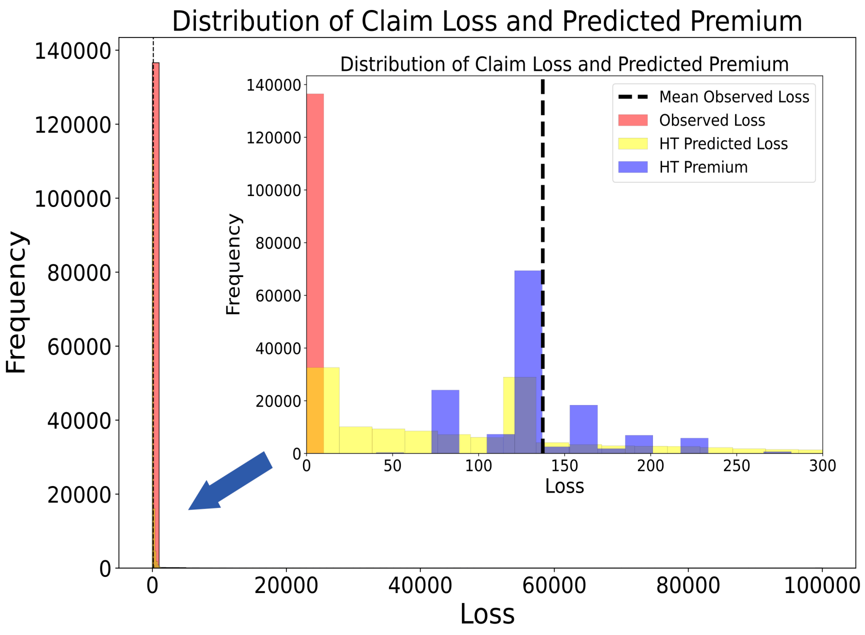

Fig. 7: Distribution of observed, HT-based prediction and premium

- HT-based premiums
  - more consistent with actuarial pricing principles

Statistics	Observed Loss	HT-based Loss	HT-based Premium
Mean	137.29	129.19	131.37
Std	2,502.25	207.77	41.33
Min	0	0	40.46
50%	0	98.77	123.38
Max	100,000	19,062.12	362.14

Tab. 3: Summary statistics

Conclusion

- HybridTree
  - An alternative of CART to capture compound insurance frequency-severity
  - Modifications allows more flexible tree growing/pruning
  - Risk adjustment as post hoc modification to serve for ratemaking purpose
- Rule-based insurance pricing
  - Extract critical decision paths/nodes
  - Reconstruct a **transparent** and **interpretable** insurance pricing model
  - Can be extended to other black box models

Thank you! Q&A

Selected Reference

- Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, New York.
- Frees, E. W. and Valdez, E. A. (2008). Hierarchical Insurance Claims Modeling. *Journal of the American Statistical Association*, 103(484):1457–146
- 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.
- Jeong, H., Valdez, E. A., Ahn, J. Y., and Park, S. C. (2021). Generalized Linear Mixed Models for Dependent Compound Risk Models. *Variance*, 14(1).

Selected Reference

- 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.
- Lee, G. Y. and Shi, P. (2019). A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics*, 87:115–129.
- 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.
- Su, X. and Bai, M. (2020). Stochastic gradient boosting frequency-severity model of insurance claims. *PLOS ONE*, 15(8):e0238000.

Appendix: Notation

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