Hybrid Tree-based Models for Insurance Claims

The second Insurance Data Science Conference - 14 June 2019

Zhiyu Quan
PhD Candidate, University of Connecticut

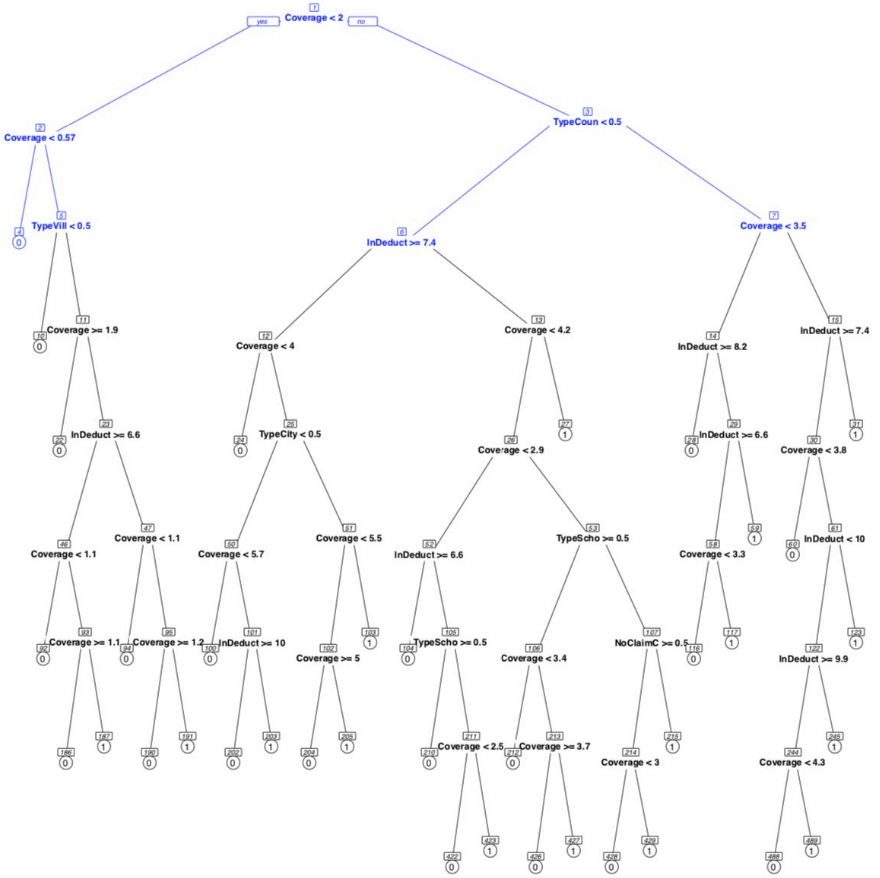
Short-term insurance and traditional approaches

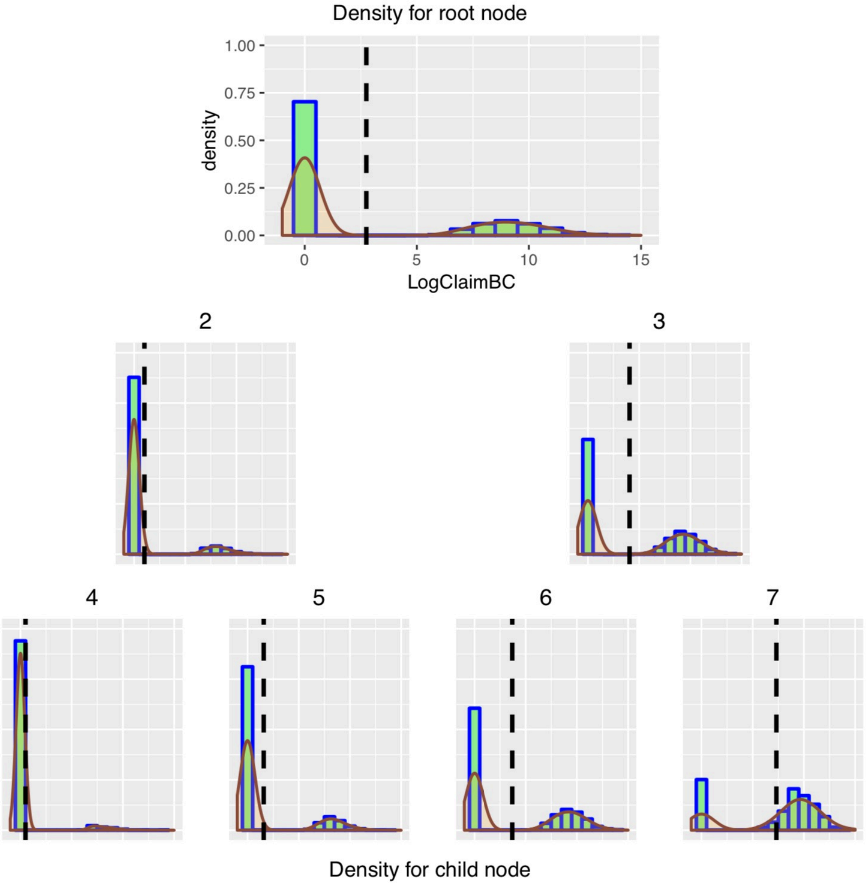
- Insurance datasets feature information about claim frequency and severity for each policy or observation.
- For most insurance claims datasets, there is typically a large proportion of zero claims that leads to imbalances that cause inferior prediction accuracy of traditional approaches.
- The two-part framework uses the Poisson-gamma approach.
  - Overlooks the internal connection between the low frequency and the subsequent low loss amount.
- Tweedie GLM (parsimonious)
  - Constant scale, or dispersion, parameter causes the mean increases with its variance.

Hybrid Tree-based Models

- The first step is the construction of a classification tree to build the probability model for frequency.
    - Binary response variable: CART, C4.5, eneralized, Unbiased, Interaction Detection and Estimation (GUIDE), Conditional Inference Trees (CTREE).
    - Count response variable: piecewise linear Poisson using CART, SUPPORT, Poisson regression using GUIDE, MOdel-Based recursive partitioning (MOB).
- In the second step, we employ linear regression model at each terminal node to build the distribution model for severity.
    - Generalized Linear Models (GLM) with various families.
    - Regression with elastic net regularization.

Visualize the change of density with binary splitting





The prediction performance on LGPIF dataset

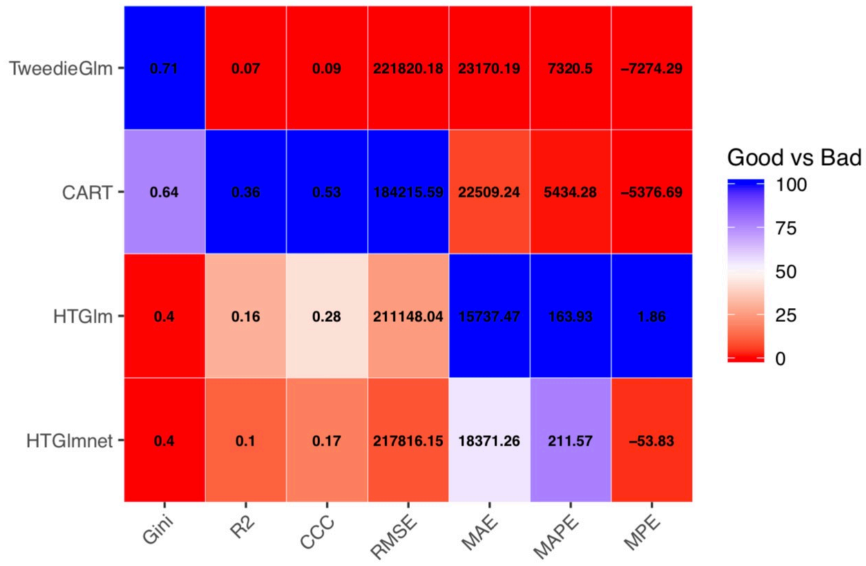


Figure 1: Model fitting.

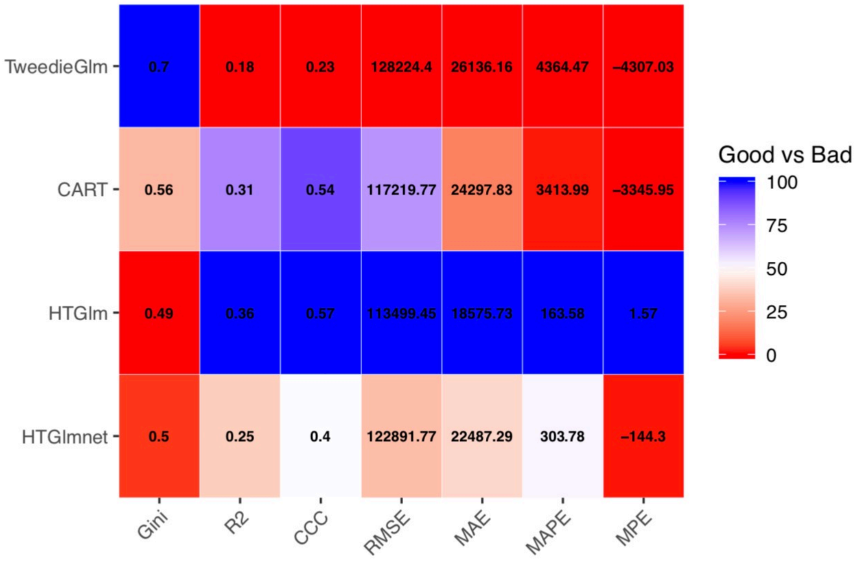


Figure 2: Model performance on test dataset.

- Compare Tweedie GLM, Regression Tree (CART), Hybrid Tree-based Model with simple GLM with Gaussian family (HTGlm), Hybrid Tree-based Models with elastic net regression (HTGlmnet) based on various validation measures.

Reference

- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Taylor & Francis Group, LLC: Boca Raton, FL.
- Xacur, O. A. Q. and Garrido, J. (2015). Generalised linear models for aggregate claims: to tweedie or not? European Actuarial Journal, 5(1):181–202.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320. Taylor & Francis Group, LLC: Boca Raton, FL.
- Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. Journal of Computational and Graphical Statistics, 17(2):492–514.

Acknowledgment and Q&A

We would like to thank the Society of Actuaries for the funding support of this research project through our Centers of Actuarial Excellence (CAE) grant on data mining.

Thank you for your attention!