Imbalanced Learning for Insurance using Modified Loss Functions in Tree-Based Models

Insurance Data Science Conference

Zhiyu (Frank) Quan

University of Illinois Urbana-Champaign

2022/06/15

Outline

- **Why** - Motivation
  - The imbalance problem in insurance loss modeling
  - Potential pitfall of CART
- **How** - Modification to loss function
  - WSSE loss function
  - Canberra loss function
- **Result** - Results on motivating examples and the simulated dataset

Motivation

- Insurance claim datasets usually contain **a high percentage of zero claims**.
- Imbalance problem:
  - Majority (zero claims), minority (non-zero large claims).
  - Standard algorithms fail to properly depict data characteristics and therefore yield poor prediction accuracy.

Imbalanced learning techniques

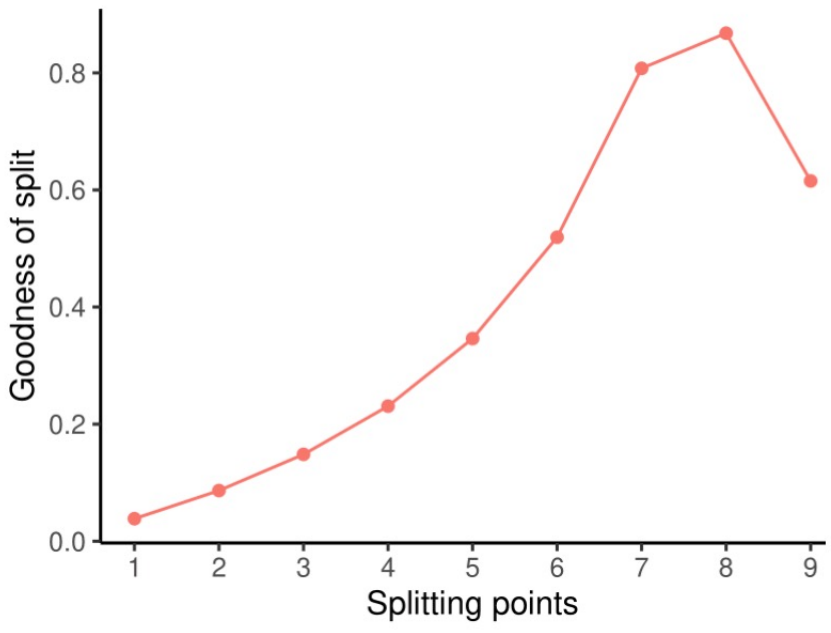
- **Resampling:** rebalance the sample space.
  - Over-sampling: adding more samples from the minority, e.g., SMOTE (Chawla et al., 2002).
  - Under-sampling: removing samples from the majority.
- **Ensemble methods:** combine weak learners to improve prediction ability.
  - Parallel-based ensembles: bagging.
  - Iterative-based ensembles: boosting, e.g., Adaboost (Freund and Schapire, 1996), TDboost (Yang et al., 2018).
- **Cost-sensitive learning:** assign different costs for different prediction errors.
  - In the real world, different misclassifications often have various interpretations.
  - Cost-sensitive learning modifies the cost of misclassification by adding penalties to misclassified predictions related to the objects of interest.

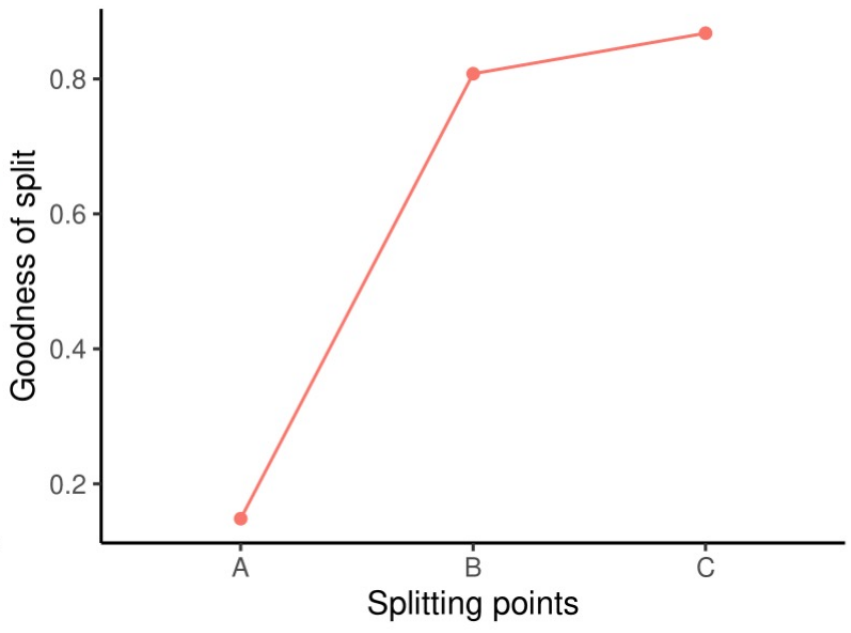
Inspiration from cost-sensitive learning

- We borrow the idea of **cost-sensitive learning** to **modify the loss function of CART**, making it more suitable for imbalanced datasets.
  - Assign different weights to zero and non-zero prediction errors,
  - Inject the **"classification" of zero and non-zero claims** into our **regression model**.
- We chose to modify a single tree for the following considerations:
  - Compared with resampling techniques, cost-sensitive learning **preserves the original distribution of the dataset**.
  - Compared to ensemble techniques, a single tree **maintains its advantage of interpretability**, and the modified single tree can also be used as a base learner in ensemble techniques.

Motivating example - a pitfall of the default split

OBS.	X_1	X_2	Y
1	1	A	0
2	2	A	0
3	3	A	0
4	4	B	0
5	5	B	0
6	6	B	0
7	7	B	0
8	8	C	1
9	9	D	2
10	10	D	3





- The default method (ANOVA) in CART **cannot separate zeros and non-zeros** as expected.
- The zeros will be **combined** with some **small** but non-zero values.
- The sum of squared errors is **greatly affected** by the prediction error of the non-zero responses.

The overview of CART algorithm and its notation

- **Step 1**: Grow a large tree.
  - Recursive binary splitting.
- **Step 2**: Prune the large tree.
  - Cost-complexity pruning.
- **Notation**:
  - Consider a dataset $(\mathbf{X}, \mathbf{y}) = \left((\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \ldots, (\mathbf{X}_N, y_N)\right)^{\mathrm{T}}$ with N observations.
  - For each i-th observation, where $i = 1, 2, \ldots, N$,
    - $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ is a vector of p explanatory variables sampled from a space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_p$,
    - y_i is a response variable sampled from a space $\mathcal{Y}$.

CART - grow a large tree

- Regression tree, denoted by $T(\mathbf{X}_i; \theta)$, is produced by **partitioning the space of the explanatory variables** $\mathcal{X}$ into M disjoint regions, which are denoted as $R_1, R_2, \ldots, R_M$.
- For each region R_m, where $m = 1, 2, \ldots, M$, **a constant** $c_m \in \mathbb{R}_+$ is assigned as a predicted value for observations falling into the region R_m.
- The regression tree is given by: for each $i = 1, 2, \ldots, N$, the predicted value

$$\hat{y}_i = T(\mathbf{X}_i; \theta) = \sum_{m=1}^{M} c_m \mathbf{1}_{R_m}(\mathbf{X}_i),$$

where $\theta = (R_1, R_2, \ldots, R_M, c_1, c_2, \ldots, c_M)$ is the vector of parameters for the regression tree, and $\mathbf{1}_{R_m}(\mathbf{X}_i) = 1$, if $\mathbf{X}_i \in R_m \subseteq \mathcal{X}$, while $\mathbf{1}_{R_m}(\mathbf{X}_i) = 0$, if $\mathbf{X}_i \notin R_m$.

CART - recursive binary splitting

- The CART algorithm identifies the optimal parameters for the regression tree via **recursive binary splittings**.
- Fix a binary splitting step $u = 1, 2, \ldots$.
- Denote $\left(\mathbf{X}^{(u)}, \mathbf{y}^{(u)}\right) = \left(\left(\mathbf{X}_1^{(u)}, y_1^{(u)}\right), \left(\mathbf{X}_2^{(u)}, y_2^{(u)}\right), \ldots, \left(\mathbf{X}_{N^{(u)}}^{(u)}, y_{N^{(u)}}^{(u)}\right)\right)^{\mathrm{T}}$ as the remaining dataset with $N^{(u)}$ observations, which serves as a parent node and depends on the former splitting steps $1, 2, \ldots, u-1$.
- In particular, when $u = 1$, $\left(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}\right) = (\mathbf{X}, \mathbf{y})$ and $N^{(1)} = N$.

CART - recursive binary splitting

- The dataset in the parent node $\left(\mathbf{X}^{(u)}, \mathbf{y}^{(u)}\right)$ is potentially **split into two daughter nodes** with respective datasets:

$$\left(\mathbf{X}^{(u,l)}, \mathbf{y}^{(u,l)}\right) = \left(\left(\mathbf{X}_1^{(u,l)}, y_1^{(u,l)}\right), \left(\mathbf{X}_2^{(u,l)}, y_2^{(u,l)}\right), \ldots, \left(\mathbf{X}_{N^{(u,l)}}^{(u,l)}, y_{N^{(u,l)}}^{(u,l)}\right)\right)^{\mathrm{T}},$$

$$\left(\mathbf{X}^{(u,r)}, \mathbf{y}^{(u,r)}\right) = \left(\left(\mathbf{X}_1^{(u,r)}, y_1^{(u,r)}\right), \left(\mathbf{X}_2^{(u,r)}, y_2^{(u,r)}\right), \ldots, \left(\mathbf{X}_{N^{(u,r)}}^{(u,r)}, y_{N^{(u,r)}}^{(u,r)}\right)\right)^{\mathrm{T}},$$

- If $X_{j^{(u)}}$ is **continuous**; there exist $j^{(u)} = 1, 2, \ldots, p$ and $s^{(u)} \in \mathcal{X}_{j^{(u)}}$ such that, for any $i = 1, 2, \ldots, N^{(u,l)}$, $X_{ij^{(u)}}^{(u,l)} \leq s^{(u)}$, while, for any $i = 1, 2, \ldots, N^{(u,r)}$; $X_{ij^{(u)}}^{(u,r)} > s^{(u)}$,
- If $X_{j^{(u)}}$ is **categorical**, there exists $s^{(u)} \in \mathcal{P}\left(\mathcal{X}_{j^{(u)}}\right)$ such that, for any $i = 1, 2, \ldots, N^{(u,l)}$, $X_{ij^{(u)}}^{(u,l)} \in s^{(u)}$, while, for any $i = 1, 2, \ldots, N^{(u,r)}$, $X_{ij^{(u)}}^{(u,r)} \notin s^{(u)}$.

CART - recursive binary splitting

- If the explanatory variable $X_{j^{(u)}}$ is continuous, define two regions of the space $\mathcal{X}_{j^{(u)}}$ by

$$R^{(u,l)}\left(j^{(u)}, s^{(u)}\right) = \left\{x_{j^{(u)}} \in \mathcal{X}_{j^{(u)}} : x_{j^{(u)}} \leq s^{(u)}\right\}, \text{ and}$$

$$R^{(u,r)}\left(j^{(u)}, s^{(u)}\right) = \left\{x_{j^{(u)}} \in \mathcal{X}_{j^{(u)}} : x_{j^{(u)}} > s^{(u)}\right\};$$

- if the explanatory variable $X_{j^{(u)}}$ is categorical, define

$$R^{(u,l)}\left(j^{(u)}, s^{(u)}\right) = \left\{x_{j^{(u)}} \in \mathcal{X}_{j^{(u)}} : x_{j^{(u)}} \in s^{(u)}\right\} = s^{(u)}, \text{ and}$$

$$R^{(u,r)}\left(j^{(u)}, s^{(u)}\right) = \left\{x_{j^{(u)}} \in \mathcal{X}_{j^{(u)}} : x_{j^{(u)}} \notin s^{(u)}\right\} = \left(s^{(u)}\right)^{\mathrm{c}}.$$

CART - loss function SSE

- The classical loss function, to determine the optimal parameters for the regression tree, is given by the **sum of squared errors (SSE)**,

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, L\left(\mathbf{y}, \widehat{\mathbf{y}}\right) = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2,$$

  where the set of all feasible vectors of parameters

$$\Theta = \{(R_1, R_2, \ldots, R_M, c_1, c_2, \ldots, c_M) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \times \cdots \times \mathcal{P}(\mathcal{X}) \times \mathcal{Y} \times \mathcal{Y} \times \cdots \times \mathcal{Y} : \cup_{m=1}^{M} R_m = \mathcal{X}, \text{ and } R_{m_1} \cap R_{m_2} = \emptyset \text{ for } m_1 \neq m_2\},$$

  in which $\mathcal{P}(\mathcal{X})$ is the power set of $\mathcal{X}$, i.e., the set of all subsets of $\mathcal{X}$.

CART - optimal parameters under SSE

- The optimal parameters $\hat{j}^{(u)}$ and $\hat{s}^{(u)}$ are given by

$$\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right) = \underset{\substack{j^{(u)}=1,2,\ldots,p; \\ s^{(u)} \in \mathcal{X}_{j^{(u)}} \text{ or } s^{(u)} \in \mathcal{P}\left(\mathcal{X}_{j^{(u)}}\right)}}{\operatorname{argmin}} \sum_{i=1}^{N^{(u,l)}} \left(y_i^{(u,l)} - \hat{c}^{(u,l)}\left(j^{(u)}, s^{(u)}\right)\right)^2 + \sum_{i=1}^{N^{(u,r)}} \left(y_i^{(u,r)} - \hat{c}^{(u,r)}\left(j^{(u)}, s^{(u)}\right)\right)^2,$$

which is also known as the **ANOVA best split**.

CART - predicted values under SSE

- The optimal parameters $\hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)$ and $\hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)$ (the predicted values at two daughter nodes) are given by:

$$\hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right) = \frac{1}{N^{(u,l)}} \sum_{i: X^{(u)}_{i\hat{j}^{(u)}} \in R^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)} y_i^{(u)} = \frac{1}{N^{(u,l)}} \sum_{i=1}^{N^{(u,l)}} y_i^{(u,l)},$$

$$\hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right) = \frac{1}{N^{(u,r)}} \sum_{i: X^{(u)}_{i\hat{j}^{(u)}} \in R^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)} y_i^{(u)} = \frac{1}{N^{(u,r)}} \sum_{i=1}^{N^{(u,r)}} y_i^{(u,r)}.$$

Actuarial modified loss function - WSSE

- Modify the classical SSE loss function by **assigning different weights** to the squared errors for observations with **zero** in the response variable and those with **non-zero** in the response variable.
- We define the following loss function, which is given by the **weighted sum of squared errors (WSSE)**:

$$
\begin{aligned}
L_{\mathrm{W}}\left(\mathbf{y}^{(u)}, \widehat{\mathbf{y}}^{(u)}\right) &= w_0^{(u)} \sum_{i:y_i^{(u)}=0} \left(y_i^{(u)} - \hat{y}_i^{(u)}\right)^2 + \left(1 - w_0^{(u)}\right) \sum_{i:y_i^{(u)}\neq 0} \left(y_i^{(u)} - \hat{y}_i^{(u)}\right)^2 \\
&= w_0^{(u)} \sum_{i:y_i^{(u)}=0} \left(\hat{y}_i^{(u)}\right)^2 + \left(1 - w_0^{(u)}\right) \sum_{i:y_i^{(u)}\neq 0} \left(y_i^{(u)} - \hat{y}_i^{(u)}\right)^2,
\end{aligned}
$$

where $u = 1, 2, \ldots$ is a binary splitting step.

- In particular, when $w_0^{(u)} = 0.5$, the WSSE loss function is reduced to the classical SSE function.

Optimal split under WSSE

- Then optimal parameters $\hat{j}^{(u)}$ and $\hat{s}^{(u)}$ are given by

$$\begin{aligned}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right) = \underset{\substack{j^{(u)}=1,2,\ldots,p; \\ s^{(u)} \in \mathcal{X}_{j^{(u)}} \text{ or } s^{(u)} \in \mathcal{P}\left(\mathcal{X}_{j^{(u)}}\right)}}{\operatorname{argmin}} & w_0^{(u)} \sum_{i: y_i^{(u,l)}=0} \hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)^2 \\ & + \left(1 - w_0^{(u)}\right) \sum_{i: y_i^{(u,l)} \neq 0} \left(y_i^{(u,l)} - \hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)\right)^2 \\ & + w_0^{(u)} \sum_{i: y_i^{(u,r)}=0} \hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)^2 \\ & + \left(1 - w_0^{(u)}\right) \sum_{i: y_i^{(u,r)} \neq 0} \left(y_i^{(u,r)} - \hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)\right)^2.\end{aligned}$$

Predicted values under WSSE

- The optimal parameters $\hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)$ and $\hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)$ (the predicted values at two daughter nodes) are given by:

$$\hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right) = \frac{\left(1 - w_0^{(u)}\right) \sum_{i:y_i^{(u,l)} \neq 0} y_i^{(u,l)}}{w_0^{(u)} N_0^{(u,l)} + \left(1 - w_0^{(u)}\right)\left(N^{(u,l)} - N_0^{(u,l)}\right)},$$

$$\hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right) = \frac{\left(1 - w_0^{(u)}\right) \sum_{i:y_i^{(u,r)} \neq 0} y_i^{(u,r)}}{w_0^{(u)} N_0^{(u,r)} + \left(1 - w_0^{(u)}\right)\left(N^{(u,r)} - N_0^{(u,r)}\right)},$$

where $N_0^{(u,l)} = \left|\left\{i = 1, \ldots, N^{(u,l)} : y_i^{(u,l)} = 0\right\}\right|$, $N_0^{(u,r)} = \left|\left\{i = 1, \ldots, N^{(u,r)} : y_i^{(u,r)} = 0\right\}\right|$, representing the number of observations with zero response in the daughter nodes.

Canberra distance

- **Canberra distance**, introduced by Lance and Williams (1966) for similarity analysis, is defined by, for any two real numbers p and q,

$$d_{\mathrm{CAN}}(p,q) = \begin{cases} 0 & \text{if } p = q = 0, \\ \frac{|p-q|}{|p|+|q|} & \text{otherwise.} \end{cases}$$

- Canberra distance is essentially the Euclidean distance being normalized by **the magnitude of the two real numbers** in the denominator.
- The Canberra distance is often used for **data scattered around the origin**, as it is **a biased measure** and is very sensitive to values close to zero. For example, $d_{\mathrm{CAN}}(0,1) = 1$ and $d_{\mathrm{CAN}}(100,101) \approx 0.005$.

(Observed, Predicted)	Squared error	Canberra	Squared Canberra
(0, 1)	$(0-1)^2 = 1$	$\frac{\lvert 0-1\rvert}{\lvert 0\rvert+\lvert 1\rvert} = 1$	$\frac{(0-1)^2}{0^2+1^2} = 1$
(100, 101)	$(100-101)^2 = 1$	$\frac{\lvert 100-101\rvert}{\lvert 100\rvert+\lvert 101\rvert} \approx 0.005$	$\frac{(100-101)^2}{100^2+101^2} \approx 0.00005$

Actuarial modified loss function - SSCE

- To be in line with the order of errors in the SSE and the WSSE, which is of squared, also define the **squared Canberra distance** by, for any two real numbers p and q,

$$d_{\mathrm{SCAN}}(p,q) = \begin{cases} 0 & \text{if } p = q = 0, \\ \frac{(p-q)^2}{p^2+q^2} & \text{otherwise.} \end{cases}$$

- We define the following loss function which is given by the **sum of squared Canberra errors (SSCE)**:

$$L_{\mathrm{C}}\left(\mathbf{y}^{(u)}, \widehat{\mathbf{y}}^{(u)}\right) = \sum_{i=1}^{N^{(u)}} d_{\mathrm{SCAN}}\left(y_i^{(u)}, \hat{y}_i^{(u)}\right),$$

where $u = 1, 2, \ldots$ is a binary splitting step.

Optimal split under SSCE

- Minimize the SSCE,

$$\left(\hat{j}^{(u)}, \hat{s}^{(u)}, \hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right), \hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)\right)$$

$$= \underset{\substack{j^{(u)}=1,2,\ldots,p; \\ s^{(u)}\in\mathcal{X}_{j^{(u)}} \text{ or } s^{(u)}\in\mathcal{P}\left(\mathcal{X}_{j^{(u)}}\right); \\ c^{(u,l)}\left(j^{(u)},s^{(u)}\right)\in\mathbb{R}_+; \\ c^{(u,r)}\left(j^{(u)},s^{(u)}\right)\in\mathbb{R}_+}}{\operatorname{argmin}} L_{\mathrm{C}}\left(\mathbf{y}^{(u)}, \widehat{\mathbf{y}}^{(u)}\right) = \underset{\substack{j^{(u)}=1,2,\ldots,p; \\ s^{(u)}\in\mathcal{X}_{j^{(u)}} \text{ or } s^{(u)}\in\mathcal{P}\left(\mathcal{X}_{j^{(u)}}\right); \\ c^{(u,l)}\left(j^{(u)},s^{(u)}\right)\in\mathbb{R}_+; \\ c^{(u,r)}\left(j^{(u)},s^{(u)}\right)\in\mathbb{R}_+}}{\operatorname{argmin}} \sum_{i=1}^{N^{(u)}} d_{\mathrm{SCAN}}\left(y_i^{(u)}, \hat{y}_i^{(u)}\right)$$

$$= \underset{\substack{j^{(u)}=1,2,\ldots,p; \\ s^{(u)}\in\mathcal{X}_{j^{(u)}} \text{ or } s^{(u)}\in\mathcal{P}\left(\mathcal{X}_{j^{(u)}}\right); \\ c^{(u,l)}\left(j^{(u)},s^{(u)}\right)\in\mathbb{R}_+; \\ c^{(u,r)}\left(j^{(u)},s^{(u)}\right)\in\mathbb{R}_+}}{\operatorname{argmin}} \sum_{i=1}^{N^{(u,l)}} d_{\mathrm{SCAN}}\left(y_i^{(u,l)}, c^{(u,l)}\left(j^{(u)}, s^{(u)}\right)\right) + \sum_{i=1}^{N^{(u,r)}} d_{\mathrm{SCAN}}\left(y_i^{(u,r)}, c^{(u,r)}\left(j^{(u)}, s^{(u)}\right)\right).$$

Properties of SSCE

- Although the optimization problem cannot be solved explicitly, the existence of the solution can be proved.
- **Lemma 1**: If $\hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)$ and $\hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)$ exist, then

$$y_{(1)}^{(u,l)} \leq \hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right) \leq y_{\left(N^{(u,l)}\right)}^{(u,l)},$$

$$y_{(1)}^{(u,r)} \leq \hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right) \leq y_{\left(N^{(u,r)}\right)}^{(u,r)},$$

where $y_{(1)}^{(u,l)}$ and $y_{(1)}^{(u,r)}$ are the smallest response values in the respective daughter nodes, while $y_{\left(N^{(u,l)}\right)}^{(u,l)}$ and $y_{\left(N^{(u,r)}\right)}^{(u,r)}$ are the largest response values in the respective daughter nodes.

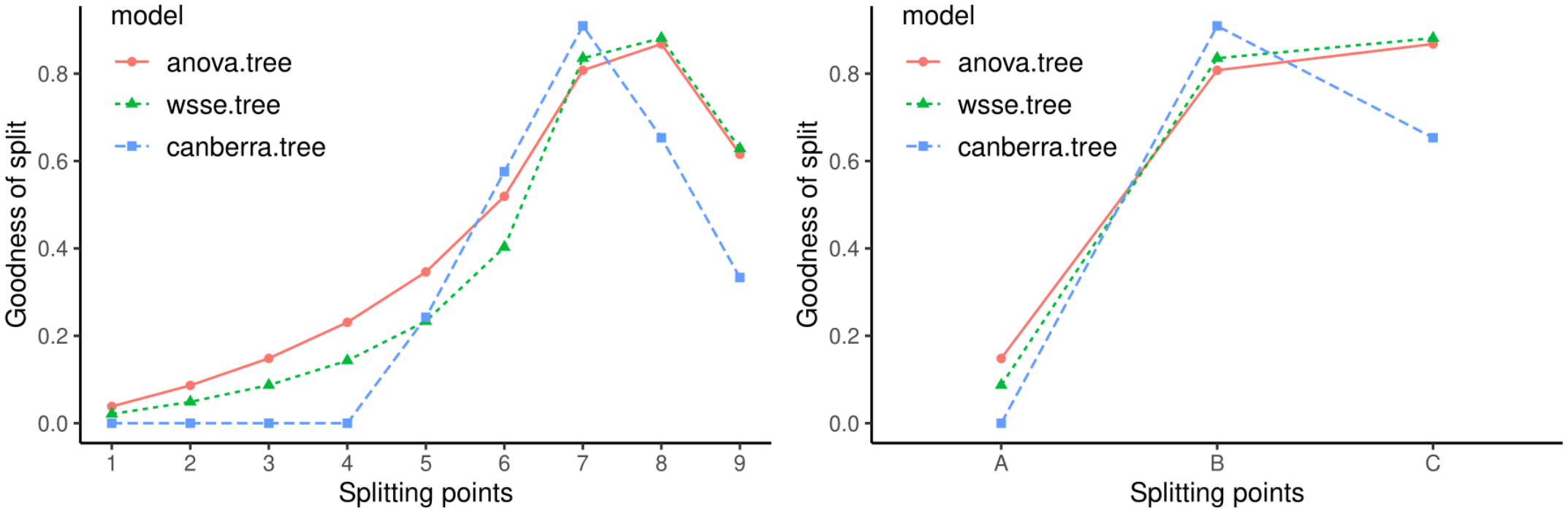
Properties of SSCE

- **Proposition 1**: The predicted values $\hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)$ and $\hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)$ exist.
- **Proposition 2**: If $N_0^{(u,l)} > \frac{1}{2}N^{(u,l)}$ (resp. $N_0^{(u,r)} > \frac{1}{2}N^{(u,r)}$), then $\hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)$ (resp. $\hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)$) is unique and must be 0; moreover, $\sum_{i=1}^{N^{(u,l)}} d_{\mathrm{SCAN}}\left(y_i^{(u,l)}, \hat{c}^{(u,l)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)\right) = N^{(u,l)} - N_0^{(u,l)}$ (resp. $\sum_{i=1}^{N^{(u,r)}} d_{\mathrm{SCAN}}\left(y_i^{(u,r)}, \hat{c}^{(u,r)}\left(\hat{j}^{(u)}, \hat{s}^{(u)}\right)\right) = N^{(u,r)} - N_0^{(u,r)}$).

Practical implementation

- We refer to the tree-based model using the WSSE loss function as the **WSSE tree model**, and the model using the SSCE loss function as the **Canberra tree model**.
- To modify the classical CART algorithm with the two proposed loss functions, we employ the rpart function in the **R package rpart** (Therneau and Atkinson, 1997).
- The package provides a **user splits option** (Therneau, 2019), which provides a way to **extend rpart and validate new methodologies**.

Results on the motivating example



- When the data contains a large proportion of zero responses, WSSE trees and Canberra trees provide **better splitting performance** than ANOVA trees.
- The Canberra tree can effectively **separate zeros and non-zeros** as expected.

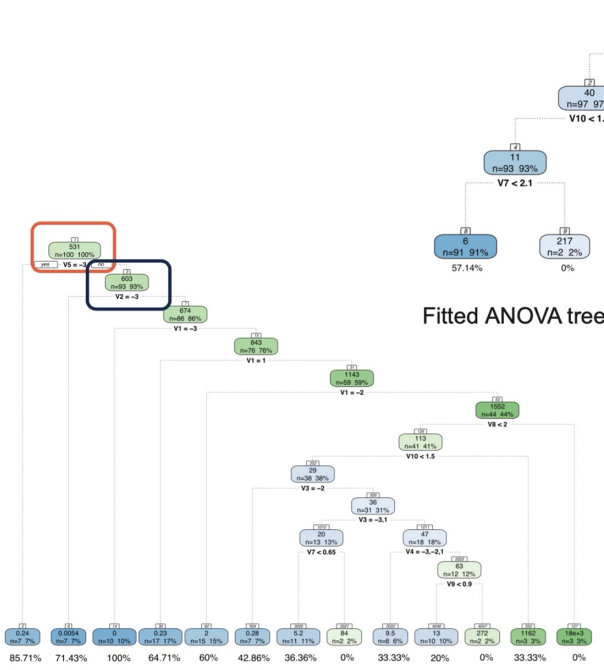
Simulation study - data generation

- To **mimic the real-life insurance datasets**, we generate the synthetic training and test datasets, with **53%** of the observations in the training dataset and **63%** in the test dataset have a zero response.
- Simulation design:
  - Explanatory variables: $\mathbf{X} = (\mathbf{X}_{categorical}, \mathbf{X}_{continuous})$.
  - $\mathbf{X}_{continuous} \sim \boldsymbol{N}_p(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{ij} = Cov(\mathbf{X}_i, \mathbf{X}_j) = (0.8)^{i-j}$. $N = 100$, $p = 10$.
  - $\mathbf{X}_{categorical}$, random sampling from the set of integers $(-3, -2, 1, 4)$, with respective probabilities of $(0.1, 0.2, 0.2, 0.5)$.
  - Linear coefficients: $\boldsymbol{\beta} = (-0.1, \underbrace{1.0, 1.0}_{\text{2 cat}}, \underbrace{0.5, 0.5}_{\text{2 cat}}, \underbrace{0}_{\text{1 cat}}, \underbrace{1.0, 1.0}_{\text{2 con}}, \underbrace{0.5, 0.5}_{\text{2 con}}, \underbrace{0}_{\text{1 con}})^T$
  - Response variable: $\mathbf{Y}$, generated from a **Tweedie GLM** framework,

$$y_i \sim Tweedie(\mu_i, \phi, \xi),$$

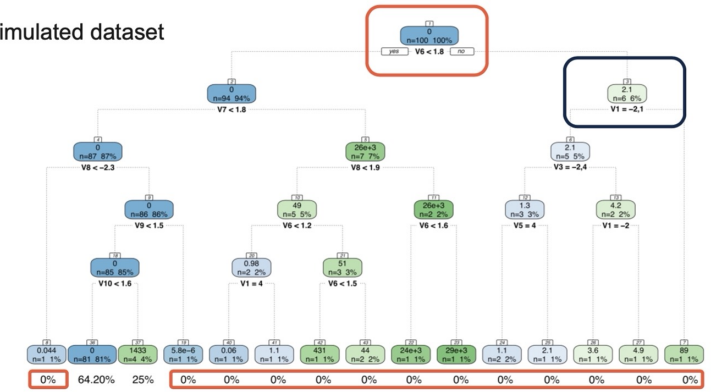
with the log link function $g(\mu_i) = \log(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}$, the dispersion parameter $\phi = 2$, and the variance power parameter $\xi = 1.7$.

Result on the simulated dataset - fitted trees



Fitted WSSE tree on simulated dataset

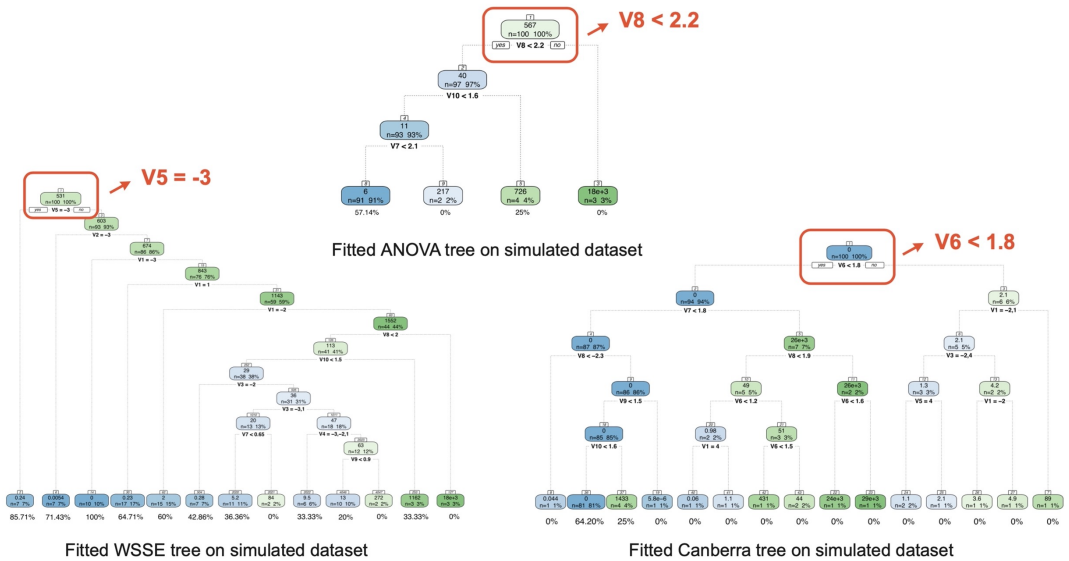
Fitted ANOVA tree on simulated dataset



Fitted Canberra tree on simulated dataset

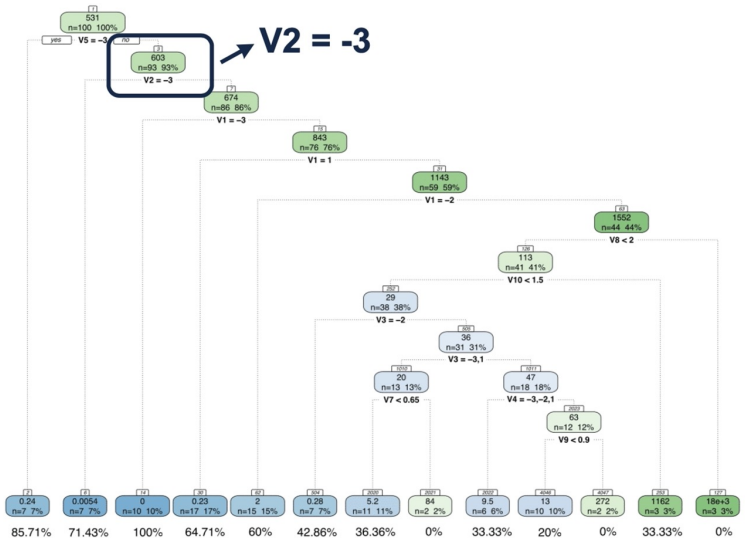
- The **overall structures** of the WSSE tree and Canberra tree models are quite **different** from that of the ANOVA tree.

Result on the simulated dataset - fitted trees

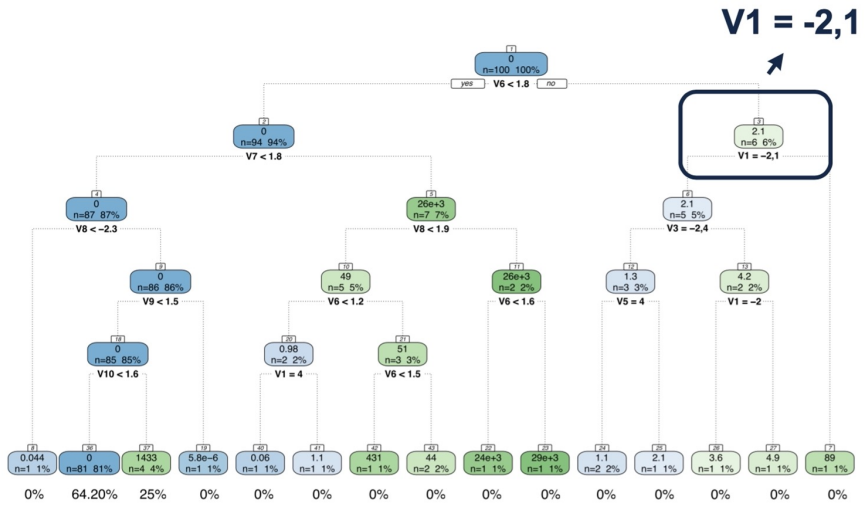


- Both the WSSE tree and Canberra tree models reveal apparent structural changes at **node 1 (the root node)**, which is the most critical split.
- In the Canberra tree model, the first split is initiated by the variable **V6**, whereas in the ANOVA tree model and the WSSE tree model, the root nodes are split by the variable **V8** and **V5** respectively.
- **V6** is **strongly correlated** with the response variable, while **V5** and **V8** are **noisy** variable or **weakly correlated** with the response variable.
- The Canberra tree model is **more effective in finding the correct explanatory variable to split** under the imbalanced problem presented in the dataset.

Results on the simulated dataset - fitted trees



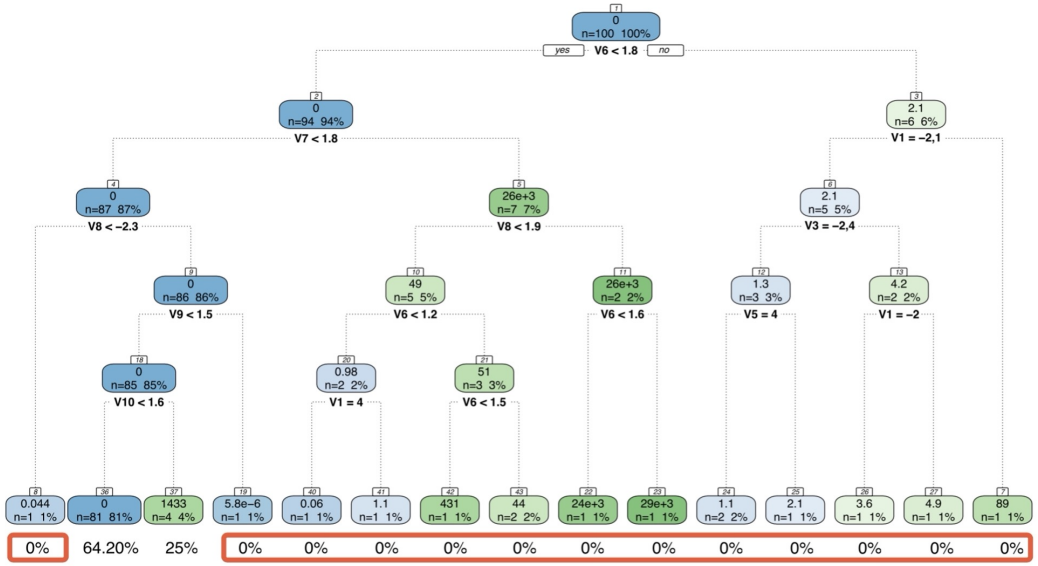
Fitted WSSE tree on simulated dataset



Fitted Canberra tree on simulated dataset

- The first few splits in the ANOVA tree are mainly determined by the **continuous** variables, while **categorical** variables, such as **V1** and **V2**, are taken into account in the WSSE tree and Canberra tree models.
- For instance, the **node 2 in the WSSE tree** is divided by the categorical variable **V2**, and the **node 3 in the Canberra tree model** is divided by the categorical variable **V1**.

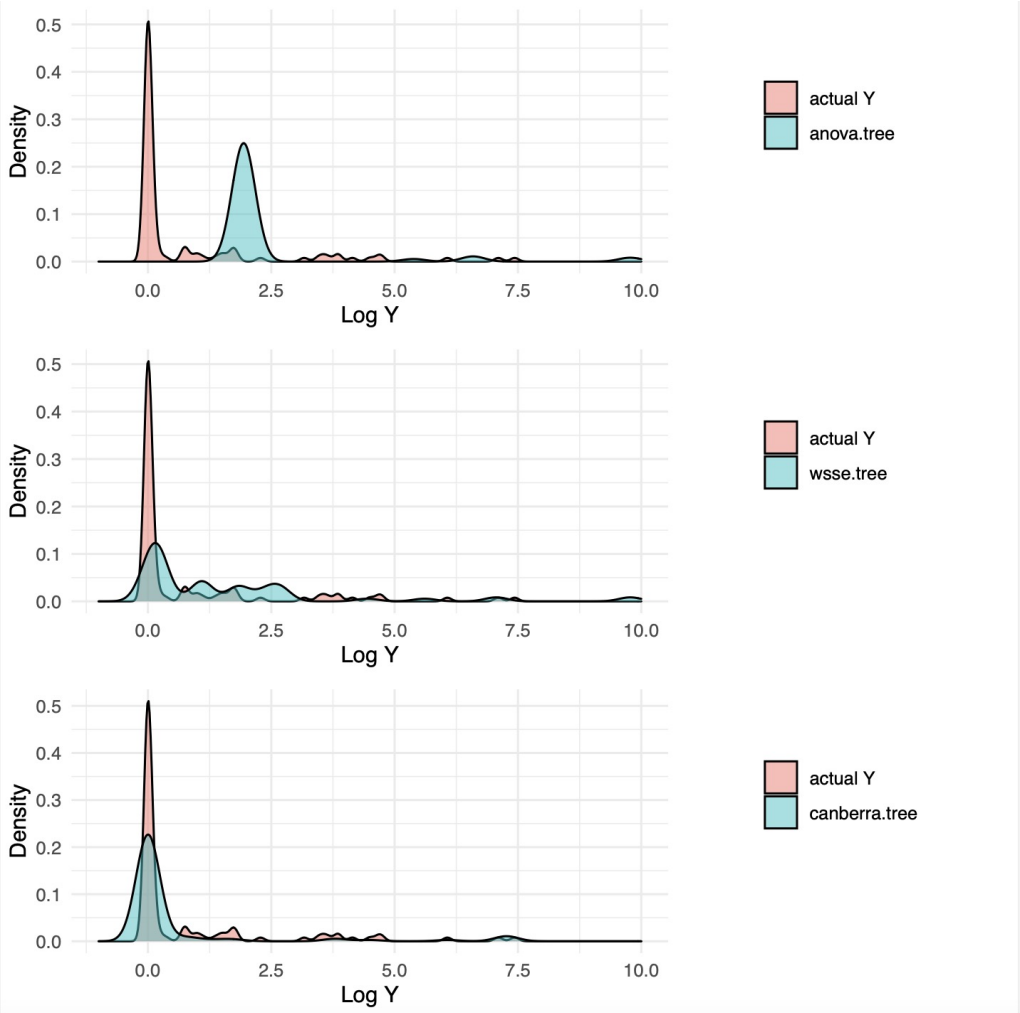
Results on the simulated dataset - fitted trees



Fitted Canberra tree on simulated dataset

- We can see from the percentages of zeros on the terminal nodes, the Canberra tree model is **more likely to have 0% or 100% zero claims** than the other two tree models, which indicates the Canberra tree model outperforms the other two models at **separating the zero and non-zero claims**.

Results on the simulated dataset - density plots



- The Canberra tree model is superior to the other two models in predicting the **zero responses**.
- The response values predicted by the ANOVA tree are **centered on a relatively small positive value**; on the other hand, the Canberra tree model is able to **identify zero claims** precisely.

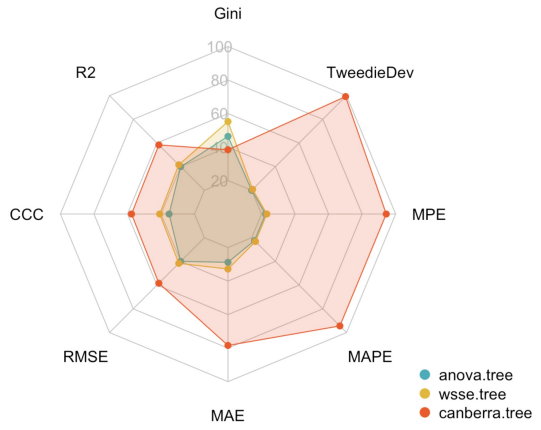
Results on the simulated dataset - heatmap

Heatmap of model performance based on training dataset

Heatmap of model performance based on test dataset

- In general, the Canberra tree model has the **best overall prediction performance**.
- Specifically, the Canberra tree model performs much better in **MAPE** and **MPE**, indicating that the Canberra tree model has a good fit for the observations with the **non-zero response**.

Robustness of simulation study - radar plot



- To ensure robustness, we re-examine these performances to each of **100 synthetic datasets**, which are still based on the same design in the simulation study, but are generated by different random seeds.
- The radar plot summarizes these records by displaying **the number of datasets in which each model performs the best under each measure**.
- The WSSE tree model is **slightly superior** to the ANOVA tree model.
- The Canberra tree model **substantially outmatches** these two tree models under all measures, except under the Gini index being exceeded by the ANOVA and WSSE tree models.

Concluding remarks

- **Motivation**:
  - The default CART is not sufficient to handle insurance datasets that contain a large number of zeros.
- **Modification**:
  - We propose two actuarial modified loss functions, namely the **weighted sum of squared error (WSSE)** and the **sum of squared Canberra error (SSCE)** loss functions, as the node impurity function under the CART framework
- **Results**:
  - The motivating and experimental examples demonstrate that the WSSE tree and Canberra tree models are **more effective in separating observations of zero responses from non-zero responses** compared to the default ANOVA method.
  - The simulation study suggests that the Canberra regression tree model **offers the best overall prediction performance**, especially when it comes to the observations with zero response.

Selected references

- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Taylor & Francis Group, LLC: Boca Raton, FL.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321–357.
- Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In Inter-national Conference on Machine Learning, volume 96, pages 148–156.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications.Expert Systems with Applications, 73:220–239.
- He, H. and Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9):1263–1284.

Selected references

- Lance, G. N. and Williams, W. T. (1966). Computer programs for hierarchical polythetic classification (“similarity analyses”).The Computer Journal, 9(1):60–64.
- Therneau, T. (2019). User written splitting functions for RPART. Technical report, MayoClinic.
- Therneau, T. M. and Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines. Technical report, Mayo Foundation.
- Yang, Y., Qian, W., and Zou, H. (2018). Insurance premium prediction via gradient tree-boosted Tweedie compound poisson models. Journal of Business & Economic Statistics, 36(3):456–470.

Q & A

Thank you for your attention!