Zero-Inflated Tweedie Boosted Trees with `CatBoost` for Insurance Loss Analytics

joint work with Banghee So, Towson University

JOCO2024 (The First Joint Colloquium of All IAA Sections)
Brussels, Belgium
23-25 September 2024

Emiliano A. Valdez, PhD, FSA
Department of Mathematics
University of Connecticut
Storrs, Connecticut

Introduction

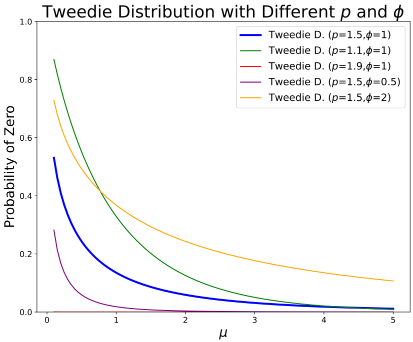
- The two-part frequency-severity model has historically been the norm.
- Since Tweedie et al. (1984), the Tweedie distribution has gained popularity as it eliminates need for separate frequency and severity models.
- Tweedie models, denoted as $\mathrm{Tw}(\mu, \phi, p)$, are defined by the following density function:

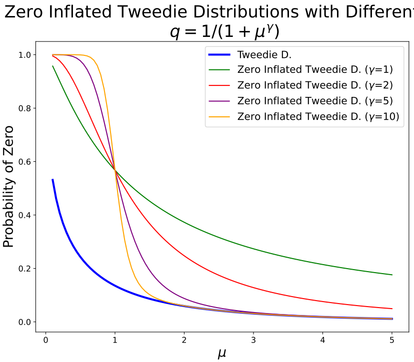
$$f_{\mathrm{Tw}}(y|\mu, \phi, p) = a(y, \phi, p) \exp\left(\frac{1}{\phi}\left(y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right)\right), \quad y \geq 0,$$

  where $a(\cdot)$ is normalizing function, $\mu > 0$ is the expected value of Y, and $\phi > 0$ represents the dispersion parameter.
- $\mathrm{Var}(Y) = \phi\mu^p$ so that p controls the relationship between variance and mean.
- We restrict the power p to $1 < p < 2$, the case of the compound Poisson-gamma model.
- When introducing predictor variables, we can consider using suitable link function.

Zero-inflation

- Tweedie distribution is largely flexible and is able to model a wide range of data including those with excess zeros (zero-inflation), right-skewness, and heavy tails, but ...





Zero-inflated Tweedie (ZITw) Distribution Model

- The ZITw model combines a point mass at zero, to help improve the accuracy of estimating μ especially when dealing with excessive zeros.
- The density function of the ZITw model can be formulated as follows:

$$
f_{\text{ZITw}}(y|\mu,\phi,p,q) = \begin{cases} q + (1-q)\cdot \exp\left(-\dfrac{1}{\phi}\dfrac{\mu^{2-p}}{2-p}\right), & \text{if } y = 0 \\[2ex] (1-q)\cdot a(y,\phi,p)\cdot \exp\left(\dfrac{1}{\phi}\left(y\cdot\dfrac{\mu^{1-p}}{1-p} - \dfrac{\mu^{2-p}}{2-p}\right)\right), & \text{if } y > 0. \end{cases}
$$

- q represents the inflation probability, indicating the degree of zero inflation.
- The expected value of Y under the ZITw model is given by $(1-q)\mu$. Thus, accurately estimating both μ and q is crucial.
- The gradient boosting framework offers techniques to achieve this effectively.

Emiliano A. Valdez, PhD, FSA

Gradient Boosting

- Gradient boosting is an ensemble technique based on concept of building a strong predictive model by combining the predictions of multiple weak learners. Friedman (2001).
- When decision trees are used as weak learners, they are called Gradient Boosted Decision Trees (GBDT).
- Given training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_1^n$, gradient boosting generates a sequence of functions $W_0, W_1, \cdots, W_T$, by minimizing the exp. value of a specified loss function, $\ell(y_i, W_t)$.
    - Each iteration trains a new weak learner to correct ensemble errors.
    - At each iteration, calculate the negative gradient (pseudo-residuals) of the loss function, which gives direction of steepest descent to minimize loss.
    - Newly trained weak learner is fitted to the pseudo-residuals; it learns to predict the errors made by the current ensemble.
    - Update ensemble by adding output to the current ensemble, with a learning rate.
    - Process is repeated for a fixed number of iterations.
    - Final prediction is the cumulative result of all weak learners combined.

Zero-inflated Tweedie Boosted Decision Trees

- We use the negative log-likelihood of the data based on the zero-inflated Tweedie distribution model.
- We use decision trees as weak learners.
- The boosted tree model assumes the logarithm of the exp. value of target variable Y, given set of features $\boldsymbol{x}$, can be effectively modeled as follows:

$$\ln \mathbb{E}(Y \mid \boldsymbol{x}) = \ln E + W_T(\boldsymbol{x}),$$

  where $\ln E$ is the offset term and $W_T(\boldsymbol{x})$ denotes the prediction score produced as:

$$W_T(\boldsymbol{x}) = w_1(\boldsymbol{x}) + w_2(\boldsymbol{x}) + \cdots + w_t(\boldsymbol{x}) + \cdots + w_T(\boldsymbol{x}).$$

- Here, $w_t(\boldsymbol{x})$ represents the prediction of the t-th tree in the gradient boosting model.
- This framework allows for a flexible and powerful modeling of complex relationships between features and target variable.

Categorical Boosting (CatBoost)

- Notable software libraries for GBDT implementation include XGBoost, LightGBM, and CatBoost.
- Increasing in popularity, CatBoost, developed by Yandex (Prokhorenkova et al., 2018), is recognized for its effectiveness in handling heterogeneous datasets, a common scenario in insurance data.
- It employs a technique known as “Ordered Target Statistic” in encoding categorical features as numerical features.
  - Replace each category with the average value of the target variable for that category.
- Additional advantages include: producing high predictive accuracy, offering scalability for large data sets, and supporting the generation of interpretative graphs that help in further understanding and explaining model results.
- Recent studies (So, 2024) have demonstrated CatBoost’s superior performance compared to its counterparts when processing insurance data.

Methodology

- In conventional zero-inflated models, training is usually conducted separately for the mean μ and the inflation probability q.
- This approach requires twice as many trees for zero-inflated Tweedie (ZITw) boosted trees compared to Tweedie (Tw) models, due to independent parameter estimation for each:

$$\ln \mu = \ln E + W_T^{mean}(\boldsymbol{x}),$$

$$\text{logit}(q) = \ln \frac{q}{1-q} = W_T^{prob}(\boldsymbol{x}).$$

Two possible approaches

- **Scenario 1** Functionally unrelated: q is not directly functionally related to μ
  - Train $W_T^{mean}(\boldsymbol{x})$ and $W_T^{prob}(\boldsymbol{x})$ separately.
- **Scenario 2** Functionally related: q is functionally linked to μ
  - Our proposed parameterization is depicted by the following equations:

$$\ln \mu = \ln E + W_T(\boldsymbol{x}),$$

$$\text{logit}(q) = \ln \frac{q}{1-q} = -\gamma(\ln E + W_T(\boldsymbol{x})).$$

This leads us to $q = \frac{1}{1+\mu^{\gamma}}$.

Adjustment of Compositional Data

- Compositional data is characterized by multiple non-negative features that sum up to a constant, typically 100% or 1.
- Due to the inherent statistical dependence among these features, transformations are often necessary to map the data onto the real Euclidean space.
- This transformation facilitates the application of traditional statistical methodologies.
- When dealing with compositional data comprising J features, denoted as $\{\boldsymbol{x}_{\cdot 1}, \boldsymbol{x}_{\cdot 2}, \ldots, \boldsymbol{x}_{\cdot J}\}$, where the features sum to 1, we refer to these features as a J-part composition.
- See Aitchison (1994) and Verbelen, Antonio, and Claeskens (2018).

Logratio transformations

- Notable transformations are the logratio methods, which include:
  - centered logratio transformation (CLR):

$$\mathrm{CLR}(j) = \ln\left(\frac{\boldsymbol{x}_{\cdot j}}{\left(\prod_i \boldsymbol{x}_{\cdot i}\right)^{1/J}}\right), \quad j = 1, 2, \ldots, J.$$

  - additive logratio transformation (ALR):

$$\mathrm{ALR}(j|d) = \ln\left(\frac{\boldsymbol{x}_{\cdot j}}{\boldsymbol{x}_{\cdot d}}\right), \quad j \neq d.$$

  - isometric logratio transformation (ILR):

$$\mathrm{ILR}(\boldsymbol{x}) = R \cdot \mathrm{CLR}(\boldsymbol{x}),$$

    where $\boldsymbol{x}$ is a $J \times n$ data matrix comprising J features, and R is a $(J-1) \times J$ matrix satisfying the condition: $RR^T = I_{J-1}$.

Models Compared

Our empirical analysis is based on a synthetic telematics dataset developed by So, Boucher, and Valdez (2021).
This dataset comprises of 100,000 policies and demonstrates a zero-inflation characteristic, with only 2,698 policies experiencing at least one claim.
For this study, a total of eight different models were trained:

1. Zero-inflated Tweedie boosted tree with scenario 1 (ZITwBT1)
2. Zero-inflated Tweedie boosted tree with scenario 2 (ZITwBT2)
3. Tweedie boosted tree (TwBT)
4. Tweedie GLM (TwGLM)
5. ALR
6. CLR
7. ILR
8. PPCA after CLR transformation

Performance Metrics

- Deviance: measures how well the predicted outcomes in a model match the observed outcomes. Lower deviance indicates better fit.
- Mean Absolute Deviation: quantifies the average absolute difference between the observed and predicted values, defined as $\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$. A lower MAD suggests higher precision.
- Vuong Test: compares likelihood functions of non-nested models. See Vuong (1989).
- Gini Index: assesses model prediction performance. Gini^a and Gini^b are two variants.

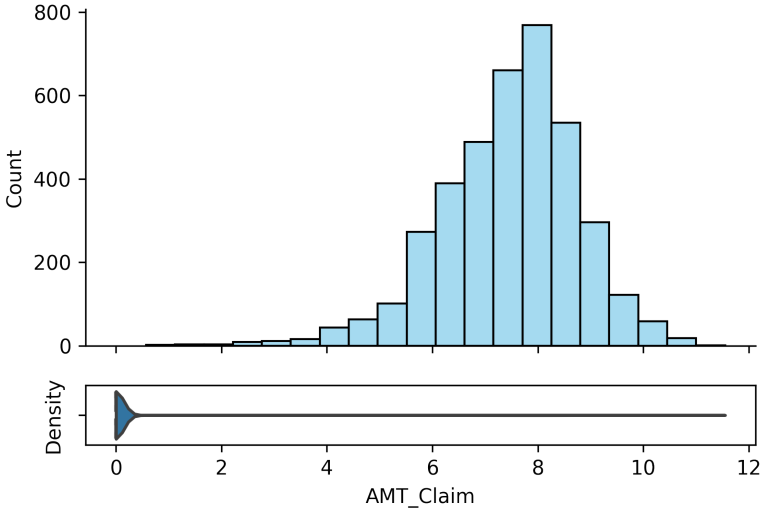
Descriptive Details of Dataset

Table 1: Variable Names and Descriptions for the Synthetic Telematics Dataset

Type	Variable	Description
Traditional	Duration	Total exposure in yearly units
	Insured.age	Age of insured driver
	Insured.sex †	Sex of insured driver: Male, Female
	Car.age	Age of vehicle (in years)
	Marital †	Marital status: Single, Married
	Car.use †	Use of vehicle: Private, Commute, Farmer, Commercial
	Credit.score	Credit score of insured driver
	Region †	Type of region where driver lives: Rural, Urban
	Annual.miles.drive	Annual miles expected to be driven declared by driver
	Years.noclaims	Number of years without any claims
	Territory †	Territorial location of vehicle: 55 labels in {11, 12, 13, . . ., 91}
Telematics	Annual.pct.driven	Annualized percentage of time on the road
	Total.miles.driven	Total distance driven in miles
	Pct.drive.xxx	Percent of driving day xxx of the week: mon/tue/. . . /sun
	Pct.drive.x hrs	Percent vehicle driven within x hrs: 2hrs/3hrs/4hrs
	Pct.drive.xxx	Percent vehicle driven during xxx: wkday/wkend
	Pct.drive.rush xx	Percent of driving during xx rush hours: am/pm
	Avgdays.week	Mean number of days used per week
	Accel.xxmiles	Number of sudden acceleration 6/8/9/. . . /14 mph/s per 1000miles
	Brake.xxmiles	Number of sudden brakes 6/8/9/. . . /14 mph/s per 1000miles
	Left.turn.intensityxx	Number of left turn per 1000miles with intensity xx: 08/09/10/11/12
	Right.turn.intensityxx	Number of right turn per 1000miles with intensity xx: 08/09/10/11/12
Response	NB_Claim	Number of claims on the given policy
	AMT_Claim	Amount of claims on the given policy

† Indicates categorical variable.

Distribution of Aggregate Claim Amounts



Model Goodness of Fit

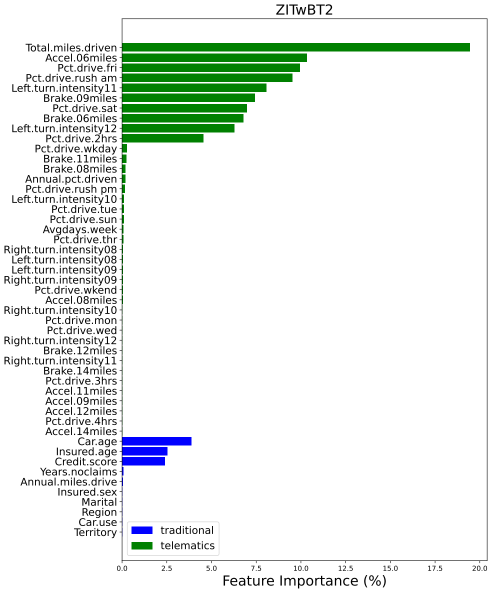
Table 2: Gini^b across 4 Models

Base Model \ Competing Model	TwGLM	TwBT	ZITwBT1	ZITwBT2
TwGLM	-	0.489	0.120	0.504
TwBT	-0.043	-	-0.275	0.266
ZITwBT1	0.695	0.598	-	0.704
ZITwBT2	0.127	0.035	-0.105	-

Table 3: Gini^b in ZITwBT2 Models with and without Compositional Data Adjustment

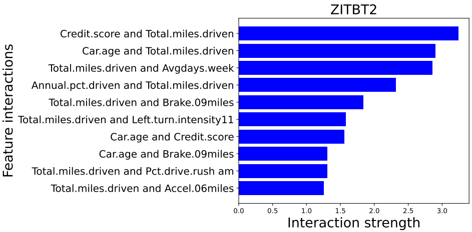
Base Model \ Competing Model		ZITwBT2	ZITwBT2			
			alr	clr	ilr	clr+PPCA
ZITwBT2		-	0.128	0.122	0.124	0.011
ZITwBT2	alr	0.160	-	0.145	0.119	0.033
	clr	0.762	0.760	-	0.755	0.735
	ilr	0.152	0.167	0.138	-	0.059
	clr+PPCA	0.335	0.349	0.342	0.304	-

Feature importance in ZITwBT2 CatBoost

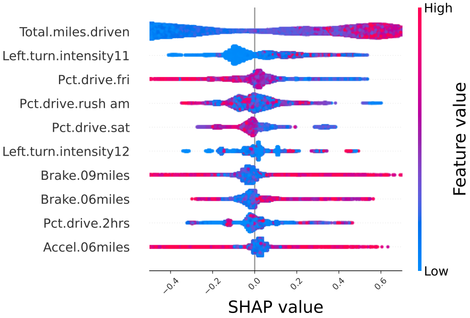


- More telematics than traditional variables are better predictors of aggregate claims.
- `Total.miles.driven` far outweigh all other variables.
- Driving maneuvers appear to be important predictors of aggregate claims.

Features Interaction and SHAP Values in ZITwBT2 `CatBoost`

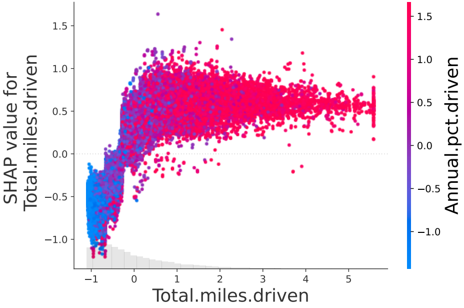


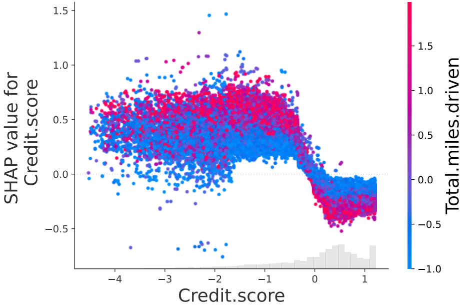
Top 10 Features Interaction Strength



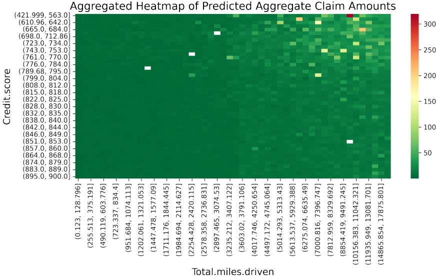
SHAP Values of Top Important Features

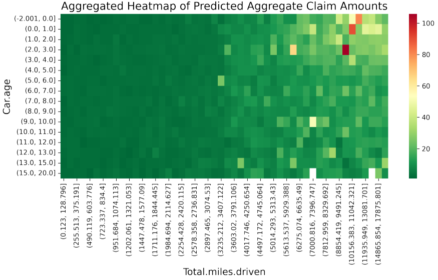
Selected Feature Interaction Through SHAP Values





Heatmaps Describing Feature Interaction with Aggregate Claim Amount





Concluding remarks

- In this paper, we applied a zero-inflated Tweedie loss function in gradient boosting with various adjustments.
  - We reparameterized the zero-inflated Tweedie loss function to express the inflation probability q as a function of μ.
  - This reparameterization led to a unified model, maximizing the use of `CatBoost` libraries.
  - This approach improves interpretation and enables better model comparison through various performance metrics.
- Our research makes significant contributions to actuarial studies as a result of:
  - Simplified interpretation and efficiency
  - Robustness to compositional data. Our model shows robustness without needing extra adjustments for compositional data, indicating superiority over GLMs.
  - Advantages of using `CatBoost` libraries.

- Thank you -