2024 Travelers University Modeling Competition

"The Traveling Salesmen"

The University of Iowa

Behrooz Khalil Loo, Nathan Munshower,

David Roth, Mahdi Saedei Kousha and Nikita Jiaswal

Outline

- What is the “Traveler’s Competition?”
- Understanding the Problem & Dataset
- Selected Models & Explanations
- Results & Conclusions

Traveler’s Analytics Case Competition

- Hosted by **Travelers**—a massive P&C U.S. insurance company
- Given a business problem where predictive analytics are needed
- 22 teams entered—we represented UIowa ☺

What's the Business Problem?

- We work at CloverShield—an insurance company with a thriving call center ☺

What's the Business Problem?

- We work at CloverShield—an insurance company with a thriving call center ☺
- **Goal: Forecast number of times *a policyholder* will call** (to reduce costs)

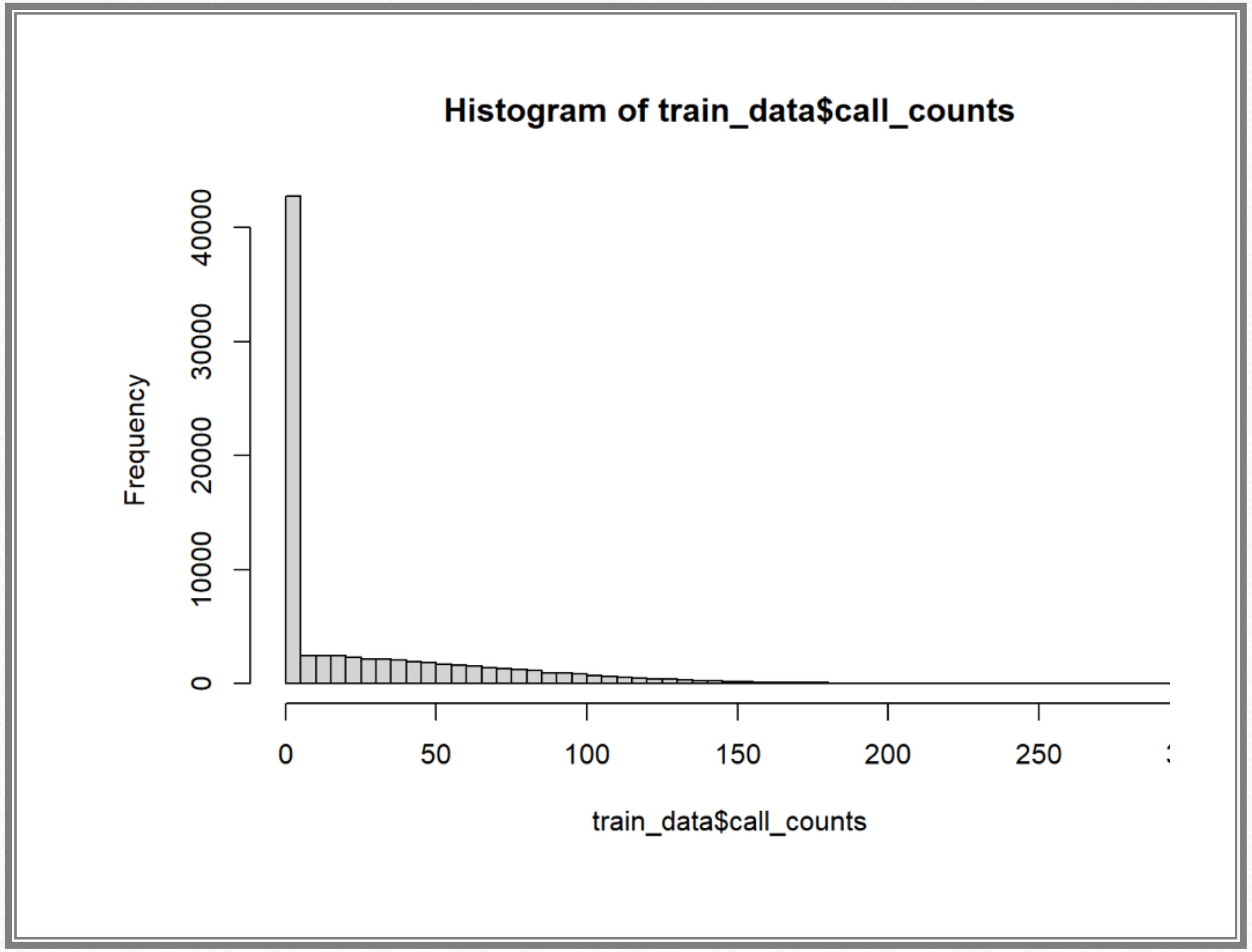
Data Description

- Variable Dictionary:
  - ~20 variables about each policyholder
    - Annual Premium Amount, Product Type, Geographic Description (e.g., rural), etc.
  - Target Variable: **call_counts (number of calls per policy)**

Understanding the Model Evaluation Metric

- **Relative Gini Index?**
  - Evaluates the **ranking ability** of predicted values.
  - Focuses on the **order** of values rather than their absolute accuracy.
- **Key Insight:**
  - **Correct prediction order** is critical: Mis-ranking one value affects the entire sequence.

$$\text{relative GINI} = \frac{\sum_{k=1}^{N}\left((k\cdot a_k)-\left(\sum_{i=1}^{k} a_i\right)\right)}{\left(\sum_{i=1}^{N} a_i\right)\cdot\left(\sum_{i=1}^{N} i\right)}$$



Challenge with Data and Metric

- **Over 50% of data have Y=0:**
  - This imbalance amplifies the importance of correctly ordering the majority class.
- **Handling Y=0:**
  - Correctly classifying data points where Y=0 helps maintain ranking integrity
  - Encourages exploring **mixture models**

Preprocessing Steps

- **Data Preprocessing:**
  - **Scaling Numerical Features**
  - **Missing Data Handling**
    - Complete Cases
    - Different Imputation Methods
- **Feature Selection:**
  - **Goal: Cut out the chaff** (multicollinear, insignificant, etc.)
  - **Utilize automatic feature selection** (CatBoost, LASSO regularization)

Initial Models Tried

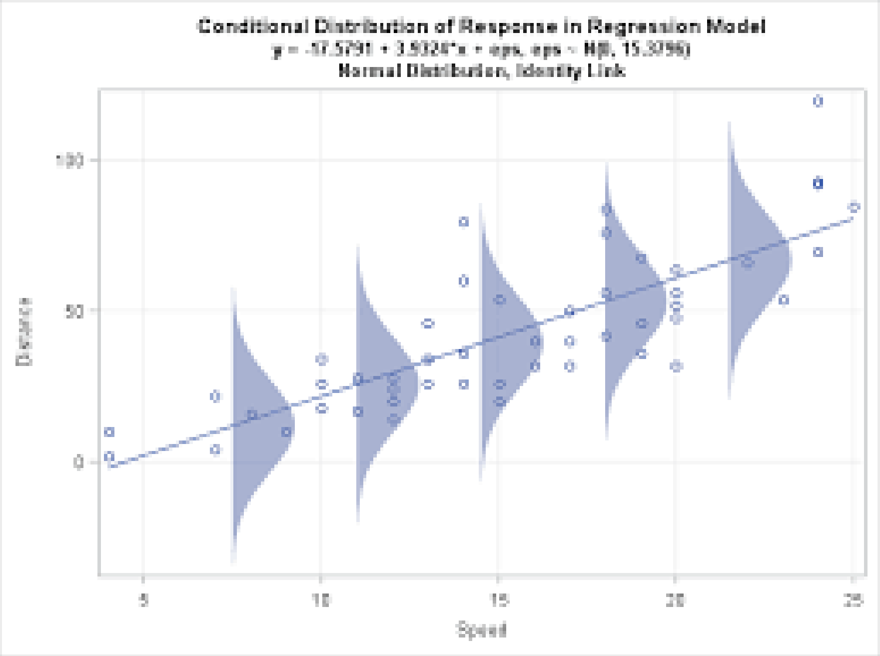
- Poisson Regression
- Random Forests
- XGBoost
- LightGBM
- Neural Network

Final Selected Models

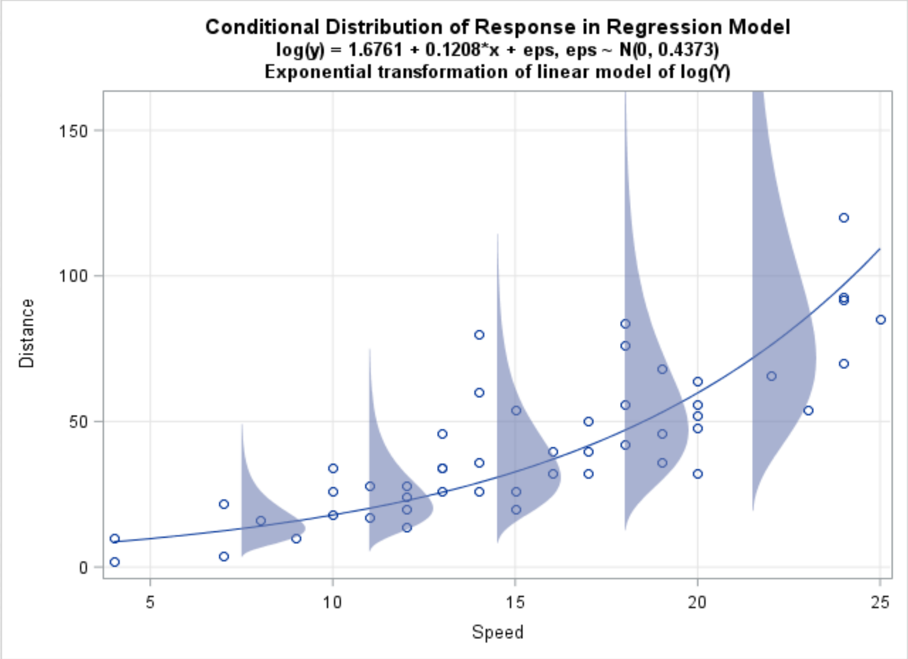
- 1. Generalized Linear Model
- 2. Gradient Tree-Boosted Two-Part Hurdle Model
- 3. Gradient Tree-Boosted Tweedie Hurdle Model

1. Generalized Linear Model: Tweedie Target Distribution w/ LASSO Regularization

Linear Model



Generalized Linear Model



Why a Tweedie Distribution?

- Tweedie distribution has a very nice property: it can effortlessly fit data that has a lot of 0 values, while still having a support on the positive real number line!
    - Sum of N exponential RVs where N ~ Poisson
- As about half of the samples are 0, it makes sense to go out of our way to account for this in the final model

Why LASSO Regularization?

- Needed a way to reduce the number of parameters – overfitting & perfect collinearity were both evident
- Model is large enough that other feature selection methods (even stepwise selection) felt unfeasible
- LASSO regularization works via iterative method (Newton's method), so we could adjust the number of steps taken to fit our computational limitations

Second Order (Interaction) Effects?

- Interested to see if any variables have interactions which are significant predictors for the number of calls made
- In this model, none appeared to give a notable increase in predictive power

- It's possible that this was a limitation of the Newton's Method approach (couldn't get close to the optimal model given very limited number of steps)

Summary of Generalized Linear Model (GLM)

- Fit a **LASSO-regularized** GLM with **Tweedie** target distribution

- **Tweedie Target Distribution:**
  - Notable mass of zero values but also many large values
- **LASSO-regularization**
  - Large number of explanatory variables
  - Interaction terms did not prove to be useful. Could be computational limitation

2. Gradient Tree-Boosted Two-Part Hurdle

- **Stage 1: Zero/Non-Zero Classification**:
  - A **CatBoost Classifier** predicts whether the target variable is zero or non-zero.
- **Stage 2: Non-Zero Count Regression**:
  - For non-zero target values, a **Negative Binomial Regressor** predicts the exact count.

3. Gradient Tree-Boosted **Tweedie** Hurdle

- **Stage 1: Zero/Non-Zero Classification**:
  - A **CatBoost Classifier** predicts whether the target variable is zero or non-zero.
- **Stage 2: Non-Zero Count Regression**:
  - For non-zero target values, a **Tweedie Regressor** predicts the exact count.

Model Evaluation

- Train/Validation Split
- Cross Validation
- Two-Stage Fine Tuning

Results

Top 3 by **Private Score**

phc.csv Complete · Behrooz · 16d ago	**0.24890**	**0.25366**
hct.csv Complete · Behrooz · 13d ago	**0.24860**	**0.25438**
ftr.csv Complete · Behrooz · 17d ago	**0.24801**	**0.25541**

Top 3 by **Public** score

predictionsGLM.csv Complete · Nathan Munshower · 5d ago · Regularized first order Tweedie GLM	**0.24677**	**0.25642**
ftr.csv Complete · Behrooz · 17d ago	**0.24801**	**0.25541**
GTBTCP.csv Complete · Behrooz · 13d ago	**0.24471**	**0.25498**

THANK YOU

Q&A