

Hawkeye Datathon

2025

Dylan Day

Natalie Schwartz

Agenda

- 1. Overview**
- 2. Data ETL**
- 3. Data Exploration**
- 4. Model Selection**
- 5. Model Deployment**
- 6. Results**

Overview

FutureBright Insurance is seeking a predictive model to estimate expected auto claim losses and improve customer risk segmentation.

Using historical exposure and claims data, we build and compare several modeling approaches such as XGBoost, transformer-based TabPFN, and composite frequency–severity models which are supported by rigorous variable reduction and model tuning.

The presentation covers the full workflow from data exploration to model evaluation and business interpretation.



Data Extract, Transform, Load

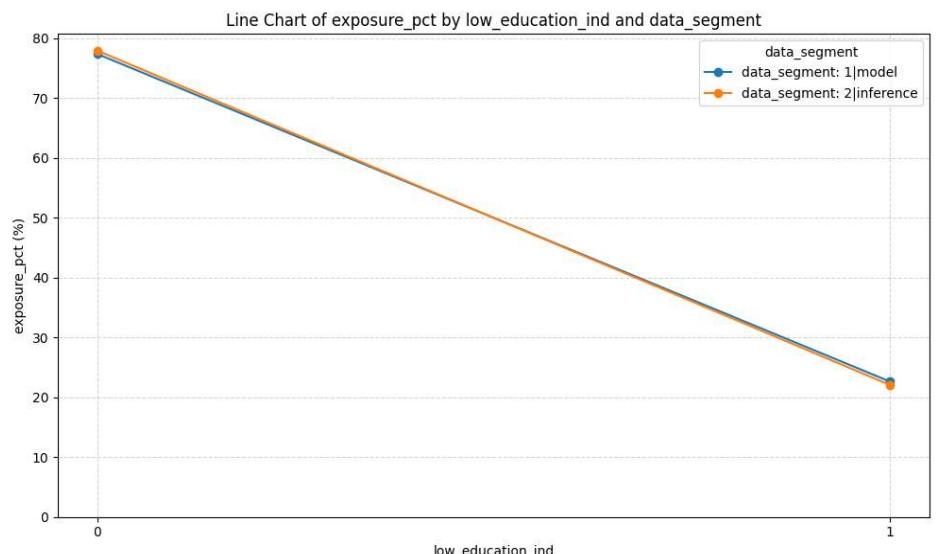
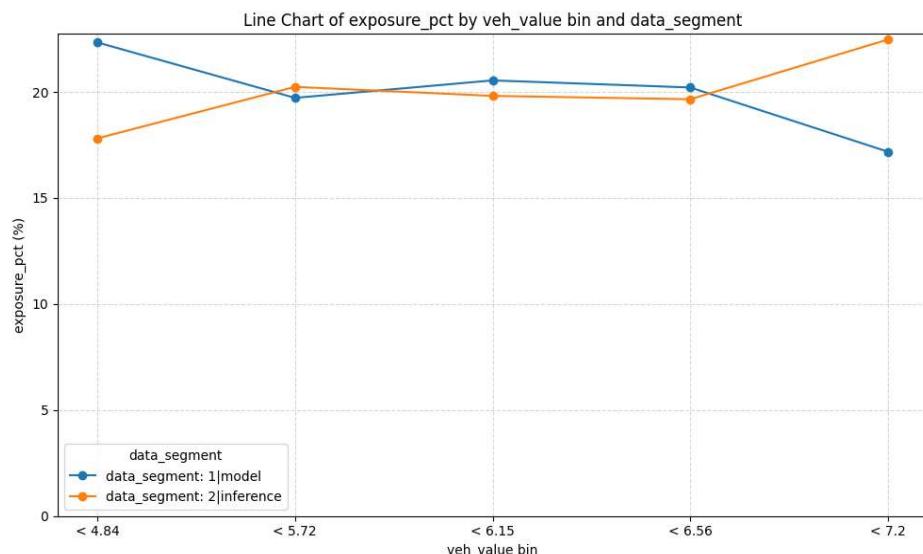
In the ETL stage, the policy-level and claim-level datasets were loaded, cleaned, and merged to create a modeling and inference dataset.

- Model and Inference data extracted and visualized to identify potential issues.
- ‘veh_value’ and ‘credit_score’ are capped at their 99th percentile for stability.
- Data is loaded into memory for further exploration.

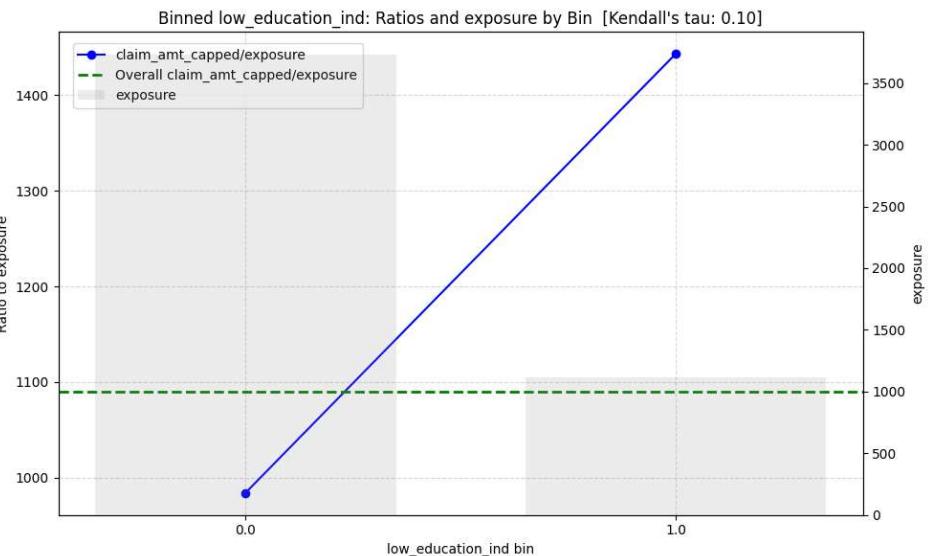
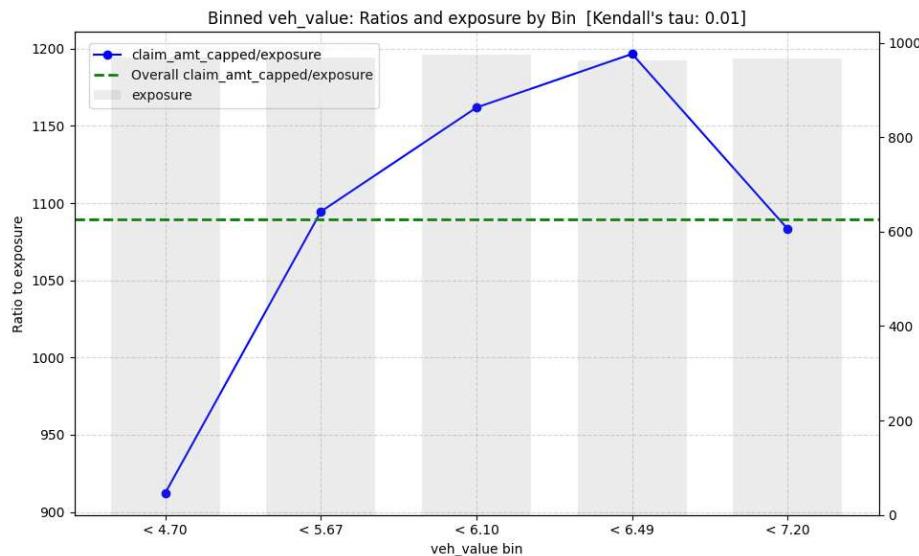
Data Exploration

- Loaded and prepared modeling and inference datasets.
- Analyzed target distributions and applied a 99th percentile severity cap to control extreme values.
- Generated diagnostic plots to understand heavy-tailed loss behavior.
- Assessed predictor stability using a consistency check across modeling and inference data.
- Ran predictiveness checks to evaluate how each predictor relates to frequency, severity, and exposure.

Data Exploration



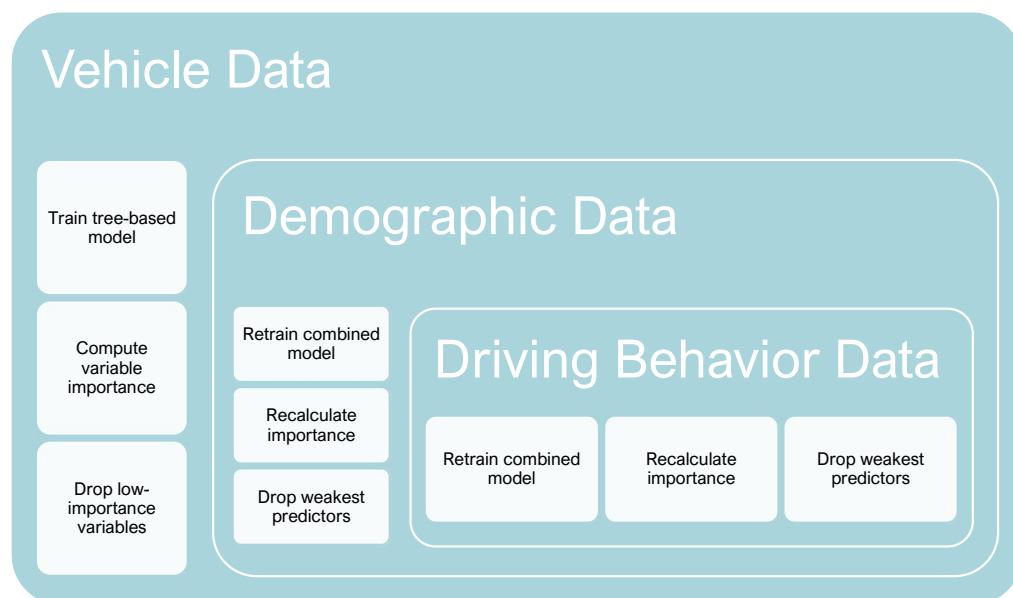
Data Exploration



Data Exploration: Variable Reduction

First Round: Variable Reduction via Tree-based Model

1. Start with the most promising data source and train a tree-based model.
2. Remove low-importance predictors, keeping only a handful of strong variables.
3. Add the next data source, retrain, and again remove weak predictors.
4. Repeat for the rest of the data sources until all sources have been screened.



Data Exploration: Variable Reduction

Second Round: Variable Reduction via VarClusHi

What VarClusHi Does:

- Groups variables into clusters based on shared variance (similar to hierarchical PCA).
- Identifies sets of predictors that carry redundant information.
- Helps reduce multicollinearity by selecting one representative variable per cluster.

How It Works:

Computes each variable's R^2 with its own cluster vs. R^2 with the nearest competing cluster.

| Variable | RS_OWN | RS_NC | RS_RATIO | Importance | Kendall Tau |
|------------------------|----------|----------|----------|------------|-------------|
| veh_value | 0.753764 | 0.105931 | 0.275410 | 0.148462 | 0.005068 |
| max_power | 0.803129 | 0.114554 | 0.222341 | 0.061292 | 0.000277 |
| veh_age | 0.753764 | 0.009796 | 0.248672 | 0.058669 | 0.021723 |
| veh_body_SUV | 0.573698 | 0.028331 | 0.438731 | 0.042644 | 0.048809 |
| driving_history_score | 0.506224 | 0.006829 | 0.497172 | 0.038845 | -0.041404 |
| veh_body_SEDAN | 0.499745 | 0.037815 | 0.519915 | 0.034991 | -0.031615 |
| low_education_ind | 0.000624 | 0.000874 | 1.000250 | 0.031098 | 0.101715 |
| credit_score | 0.003006 | 0.001924 | 0.998916 | 0.028970 | -0.060454 |
| time_driven_6am - 12pm | 0.834657 | 0.006283 | 0.166388 | 0.015685 | 0.009180 |
| veh_body_PANVN | 0.027737 | 0.002069 | 0.974279 | 0.012804 | -0.007882 |

Model Selection: Overview

Built Base Model

- Constructed a Tweedie-based XGBoost model to predict claim cost per exposure.
- Generated predictions, evaluated performance (Top-Lift, Gini, RMSE, MAE, R²).

Hyperparameter Tuning of Tree-Based model

- Tuned parameters in sequential blocks for interpretability and efficiency.
- Performed XGBoost CV for each grid and selected best combination by lowest RMSE.
- Trained a fully tuned model and re-evaluated using validation metrics.

TabPFN

- Implemented TabPFNRegressor on a subsample (1,000 train / 500 val) due to CPU constraints.
- Evaluated model performance.

Composite Poisson-Gamma Model

- Poisson regression for claim frequency
- Gamma regression for claim severity (only on positive losses)
- Combined frequency × severity to obtain total loss predictions.

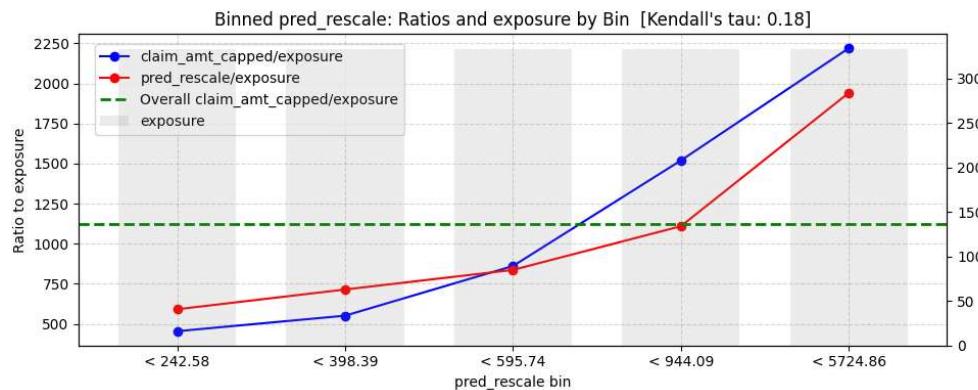
Model Comparisons and Conclusions

- Compared all models across validation metrics (RMSE, MAE, Gini, R², Top-Lift).
- Summarized strengths/weaknesses of Linear/GLM, Tree-based, Deep learning/TabPFN models
- Identified where each model type is most

Model Selection: Tree-Based Models

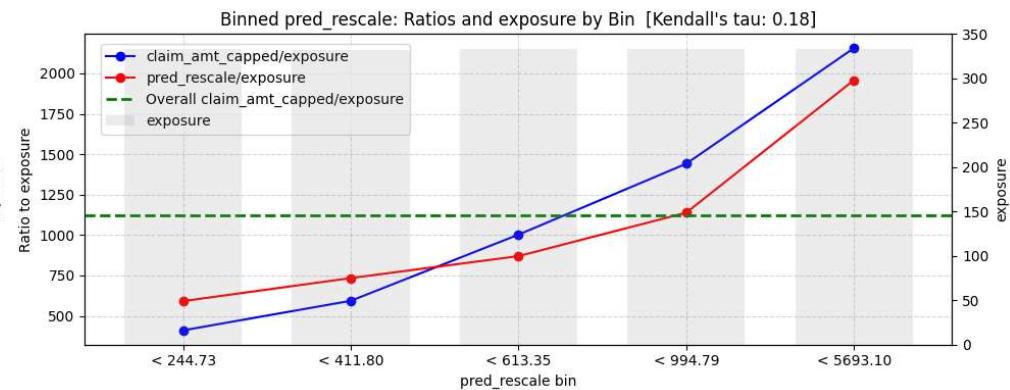
XGBoost Base Model

- Built an initial Tweedie-based XGBoost model using default or lightly-adjusted hyperparameters.
- Provided strong baseline metrics with minimal tuning, capturing nonlinearities and interaction effects automatically



XGBoost Tuned Model

- Improved the base model through staged hyperparameter tuning (learning rate, subsampling, tree depth, regularization).
- Achieved lower RMSE and higher R², indicating better generalization and more stable predictions.



Model Selection: Deep Learning/TabPFN

TabPFN

- Transformer-based model pretrained on millions of synthetic tabular datasets, enabling zero-shot prediction with no training or tuning.
- Designed for small, fully numeric, complete datasets, where it delivers fast and competitive performance.
- Acts as a universal learner by using in-context learning rather than gradient-based training.

Results

- In this project, TabPFN was applied to a small subset (1,000 training / 500 validation) due to compute constraints.
- On this subset, it produced higher RMSE and MAE and negative R², performing worse than XGBoost and the composite model.
- Still captured some predictive signal, achieving moderate Gini values, despite using less than 10% of the full data.
- RMSE: 2504.53, MAE: 1156.35, R²: -0.1021, Gini: 0.4625

Model Selection: Composite Model

What It Is

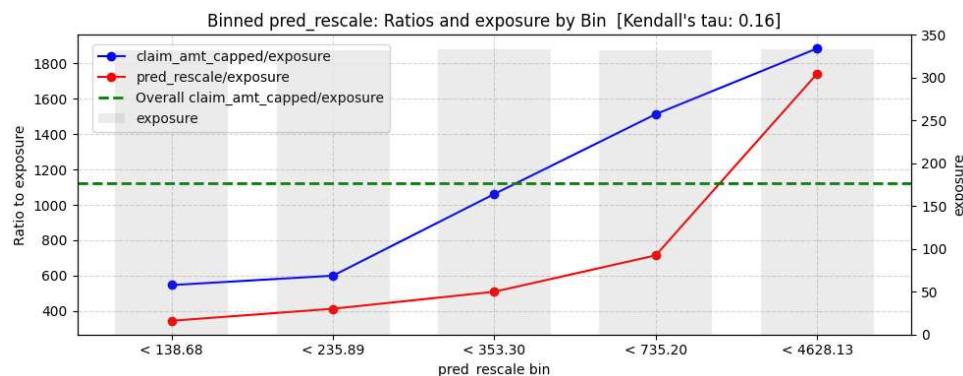
- A traditional actuarial modeling framework that separates claim frequency and claim severity.
- Uses Poisson regression to model the claim frequency.
- Uses Gamma regression (on positive losses) to model the claim severity.
- Combines the two components to produce expected total loss = frequency × severity.

Why It's Used

- Mirrors the structure used in real insurance pricing, making results easy to justify to stakeholders.
- Useful when modeling skewed, zero-inflated loss data where frequency and severity behave differently.

Results

- Achieved RMSE = 2341.77, MAE = 867.54, R² = 0.03, and Gini = 0.34.
- Performed reasonably well, but both the base and tuned XGBoost models achieved lower RMSE/MAE and Gini values higher by ~0.2, showing significantly stronger ranking ability.



Model Selection: Comparison

Tree-Based Models

Pros

- Strong predictive accuracy on tabular data
- Captures nonlinearities & interactions automatically
- Robust to outliers and skewed loss distributions
- Tuned model improves stability & generalization

Cons

- Requires tuning for best performance
- Harder to interpret than GLMs
- Training can be computationally heavier than linear models

TabPFN

Pros

- No training or hyperparameter tuning needed
- Extremely fast inference
- Captures complex patterns through pretraining

Cons

- Works best on *small datasets*
- Sensitive to input encoding and sampling

Composite Model

Pros

- Highly interpretable and industry-standard
- Separates frequency and severity for clarity
- Simple, transparent, and easy to communicate

Cons

- Limited ability to capture nonlinear interactions
- Predictive accuracy generally lower than ML models
- Relies heavily on correct distributional assumptions

- Tuned XGBoost Tweedie Regression model is the best model for predictability.
- The Composite Model is the best for interpretability.

Model Performance Summary

- XGBoost Tweedie clearly dominates TabPFN and the composite model.
 - TabPFN has the highest errors and even a negative R²
 - Composite Poisson-Gamma improves MAE but still lags in RMSE and Gini
- Base vs tune XGBoost: only incremental changes, but tuned has the best overall fit
 - Tuning slightly reduces RMSE and improves R², with essentially the same Gini as the base model, showing the base model was already strong.
- Final choice for deployment: tuned XGBoost Tweedie

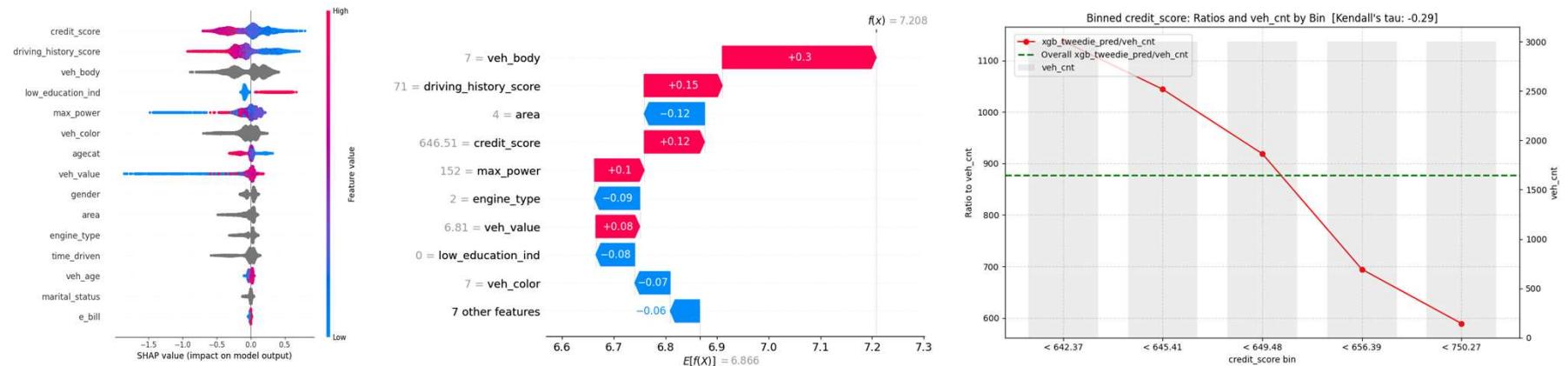
| Model | RMSE | MAE | R ² | Gini |
|-------------------------|---------|---------|----------------|--------|
| Base XGBoost Tweedie | 2336.32 | 695.18 | 0.0319 | 0.5656 |
| Tuned XGBoost Tweedie | 2327.12 | 746.63 | 0.0395 | 0.5631 |
| TabPFN | 2504.53 | 1156.35 | -0.1021 | 0.4625 |
| Composite Poisson-Gamma | 2341.77 | 867.54 | 0.0274 | 0.3423 |

Model Deployment:

Once the final model was selected, we operationalized it to generate predictions on the unseen inference dataset using the following workflow:

- Loaded the cleaned inference dataset and finalized modeling objects.
- Applied consistent preprocessing steps to inference data.
- Loaded the trained tuned XGboost model.
- Generated expected loss predictions at the policy level.
- Assigning risk tiers based on modeled expected loss
- Provided explanation outputs for business users through top-reason summaries and SHAP

Model Deployment: SHAP & Top Reasons



- SHAP identified global drivers of risk (credit, driving history, vehicle body)
- Waterfall plots show why an individual customer's risk is high or low (example: Quote 55 bin: 8)
- Top Reasons confirm monotonic patterns (higher credit → lower predicted loss)

Model Deployment: SHAP & Top Reasons

Business Results: Insights

- Risk is driven by vehicle + behavior more than demographics
 - The model shows higher risk for SUV's, high value, high-powered vehicles and worse driving history/ low credit scores
 - Hybrid engines, higher education, and some areas reduce risk.
- Example quote (ID 55) is slightly above average risk.
 - The waterfall plot shows this customer's SUV body, driving_history_score, credit_score, max_power, and veh_value push expected loss about 5% above the portfolio average.
 - This is partly offset by living in a safer area, driving a hybrid, and having higher education
- Recommendations
 - Underwriters should prioritize reviewing quotes where modeled risk is elevated due to vehicle attributes, driving history, or credit indicators. Consider pricing surcharges or coverage modifications

Final Takeaways

- Tuned XGBoost Tweedie is the best-performing model, showing the strongest ranking ability and lowest error metrics
- Model is operationalized and explainable, producing policy-level predictions, risk tiers, and transparent explanations.
- Risk is primarily driven by vehicle attributes and driving behavior, not demographics.
- Underwriters can take action by focusing review or pricing adjustments on SUV's, high-powered vehicles, low credit scores, and poor driving histories.
- Prototype is ready for next steps, such as calibration, integration into rating engines, and ongoing monitoring.

END