Predicting Insurance Claim Costs: A Data-Driven Approach for Premium Development

OBJECTIVE

Analyse claim frequency and severity using machine learning to develop accurate cost estimates and quantify uncertainty.

ABOUT DATASET

Data preparation and Cleaning

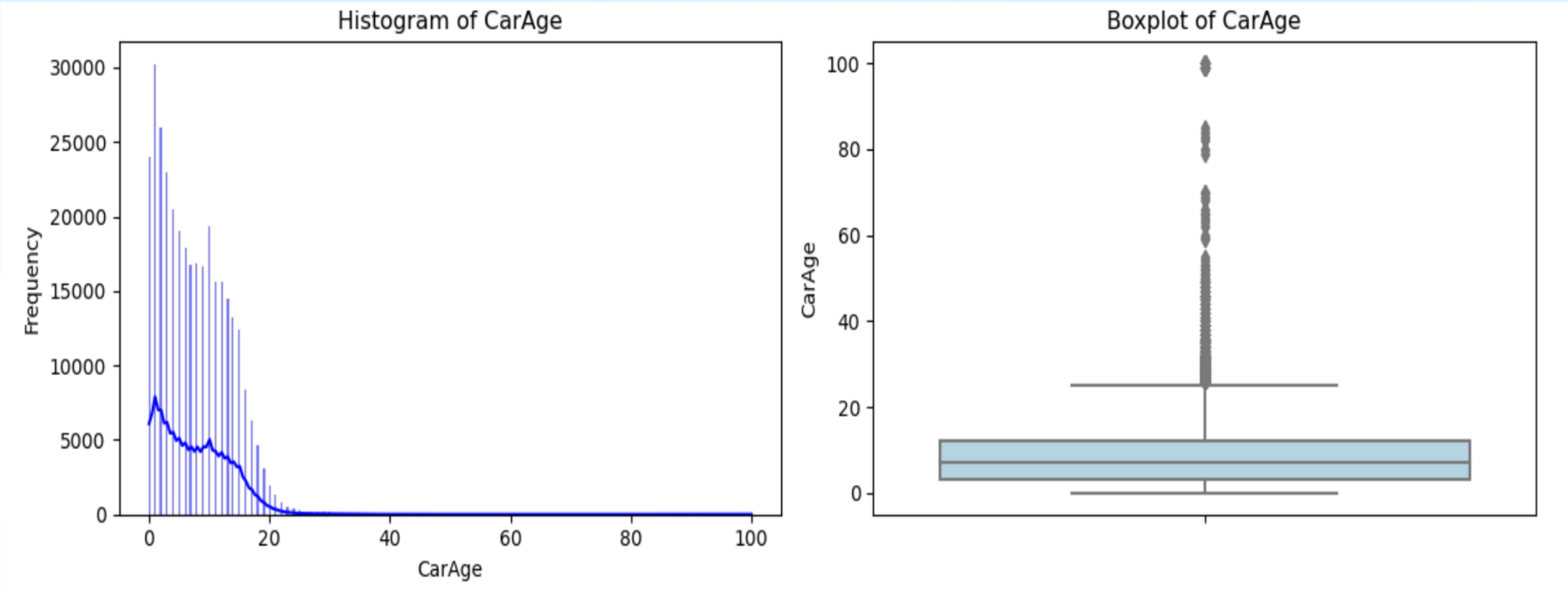
- The merged dataset has 10 columns with 330535 obsearvations. The target variable to be predicted is PurePremium.
- Risk factors: (e.g: DriverAge, CarAge, Power)
- Categorical attributes: (e.g: Region, Brand, Gas)
- Removed column: (Unnamed: 0, Exposure)

Exploratory Data Analysis

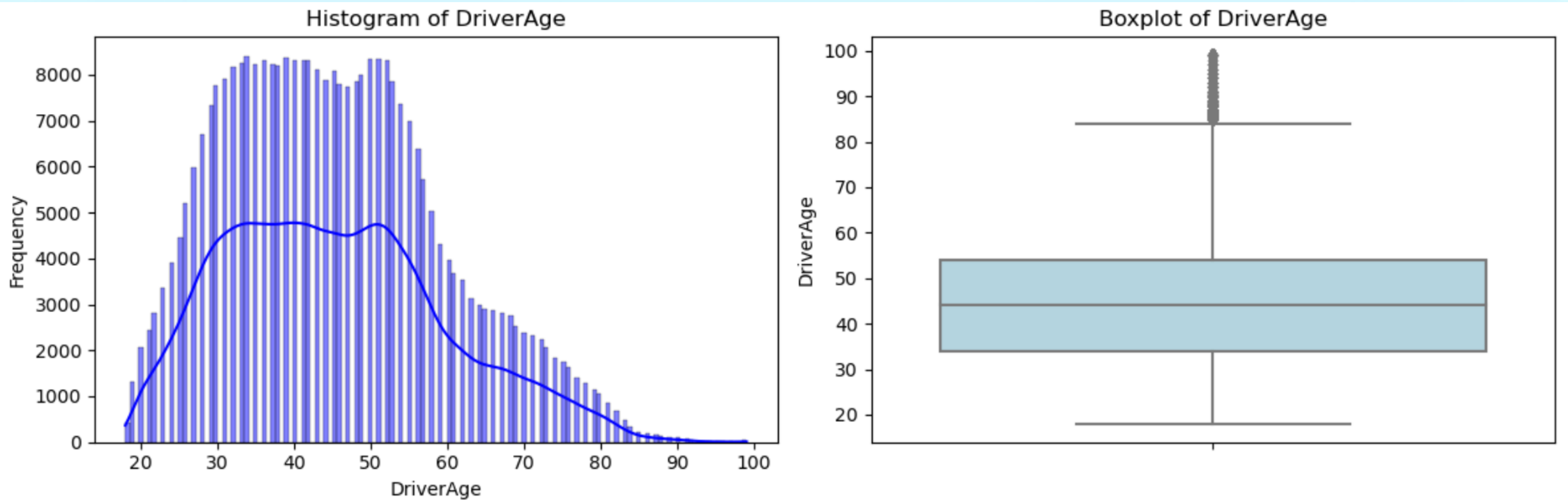
Exploratory Analysis Overview

- Data preparation and cleaning
- Univariate Data Analysis
- Bivariate Data Analysis

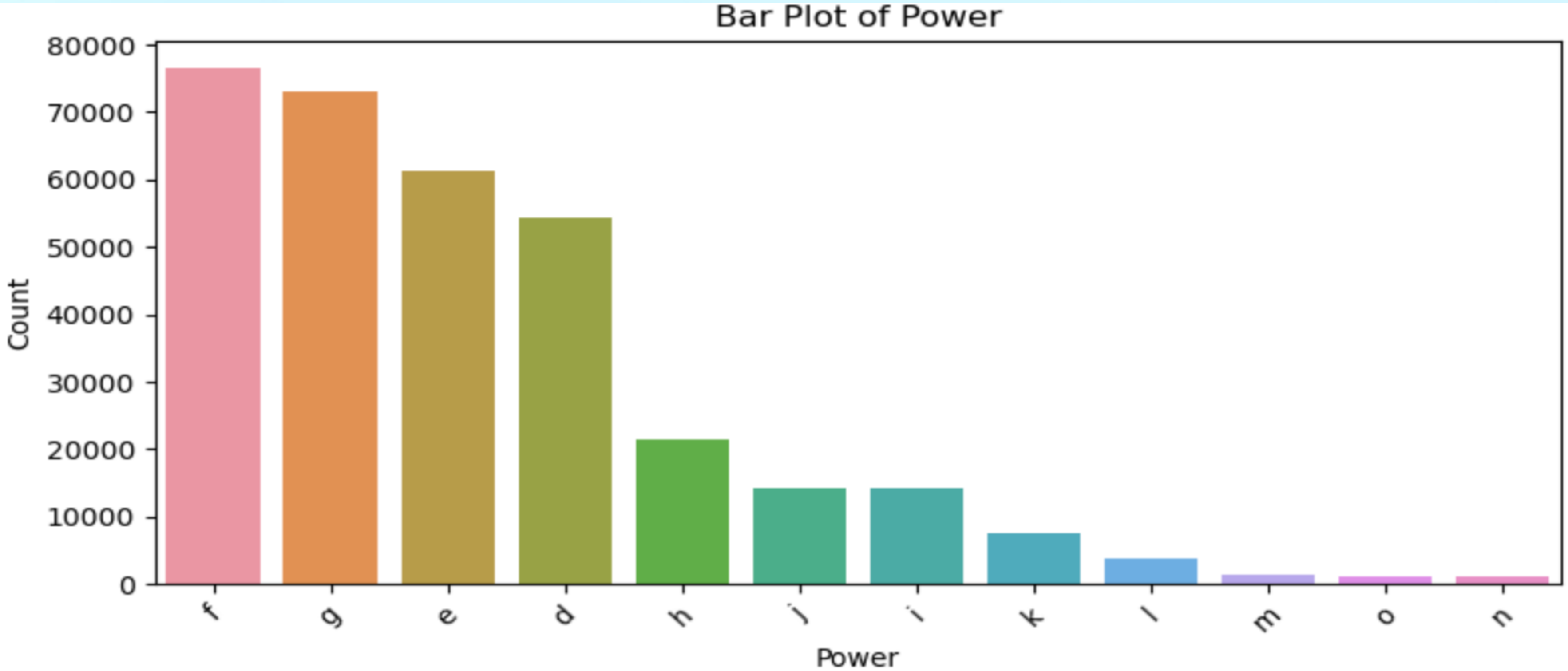
Univariate Analysis



Univariate Analysis



Univariate Analysis

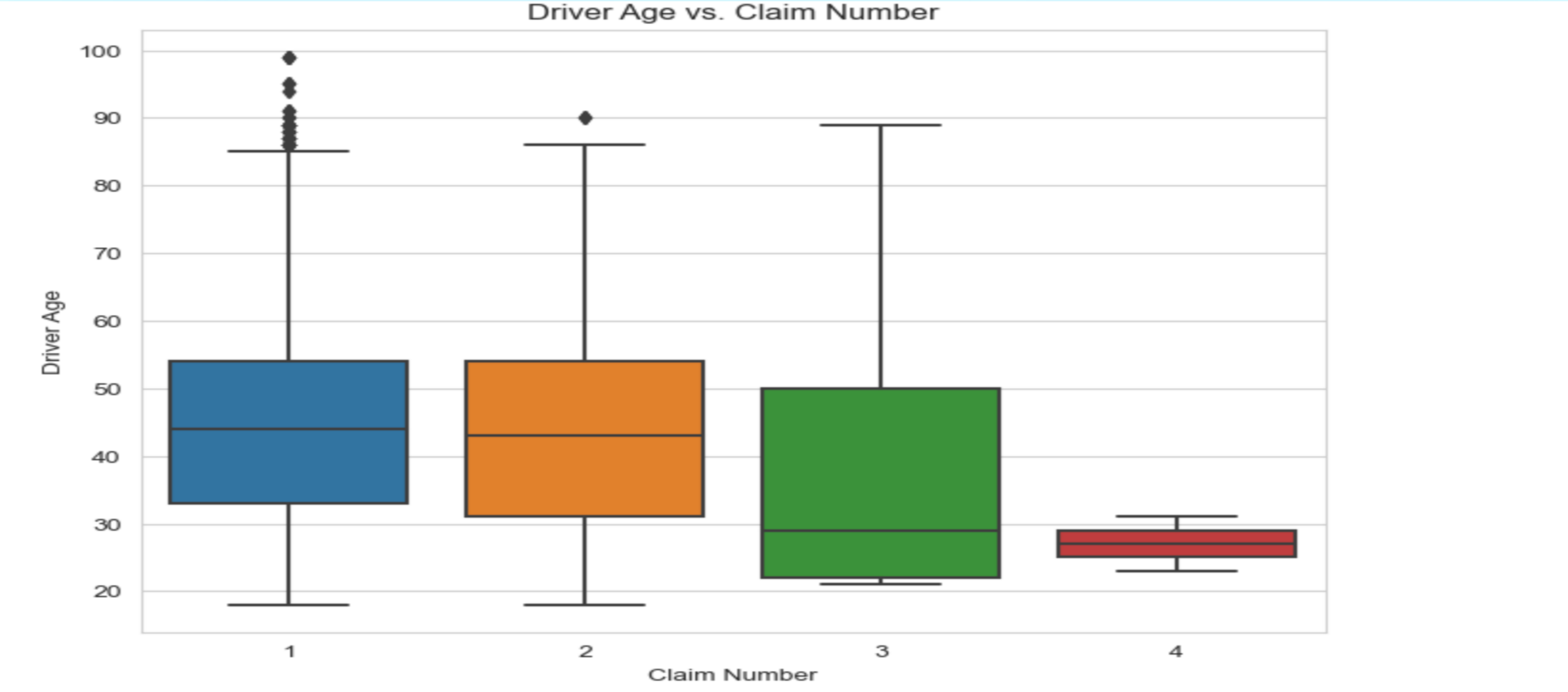


Bivariate Analysis

- According to the plot, there exist significant dependence between variables
- In this section we aim to illustrate the bivariate plot between variables.

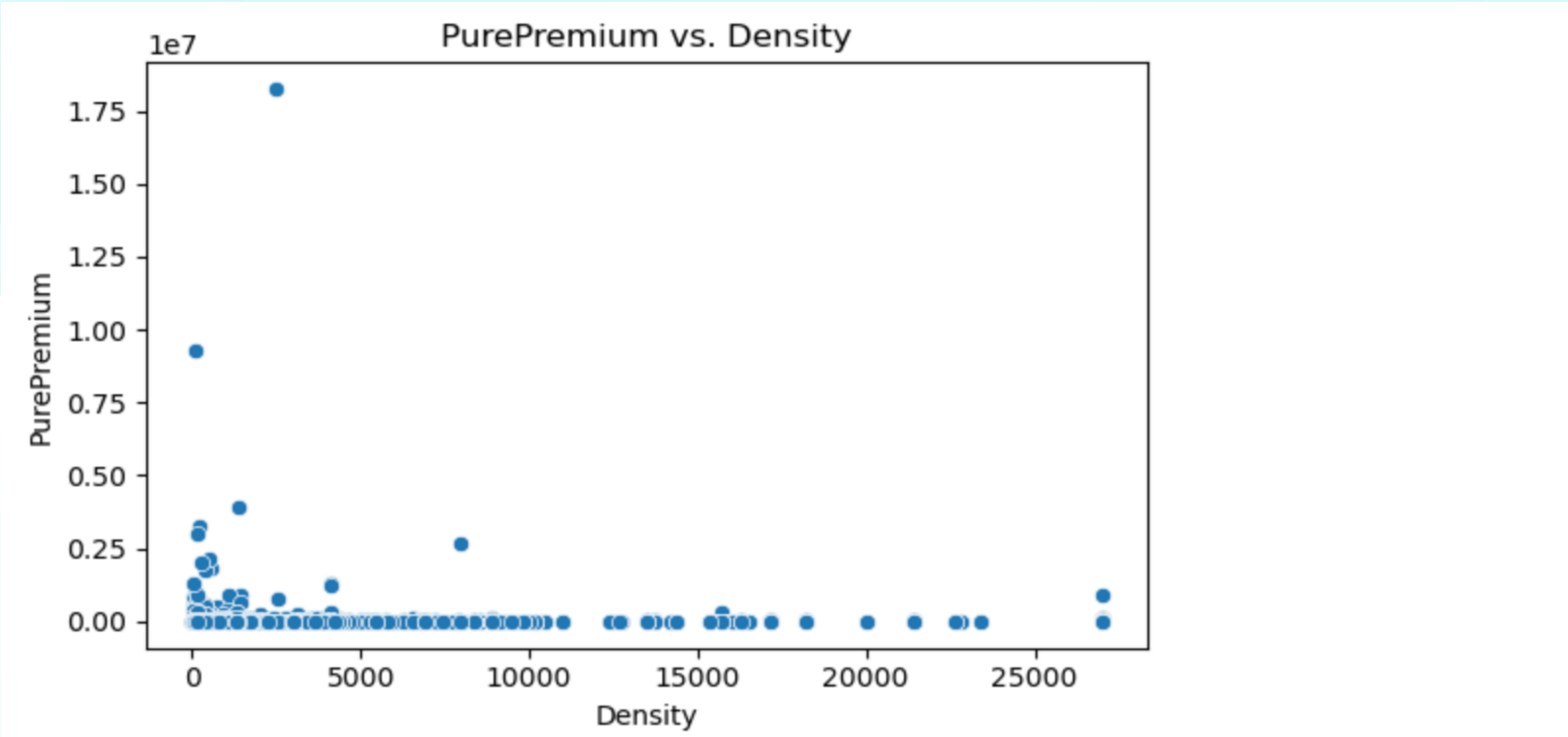
Bivariate Analysis

Driver Age VS Claim Number



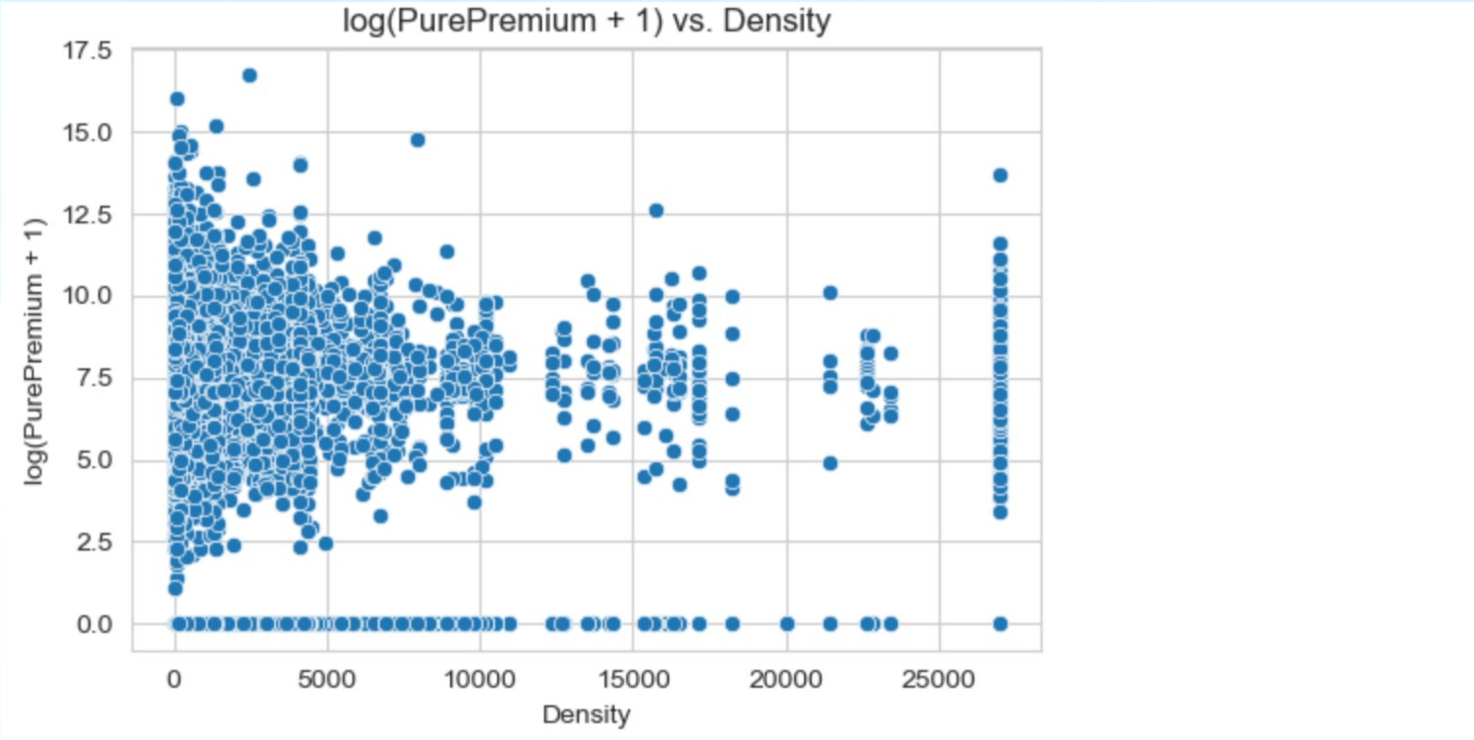
Bivariate Analysis

Pure Premium VS Density



Bivariate Analysis

logged Premium VS Density



COMPOUND MODELING

TWEEDIE REGRESSION

MODEL SUMMARY AND UNCERTAINITY QUANTIFICATION

```
=== Tweedie Regressor Model Summary ===
Best Hyperparameters: {'power': 1.7, 'alpha': 0.01}
Training Time: 44.18 seconds
Model Type: TweedieRegressor (Generalized Linear Model)
Link Function: log
Number of Features: 40
Intercept: 4.0605
Sample Coefficients (first 5):
  Feature 0: -0.0836
  Feature 1: -0.0701
  Feature 2: 0.3608
  Feature 3: -0.0010
  Feature 4: -0.0423
Total Number of Coefficients: 40
```

Single Tweedie Regressor Performance with Uncertainty

RMSE = 251.70

MAE = 92.56

R-squared = -0.001

Training Time: 44.18 seconds

	Actual	Predicted	Lower CI (95%)	Upper CI (95%)
0	0.0	41.138110	28.567097	53.709122
1	90.0	52.080292	37.935883	66.224701
2	0.0	81.306693	63.633654	98.979732
3	0.0	85.449402	67.331721	103.567083
4	0.0	86.925910	68.652369	105.199451

Prediction Time (including conformal intervals): 0.01 seconds

Baseline Claim Cost Prediction Model

Separate Tweedie Approach (Frequency & Severity)

- **Frequency Model (Power = 1.5)**
- **Severity Model (Power = 1.5, Alpha = 0.5)**

Metric	Value
Root Mean Squared Error	1314.50
Mean Absolute Error	152.49
R-squared	-0.00

Tweedie Grid Searched Model with Log Transformations

Enhanced Composite Model – Frequency & Severity with Hyperparameter Tuning

- **Frequency Model:** grid search tuning of power and alpha.
- **Severity Model:** log-transformed severity with similar hyperparameter tuning.

Metric	Value
Root Mean Squared Error	1314.46
Mean Absolute Error	118.03
R-squared	-0.00

Component	Best Parameters	
Frequency Model	alpha: 5	power: 1.2
Severity Model	alpha: 0.5	power: 1.2

Enhanced Tweedie Regressor with Uncertainty

Polynomial Features & Randomized Hyperparameter Search (Compound)

➢Data Preprocessing & Feature Engineering:

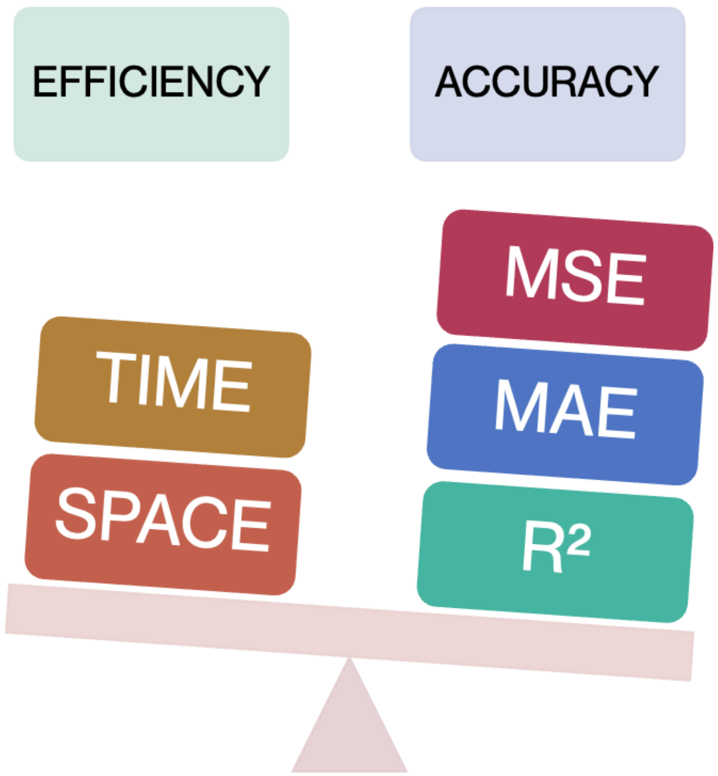
- Capped extreme ClaimAmount at the 99.5th percentile
- Applied StandardScaler and PolynomialFeatures (degree=2)

➢Tweedie Regression with Uncertainty:

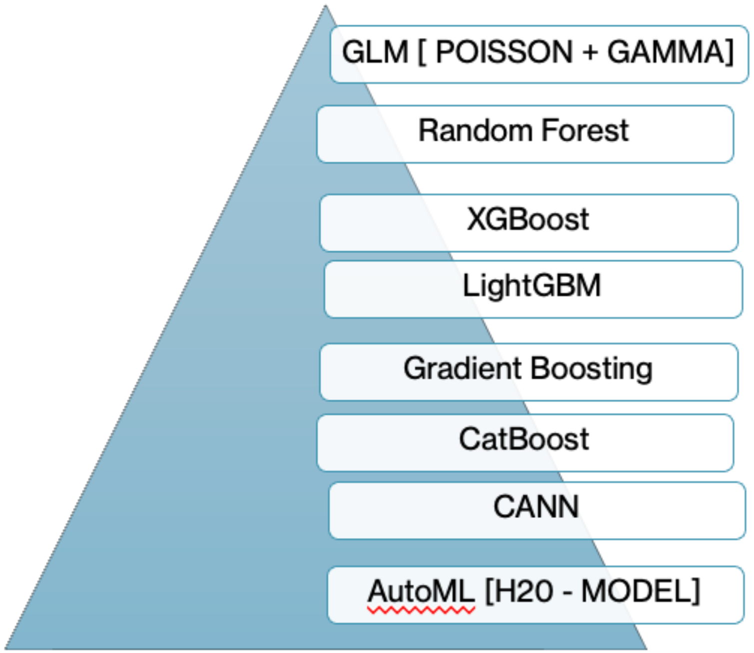
- Hyperparameter tuning via RandomizedSearchCV (alpha=1.7)
- Estimated uncertainty with approximate standard errors and 95% confidence intervals

Metric	Value
RMSE	251.70
MAE	92.56
R-squared	-0.001

WHAT DEFINES A GOOD MODEL?



SEPARATE MODELLING



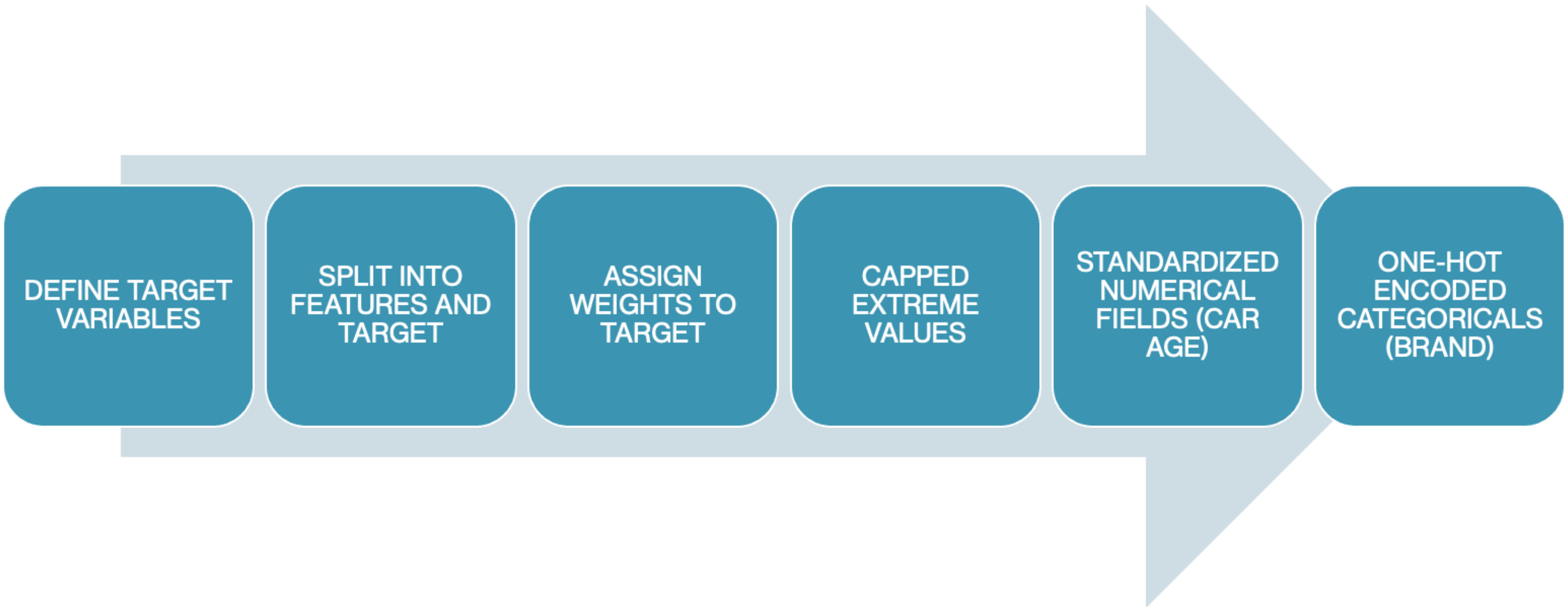
Predicted frequency and severity individually, then multiplied for PurePremium.

COMPOUND MODELLING

TWEEDIE REGRESSION

Direct PurePremium prediction

SETTING UP DATA FOR MODELING



HOW WE EVALUATED SUCCESS

Metrics: R^2 (fit quality), MSE/MAE (error magnitude), Training Time (efficiency).

Uncertainty: Used four methods for Tweedie – Conformal (coverage-guaranteed), Parametric (normality-based), Bootstrapped (non-parametric), Quantile (distribution-free).

Goal: Balance Accuracy (better predictions) and Efficiency (faster models) while quantifying risk.

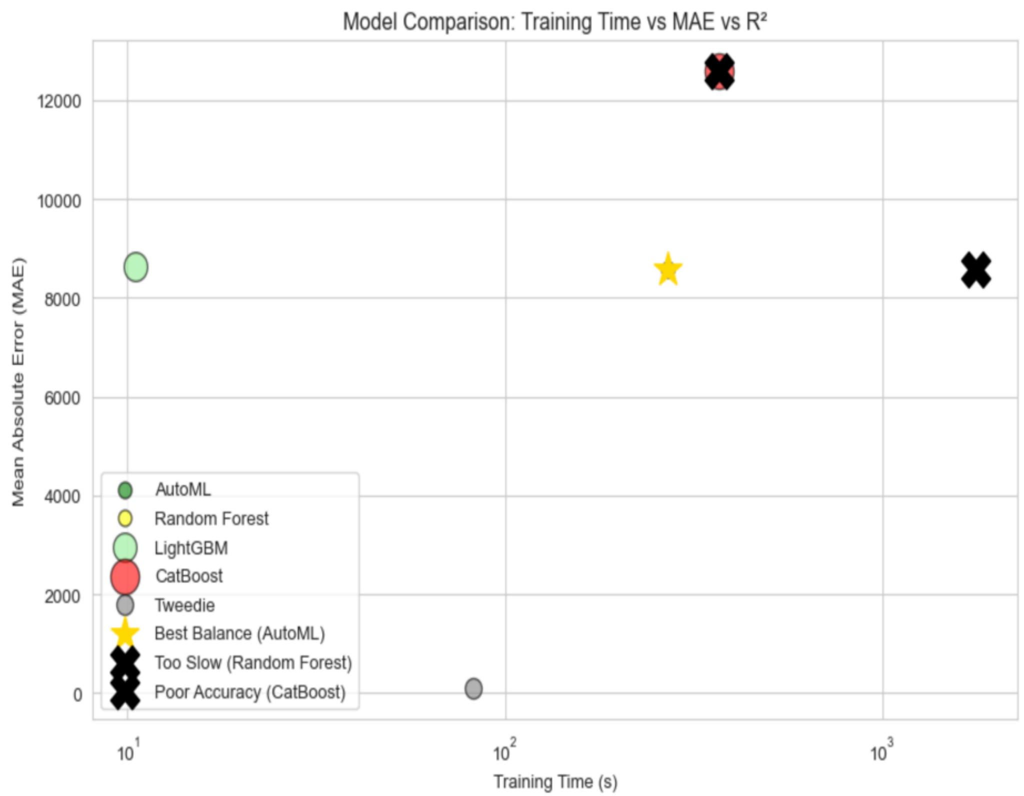
Best Ensemble Model – Why AutoML(h20 model) Wins?

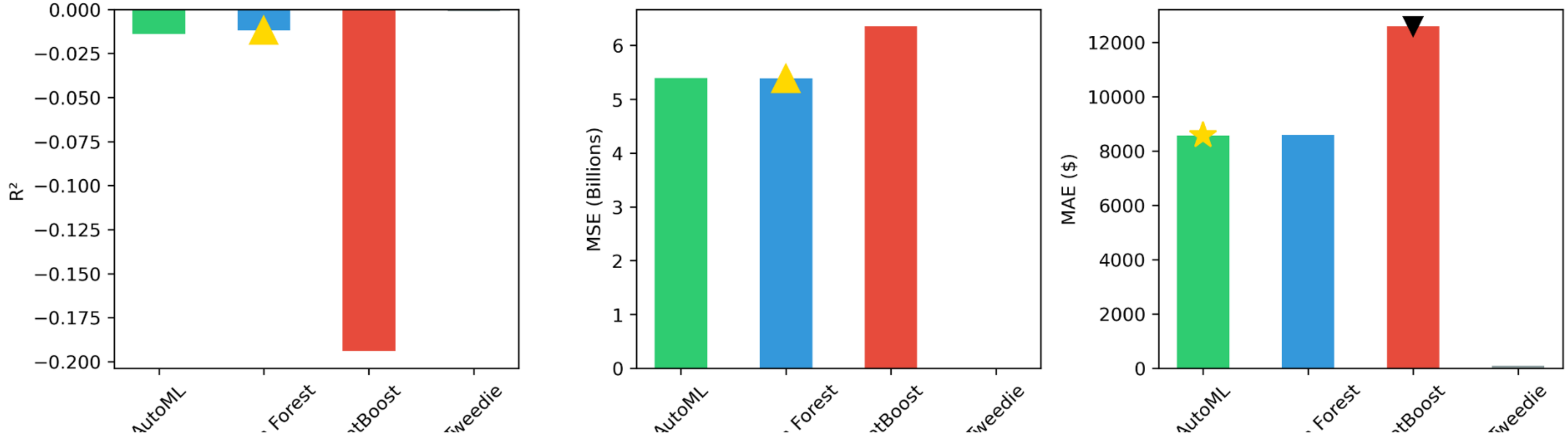
AutoML shines: Green bubble at 270s with lowest MAE, starred for balance.

Random Forest lags: Yellow bubble at 1,772s, thumbs-down for slow training.

CatBoost fails: Red bubble with high MAE, thumbs-down for poor accuracy.

Alternatives: LightGBM (fast at 10.46s), Tweedie (high MAE at 82.51s).

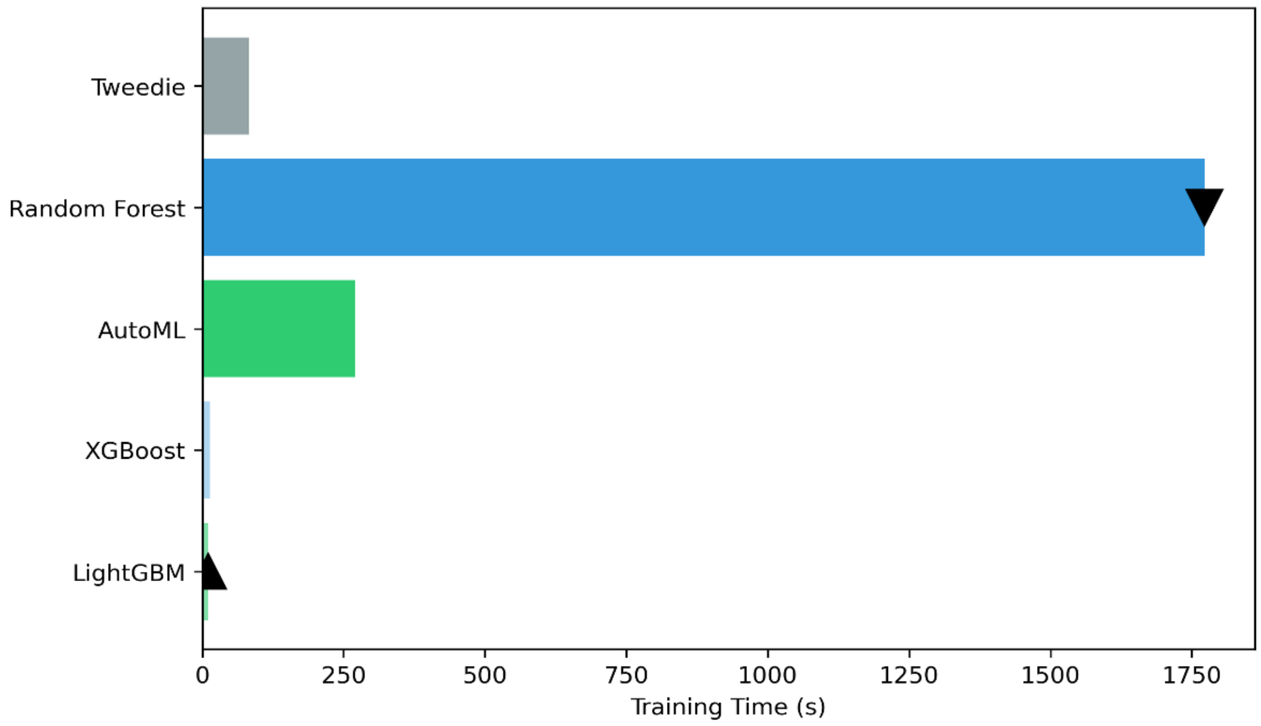




SEPARATE MODELING – ACCURACY INSIGHTS

- AutoML leads in consistency with the lowest average error (MAE), marked by a star.
- Random Forest captures trends slightly better (R^2, MSE), indicated by crowns.
- CatBoost struggles significantly across all metrics, shown with a down arrow.
- Tweedie's compound approach underperforms, especially in MAE, highlighting data challenges.

Separate Modeling – Efficiency Insights

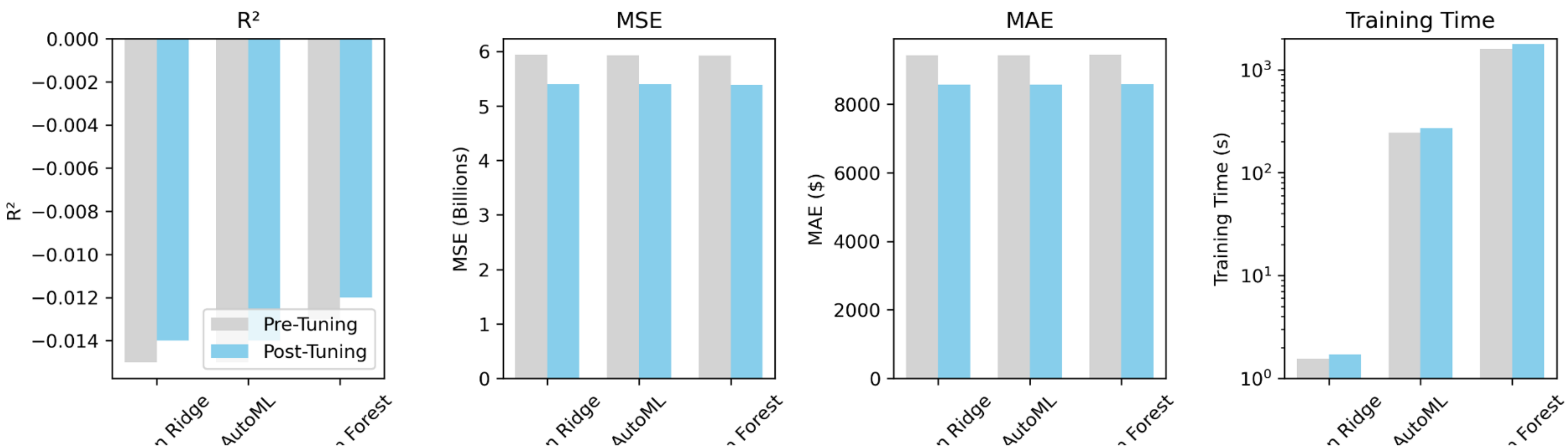


- LightGBM trains fastest, marked by a (^), ideal for quick deployment.
- Random Forest is slowest, indicated by a (v), impractical for frequent updates.
- AutoML and Tweedie offer moderate training times, balancing speed and performance.
- Efficiency varies widely, impacting real-time pricing feasibility.

PRE AND POST HYPERPARAMETER TUNING

	Model	R^2	MSE	MAE	Training Time (s)
0	GLM (Poisson + Gamma)	-0.014	5396987842.937	8595.863	6.198
1	Random Forest	-0.012	5386732416.964	8583.659	1772.506
2	XGBoost	-0.014	5396349041.145	8633.439	13.422
3	LightGBM	-0.014	5396815478.004	8628.241	10.461
4	Gradient Boosting	-0.018	5417156062.812	8660.580	59.614
5	CatBoost	-0.194	6353509005.337	12579.032	369.903
6	Bayesian Ridge	-0.014	5396082380.925	8582.585	1.723
7	CANN	-0.014	5397472910.623	8591.836	921.397
8	AutoML	-0.014	5395772664.472	8572.259	270.323

	Model	R^2	MSE	MAE	Training Time (s)
0	Bayesian Ridge	-0.014	5396044946.304	8581.497	248.145
1	Random Forest	-0.013	5392066460.035	8606.918	1684.970
2	AutoML	-0.014	5395772664.472	8572.259	262.456



- AutoML's consistency improved, with a 10% drop in average error (MAE).

Impact of Hyperparameter Tuning – Top 3 Models

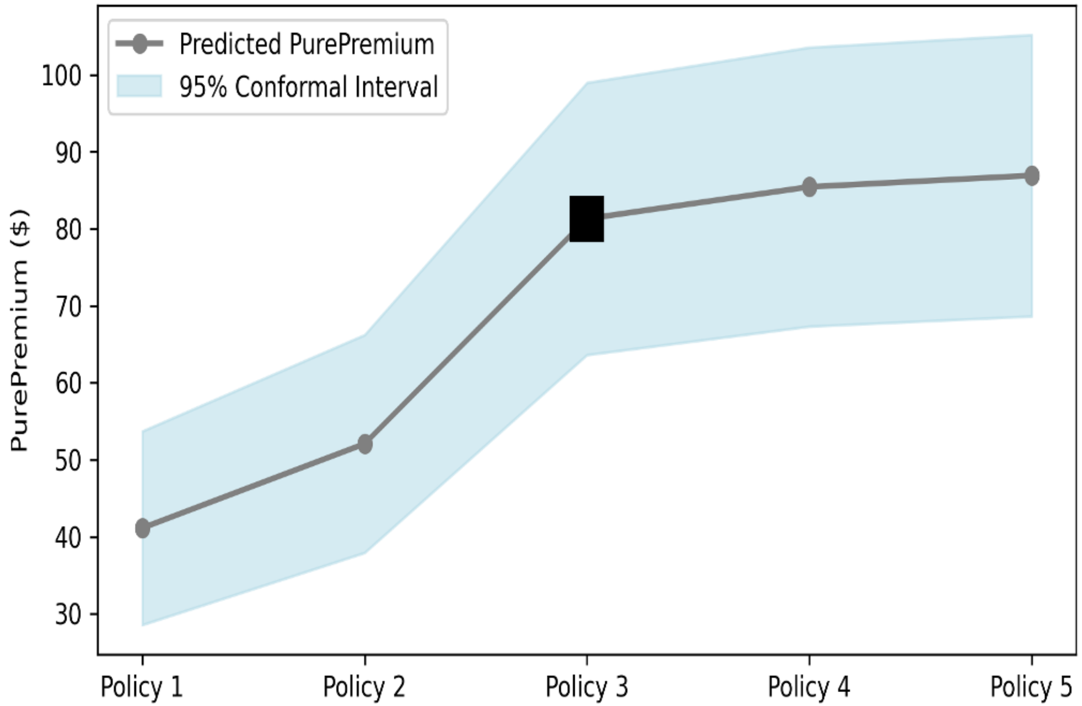
- Random Forest captured trends better, with gains in R^2 and MSE.

- Bayesian Ridge's training time increased slightly, but remains fastest.

- Tuning boosted accuracy by 10–20%, with a trade-off in efficiency.

Compound Modeling – Risk Assessment with Tweedie

- Tweedie predicts rising PurePremium across policies, reflecting varying risk levels.
- 95% Conformal intervals, marked by a shield, ensure reliable risk ranges for planning.
- Intervals widen for costlier policies, indicating higher uncertainty in larger claims.
- Outperforms other interval methods (Parametric, Bootstrapped, Quantile) for coverage.



SUMMARY OF OUR FINDINGS

TWEEDIE REGRESSION WITH UNCERTAINITY QUANTIFICATION

Metric	RMSE	MAE	R-squared	Training Time (s)	Prediction Time (s)
Value	251.70	92.56	-0.001	82.51	0.01

COMPOUND MODELING (TOP 3 BEST PERFORMING) WITH UNCERTAINITY

(a) Top 3 Models After Hyperparameter Tuning with Uncertainity Quantification

Model	R^2	MSE	MAE	Training Time (seconds)	Prediction Time (seconds)
AutoML	**-0.014**	**5395772664.472**	**8572.259**	**262.456**	**1.64**
Bayesian Ridge	-0.014	5396835904.628	8583.953	248.145	0.61
Random Forest	-0.013	5396715149.359	8574.722	2089.18	4.83

THANK YOU!