Executive Summary — Claims Severity Analytics

Objective

Develop a claims severity modeling framework to identify loss drivers, quantify tail risk, and support severity-aware insurance decision-making.

Data

The analysis uses a P&C claims dataset with ~188,000 observations and 132 features, combining categorical segmentation variables with continuous risk measures. Claim severity (loss) serves as the modeling target.

Exploratory Findings

- Claims severity is highly right-skewed, requiring transformation
- The top 1% of claims contribute ~6% of total loss
- Individual features show weak linear correlation with severity
- Segmentation reveals meaningful severity differences across groups

Modeling Approach

- Log-transformed severity modeling
- Ridge regression as a stable linear baseline
- Random Forest regression to assess non-linear effects and ranking behavior
- Evaluation using both accuracy metrics and decile-based loss concentration

Results

- Ridge regression achieves strong predictive accuracy ($R^2 \approx 0.52$)
- Random Forest does not improve global error metrics but excels at ranking
- Top 10% of predicted claims account for ~26% of total loss
- Top 20% account for ~42% of total loss

Business Value

- Linear models support pricing, reserving, and portfolio analysis
- Ranking-based models enable claims triage and targeted review
- Loss concentration metrics provide actionable operational insight

Conclusion

Effective claims severity modeling balances accuracy, interpretability, and operational usefulness. Ranking performance and loss concentration often deliver greater business impact than marginal improvements in prediction error.