USC Marshall
School of Business

Insurance Loss Analytics Technical Report

DSO 530: Applied Modern Statistical Learning Methods
Spring 2025

Course Instructor:
Paromita Dubey, Assistant Professor
Data Sciences and Operations Department

EXECUTIVE SUMMARY

This project tackles a core challenge in the insurance industry: setting premiums that are fair, risk-based, and profitable—while avoiding adverse selection. Using a provided zero-inflated insurance dataset, we developed machine learning models to predict Loss Cost (LC), Historically Adjusted Loss Cost (HALC), and Claim Status (CS), enabling more accurate underwriting and pricing decisions.

A key innovation in this project was the two-stage prediction approach: first predicting LC, then combining original features and predicted LC to estimate HALC. This design improved overall stability and predictive accuracy while addressing data sparsity in claim events.

We tested a broad set of regression and classification models across tasks. LightGBM delivered the best performance for LC and HALC (MSEs of 500.82 and 1019.12), while XGBoost led CS prediction with a ROC-AUC of 0.74. Although Tweedie Regression aligned well with the theoretical data structure, it lacked the flexibility to capture complex interactions. Neural Networks showed promise but were limited by data size and interpretability.

These models enable real-time pricing, precise customer segmentation, fraud detection, and portfolio risk optimization—contributing to smarter underwriting and improved profitability. However, the study has limitations. The absence of temporal validation may limit long-term generalizability, and the complexity of ensemble and neural models raises interpretability concerns, particularly in regulated settings. Future work should explore time-based validation, integrate external data sources, and apply explainability techniques to enhance model transparency and adoption.

In summary, our findings demonstrate the power of ensemble learning to modernize insurance pricing, offering scalable, data-driven solutions that balance fairness, accuracy, and operational impact.

INTRODUCTION

The biggest challenge in the insurance industry is setting premiums that are fair for policyholders and profitable for providers. Mispricing can lead to risk exposure, financial loss, and adverse selection for the insurer. For instance, by overcharging low-risks individuals or undercharging high-risk ones, it can lead to adverse selection and profitability loss.

This project aims to address that problem through predictive models that calculate precise insurance loss, which will improve performance, reduce risk, and facilitate stronger portfolio performance. It is relevant to three major stakeholders: policyholders, policymakers, and insurers, who need reliable tools to perform sustainable operations. By employing Tweedie-based machine learning, we aim to bring more transparency, fairness, and profitability to the insurance market.

The analysis is structured around two key tasks. **Task 1** estimates the expected size of the loss if a claim occurs. We predict two important targets. First, the **Loss Cost per Exposure Unit, or LC**, measures the average claim size for a policyholder. Second, the **Historically Adjusted Loss Cost, or HALC**, adjusts

LC by how frequently claims happen historically. From a business perspective, this helps determine the financial impact when a risk materializes, and supports setting fair and sustainable premiums. **Task 2** predicts whether a loss event will occur. Here, we model **Claim Status (CS)**, where 1 means a claim is made and 0 means no claim. This task helps the business assess the risk of any claim happening, which is critical for understanding overall exposure and customer segmentation.

METHODOLOGY

1. Dataset

We worked with a training set of over 37,000 policy records across 28 variables, and a test set of about 15,700 records. The variables include numerical features like vehicle value and claim count, categorical variables like fuel type and risk type, and several datetime fields.

2. Data Preprocessing

We began by converting all relevant date fields (X.2–X.6) into proper datetime formats and engineered meaningful features such as the policyholder's age (X.3, X.5), policy duration (X.4, X.2), driver experience (X.3, X.6), vehicle age as of 2019 (X.22), and policy dropped (X.10, X.9). Categorical variables (X.7, X.13, X.19, X.20, X.21) were label-mapped to descriptive string values to enhance interpretability. For downstream modeling, we applied one-hot encoding to convert all categorical variables into dummy variables.

For missing values in X.27 (fuel type), we applied supervised classification models using relevant features, with XGBoost delivering the best balance between accuracy and generalization—especially in capturing the minority class. Imputed values were then integrated back into the training set.

To address multicollinearity, we performed correlation analysis and removed highly correlated features such as X.10, Vehicle_age, X.24, and other engineered variables. Outliers were handled by binning target variables (X.15–X.18) into logical ranges, allowing us to detect and exclude extreme values, thus improving model stability. Finally, we dropped non-informative or redundant columns (X.1, X.2–X.6, and X.22) after extracting relevant components, resulting in a cleaner and more focused dataset for modeling.

3. Exploratory Data Analysis (EDA)

To support our modeling strategy for predicting LC, HALC, and CS, we performed exploratory data analysis on both numerical and categorical features in the dataset.

Among the numerical variables, many exhibited pronounced right skewness. For instance, the total claim cost (X.15) contained a large number of zeros, suggesting that most policyholders did not file any claims. However, a small group of records showed exceptionally high claim values, introducing notable outliers. HALC, which is derived by dividing X.15 by X.16 (exposure) and scaling the result by X.18 (claim frequency), followed a similarly skewed pattern. Additionally, X.18 itself was concentrated near zero, reinforcing the rarity of claim events (Figure A1).

As for categorical features, we focused on vehicle type (X.19), region (X.20), and fuel type (X.27). The data was largely composed of passenger cars, with smaller proportions of vans, motorcycles, and agricultural vehicles. In terms of geography, most policyholders were from rural areas. Notably, the distribution of fuel types was relatively balanced between diesel and petrol, which is favorable for training classification models (Figure A2).

4. Models

4.1. Regression

We evaluated models to predict LC and HALC, addressing zero-inflated, right-skewed distributions. All models were trained under a Tweedie loss with a tuned power of 1.3, using scaled features and 5-fold cross-validation for robust evaluation. A two-stage strategy was applied: first predicting LC, then combining original features and predicted LC to estimate HALC.

- **Tweedie Regression** is a generalized linear model assuming a compound Poisson-Gamma distribution, ideal for skewed, non-negative targets. It provides interpretable coefficients but struggles with complex nonlinear patterns.
- **LightGBM** is a fast, memory-efficient boosting model that handles high-cardinality features and missing data. It captures nonlinear patterns effectively but may overfit and is less interpretable than Tweedie Regression.
- **XGBoost** is a high-accuracy ensemble model that models zero-inflated, skewed distributions and reduces overfitting through regularization. It requires more training time and offers lower interpretability.
- **Neural Networks** is a flexible, nonlinear model trained with Tweedie loss to capture complex hidden patterns. It achieves strong accuracy but requires careful tuning and lacks interpretability.

4.2. Classification

After evaluating all regression models, we shifted focus to the binary classification task of predicting Claim Status (CS), where CS = 1 if a new policyholder files a claim and 0 otherwise. The CS target was created from X.16 (total claims in the current year). To address class imbalance, we experimented with SMOTE (Synthetic Minority Over-sampling Technique); however, it biased predictions toward Class 1 on the test set, so it was excluded from the final approach. All features were scaled, and models were evaluated using 5-fold cross-validation for robustness and consistency across data splits.

- **Logistic Regression** was trained with a maximum of 1000 iterations and a fixed random state. A threshold of 0.5 was used to predict class labels.
- **XGBoost** was configured with a learning rate of 0.05, 300 estimators, a max depth of 4, and subsample and colsample rates of 0.8. The objective was set to 'binary:logistic' and log loss was used for evaluation.
- **Random Forest** used 200 estimators, a maximum depth of 6, and class weights set to 'balanced' to mitigate class imbalance.
- **Neural Networks** were built with an input layer (128 neurons, ReLU activation), two hidden layers (95 and 64 neurons, ReLU), a 30% dropout layer, and an output layer with sigmoid activation. The model was trained with the Adam optimizer (learning rate 0.0001) and evaluated using Tweedie loss and AUC.

RESULTS & DISCUSSION

To address the challenge of accurately pricing insurance policies and mitigating adverse selection, we developed models to predict loss costs and claim status.

1. Regression (LC and HALC)

Performance was evaluated using Mean Squared Error (MSE). LightGBM outperformed other models, achieving the lowest MSE for both LC (500.82) and HALC (1019.12), highlighting its ability to handle non-linear patterns and complex interactions. Tweedie Regression performed reasonably well, particularly for LC. XGBoost showed slightly higher errors, especially for HALC.

Models	LC	HALC
Tweedie Regression	502.30	1020.02
LightGBM	**500.82**	**1019.12**
XGBoost	503.78	1025.16
Neural Networks	502.45	1019.48

2. Classification (CS)

ROC-AUC was used to evaluate performance. XGBoost achieved the highest score (0.74), indicating excellent class separation. Neural Networks performed well but were limited by data size and interpretability, followed closely by Random Forest.

Models	ROC-AUC
Logistic Regression	0.71
XGBoost	**0.74**
Random Forest	0.73
Neural Networks	0.73

Visualizations of predicted vs. actual values and ROC curves (Figures A3 and A4) are used to support model comparison and evaluate classification performance.

These findings highlight the effectiveness of ensemble models in solving real-world insurance prediction problems. These models were thus deployed for final predictions on the test dataset, balancing accuracy, generalization, and computational efficiency.

3. Unexpected Outcomes and Limitations

Despite being theoretically suited for zero-inflated loss data, the Tweedie Regression underperformed compared to ensemble models—likely due to its fixed mean-variance assumption and limited ability to capture complex interactions. Neural Networks, while competitive, were constrained by limited data and lower interpretability, reducing their practical advantage.

Limitations of the study include:

- No temporal validation, which may affect performance in time-sensitive applications.
- Model interpretability remains a concern for adoption in regulated settings, where simpler models may still be favored.

4. Business Implications

The findings of this project have direct operational and strategic impact in the insurance domain. Accurate predictions of loss cost and claim status empower insurers to improve underwriting precision, enable risk-based pricing, and reduce adverse selection. LightGBM's regression performance supports more granular premium setting, while XGBoost's classification accuracy facilitates claims triaging, fraud detection, and resource optimization.

The selected models offer a scalable, data-driven foundation for real-time decision-making and can be seamlessly integrated into pricing engines and underwriting workflows. Moreover, their deployment enhances competitive advantage by enabling personalized pricing, efficient risk segmentation, and dynamic portfolio management—ultimately driving profitability, fairness, and customer retention in a highly competitive market.

CONCLUSION

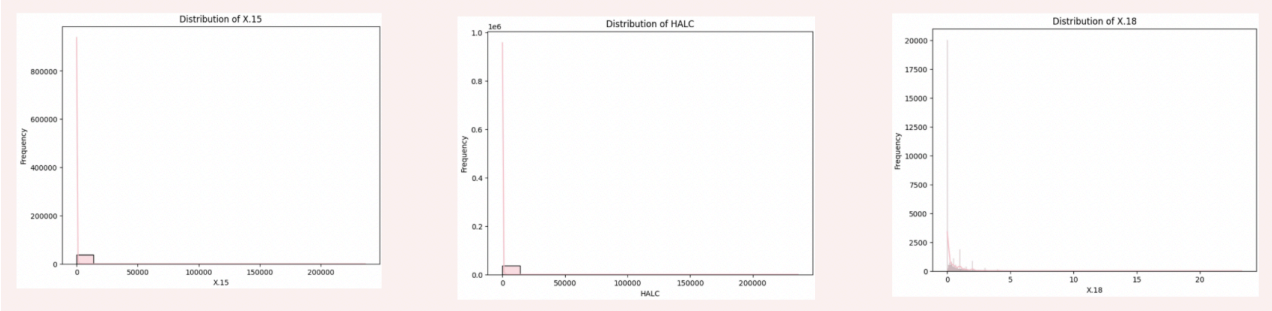
This project addressed a core insurance challenge: setting fair, risk-based premiums to avoid adverse selection and financial imbalance. Using provided data, we developed models to predict both claim losses (LC and HALC) and claim occurrence (CS).

LightGBM achieved the lowest MSE for regression tasks, while XGBoost delivered the highest ROC-AUC for classification—both outperforming traditional methods on skewed, zero-inflated data. These models were selected for deployment based on their accuracy, robustness, and ability to handle complex, non-linear relationships with minimal preprocessing. The results highlight the value of machine learning in enhancing pricing precision, risk segmentation, and overall portfolio health.

This study demonstrates the effectiveness of tree-based ensemble models (LightGBM, XGBoost) in handling complex, skewed insurance data, enabling risk-based pricing, claims forecasting, and portfolio optimization. The developed pipelines are scalable for real-time deployment and can be further enhanced through temporal modeling, external data integration, and adaptive learning to improve long-term accuracy and decision-making. By bridging predictive accuracy with business impact, this work shows how machine learning strengthens actuarial practices, helping insurers set fairer premiums, manage risk proactively, and build a more dynamic, data-driven underwriting process.

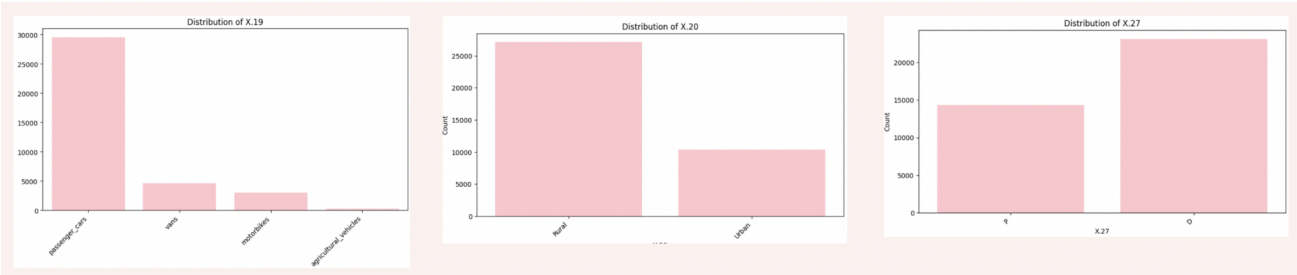
APPENDIX

Figure A1. Distributions of Key Numerical Features for Predicting LC, HALC, and CS



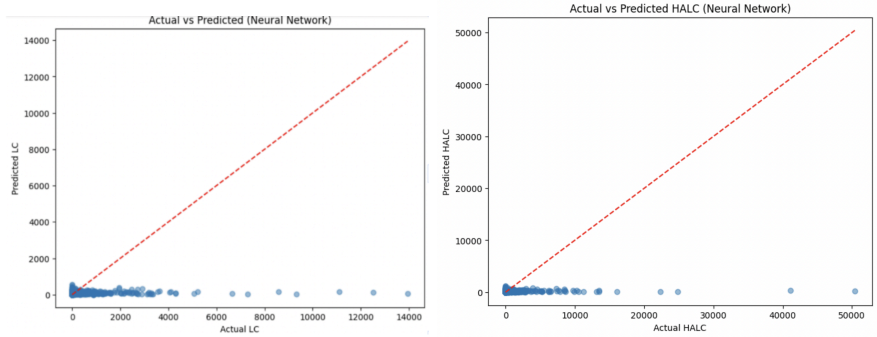
Distributions of X.15 (Total Claim Cost), HALC (X.15/X.16 × X.18), and X.18 (Claim Frequency). All exhibit heavy right skew and a sharp concentration near zero. This indicates that the majority of policyholders either do not file claims or do so infrequently, supporting the use of models designed to handle zero-inflated and skewed data.

Figure A2. Distributions of Key Categorical Features



Distributions of X.19 (Vehicle Type), X.20 (Region Type), and X.27 (Fuel Type). The dataset is primarily composed of passenger cars, rural policyholders, and diesel vehicles. Of these, X.27 shows the most balanced class distribution, which is advantageous for building effective classification models.

Figure A3. Predicted vs. Actual Values for LC and HALC



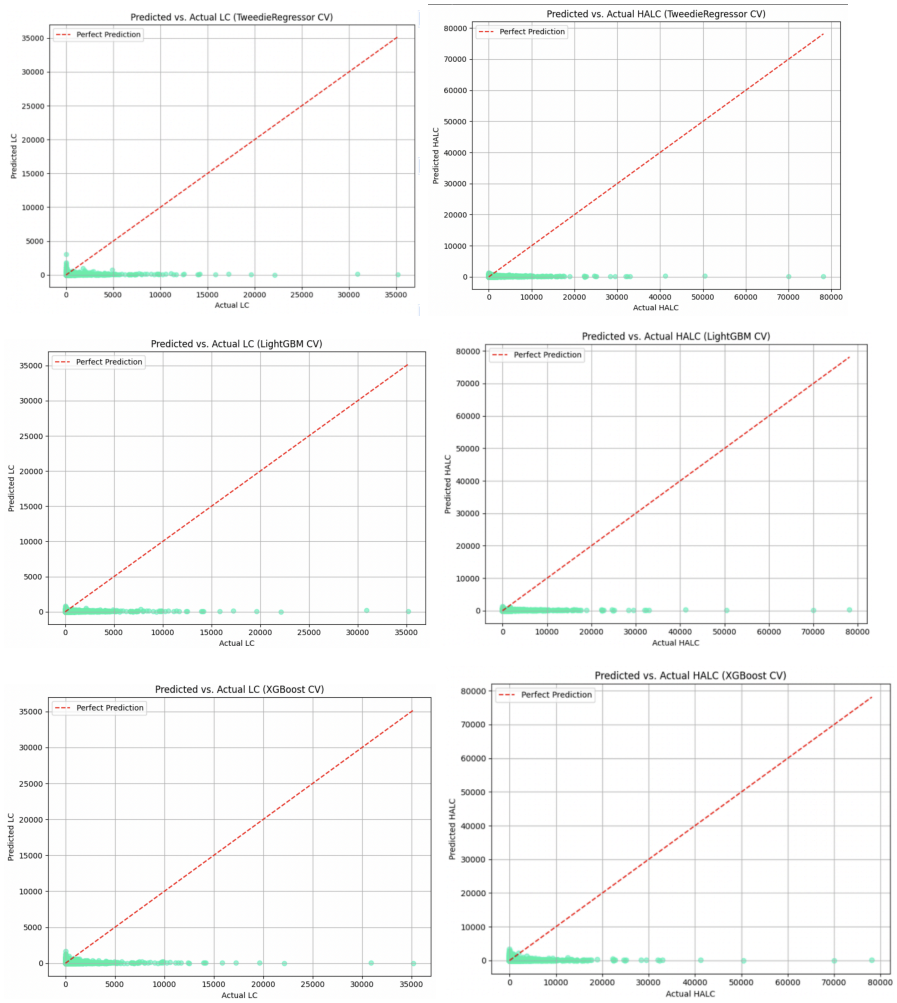
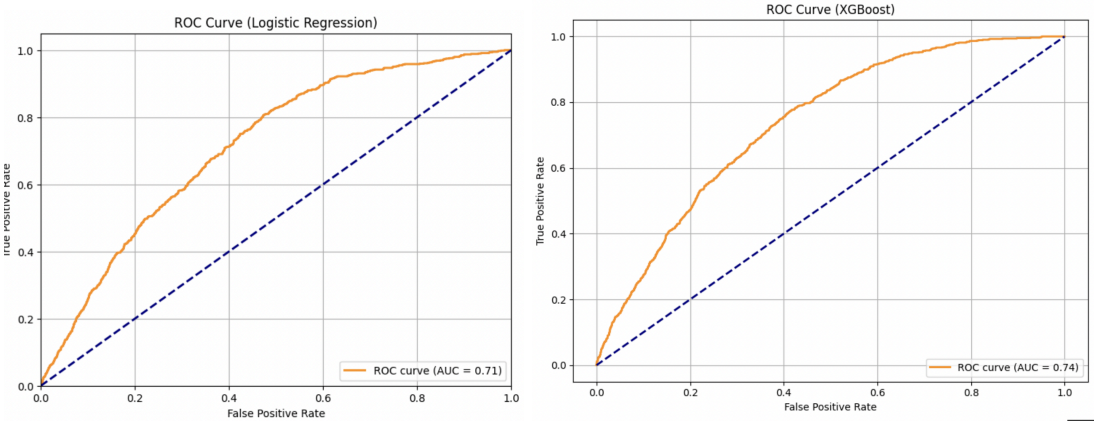
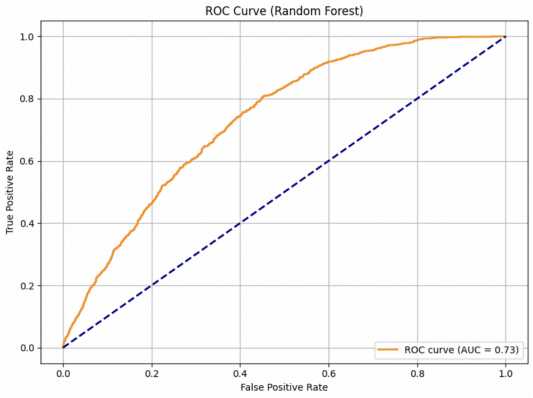
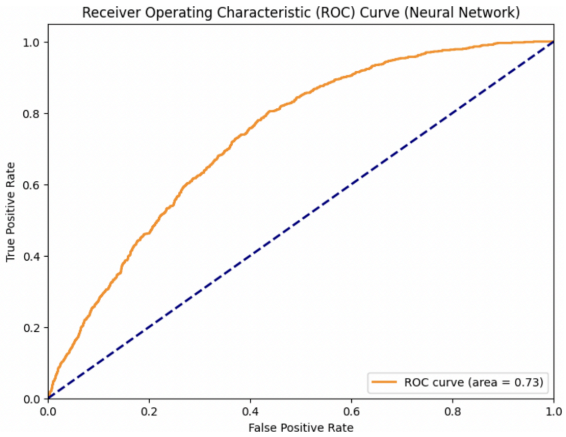


Figure A4. ROC-AUC Curves




ROC Curve (Random Forest)
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.73)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver Operating Characteristic (ROC) Curve (Neural Network)
True Positive Rate
False Positive Rate
ROC curve (area = 0.73)
0.0
0.2
0.4
0.6
0.8
1.0