Insurance Loss Analytics Report

By Group 23:
Gihyeon Kwon (8766526948), Gray Underhill (7582020489),
Heng-Yao Cheng (8985680635), Morgan Budidjaja (7748057449),
Raymond Moy (6106900274), Xingjian Shao (2451945705)

Contact Person: Morgan Budidjaja
budidjaj@usc.edu

DSO530: Applied Modern Statistical Learning Methods
Section: 20251_16219
Professor Paromita Dubey
May 8, 2025

Executive Summary

Setting accurate premiums is a critical challenge in the insurance industry. Mispricing can lead to adverse selection, where profitable, lower-risk policyholders leave for better-priced competitors while higher-risk individuals remain. This imbalance threatens financial performance and portfolio stability.

On the other hand, mispricing the premium too low will result in less profitability of the company as well as losing the financial buffer to the risk of high impact accidents. An accurate prediction of the claim losses ensures competitive pricing following National Association of Insurance Commissioners (NAIC)'s regulation on Risk-Based Capital (RBC).

This project aims to build data-driven predictive models to support more accurate and risk-sensitive pricing in auto insurance. Specifically, we forecast claim losses through two target variables: Loss Cost per Exposure Unit (LC) and Historically Adjusted Loss Cost (HALC). These continuous outcomes are evaluated using Mean Squared Error (MSE) to assess the model's performance in estimating expected claim severity.

In addition, we develop a binary classification model to predict the likelihood that a policyholder will file a claim. This probability-based risk assessment enables targeted pricing strategies and better customer segmentation. The classification performance is measured using the Receiver Operating Characteristic Area Under the Curve (ROC-AUC), reflecting the model's ability to distinguish between claim and non-claim cases.

Initially attempting machine learning models like multiple linear regression and decision trees, we soon realized limitations with these simple models. We decided to use a strong boosting model, XGBoost, which performed the best compared to other ML models including deep learning models.

For Task 2, resulting in a ROC-AUC of 0.85 for classification. The end result was after the hyperparameter tuning, ensuring the model would perform better on unseen test data. The result of Task 1 is a MSE of 410K for LC and a MSE of 1.37M for HALC on our dataset. Using a two-step approach, first classifying the profile into CS = 1 and 0, we ran a separate XGBoost Regressor model that only trained on profiles with a 1 claim status.

In future processing, we aim to improve our model's ability to generalize on unseen test data. Adding interaction terms, more advanced feature engineering depending on SHAP analysis and two step modeling for the tail risk are some of the next steps for this project.

These predictive models help insurers segment policyholders more effectively, identify high-risk profiles early, and comply with solvency standards such as the NAIC's RBC guidelines. Ultimately, this improves premium adequacy, minimizes underwriting losses, and strengthens financial resilience in an increasingly competitive market.

Data Pre-Processing Steps [1]

1. Compute the variables ‘LC’, ‘HALC’, and ‘CS’ which will be used as the response variables in different models.
    a. LC = Total Cost of Claims, Current Year / Total number of claims, current Year
    b. HALC = LC / Ratio of the number of claims filed to the total year
    c. CS = 1 if Total number of claims, current Year > 0 else 0
2. Deal with null or missing values using either 0’s, the mean, or the mode as appropriate according to the nature of the column.
3. Added dummy variables into the dataset for the relevant categorical features using one-hot encoding.
4. Transform date columns into new columns with more useful information: policy duration in days, time since last renewal in days, age of the driver, and age of the driver’s license.
5. Drop the necessary columns for the datasets that will be used to predict ‘LC’, ‘HALC’, and ‘CS’.

Task 1: LC & HALC [2]

Linear Regression

We began our modeling process with multiple linear regression to establish a performance baseline. This approach offers straightforward implementation and produces easily interpretable coefficients. However, the model exhibited poor performance on our dataset. Diagnostic checks revealed several violations of linear regression assumptions, including non-normal distribution of residuals, multicollinearity among predictors, and heteroscedasticity (non-constant variance of errors). Linear regression also assumes a linear relationship that is often unrealistic in the context of insurance claim data which is inherently skewed and influenced by complex, nonlinear interactions. Consequently, while useful for interpretation, the linear model lacked the flexibility needed to produce accurate predictions.

Tweedie XGBoost with Cross-Validation and Hyperparameter Tuning

To model the highly skewed and zero-inflated insurance claims data, we implemented a Tweedie XGBoost model, well-suited for targets like Loss Cost (LC) and HALC. Starting with a base configuration (variance power = 1.5), we performed 5-fold cross-validation and achieved a cross-validated RMSE of 742.46 at the 115th boosting round for LC.

We then conducted a randomized hyperparameter search over 150 combinations, tuning parameters such as learning rate, tree depth, regularization terms, and Tweedie variance power. The best configuration (variance power = 1.1, depth = 6, learning rate = 0.05) achieved a comparable RMSE of 754.68, confirming the model's consistency and resilience across settings.

The same process was applied to HALC, where the base model yielded an RMSE of 1446.20, and the tuned model achieved 1511.43. Although R^2 remained low due to high variance in the targets, the model consistently captured the core signal better than simpler baselines. Tweedie XGBoost struck a strong balance between predictive power and interpretability, aided by SHAP-based feature explanations.

LightGBM Tweedie with Cross-Validation and Hyperparameter Tuning

We also explored LightGBM, favored for its speed, efficiency, and scalability on structured datasets. Using a similar Tweedie objective and hyperparameter set (variance power = 1.1, learning rate = 0.04), we ran 5-fold cross-validation after standardizing features.

For LC, LightGBM reached a cross-validated RMSE of 743.13 at boosting round 128—nearly identical to XGBoost but with faster training and lower computational cost. For HALC, the model achieved a best RMSE of 1448.04, again closely mirroring XGBoost's performance. While it didn't outperform XGBoost in accuracy, LightGBM's training speed and ease of deployment make it an appealing choice for production-scale modeling.

Outlier Removal Trial

To assess model robustness, we experimented with removing outliers using a ±2.5 z-score threshold on LC (Figures 1) and HALC (Figures 2), applying this filter only to the training and validation sets. This preprocessing led to improved validation performance across models: linear regression reached 158.08 (LC) and 314.20 (HALC), random forest achieved 154.08 (LC) and 308.93 (HALC), and Tweedie XGBoost obtained 157.03 (LC) and 312.35 (HALC).

However, when tested on the full, unfiltered dataset, all models experienced significant performance drops. For example, Tweedie XGBoost's RMSE rose to 882.37 for LC and 1484.26 for HALC, revealing that trimming outliers reduced generalizability. Given the importance of capturing rare, high-loss events in insurance modeling, we ultimately trained our final models on the complete dataset to ensure real-world applicability.

Task 2: CS

Logistic Regression

For predicting the occurrence of a claim, we applied logistic regression as a baseline model due to its ease of interpretation. Running this basic model with no tuning, we obtained a 73% prediction accuracy and ROC-AUC value of 0.79 on our test dataset. This left room for plenty of improvement with more complex models and serves as a helpful point of comparison.

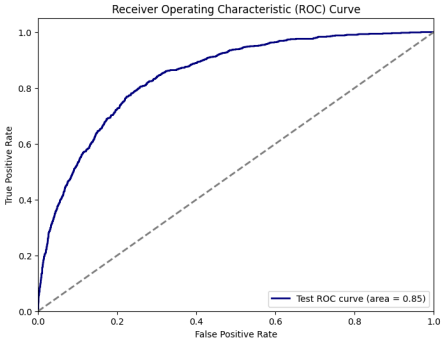
Neural Network Classification

After testing a basic logistic regression model, we moved on to a neural network classifier with two hidden layers of 128 and 64 neurons, using a sigmoid activation for the output layer. The model was trained with a binary cross-entropy loss function, and we applied dropout along with early stopping callbacks to prevent overfitting. Logistic regression provided a solid baseline, but given the complexity and non-linearity present in the data, it struggled to capture nuanced patterns. In contrast, the neural network was able to model more complex interactions between features, which led to a significant improvement. Our best ROC-AUC score increased to 0.835, making it a more effective tool for classifying claim status in this context.

Final Model for Task 2 [3]

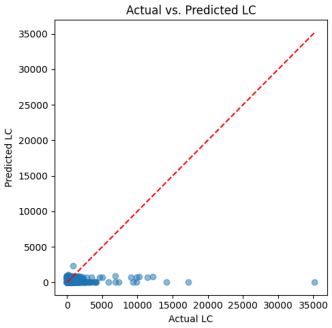
We applied XGBoost for the classification task and conducted extensive hyperparameter tuning to improve the model's predictive performance. Specifically, we tuned parameters such as learning rate, maximum tree depth, number of estimators, subsample ratio, and column

subsampling. These parameters were chosen because they directly impact the model's ability to balance bias and variance. Learning rate and tree depth help control how aggressively the model learns, while subsample and colsample_bytree introduce regularization to reduce overfitting and improve generalization.

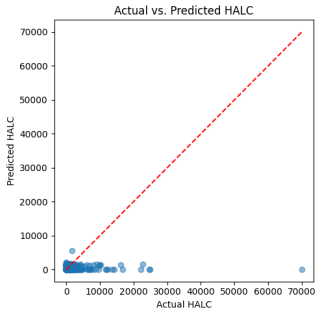


It handles imbalanced data well, supports missing values, and can capture complex nonlinear relationships with minimal preprocessing. These strengths align well with the nature of insurance data, which often includes a mix of categorical and continuous variables, missing values, and skewed distributions. For insurance companies, XGBoost offers a powerful yet practical tool for identifying policyholders who are likely to make a claim. With a final ROC-AUC score of 0.849, it outperformed all other classification models in our approaches.

Final Creative Approach for Task 1 [4]



The distributions of LC and HALC are highly skewed, with most values being zero. To address this, we implemented a **creative two-step approach.** First, we used the CS classification model to estimate the probability of a claim. Then, we trained the XGBoost regressor from Task 1 using only the subset of training data where CS equals 1, focusing on rows where a claim has occurred. After applying the classification model to the entire dataset, we used a threshold on the predicted probabilities to decide which rows should receive LC and HALC predictions. For rows below the threshold, we assumed both LC and HALC would be zero. By tuning this threshold (Figure 3), we achieved an LC RMSE of 640.87, a HALC RMSE of 1172.55, and a CS ROC-AUC of 0.849. This method avoids applying a single model across a skewed dataset dominated by zeros and outliers. Instead, it separates classification from regression, reducing error and improving the accuracy of claim-based cost predictions for insurance companies.



Additional Notes

The final CS model was tuned based on ROC-AUC, while the regressors were evaluated using MSE. Since the regression models were trained only on rows with CS = 1, zero-inflation was no longer a concern, allowing us to switch the loss function from Tweedie to squared error for better fit. While our classification model could have been further optimized for recall to capture more claim cases, we focused on ROC-AUC and MSE to align with the evaluation criteria. Initially, we also considered using clustering methods to identify groups with similar claim behavior, but opted for a supervised classification approach due to its clearer alignment with predictive

accuracy and business interpretability. This trade-off balanced classification precision with regression accuracy for overall performance in real-world insurance decision-making.

Model Interpretation [5]

Task 1 & 2 SHAP

The SHAP analyses (Figure 4) for the LC, HALC, and CS models reveal consistent patterns in the features most predictive of loss severity and claim likelihood, offering meaningful insights for underwriting and pricing decisions. Shorter time to next renewal and higher net premium were strong indicators of increased risk across all models. Features like total policies, contract age, and vehicle value helped differentiate between more stable, long-term policyholders and those with newer or higher-value contracts. The HALC model emphasized behavioral and demographic signals such as nonpayment history and age of insured, reflecting its focus on historical claim frequency. The CS model highlighted similar trends, reinforcing the role of customer behavior in predicting claims. These findings support a more personalized pricing strategy that reduces adverse selection and improves reserve planning by aligning premiums more closely with individual policyholder risk.

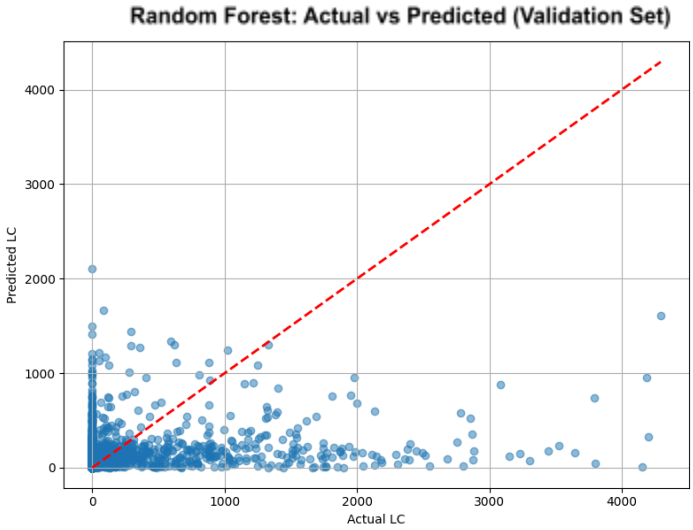
Business Implications

In the insurance industry, machine learning is used as a complementary tool to estimate the financial capital needed to reserve for covering risk. The company sums up the total estimated HALC for the upcoming year, takes the 99% quantile upper extreme values for the population, and reserves capital to cover any disastrous accidents. The Law of Large Numbers (LLM) is the core to making the prediction. Each policyholder faces random risk (accident, death, loss, etc), but aggregating the risk for a large number of sample policyholders, the total average will be closer to the real population average. However, we need to verify our model performance on the test data, to check the generalization ability of the model on unforeseen data, so we can create real-life impact with our machine learning model.

Conclusion

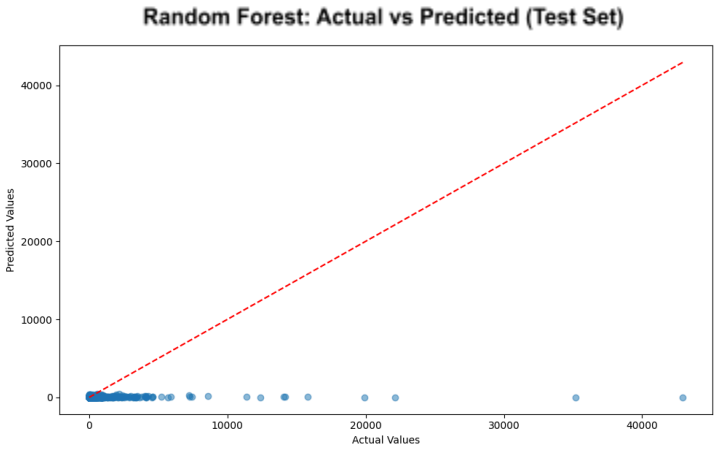
Accurately forecasting insurance claims is vital for maintaining pricing competitiveness, managing risk exposure, and complying with regulatory capital requirements. We began with baseline models such as linear regression, which offered interpretability but were limited by its inability to capture complex, nonlinear patterns from the insurance data. After transitioning to more advanced methods, Tweedie XGBoost with Cross-Validation and Hyperparameter Tuning and a classification-regression two-step model, we significantly improved predictive performance. Our final models demonstrated strong predictive power for claim likelihood and reducing loss prediction error through targeted modeling of claim-positive cases. SHAP analysis further enabled transparency, highlighting actionable drivers like renewal timing, policyholder behavior, and premium value. These insights could directly inform underwriting decisions, refine pricing strategies, and support any capital allocations. The modeling framework is scalable and adaptable, offering insurers a practical toolset for improving profitability and portfolio stability in a volatile market.

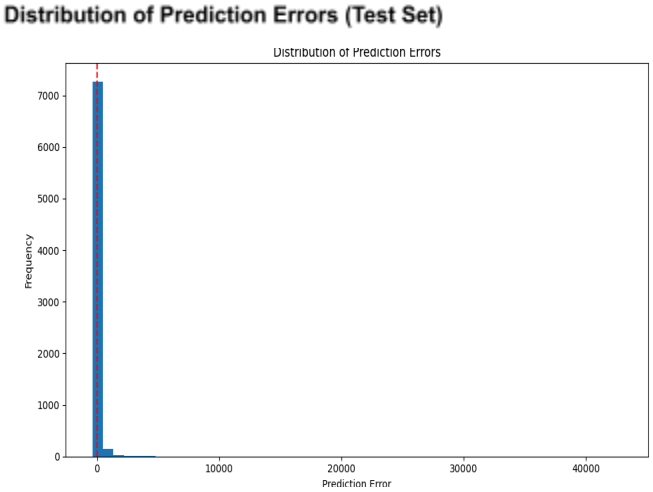
Appendix
Outlier Removal Approach (LC) (Figures 1)



On the filtered validation set, the model tends to underpredict high LC values and clusters heavily around low to mid-range predictions. While performance appears relatively controlled within the expected LC range (due to outlier removal), the scatter remains wide, suggesting variability in prediction precision even within the trimmed data.

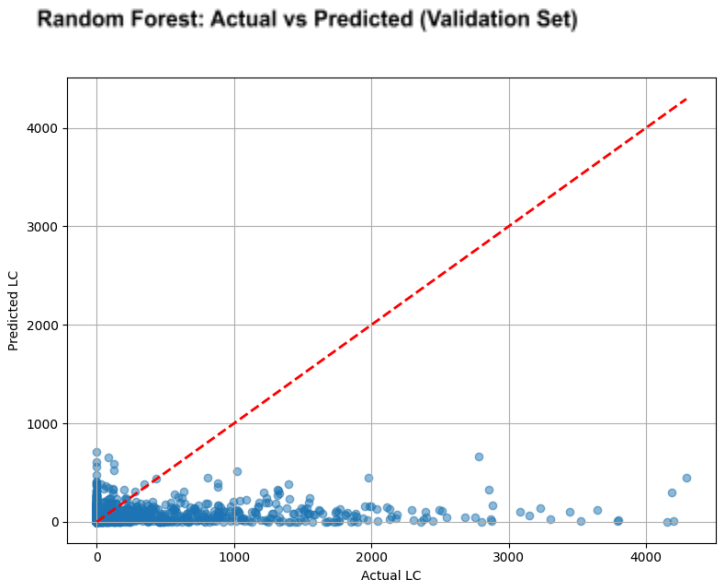
When the same model is applied to the untouched test set—which includes the full range of LC values including extreme cases—the predictive weakness becomes clear. The model fails to capture the magnitude of high LC values, with most predictions collapsing toward the lower end. This confirms the issue of model fragility when trained only on filtered data, especially in domains like insurance where outliers are common and materially important.





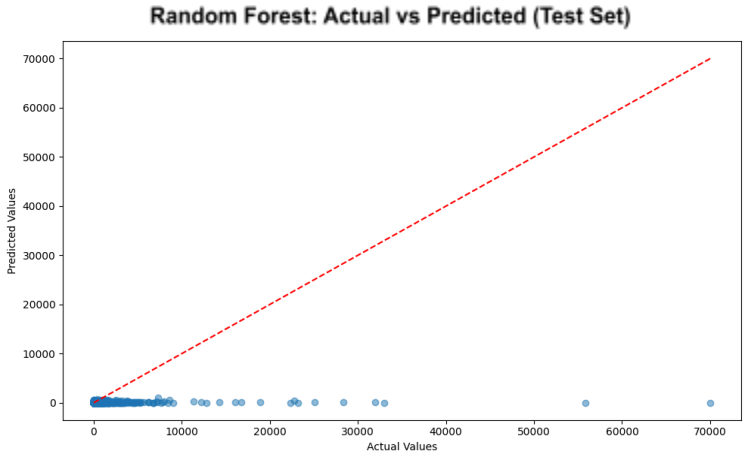
The histogram of prediction errors on the test set shows a heavy right skew. The vast majority of predictions incur small errors, but there are notable spikes in large errors—reflecting the model's inability to predict extreme losses. These few but significant underpredictions could have serious financial consequences in real-world insurance applications.

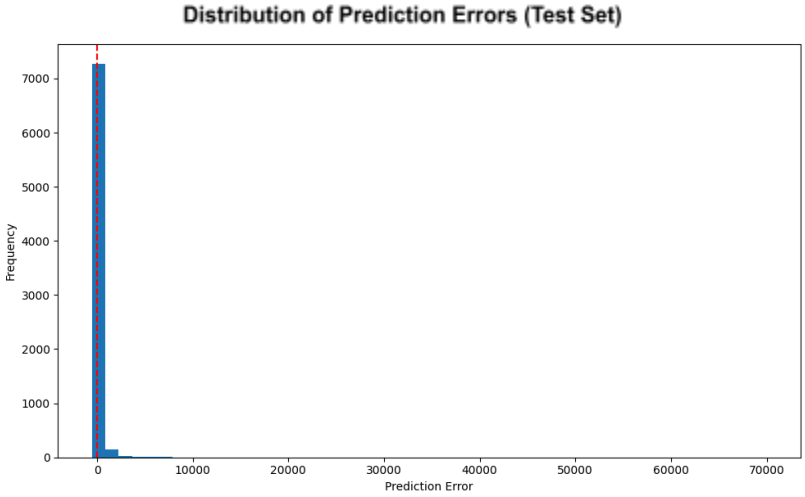
Outlier Removal Approach (HALC) (Figures 2)



On the trimmed validation set, HALC predictions mirror those of LC in their tendency to underestimate higher values, with dense clustering around lower predicted values. However, HALC exhibits even less vertical spread, suggesting the model struggles more with precision, likely because it's capturing both severity and frequency effects without fully modeling their interaction.

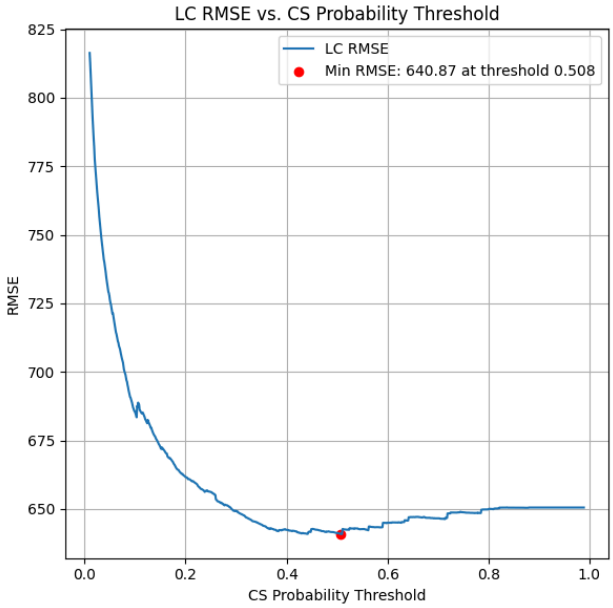
When the model is exposed to the full HALC distribution, prediction quality deteriorates notably. Many extreme actual values receive overly compressed predictions, with a visible gap between predicted and actual values for the most expensive cases. This indicates limited generalization capacity when faced with out-of-distribution or high-leverage samples.

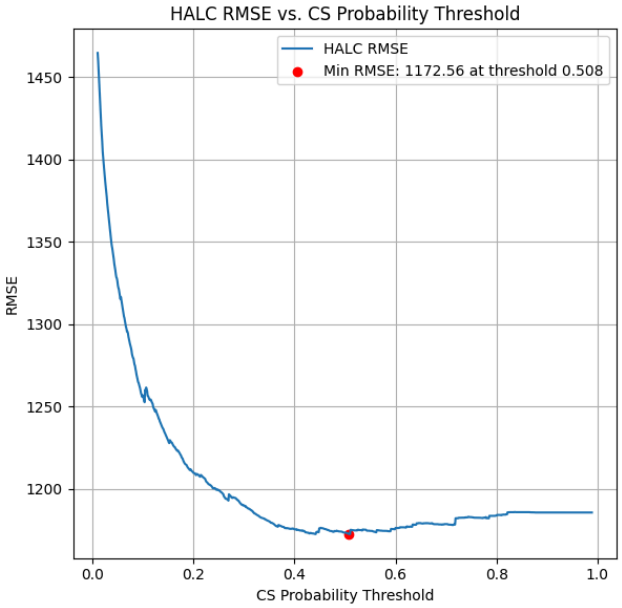




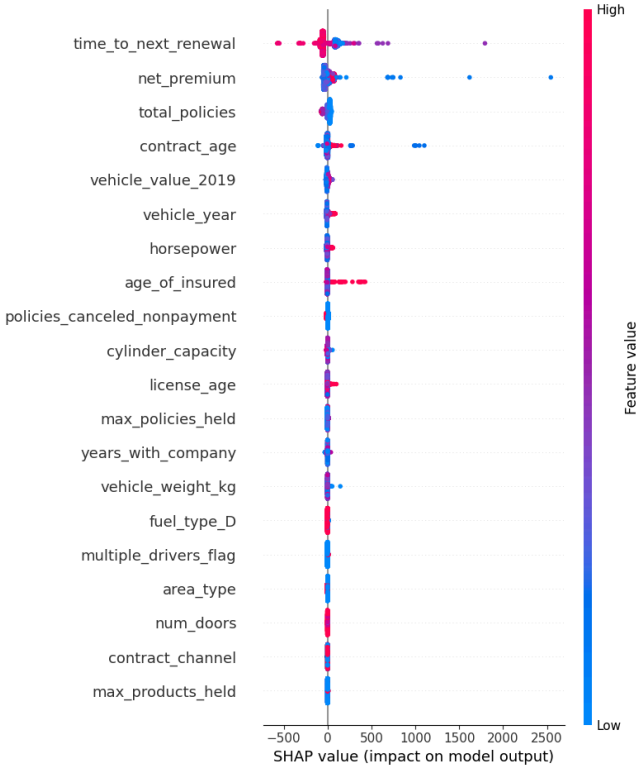
The histogram again reveals an extremely right-skewed distribution of prediction errors, though the spikes are more severe in HALC than LC. This reinforces the idea that the model is disproportionately misestimating the highest-risk policyholders, which poses a substantial challenge for pricing or reserving in actuarial contexts.

LC/HALC Threshold Tuning (Figure 3)





SHAP Task 1 (Figure 4)



SHAP Task 2