May 8th, 2025 DSO 530

Insurance Loss Analytics Group Project

Group 21

Romina Fareghbal: ID # 7422900176
Carla Francois: ID # 6212063149
Fangdi (Flora) Zhai: ID # 1906623109
Weiqi Huang: # ID 8486683680, weiqihua@usc.edu
Shahriar Rahman: ID # 9432816739
Mansour Almubaraki: ID # 5910674776

USC Marshall
School of Business

Professor Paromita Dubey

Executive Summary

Introduction

This project examines how predictive modeling can help insurance companies improve pricing accuracy and risk management. Using real-world policy data, our team built machine learning models to predict two key outcomes: how much a policyholder may cost and whether they are likely to file a claim.

The Challenge

Insurance firms often face challenges in pricing policies correctly. Underpricing can lead to financial loss, while overpricing can drive away low-risk customers. Moreover, not knowing which customers are likely to file claims can increase reserve misallocations and fraud exposure.

Our Solution

In Task 1, we engineered Loss Cost (LC) and Historically Adjusted Loss Cost (HALC) targets and tested various models. LightGBM emerged as the top-performing model, with RMSE values of 283.37 for LC and 545.27 for HALC. We used SHAP to explain predictions and identify the most influential features. In Task 2, we built a classification model to predict Claim Status (CS), using metrics like AUC, precision, recall, and F1 score. After testing logistic regression, Random Forest, Neural Network, XGBoost, and LightGBM, we again selected LightGBM based on performance (AUC = 0.76). We handled class imbalance with scaled weights and tuned the classification threshold using the F1 score. SHAP confirmed top features including time since last renewal, policy duration, and driving experience.

Findings & Business Impact

Across both tasks, time since last renewal consistently emerged as the most important predictor, reinforcing its value as a behavioral signal of customer loyalty and risk. In Task 1, HALC was influenced more by historical behavior and payment patterns, while LC was shaped by operational characteristics such as vehicle age and driving experience. In Task 2, our LightGBM classification model achieved a test AUC of 0.76, with driving experience, policy duration, and time since last renewal again ranking as the most influential features. These results support the model's ability to detect likely claimants with high recall and balanced precision.

Conclusion & Next Steps

By combining predictive performance with interpretability, this project delivers a practical, end-to-end solution for risk-based decision-making. Our approach improves pricing fairness, enhances risk visibility, and equips insurers with tools to act before claims are filed. Future steps include integrating claim probability with cost tiers to build pricing bands and expanding the approach to other insurance products.

Data Cleaning

We cleaned and transformed the raw dataset to support accurate predictive modeling. We created three target variables: Claim Status (CS) which is a binary indicator of whether a claim was filed; Loss Cost (LC) which is average cost per claim for policies with claims; and Historically Adjusted Loss Cost (HALC) which is LC scaled by historical adjustment ratios to reflect long-term risk exposure. We engineered features such as Vehicle Age (from registration year), Driver Age (from birth year), and Driving Experience (from license issue year) using 2020 as a reference year to align with market valuation as of December 31, 2019. We also created Policy Duration and Time Since Last Renewal by calculating differences between contract dates. All date fields were converted to proper datetime format. We dropped columns that were either used to derive the target variables (like X.15, X.16, X.18) or were irrelevant to modeling such as internal ID columns, exact dates after conversion. Categorical variables such as region, fuel type, and sales channel were one-hot encoded for use in machine learning models. We also validated and corrected inconsistent or illogical date values, such as negative durations or license issue dates occurring before birth dates.

To reduce the impact of extreme outliers on model training, we removed the top and bottom 0.1% of values from both LC and HALC, dropping a total of 74 rows (approximately 0.2% of the data). This step was necessary because the dataset contains a few extremely high claim costs that could skew the model and inflate RMSE. Overall, these steps helped us clean the data in a way that keeps it realistic and useful for predicting customer risk, setting better prices, and improving how claims are handled.

Task 1

To begin, we implemented a Generalized Linear Model (GLM) using the Tweedie distribution due to the zero-inflated nature of the target variables. While GLM served as a solid baseline, we subsequently tested more advanced models, including Random Forest, Gradient Boosting, LightGBM, and a simple Neural Network. Model performance was evaluated using Root Mean Squared Error (RMSE) for both LC and HALC. LightGBM consistently outperformed other models across both metrics (**Figure 1A**). We validated this result using 5-fold cross-validation, confirming LightGBM's superior performance. As a result, LightGBM was selected as our final model (**Figure 1B)**.

We further explored model improvements through feature selection and log transformation. By retaining only the top predictive features and removing less informative columns, we observed a slight improvement in RMSE. However, log transformation of the target variables did not enhance performance and was therefore excluded from the final pipeline.

Hyperparameter tuning was performed using grid search on LightGBM's key parameters. The optimal configuration—100 estimators, maximum tree depth of 7, and a learning rate of 0.05—produced the lowest RMSE (LC 283.37, HALC 545.27) on both targets and outperformed default settings.

To enhance model interpretability, we applied SHAP (Shapley Additive Explanations) to our final LightGBM models. For both LC and HALC, **time_since_last_renewal_days** emerged as the most influential feature by a significant margin. Other consistently important predictors included X.12, X.14, X.8, and policy_duration_days. (**Figure 1C**) The SHAP plots visually confirmed the stability and relevance of these variables across both outcomes, offering valuable insights for domain stakeholders.

To further validate our model's robustness, we analyzed residual patterns and prediction accuracy using diagnostic plots (**Figure 1D**). These graphs confirmed that our LightGBM model is not only accurate based on RMSE, but also well-behaved across the full range of policyholders. Most errors fall within a reasonable margin, with only a few high-cost cases showing large deviations, a known challenge in insurance modeling due to claim unpredictability.

Innovation Beyond Question - Task 1

To complement our model evaluation metrics, we created Risk Profile Cards using SHAP values to break down the prediction for individual policyholders (**Figure 1E**). These cards highlight the top three features driving the HALC prediction for each customer, enabling actionable insights for underwriters. For example, Policyholder #25 showed high predicted HALC driven by prior policy behavior and lack of renewal activity, while Policyholder #10 had lower predicted cost due to a longer customer relationship and no cancellations.

In addition, we computed a prediction summary using a ±$100 tolerance band (**Figure 1F**). The LightGBM model accurately predicted HALC within this range for 4,935 customers (approximately 66% of the validation set), with modest over and under predictions. This diagnostic supports the model's consistency and adds a real-world interpretation layer that bridges statistical performance and operational decision-making.

Task 2

The second part of the modeling project focuses on building a classification model to predict the claim status (CS) of each insurance customer. Our approach involves selecting the best-performing model, tuning its hyperparameters, and applying the final model to generate predictions on the test set. While AUC (Area Under the Curve) is used as the primary optimization metric, we also evaluate recall, precision, and F1 score, given the importance of achieving a high recall rate in the insurance industry to minimize missed claims.

We began by selecting the best model using all available features. As a baseline, we first implemented a simple logistic regression model due to its simplicity and interpretability. However, because logistic regression struggles with capturing non-linear relationships, we subsequently tested more advanced models, including Random Forest, XGBoost, Neural Network, and LightGBM (**Figure 2A**). Noticing that the dataset was imbalanced, we adjusted each model accordingly to better handle the class imbalance, increase AUC, and prevent overfit (**Figure 2B**). We then performed 5-fold cross-validation to tune the hyperparameters for each model. Among all candidates, LightGBM achieved the highest ROC-AUC score, making it the top-performing model.

For further tuning of the LightGBM model, we experimented with feature selection using SHAP values (**Figure 2C**). However, this did not lead to a significant improvement in AUC, so we decided to retain the original feature set. To better address class imbalance, we applied the scale weight parameter based on the distribution of the target classes. Finally, we optimized the classification threshold using the F1 score, aiming to strike the best balance between recall and precision.

After completing hyperparameter tuning, the final LightGBM model was configured with the following parameters: number of estimators set to 1000, learning rate set to 0.01, number of leaves set to 18, minimum data in leaf set to 100, maximum depth set to 5, and scale position weight applied. These settings were chosen to balance model complexity, learning stability, and class imbalance. The model achieved a training AUC of 0.84 and a testing AUC of 0.76. While there is a slight indication of overfitting, the difference between training and testing performance remains within an acceptable range for this classification task.

Innovation Beyond Question - Task 2

In Task 2, we innovated by experimenting with a range of advanced machine learning models, including Neural Networks, LightGBM, Random Forest, and XGBoost. This diverse model selection allowed us to compare performance across different algorithms and identify the most accurate for predicting Claim Status. By incorporating multiple models, we ensured that we captured a wide range of patterns in the data and optimized the prediction process. This comprehensive approach not only improved model accuracy but also enhanced our understanding of how different algorithms handle class imbalance and model drift. These innovations provide deeper insights into risk assessment and contribute to more robust, data-driven decision-making in the insurance industry.

In addition to algorithm experimentation, we fine-tuned the final LightGBM model by adjusting class weights to address the significant imbalance in claim outcomes, an essential step given the rarity of insurance claims. We also optimized the classification threshold using the F1 score rather than the default 0.5 cutoff, ensuring a better balance between catching actual claimants (recall) and avoiding false alarms (precision).

To interpret and validate our model's behavior, we used SHAP values to examine which features consistently influenced predictions. Time since last renewal, policy duration, and driving experience emerged as top predictors. This level of transparency was essential not only for model evaluation but also for building trust among stakeholders who would use these predictions in underwriting decisions.

Together, these enhancements allowed us to deliver a model that’s not only accurate but also explainable, scalable, and aligned with real-world insurance needs.

Conclusion

Through this project, we built two machine learning models that enhance insurers' ability to manage both the frequency and severity of claims. The loss cost prediction model enables more accurate premium pricing by quantifying expected risk at the individual level, while the claim classification model supports early detection of likely claimants, allowing for targeted interventions.

By selecting LightGBM for both tasks and applying techniques like SHAP for interpretability, hyperparameter tuning for performance, and threshold optimization for decision control, we ensured the models are both powerful and practical. Beyond predictive performance, we addressed real-world challenges such as class imbalance, feature leakage, and business relevance of input variables.

These models represent a scalable, data-driven approach to underwriting and portfolio management. They not only improve financial outcomes but also support fairer pricing, customer segmentation, and more proactive risk strategies, driving long-term business value.

Key Challenges and Model Applicability

In insurance claim prediction, accuracy takes precedence, as accurate loss cost estimates are essential for pricing, reserving, and maintaining portfolio profitability. However, interpretability remains important, especially for regulatory compliance and gaining stakeholder trust. Tools like SHAP help balance these needs by offering insight into model behavior without compromising performance.

Building predictive models for insurance claims involved addressing two major hurdles. First, class imbalance was a significant issue, as claims are relatively rare events. To prevent the model from over-predicting "no claim," we employed strategies like adjusting class weights and resampling techniques to balance the dataset. Second, feature leakage and policy drift were carefully managed. We ensured that no future information was inadvertently included in the training data, preserving the model's integrity. Additionally, we stayed vigilant to potential changes in business policies over time to avoid skewing the model's predictions.

From a business standpoint, variable selection is critical. Including too many variables (especially those that are unstable, irrelevant, or hard to justify) can reduce model reliability and stakeholder confidence. Focusing on features that are meaningful and actionable, such as **policy_duration_days, driving_experience, and time_since_last_renewal_days**, ensures that the model aligns with real-world insurance decision-making and improves maintainability over time.

This model is specifically designed for property and casualty (P&C) insurance, and therefore cannot be directly applied to life insurance. Life insurance operates on different timeframes, risk factors, and data structures. To use this model in life insurance, significant retraining with domain-specific data would be required. However, the model can inform reinsurance strategies by identifying high-risk customer segments, optimizing reinsurance treaties, and supporting broader risk management decisions.

Appendix

Task 1:

	Model	LC MSE	LC RMSE	HALC MSE	HALC RMSE
0	GLM (Tweedie)	84680.950000	291.000000	329071.410000	573.650000
1	Random Forest	88761.470000	297.930000	339787.000000	582.910000
2	Gradient Boosting	83610.060000	289.150000	324290.070000	569.460000
3	LightGBM	82999.900000	288.100000	323034.690000	568.360000
4	XGBoost	83034.960000	288.160000	324819.540000	569.930000
5	Neural Network	85931.460000	293.140000	324680.990000	569.810000

Figure 1A: Summary table for all models

	Model	LC MSE	LC RMSE	HALC MSE	HALC RMSE
3	LightGBM	80876.470000	284.190000	301875.160000	548.740000
2	Gradient Boosting	82225.260000	286.560000	302763.040000	549.490000
0	GLM (Tweedie)	83252.160000	288.270000	308180.610000	554.360000
5	Neural Network	85930.290000	292.950000	310842.530000	556.820000
1	Random Forest	87007.950000	294.840000	322730.070000	567.610000
4	XGBoost	88265.930000	296.920000	332165.420000	575.890000

Figure 1B: Summary table for all models after cross-validation

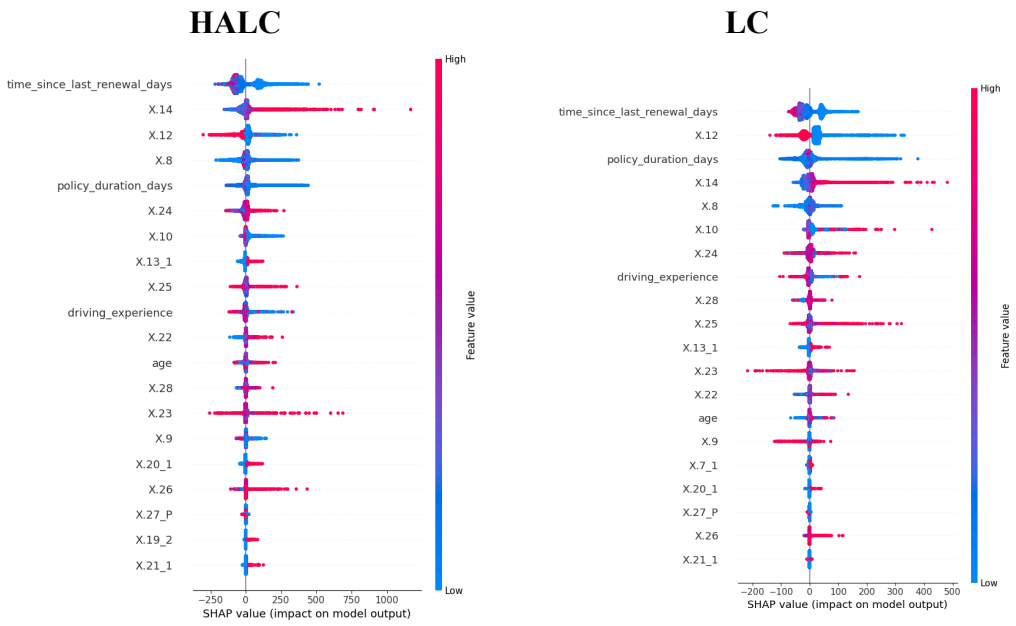

Figure 1C: SHAP - LC and HALC

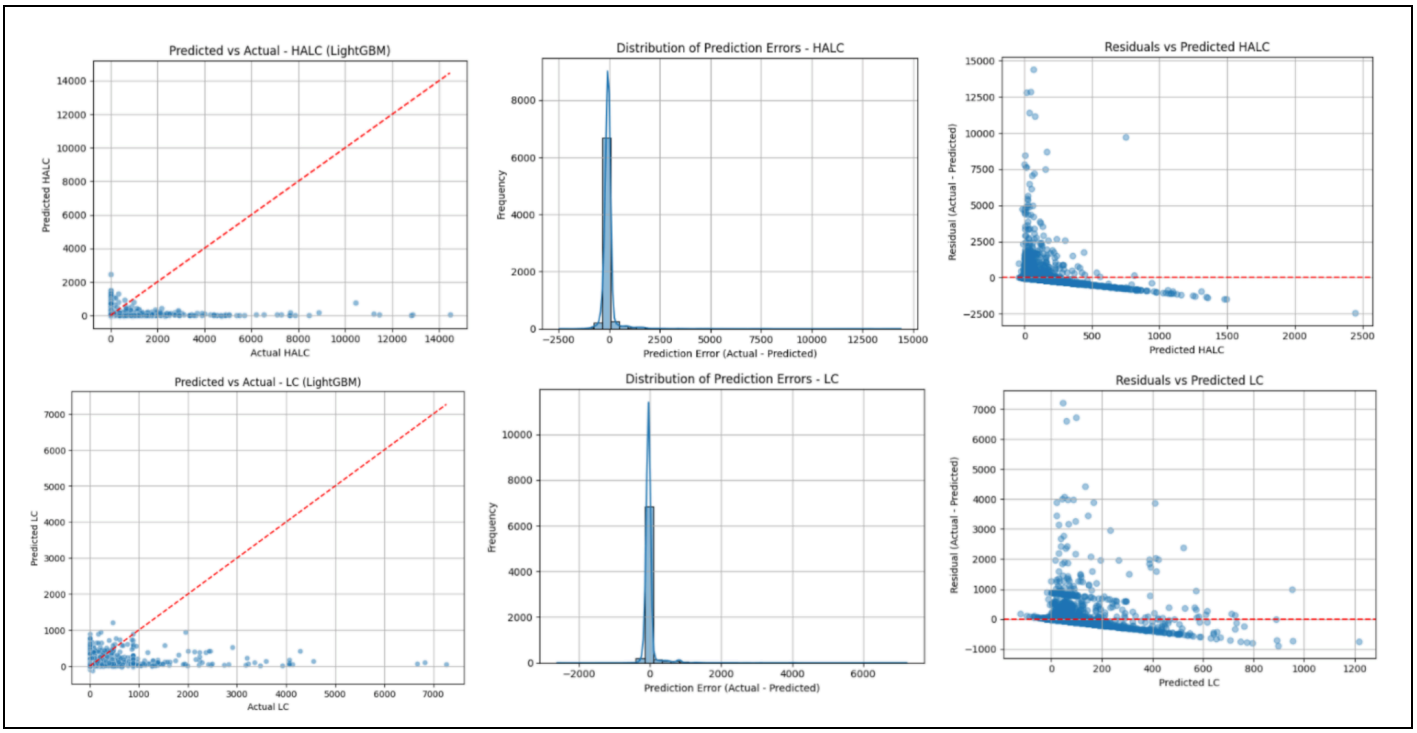

Figure 1D: LightGBM Diagnostic Plots for LC and HALC Predictions

```
========================================
Risk Profile — Policyholder #10
----------------------------------------
Predicted HALC:  $15.16
Actual HALC:     $0.00
----------------------------------------
Top Feature Drivers:
                    Feature Value  SHAP Impact
time_since_last_renewal_days  2557   -95.368409
                       X.12     0    17.411995
                        X.8    13   -10.041212
========================================

========================================
Risk Profile — Policyholder #25
----------------------------------------
Predicted HALC:  $469.11
Actual HALC:     $0.00
----------------------------------------
Top Feature Drivers:
                    Feature Value  SHAP Impact
                       X.26     6    88.446288
time_since_last_renewal_days     0    76.314466
                        X.8     1   -60.169905
========================================
```

Figure 1E: Risk Profile Cards: Customer-Level SHAP Interpretation

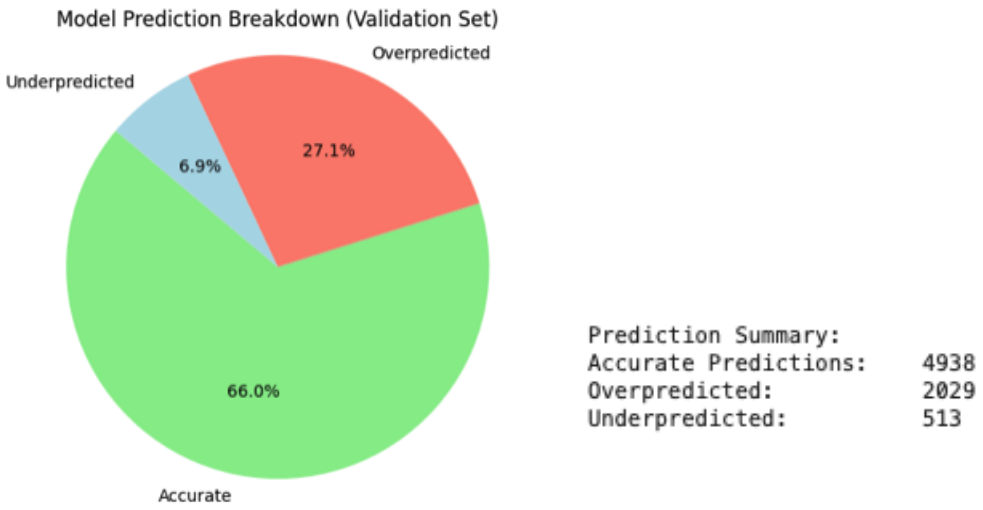

Figure 1F: Model Prediction Accuracy Breakdown (HALC Validation Set)

Task 2:

Model	**Purpose of Tuning**
Logistic Regression	- **C**: Adjusts regularization strength to prevent overfitting - **penalty**: L1 can drop useless features; L2 smooths the model
Random Forest	- **n_estimators**: Adjust to increase AUC - **max_depth**: Limits overfitting by restricting tree growth - **min_samples_leaf**: Prevents overly specific splits (overfit) - **max_features**: Add randomness to improve generalization
Neural Network	- **units / num_layers**: Control model complexity to avoid overfit and increase performance - **dropout_rate**: Regularization to reduce overfitting - **learning_rate**: Adjust to increase AUC - **batch_size, epochs**: Tune training stability
XGBoost	- **max_depth, min_child_weigh**t: Reduce model complexity lead to less overfit - **subsample, colsample_bytree**: Add randomness to improves generalization - **learning_rate**: Adjust to increase AUC - **n_estimators**: Controls boosting rounds prevent overfit and underfit
LightGBM	- **num_leaves, max_depth**: Control complexity prevents overfit. - **min_child_samples**: Stop the model from making too-specific rules - **learning_rate**: Adjust to increase AUC

Figure 2A: Hyperparameter Overview

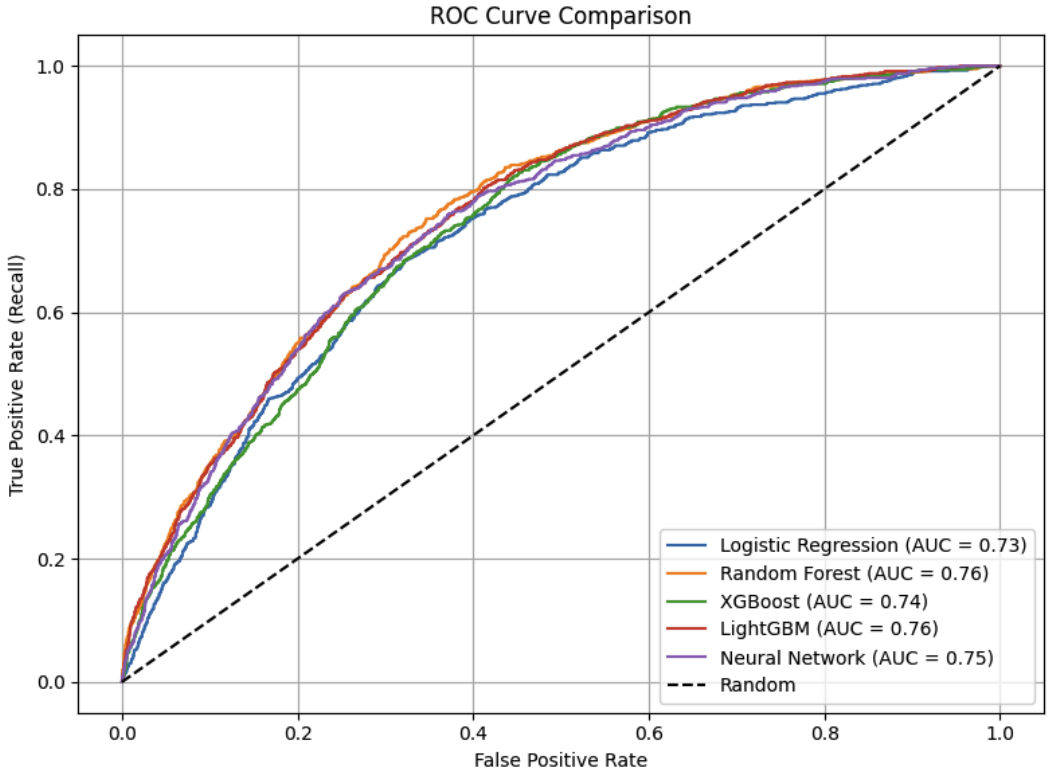

Figure 2B: ROC-AUC for all model before tune

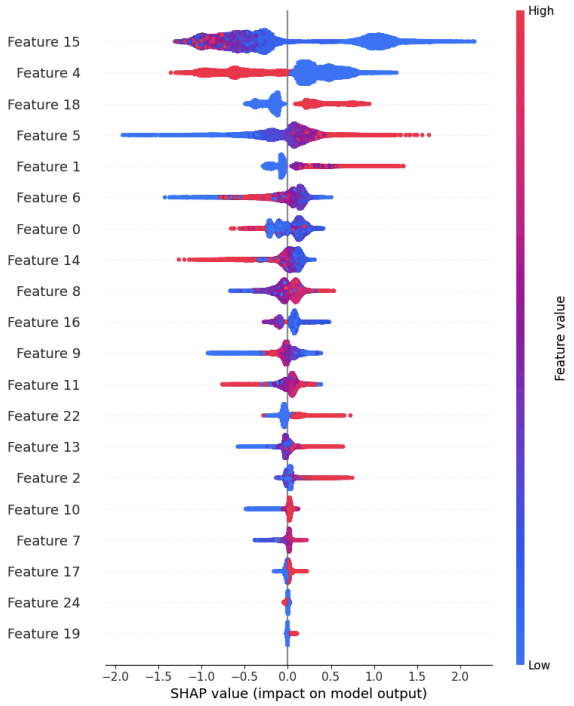

Figure 2C: SHAP - CS