Predicting Motor Insurance Policy Lapse: A Multi-Stage Pipeline Integrating Machine Learning, Survival Analysis, and Claims Behaviour

1st Karlene Philanders
Department of Computing and Information Technology
University of the West Indies
St. Augustine, Trinidad
karlene.philanders@my.uwi.edu

2nd Patrick Hosein
Department of Electrical and Computer Engineering
University of the West Indies
St. Augustine, Trinidad
patrick.hosein@uwi.edu

***Abstract*—Motor insurance lapse prediction is critical for insurer retention strategy, yet most studies treat lapse either as a static binary outcome or a survival process in isolation. This paper presents a four-stage integrated framework combining binary classification, lapse timing regression, classical survival analysis, and survival-aware machine learning to model policy lapse. We use a publicly available portfolio of 105,555 annual policy records from a Spanish non-life motor insurer, enabling full reproducibility. Random Forest and Gradient Boosting classifiers achieved ROC-AUC scores of 0.66, modestly outperforming Logistic Regression (0.61). In the survival stage, once features that mechanically encode the time axis are excluded, a Random Survival Forest attains the strongest discrimination (C-index = 0.801), narrowly ahead of a regularised Cox Proportional Hazards model (0.781), with Gradient Boosting Survival Analysis lower (0.711); the survival-aware forest and the classical Cox model are thus broadly comparable, both clearly ahead of gradient-boosted survival. Contract duration, policy seniority, and claims intensity rate emerged as the most consistent predictors across the classification and timing stages, all measured using only information available to the insurer at the renewal decision point. SHAP analysis confirmed that claims intensity elevates lapse probability among newer policyholders. These findings support proactive, tenure-aware retention strategies and show that, with leakage-free features, survival-aware machine learning and well-specified classical survival models perform comparably.**

***Keywords*—Motor Insurance, Survival Analysis, Cox Proportional Hazard, Machine Learning, Survival-Aware Machine Learning, Policy Lapse Prediction.**

I. Introduction

Policy lapse is the failure of a policyholder to renew their motor insurance contract and it represents a dual threat to insurers: immediate loss of premium revenue and erosion of long-term customer lifetime value. In competitive motor insurance markets, where contracts renew annually and switching costs are low, understanding both whether and when policyholders lapse is critical for effective retention strategy.

Existing studies predominantly treat lapse as either a static binary classification problem or a survival process examined in isolation. Claims experience is the most direct touchpoint between insurer and policyholder and is rarely examined as a temporal driver of lapse hazard within a unified modelling framework. Furthermore, direct empirical comparisons between survival-aware machine learning and classical survival methods in motor insurance remain limited.

This paper addresses these gaps with four contributions:

- A four-stage integrated pipeline combining binary classification, lapse timing regression, survival-aware machine learning, and classical survival analysis applied to a publicly available motor insurance dataset, with all features constructed to respect the information available at the prediction point.
- Empirical evidence that claims intensity rate, not raw claims volume, is the sole independent claims predictor of lapse hazard in a multivariate survival framework.
- A cross-stage feature importance synthesis confirming the consistent dominance of policy tenure (seniority and contract duration) and claims intensity across the modelling stages.
- A demonstration that naive duration-based features can manufacture a spurious advantage for classical survival models, and that once such features are removed a survival-aware Random Survival Forest and a regularised Cox model are comparable, both outperforming gradient-boosted survival analysis.

II. Related Work

Early lapse modelling relied on logistic regression and generalised linear models, identifying premium, tenure, payment frequency, and claims history as recurrent predictors of policy discontinuation [1]; Kiesenbauer [9] confirmed the influence of product type and contract age, and Jeong et al. [10] used association rule mining to surface interpretable lapse patterns. Such models are interpretable but treat lapse as a static binary event, ignoring right censoring and the temporal structure of renewal. Survival methods address this by modelling when lapse occurs: Pavlov and Mihova [2] showed lapse risk is strongly duration-dependent and clusters around renewals, and Aleandri and Eletti [3] found lapse dynamics evolve over time and are better captured with time-varying covariates. These approaches handle censoring but typically rely on the

Cox model, whose semi-parametric assumptions may not hold when covariate effects vary across renewal cycles. Machine learning has been applied to capture non-linear effects and interactions: Azzone et al. [4] and Manteigas and António [5] improved on logistic regression in life and mortgage life insurance, and Mogusu et al. [11] reported similar gains in an emerging-market setting. Class imbalance is a recurring challenge, addressed with synthetic minority oversampling [8] and cost-sensitive learning, though aggregate discrimination gains do not always translate into better minority-class detection, the operationally relevant objective. Hybrid survival-aware methods combine these strengths: tree-based and gradient-boosted survival models let hazard vary flexibly without proportional-hazards constraints [12], but tend to beat Cox mainly with strong non-proportional hazards, many covariates, and large samples. In motor insurance specifically, Spiteri and Azzopardi [6] combined random forest classification with Kaplan-Meier analysis on 72,445 policyholders, but used neither survival-aware machine learning nor a cross-stage comparison of predictors, and much of the literature reports classification and survival results separately. We address these gaps by integrating classification, regression, classical survival analysis, and survival-aware machine learning in one pipeline, with a cross-stage feature-importance synthesis, applied to motor insurance where annual renewals and frequent switching make lapse timing especially salient.

III. Methodology

The modelling framework consists of four interrelated stages. A temporal 80:20 train-test split was applied throughout, assigning earlier records to training and more recent policies to testing to preserve forecasting integrity and avoid look-ahead bias.

A. Dataset

We use the publicly available, anonymised motor vehicle insurance portfolio released by Segura-Gisbert, Lledó, and Pavía [13]. Each record represents one annual policy-year for a given policyholder, since contracts renew automatically each year until cancellation, a single policy contributes one record per renewal cycle it survives.

The overall lapse rate is 20.4%, confirming substantial class imbalance. After date parsing and deduplication to one record per policy renewal cycle, 45,962 records were used for classification modelling and 44,435 for survival modelling. Of the survival subset, 30,253 records (68.1%) were right-censored, reflecting policies still active at the end of the observation window in December 2018. Censoring was explicitly incorporated into all survival model likelihoods. Missing numeric values were imputed using median imputation.

B. Variables and Information Availability

All predictors are measured as of each policy's most recent completed renewal, the point at which an insurer would decide whether to intervene, so none draws on information that postdates that decision; the cited dataset [13] documents the underlying fields. Whether and when a policy lapsed are the prediction targets. Contract duration is the tenure accrued by that observation point rather than the time to eventual lapse, so it introduces no look-ahead, the one exception being the Stage 2 timing regression, where it overlaps the target by construction and is read descriptively.

C. Feature Engineering

Features were grouped into five categories: age and tenure variables (policyholder age, driving experience), contract duration variables (elapsed contract duration, days since and until renewal), claims behaviour variables (claims frequency, average cost, claims intensity rate), and premium and vehicle characteristics. The key engineered features are the claims intensity rate, contract duration, seniority, and annual premium. The claims intensity rate normalises historical claims volume by exposure, capturing relative claiming behaviour rather than absolute frequency, and corresponds to the source dataset's pre-computed claims-history ratio:

$$\text{Claims intensity rate} = \frac{\text{historical claim count}}{\text{contract duration (years)}} \quad (1)$$

Because both numerator and denominator are accumulated only up to the observation point, this rate is fully known to the insurer at renewal. This distinction proves central to the survival analysis findings, where exposure-adjusted intensity, rather than absolute volume, retains independent predictive significance.

D. Stage 1: Binary Classification

Three classifiers were trained: Logistic Regression as a transparent actuarial baseline, Random Forest, and Gradient Boosting. SMOTE oversampling was applied strictly within five-fold cross-validation folds to address class imbalance without inflating validation estimates. Models were evaluated on ROC-AUC and recall, with recall prioritised given the operational cost of missed lapse events. SHAP values were computed for the two ensemble models to provide directional, individual-level feature explanations.

E. Stage 2: Lapse Timing Regression

Applied exclusively to the subset of policies with confirmed lapse dates, this stage estimates the expected duration from policy inception to lapse. Three models were evaluated: Linear Regression on raw duration, log-transformed Linear Regression, and a Random Forest Regressor. Performance was assessed using MAE, RMSE, and R^2. Because elapsed contract duration and the time-to-lapse target share a common construction for lapsed policies, this stage is reported as a structural diagnostic of lapse timing rather than as an operationally deployable forecast.

F. Stage 3: Survival-Aware Machine Learning

Random Survival Forest (RSF) and Gradient Boosting Survival Analysis (GBSA) were trained on the full portfolio including censored observations. Both models capture nonlinear hazard structures without imposing proportional hazards assumptions. Performance was assessed using the concordance index (C-index).

G. Stage 4: Classical Survival Analysis

Kaplan-Meier estimation was applied to the full portfolio and to subgroups stratified by claims history, area, distribution channel, and risk type. A regularised Cox Proportional Hazards model was fitted using L1 and L2 penalisation to address multicollinearity among engineered features. Proportional hazards assumptions were validated using Schoenfeld residual tests. The Cox model specifies the hazard for policy i as

$$h_i(t) = h_0(t)\exp(\beta^\top x_i), \qquad (2)$$

where $h_0(t)$ is the baseline hazard, x_i the covariate vector for policy i, and β the coefficient vector. Discrimination was assessed using the concordance index (C-index), the proportion of comparable policy pairs whose predicted risk ordering matches their observed lapse ordering. Covariates that are deterministic functions of the analysis-time dates (elapsed contract duration and the days since and until renewal) were excluded from the survival covariate sets: for censored policies the survival time equals their sum, so their inclusion lets a model reconstruct the target rather than predict it.

H. Model Configuration

The principal hyperparameters were as follows. Logistic Regression used balanced class weights. The Random Forest classifier used 300 trees at maximum depth 8 with balanced class weights; the Gradient Boosting classifier used 300 trees, learning rate 0.05, maximum depth 3, subsample 0.8, and a minimum of 20 samples per leaf. The Random Forest regressor used 100 trees at maximum depth 20. The Random Survival Forest used 100 trees with a minimum of 10 samples per leaf and $\sqrt{p}$ features per split; Gradient Boosting Survival Analysis used 100 trees, learning rate 0.1, maximum depth 2, and subsample 0.5. The regularised Cox model used an elastic-net penalty (strength 0.1, ℓ_1 ratio 0.1). Unlisted settings used library defaults (scikit-learn; scikit-survival for the survival learners).

IV. Results

A. Binary Classification

Table I summarises classifier performance on the held-out test set. Random Forest achieved the highest ROC-AUC (0.659), marginally outperforming Gradient Boosting (0.658) and Logistic Regression (0.610), providing partial support for the hypothesis that ensemble methods outperform traditional models. However, at the default 0.50 threshold, Logistic Regression achieved the highest recall (0.299), while Gradient Boosting, despite its superior accuracy (0.801), recorded the lowest recall (0.189). This reversal reflects the tendency of gradient boosting to favour the majority class under imbalance, and highlights that ROC-AUC superiority does not translate directly to operational minority-class detection. In practical retention settings, false negatives, policyholders incorrectly classified as unlikely to lapse, are particularly costly, as they represent missed intervention opportunities and foregone premium revenue. This asymmetry justifies prioritising recall over aggregate accuracy when the operational objective is to flag at-risk policyholders for proactive contact rather than to classify the majority class correctly. SHAP analysis confirmed

TABLE I
Classification Model Performance

Model	AUC	Recall	F1	Accuracy
Logistic Regression	0.610	**0.299**	0.264	0.737
Random Forest	**0.659**	0.292	**0.282**	0.766
Gradient Boosting	0.658	0.189	0.231	**0.801**

that the historical claim count, contract duration, and seniority are the top three predictors across both ensemble models. High values of contract duration and seniority consistently produce negative SHAP contributions, reducing predicted lapse probability, while a high historical claim count produces positive contributions that are amplified among lower-seniority policyholders (Fig. 1).

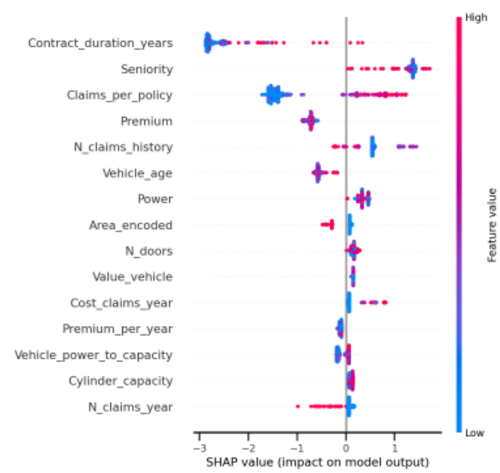

Fig. 1. Gradient Boosting SHAP beeswarm plot.

B. Lapse Timing Regression

The lapse duration distribution is strongly right-skewed, with a median of approximately 2.0 years. Linear regression on raw duration attains a near-perfect in-sample fit ($R^2 = 0.957$, MAE $= 0.287$ yr), with the log-transformed and random-forest variants lower ($R^2 = 0.787$ and 0.873). This apparent fit should be read as a caveat rather than a predictive result: for lapsed policies, elapsed contract duration and the time-to-lapse target are constructed from the same dates and therefore overlap mechanically, so this figure reflects that structural identity rather than genuine out-of-sample forecasting skill. We accordingly treat Stage 2 as a descriptive characterisation of the structural determinants of lapse timing rather than as a deployable forecasting model. Nevertheless, Stage 2 was retained because it provides behavioural insight into the temporal clustering of lapse events, identifying critical retention windows that complement the classification and survival stages. Interpreted on these terms, the clustering of predicted lapses around early renewal periods confirms that the first renewal cycle is the most critical retention window.

C. Survival Analysis Results

Table II compares concordance indices across the three survival-stage models, all fitted on the leakage-free covariate set described in Section III. The Random Survival Forest achieves the strongest discrimination (C-index $=$ 0.801), marginally ahead of the regularised Cox Proportional Hazards model (0.781), while Gradient Boosting Survival Analysis trails (0.711). Given a single temporal split without confidence intervals, the forest and the Cox model are best regarded as comparable, both clearly ahead of gradient-boosted survival. A survival-aware method thus gives the best discrimination, but its margin over a well-specified Cox model is small. An earlier specification retaining the duration-derived covariates inflated the Cox C-index to 0.835 and reversed the ranking; that apparent classical advantage vanished once those covariates were removed, showing how sensitive survival comparisons are to feature construction.

TABLE II
SURVIVAL MODEL CONCORDANCE INDEX COMPARISON

Model	C-index
Random Survival Forest	**0.801**
Cox Proportional Hazards	0.781
Gradient Boosting Survival Analysis	0.711

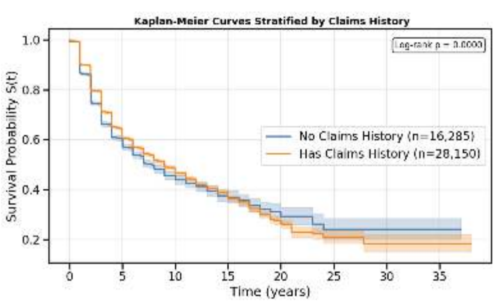

Fig. 2. Kaplan-Meier survival curves stratified by claims history.

The Kaplan-Meier analysis revealed a median policy survival time of 8.93 years, with the steepest survival decline occurring within the first five years of the insurance relationship (Fig. 2). Policyholders with prior claims history exhibited significantly lower survival probabilities throughout the observation period (log-rank $p < 0.001$), confirming that claims experience shapes survival trajectories.

In the leakage-free multivariate model (concordance 0.78; covariates standardised, so hazard ratios are per standard-deviation increase), policy seniority is the dominant predictor and strongly protective (HR $= 0.50$, $p < 0.005$), with longer driving experience mildly protective (HR $= 0.92$). Elevated hazard is associated with the distribution channel (HR $= 1.15$), exposure-adjusted premium (HR $=$ 1.26), vehicle age (HR $= 1.06$), and the number of policies held (HR $= 1.05$), all $p < 0.005$ (Fig. 3). Among the claims variables, only the exposure-adjusted claims intensity rate retains independent significance (HR $=$ 1.09, p $<$ 0.005); raw claim counts, cost, and claims-per-policy are non-significant once other covariates are controlled, indicating the claims–lapse link operates through relative intensity rather than absolute volume. Schoenfeld tests were non-significant ($p > 0.05$) for the principal predictors, supporting proportional hazards.

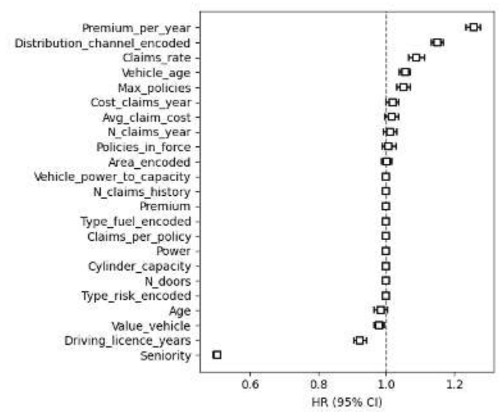

Fig. 3. Cox Proportional Hazards forest plot (fixed model). HR < 1 reduces, HR > 1 elevates lapse hazard.

D. Cross-Stage Feature Synthesis

A consistent picture emerges across the pipeline. In the classification and lapse-timing stages the strongest predictors are policy seniority, contract duration, and the historical claim count, while in the survival models seniority and the claims intensity rate are the most influential covariates. The maturity of the insurance relationship is therefore the most robust dimension of lapse behaviour, recurring as a dominant signal regardless of the analytical framework. This synthesis is qualitative: feature importance is defined differently in each model family, Gini-impurity reduction for the classifiers, gain-based scores for gradient boosting, and standardised hazard ratios for the Cox model, so the frameworks agree on which features matter even though the magnitudes are not directly comparable across models.

V. DISCUSSION

Survival-aware machine learning and Cox models are comparable once leakage is removed: With the duration-derived covariates excluded, the Random Survival Forest gives the best discrimination but leads the Cox model only marginally (0.801 vs 0.781), both clearly ahead of gradient-boosted survival. Cox remaining competitive is consistent with this dataset: the proportional hazards assumption is approximately satisfied, the sample is of moderate size, and the features are largely linear in effect, conditions under which survival-aware methods gain little over a well-specified Cox model [3]. The more important lesson is methodological: the apparent 0.10 Cox advantage in our initial specification was an artefact of covariates that encode the survival time and reversed under correction. Concordance gains should therefore be scrutinised for time-axis leakage before being attributed to model choice.

Claims rate versus raw counts: The attenuation of raw claims variables in the multivariate Cox model, with only the claims intensity rate retaining significance, indicates that the claims-lapse relationship operates primarily through relative claiming intensity rather than absolute volume. Under bonus-malus pricing systems, it is the rate of claiming relative to tenure that most directly triggers premium adjustments at renewal, providing a theoretically grounded mechanism linking claims intensity to lapse behaviour [7]. Insurers whose retention systems track raw claims counts rather than exposure-adjusted rates may therefore be monitoring the wrong signal.

Recall versus AUC trade-off: Ensemble models rank policyholders better globally (AUC), but Logistic Regression flags more actual lapsers at the default threshold (recall). Where a missed lapse costs more than an unnecessary intervention, thresholds should be calibrated to asymmetric misclassification costs rather than left at 0.50.

Practical implications: Three recommendations follow: track exposure-adjusted claims rate rather than raw counts when flagging risk; concentrate retention effort in the first five policy years, timed ahead of renewals; and prioritise for proactive contact the policyholders who combine low seniority, high claims rate, and rising premiums.

Limitations: The analysis used data from a single Spanish non-life insurer, observed between 2015 and 2018; generalisability to other markets, distribution structures, and regulatory regimes (for example, those with different bonus-malus rules or renewal conventions) therefore remains to be established. The dataset also lacked claim-type attribution (at-fault versus not-at-fault) at the portfolio level and behavioural variables such as competitor pricing, constraining causal interpretation of the claims-lapse association.

VI. Conclusion

This paper presented a four-stage framework for motor insurance lapse prediction combining binary classification, lapse timing regression, classical survival analysis, and survival-aware machine learning. Unifying these perspectives characterises not only whether policyholders lapse but also when, and how claims experience shapes that timing.

On leakage-free features, the Random Survival Forest achieved the strongest concordance (C-index = 0.801), marginally ahead of the regularised Cox Proportional Hazards model (0.781) and well ahead of gradient-boosted survival analysis (0.711). Rather than a blanket superiority of either paradigm, the result shows that a survival-aware method and a well-specified classical model are comparable at this scale and, crucially, that an earlier apparent Cox advantage was an artefact of features that encode the survival time. Matching model complexity to the data, and validating features against look-ahead leakage, matter more than the classical-versus-ML dichotomy. Claims intensity rate, rather than raw claims volume, was the independent claims predictor in the survival framework, with direct implications for how insurers should instrument retention monitoring systems. Contract duration and policy seniority were the most consistent predictors across the classification and timing stages, confirming that the maturity of the insurance relationship is a primary organising dimension of policyholder retention behaviour.

Beyond these findings, the cross-stage synthesis shows that reconciling predictor importance across frameworks yields a more internally consistent account of lapse determinants than reporting each in isolation, offering a template for other retention contexts.

Future work should add time-varying premium covariates, claim-type attribution (at-fault versus not-at-fault), and multi-insurer data to test whether survival-aware machine learning gains emerge at larger scale, and should link the lapse-timing predictions to a retention cost-benefit model to quantify the return on targeted intervention.

Acknowledgment

The dataset analysed in this study is publicly available from Mendeley Data at https://doi.org/10.17632/5cxyb5fp4f.2 and is described in detail by Segura-Gisbert et al. [13].

References

[1] M. Guillén, A. M. Perez, and M. Alcañiz, “A logistic regression approach to estimating customer profit loss due to lapses in insurance,” SSRN Working Paper, 2011.
[2] V. Pavlov and V. Mihova, “An application of survival model in insurance,” *AIP Conference Proceedings*, vol. 2025, p. 030005, 2018.
[3] M. Aleandri and A. Eletti, “Modelling dynamic lapse with survival analysis and machine learning in CPI,” *Decisions in Economics and Finance*, vol. 44, pp. 37–56, 2021.
[4] M. Azzone, E. Barucci, G. G. Moncayo, and D. Marazzina, “A machine learning model for lapse prediction in life insurance contracts,” *Expert Systems with Applications*, vol. 191, p. 116261, 2022.
[5] C. Manteigas and N. António, “Understanding and predicting lapses in mortgage life insurance using a machine learning approach,” *Expert Systems with Applications*, vol. 255, p. 124753, 2024.
[6] M. Spiteri and G. Azzopardi, “Customer churn prediction for a motor insurance company,” in *Proc. 13th Int. Conf. Digital Information Management (ICDIM)*, pp. 173–178, 2018.
[7] J. Lemaire, *Bonus-Malus Systems in Automobile Insurance*. Springer, 1995.
[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[9] D. Kiesenbauer, “Main determinants of lapse in the German life insurance industry,” *North American Actuarial Journal*, vol. 16, no. 1, pp. 52–73, 2012.
[10] H. Jeong, G. Gan, and E. A. Valdez, “Association rules for understanding policyholder lapses,” *Risks*, vol. 6, no. 3, p. 69, 2018.
[11] D. K. Mogusu, O. Ngesa, and A. Kombo, “Prediction of the likelihood of policy lapse using machine learning models: A case study of a life insurance company operating in Kenya,” *Journal of Information and Technology*, vol. 9, no. 1, pp. 105–125, 2025.
[12] S. Loisel, P. Piette, and C.-H. J. Tsai, “Applying economic measures to lapse risk management with machine learning approaches,” *ASTIN Bulletin: The Journal of the IAA*, vol. 51, no. 3, pp. 839–871, 2021.
[13] J. Segura-Gisbert, J. Lledó, and J. M. Pavía, “Dataset of an actual motor vehicle insurance portfolio,” *European Actuarial Journal*, vol. 15, no. 1, pp. 241–253, 2025. [Dataset], doi: 10.17632/5cxyb5fp4f.2.