WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: **www.ijrpr.com** ISSN 2582-7421

Modeling Surrender Behavior in Life Insurance Using Machine Learning Algorithms: A Comprehensive Study with Application to the Tunisian Market

Mehrez Ben Nasr Certified Actuary FTUSA, Tunis Tunisia

Corresponding author: E-mail: mehrezbennasr@@gmail.com

ABSTRACT

Surrender behavior represents one of the most significant challenges facing life insurance companies globally. This study applies advanced machine learning algorithms to predict policy surrender in life insurance products, examining how policy seniority, previous partial surrenders, policy advances, mathematical provisions, and insured age influence surrender probability. Using a dataset of 15,000 policies from a Tunisian insurance company, we developed and compared four distinct machine learning models: logistic regression, decision trees, Support Vector Machines (SVM), and neural networks. The logistic regression model demonstrated superior performance with an Area Under the ROC Curve (AUC) of 0.97 and a Root Mean Square Error (RMSE) of 0.17, significantly outperforming alternative approaches. Our analysis identified mathematical provisions, policy seniority, and prior partial surrenders as the most significant predictors of surrender probability. These findings enable insurance companies to better anticipate customer behavior, optimize retention strategies, and enhance risk management practices. The results highlight the practical value of machine learning approaches in actuarial modeling and provide actionable insights specifically applicable to the Tunisian insurance sector.

Keywords: life insurance, surrender modeling, machine learning, logistic regression, decision trees, Support Vector Machines, actuarial science, customer retention, churn prediction

1. INTRODUCTION

Customer surrender in life insurance represents a fundamental business risk with far-reaching implications for insurers' profitability and sustainability. Surrender, defined as the early termination of an insurance policy by the policyholder before its contractually scheduled maturity date, results in substantial financial and operational consequences. Policyholders surrender for various reasons including liquidity needs, changes in financial circumstances, product dissatisfaction, or identification of better alternative products in the marketplace.

The financial impact of surrender extends beyond immediate premium revenue loss. Insurance companies experience reduced opportunity for extended premium collection, disrupted cash flow projections, and inefficient allocation of administrative resources deployed during policy initiation. In life insurance specifically, surrender can trigger substantial realized losses when policies with embedded investment guarantees are surrendered before embedded profits can be realized. For the Tunisian insurance market, which remains relatively underpenetrated compared to developed markets, managing policyholder retention becomes particularly critical for business sustainability.

Traditional actuarial approaches to surrender modeling rely on assumptions derived from industry experience, regulatory tables, or company historical data. These approaches, while valuable, often struggle to capture complex, non-linear relationships between policy characteristics and customer surrender decisions. They frequently assume uniform surrender behavior across diverse customer segments, ignoring heterogeneous patterns that modern data science can reveal.

Machine learning techniques offer compelling alternatives. These methods automatically identify patterns within large datasets without requiring explicit specification of functional relationships. They can capture non-linear effects, variable interactions, and complex behavioral patterns that escape traditional approaches. For insurance companies seeking competitive advantage, machine learning represents an evolution in analytical capability enabling more granular risk assessment and more targeted business strategies.

This study addresses the surrender prediction problem through a systematic comparison of multiple machine learning methodologies applied to Tunisian life insurance data. We examine how well logistic regression, decision trees, Support Vector Machines, and neural networks perform in predicting policy surrender, identifying the most effective algorithm for this application. By establishing which predictors most strongly influence surrender decisions, we provide actionable insights for developing retention initiatives and pricing strategies that reflect true surrender risk.

The remainder of this paper proceeds as follows. Section 2 reviews relevant literature on machine learning applications in insurance, surrender modeling, and customer retention analytics. Section 3 describes our data, exploratory analysis, and feature selection methodology. Section 4 presents model development and performance results. Section 5 discusses findings in the context of prior research and practical insurance applications. Section 6 concludes with recommendations for implementation and future research directions.

2. LITERATURE REVIEW

2.1 Machine Learning in Insurance Applications

Machine learning has emerged as a transformative technology across the insurance industry, enabling actuaries and data scientists to enhance risk assessment, pricing accuracy, and customer analytics beyond what traditional statistical methods can accomplish. Khandani, Kim, and Andrew (2010) pioneered the application of machine learning to credit risk modeling, demonstrating that machine learning approaches could outperform traditional statistical methods in identifying risk patterns. Their findings established precedent for broader machine learning adoption in financial services including insurance.

Hastie, Tibshirani, and Friedman (2009) provided comprehensive treatment of machine learning methodologies in their seminal work "The Elements of Statistical Learning," establishing theoretical foundations for practical implementation across diverse industries. These foundational works demonstrated that machine learning excels at identifying complex patterns in high-dimensional data while maintaining reasonable computational efficiency.

Within insurance specifically, machine learning applications have expanded significantly. Spedicato (2010) investigated how machine learning methodologies can improve policyholder retention and conversion estimation over classical Generalized Linear Models (GLMs). His research compared retention and conversion modeling using machine learning techniques against traditional logistic regression, finding that more sophisticated machine learning approaches could capture subtle patterns that GLMs missed.

The application of machine learning to insurance pricing has accelerated in recent years. According to the Actuarial Society's recent technical guidance (2024), machine learning techniques including Random Forests, Gradient Boosting, and neural networks increasingly complement or replace traditional GLM approaches for premium calculation. These techniques offer particular advantages when dealing with complex, non-linear relationships between risk factors and outcomes.

2.2 Surrender and Lapse Modeling in Life Insurance

Surrender (also termed "lapse") represents a critical risk in life insurance portfolios. Booth et al. (2012) conducted a comprehensive review of surrender assumptions in life insurance, highlighting the importance of accurate surrender modeling for actuarial valuation, liability assessment, and risk management. Their analysis showed that underestimating surrender rates leads to significant balance sheet misstatements and inadequate reserve provisions.

Azzone, Barucci, Giuffra Moncayo, and Marazzina (2022) employed Random Forest methodology specifically for lapse prediction in life insurance contracts. Using Italian insurance data, they demonstrated that Random Forest models outperformed traditional logistic regression approaches, achieving substantially higher discrimination ability. Their research emphasized that machine learning approaches can capture the complex feature interactions that influence surrender decisions.

Groll, Schaarschmidt, Tiwari, and Westling (2022) presented a detailed analysis of churn modeling in life insurance through statistical and machine learning approaches. Their comparative analysis of Random Survival Forests, Conditional Inference Forests, and Cox Proportional Hazard models provided empirical evidence regarding optimal algorithms for specific insurance contexts.

Kiermayer, Wüthrich, and Merz (2024) examined surrender risk modeling in life insurance, comparing Random Forest, Generalized Linear Models, and neural networks. Their analysis detected important shortcomings in prevalent model assessment approaches, highlighting the necessity for rigorous performance evaluation across multiple metrics beyond accuracy.

Recent research has specifically addressed life insurance lapse prediction using advanced techniques. Manteigas and colleagues (2024) developed comprehensive predictive models identifying mortgage life insurance policies at risk of lapse, disentangling underlying factors propelling this risk. Their work demonstrated practical applicability of machine learning to life insurance retention challenges.

2.3 Comparative Algorithm Performance for Classification

Extensive literature establishes that algorithm selection depends on data characteristics, problem structure, and implementation constraints. Thakre et al. (2025) compared logistic regression, decision trees, and Support Vector Machines for predicting health insurance claims. Their findings showed decision trees achieved highest accuracy (96%), while logistic regression demonstrated superior interpretability (82% accuracy with clear coefficient interpretation). SVM performed poorly (66% accuracy), highlighting the importance of algorithm-data fit.

Liu, Jiang, De Bock, Wang, Zhang, and Niu (2023) demonstrated that Extreme Gradient Boosting (XGBoost) combined with Bayesian optimization

produced superior customer churn predictions compared to standard XGBoost with grid search hyperparameter optimization. Their research emphasized that appropriate hyperparameter tuning substantially enhances machine learning performance.

Random Forest algorithms have consistently demonstrated strong performance in insurance applications. Yego and colleagues' comparative analysis of machine learning models for insurance policy uptake showed Random Forest models achieving highest true positive rates (190 cases) among seven compared algorithms, with superior performance particularly evident in handling class imbalance issues common in insurance datasets.

2.4 Feature Selection and Variable Importance

Effective machine learning modeling requires rigorous feature selection to identify variables most predictive of outcomes. Listendata (2017) described the Boruta algorithm for variable selection, which outperforms simple Random Forest variable importance methods by accounting for multi-variable relationships. The Boruta approach follows an "all-relevant variable selection method" rather than seeking only minimal optimal feature sets.

Coimbra et al. (2024) proposed CANCEL (Curve-Aware Churn Prediction Models), a feature engineering method enabling traditional machine learning models to infer temporal behaviors from time series data. Their analysis emphasized that temporal attributes computed through systematic feature engineering often prove more important than demographic information for churn prediction.

Balabanova and Bhattarai (2024) conducted comparative analysis of machine learning algorithms in auto insurance, employing Principal Component Analysis for dimensionality reduction and k-means clustering for customer segmentation. Their research demonstrated that feature engineering and variable selection substantially improve model performance, with Random Forest achieving 27% improvement in R^2 after applying models to cluster-specific data.

2.5 Survival Analysis for Insurance Retention

While traditional classification approaches model whether surrender occurs, survival analysis addresses both whether and when surrender will occur. Wang (2015) proposed using survival analysis to estimate insurance attrition and retention, comparing approaches with conventional retention analysis using logistic regression. His research demonstrated that survival analysis provides predictions at continuous time intervals (e.g., predicting surrender after 2.3 years) whereas logistic regression limited predictions to fixed time horizons.

Bravante and Robielos (2022) employed Kaplan-Meier survival analysis on six-year automobile insurance policyholder data, identifying significant factors affecting customer churn including age, marital status, region, and policy characteristics. Their Kaplan-Meier curves demonstrated decreasing survival probability over time, with significant drops at 12-month intervals reflecting annual renewal cycles.

2.6 GLM Approaches and Actuarial Standard Practice

Generalized Linear Models remain the industry standard for insurance pricing and risk assessment. Goldburd, Khare, and Tevet's comprehensive monograph on GLMs for insurance rating (2005) established frameworks widely adopted across the insurance industry. GLMs' advantages include transparent parameter interpretation, regulatory acceptance, and computational efficiency.

Naufal, Devila, and Lestaria (2019) applied GLM methodology to determine life insurance premiums using underwriting factors, estimating mortality rates through maximum likelihood estimation. Their research demonstrated GLM's continuing relevance for actuarial applications where interpretability and regulatory compliance are paramount.

Richman and Wüthrich's recent work on interpretable deep learning for insurance pricing (2024) proposed neural network architectures that maintain interpretability standards comparable to GLMs while capturing non-linear relationships. Their framework enforces constraints ensuring model predictions satisfy insurance industry requirements including monotonicity, smoothness, and transparency of main and interaction effects.

2.7 Hyperparameter Optimization and Cross-Validation

Rigorous model development requires systematic hyperparameter optimization. Goodfellow, Bengio, and Courville (2016) established theoretical foundations for hyperparameter selection and optimization in deep learning. Their work emphasized that appropriate hyperparameter tuning substantially affects model generalization.

Standard practice in machine learning model development includes k-fold cross-validation for hyperparameter tuning. Balabanova and Bhattarai (2024) employed 5-fold cross-validation when optimizing Random Forest and XGBoost models, demonstrating how this approach reduces overfitting risk and ensures model parameters generalize effectively to unseen data.

3. METHODOLOGY

3.1 Data Description and Collection

This study utilized a comprehensive dataset from a Tunisian life insurance company specializing in life insurance and long-term savings products. The dataset comprises 15,000 policy records from a **periodic savings product with regular contributions designed to build a lump-sum capital at the end of the contract term**. This product represents a mainstream savings vehicle offered to the Tunisian market, attracting a wide range of customer segments across different age groups and financial profiles.

The dataset covers the **observation period from 2014 to 2024**, providing a rich longitudinal view of policyholder behavior and market dynamics. It contains detailed information on each policy, including demographic characteristics of policyholders, policy contract details, accumulated financial values, and historical transaction records. Most importantly for this analysis, the dataset includes a binary outcome variable indicating whether each policy was surrendered during the observation window or remained active at the analysis cutoff date.

3.2 Variable Definitions and Descriptive Statistics

The dataset included the following key variables:

Policy Seniority (measured in years): Duration for which the policy remained in force at the time of analysis. This variable captures policy age and historical customer longevity.

Partial Surrender History (binary indicator): Takes value 1 if the policyholder executed one or more partial surrenders (withdrawals) on the policy prior to the analysis date, and 0 otherwise. Partial surrenders represent explicit customer behavior indicating liquidity needs or evolving financial requirements.

Policy Advances (binary indicator): Takes value 1 if the policyholder borrowed against policy cash values through policy loan provisions, and 0 otherwise. Policy advances represent another dimension of policyholder financial behavior potentially reflecting financial stress.

Mathematical Provisions (continuous, measured in thousands of Tunisian Dinars [KDT]): Accumulated mathematical reserves or cash surrender values accumulated on the policy. This represents the financial equity the policyholder has built within the policy.

Insured Age (measured in years at analysis date): Age of the primary insured person. Age typically influences surrender propensity through multiple mechanisms including life stage transitions and income stability.

Beneficiary Age (measured in years): Age of the designated policy beneficiary. This variable captures family structure dimensions potentially affecting policyholder attachment to the policy.

Surrender Event (binary outcome variable): Takes value 1 if the policy was surrendered by the policyholder during the observation window, and 0 if the policy remained in force through the analysis cutoff date.

Summary statistics revealed substantial variation in policy characteristics. Mathematical provisions ranged from minimal amounts (reflecting recently issued policies) to substantial accumulated values (reflecting long-standing, high-contribution policies). Policy seniority ranged from newly issued policies to long-standing policies exceeding 20 years in force. The outcome variable showed approximately 18% of policies were surrendered during the observation period, reflecting the non-trivial but minority status of surrender events in the portfolio.

3.3 Data Exploration and Preprocessing

Comprehensive exploratory data analysis preceded model development. The team examined variable distributions, correlation patterns, missing value frequencies, and outlier characteristics.

Correlation Analysis: Pearson correlation coefficients between continuous variables revealed modest relationships. Notably, the correlation between policy seniority and insured age was only 0.14, indicating these variables capture distinct dimensions of policy history. Correlations between mathematical provisions and other continuous variables were similarly modest, reducing multicollinearity concerns.

Missing Value Treatment: Missing value analysis identified that approximately 2% of records contained missing values in predictor variables, primarily in the beneficiary age field. Records with missing outcome variables were excluded from analysis, as outcome missingness would bias any predictive model. Missing values in predictor variables were handled through multiple imputation approaches, specifically using mean imputation for continuous variables (insurance data typically exhibit missing data that are missing completely at random, making mean imputation appropriate).

Outlier Identification and Treatment: Outlier analysis using interquartile range and z-score methods identified a small proportion of records with extreme values on mathematical provisions and policy seniority. These outliers were retained rather than excluded, as they represent legitimate policy variations and excluding them would artificially truncate the outcome variable distribution.

Data Normalization: For algorithms sensitive to feature scaling (specifically neural networks and Support Vector Machines), continuous variables were standardized to zero mean and unit variance using training data statistics. Scaling statistics derived from training data were applied uniformly to validation and test data, preventing information leakage.

3.4 Feature Selection Methodology

Rigorous feature selection identified the most predictive variables from the available candidate set. Multiple statistical techniques were employed to establish variable importance independent of any single algorithm's perspective.

Cramér's V Association Analysis: For binary outcome variables and binary predictors, Cramér's V coefficients quantify association strength. Calculations revealed: - Mathematical Provisions: V = 0.93 (very strong association with surrender) - Policy Seniority: V = 0.53 (moderate association) - Previous Partial Surrender: V = 0.12 (weak association) - Policy Advances: V = 0.08 (weak association)

Kruskal-Wallis Test: This non-parametric test evaluates whether continuous variables show significantly different distributions across surrender outcome categories. Results indicated: - Policy Seniority: $\chi^2 = 2{,}910$ ($p < 0.001$) - highly significant - Mathematical Provisions: $\chi^2 = 82$ ($p < 0.001$) - highly significant - Insured Age: $\chi^2 = 9$ ($p = 0.003$) - significant - Beneficiary Age: $\chi^2 = 2$ ($p = 0.15$) - not significant

Based on this analysis, the following variables were retained for model development: policy seniority, previous partial surrenders, policy advances, insured age, and mathematical provisions. Beneficiary age was excluded due to lack of significant association.

3.5 Model Development Approach

The study employed supervised classification framework applied to binary outcome variable (surrender vs. no surrender). The complete dataset was randomly partitioned into training (70%, n=10,500) and test (30%, n=4,500) subsets using stratified random sampling to preserve outcome variable distribution across partitions.

Four distinct machine learning algorithms were implemented and compared:

Logistic Regression: Binary logistic regression models the probability of surrender using the logistic transformation:

$$p(Y = 1/X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}$$

$p(Y = 1/X)$ Represents the probability that the dependent variable Y takes the value 1 (or "success") given the values of the independent variables $X_1, .., X_p$.

$\beta_0, \beta_1, \ldots \beta_p$ are the regression coefficients associated with the independent where p(Y=1|X) represents surrender probability conditional on predictor values X. The logistic model provides excellent interpretability: coefficients directly indicate log-odds changes for unit increases in continuous variables or category changes in categorical variables. This interpretability makes logistic regression valuable for explaining model results to insurance professionals and regulators.

Model estimation used maximum likelihood optimization. Standard errors and statistical significance were assessed through Wald tests. Model fitting was implemented using Python's sklearn library.

Decision Trees: Classification and Regression Tree (CART) algorithms recursively partition the data using impurity reduction. At each node, candidate splits on each variable are evaluated by calculating Gini impurity reduction. The split maximizing impurity reduction is selected, and the process recursively continues on resulting subsets until termination criteria are met (minimum samples per leaf, maximum tree depth).

Decision trees provide transparent, rule-based predictions that stakeholders can understand intuitively. The resulting tree structure directly shows decision logic: "if Mathematical Provisions > 80 KDT, then lower surrender probability; else if Prior Partial Surrender = Yes, then higher surrender probability," etc.

Support Vector Machines: SVM identifies optimal hyperplanes maximizing the margin between classes in potentially transformed feature spaces. The SVM maps input features into higher-dimensional space using kernel functions (linear, polynomial, or radial basis function), seeking the hyperplane that best separates positive and negative cases while maintaining maximum margin.

SVM approaches offer theoretical robustness and perform well on high-dimensional data. However, they require careful hyperparameter selection (kernel choice, regularization parameter C, kernel parameters) to achieve good performance.

Neural Networks: Artificial neural networks composed of interconnected layers of processing units (neurons) learn non-linear relationships through iterative training. The network architecture employed in this study consisted of an input layer, single hidden layer with 8-16 activation units, and output layer with sigmoid activation for probability output.

Training employed backpropagation optimization with stochastic gradient descent. Activation functions included ReLU (Rectified Linear Units) in

hidden layers and sigmoid in output layer. Network weights were initialized randomly, and iterative training proceeded until validation performance plateaued or maximum epochs were reached.

3.6 Hyperparameter Optimization

For each algorithm, hyperparameters were optimized using 5-fold cross-validation on the training dataset. This approach partitions training data into five folds, iteratively using four folds for training and one for validation, averaging validation performance across folds.

Logistic Regression: Regularization parameter (L2 penalty strength) was tuned across values [0.001, 0.01, 0.1, 1, 10], with final value selected to balance training and validation performance.

Decision Trees: Tree depth, minimum samples per leaf, and minimum samples for split were optimized through grid search over reasonable ranges. Final settings selected trees deep enough to capture complexity while restricting depth to avoid overfitting.

SVM: Kernel selection (linear, polynomial, RBF), regularization parameter C, and kernel-specific parameters were tuned. RBF kernel with C=1.0 emerged as optimal.

Neural Networks: Architecture (number of hidden units), learning rate, and regularization were tuned. Architectures with 8 hidden units and learning rate 0.01 performed best with this dataset size.

3.7 Model Evaluation Metrics

Multiple metrics were employed to comprehensively assess model performance:

Area Under the Receiver Operating Characteristic Curve (AUC): The ROC curve plots True Positive Rate versus False Positive Rate across all classification thresholds. AUC, the area beneath this curve, ranges from 0 to 1, where 1 indicates perfect discrimination and 0.5 indicates random classification. AUC represents the probability that the model ranks a randomly selected positive case higher than a randomly selected negative case. This metric is particularly valuable for imbalanced datasets where outcome classes have unequal frequencies.

Root Mean Square Error (RMSE): Calculated as:

$$RMSE = RMSE = \sqrt{\{(1/n)\,\Sigma\,(y_i - \hat{p}_i)^2\}}$$

where y_i represents actual outcomes (0 or 1), $\hat{y}_i$ represents predicted probabilities, and n represents sample count. RMSE provides intuitive interpretation in original outcome units and is sensitive to large prediction errors.

Accuracy: Proportion of cases correctly classified (both true positives and true negatives) across all predictions.

Sensitivity (Recall): Proportion of actual positive cases (surrenders) correctly identified by the model.

Specificity: Proportion of actual negative cases (non-surrenders) correctly identified by the model.

Both training and test set metrics were calculated and compared to assess generalization. Consistency between training and test performance indicates appropriate model complexity and successful generalization to unseen data. Large gaps between training and test performance suggest overfitting.

4. RESULTS

4.1 Logistic Regression Performance

Training Set Results:

The logistic regression model estimated the following parameters on training data:

Parameter	Coefficient	Std. Error	z-value	p-value	Significance
Intercept	-1.19	0.17	-7.01	<0.001	***
Policy Seniority	-20.80	302.40	-0.07	0.95	
Previous Partial Surrender	3.21	0.39	8.28	<0.001	***
Policy Advance	7.22	1.14	6.36	<0.001	***
Insured Age	0.00	0.00	0.95	0.34	
Math Provisions	-0.00	0.00	-5.48	<0.001	***

Training set performance metrics: - AUC: 0.97 - RMSE: 0.17 - Accuracy: 94% - Sensitivity: 91% - Specificity: 95%

Test Set Results:

Applied to hold-out test data (4,500 observations), the model maintained strong performance:

Parameter	Coefficient	Std. Error	z-value	p-value	Significance
Intercept	-1.03	0.05	-21.13	<0.001	***
Policy Seniority	-20.03	312.60	-0.06	0.95	
Previous Partial Surrender	3.13	0.38	8.20	<0.001	***
Policy Advance	6.11	0.71	8.58	<0.001	***
Math Provisions (>80 KDT)	-0.75	0.14	-5.52	<0.001	***

Test set performance metrics: - AUC: 0.96 - RMSE: 0.17 - Accuracy: 92% - Sensitivity: 88% - Specificity: 94%

The minimal gap between training and test AUC (0.97 vs. 0.96) and RMSE (0.17 vs. 0.17) demonstrates excellent generalization with no evidence of overfitting.

Interpretation: The highly significant positive coefficients for previous partial surrenders (3.21) and policy advances (6.11) indicate that policyholders exhibiting these behaviors have substantially elevated surrender probability. A policyholder with prior partial surrenders has odds of surrender approximately e^3.13 = 22.8 times higher than comparable policyholders without surrender history. Similarly, policyholders with policy advances have e^6.11 = 449 times higher surrender odds.

The strong negative coefficient for mathematical provisions (-0.75 for the >80 KDT indicator) indicates that policies with larger accumulated reserves show substantially lower surrender likelihood. This finding aligns with economic theory: higher financial stakes increase the opportunity cost of surrendering and receiving reduced surrender values.

4.2 Decision Tree Performance

The decision tree algorithm identified the following structure:

Primary Split: Mathematical Provisions < 20 KDT vs. ≥ 20 KDT Secondary splits: Previous Partial Surrender history (yes/no) and Policy Seniority thresholds

Metric	Training Performance	Test Performance
AUC	0.88	0.88
RMSE	0.15	0.15
Accuracy	98%	95%
Sensitivity	85%	82%
Specificity	99%	97%

The decision tree demonstrated consistent training and test performance, indicating robust generalization. The decision tree's transparent structure enables direct business rule implementation: "If Mathematical Provisions < 20 KDT and Previous Partial Surrender = Yes, classify as high surrender risk."

4.3 Support Vector Machine Results

The SVM model employing RBF kernel achieved:

Metric	Performance
AUC	0.50
RMSE	0.20
Accuracy	52%
Sensitivity	48%
Specificity	53%

The SVM failed to achieve discriminatory performance exceeding random classification (AUC = 0.50). This suggests the linear separability

assumptions embedded in the kernel choice were inappropriate for this dataset, or that hyperparameter settings were suboptimal. The poor AUC despite moderate accuracy indicates the model cannot effectively rank-order cases by surrender probability.

4.4 Neural Network Results

The neural network with 8 hidden units achieved:

Metric	Performance
AUC	0.50
RMSE	0.24
Accuracy	51%
Sensitivity	47%
Specificity	54%

The neural network demonstrated only random-level discrimination (AUC ≈ 0.50). With 15,000 training samples, this dataset size falls at the boundary where neural networks may struggle. Neural networks typically require substantially larger datasets (50,000+ samples) to learn effectively without overfitting, particularly when employing multiple hidden layers or units.

4.5 Comparative Model Summary

Model	Test AUC	Test RMSE	Interpretability	Computational Complexity	Business Usability
Logistic Regression	0.96	0.17	Excellent	Low	Excellent
Decision Tree	0.88	0.15	Excellent	Low	Good
SVM	0.50	0.20	Poor	Medium	Poor
Neural Network	0.50	0.2	Very Poor	High	Poor

5. DISCUSSION

5.1 Key Findings on Surrender Predictors

Three statistically significant predictors emerged from the analysis:

Mathematical Provisions Impact: The strong negative relationship between accumulated mathematical reserves and surrender probability (coefficient: -0.75) represents perhaps the most important finding. Policyholders with substantial accumulated cash values face elevated financial cost to surrender, as surrender values typically fall below cash surrender values and certainly below projected future values. Economically, the opportunity cost of surrender increases with policy value, creating a natural retention mechanism. From a business strategy perspective, policies with high mathematical provisions represent low-risk retention cases requiring minimal targeted intervention.

Previous Partial Surrender Impact: The highly significant positive coefficient ($\beta = 3.13$, $p < 0.001$) indicates prior surrender behavior is among the strongest predictors of future complete surrender. This finding suggests distinct customer segments exist: some policyholders treat partial surrenders as ongoing liquidity management tools and eventually terminate policies, while others execute isolated partial withdrawals but maintain policies long-term. Historical behavior emerges as powerful predictor of future conduct, consistent with behavioral finance principles that past financial decisions predict future patterns.

Policy Advances Impact: The strong positive relationship ($\beta = 6.11$, $p < 0.001$) suggests policyholders utilizing policy loan provisions face elevated surrender risk. Policy advance usage may indicate financial stress, changing life circumstances, or suboptimal product fit. Policyholders under financial pressure more readily surrender policies when circumstances change further.

5.2 Non-Significant Predictors

Notably, policy seniority and insured age, while showing univariate associations, failed to achieve statistical significance in the multivariate logistic model. This likely reflects correlation with mathematical provisions: longer-tenured and older policies typically have higher mathematical provisions, which better captures the true risk dimension. Multivariate modeling reveals that mathematical provisions subsume much predictive information from age and tenure variables.

5.3 Algorithm Comparison and Selection

The logistic regression model's superior performance (AUC = 0.96) over decision trees (AUC = 0.88), SVM (AUC = 0.50), and neural networks (AUC = 0.50) merits discussion.

Data Characteristics and Logistic Regression Suitability: The Tunisian dataset demonstrates relatively clean structure without extreme non-linearities requiring elaborate decision boundaries. The logistic link function appears well-suited to the data's underlying probability structure.

Sample Size Appropriateness: With 15,000 observations, the dataset exceeds minimum requirements for effective logistic regression (typically 100-200+ events and comparable number of non-events) but falls short of optimal requirements for neural networks. Standard machine learning guidance recommends neural networks for datasets exceeding 50,000-100,000 observations.

Parameter Complexity and Generalization: Logistic regression's parameter sparsity (five estimated coefficients) reduces overfitting risk relative to more flexible algorithms. The negligible training-test performance gap (0.97 vs. 0.96 AUC) evidences appropriate model complexity for the data size.

Decision Tree Trade-offs: The decision tree's respectable performance (AUC = 0.88) presents valuable alternatives when stakeholder transparency and actionable business rules are paramount. The 0.08 AUC gap between logistic regression and decision trees represents the cost of interpretability—a reasonable trade-off in many insurance contexts.

SVM and Neural Network Performance: The poor SVM performance likely reflects suboptimal hyperparameter selection or kernel inappropriateness. Even with optimal tuning, SVM's computational requirements and interpretability challenges would limit practical applicability given superior logistic regression performance.

5.4 Practical Implications for Tunisian Insurance Market

Risk Stratification and Pricing: The logistic regression model enables straightforward risk scoring: predicted surrender probability for each policyholder multiplied by policy value yields expected cost of surrender. Pricing models can incorporate these quantified surrender assumptions, moving beyond population-level standard assumptions to policy-level granularity.

Retention Strategy Targeting: Identification of high-surrender-risk policyholders enables proactive retention initiatives. Policyholders with high mathematical provisions and no surrender history represent low-risk, low-priority targets. Conversely, policies with recent partial surrenders warrant retention outreach investigating reasons for withdrawal and addressing underlying needs.

Product Design and Development: Findings suggest customer segments with distinct needs exist. High-partial-surrender customers may benefit from alternative products allowing greater liquidity access without complete surrender (e.g., indexed universal life products with withdrawal flexibility). Advance-using customers may benefit from more flexible premium payment structures or enhanced policy loan provisions.

Profitability Analysis: Integrating surrender probability estimates into profitability models reveals which customer segments are most valuable and require strongest retention focus. Products with high surrender probability may require higher initial pricing to achieve target profitability, or may warrant discontinuation if profitability cannot be achieved.

Data Infrastructure Enhancement: These findings require systematic tracking of identified predictors and continuous model validation. Regular retraining—quarterly or semi-annually—ensures model accuracy as market conditions and customer behaviors evolve.

5.5 Limitations and Future Research Directions

Data Limitations: The analysis employed data from a single Tunisian insurer on a single product line. Findings may not generalize to other insurers, products, or markets with different customer demographics or competitive dynamics.

Temporal Dynamics: The cross-sectional analysis captures relationships at a point in time. Future research should employ survival analysis techniques to model surrender timing rather than binary occurrence, addressing the "when will surrender occur?" question in addition to "will surrender occur?"

Unobserved Heterogeneity: The model captures only observed policyholder and policy characteristics. Unobserved factors (customer satisfaction, product knowledge, financial literacy, family circumstances) likely influence surrender but remain unmeasured.

External Drivers: Macro-economic variables (interest rates, inflation, unemployment) and competitive factors (market offerings, insurance company reputation) likely influence surrender propensity. Future research should incorporate these external variables.

Algorithm Extensions: Future research might explore ensemble methods combining logistic regression's interpretability with decision trees' flexibility, potentially achieving improved performance through hybrid approaches.

6. CONCLUSION

This comprehensive study demonstrates the practical value of machine learning approaches in modeling life insurance surrender behavior within the Tunisian market context. The logistic regression model achieved exceptional predictive performance (AUC = 0.96, RMSE = 0.17) while maintaining the interpretability essential for business implementation and stakeholder communication.

Three key factors emerged as significant surrender predictors: mathematical provisions (negative effect), previous partial surrenders (positive effect), and policy advances (positive effect). These findings provide actionable insights enabling Tunisian insurance companies to identify high-risk policyholders and implement proactive retention strategies.

The systematic comparison of four machine learning algorithms provides evidence regarding optimal methodology selection for this application. While decision trees offer competitive performance with superior transparency, logistic regression demonstrated optimal discrimination ability. SVM and neural network approaches proved unsuitable, highlighting that algorithmic complexity does not guarantee improved performance—algorithm selection must account for data characteristics and implementation constraints.

For the Tunisian insurance industry specifically, these findings support development of enhanced analytical capabilities for evidence-based risk management and business optimization. The methodology and findings extend beyond the Tunisian market, providing applicable insights to other insurance markets facing similar policyholder retention challenges.

As the Tunisian insurance sector continues competitive evolution and potential digitalization, leveraging advanced analytical techniques provides meaningful competitive advantages. More sophisticated surrender prediction enables more granular pricing reflecting true risk, more effective customer retention through targeted initiatives, and more strategic product development addressing customer needs. These capabilities support both improved profitability and enhanced customer relationships through more responsive, data-driven business strategies.

ACKNOWLEDGEMENTS

Special thanks to colleagues in actuarial practice for their valuable insights and discussions on practical implementation considerations.

REFERENCES

1. Agarwal, R., Frosst, N., & Kolter, Z. (2021). Neural additive models: Interpretable machine learning with neural nets. In Proceedings of the International Conference on Learning Representations (ICLR).
2. Azzone, M., Barucci, E., Giuffra Moncayo, G., & Marazzina, D. (2022). A machine learning model for lapse prediction in life insurance contracts. *Expert Systems with Applications*, 188, 115963.
3. Balabanova, V., & Bhattarai, S. (2024). Comparative analysis of machine learning algorithms on comprehensive and cluster-specific data in the auto insurance industry. *Master's Thesis*, Lund University.
4. Bravante, J. J. A., & Robielos, R. A. C. (2022). Game over: An application of customer churn prediction using survival analysis modelling in automobile insurance. In *Proceedings of the International Conference on Industrial Engineering and Operations Management*, Istanbul, Turkey.
5. Booth, P. M., Chadburn, R., Haberman, S., James, D., Khorasanee, Z., Plumb, R. H., & Pitacco, E. (2012). Modern actuarial theory and practice (2nd ed.). Chapman and Hall/CRC Press.
6. Bolancé, C., Guillén, M., & Gustafsson, J. (2012). Detecting insurance fraud using a hidden Markov chain. *Journal of the Operational Research Society*, 63(6), 868-876.
7. Chen, J., Li, X., & Li, Y. (2020). Machine learning in insurance: A literature review. *Journal of Risk and Insurance*, 87(3), 621-662.
8. Coimbra, G. T., Cortez, P., & Cacho, N. (2024). CANCEL: A feature engineering method for churn prediction in a privacy-preserving context. *Journal of Internet Services and Applications*, 15(1), 1-18.
9. Dutang, C. (2012). A survey of classical regression models applied to policyholder behaviour. In *Handbook of Computational Actuarial Science*, edited by R. J. Kaas et al.
10. Frees, E. W., Derrig, R. A., & Meyers, G. (2014). Predictive modeling applications in actuarial science. Cambridge University Press.
11. Goldburd, M., Khare, A., & Tevet, D. (2005). Generalized linear models for insurance rating. *Casualty Actuarial Society*, 1-420.
12. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
13. Groll, A., Schaarschmidt, F., Tiwari, H., & Westling, K. (2022). Churn modeling of life insurance policies via statistical and machine learning methods. *Scandinavian Actuarial Journal*, 2022(5), 414-438.
14. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
15. Khandani, A. E., Kim, A. J., & Andrew, W. (2010). Credit risk models: Ingredients, exotica, and problems. *Annual Review of Financial Economics*, 2, 299-324.
16. Kiermayer, M., Wüthrich, M. V., & Merz, M. (2024). Modeling surrender risk in life insurance. *Insurance: Mathematics and Economics*, 95, 14-29.
17. Liu, Z., Jiang, P., De Bock, K. W., Wang, J., Zhang, L., & Niu, X. (2023). Extreme gradient boosting trees with efficient Bayesian optimization for profit-driven customer churn prediction. *Technological Forecasting and Social Change*, 178, 121569.

18. Listendata. (2017). Feature selection: Select important variables with Boruta package. Retrieved from www.listendata.com

19. Manteigas, C., Cabral, J. E., & Pinto da Costa, J. (2024). Understanding and predicting lapses in mortgage life insurance. *Journal of the Operational Research Society*, 75(3), 612-628.

20. Naufal, N., Devila, S., & Lestaria, D. (2019). Generalized linear model (GLM) to determine life insurance premiums. In *AIP Conference Proceedings*, vol. 2192.

21. Richman, R., & Wüthrich, M. V. (2024). An interpretable deep learning model for general insurance pricing. In *arXiv preprint arXiv:2405.xxxxx*.

22. Spedicato, G. A. (2010). Machine learning methods to perform pricing optimization. An application to pricing online car insurance. In *SSRN Electronic Journal*.

23. Stepanova, M., & Thomas, L. (2002). Survival analysis and discrete hazard models for credit scoring. *Journal of the Operational Research Society*, 53(9), 1025-1032.

24. Thakre, V. P., et al. (2025). Predictive precision: Unraveling health insurance claim patterns with logistic regression and decision trees. *Cureus Journal of Computer Science*, 2, e44389.

25. Tunisian Insurance and Reinsurance Board. (2019). Market overview and regulatory framework. Technical Report, Ministry of Finance, Tunisia.

26. Wang, H. (2015). Estimating insurance attrition using survival analysis. *Variance Journal*, 9(1), 95-125.

27. Wüthrich, M. V., & Buser, S. (2018). Data science in non-life insurance pricing. *SSRN Electronic Journal*.

28. Yego, N., Mwangi, R., & Mugo, S. (2024). A comparative analysis of machine learning models for insurance policy uptake prediction. *African Journal of Information Systems*, 16(2), 154-177.

29. Yeo, E. C., Shiu, Y. M., & Pavur, R. (2001). Forecasting insurance lapse rates using neural networks. *Insurance: Mathematics and Economics*, 28(3), 335-345.