Journal of Money and Economy
Vol. 20, No. 4, Fall 2025
pp. 649-667

Original Research Article

Predicting Life Insurance Policyholder Churn in Iran Using Machine Learning: A Transparent and Actionable Framework

Ghadir Mahdavi[*] **Ramin Heidarzadeh Azar**[†]
Reza Ofoghi[‡]

Received: 08 Oct 2025 Approved: 18 Nov 2025

This study tackles the challenge of customer churn in life insurance, which leads to substantial financial losses. It introduces a transparent, reproducible, and leakage-free machine learning framework designed to identify at-risk policyholders accurately and efficiently.

Using 20,000 anonymized Iranian life insurance policies with a churn rate of 26%, the study develops a complete modeling pipeline. The pipeline includes imputation, standardization, and encoding steps performed strictly within the training process to prevent data leakage. Three model types—logistic regression, random forest, and XGBoost—were trained and evaluated using stratified cross-validation, F1-optimized thresholds, and performance metrics such as ROC-AUC, F1, and precision.

Results show that all models perform similarly (ROC-AUC ≈ 0.70), with logistic regression achieving the highest F1 score (0.66) and XGBoost offering the best precision (0.68). The top-ranked predictions captured about 69% of churners, demonstrating strong operational potential. Policyholders with major payment delays were identified as the most churn-prone and easily detectable group. Overall, the findings confirm that transparent and interpretable machine learning models can effectively balance accuracy, simplicity, and practicality—supporting data-driven, value-focused customer retention strategies in the life insurance industry.

Keywords: Customer Churn Prediction, Life Insurance, Machine Learning, Data Leakage Prevention, Retention Strategy
JEL Classification*:* G22, M31, C53

[*] ECO College of Insurance, Allameh Tabataba'i University, Tehran, Iran; Mahdavi@atu.ac.ir (Corresponding Author)
[†] Ph.D Candidate, Allameh Tabataba'i University; ramin.heidarzadeh.phd@gmail.com
[‡] ECO College of Insurance, Allameh Tabataba'i University, Tehran, Iran; Ofoghi@atu.ac.ir

1 Introduction

Customer churn in service industries, especially in life insurance, has costly and multidimensional consequences, since contracts are long-term, cash flows depend on distant horizons, and replacing each customer entails significant expenses for acquisition, training, and after-sales services (Bansal et al., 2005; Kotler & Keller, 2016).

In the Iranian market, the intensity of competition, the prevalence of digital channels, and the sensitivity of demand to economic conditions increase the likelihood of switching and cancellation by policyholders. Therefore, early identification of high-risk individuals for customer retention has become a strategic necessity (Shahroodi et al., 2024).

The consumer behavior literature indicates that the decision to stay or leave results from the interplay of service quality, perceived value, and the customer-organization relationship. These factors justify the design of early warning systems and intervention programs (Bhattacherjee, 2001; Bansal et al., 2005). In this context, data-driven analytics can timely reveal behavioral and experiential signals and enable targeted planning to prevent churn (Ngai et al., 2009).

Over the past decade, machine learning has become the core of churn prediction, and systematic reviews have mapped the landscape of methods, data challenges, and organizational deployment considerations (Geiler et al., 2022; Manzoor et al., 2024). From an algorithmic perspective, both transparent and linear models such as logistic regression and tree-based algorithms such as random forest and XGBoost are widely used, and the choice among them depends on the trade-off between transparency, efficiency, and operational requirements (Vafeiadis et al., 2015; Hanafy and Ming, 2021).

In the insurance industry, research shows that the quality of features, experimental design, and attention to data localization determine the success of models, and algorithm choice alone is insufficient (Groll et al., 2022).

Before modeling, data engineering and preprocessing play a central role. Standard transformations such as imputing missing values, standardization for scale-sensitive models, and one-hot encoding for categorical variables create consistent and usable inputs for different algorithms (Alpaydin, 2020).

Moreover, adherence to the principle of data leakage prevention is essential; that is, all transformations must be fitted within the training cycle and then applied to validation and test sets with the same mapping. Any direct

or proxy identifiers must be removed so that the model does not gain unintended access to target information (Alpaydin, 2020).

In churn model evaluation, reliance on problem-appropriate metrics is crucial.

The F1 score, as a combination of precision and recall under class imbalance, serves as an appropriate indicator of operational performance, since it balances reducing false alarms with capturing true cases (Vafeiadis et al., 2015). In addition, the Area Under the ROC Curve (ROC-AUC) and the Area Under the Precision–Recall Curve (PR-AUC) illustrate separability across thresholds, while the Brier score assesses the quality of predicted probabilities for use in probability-based decision rules (Imani and Arabnia, 2023). Such a set of metrics helps managers evaluate models not only in terms of ranking ability but also regarding "convertibility to decision" (Kotler & Keller, 2016).

From an operational standpoint, the output of each predictive system must be translated into actionable tools. The most common format is ranking and decile analysis, which allows the extraction of lift and cumulative gain charts and the planning of communication campaigns in accordance with available contact capacity or budget (Groll et al., 2022; Kotler & Keller, 2016). Thus, even if it is not possible to contact all customers, focusing on the top deciles of the ranking can improve the efficiency of retention programs (Ngai et al., 2009).

Given this theoretical and practical background, the objective of this paper is to present a coherent framework for churn prediction in life insurance that adheres to the principles of reproducibility and unbiased evaluation and remains aligned with operational needs from start to finish (Dietterich, 2000).

In this framework, the insurance dataset is stratified into training/validation/test sets; preprocessing is implemented as an integrated pipeline; cross-validation is used for hyperparameter tuning; the decision threshold is determined based on the maximum F1 score on the validation set; and finally, a single-pass evaluation is conducted on the test set to obtain an unbiased estimate of generalization (Imani & Arabnia, 2023). The selected algorithms, logistic regression, random forest, and XGBoost, are compared to separate the effect of algorithm choice from the effects of preprocessing and data (Alpaydin, 2020; Vafeiadis et al., 2015).

The contributions and innovations of this study can be summarized in four aspects:

1) Emphasis on preventing data leakage and implementing transformations within the learning cycle to ensure unbiased evaluation;

2) Reporting a comprehensive set of discriminative and probabilistic metrics (ROC-AUC, PR-AUC, Brier score) alongside operational metrics (F1, precision, recall) to support decision-making;
3) Translating model outputs into actionable operational tools, including decile segmentation, lift, and cumulative gain for designing retention campaigns under limited capacity;
4) Integrating explainability and segment stability into the evaluation process so that retention interventions are defined transparently and evidence-based (Adadi & Berrada, 2018; Groll et al., 2022; Manzoor et al., 2024).

Life-insurance voluntary lapse (customer churn) reduces profitability and disrupts servicing. Our goal is to predict which active policyholders are at risk of churn in the next period and rank them for targeted retention, using a leakage-free ML pipeline that is transparent and operational. The problem is characterized by class imbalance, potential data leakage from future-informed variables, and the need for interpretable drivers to justify actions.

In Section 2, we review previous work relevant to our study, highlighting gaps in the literature. Section 3 describes the dataset, its splitting and preprocessing, the models employed, and the tuning and evaluation metrics used. Section 4 presents the experimental results and operational analyses, while Section 5 discusses the implications of our findings and concludes the paper.

2 Theoretical and Empirical Foundations

The churn literature highlights three conceptual roots for customer departure: continuous evaluation of service quality, judgment about the received value relative to cost, and the dynamics of the customer–organization relationship. These three dimensions, within the frameworks of usage continuance models and service-switching structures, create mechanisms through which signs of dissatisfaction and behavioral frictions evolve into the intention to leave (Bhattacherjee, 2001; Bansal et al., 2005). In service organizations, this logic is translated, within the context of customer relationship management, into the processes of early recognition, preventive intervention, and outcome evaluation. Transactional, interactional, and satisfaction data serve as the raw materials for these processes (Ngai et al., 2009).

From a methodological perspective, systematic reviews show that different algorithmic families possess distinct advantages: linear models are renowned for their transparency and interpretability, whereas tree-based and ensemble

methods typically compete in terms of ranking accuracy (Geiler et al., 2022; Vafeiadis et al., 2015). In the insurance industry, empirical evidence indicates that the quality of features, their cleaning and representation, and the evaluation design (including strict separation of training/validation/test sets and systematic prevention of leakage) are just as influential as the choice of algorithm (Groll et al., 2022; Hanafy & Ming, 2021). Therefore, credible empirical foundations rely on two principles: integrity of the data pipeline and discipline in evaluation.

From the data perspective, the literature recommends combining layers of demographic information, insurance product characteristics, payment behaviors, and experience/satisfaction indicators; these layers complement one another and together provide a more realistic estimation of churn risk (Ngai et al., 2009). Standard procedures, such as imputation of missing values, standardization for scale-sensitive models, and one-hot encoding for categorical variables, not only enhance algorithmic performance but also allow fair comparison among models (Alpaydin, 2020). The key principle is to execute all these transformations within the learning cycle and transfer the learned mapping to validation and test data, ensuring that no information from the future leaks into the past (Alpaydin, 2020).

In performance measurement, the literature emphasizes a set of metrics, each covering a different dimension of the problem. F1, precision, and recall answer how many at-risk customers are correctly identified at the selected operational threshold (Vafeiadis et al., 2015). Threshold-independent metrics such as ROC-AUC and PR-AUC provide an overall picture of separability and are useful for algorithmic family comparisons (Geiler et al., 2022). ROC-AUC measures how well the model ranks positives above negatives across all possible classification thresholds: the ROC curve plots True Positive Rate vs. False Positive Rate, and its area (AUC) equals the probability a randomly chosen churner is scored higher than a non-churner (0.5 = random, 1.0 = perfect). PR-AUC summarizes the Precision–Recall trade-off over all thresholds and is especially informative under class imbalance, emphasizing how many of the predicted churns are correct (precision) and how many actual churns are captured (recall). Together, these threshold-independent curves provide an overall view of separability and performance stability across operating points. Additionally, the Brier score measures the quality of probabilistic outputs and is particularly important for value-based decisions (e.g., prioritization based on “probability × policy value”) (Imani & Arabnia, 2023).

The next essential step is to translate the model output into operational artifacts.

Practical experience in service industries has shown that population ranking and decile analysis form the foundation for retention campaign design: lift and cumulative gain charts intuitively demonstrate what proportion of at-risk customers are concentrated in the top-ranked segments and what level of coverage can be achieved with a given budget (Kotler & Keller, 2016).

This capacity-oriented approach aligns more closely with organizational realities (limited contact channels and per-contact costs) than reliance on a single accuracy number or fixed threshold (Groll et al., 2022).

The organizational adoption of predictive systems depends on their explainability and stability. Even in scenarios where tree-based or ensemble methods exhibit superior discrimination, managers require answers to "why" in order to design targeted retention interventions. Data-driven explanatory methods such as permutation importance and explainable artificial intelligence (XAI) frameworks clarify the pathways through which variables like customer satisfaction, payment delay/regularity, and product features affect outcomes, helping to avoid unfair or ineffective decisions (Adadi & Berrada, 2018; Manzoor et al., 2024). Alongside explanation, segment stability testing (assessing model performance within subsegments defined by payment behavior, retention period, or age) is necessary to ensure that threshold policies and messaging are aligned with the actual risk level of each segment (Groll et al., 2022).

Accordingly, the theoretical and empirical foundations of churn in life insurance converge toward an operational framework in which:

1) Multilayered data are transformed into a standardized representation through a consistent, leakage-free pipeline;
2) Hyperparameter tuning is performed via cross-validation, and the decision threshold is selected based on goal-aligned metrics;
3) Both discriminative and probabilistic metrics are reported together to guarantee convertibility into decisions; and
4) Model outputs are rewritten in the language of operations (deciles, lift, and gain), and finally, based on explanation and segment stability, targeted retention interventions are designed (Dietterich, 2000; Kotler & Keller 2016; Manzoor et al., 2024).

This framework provides the theoretical foundation required for the present research methodology and enables fair evaluation of various algorithmic approaches within the constraints of real-world organizational settings (Geiler et al., 2022; Ngai et al., 2009).

Building on prior work that stresses operational decision-making under class imbalance, leakage control, and interpretability, we translate the theoretical points into concrete design choices. Because churn is rare, we use stratified splits and emphasize PR-AUC alongside ROC-AUC to reflect ranking quality under imbalance. To avoid data leakage and obtain unbiased generalization, all preprocessing (imputation/encoding) fit inside cross-validation folds, and we report a single, untouched test-set result. Since managers act on ranked risk lists with limited outreach capacity, we select a single operating threshold from cross-validation and report precision/recall/F1 at that point, plus recall at k% for top-segment targeting. Together, these methodological choices operationalize the theoretical requirements and enable fair, real-world evaluation of competing algorithms.

3 Data Methodology

3.1 Data and Definition of the Target Variable

The study population consists of 20,000 anonymized life insurance policies from an Iranian company. Four families of features are employed: demographic, policy-related, behavioral/transactional, and experience/satisfaction indicators. The target variable, *churn*, is an operational binary label applied consistently across all data segments, with a prevalence of about 26%. To prevent unintended access of the model to the label, all direct and proxy identifiers (e.g., sequential keys, etc.) were removed prior to modeling (Alpaydin, 2020).

3.2 Experimental Design and Data Splitting

The data are stratified and divided in a 70/15/15 ratio into training/validation/test sets, with fixed random seeds to ensure reproducibility. The test set is used only once, after model/parameter selection, for an unbiased estimation of generalization. This evaluation discipline aligns with classical recommendations regarding data separation and bias control (Dietterich, 2000).

3.3 Preprocessing and Leakage-Free Pipeline

All transformations are implemented within a single pipeline and *inside* the learning cycle to prevent data leakage. For numerical features, median imputation is combined with a “missing data” indicator, and for scale-sensitive models, standardization is applied. Categorical features are represented with one-hot encoding, and an “unknown/unseen” level is defined

to handle new categories at deployment. All models receive exactly the same transformed matrix to ensure fair comparison (Alpaydin, 2020).

We convert our categorical fields to numeric form using encodings matched to each variable's semantics. Marital status is nominal, so we apply one-hot encoding (a separate 0/1 column per status) to avoid imposing any artificial order. Gender and Additional-coverage purchase (yes/no) are binary and are represented as 0/1 indicators (functionally equivalent to a single-column one-hot). Contract length/term is ordinal (longer terms represent a higher level on the same scale); when stored as categories (e.g., "5y, 10y, 15y"), we use ordinal encoding that respects this natural order, or we keep it as an integer year count if available. The remaining variables are numeric and therefore do not require categorical encoding (they undergo the numeric preprocessing described earlier). All encoders are fitted inside each cross-validation fold and then applied to that fold's validation split and to the final test set to prevent data leakage. This preserves the class balance and ensures the model sees only training-fold information when learning the encoding.

3.4 Models and Hyperparameter Tuning

Three algorithmic families are compared within the same environment: logistic regression, random forest, and XGBoost (as gradient boosting). A small-scale grid/random search is conducted using stratified 3-fold cross-validation, and the selection criterion is the F1 score on validation at each model's operating point (Dietterich, 2000).

Hyperparameters are tuned with a small-scale grid/random search evaluated by 3-fold stratified cross-validation. In k-fold cross-validation, the data are split into k equal folds; the model is trained on k–1 folds and evaluated on the held-out fold, rotating so each fold serves once as validation, and the scores are averaged. Stratification preserves the class ratio (churn vs. non-churn) inside every fold, which is crucial under imbalance. We use k = 3 at this stage to balance robustness with computation time during the initial sweep; this yields stable, preventing data leakage estimates while keeping the search tractable.

All algorithms are implemented as single pipelines in which preprocessing (imputation/encoding) is trained only on the training folds and then applied to validation/test data to prevent leakage. We evaluate three classifiers: logistic regression, random forest, and XGBoost. For logistic regression we vary C (inverse regularization strength) and penalty/solver compatible pairs; for random forest we vary n_estimators, max_depth, and min_samples_leaf (with class_weight='balanced'); for XGBoost we vary n_estimators, max_depth,

learning_rate, subsample, colsample_bytree, and scale_pos_weight to address class imbalance. We set fixed random seeds and keep the same folds for all models to ensure a like-for-like comparison.

3.5 Operating Point Selection and Metrics

First, the probability threshold is swept across the range 0–1, and the threshold that maximizes the F1 score on the validation set is chosen. Then, the selected model is retrained on the combined training+validation sets and evaluated only once on the test set (Imani & Arabnia, 2023). Reported metrics include: F1 (primary), precision, recall, ROC-AUC, PR-AUC, and Brier score for probability quality. In addition to the confusion matrix at the chosen threshold, ROC and precision–recall curves are plotted for threshold-independent comparison (Vafeiadis et al., 2015).

3.6 Explainability, Calibration, and Probability Quality Control

Permutation importance is applied to the original feature space to interpret feature roles by estimating the performance drop caused by shuffling each variable. Probability quality is assessed using the Brier score and calibration plots; if miscalibration is detected, isotonic regression or Platt scaling (on the validation set) is applied and then verified on the test set (Adadi & Berrada, 2018; Imani & Arabnia, 2023).

3.7 Operational Outputs and Ranking Metrics

To translate model outputs into actionable decisions, scores are converted into rankings, and the population is divided into deciles. Lift and cumulative gain charts are used to evaluate the concentration of risk in top deciles, and contact lists are prepared for 5%, 10%, and 20% cutoffs to match contact capacity constraints. This procedure aligns with recommendations in the CRM and insurance literature, creating a direct link between machine learning metrics and campaign planning (Kotler & Keller, 2016; Groll et al., 2022).

3.8 Segment Stability Testing and Reliability of Balance

Model performance is examined within meaningful subsegments (e.g., payment delay severity, contract duration, age groups). For precision/recall rates, Wilson confidence intervals are reported so that uncertainty in smaller segments is considered in decision-making. Additionally, bootstrap resampling is conducted on the test set for key metrics to assess the robustness of results against random sampling variation (Groll et al., 2022).

4 Research Findings

In this study, eight key features are used, covering four informational groups. From the customer experience dimension, the satisfaction score reflects perception and perceived service quality. From the financial behavior layer, payment delay is recorded as the number of delayed installments. From the product aspect, the purchase of additional coverage/rider indicates whether the policyholder has added an extra component to the contract. Among demographic features, marital status, age, and gender are included. In the time and contract dimension, contract length specifies the nominal policy duration at inception, and elapsed time indicates how long the policy has been active.

It is expected that satisfaction score and payment delay provide stronger churn signals, while demographic variables typically play a complementary role. Temporal features (contract length and elapsed time) help explain the customer life cycle and may interact with payment and satisfaction behaviors.

For data preparation, numerical features (satisfaction score, age, contract length, elapsed time, and numeric form of payment delay) undergo missing-value imputation, and when required, standardization for scale-sensitive models. Categorical features such as marital status, gender, and additional coverage purchase are represented through one-hot encoding to be usable across all models.

In this section, results are reported and interpreted based on the selected operating point (decision threshold determined via validation). The structure of presentation follows this sequence:

- final model performance on the test data,
- ranking quality for operational execution via decile precision,
- decile/lift/cumulative gain analysis to estimate coverage at population cutoffs,
- segment analysis focusing on payment delay severity,
- uncertainty estimation via bootstrap confidence intervals,
- and feature importance interpretation using permutation importance.

4.1 Final Model Performance on the Test Data

At first glance, all three models exhibit a ROC-AUC around 0.7, indicating similar overall discriminative power. Table 1 shows that the main differences lie in the balance between precision and recall, and consequently in the F1 score.

- Logistic regression achieves precision = 0.62 and recall = 0.71, yielding the highest F1 = 0.66.

- This means at the selected operating point, it identifies churn cases with relatively good coverage while controlling unproductive contacts.
- Random forest gives precision = 0.640, recall = 0.630, F1 = 0.630, offering a more balanced profile, suitable when both error types carry equal importance.
- XGBoost yields precision = 0.68, recall = 0.56, F1 = 0.61, meaning that when contact capacity is limited or contact cost is high, this model produces more precise but narrower targeting.

The Brier score ≈ 0.22 for all three indicates adequately calibrated probabilities, suitable for value-based rules (e.g., prioritization by *probability* × *policy value*).

In summary:

- For broader churn coverage, choose logistic regression;
- For higher contact accuracy, prefer XGBoost;
- For a middle ground, random forest is appropriate.

Table 1
Final test scoreboard at the selected operating point

Model	ROC-AUC	F1	Recall	Precision
Logistic Regression	0.70	0.66	0.71	0.62
Random Forest	0.69	0.63	0.63	0.64
XGBoost	0.70	0.61	0.56	0.68

Source: Research Findings

4.2 Ranking Quality for Execution: Precision by Population Decile

Decile precision in Table 2 shows how risk concentrates at the top of the score list. In the selected model, Decile 1 has precision = 0.80, meaning that if only the top 10% of customers are contacted, 4 out of 5 are actual churners. Precision gradually decreases across deciles to 0.40 in the 10th decile.

This stepwise pattern provides two operational insights:

1) For capacity-constrained campaigns (e.g., 5–10% of the population), focusing on deciles 1–2 yields maximum return.
2) Expanding outreach to decile 3 still adds value but with diminishing efficiency beyond mid-deciles.

Hence, decision-makers can select a decile cutoff that optimizes the trade-off between the number of contacts and success rate.

Table 2
Precision by population decile (Selected Model; Test Set)

Decile	Precision
1–10%	0.80
11–20%	0.72
21–30%	0.67
31–40%	0.64
41–50%	0.60
51–60%	0.57
61–70%	0.53
71–80%	0.49
81–90%	0.45
91–100%	0.40

Source: Research Findings

All calculations use the held-out test set only. Estimation steps for the decile table are as follows:

1) Score: We applied the finalized model to the test set to obtain predicted churn probabilities for each policy.
2) Rank: We sorted policies descending by predicted probability.
3) Bin into deciles: We split the ranked list into 10 equal-count bins (10% each) using probability quantiles. Ties at a cutoff were assigned to the higher (riskier) decile.
4) Compute precision per decile: For each decile, we counted how many customers actually churned within that decile and divided by the total number of customers in the same decile. This ratio (shown as a number between 0 and 1) is the precision for that decile.

4.3 Deciles, Lift, and Cumulative Gains: Coverage Estimation by Population Cutoffs

Table 3 summarizes key decision-making ratios. Decile 1 with Lift = 2.09 indicates a churn rate over twice the base rate, thus offering high efficiency if targeted. Cumulative recall rises progressively:

- Decile 1 alone covers 21% of churns;
- Adding Decile 2 raises coverage to 36%;
- With Decile 3, nearly 49% are covered;
- The top 50% (Deciles 1-5) capture ≈69% of all churns.

These values serve as budgeting tools: e.g., with capacity for 20% of customers, about 36% of churn can be addressed; with 50% capacity, nearly two-thirds.

Table 3
Deciles, lift and cumulative gains

Decile	Lift	Cumulative Recall
1	2.09	0.21
2	1.50	0.36
3	1.28	0.49
4	1.03	0.59
5	1.01	0.69
6	0.65	0.76
7	0.73	0.83
8	0.69	0.90
9	0.62	0.96
10	0.39	1.00

Source: Research Findings

4.4 Segment Analysis: Payment Delay Severity

A clear pattern emerges: the greater the delay, the higher the churn probability and the easier the detection.

- In the no-delay group (0), churn rate ≈ 22.3%, with moderate model performance (precision ≈ 0.35, recall ≈ 0.50). Here, identifying churners is harder; thus, a stricter threshold is advisable to reduce false contacts.
- In the high-delay group (3+), intrinsic churn ≈ 46%, and the model performs very well (precision ≈ 0.471, recall ≈ 0.948).

Managerial implication: Defining segment-based thresholds

- Softer in high-risk segments to maximize coverage;
- Stricter in low-risk segments to avoid wasted contacts.

Messaging should also differ:

- For high-risk groups, focus on payment facilitation (active reminders, installment options);
- For low-risk groups, emphasize experience enhancement and loyalty. Confidence intervals reveal where differences are statistically stronger; wider intervals in smaller segments suggest cautious decision-making and further data collection.

4.5 Uncertainty at the Selected Operating Point: Bootstrap

As shown in Table 4, 95% confidence intervals for key metrics (F1, AUCs, precision, recall, Brier) are narrow. For example:

- F1 fluctuates between 0.445–0.501,
- ROC-AUC between 0.658–0.705.

This narrowness implies stable performance; slight sampling changes will not alter conclusions. Operationally, such robustness allows reliable planning without major concern for randomness. Alignment between PR-AUC and point metrics confirms well-chosen threshold calibration.

Table 4
Bootstrap 95% confidence intervals (Selected Operating Point).

Metric	Value	CI 2.5%	CI 97.5%
F1	0.4728	0.4450	0.5014
ROC-AUC	0.6812	0.6584	0.7048
PR-AUC	0.4408	0.4052	0.4798
Precision	0.3806	0.3518	0.4103
Recall	0.6239	0.5887	0.6587
Brier	0.2227	0.2176	0.2276

Source: Research Findings

4.6 Final Explainability: Permutation Importance of Features

According to Table 5, "Satisfaction Score" and "Payment Delay" produce the largest F1 drops upon permutation (≈0.055 and ≈0.037). Hence, disturbances in experience or payment data substantially weaken prediction. Product variables (e.g., additional coverage) add modest, complementary value, while purely demographic variables contribute minimally once main signals are included. Managerial takeaway: To improve the model, invest in enriching experience data (periodic surveys, service contact metrics) and granular payment histories (delay patterns, days to due date, number of reminders).

Table 5
Permutation importance (ΔF1 ± SD)

Feature	Mean ΔF1	SD
Satisfaction Score	0.05511	0.00993
Payment Delay	0.03680	0.00515
Additional Coverage Purchase	0.01143	0.00459
Marital Status	0.00937	0.00355
Contract Length	0.00084	0.00052
Age	0.00062	0.00001
Elapsed Time	0.00045	0.00049
Gender	0.00028	0.00071

Source: Research Findings

4.7 Summary of Empirical Results

The findings are:

- All three models show comparable discriminative power, but differ in precision–recall balance;
- Top deciles contain strong risk concentration, offering significant operational leverage;
- Lift/gain charts demonstrate how covering 10–20% of the population can target a large share of churn;
- Segment analysis confirms that payment delay severity is both a risk signal and an actionable lever;
- Bootstrap stability ensures field reliability;
- Permutation analysis clarifies improvement paths (enriching experience and payment data).

Practically, a "decile cut and segment-based threshold" policy, combined with cost-effective channels, can meaningfully raise retention success without altering model complexity.

5 Discussion and Conclusion

5.1 Summary of Findings

We presented an integrated, leakage-free framework for predicting churn in life insurance. Three benchmark algorithms, logistic regression, random forest, and XGBoost, were compared at the selected operating point on the test data. The findings are clear: the three models exhibit similar discriminative ability, but their operational profiles differ; i.e., each displays a distinct balance between precision and recall, and must be chosen based on business objectives.

The ranking analysis revealed strong risk concentration in top deciles, and lift/gain charts demonstrated that targeting a small portion of the population (e.g., 10–20%) can capture a substantial share of churn cases.

Segment analysis showed that segments with higher payment delays have both higher intrinsic churn risk and better model detectability. In terms of feature importance, satisfaction score and payment delay play the leading roles in churn prediction, while demographic variables act as complementary features.

5.2 Managerial and Policy Implications

The results offer several clear messages for operational decision-making:

1) The operating-point scoreboard on the held-out test set shows the precision–recall trade-off; when the threshold is lowered, recall rises at the cost of precision, so choose higher-recall for broad coverage or higher-precision when contact capacity or costs are constrained. Therefore, threshold selection must align with contact capacity and action cost:
   - If the goal is wider churn coverage, choose a point with higher recall. It is because validation scans around the operating point show recall increases as the threshold is relaxed with acceptable precision, matching the campaign goal of maximum coverage;
   - If contact cost is high or capacity limited, select a higher-precision model and threshold. It is because at the selected operating point, precision is materially above the base rate, which limits wasted contacts; tighter thresholds further raise precision when budgets are constrained.
2) The decile precision table concentrates churners in the top bins (large lift over baseline) with a clear step-down pattern, so targeting by deciles maximizes returns. Therefore, ranking and decile segmentation should be the basis for campaign design:
   - For short lists, focus on deciles 1 and 2 for maximum return. It is because deciles 1 and 2 have the highest precision and lift, well above baseline, so they deliver the most churners per contact;
   - For larger scopes, extend to decile 3-5. It is because precision in deciles 3–5 remains above the population rate, keeping positive marginal return as scope expands.
3) Implement segment-based policies:
   - In high-delay segments, adopt softer thresholds and payment-facilitation messages (e.g., active reminders, installment plans, rescheduled due dates). It is because payment delay ranks among the top importance drivers exhibit higher churn prevalence;
   - In low-risk segments (no or low delay), apply stricter thresholds and experience-focused interventions (e.g., satisfaction follow-ups, value-added offers). It is because lower-risk segments sit in lower deciles with precision near baseline; raising thresholds conserves budget there, while softer service touches maintain satisfaction.
4) Given the strong role of satisfaction, invest in customer experience enhancement, which both reduces churn probability and enriches model signals. It is because satisfaction score contribute meaningfully in the permutation-importance and attribution results.

5.3 Deployment, Monitoring, and Data Ethics

For operational deployment, several key steps are recommended:

1) Conduct a small A/B pilot to measure the real effect of contact lists and suggested messages;
2) Perform periodic calibration of thresholds and probabilities, since portfolio composition and customer behavior evolve over time;
3) Continuously monitor metrics (precision, recall) and adjust thresholds and decile cutoffs gradually;
4) Establish a feedback loop between operations/contact teams and data teams so that reasons for success or failure of contacts are returned as new labels.

Regarding ethics and fairness, ensure:

- Data anonymization,
- Restricted access, and
- Non-discrimination among groups.

Using simple explanatory reports (e.g., top five influential variables per prediction) also improves organizational acceptance.

5.4 Conclusion

In summary, the proposed framework demonstrates that, through a reproducible and operationally aligned pipeline, customer churn in life insurance can be predicted with suitable accuracy and transparency, and translated into actionable decisions. Top-ranked deciles provide short-term high-return intervention opportunities, and the key features, experience and payment behavior, guide the design of retention programs.

With continuous monitoring, periodic calibration, and segment-based policies, these benefits can be maintained and even strengthened over time. This approach provides a practical foundation for integrating data science with decision-making in life insurance portfolios and can be directly applied to annual retention planning and customer communication budgeting.

Our findings are consistent with prior churn studies. For example, Imani and Arabnia (2023), working on a public telecom dataset, report very high discrimination and F1 after Optuna hyperparameter tuning and SMOTE variants. Their protocol optimizes on resampled training distributions and reports cross-validated means, which tends to inflate threshold-dependent metrics. By contrast, we avoid resampling, fix the operating threshold from cross-validation, and report a single pass on an untouched test set. Even so, the model ranking matches theirs: tree-based ensembles and well-regularized linear models form the top tier.

Similarly, Vafeiadis et al. (2015) achieve strong performance on the classic UCI telecom churn data (boosted SVM, very high accuracy and F-measure) under Monte-Carlo cross-validation. That dataset is cleaner, less sparse, and evaluated primarily with accuracy, a metric that can look optimistic under imbalance. Our study emphasizes PR-AUC and threshold-specific precision/recall/F1, which are more informative for rare-event targeting; under those criteria our discrimination is more modest, but the operational lift pattern, a steep concentration of churners in the top ranks, mirrors theirs.

Finally, Lalwani et al. (2022) report AUC in the mid-0.8s and ~82% accuracy for AdaBoost/XGBoost on telecom churn with k-fold tuning. Our ROC-AUC is lower, consistent with a harder domain and a conservative evaluation protocol, yet we observe the same decile behavior they imply: the top deciles contain a disproportionate share of churners, supporting decile-based campaign design. Taken together, these comparisons suggest our approach reproduces well-known regularities (model ordering, ranked-list lift) under more deployment-like conditions.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable AI. *IEEE access*, 6, 52138–52160.

Alpaydin, E. (2020). *Introduction to machine learning* (4th ed.). MIT Press.

Bansal, H. S., Taylor, S. F., & St. James, Y. (2005). "Migrating" to new service providers: Toward a unifying framework of consumers' switching behaviors. *Journal of the academy of marketing science*, 33 (1), 96–115.

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model 1. *MIS Quarterly*, 25 (3), 351–370.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *In International workshop on multiple classifier systems (pp. 1-15)*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Geiler, L., Affeldt, S., & Nadif, M. (2022.) Machine learning methods for churn prediction: A survey. *International Journal of Data Science and Analytics*, 14 (۳), 217–242.

Groll, A., Wasserfuhr, C., & Zeldin, L. (2022.) Churn modeling of life insurance policies via statistical and machine learning methods: Analysis of important features. *arXiv preprint arXiv:2202.09182*.

Hanafy, M., & Ming, R. (2021). Machine learning approaches for handling auto insurance big data. *Risks*, 9 (2), 42.

Imani, M., & Arabnia, H. R. (2023). Hyperparameter optimization and sampling techniques for customer churn prediction. *Technologies*, *11 (6), 167*.

Kotler, P., & Keller, K. L. (2016). *Marketing management* (15th ed.). Pearson.

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2), 271-294.

Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). Machine learning for customer churn prediction: A practitioner-oriented review. *IEEE Access*, 12, 70434–70449.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36 (2), 2592–2602*.

Shahroodi, K., Avakh Darestani, S., Soltani, S., & Eisazadeh Saravani, A. (2024). Developing strategies to retain organizational insurers using a clustering technique: Evidence from the insurance industry. *Technological Forecasting and Social Change*, *201*, *123217*.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. Ch. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55, 1–9.