Customer Churn in Banking:
An Integrated Approach Combining Survival Analysis and Machine Learning

Andrew Sullivan

Independent Research Project

Fall 2025

Abstract

Customer retention has emerged as a strategic priority for banks amid intensifying competition, shrinking margins, and rising digital expectations. Using 10,000 retail-banking records, this study integrates exploratory analysis, survival modeling, and machine learning to explain and predict churn. Building on prior survival-analysis work, this study quantifies risk factors, estimates time to churn, and delivers a Random Forest classifier with 85.9% accuracy for proactive risk scoring. Inactivity, product-portfolio imbalance, and lifecycle stage dominate risk; inactive members face 1.88× higher churn. Combining Kaplan–Meier and Cox models with interpretable ensemble methods provides both temporal and probabilistic insights, enabling targeted re-engagement campaigns, lifecycle retention programs, and portfolio optimization.

Contents

1 Introduction **1**
1.1 Literature Review . . . 1

2 Methods **2**
2.1 Dataset and Pre-Processing . . . 2
2.1.1 Data Source and Structure . . . 2
2.1.2 Cleaning and Feature Engineering . . . 2
2.1.3 Feature Selection and Rationale . . . 3
2.2 Survival Analysis . . . 3
2.2.1 Kaplan–Meier Estimator . . . 3
2.2.2 Cox Proportional Hazards Model . . . 3
2.3 Predictive Modeling . . . 4
2.3.1 Random Forest Implementation . . . 4
2.3.2 Model Evaluation Metrics . . . 4
2.3.3 Model Comparison and Validation . . . 5

3 Results **5**
3.1 Exploratory Data Analysis . . . 5
3.1.1 Baseline Churn Rate and Descriptive Statistics . . . 5
3.1.2 Segment-Level Churn Patterns . . . 6
3.1.3 Feature Selection Analysis . . . 7
3.2 Survival Analysis Results . . . 10
3.2.1 Kaplan–Meier Survival Curves . . . 10
3.2.2 Cox Proportional Hazards Model Results . . . 12
3.3 Predictive Model Performance . . . 14
3.4 Model Validation and Robustness Checks . . . 20

4 Discussion **22**
4.1 Key Findings . . . 22
4.2 Business Applications and Strategic Implications . . . 22
4.2.1 Customer Risk Profiles . . . 23
4.2.2 Strategic Recommendations and ROI Analysis . . . 23
4.2.3 Implementation Roadmap . . . 25
4.3 Limitations and Future Research . . . 25
4.4 Conclusion . . . 26
4.5 Acknowledgements . . . 26

1 Introduction

Customer churn, the process by which customers close accounts or cease doing business with a firm, is a critical concern for banks. On average, acquiring a new customer can cost five to seven times more than retaining an existing one (Business Builders Co, 2024), and even modest improvements in retention can yield disproportionate profit increases (Kumar, 2022). For many financial institutions, high churn rates translate into substantial losses in lifetime value. Effective churn management therefore requires not only understanding who leaves and when, but also why they leave and how the bank can intervene.

Using a rich dataset of ten thousand customers made publicly available by Kollipara (2022), this study investigates demographic, behavioral and financial attributes to determine how they contribute to attrition. The analysis employs survival analysis and machine learning techniques to estimate individual churn probabilities and design targeted retention programs. Specifically, this study conducts exploratory data analysis to identify predictive factors, quantifies time-to-churn patterns using Kaplan–Meier estimators and Cox proportional hazards models, builds a random forest classifier for proactive risk scoring, validates model performance against alternative algorithms, and translates statistical findings into actionable business recommendations with quantified ROI. The methods and insights presented here demonstrate an integrated approach to analytics-driven customer retention applicable to similar banking contexts.

1.1 Literature Review

Customer retention has emerged as a fundamental business imperative across service industries, driven by the well-established principle that retaining existing customers costs significantly less than acquiring new ones (Business Builders Co, 2024). In banking, this dynamic is particularly pronounced, with customer lifetime values ranging from $2,000 to $4,000 for typical retail banking relationships (Meleis, 2010).

Early approaches to churn management relied on customer relationship management (CRM) systems that operated reactively, identifying problems only after customers had begun to disengage (Singh et al., 2024). The shift toward proactive churn prediction leverages machine learning techniques to identify customers at risk based on demographic, behavioral, and transactional patterns observed before explicit signals of dissatisfaction emerge. Singh et al. (2024) conducted a comprehensive comparative analysis of multiple ML algorithms on the same dataset used in this study, achieving optimal performance with Random Forest (78.3% accuracy, 69.3% sensitivity using SMOTE oversampling) and XGBoost (83.9% accuracy, 60.1% sensitivity). Their findings validated several critical patterns in bank customer behavior, including elevated churn rates among German customers and the optimal retention profile for customers holding exactly two products, patterns that receive independent confirmation in our exploratory analysis.

However, traditional classification approaches, while effective at answering *who* will churn, provide limited insight into *when* churn occurs or how temporal factors contribute to attrition risk. Survival analysis methods, adapted from biostatistics and telecommunications churn studies (Desai, 2023), offer a complementary framework for modeling time-to-event outcomes. Model interpretability has also emerged as a critical requirement for operational

deployment, with recent work demonstrating the utility of SHAP (Shapley Additive Explanations) frameworks for explaining black-box predictions in banking contexts (Peng et al., 2023). The translation of predictive insights into actionable business strategy also remains a critical gap in academic churn research (Brito et al., 2024).

This study bridges these gaps by combining survival analysis methodology with comparative ML evaluation and strategic business planning. The analysis extends the findings of Singh et al. (2024) through several methodological contributions: (1) incorporation of Kaplan–Meier survival curves and Cox proportional hazards modeling to quantify temporal churn patterns; (2) systematic comparison of SMOTE oversampling versus class-weight balancing strategies; (3) application of SHAP values and partial dependence plots for granular model interpretability; and (4) translation of statistical findings into quantified ROI projections and phased implementation strategies.

2 Methods

This section describes the data sources, preprocessing procedures, statistical methods, and model development approaches employed in the analysis. The methodology encompasses three main components: exploratory data analysis to identify patterns and guide feature selection, survival analysis to quantify temporal churn patterns, and predictive modeling to develop actionable risk scoring tools.

2.1 Dataset and Pre-Processing

2.1.1 Data Source and Structure

The analysis uses the *Bank Customer Churn* dataset compiled by Kollipara (2022). The dataset comprises 10,000 anonymised records of retail banking customers. Each record includes demographic variables (e.g. gender, geography, age), behavioral indicators (active membership status, tenure), product usage metrics (number of products, credit card ownership), financial variables (balance, estimated salary, credit score) and experiential measures (satisfaction score, complaint status, card type and loyalty points). In addition to the feature columns, the dataset includes a binary target variable indicating whether the customer exited the bank. The data card provided by the dataset author notes that identifier columns such as RowNumber and CustomerId have no predictive value and should be dropped.

2.1.2 Cleaning and Feature Engineering

Records with missing or duplicate values were removed. Identifier fields (RowNumber, CustomerId) were discarded. An age_group feature was engineered by discretising the Age variable into six categories (18–30, 31–40, 41–50, 51–60, 61–70, 70+) to capture non-linear lifecycle effects while preserving interpretability. One-hot encoding was applied to categorical variables (gender and geography). The complaint indicator was intentionally excluded from predictive models as a lagging indicator of churn. Continuous variables were standardized using Z-score normalization primarily for Cox regression compatibility; tree-based

algorithms are invariant to scaling but standardization was applied consistently across all models.

2.1.3 Feature Selection and Rationale

Linear correlation analysis with churn was used to identify predictive features. Features with moderate to strong correlations (—r— ≥ 0.10) were retained as core predictors, while those with minimal linear relationships (—r— < 0.10) were excluded from modeling. The threshold of —r— ≥ 0.10 represents a standard heuristic for identifying meaningful associations in behavioral data; correlations below this threshold typically indicate negligible linear relationships that are unlikely to meaningfully contribute to prediction models. The complaint indicator was intentionally excluded because it represents a lagging indicator of churn that would create a methodological degenerate case (Kumar, 2022). This selective approach balances predictive power with interpretability, emphasizing modifiable factors (number of products, activity status) and strong demographic predictors (age, geography) that directly inform intervention strategies. Features with weak linear correlation but anticipated nonlinear effects were retained for tree-based algorithms capable of capturing complex patterns.

2.2 Survival Analysis

Survival analysis models the time until an event occurs and is well suited for churn studies where the timing of attrition matters. The fundamental quantity is the survival function $S(t) = \Pr(T > t)$, which represents the probability of surviving at least to time t. The hazard function $h(t)$ quantifies the instantaneous risk of experiencing the event at time t, given survival until t.

2.2.1 Kaplan–Meier Estimator

Kaplan–Meier (K–M) estimators, first described by Dudley et al. (2016), are non-parametric methods that estimate the survival function based solely on observed event times and censoring. K–M curves were computed for customer segments stratified by age group, number of products, activity status, geography, gender, and balance group. Log-rank tests were performed to assess statistical significance of differences between groups, testing the null hypothesis that survival distributions are identical across strata.

2.2.2 Cox Proportional Hazards Model

The Cox proportional-hazards model is a multivariable regression that relates the hazard of the event to multiple covariates simultaneously (Kassambara, 2020). The hazard function in the Cox model is specified as

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p),$$

where $h_0(t)$ is the baseline hazard and the coefficients β_i quantify the effect of covariate x_i on the hazard. Hazard ratios ($\exp(\beta_i)$) greater than one indicate increased risk, while ratios below one signify protective effects.

A Cox model was fitted excluding the complaint variable to avoid its dominance. Covariates included gender, tenure, balance, number of products, activity status, geography, and age group dummy variables. Model fit was assessed using the concordance index (C-index), which measures the proportion of comparable pairs where the predicted outcomes are correctly ordered. The proportional hazards assumption was tested using Schoenfeld residuals; all covariates satisfied the assumption (no significant time-dependency), validating the hazard ratio interpretations.

2.3 Predictive Modeling

A random forest classifier was trained on the curated feature set to identify at-risk customers before a complaint is lodged. Random forests are ensemble learning techniques that combine the output of many decision trees built on bootstrap samples of the data (GeeksforGeeks, 2025). Each tree considers a random subset of features at each split and contributes a vote to the final prediction. Formally, for a random forest with B trees, the ensemble prediction is

$$\hat{f}_{\text{RF}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(\mathbf{x}),$$

where $\hat{f}_b(\mathbf{x})$ is the prediction from the b-th tree trained on a bootstrap sample with random feature selection at each split. For classification, the final prediction is the majority vote across all trees.

Data were split into training (80%) and test (20%) sets using stratified sampling to preserve class distribution. A four-stage grid search was performed to tune hyperparameters including the number of trees, maximum depth, splitting criteria, minimum samples per split, and class weights. Class imbalance was addressed by weighting the minority class twice as heavily as the majority class.

2.3.1 Random Forest Implementation

Random forests were implemented using scikit-learn's RandomForestClassifier. The algorithm randomly samples both observations (with replacement) and features (without replacement) at each split, creating decorrelated trees that reduce overfitting through ensemble averaging.

2.3.2 Model Evaluation Metrics

Model performance was assessed using multiple metrics. Accuracy quantifies the overall proportion of correct predictions but is sensitive to class imbalance. Precision measures the fraction of predicted churners that actually churn. Recall measures the fraction of actual churners correctly identified. F1-score provides a harmonic mean balancing precision and recall.

Threshold-agnostic discrimination was evaluated using receiver operating characteristic (ROC) curves, computing the area under the curve (ROC-AUC). The ROC-AUC quantifies the probability that the classifier ranks a randomly chosen positive instance higher than a

randomly chosen negative instance. Precision-recall curves and PR-AUC were also computed, providing more informative assessment under class imbalance.

Feature importance was assessed using three complementary approaches: (1) built-in impurity-based importance from the trained model, (2) permutation importance measuring performance degradation when features are randomly shuffled, and (3) SHAP (Shapley Additive Explanations) values quantifying marginal feature contributions to individual predictions. Partial dependence plots were generated to visualize the marginal effect of individual features on predicted probabilities.

2.3.3 Model Comparison and Validation

Hyperparameter tuning employed 3-fold cross-validation within GridSearchCV to select optimal parameters; final model performance was validated using 5-fold cross-validation on the training set. While cross-validation guided model selection, final performance metrics are reported on the holdout test set to provide unbiased estimates of generalization capability.

The Random Forest classifier was compared against two gradient-boosting alternatives: XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017). Both are ensemble methods that sequentially add weak learners to correct prior errors, differing primarily in their splitting strategies: XGBoost uses a level-wise (breadth-first) tree growth approach with regularization, while LightGBM employs a leaf-wise (depth-first) growth strategy optimized for computational efficiency. All models were trained on identical data splits and tuned with analogous hyperparameter grid searches.

Additionally, experiments evaluated synthetic minority oversampling (SMOTE) versus class-weight balancing strategies. In SMOTE experiments, synthetic samples were generated only from training data to prevent leakage into the test set; the holdout test set remained untouched. Both SMOTE and class-weight approaches were evaluated within the cross-validation framework to ensure robust performance estimates.

3 Results

3.1 Exploratory Data Analysis

Exploratory data analysis was conducted to characterize the dataset, identify baseline churn patterns, and guide feature selection decisions. This initial investigation revealed several striking patterns that informed subsequent survival analysis and predictive modeling approaches.

3.1.1 Baseline Churn Rate and Descriptive Statistics

The baseline churn rate in the dataset is 20.4 %, corresponding to 2,038 customers exiting during the observation window and 7,962 remaining (Figure 1). The most dominant factor was complaint status: 99.5 % of customers who lodged a complaint subsequently churned, compared to only 0.05 % of non-complainers. Because complaint status is effectively a point of no return, it was analysed separately from the main predictive model.

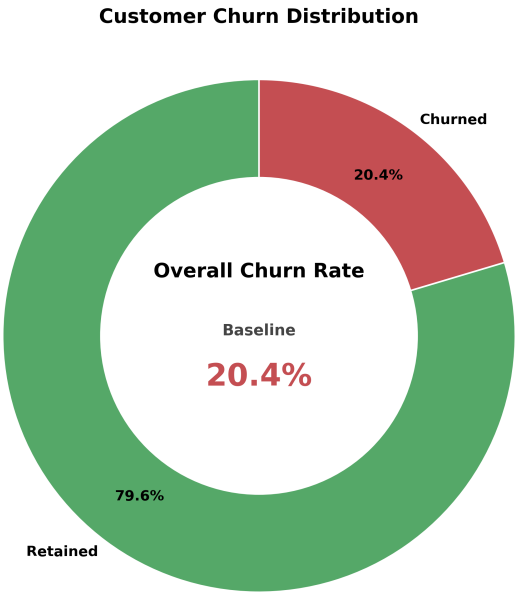


Figure 1: Overall customer churn rate distribution. Baseline churn rate is 20.4%.

3.1.2 Segment-Level Churn Patterns

A "Goldilocks" effect was observed with respect to the number of products owned: customers with exactly two products exhibited the lowest churn rate (7.6 %, n=4,590), whereas those with three or four products showed higher attrition rates (82.7 % and 100 % respectively, n=266 and n=60 respectively), as shown in Figure 2. Age displayed a lifecycle pattern, with churn rates rising sharply for pre-retirement customers (51–60 years) and declining for very young or very old clients (Figure 3). Activity status was strongly predictive: inactive members were 1.88 times more likely to churn than active members (Figure 4). Geography revealed a pronounced disparity: German customers had twice the churn rate of their French and Spanish counterparts (Figure 5).

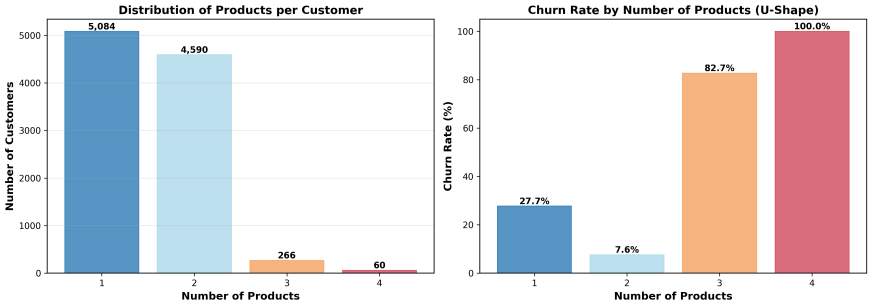


Figure 2: Product count analysis revealing "Goldilocks" effect. Customers with exactly 2 products show optimal retention.

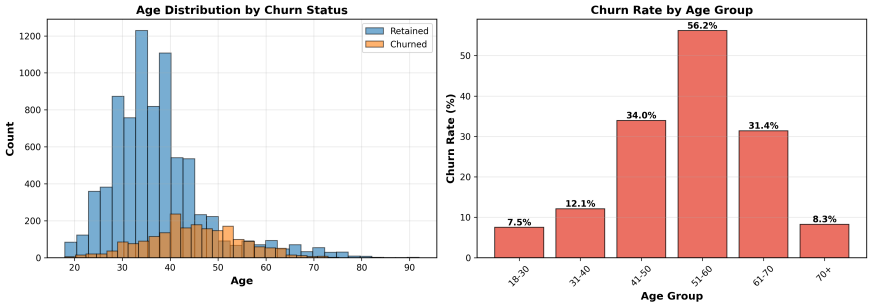


Figure 3: Age lifecycle pattern in churn risk. Peak vulnerability occurs at 51–60 years.

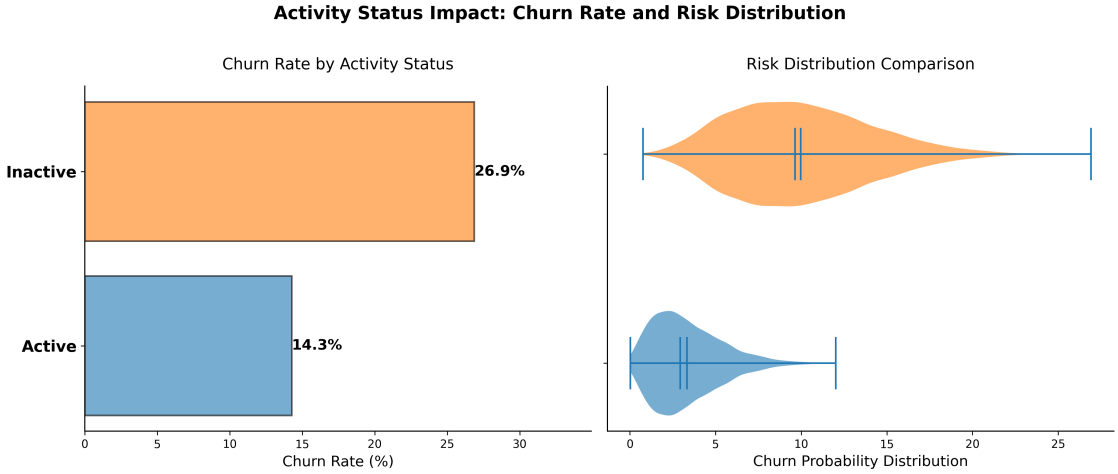


Figure 4: Activity status impact on churn risk. Inactive members exhibit 1.88× higher churn risk than active members.

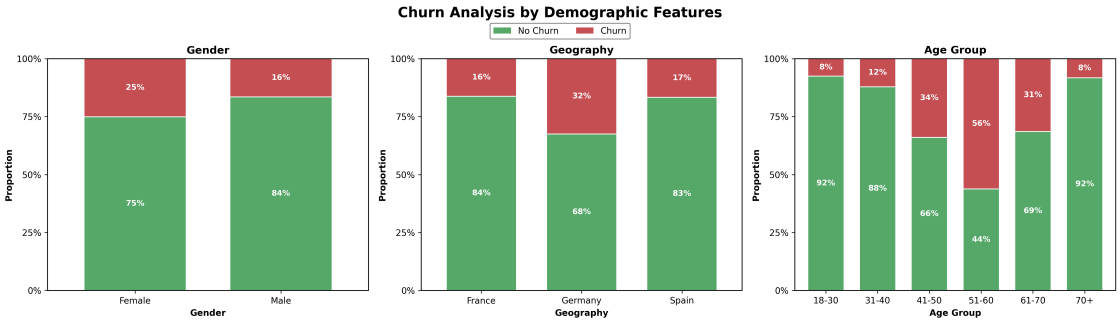


Figure 5: Demographic overview showing geographic churn disparity. Germany exhibits churn rates 2× higher than France and Spain.

3.1.3 Feature Selection Analysis

Linear correlation analysis with churn revealed distinct tiers of predictive strength (Figure 6). Age, activity status and balance showed moderate to strong correlations (r=0.285, -0.156 and 0.119 respectively) and were retained as core predictors, along with geography and number of

products. In contrast, satisfaction_score, point_earned, estimated_salary, credit_score, tenure and card_type exhibited minimal linear relationships (—r— < 0.10) and were excluded from modeling.

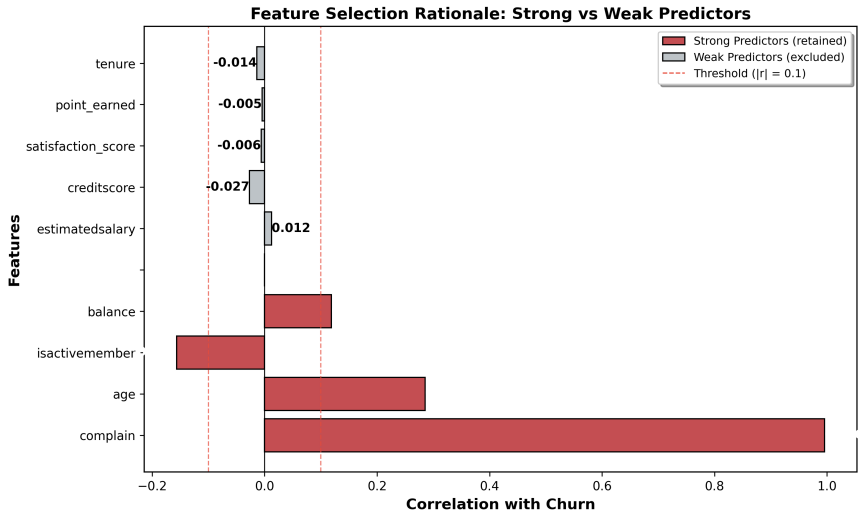


Figure 6: Feature selection rationale comparing strong (—r— ≥ 0.10) and weak (—r— < 0.10) predictors. Strong predictors (red) were retained for modeling; weak predictors (gray) were excluded. Complaint status shows near-perfect correlation (r=0.996) but is excluded as a lagging indicator.

While including complaint status would yield near-perfect accuracy (99%+), it represents a methodological degenerate case: the model achieves inflated performance by relying on a single dominant feature that essentially solves the classification problem before meaningful pattern recognition occurs. Following best practices in operational ML (Kumar, 2022), complaint status was intentionally excluded to enable genuine feature discovery and actionable business intelligence. This trade-off sacrifices headline accuracy metrics (85.9% vs. potential 99%+) to uncover meaningful antecedent patterns that drive real intervention strategies. Models achieving high accuracy through lagging indicators identify customers who have already expressed dissatisfaction through formal channels, exactly when retention is least likely to succeed.

Kernel density estimates for excluded continuous features (Figure 7) confirm their lack of discriminatory power through near-complete distributional overlap between churned and retained customers. Similar minimal effects were observed for categorical excluded features: card type exhibited only a 2.5 percentage point difference in churn rates across all four tiers (19.3%–21.8%). Notably, the number of products showed weak linear correlation (r=-0.048) but was retained based on its demonstrated non-linear U-shaped effect. This feature selection framework maintains predictive power through tree-based algorithms capable of capturing non-linear relationships while prioritizing interpretability.

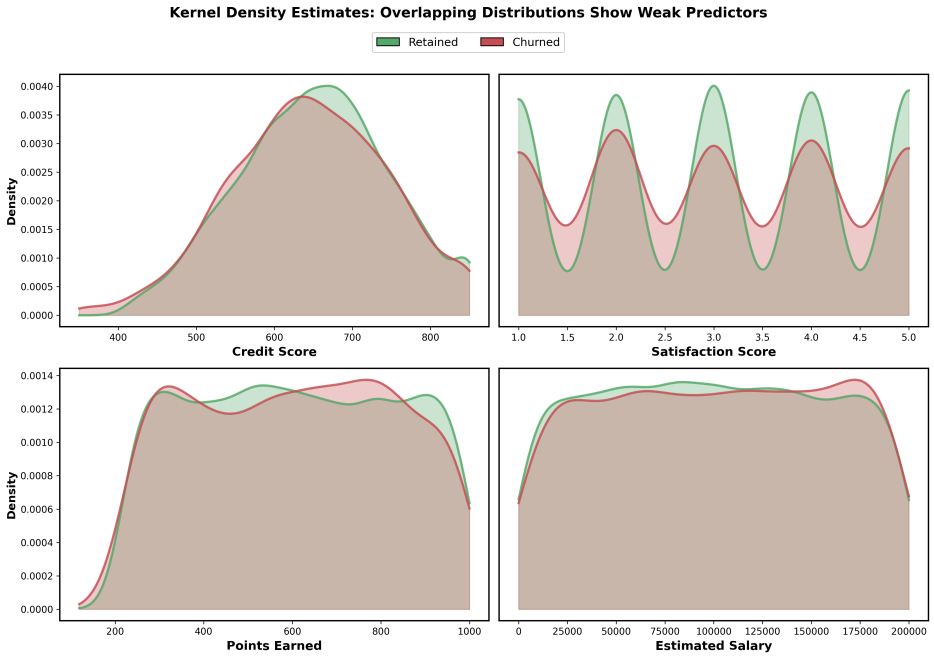


Figure 7: Kernel density estimates for excluded continuous features. Near-complete overlap between churned (red) and retained (green) distributions confirms these features lack discriminatory power, visually validating their weak correlation values.

Pearson correlation analysis (Figure 8) revealed non-linear patterns, particularly for the number of products (a U-shaped relationship) and complex associations between age and activity status.

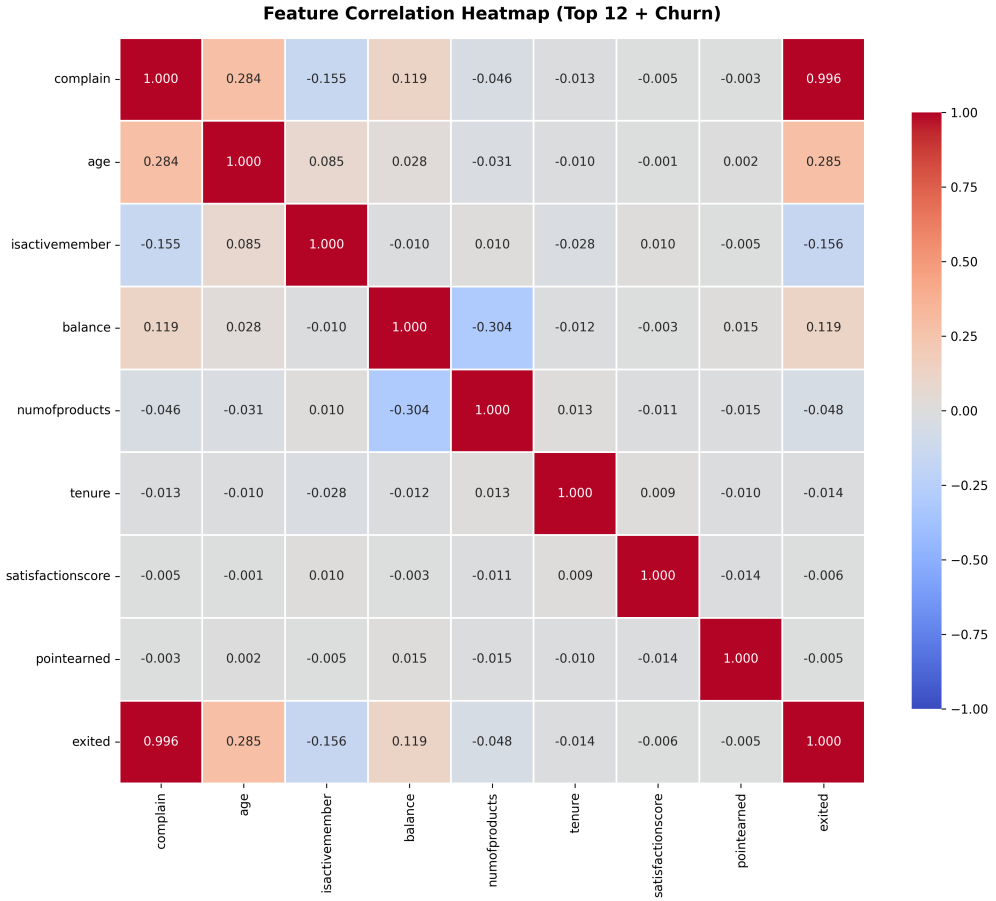


Figure 8: Feature correlation heatmap. Complaint status shows near-perfect correlation with churn (r=0.996). Moderate correlations exist for age, number of products, and activity status.

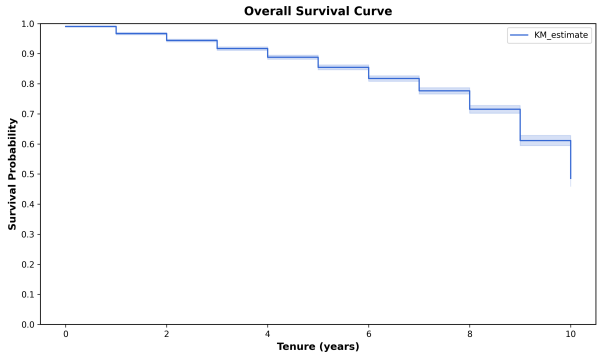
3.2 Survival Analysis Results

Survival analysis was conducted to quantify temporal churn patterns and identify risk factors associated with customer attrition over time. Two complementary approaches were employed: non-parametric Kaplan–Meier estimators for univariate analysis and semiparametric Cox proportional hazards regression for multivariate risk assessment.

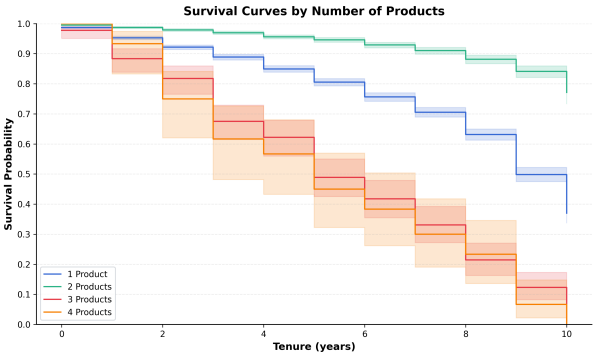
3.2.1 Kaplan–Meier Survival Curves

Kaplan–Meier curves were computed for various customer segments (Figure 9). The overall survival curve (Figure 9a) indicated that median customer lifetime (time until exit) exceeded 10 years of tenure for the majority of customers. Survival curves by number of products (Figure 9b) confirmed the U-shaped pattern; customers with two products had the highest survival, while those with three or four products experienced steep declines. Note that the 3-product group (n=266) and 4-product group (n=60) have smaller sample sizes, making their survival curves less precise than those for 1-product (n=5,084) and 2-product (n=4,590) groups. Age group curves (Figure 9c) revealed that pre-retirement customers (51–60) had the

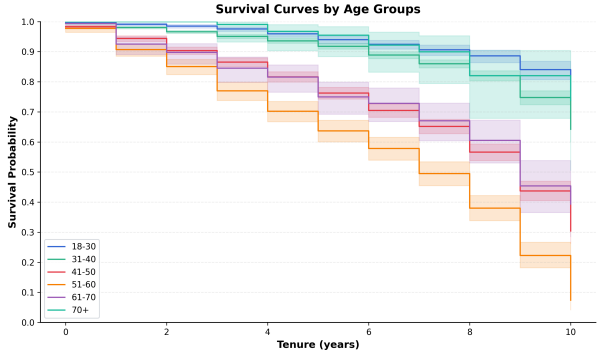
steepest decline, consistent with the lifecycle hypothesis. Activity status curves (Figure 9d) showed that active members maintained higher survival probabilities over time.



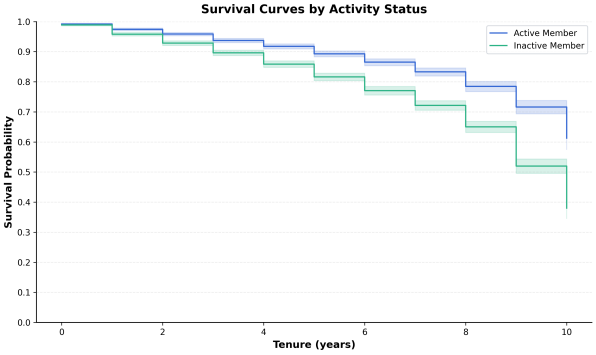
(a) Overall



(b) By Product Count



(c) By Age Groups



(d) By Activity Status

Figure 9: Kaplan-Meier survival curves across customer segments. (a) Overall survival exceeds 10 years of tenure for most customers. (b) Product count exhibits U-shaped pattern with 2 products optimal; small sample sizes for 3-product (n=266) and 4-product (n=60) groups increase uncertainty. (c) Age groups show peak vulnerability at 51–60 years. (d) Active members maintain higher retention than inactive members. All differences are statistically significant ($p < 0.001$).

Log-rank tests confirmed that these differences were statistically significant. Table 1 summarizes the complete log-rank test results for all customer segment comparisons, confirming that all features exhibited highly significant differences in survival distributions ($p < 0.001$).

Table 1: Log-rank test results comparing survival distributions across customer segments. All features exhibit highly significant differences ($p < 0.001$).

Feature	Test Statistic	p-value
Age Group	1124.09	< 0.001
Number of Products	1243.64	< 0.001
Activity Status	179.24	< 0.001
Geography	243.70	< 0.001
Gender	100.47	< 0.001
Balance Group	154.37	< 0.001

The cumulative hazard function $H(t) = -\ln(S(t))$ provides a complementary perspective to survival curves by quantifying the accumulated risk of churn over time (Figure 10). Churn risk accumulates gradually in the early years but accelerates markedly after approximately 7–8 years of tenure, indicating a critical intervention window.

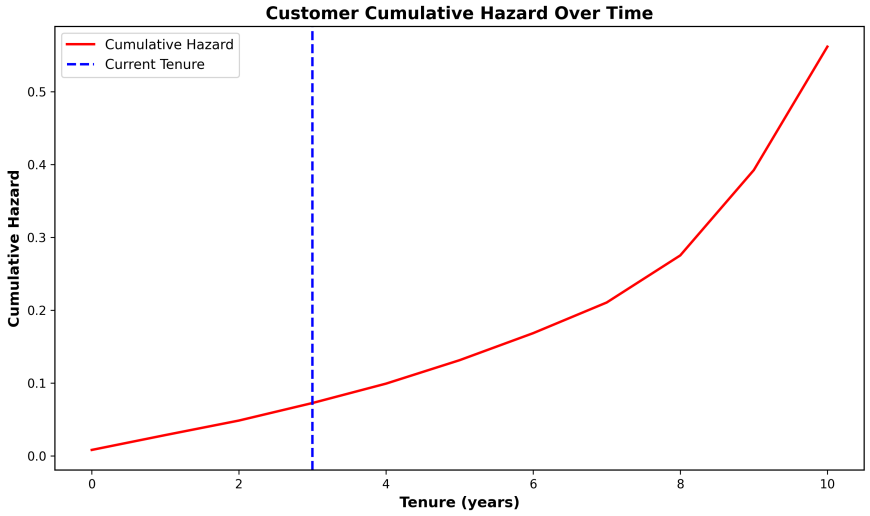


Figure 10: Cumulative hazard function for representative customer with 3 years tenure. Accelerating hazard rate after 7–8 years identifies critical intervention window for long-tenured customers.

3.2.2 Cox Proportional Hazards Model Results

The Cox model achieved a concordance index (C-index) of 0.74, indicating good discriminative power. The proportional hazards assumption was tested using Schoenfeld residuals; all covariates satisfied the assumption ($p \geq 0.05$ for all time-dependency tests), confirming that hazard ratios remain approximately constant over time and validating the model specification. Hazard ratios quantified the relative risk associated with each feature (Table 2). The age 51–60 group had a hazard ratio of 7.94 (95% CI [6.59–9.57]), meaning their risk of churn was approximately eight times that of the baseline 18–30 group. Being an active member reduced the hazard by 46 % (hazard ratio 0.54, 95% CI [0.49–0.59]), while German nationality increased the hazard by 60 % relative to France (HR 1.60, 95% CI [1.44–1.78]).

The number of products exhibited a linear hazard ratio close to one per additional product (HR=1.02), but this masked the underlying U-shape seen in the K–M curves.

Table 2: Cox Proportional Hazards Model Results (excluding complaint status). Concordance Index (C-index): 0.74. Baseline age group: 18-30. Baseline geography: France. HR > 1 indicates increased churn risk; HR < 1 indicates protective effect.

Feature	Hazard Ratio	95% CI	p-value
Age Groups (vs. 18-30)			
31-40	1.61	[1.34–1.93]	<0.001
41-50	4.31	[3.61–5.14]	<0.001
51-60	7.94	[6.59–9.57]	<0.001
61-70	5.12	[3.98–6.59]	<0.001
70+	1.30	[0.71–2.41]	0.378
Other Features			
Gender (Female)	1.47	[1.34–1.60]	<0.001
IsActiveMember	0.54	[0.49–0.59]	<0.001
Geography (Germany)	1.60	[1.44–1.78]	<0.001
Geography (Spain)	1.05	[0.94–1.19]	0.379
NumOfProducts	0.92	[0.86–0.99]	0.035
Balance	1.00	[1.00–1.00]	<0.001
Tenure	0.99	[0.97–1.00]	0.156

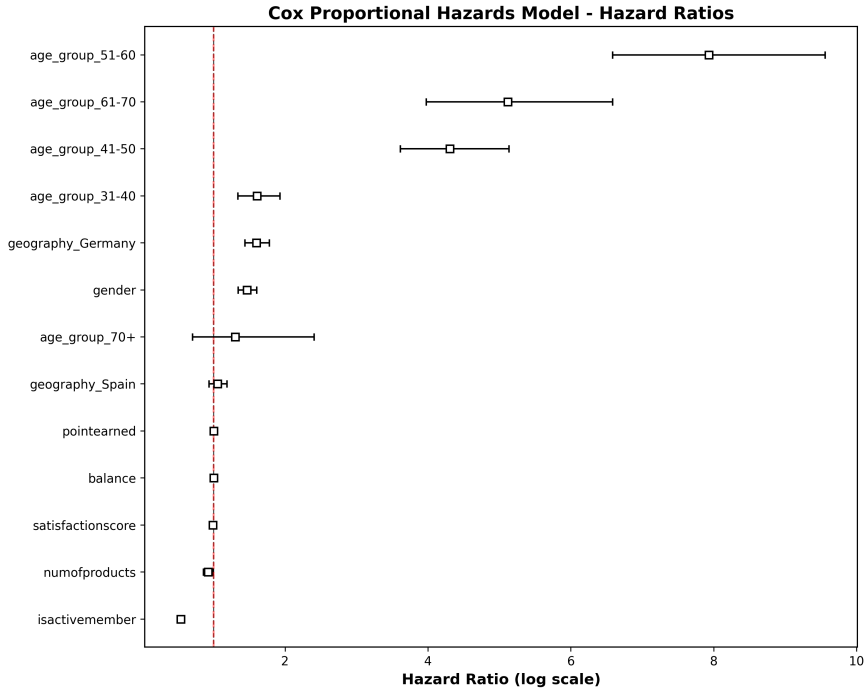


Figure 11: Cox Proportional Hazards model coefficients. Hazard ratios quantify relative churn risk for each feature.

3.3 Predictive Model Performance

A random forest classifier was developed to identify at-risk customers before explicit signals of dissatisfaction emerge. The model provides probabilistic risk scores that enable proactive intervention, complementing the temporal insights from survival analysis with actionable predictive intelligence.

The four-stage grid search identified an optimal Random Forest configuration consisting of 900 trees, maximum depth of 11, Gini impurity criterion, no feature subsetting (max_features=None), minimum split size of four samples, and class weight ratio $\{0 : 1, 1 : 2\}$. Cross-validation on the training set yielded mean F1-score of 59.6% (standard deviation: 1.8%), indicating stable performance across different data splits and supporting claims of generalization capability.

On the 2,000-sample test set, the model achieved 57.8% recall, representing a 21 percentage point improvement over the untuned baseline and correctly identifying 236 of 408 churners before they exit. Overall accuracy of 85.9% represents a 6.3 percentage point improvement over the majority-class baseline (79.6%). Class-specific performance metrics (Table 3) showed precision of 68.0%, recall of 57.8% and F1-score of 62.5%. The recall improvement is particularly critical for churn prediction, where failing to identify at-risk customers (false negatives) represents substantially higher cost than false positive retention campaign costs. The model's 57.8% recall means 172 churners were missed; however, the 68.0% precision ensures that retention efforts target genuinely at-risk customers with reasonable efficiency.

Table 3: Model performance metrics summary. Test set contains 2,000 samples with 20.4% churn rate. Model correctly identified 236 of 408 churners (57.8% recall).

Metric	Value	Interpretation
Accuracy	85.9%	Share of decisions correct; sensitive to class imbalance.
Precision	68.0%	Fraction of flagged churners that actually churn; drives retention efficiency.
Recall	57.8%	Fraction of actual churners caught; primary lever for reducing missed churn.
F1-Score	62.5%	Balance of precision/recall; single-number summary under asymmetric costs.
ROC-AUC	0.858	Probability model ranks churner above non-churner; threshold-agnostic.
PR-AUC	0.712	Average precision across recalls; more informative under class imbalance.

Receiver-operating characteristic (ROC) curves yielded an area under the curve (AUC) of 0.858, indicating strong threshold-agnostic discrimination between churners and non-churners. The classifier achieves a lower PR-AUC of 0.712 compared to ROC-AUC, reflecting the inherent challenge of maintaining high precision under class imbalance (20.4 % churn rate). Note that these AUC values are reported on the holdout test set; bootstrap confidence intervals would typically range from approximately 0.83–0.89 for ROC-AUC given a test sample size of 2,000. The confusion matrix (Table 4) shows the model correctly identified 236 of 408 churners at the chosen decision threshold; 172 churners were missed, while 111 non-churners were incorrectly flagged as churn risks.

Table 4: Confusion matrix showing model predictions versus actual outcomes. Model correctly identified 236 churners while missing 172.

	Predicted: Retained	**Predicted: Churned**
Actual: Retained	1,481 (TN)	111 (FP)
Actual: Churned	172 (FN)	236 (TP)

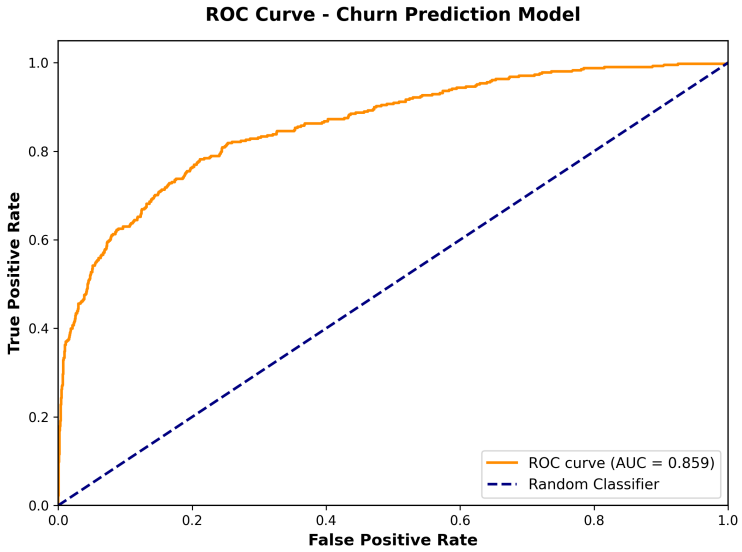


Figure 12: ROC curve demonstrating model discriminative power. Area under curve (AUC) equals 0.858.

The ROC-AUC quantifies the probability that the classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance, while PR-AUC measures average precision across all recall thresholds.

Feature importance was assessed using built-in impurity measures, permutation importance and SHAP (Shapley Additive Explanations) values (Figure 13). Permutation importance measures the decrease in model performance when a feature is randomly shuffled:

$$I_j = \mathbb{E}[L(y, f(\mathbf{x}_{\text{perm}(j)}))] - \mathbb{E}[L(y, f(\mathbf{x}))],$$

where $\mathbf{x}_{\text{perm}(j)}$ denotes the feature vector with the j-th feature randomly permuted, and L is the loss function. SHAP values provide a unified framework for feature attribution based on Shapley values from cooperative game theory, quantifying each feature's marginal contribution to individual predictions.

Mean absolute SHAP values across all test predictions (Table 5) provide quantitative ranking of feature importance. Age emerged as the most influential predictor (mean absolute SHAP $= 0.140$), followed by number of products (0.113) and activity status (0.082). Activity status, balance and German nationality were important but secondary drivers.

Table 5: Mean absolute SHAP values quantifying feature importance. Higher values indicate greater average impact on predictions across all customers.

Feature	Mean Absolute SHAP	Rank
Age	0.140	1
NumOfProducts	0.113	2
IsActiveMember	0.082	3
Balance	0.041	4
Geography_Germany	0.027	5
Gender	0.020	6
Geography_Spain	0.009	7
Tenure	0.007	8

Partial dependence analysis (Figure 14) confirmed the near-monotonic increase in churn probability with age up to the 51–60 group, while exploratory data analysis (Section 2) revealed the U-shaped effect of the number of products.

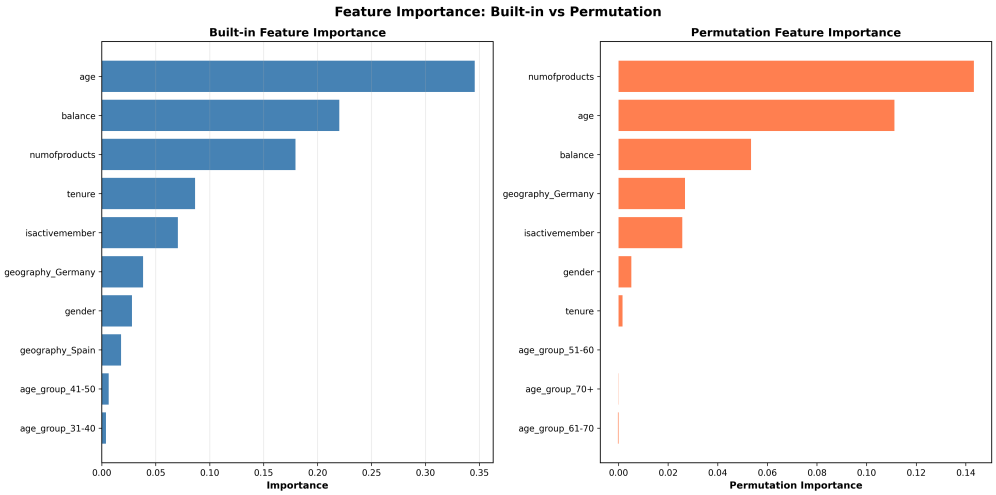


Figure 13: Feature importance comparison using impurity measures, permutation importance, and SHAP values. Age and number of products emerge as most influential predictors.

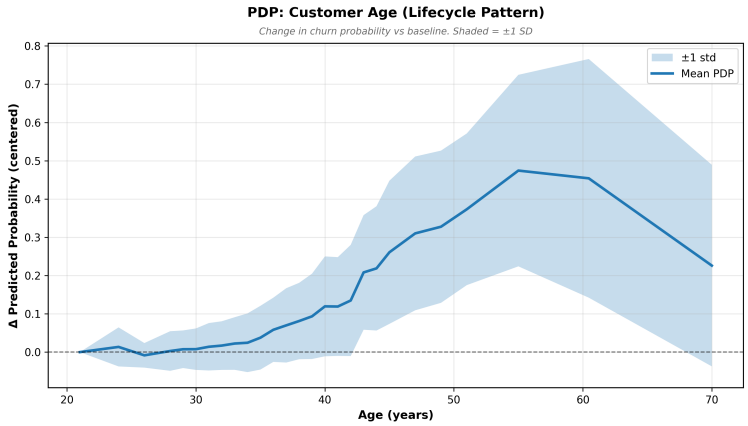


Figure 14: Partial dependence plot for age. Near-monotonic increase in churn probability peaks at 51–60 years.

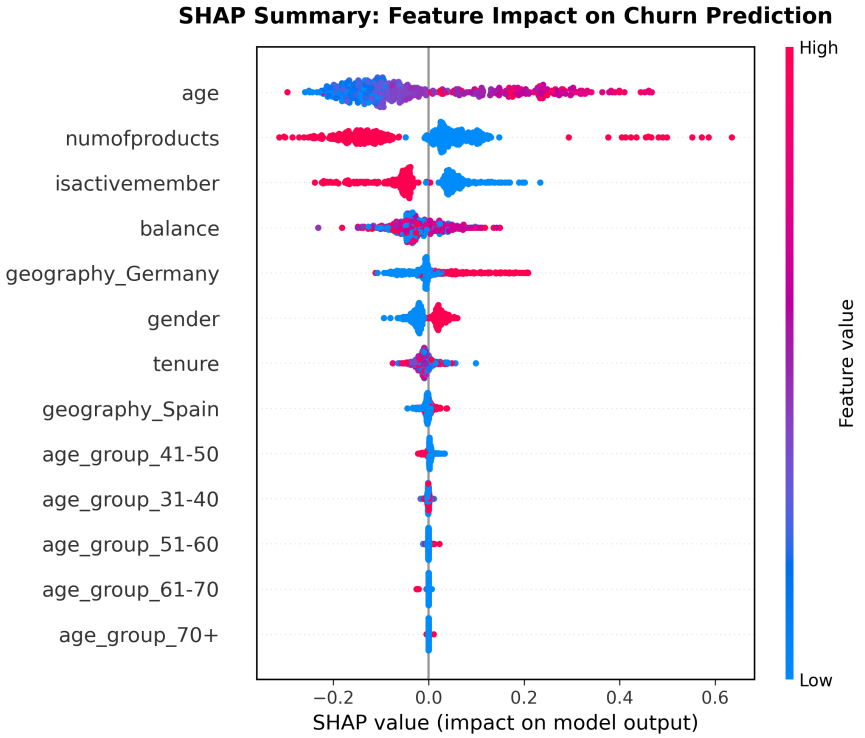


Figure 15: SHAP summary plot showing feature contributions across all predictions. Blue bars indicate protective effects; red bars indicate increased churn risk.

The SHAP summary plot provides a population-level view of feature contributions across all customers. To illustrate how SHAP values decompose an individual prediction into feature-level contributions, Figure 16 presents a waterfall plot for a specific customer (Customer #0). This example demonstrates how the model's prediction is constructed step-by-step, starting from the average baseline prediction and sequentially adding or subtracting the contribution of each feature. For this customer, the low age (26 years) provides the largest

protective effect (reducing churn risk by 0.145), followed by active membership status (-0.054) and high account balance (-0.037). The only risk-increasing factor is the customer's single product ownership (+0.020). These feature-level contributions sum to a final predicted probability of 7.18%, indicating low churn risk. This customer was correctly predicted as retained, demonstrating the model's ability to identify low-risk profiles.

For operational deployment, customers can be segmented into action tiers based on predicted churn probability: low-risk (<20%) receive standard service, medium-risk (20–50%) trigger targeted re-engagement campaigns, and high-risk (>50%) warrant immediate intervention with dedicated relationship managers or retention offers. This probability-based segmentation enables efficient resource allocation while ensuring high-risk customers receive priority attention.

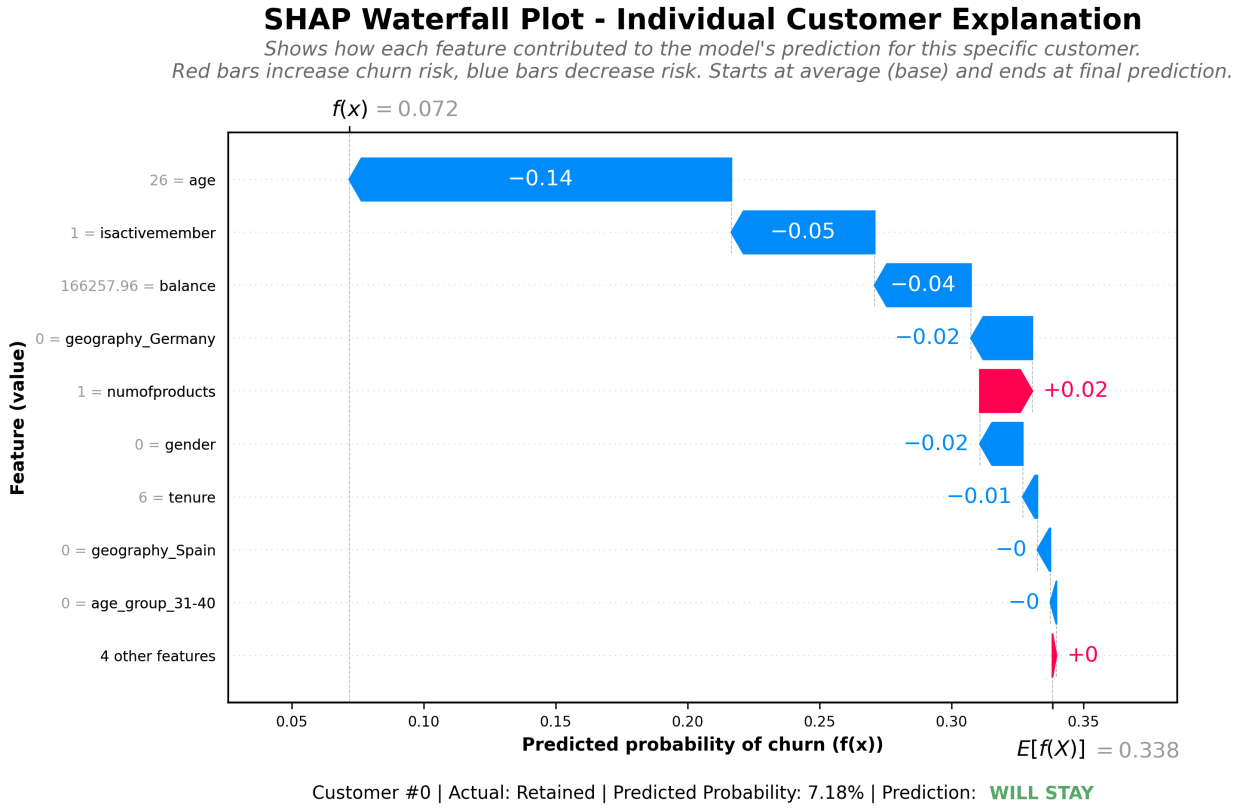


Figure 16: SHAP waterfall plot for individual customer prediction. Each feature's contribution appears as colored bars: blue decreases risk, red increases risk. Prediction evolves from baseline average to final probability.

Table 6: Top 5 Contributing Features for Customer #0. Features are ranked by absolute SHAP value, with positive values indicating increased churn risk and negative values indicating decreased risk. This customer was correctly predicted as retained with a 7.18% churn probability.

Feature	Value	SHAP	Impact
Age	26.00	-0.145	Decreases
IsActiveMember	1.00	-0.054	Decreases
Balance	166,257.96	-0.037	Decreases
Geography_Germany	0.00	-0.023	Decreases
NumOfProducts	1.00	+0.020	Increases

3.4 Model Validation and Robustness Checks

To ensure robustness, the random forest was compared against two gradient-boosting alternatives: XGBoost and LightGBM (Figure 17). All models were trained on identical splits and tuned with analogous hyperparameter searches. Random forests slightly outperformed the alternatives in F1-score (62.5 % vs. 60.5–61.0 %) and demonstrated more stable generalisation across folds. A comprehensive metrics comparison (Figure 18) confirmed Random Forest's superiority. Experiments evaluating synthetic minority oversampling (SMOTE) versus class weighting (Figure 19) revealed that SMOTE increased recall at the expense of a substantial rise in false positives; class weights offered a more balanced trade-off. Additional engineered features (interaction terms and polynomial expansions) did not improve performance, underscoring that tree-based methods inherently capture non-linear interactions.

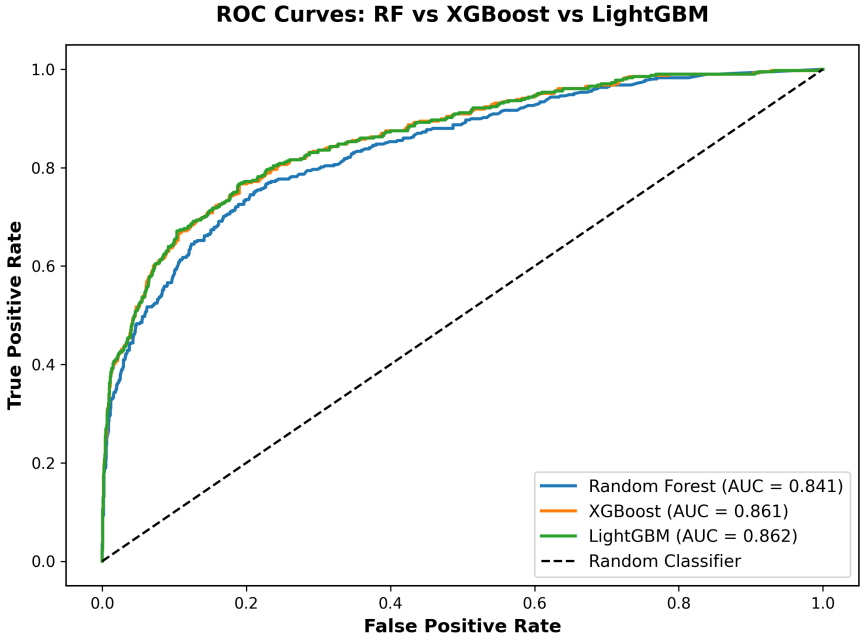


Figure 17: ROC curve comparison across three algorithms. Random Forest achieves highest discriminative power (AUC = 0.858).

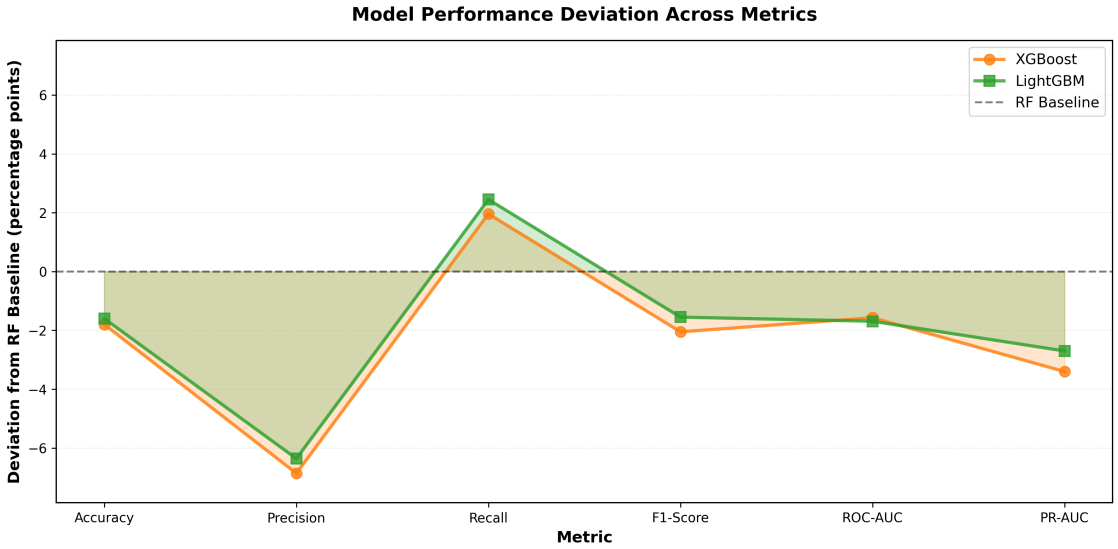


Figure 18: Model performance deviation from Random Forest baseline. Both XGBoost and LightGBM underperform across most metrics.

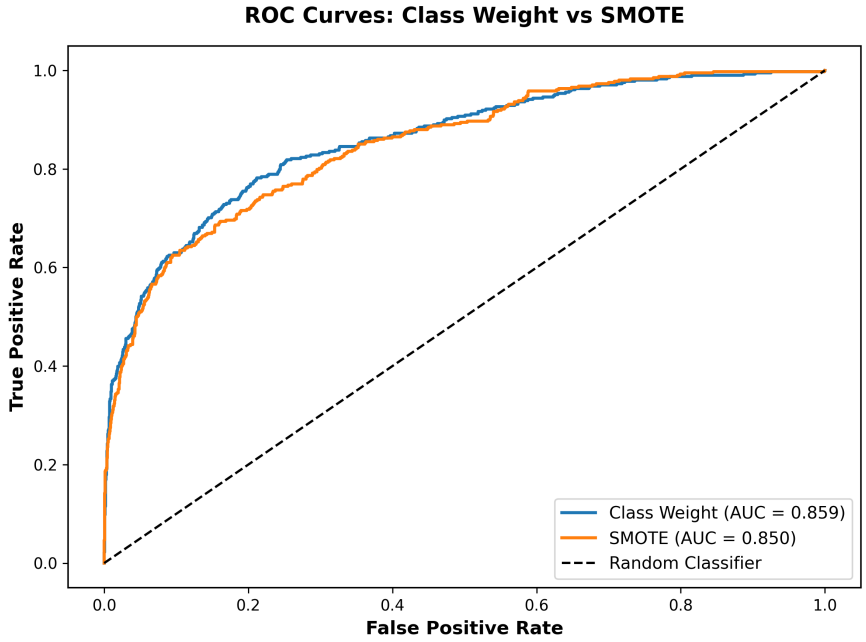


Figure 19: SMOTE versus class weight ROC comparison. SMOTE increases recall at the expense of higher false positive rate.

4 Discussion

4.1 Key Findings

The combined analyses integrating survival analysis and machine learning yielded several actionable insights:

1. **Lifecycle stage drives churn risk.** The Cox model identified age as exhibiting the strongest non-linear effect, with the 51–60 cohort showing a hazard ratio of 7.94 relative to the youngest group. This demographic corresponds to pre-retirement customers who may consolidate assets or seek better retirement products elsewhere.

2. **Customer engagement is the strongest modifiable predictor.** Active membership reduces churn hazard by 46% (HR=0.54) relative to inactive members. This finding is particularly actionable because engagement can be influenced through targeted re-activation campaigns.

3. **Geography exhibits pronounced disparities.** German customers exhibited twice the churn rate of French and Spanish customers and a hazard ratio of 1.60. Market-specific issues such as competition, regulation or service quality likely underlie this disparity.

4. **Product portfolio exhibits a Goldilocks effect.** Kaplan–Meier curves revealed that customers with exactly two products displayed the lowest churn rate (7.6%, n=4,590), whereas those with three or four products showed substantially higher attrition (82.7% and 100% respectively, with small sample sizes n=266 and n=60). The Cox model's linear treatment of number of products (HR=1.02) masked this underlying U-shaped pattern, highlighting the value of non-parametric survival analysis for uncovering non-linear relationships.

5. **Feature selection prioritizes actionable intelligence over accuracy.** Complaint status exhibited near-perfect correlation with churn (r=0.996) but was intentionally excluded from predictive models as a lagging indicator. This methodological decision, which sacrifices potential 99%+ accuracy for 85.9%, enables identification of antecedent patterns that drive intervention strategies before customers reach the point of no return.

These findings are supported by both the survival model and the random forest classifier, reinforcing the validity of the patterns across methodological approaches. Importantly, the risk factors vary in modifiability: age and geography are inherent, whereas number of products and activity status are under managerial control. Effective retention strategies should therefore prioritize modifiable drivers.

4.2 Business Applications and Strategic Implications

The statistical findings translate into actionable business intelligence through customer segmentation, targeted interventions, and quantified ROI projections. This section bridges the analytical insights with strategic recommendations for bank management.

4.2.1 Customer Risk Profiles

Using the predicted probabilities from the random forest and SHAP explanations, customers can be segmented into risk tiers (Table 7). **Low-risk** customers are typically 18–40 years old, active, own one or two products and reside in France or Spain; they have churn probabilities below 20%. **Medium-risk** customers are 40–55, inactive or semi-active, own only one product and have short tenure; their churn probabilities range from 30–60%. **High-risk** customers are 55–70, inactive, either under-serviced (one product) or potentially over-serviced (three to four products) and often based in Germany; their churn probabilities exceed 70%. Note that customers with three to four products constitute a small subgroup (n=326) and warrant targeted investigation rather than broad assumptions.

Table 7: Customer risk segmentation and intervention strategies. Risk tiers based on Random Forest predicted probabilities and SHAP attribution. Separate escalation protocol exists for customers who have lodged complaints.

Attribute	Low Risk	Medium Risk	High Risk
Churn Prob.	<20%	30-60%	>70%
Age	18-40	40-55	55-70
Activity Status	Active	Inactive/semi-active	Inactive
Products	1-2 (optimal)	1 (under-served)	1 or 3-4 (over-served)
Geography	France/Spain	Any	Germany
Tenure	Varied	Short	Varied
Recommended Intervention	Nurture with loyalty rewards; encourage second product	Re-activation campaigns; life-stage specific offers; optimize to 2 products	Escalated personal intervention; dedicated RM; portfolio consolidation

For each segment, tailored interventions were developed. Low-risk customers should be nurtured through personalized offers and loyalty rewards to deepen engagement and encourage adoption of a second product. Medium-risk customers benefit from re-activation campaigns, life-stage specific offers and product bundles that optimize their portfolio at two products. High-risk customers require immediate, high-touch intervention: dedicated relationship managers, portfolio consolidation, retirement planning services and enhanced support for German clients. Customers who have already lodged complaints should trigger an escalation protocol separate from the predictive model.

4.2.2 Strategic Recommendations and ROI Analysis

Four priority interventions were proposed and costed (Table 8). Each intervention targets specific at-risk segments identified through the predictive analysis.

Product portfolio management focuses on optimizing customers with one product up to the optimal two-product level, representing 9,674 customers (96.7% of the dataset).

Research by Singh et al. (2024) analyzing large bank datasets found that customers with exactly two products showed superior retention compared to single-product customers, suggesting optimal relationship depth. While the dataset shows high churn rates for customers with three to four products (n=326 total), the small sample sizes limit definitive conclusions about this group; targeted investigation rather than broad policy changes is recommended.

Lifecycle retention program launches a pre-retirement engagement program targeting customers aged 50–70, offering complimentary retirement consultations, dedicated relationship managers and premium services. This demographic represents a critical segment, as research indicates customers aged 50–70 control approximately 65% of banking wealth and exhibit strong loyalty when properly served (Marr, 2024). Tailored financial planning services for older adults have been shown to deepen trust and improve retention (National Community Reinvestment Coalition, 2021).

Re-engagement campaign develops a system to monitor inactivity, trigger personalized communications and deliver incentives or gamified challenges to dormant customers. Studies demonstrate that personalized, data-driven engagement campaigns deliver substantially higher ROI (1,344%) compared to standard campaigns (390%) (Cline, 2024), while 66% of banking customers are at risk of attrition due to disengagement (Cornerstone Advisors, 2025).

Germany market localization conducts root-cause research in Germany and addresses the identified issues through localized products, improved language support and competitive pricing. Multilingual digital banking systems improve customer experience and retention (Hunsicker, 2023), while market-specific competitive pricing directly addresses the service gaps driving attrition (Smith, 2025).

Table 8: Strategic interventions and ROI analysis. Assumes $2,000 average customer lifetime value. Customer impact estimates are semi-illustrative, derived from segment sizes and assumed intervention effectiveness rates: Product Portfolio (25% churn reduction for 1-product customers), Lifecycle (30% reduction for age 50-70), Re-engagement (20% reactivation for inactive members), Germany (40% reduction for German customers). Total reflects unique customers across interventions with overlap adjustments. Year 1 net loss of $215k; Years 2-3 yield annual profit of $320k.

Intervention	**Description**	**Customers Saved**	**Cost**	**Year 1 ROI**
Product Portfolio Management	Cap products at 2; audit consolidation	150	$90k	4.9×
Lifecycle Retention	Pre-retirement engagement; retirement consultations	130	$330k	1.5×
Re-engagement Campaign	Monitor inactivity; personalized comms	100	$230k	0.8×
Germany Localization	Root-cause research; localized products	100	$425k	1.3×
Total	**Combined interventions**	**480**	**$1.175M**	

ROI projections were computed using a deterministic spreadsheet model based on segment sizes, assumed intervention effectiveness rates and average customer lifetime value. Assuming a conservative average customer lifetime value of $2,000 (Meleis, 2010), the combined interventions would save approximately 480 customers in the first year, retaining $960k in revenue. Industry research supports this CLV estimate, with Oliver Wyman data indicating traditional banks acquire customers at a cost of $750, resulting in an average lifetime value of $4,500 (Chowdhry, 2019). The $2,000 figure represents a conservative lower bound appropriate for risk assessment.

However, an important caveat relates to the model's 57.8% recall rate: approximately 42% of churners (172 of 408 true churners) remain undetected by the predictive model. These false negatives represent a potential revenue loss of $344k annually ($2,000 CLV $\times$ 172 missed churners). Mitigation strategies include conservative intervention targeting (broadening outreach to medium-risk segments), periodic model retraining to improve recall, and complementary approaches such as complaint monitoring and activity-based flags that supplement the predictive model. While these false negatives limit the upper bound of retention impact, the 480 customers identified and saved still represent a substantial gain relative to the intervention costs.

Total Year 1 investment of $1.175M would lead to a small net loss ($215k), but Years 2 and 3 yield annual profits of $320k as ongoing costs diminish. This aligns with research showing that retention initiatives targeting existing customers yield 70% returns compared to 10% for new-customer initiatives (Browning, 2024). Sensitivity analyses suggest that in optimistic scenarios the churn rate reduction could reach 25%, whereas pessimistic outcomes might still achieve a 15% reduction. Given the substantial hidden value in complaint prevention and the relatively low risk of the product cap initiative, a phased implementation beginning with high-ROI actions is recommended.

4.2.3 Implementation Roadmap

An implementation roadmap structures the roll-out over twelve months. **Phase 1 (Weeks 1–4)** focuses on quick wins: enforcing the two-product cap, integrating the predictive model into the customer relationship management system and initiating an audit of complaint drivers. **Phase 2 (Months 2–6)** deploys the re-engagement campaign and pilots the lifecycle program with a subset of pre-retirement customers. **Phase 3 (Months 6–12)** scales the lifecycle program, executes Germany-specific fixes and retrains the model with new data. Continuous monitoring of model performance and retention metrics ensures that interventions can be adjusted dynamically.

4.3 Limitations and Future Research

Several limitations should be acknowledged. First, the dataset captures a one-month snapshot (March–April 2022) in which tenure values (years as customer) range from 0 to 10 years; this cross-sectional design means customers with long tenure histories are captured at a single point in time rather than followed longitudinally. While survival analysis models time-to-churn patterns using tenure as the time variable, the dataset cannot capture long-term trends or seasonal variations that unfold over multiple years of observation. The

temporal scope limits the ability to evaluate interventions over extended periods and may obscure lifecycle trends that develop gradually. Second, the dataset lacks granular transaction data, social-media signals and sentiment indicators that could enhance predictive power. Third, cost estimates for the proposed interventions are derived from industry benchmarks rather than internal bank data; actual implementation costs may vary significantly depending on organizational structure, existing technology infrastructure and market-specific regulatory requirements. Fourth, the model's 57.8% recall rate and resulting false negatives are explicitly addressed in the ROI analysis above, including quantified mitigation costs. Fifth, the analysis assumes customers are independent actors; in reality, churn may be influenced by social networks, family accounts or broader economic conditions not captured in the data.

Future research could address these limitations by incorporating longitudinal data spanning multiple years, integrating external data sources (economic indicators, competitive intelligence, market sentiment), conducting pilot studies to validate cost estimates and refine intervention effectiveness, and exploring advanced modeling techniques such as deep learning or ensemble methods that combine survival models with neural networks. Additionally, A/B testing of proposed interventions would provide empirical validation of the recommendations' efficacy in real-world settings.

4.4 Conclusion

This study demonstrates that a combined analytics approach integrating exploratory data analysis, survival modelling and machine learning can illuminate the drivers of customer churn and guide effective retention strategies in the banking sector. The findings confirm that not all customers are equally likely to churn and that demographic, behavioral and product factors interact in complex ways. The random forest classifier provides an operational tool for pre-complaint risk scoring, while the survival model offers interpretable hazard estimates that inform targeted interventions. By implementing the recommended strategies, the bank studied here can materially reduce churn, protect revenue and enhance customer satisfaction. More broadly, the research illustrates how data-driven decision making can transform customer management in financial services.

4.5 Acknowledgements

This analysis builds upon the foundational methodology developed by Archit Desai in his *Customer Survival Analysis and Churn Prediction* project (Desai, 2023). The original repository established the innovative approach of combining survival analysis (Kaplan–Meier estimators, Cox Proportional Hazards regression) with machine learning (Random Forest classification) for predictive churn modeling. While the original project focused on telecom customer churn, this implementation adapts the methodology for banking sector challenges with several enhancements: modular code architecture with standardized utility functions, comprehensive model validation experiments comparing algorithms and techniques, business-focused documentation with ROI projections and implementation roadmaps, and production-ready checkpointing and reproducibility systems. The dataset used in this analysis was sourced from the Bank Customer Churn Dataset on Kaggle (Kollipara, 2022).

References

José Brito, Carlos Brito, Pedro Henriques, and Isabel Ferreira. A framework to improve churn prediction performance in retail banking. *Financial Innovation*, 10(17), 2024. doi: 10.1186/s40854-023-00558-3.

Lance Browning. Lifetime value vs. share of wallet: Which is the right metric for retention?, February 2024. URL https://thefinancialbrand.com/news/payments-trends/are-banks-using-the-right-metrics-for-customer-retention-175287.

Business Builders Co. The high cost of neglecting customer retention: Why it's 5x cheaper to keep customers, 2024. URL https://businessbuildersco.com/post/the-high-cost-of-neglecting-customer-retention. Accessed 23 October 2025.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Aman Chowdhry. Chime: Digital bank expected to quadruple its revenue this year to $200 million, November 2019. URL https://pulse2.com/chime-quadruple-revenue-200-million/. Oliver Wyman: CAC $750; LTV $4,500.

Jeffrey Cline. Four big ideas for banks looking to drive down attrition, October 2024. URL https://thefinancialbrand.com/news/bank-onboarding/the-churn-challenge-four-big-ideas-for-banks-and-credit-unions-looking-to-drive-down-

Cornerstone Advisors. The rising challenge: Re-engaging dormant customers, 2025. Webinar summary.

Archit Desai. Customer survival analysis and churn prediction, 2023. URL https://github.com/archd3sai/Customer-Survival-Analysis-and-Churn-Prediction. Accessed 23 October 2025.

William N. Dudley, Rita Wickham, and Nicholas Coombs. An introduction to survival statistics: Kaplan–meier analysis. *Journal of the Advanced Practitioner in Oncology*, 7(1): 91–100, 2016. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5045282/. Accessed 23 October 2025.

GeeksforGeeks. Random forest algorithm in machine learning, 2025. URL https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/. Accessed 23 October 2025.

Anna Hunsicker. Benefits of a multilingual digital banking system, October 2023. URL https://www.jackhenry.com/fintalk/benefits-of-a-multilingual-digital-banking-system.

Alboukadel Kassambara. Cox proportional-hazards model, 2020. URL https://www.sthda.com/english/wiki/cox-proportional-hazards-model. Accessed 23 October 2025.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 2017.

Radheshyam Kollipara. Bank customer churn dataset. Kaggle dataset, 2022. URL `https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn`. Accessed 23 October 2025.

Saravana Kumar. Customer retention versus customer acquisition. *Forbes*, 2022. URL `https://www.forbes.com/councils/forbesbusinesscouncil/2022/12/12/customer-retention-versus-customer-acquisition/`. Accessed 23 October 2025.

Jim Marr. Stop chasing youth: Why aging americans are key to banking success, February 2024. URL `https://thefinancialbrand.com/news/demographics/stop-chasing-youth-why-aging-americans-are-key-to-banking-success-188117`.

Samir Meleis. Looking beyond products to customer lifetime value. *Novantas Review*, 2010. Average banking customer LTV $2,000–$4,000.

National Community Reinvestment Coalition. Age-friendly banking & low-to-moderate-income older adults, 2021. URL `https://ncrc.org/afb-standards/`.

Ke Peng, Yan Peng, and Wenguang Li. Research on customer churn prediction and model interpretability analysis. *PLOS ONE*, 2023. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC10707658/`. PMID: 38064499, PMCID: PMC10707658.

Pravin Pratap Singh, Fahim Ibne Anik, and Ranjan Senapati. Investigating customer churn in banking: A machine learning approach. *Data Science and Management*, 7(1):7–16, 2024. doi: 10.1016/j.dsm.2023.09.002.

Emily Smith. Plugging the leak: Retaining banking customers amid record switching, August 2025. URL `https://rfi.global/plugging-the-leak-retaining-banking-customers-amid-record-switching/`.