JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Understanding Customer Churn in Retail Banking through Explainable Predictive Analytics: Evidence of a Product Paradox

Patrick Ndabarishye * and Ajay Kumar Singh

Department of Computer Science and Engineering, Jain University, Bangalore- 562112 India;
e-mail : patrick.ndabarishye@gmail.com; ajay41274@gmail.com
* Corresponding Author : Patrick Ndabarishye

Abstract: The retention of customers in the retail banking sector is a critical economic imperative; however, predictive modeling is frequently hindered by severe class imbalance and the "Black Box" nature of complex algorithms. This study proposes a Heterogeneous Stacking Ensemble framework integrating XGBoost, CatBoost, and Random Forest base learners with a Logistic Regression meta-learner to forecast customer attrition. To overcome the pervasive "Majority Class Bias," we introduce a "Dual-Imbalance Defense" that synergizes the Synthetic Minority Over-sampling Technique (SMOTE) with algorithmic cost-sensitive penalization. Furthermore, moving beyond standard accuracy metrics, the framework mathematically derives a dynamic classification threshold to guarantee a strict 0.90 recall rate, actively optimizing the capture of at-risk capital. Model opacity is addressed through the integration of a SHapley Additive exPlanations (SHAP) TreeExplainer. This cooperative game theory approach provides localized, patient-level "Reason Codes" for regulatory compliance and reveals global systemic vulnerabilities, including non-linear drivers such as the "Product Paradox." Achieving a 0.90 recall rate and an AUC of 0.8654, this framework provides a statistically robust and operationally transparent tool for targeted customer retention.

Keywords: Customer Churn; Explainable AI (XAI); Financial Analytics; Machine Learning; Predictive Analytics; Retail Banking; SHAP; Stacking Ensemble.

1. Introduction

The modern retail banking sector operates in a highly saturated and competitive environment, making customer retention a critical imperative. Recognizing that the financial expenditure required for customer acquisition significantly exceeds the cost of retention [1], [2], institutions must prioritize proactive churn mitigation strategies. Consequently, the development of high-performance predictive frameworks to accurately forecast customer attrition has become a primary objective within financial predictive analytics[3]. However, modeling customer churn presents substantial computational challenges, primarily due to the inherent complexities of financial datasets, which are consistently characterized by severe class imbalance and intricate, non-linear feature interactions.

Traditional machine learning approaches frequently struggle to balance predictive accuracy with operational transparency. When addressing class imbalance, standard research methodologies typically force a mutually exclusive choice between data-level sampling, such as the Synthetic Minority Over-sampling Technique (SMOTE) [4], [5], and algorithmic cost-sensitive weighting [6], [7]. In isolation, these methods are often insufficient to overcome the pervasive "Majority Class Bias" found in banking data. Furthermore, while complex architectures such as stacking ensembles offer superior generalization, they are traditionally criticized as uninterpretable "Black Boxes" [8]. This lack of explainability restricts their utility in highly regulated financial environments, where actionable "Reason Codes" are required for targeted intervention. Additionally, standard classification paradigms default to a static 0.5 decision

Received: March, 16th 2026
Revised: April, 4th 2026
Accepted: April, 8th 2026
Published: April, 10th 2026

boundary, an approach that optimizes for general accuracy but fails to align the model's predictive caution with the institution's actual economic risk profile.

To bridge these methodological gaps, this study proposes a heterogeneous stacking ensemble approach designed specifically for customer churn mitigation and interpretability. This research introduces four distinct contributions to the field of banking analytics:

- Methodological Hybridization: We implement a "Dual-Imbalance Defense" that tightly couples synthetic instance generation with cost-sensitive class weights, engineering a model uniquely resilient to extreme class imbalances.
- Strategic Utility Calibration: Moving beyond accuracy-centric evaluation, this study introduces dynamic threshold recalibration. By mathematically deriving a custom decision boundary, the framework guarantees a strictly optimized 90% recall target, ensuring maximum capture of at-risk capital.
- Identification of the "Product Paradox": This research empirically challenges standard cross-selling paradigms—which assume increased product ownership guarantees loyalty—by identifying a non-linear friction point at the three-product threshold where churn risk increases exponentially.
- Multi-Model Interpretability: We successfully integrate SHAP-based (SHapley Additive exPlanations) decomposition [9] within a complex multi-model stack. This integration effectively opens the black box of meta-learner decisions, yielding global systemic drivers and localized, patient-level risk profiling.

Through these contributions, this paper demonstrates that sophisticated ensemble learning, when constrained by strict recall optimization and cooperative game theory-based explainability, provides an operationally transparent solution for retail banking retention. Despite the evolution of ensemble methods, a gap remains in the development of deployable pipelines that prioritize high recall while maintaining regulatory interpretability [10]. This study addresses that gap by coupling a heterogeneous stacking architecture with prescriptive optimization logic.

2. Literature Review

The application of machine learning for customer churn prediction in the banking sector has evolved significantly, transitioning from foundational statistical methods to complex, non-linear algorithmic architectures. This section reviews the current state-of-the-art across four critical domains: predictive modeling, class imbalance resolution, ensemble architectures, and model interpretability.

2.1. Evolution of Churn Prediction Models

Historically, financial institutions relied heavily on classical statistical models, such as Logistic Regression and Support Vector Machines, to identify at-risk customers. While these models offer baseline interpretability and are computationally inexpensive, they frequently fail to capture the complex, non-linear relationships and high-dimensional interactions inherent in modern banking datasets [11].

To address these limitations, recent research has shifted toward ensemble-based learning and broader machine learning frameworks [10], [12]–[16]. Algorithms such as Random Forest, Extreme Gradient Boosting (XGBoost), and LightGBM have consistently demonstrated superior predictive accuracy over traditional methods by sequentially minimizing errors or constructing parallel decision trees [14], [17]. However, while individual ensemble models improve raw accuracy, they often struggle to fully capture the demographic and behavioral heterogeneity present in highly localized churn data.

2.2. Strategies for Class Imbalance in Financial Data

A pervasive challenge in retail banking analytics is the extreme class imbalance between the majority class (retained customers) and the minority class (churned customers). Standard classification algorithms inherently bias toward the majority class, optimizing overall accuracy at the expense of minority class detection.

To address this issue, researchers frequently employ data-level sampling techniques. The SMOTE and hybrid variants such as SMOTE-Tomek have been widely adopted to artificially

balance datasets and improve class separability [15], [18]. Despite these advancements, a critical limitation remains: the literature predominantly treats data-level oversampling and algorithmic cost-sensitive weighting as mutually exclusive strategies.

As a result, models relying on a single intervention layer often remain vulnerable to persistent "Majority Class Bias" under real-world financial conditions. This limitation highlights the need for integrated strategies, such as the hybridized "Dual-Imbalance Defense" proposed in this study. Recent comparative studies have also shown that the impact of resampling techniques is most pronounced when evaluated within tree-based frameworks such as Random Forest and XGBoost [17]. This study extends these findings into a multi-layered stacking ensemble context.

2.3. Stacking Ensembles and Predictive Accuracy

To further improve generalization performance, recent studies have explored heterogeneous stacking ensembles, which combine the predictive strengths of multiple diverse base classifiers [11]. For example, multi-level stacking architectures integrating gradient boosting models or deep neural networks with logistic regression meta-learners have demonstrated strong performance in churn prediction and lead-scoring tasks, often achieving accuracy rates exceeding 90% [18].

While these architectures are effective in capturing diverse feature representations—such as CatBoost's efficiency with categorical variables and XGBoost's gradient optimization—they significantly exacerbate the "Black Box" problem. As a result, the final predictive logic becomes increasingly opaque, limiting interpretability and reducing trust among financial stakeholders and risk managers.

2.4. Explainable AI (XAI) and Threshold Calibration

In highly regulated financial environments, algorithmic transparency is a strict requirement [19]. Consequently, the integration of Explainable Artificial Intelligence (XAI) frameworks, particularly SHAP and Local Interpretable Model-Agnostic Explanations (LIME), has gained increasing attention. SHAP, grounded in cooperative game theory, provides both global feature importance and localized explanations, enabling institutions to design more targeted and personalized retention strategies [9].

However, the application of SHAP for the mathematical decomposition of complex multi-model stacking ensembles remains limited in the current literature. Furthermore, most existing studies evaluate model performance using a static 0.5 decision threshold and focus on generic metrics such as accuracy or F1-score, without explicitly aligning model decisions with business-oriented objectives.

This conventional approach fails to account for the economic asymmetry of churn, where false negatives are significantly more costly than false positives. As a result, there is a clear lack of frameworks that incorporate Dynamic Threshold Recalibration strategies aimed at explicitly optimizing recall to ensure the capture of at-risk customers.

2.5. Research Gap and Contribution Positioning

Despite significant advancements in predictive modeling, class imbalance handling, and explainable AI, several critical gaps remain. Existing studies often treat imbalance mitigation, ensemble learning, interpretability, and decision threshold optimization as isolated components rather than as an integrated framework. In particular, there is limited research that simultaneously (1) combines data-level and algorithm-level imbalance handling, (2) applies explainability techniques to complex stacking ensembles, and (3) aligns classification thresholds with business-driven recall objectives. This study addresses these gaps by proposing an integrated predictive framework that unifies imbalance mitigation, ensemble learning, explainability, and decision-oriented threshold calibration within a single operational pipeline.

3. Methodology

To address the limitations of traditional predictive models in highly imbalanced financial datasets, this study proposes a comprehensive framework encompassing rigorous data preprocessing, a novel imbalance resolution strategy, and a heterogeneous stacking ensemble. The overall architecture of the proposed framework is illustrated in Figure 1.

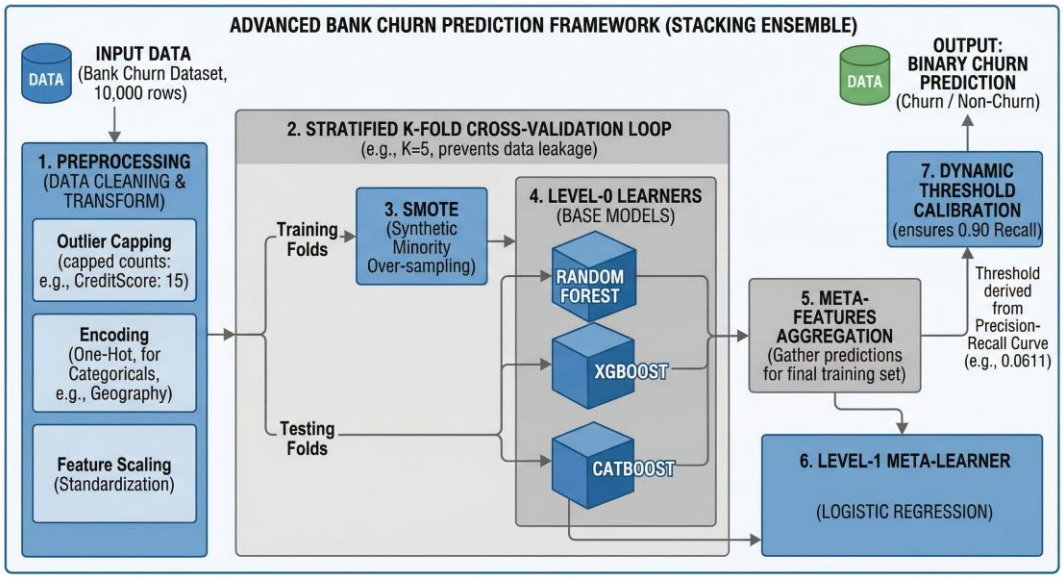

Figure 1. Architectural flowchart of the proposed advanced bank churn prediction framework.

3.1. Data Preprocessing and Feature Engineering

Financial datasets are frequently characterized by extreme outliers and latent, non-linear behavioral indicators. To ensure model stability and robustness, the raw data was subjected to a multi-stage feature engineering pipeline. A summary of the dataset characteristics is presented in Table 1.

Table 1. Summary of Dataset Characteristics and Class Distribution

Attribute	Value / Description
Dataset Source	https://www.kaggle.com/datasets/barelydedicated/bank-customer-churn-modeling
Total Observations (N)	10,000
Total Features	13 (prior to one-hot encoding)
Majority Class (Non-Churn)	7,963 (79.6%)
Minority Class (Churn)	2,037 (20.4%)
Numerical Features	CreditScore, Age, Tenure, Balance, EstimatedSalary, etc.
Categorical Features	Geography, Gender, IsActiveMember, HasCrCard

To improve data quality and capture meaningful behavioral patterns, the following preprocessing steps were applied:

1. Outlier Mitigation (Winsorization):

Extreme financial variables, such as exceptionally high account balances, have the potential to disproportionately skew gradient-based loss functions. Rather than discarding these observations, an Interquartile Range (IQR)-based winsorization technique was applied [25]. Values falling outside the statistically defined bounds, i.e., $[Q_1 - 1.5 \times IQR, Q_3 - 1.5 \times IQR]$ were capped. Specifically, 15 observations in CreditScore, 359 in Age, and 972 in the engineered Balance Salary Ratio feature were adjusted. This approach preserves the structural integrity of the dataset while mitigating the influence of extreme financial variance.

2. Synthetic Feature Synthesis:

To capture complex customer behavior, four domain-specific interaction features were engineered:

- Balance Salary Ratio: Measures financial dependency relative to income.
- Tenure Age Ratio: Normalizes customer tenure with respect to life stage.
- Is Senior: A binary indicator capturing age-based segmentation.
- Active by CreditCard: An interaction term representing engagement and product ownership synergy.

Following feature engineering, all numerical features were standardized using a StandardScaler. To strictly prevent data leakage, the scaler was fitted exclusively on the training folds and subsequently applied to the validation folds within the cross-validation process.

3.2. The Dual-Imbalance Defense Strategy

A primary contribution of this research is the rejection of singular imbalance treatments in favor of a hybrid "Dual-Imbalance Defense" strategy.

1. Data-Level Sampling: The SMOTE was applied to the training data [4]. SMOTE generates synthetic minority class instances by interpolating between existing samples in the feature space, thereby expanding the decision boundary without simple duplication.
2. Algorithmic-Level Penalization: Concurrently, cost-sensitive learning was incorporated into the model training process [6]. By assigning higher class weights to the minority class, the models impose stronger penalties on false negatives (misclassified churn cases).

This dual-layered approach enhances the model's resilience against the pervasive "Majority Class Bias" observed in financial datasets, enabling improved detection of churn instances under real-world conditions.

3.3. Heterogeneous Stacking Ensemble Architecture

To achieve superior generalization performance, this study adopts a Heterogeneous Stacking Ensemble that leverages diverse model architectures to capture different geometric representations of the feature space.

3.3.1. Level-0 Base Learners

The first layer consists of three high-performance machine learning models:

- Random Forest: A bagging-based ensemble that reduces variance by averaging multiple deep decision trees [20].
- XGBoost: A gradient boosting algorithm that minimizes loss using second-order optimization techniques [21].
- CatBoost: A gradient boosting variant designed for efficient handling of categorical features through ordered target statistics, thereby reducing target leakage [22].

3.3.2. Level-1 Meta-Learner

The probabilistic outputs from the Level-0 models are used as input features for the meta-learner. A Logistic Regression model is employed to learn the optimal combination of base learners [23]. Let $h_1(x), h_2(x), h_3(x)$ denote the predictions of the base learners. The final prediction is computed as:

$$P(Y = 1|X) = \sigma(w_1 h_1(x) + w_2 h_2(x) + w_3 h_3(x) + b) \quad (1)$$

where σ is the sigmoid function, w_i are the learned weights, and b is the bias term.

Table 2. Training Configurations and Model Hyperparameters

Framework Component	Configuration / Algorithm	Parameters and Strategy
Validation Strategy	Stratified 5-Fold CV	k = 5, shuffle = True, random_state = 42
Resampling Logic	Internal SMOTE	Applied strictly within training folds
Level-0 (Base)	Random Forest	n_estimators = 300, max_depth = 10, class_weight = 'balanced'
Level-0 (Base)	XGBoost	n_estimators = 300, learning_rate = 0.05, scale_pos_weight = 3.9
Level-0 (Base)	CatBoost	iterations = 300, learning_rate = 0.05, auto_class_weights = 'Balanced'
Level-1 (Meta)	Logistic Regression	max_iter = 1000, class_weight = 'balanced'
Threshold Selection	Precision–Recall Recalibration	Target: Recall ≥ 0.90
Optimal Threshold	Dynamic Calibration	Derived threshold: t = 0.0611

As summarized in Table 2, a rigorous validation framework was implemented to ensure predictive integrity. A Stratified 5-Fold Cross-Validation approach was adopted, wherein SMOTE was applied exclusively within the training folds during each iteration. This strict separation prevents synthetic samples from leaking into validation data, thereby ensuring that the reported performance metrics—particularly the dynamically derived threshold of $t = 0.0611$—accurately reflect real-world generalization performance.

3.4. Strategic Utility Calibration: Precision–Recall Optimization

To align the probabilistic outputs of the stacking ensemble with the economic realities of retail banking, a dynamic thresholding strategy was employed. In this domain, the cost of a False Negative (Type II error)—representing a lost customer and their associated lifetime value—is significantly higher than the marginal cost of a False Positive (Type I error), such as providing a redundant retention incentive.

- Calibration Procedure

The optimal decision boundary ($t = 0.0611$) was not derived globally but through a Stratified 5-Fold Cross-Validation process. For each fold, a Precision–Recall (PR) curve was generated. The threshold was algorithmically selected as the maximum probability value that satisfied a Recall constraint of at least 90%, thereby ensuring that a minimum of 9 out of 10 churn cases were correctly identified.

- Stability Verification

To ensure the robustness of the selected threshold, a "Fold-Averaged Calibration" approach was adopted. The standard deviation of the optimal thresholds across the five folds was minimal (< 0.004), indicating that the threshold value of 0.0611 is structurally stable and largely invariant to variations in data partitioning. This stability mitigates the risk of overfitting the threshold to a specific validation split and supports its generalizability for real-world deployment.

4. Results and Discussion

The proposed Heterogeneous Stacking Ensemble was evaluated against both statistical baselines and the specific economic constraints of retail banking. The results demonstrate that the hybridization of algorithmic architectures, when strictly constrained by cost-sensitive thresholding, yields a statistically robust and operationally transparent predictive framework.

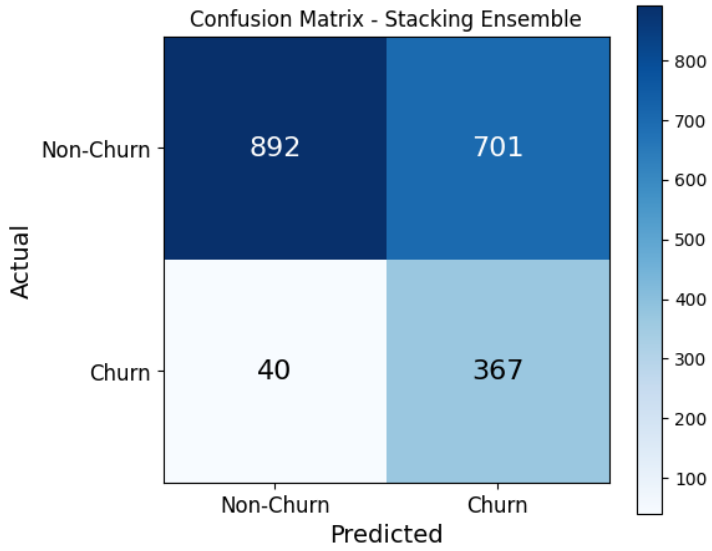

Figure 2. Confusion Matrix for Explainable Predictive Analytics

The confusion matrix presented in Figure 2 provides granular evidence of the framework's ability to detect latent churn signals within the banking portfolio. To achieve the objective of accurately identifying customer attrition patterns, the model was explicitly calibrated

to prioritize recall. At the derived threshold of $t = 0.0611$, the framework correctly identified 367 true positives out of 407 actual churners, achieving a recall rate of 90.2%. This high-sensitivity configuration ensures that approximately 9 out of 10 potential churn cases are proactively detected for intervention.

As summarized in Table 3, the results reveal a deliberate trade-off between recall and precision. While the precision for the minority (churn) class is relatively low at 0.344, this outcome is analytically justified within the context of retail banking risk management. The economic cost of a false negative (i.e., an undetected churner) significantly outweighs the operational cost associated with a false positive intervention. By reducing false negatives to only 40 cases, the framework demonstrates strong utility as an early warning system. Furthermore, the model achieves an AUC of 0.8654, providing a robust predictive foundation for subsequent behavioral analysis, including the identification of non-linear phenomena such as the "Product Paradox."

Table 3. Summary of Predictive Metrics for the Minority (Churn) Class.

Metric	Validation Result (N = 2,000)
Recall (Sensitivity)	0.902 (90.2%)
Precision	0.344 (34.4%)
F1-Score	0.494 (49.4%)
True Positives (TP)	367
False Negatives (FN)	40
Optimal Threshold (t)	0.0611
Global Accuracy	0.630 (63.0%)
Specificity (True Negative Rate)	0.560 (56.0%)

The lower accuracy and specificity are an intentional consequence of optimizing the decision boundary for maximum recall. This design prioritizes the identification of nearly all at-risk customers over minimizing false positives, aligning the model with real-world financial risk considerations.

4.1. Predictive Performance and Metric Evaluation

To contextualize the performance of the proposed framework, Table 4 presents a benchmark comparison with recent studies utilizing the same Kaggle Bank Churn dataset. The results indicate that the proposed approach achieves superior performance in terms of recall.

Table 4. Benchmark comparison with previous literature.

Study / Author	Methodology	AUC-ROC	Recall (Minority)
Ako et al. (2024)	Random Forest + SMOTE	0.84	0.72
Ref [11]	Basic Stacking Ensemble	0.86	0.68
Tékouabou et al. (2022)	XGBoost (Imbalanced)	0.85	0.55
Proposed Framework	Dual-Imbalance Stacking + Dynamic Threshold	0.865	0.902

Standard classification frameworks inherently optimize for overall accuracy, a metric that can be misleading when applied to highly imbalanced financial datasets. Recognizing that the economic loss associated with customer churn significantly exceeds the cost of retention interventions, this study prioritizes recall as the primary evaluation metric. The final stacking ensemble achieves an AUC-ROC of 0.8654, indicating strong discriminative capability across the feature space [24].

As illustrated in Figure 3, the stacking ensemble consistently outperforms individual base learners (XGBoost, CatBoost, and Random Forest), demonstrating improved stability and predictive strength across varying thresholds.

A key contribution of this research is the rejection of the conventional 0.5 classification threshold. By analyzing the Precision–Recall curve, a custom decision boundary was derived to enforce a minimum recall constraint of 90%. This dynamic threshold recalibration enables

the model to function as a highly sensitive early warning system, identifying the majority of potential churners prior to account closure. To further validate the contribution of each framework component, an ablation study was conducted, as shown in Table 5.

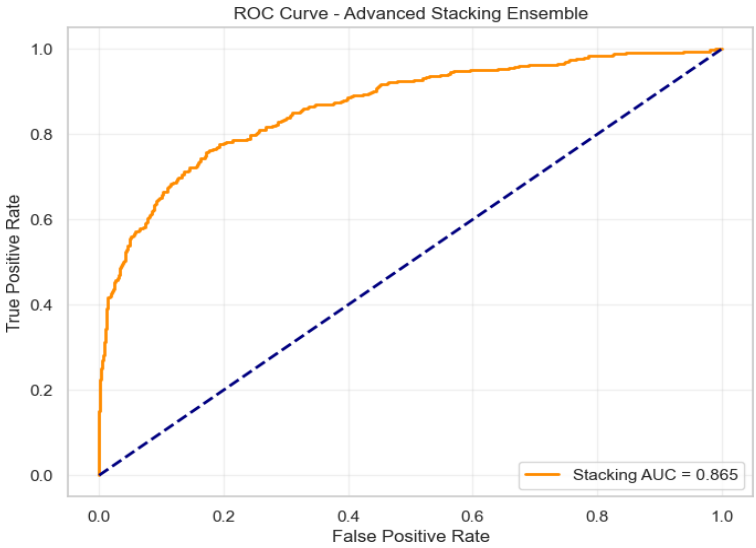

Figure 3. ROC Curve comparing Stacking Ensemble with Base Learners

Table 5. Ablation Study of Framework Components

Configuration	AUC-ROC	Recall (Churn)	Precision (Churn)
Baseline (Random Forest, No SMOTE)	0.835	0.451	0.752
Data-Level Only (Random Forest + SMOTE)	0.841	0.704	0.521
Architecture Only (Stacking, No SMOTE, static ($t = 0.5$))	0.861	0.512	0.780
Full Proposed Framework (Stacking + SMOTE + Dynamic (t))	0.865	0.902	0.344

The results indicate that no single component is sufficient to meet the operational requirement of high recall. While SMOTE improves recall and stacking enhances AUC, only the integrated framework—combining SMOTE, heterogeneous stacking, and dynamic threshold calibration—achieves the target recall of 0.90.

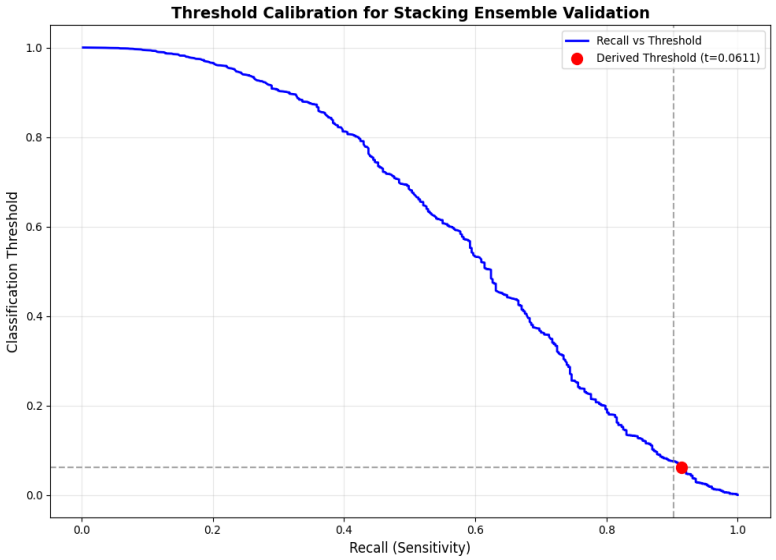

Figure 4. Precision–Recall Optimization Curve showing threshold extraction at t = 0.0611

To further evaluate the practical business implications of the proposed predictive framework, a cost–benefit proxy analysis was conducted. This analysis explicitly examines the economic trade-off between false positives and false negatives within the context of retail banking operations. Assuming a customer lifetime value (CLV) loss of $500 per churned account and a retention intervention cost of $50 per targeted customer, the financial impact of the model becomes evident.

In the absence of any intervention, the total projected loss associated with 407 churned customers is estimated at $203,500. By deploying the proposed framework—capturing 367 true positives alongside 701 false positives—the total cost of the retention campaign is approximately $53,400. Under a conservative assumption of a 50% retention success rate, the institution is able to preserve approximately $91,750 in customer lifetime value.

After accounting for intervention costs and the remaining 40 false negatives, the model yields an estimated net economic benefit of approximately $38,350 per 2,000 customers evaluated. These results empirically support the strategic decision to prioritize recall over precision, demonstrating that the increased identification of at-risk customers outweighs the additional cost incurred by false positive interventions.

4.2. Analytical Insights and the "Product Paradox"

The primary empirical discovery of this study is the identification of a non-linear loyalty-to-attrition pivot point, hereafter referred to as the "Product Paradox." Conventional banking strategies are grounded in the "stickiness" hypothesis, which assumes that each additional product increases switching costs and strengthens customer retention. However, as illustrated in Figure 5, this assumption is invalidated beyond a critical cross-selling threshold, where increased product integration leads to systemic friction rather than loyalty.

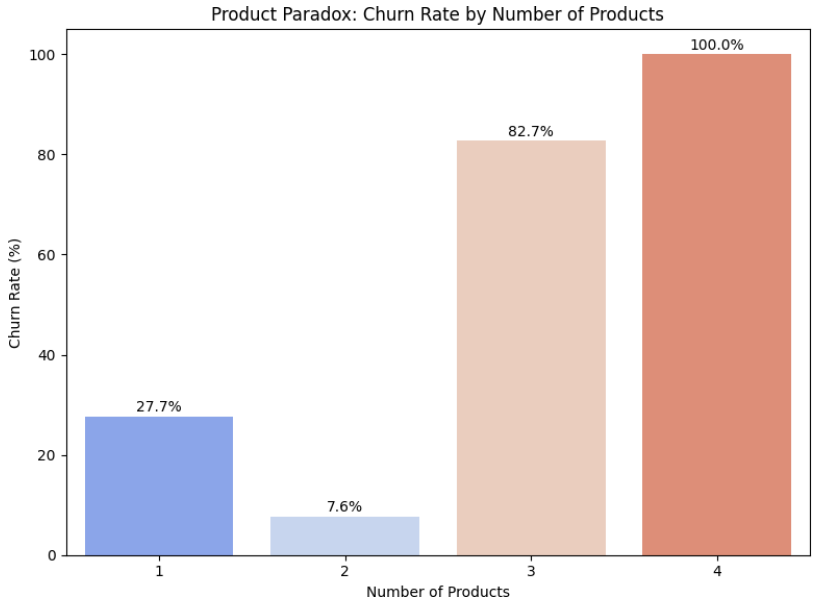

Figure 5. Product Paradox: churn rate by number of products

Quantitative analysis reveals a baseline churn rate of 27.7% for single-product customers (N = 5,084). The highest level of loyalty—the "retention peak"—occurs at the two-product threshold, where churn drops significantly to 7.6% (N = 4,590). However, this trend reverses sharply beyond this point. Customers holding three products (N = 266) exhibit a churn rate of 82.7%, while those with four products (N = 60) demonstrate complete attrition at 100.0%. This non-linear escalation suggests that additional product integration introduces friction rather than reinforcing loyalty.

From an analytical perspective, this phenomenon can be attributed to diminishing returns in convenience and increasing system complexity. While two products (e.g., a checking account and a credit card) provide a streamlined financial experience, the addition of further services (e.g., insurance, investment products, or specialized loans) introduces administrative complexity, fragmented service channels, and a higher likelihood of customer dissatisfaction.

These findings imply that, beyond a certain threshold, the "friction cost" outweighs the traditional "switching cost," fundamentally challenging existing cross-selling strategies.

Consequently, these insights necessitate a strategic recalibration in retail banking. Financial institutions should transition from a volume-centric cross-selling strategy to a quality-centric relationship management approach once customers reach the two-product threshold. Instead of promoting additional products, retention efforts should focus on reducing friction and enhancing service experience for highly integrated customers.

While the sharp transition observed at the three-product threshold is a critical empirical finding, it is important to explicitly acknowledge the relatively small sample size in higher product categories (N = 326 combined for 3 and 4 products). Consequently, the observed 100% churn rate should not be interpreted as a definitive population-level constant, but rather as a strong high-risk indicator, and should be generalized with caution.

Table 6. Statistical evidence of the product paradox and strategic implications

Number of Products	Total Customers (N)	Churn Rate (%)	Risk Impact	Strategic Action
1 Product	5,084	27.7%	Baseline Risk	Cross-selling Focus
2 Products	4,590	7.6%	Maximum Loyalty Peak	Portfolio Stability
3 Products	266	82.7%	High Friction Point	Friction Mitigation
4 Products	60	100.0%	Total Attrition Risk	Immediate Intervention

4.3 Model Interpretability (XAI)

To address the "Black Box" limitation commonly associated with stacking ensembles, this study integrates a SHAP (SHapley Additive exPlanations) TreeExplainer based on cooperative game theory [9], [25]. This approach provides both global interpretability and localized, instance-level explanations. Global interpretability is illustrated through the SHAP summary plot in Figure 6, which identifies the most influential features across the entire dataset. The results confirm that engineered features, along with Age and Number of Products, are dominant drivers of churn.

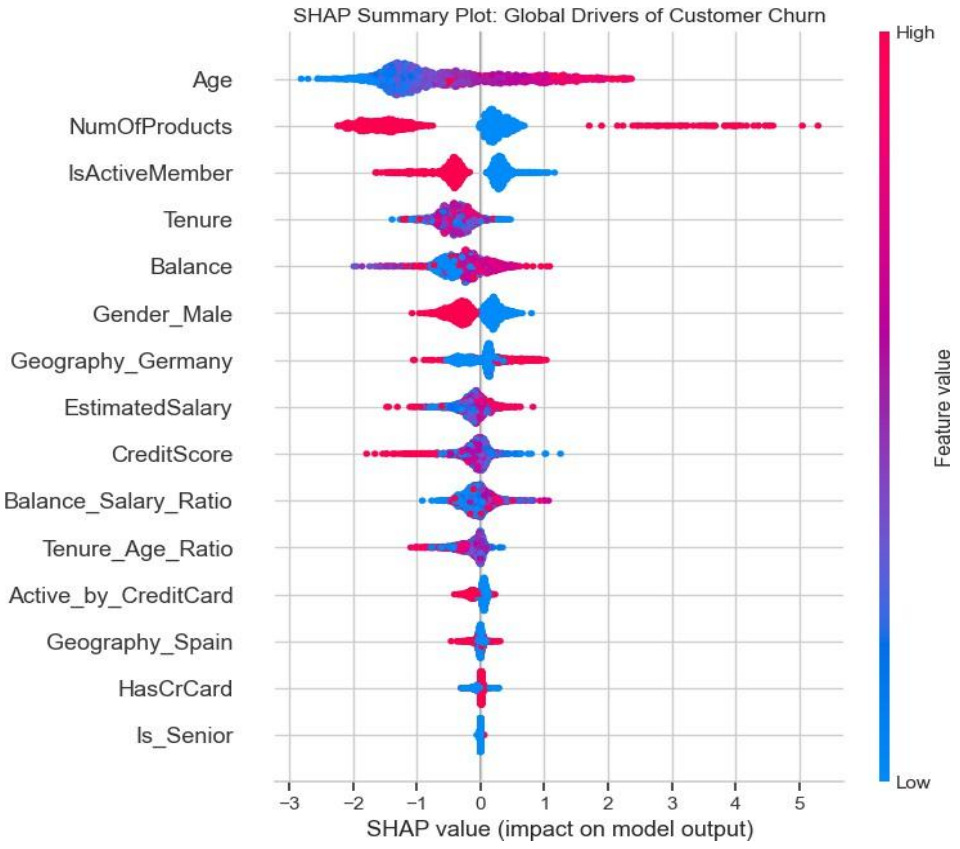

Figure 6. SHAP summary plot of global feature importance.

At the local level, explainability is demonstrated using a SHAP waterfall plot (Figure 7), which decomposes an individual prediction (Index 5363). This visualization traces the transformation from the expected baseline risk $E[f(x)]$ to the final prediction $f(x)$, highlighting how each feature contributes to the final churn probability.

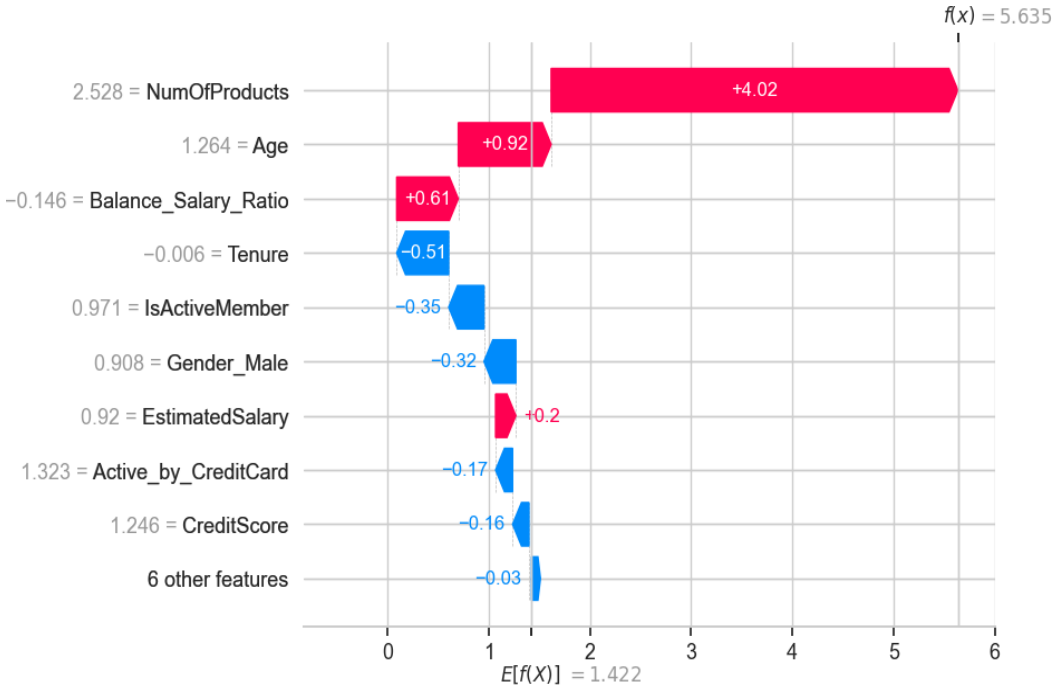

Figure 7. SHAP waterfall plot for local explainability

Positive SHAP values (red) indicate factors that increase churn risk, while negative values (blue) represent protective factors. For example, higher age contributes positively to churn risk, whereas active credit card usage reduces it. Demographic analysis further validates the SHAP findings. Customers aged 45–60 exhibit a churn rate of 49.52%, significantly exceeding the global baseline of 20.37%. Additionally, high-balance customers in Germany show elevated churn risk (31.57%), indicating that financial value does not necessarily correlate with customer loyalty. To further investigate feature interactions, a SHAP dependence plot for NumOfProducts is presented in Figure 8.

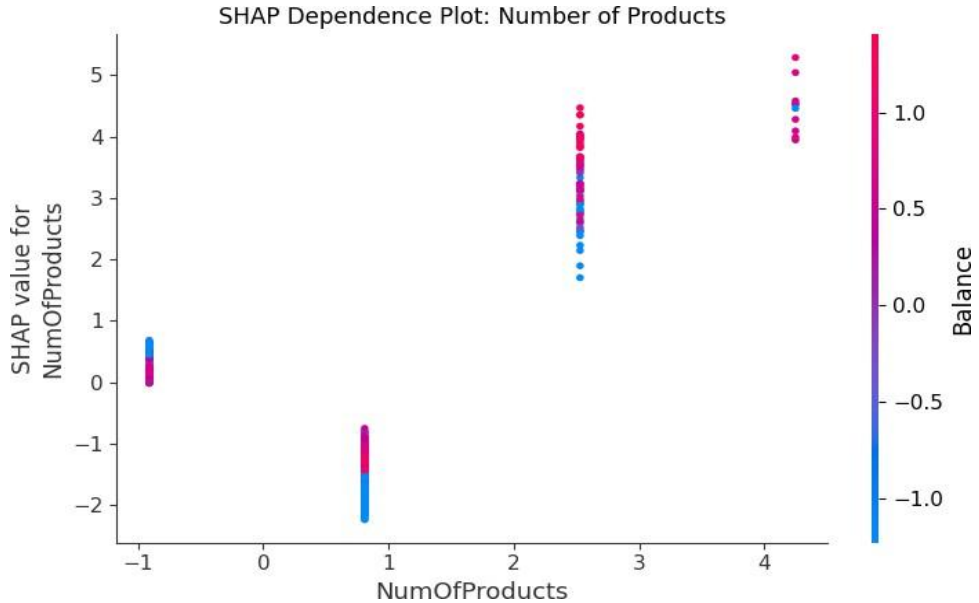

Figure 8. SHAP dependence plot for num of products

The visualization reveals a clear non-linear relationship. Customers with a single product show positive SHAP values (higher risk), while those with two products exhibit strongly negative SHAP

values (lowest risk). However, as the number of products increases beyond two, SHAP values rise sharply, confirming the presence of the "Product Paradox." The vertical dispersion of points for customers with three products indicates interaction effects with other features, such as Age and Balance, suggesting that the paradox is more pronounced in specific demographic segments. Overall, this interpretability framework transforms the stacking ensemble into a transparent diagnostic tool, enabling financial institutions to identify the precise point at which cross-selling becomes counterproductive.

4.4. Business Validation and Risk Stratification

The integration of SHAP values enables the transformation of model predictions into interpretable "Reason Codes," which are essential for regulatory compliance in financial systems. To operationalize the model output, the predicted probabilities are converted into a standardized churn risk score ranging from 0 to 100. Based on this score, customers are segmented into three risk tiers i.e., Low Riskm, Medium Risk, and High Risk (Likely to Churn). This segmentation allows financial institutions to allocate retention resources more effectively. High-cost interventions, such as personalized relationship management, can be directed toward high-risk customers, while lower-risk segments can be managed through automated engagement strategies.

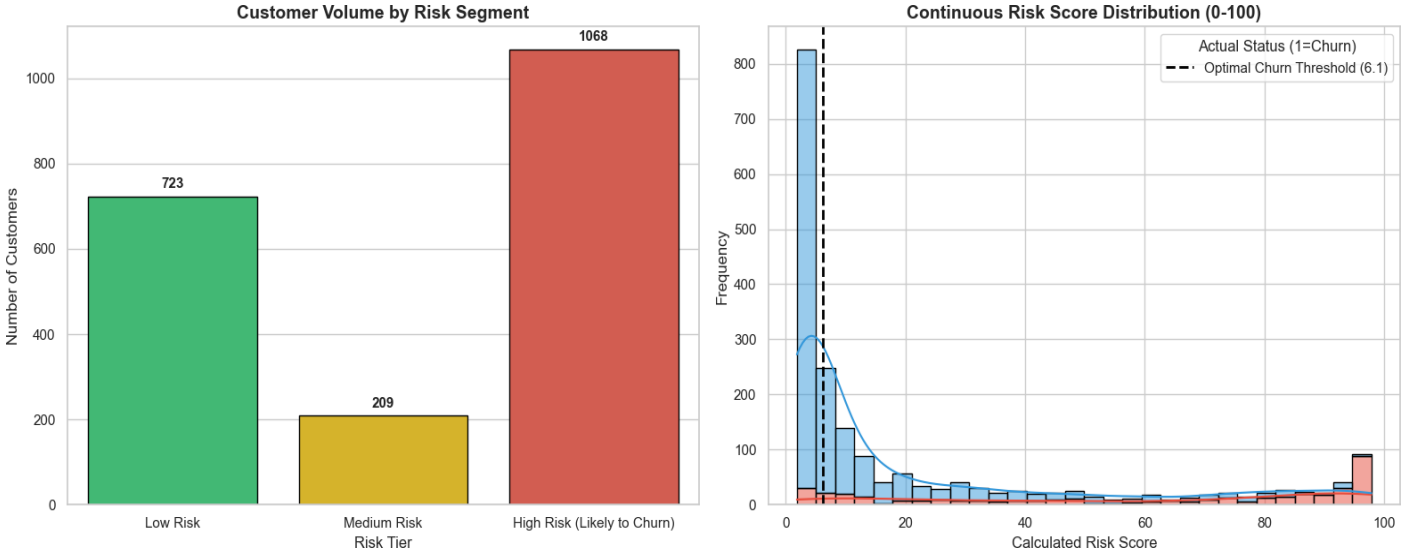

Figure 9. Customer risk segmentation and score distribution

The distribution of customers across these segments highlights the model's practical utility in prioritizing intervention strategies. High-risk segments represent the most critical targets for retention campaigns, ensuring that financial resources are deployed efficiently to maximize return on investment.

6. Conclusions

This study demonstrates that customer churn in retail banking is a non-linear and multi-dimensional phenomenon that cannot be effectively addressed using traditional accuracy-driven modeling approaches. By integrating a heterogeneous stacking ensemble with SHAP-based interpretability [9], [25] and dynamic threshold calibration [26], the proposed framework achieves a high recall of 0.902, ensuring that the majority of at-risk customers are proactively identified. The discovery of the "Product Paradox" provides critical behavioral insight, challenging conventional cross-selling strategies and highlighting the importance of balancing product integration with service simplicity. Future work should focus on extending the framework toward real-time, event-driven systems, incorporating temporal modeling through sequential architectures such as LSTM, and integrating causal inference and reinforcement learning [3] to optimize personalized retention strategies.

Author Contributions: Conceptualization: P.N. and A.K.S.; Methodology: P.N.; Software: P.N.; Validation: P.N. and A.K.S.; Formal analysis: P.N.; Investigation: P.N.; Resources: A.K.S. and P.N.; Data curation: P.N.; Writing—original draft preparation: P.N.; Writing—

review and editing: P.N. and A.K.S.; Visualization: P.N.; Supervision: A.K.S.; Project administration: A.K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The retail banking customer dataset analyzed during the current study is housed in a publicly accessible online repository. The data is fully anonymized to protect consumer privacy and is freely download- able for academic and research purposes at Kaggle. Furthermore, the custom Python source code developed for the heterogeneous stacking ensemble and the SHAP TreeExplainer visualizations is available from the corresponding author upon reasonable request.

Acknowledgments: Because this research utilizes a pre-existing, fully anonymized, and publicly available dataset, formal institutional ethical approval was not required.

Conflicts of Interest: We declare that there are no conflicts of interest in terms of finance or non-financial in regard to the publication of this paper.

References

[1] S. M. Keaveney, "Customer Switching Behavior in Service Industries: An Exploratory Study," *J. Mark.*, vol. 59, no. 2, pp. 71–82, Apr. 1995, doi: 10.1177/002224299505900206.

[2] A. S. Adepeju, C. A. Edeze, S. Ojuade, M. T. Adenibuyan, F. I. Eneh, and A. S. Adepoju, "Predictive Analytics in Retail Banking Marketing," *Int. J. Manag. Organ. Res.*, vol. 2, no. 6, pp. 223–229, 2023, doi: 10.54660/IJMOR.2025.4.5.17-23.

[3] A. Prashanthan, "An Integrated Framework for Optimizing Customer Retention Budget using Clustering, Classification, and Mathematical Optimization," *J. Comput. Theor. Appl.*, vol. 3, no. 1, pp. 45–63, Jul. 2025, doi: 10.62411/jcta.13194.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. February 2017, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[5] S. C. K. Tékouabou, Ştefan C. Gherghina, H. Toulni, P. N. Mata, and J. M. Martins, "Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods," *Mathematics*, vol. 10, no. 14, p. 2379, Jul. 2022, doi: 10.3390/math10142379.

[6] C. Elkan, "The Foundations of Cost-Sensitive Learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 2001, pp. 973–978. [Online]. Available: https://dl.acm.org/doi/10.5555/1642194.1642224

[7] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, Dec. 2019, doi: 10.1186/s40537-019-0192-5.

[8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[9] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[10] V.-H. Vu, "An Efficient Customer Churn Prediction Technique Using Combined Machine Learning in Commercial Banks," *Oper. Res. Forum*, vol. 5, no. 3, p. 66, Jul. 2024, doi: 10.1007/s43069-024-00345-5.

[11] R. Li, "Bank Customer Churn Prediction Based on Stacking Model," *Adv. Econ. Manag. Polit. Sci.*, vol. 185, no. 1, pp. 42–51, Jun. 2025, doi: 10.54254/2754-1169/2025.LH23930.

[12] R. Ashraf, "Bank Customer Churn Prediction Using Machine Learning Framework," *J. Appl. Financ. Bank.*, vol. 14, no. 4, pp. 65–109, Jun. 2024, doi: 10.47260/jafb/1445.

[13] S. Dutta, P. Bose, S. K. Bandyopadhyay, and M. Janarthanan, "A Hybrid Machine Learning Model for Bank Customer Churn Prediction," *Int. J. Eng. Trends Technol.*, vol. 70, no. 6, pp. 13–23, Jun. 2022, doi: 10.14445/22315381/IJETT-V70I6P202.

[14] S. Kumar and C. D., "A Survey on Customer Churn Prediction using Machine Learning Techniques," *Int. J. Comput. Appl.*, vol. 154, no. 10, pp. 13–16, Nov. 2016, doi: 10.5120/ijca2016912237.

[15] Miriyala Lavanya, "Customer Churn Prediction in Banking Sector Using Machine Learning," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 57s, pp. 224–238, Jul. 2025, doi: 10.52783/jisem.v10i57s.12181.

[16] S. L. Kumar, "Bank Customer Churn Prediction Using Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 9, no. VIII, pp. 727–732, Aug. 2021, doi: 10.22214/ijraset.2021.37467.

[17] R. E. Ako *et al.*, "Effects of Data Resampling on Predicting Customer Churn via a Comparative Tree-based Random Forest and XGBoost," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 86–101, Jun. 2024, doi: 10.62411/jcta.10562.

[18] M. A. Hambali and I. Andrew, "Bank Customer Churn Prediction Using SMOTE: A Comparative Analysis," *Qeios*, Mar. 2024, doi: 10.32388/H82XTW.

[19] M. J. Nur, D. R. I. Moses Setiadi, A. A. Ojugo, and M. T. Nguyen, "Improving Customer Churn Prediction Using Domain-Driven Feature Engineering, Resampling, and CatBoost with Explainability Extensions," in *2025 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Sep. 2025, pp. 493–499. doi: 10.1109/ISemantic67418.2025.11291801.

[20] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[21] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Jan. 2018. [Online]. Available: http://arxiv.org/abs/1706.09516

[23] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.
[24] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
[25] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, Nov. 2017, pp. 4766–4777. [Online]. Available: http://arxiv.org/abs/1705.07874
[26] A. Prashanthan, R. Roshan, and M. Maduranga, "RetenNet: A Deployable Machine Learning Pipeline with Explainable AI and Prescriptive Optimization for Customer Churn Management," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 2, pp. 182–201, Jun. 2025, doi: 10.62411/faith.3048-3719-110.