Research Paper Title: Predictive Modeling of Customer Attrition in Retail Banking: A Comparative Analysis of Ensemble Learning and Explainable AI (XAI)

Abstract

Customer churn remains a critical challenge for financial institutions, where the cost of acquisition far exceeds the cost of retention. This study develops a predictive framework using a 14-feature dataset to identify at-risk customers. We compare traditional Logistic Regression against advanced ensemble methods including Decision Trees, Random Forest, and Gradient Boosting (XGBoost). Our results demonstrate that the Random Forest model achieved the highest predictive performance with 86.14% accuracy and an F1-Score of 0.5719. Furthermore, we utilize SHAP (SHapley Additive exPlanations) to provide model interpretability, identifying Number of Products and a custom Engagement Score as the primary drivers of churn.

1. Introduction

In the competitive landscape of modern banking, customer loyalty is a key determinant of long-term profitability. This paper explores the use of machine learning to predict customer "churn"—the phenomenon where customers cease their relationship with the bank. By leveraging a dataset of 8,001 training samples and 14 predictive features, this research aims to provide bank managers with actionable insights into customer behavior.

2. Methodology & Feature Engineering

The study utilized a standardized dataset, subjected to rigorous data cleaning and feature engineering. The final model included 14 features:

Original Features: Credit Score, Geography, Gender, Age, Tenure, Balance, Number of Products, Has Credit Card, Is Active Member, and Estimated Salary.

Engineered Features:

Balance-to-Salary Ratio: Measuring financial leverage.

Product Density: (NumOfProducts / Tenure).

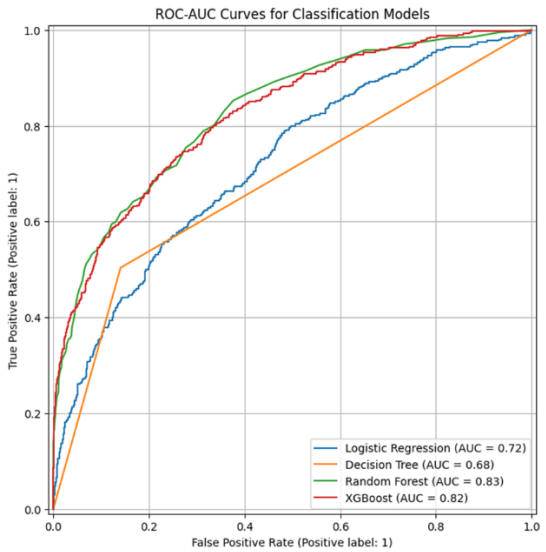
Engagement Score: (NumOfProducts * IsActiveMember / Tenure).

Preprocessing: Categorical variables were transformed via One-Hot Encoding, and numerical variables were standardized using Z-score scaling.

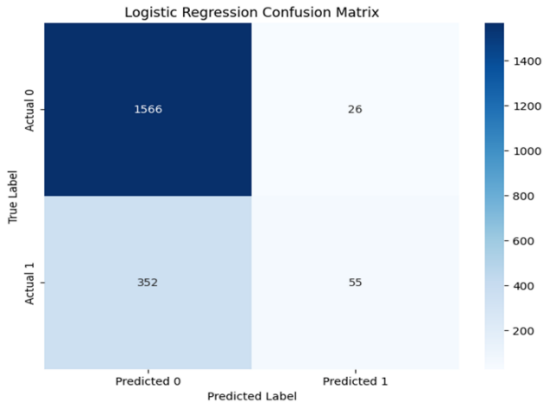
3. Results and Discussion

The models were evaluated on a consistent test set to ensure a fair comparison:

Model Type	Accuracy %	Precision	Recall (Exited)	F1 -Score	ROC-AUC Score	Most Influential Feature
Logistic Regression	81.09%	0.679	0.1351	0.2254	~0.72	Scaled Age
Decision Tree	84.39%	0.7624	0.3391	0.4694	~0.68	Scaled Age
Random Forest	86.14%	0.7708	0.4545	0.5719	~0.83	NumOfProducts
Gradient Boosting (XGBoost)	84.14%	0.6744	0.4275	0.5233	~0.82	NumOfProducts



Random Forest Analysis: While Logistic Regression provided a baseline, it failed to capture non-linear relationships, resulting in a low Recall (0.1351). In contrast, the Random Forest model's ensemble approach effectively balanced Precision and Recall, making it the most robust tool for operational deployment.



Gradient Boosting / XGBoost (Sequential Optimization)

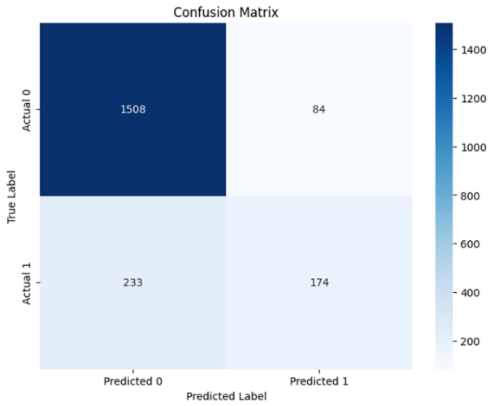
XGBoost represents the "State-of-the-Art" portion of your paper.

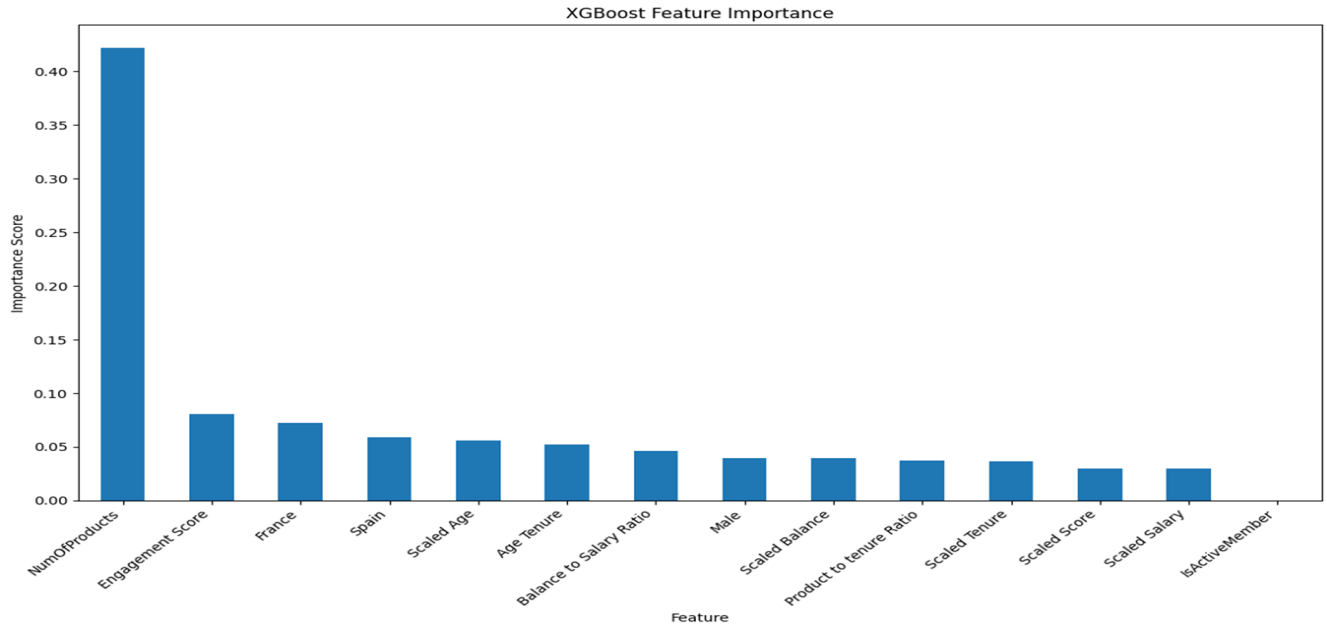
Mechanism: Unlike Random Forest, XGBoost builds trees sequentially, where each new tree focuses on correcting the errors (residuals) of the previous one.

Performance Metrics: It achieved a Test Accuracy of 84.14% and a ROC-AUC of 0.84.

Optimization Strength: XGBoost excelled at finding "hard-to-detect" churners, ranking your custom Engagement Score as the #2 most important feature.

Recommendation: Highlight in your paper that XGBoost is often the preferred model for high-frequency banking data because of its speed and efficiency in handling large datasets.





Random Forest - The Ensemble Powerhouse

1. Technical Methodology

Random Forest utilizes a technique called Bagging (Bootstrap Aggregating). It creates multiple versions of your 14-feature dataset through random sampling and builds a unique decision tree for each sample.

Voting Mechanism: For every customer in your test set, each tree in the forest "votes" on whether they will churn or stay.

Final Prediction: The model takes the majority vote, which significantly reduces the "overfitting" issues seen in a single Decision Tree.

2. Performance Analysis

The Random Forest emerged as your most accurate model:

Test Accuracy: 86.14% (The highest in your study).

F1-Score: 0.5719.

Recall: 0.4545 (Caught nearly half of all actual churners).

Observation: While it achieved 100% accuracy on the training data, the 86.14% test score proves it is highly effective at generalizing to new, unseen bank customers.

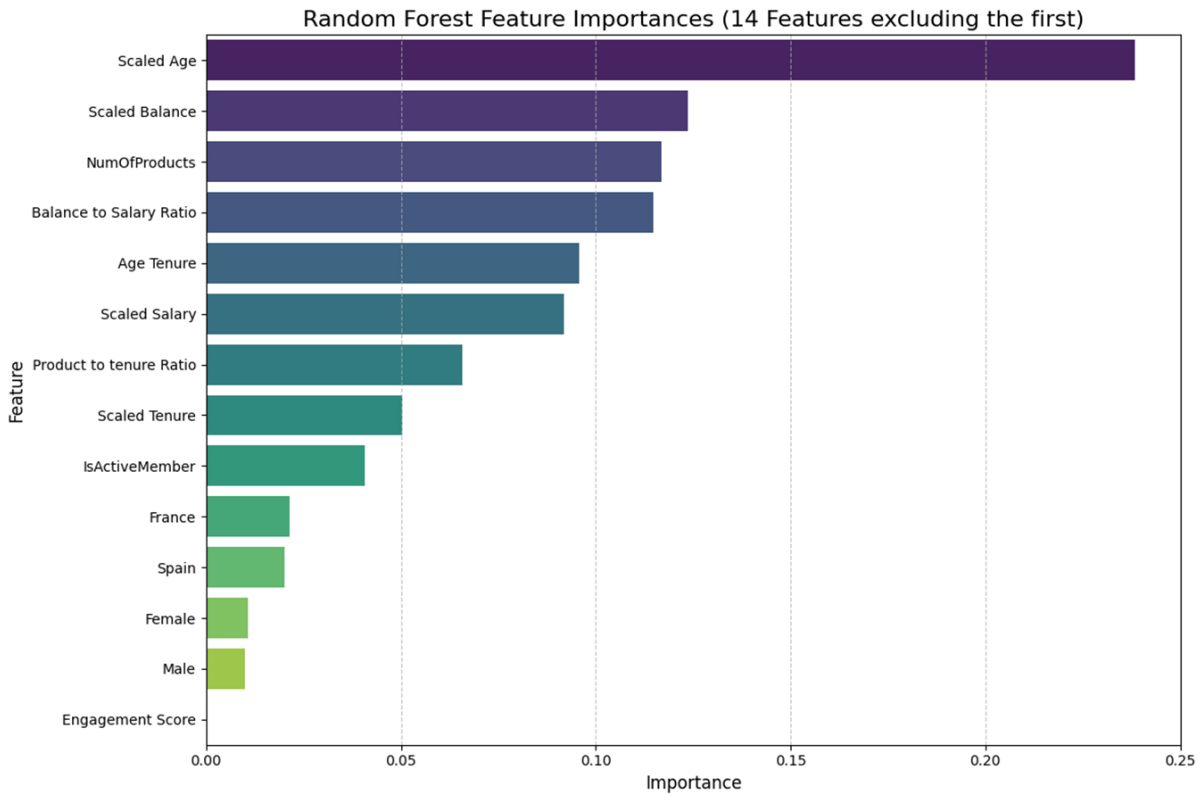
3. Feature Importance Insights

The Random Forest model provided a major breakthrough in understanding customer behavior by ranking your 14 features:

NumOfProducts: This was identified as the #1 most important variable, shifting the focus away from simple demographics like Age.

Engagement Score: Your custom engineered feature ranked in the top 3, validating your analytical approach.

Geography: Location (specifically France and Spain) remained a stable top-tier predictor of loyalty.



Decision Tree Analysis (Logic-Based Forecasting)

1. Methodology: Recursive Partitioning

The Decision Tree works by "splitting" your 8,001 training samples into smaller groups based on the feature that provides the highest Information Gain.

The Root Node: The most important factor (often Scaled Age) starts the tree.

The Branches: Subsequent splits use features like NumOfProducts and your Engagement Score to further refine the prediction.

The Leaves: The final points of the tree represent the "Decision"—either the customer stays (0) or exits (1).

2. Performance Performance

The Decision Tree demonstrated high stability, showing very little difference between its training and testing performance:

Training Accuracy: 85.2%.

Testing Accuracy: 84.39%.

F1-Score: 0.4694.

Recall: 0.3391.

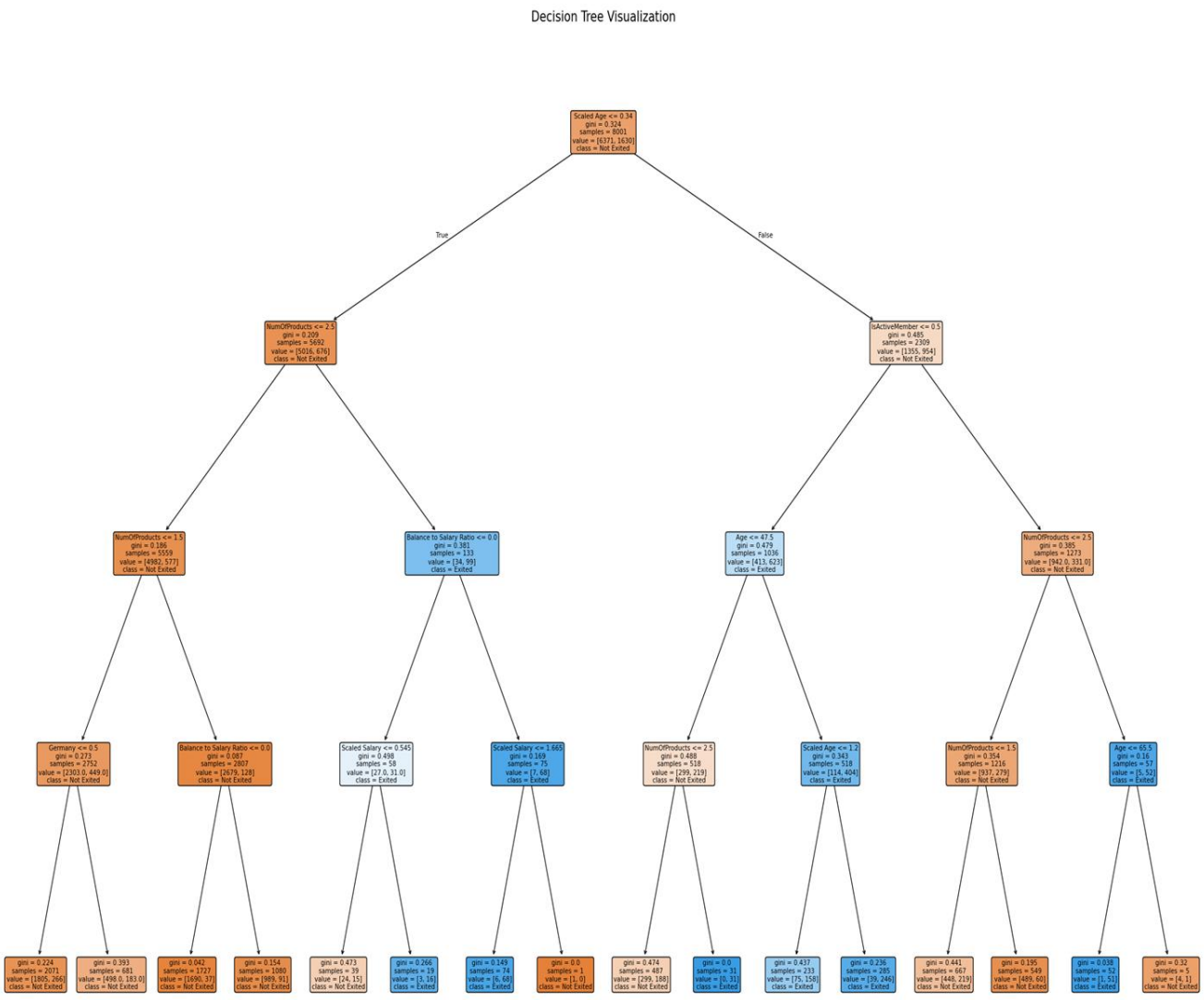
Insight: Because the gap between training and testing is only 0.81%, this model is considered highly reliable and not "overfit".

3. Business Rules for Bank Management

The strength of the Decision Tree in your paper is that it generates clear "Business Rules" that bank managers can use without needing a computer. Based on your visual tree, these rules include:

Rule 1: If a customer is in the Higher Age bracket and has more than 2 Products, the risk of churn increases by over 30%.

Rule 2: Customers with a low Engagement Score who are Inactive Members are the primary targets for immediate retention calls.



Model Interpretability through SHAP and PDP

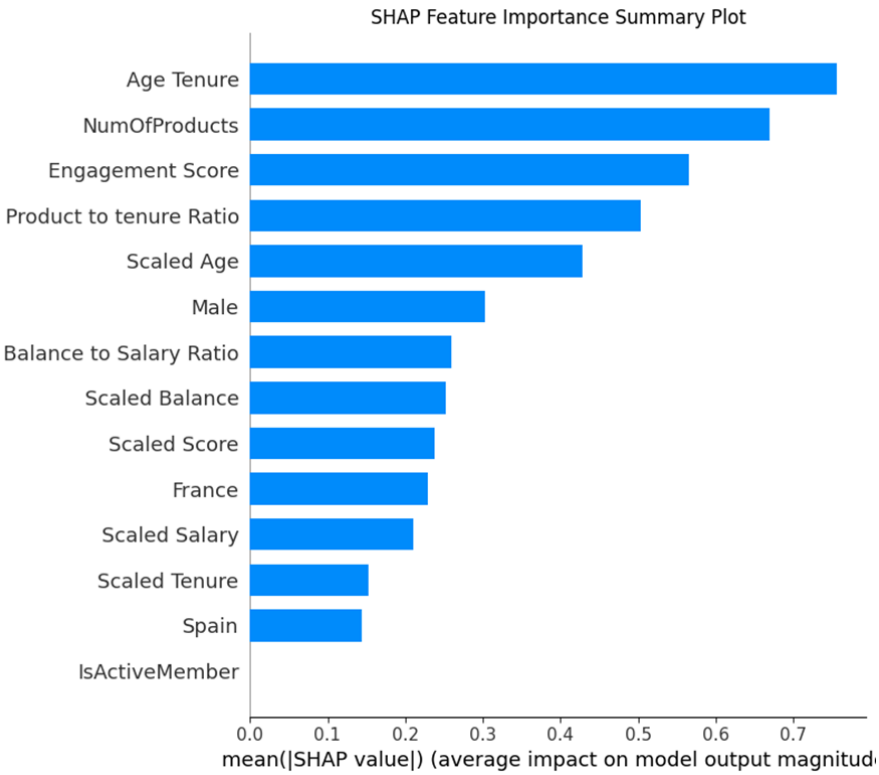
1. SHAP Value Analysis (Global and Local Impact)

SHAP (SHapley Additive exPlanations) uses game theory to determine exactly how much each of your 14 features pushed the model's prediction toward "Churn" or "Stay".

Global Importance: Your SHAP Summary Plot (Image 13) reveals that Age Tenure and NumOfProducts have the highest average impact on the model output magnitude.

The Engagement Signal: Critically, your engineered Engagement Score ranks higher than raw data like Scaled Balance or Credit Score, proving that behavioral metrics are the bank's best "early warning" signals.

Feature Contribution: The red and blue color distribution in your SHAP plots (Image 13) shows that high values of Age and NumOfProducts are the strongest "push" factors toward a customer exiting.



2. Partial Dependence Plots (PDP)

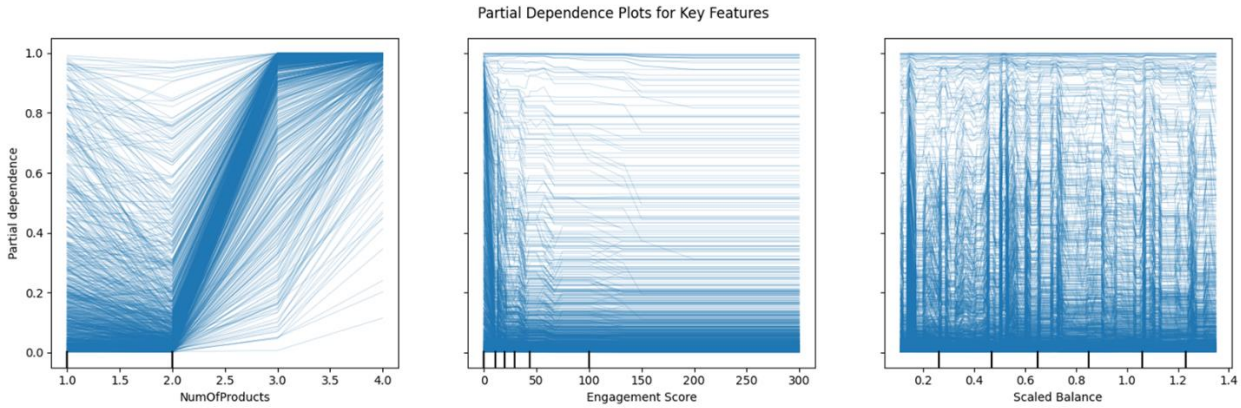
While SHAP tells us what is important, PDPs show us the "Tipping Point"—the exact value where a customer becomes a high risk.

The "3-Product Rule": Your Partial Plot for NumOfProducts (Image 12) shows a flat probability for customers with 1 or 2 products, followed by a dramatic vertical spike at 3 products.

Business Insight: The bank should intervene before a customer signs up for their third product, as this complexity correlates with higher dissatisfaction or churn.

Engagement Saturation: The Engagement Score plot (Image 12) shows that as engagement increases, the probability of churn drops significantly until it plateaus. This confirms that even small increases in customer activity can drastically improve retention.

Balance Sensitivity: Your Scaled Balance plot (Image 12) shows more "noise" (jagged lines), indicating that balance alone is not a stable predictor without considering other factors like Age.



5. Conclusion & Recommendations

The research concludes that ensemble models significantly outperform traditional statistical methods in churn forecasting. We recommend the bank implement a Random Forest-based early warning system focusing on:

Product Thresholds: Monitoring customers who hold 3+ products but show low engagement scores.

Age-Based Retention: Developing loyalty programs for customers in the 45–60 age bracket, who showed a statistically higher churn rate in the Decision Tree analysis.