Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Explainable Artificial Intelligence Framework for Customer Churn Management and Retention Decision Support in Service Operations

Muzamil Mohib

muzamilmohib@outlook.com

Nanjing University of Information Science and Technology

Research Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11767677/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Explainable Artificial Intelligence Framework for Customer Churn Management and Retention Decision Support in Service Operations

Abstract

Customer churn poses a significant challenge in service operations engineering, directly impacting revenue stability, resource allocation, and long-term system profitability. While machine learning models can predict churn, most existing approaches lack sufficient explainability and fail to translate predictions into actionable operational decisions. This study proposes an integrated explainable artificial intelligence framework that combines gradient boosting prediction (XGBoost), SHAP-based local and global explanations, probability-driven risk segmentation, and counterfactual simulation for retention strategy evaluation. The framework is rigorously validated on three publicly available real-world customer churn datasets from the telecommunications and banking sectors (IBM Telco, BigML Orange Telecom, and Banking Churn Modeling). Three models Logistic Regression, Random Forest, and XGBoost were trained and evaluated using accuracy, precision, recall, F1-score, AUC-ROC, and AUC-PR. Results demonstrate that model performance is domain-dependent, with XGBoost and Random Forest showing strong predictive capability across datasets. SHAP analysis identifies domain-specific churn drivers such as contract type and tenure in telecommunications, usage intensity and service calls in the Orange Telecom dataset, and customer age, product ownership, balance, and activity status in banking. Risk segmentation effectively distinguishes low-, medium-, and high-risk customer groups, with actual churn rates increasing monotonically across segments. Counterfactual simulations estimate substantial churn risk reduction (up to 66 percentage points) under feasible operational interventions such as contract upgrades, service bundles, and customer engagement actions. The proposed framework advances intelligent decision-support systems for service operations by bridging predictive modeling, explainability, and prescriptive analytics, enabling more efficient resource allocation, capacity planning, and retention interventions in real-world engineering environments. All datasets and code are publicly accessible to ensure research replicability.

Final keywords

Explainable artificial intelligence; Customer churn prediction; SHAP; XGBoost; Risk segmentation; Counterfactual simulation; Decision support system; Service operations engineering; Retention optimization

1. Introduction

Customer churn refers to the situation in which customers discontinue their relationship with a firm, stop purchasing its products, or stop using its services. It is a critical performance indicator for service-oriented, subscription-based, and relationship-driven businesses because a high churn rate may indicate customer dissatisfaction, weak service quality, poor relationship management, or increasing competitive pressure (Ahn et al., 2020; Zaghloul et al., 2025). In industries such as telecommunications and banking, where customers can often switch providers with relatively low effort, churn directly affects revenue stability, customer lifetime value, profitability, and long-term competitive advantage (Ascarza, 2018; Lalwani et al., 2022). Therefore, organizations need analytical systems that can identify customers likely to churn, explain the reasons for their risk, and

support targeted retention decisions before customer loss occurs.

Customer churn prediction has become an important application area for machine learning, business intelligence, and decision-support systems. Predictive models can help firms identify high-risk customers in advance and allocate retention resources more efficiently. Recent studies have applied a wide range of machine learning, ensemble learning, deep learning, and hybrid intelligent techniques to improve churn prediction performance. For example, Haddadi et al. (2024) examined churn prediction in imbalanced datasets using resampling methods, while Rao et al. (2024) proposed a multi-strategy collaborative processing method for imbalanced customer churn classification. Zhu et al. (2023) developed a bagging-based selective ensemble model for churn prediction under class imbalance, and Jiang et al. (2023) proposed a profit-driven hybrid model based on cost minimization and return maximization. In the telecommunications domain, Wang et al. (2024) introduced an FCLCNN-LSTM model for churn risk assessment, while Haridasan et al. (2025) applied gazelle optimization and conditional variational autoencoder techniques for telecom user service recommendation based on churn analysis. These studies demonstrate the growing importance of intelligent systems for customer churn prediction and retention analytics.

Despite these advances, predictive accuracy alone is not sufficient for practical churn management. In real business environments, managers need to know not only whether a customer is likely to churn, but also why the customer is at risk and which retention action may reduce that risk. This creates a need for explainable artificial intelligence in customer churn prediction. Explainable AI techniques improve transparency, trust, and accountability by helping users understand how complex machine learning models generate their predictions (Abusitta et al., 2024). Counterfactual explanation methods are especially relevant for decision support because they can indicate how a prediction may change under alternative feature conditions or intervention scenarios (Balbino et al., 2023). In churn management, these methods can help identify actionable drivers, including contract type, tenure, service usage, customer service calls, product ownership, activity status, and customer engagement indicators.

Customer segmentation is another important component of effective churn management. Firms cannot apply the same retention strategy to all customers because customers differ in value, behavior, service usage, needs, and risk levels. Segmentation allows customers to be divided into meaningful groups according to behavioral, transactional, demographic, or risk-based similarities (Anitha & Patil, 2022; Rungruang et al., 2024). Recent studies have emphasized the role of segmentation in decision support and targeted marketing. Rungruang et al. (2024) proposed a hierarchical customer segmentation approach based on the RFM model and formal concept analysis, while Barrera et al. (2024) developed a multiple-criteria decision-support system for customer segmentation using a sorting outranking method. Zaghloul et al. (2025) further demonstrated the value of combining segmentation and churn prediction through a hybrid RFM, K-means, and deep neural network framework for online retail. However, segmentation becomes more useful for managerial decision-making when it is connected with predicted churn probability, model explanations, and retention strategy simulation.

Although the existing literature has made substantial progress in churn prediction, imbalanced classification, segmentation, and explainable AI, several important gaps remain. First, many churn studies focus on a single dataset or one business domain, which limits the generalizability of the findings across industries. Second, most studies emphasize predictive performance metrics such as accuracy, F1-score, AUC-ROC, and AUC-PR, but pay less attention to whether predicted probabilities can be translated into meaningful customer risk groups. Third, although explainable AI methods are increasingly used, they are often applied only as post-hoc feature-ranking tools and are not fully connected to validated risk segmentation or retention strategy design. Fourth, few studies integrate churn prediction, SHAP-based explainability, probability-based risk segmentation, and counterfactual retention strategy simulation into a single decision-support framework. As a result, the link between predictive analytics and practical retention action remains underdeveloped.

To address these limitations, this study proposes an explainable multi-dataset churn management framework that integrates machine-learning prediction, SHAP-based explainability, probability-based risk segmentation, and counterfactual retention-strategy simulation. The study uses three publicly available churn datasets: IBM Telco, Orange Telecom, and Banking Churn. These datasets allow the proposed framework to be evaluated across both telecommunications and banking contexts, supporting within-domain validation in telecommunications and cross-industry validation between telecommunications and financial services. Logistic Regression, Random Forest, and XGBoost are trained and evaluated using a consistent modeling pipeline. The models are assessed using accuracy, precision, recall, F1-score, AUC-ROC, and AUC-PR to account for the imbalanced nature of churn prediction. SHAP analysis is then applied to identify the most influential churn drivers, predicted probabilities are used to classify customers into low-, medium-, and high-risk groups, and counterfactual simulations are conducted to estimate the potential churn-risk reduction associated with selected managerial interventions.

The following research questions guide the study:

RQ1: Can machine learning models provide robust churn prediction performance across telecommunications and banking datasets?

RQ2: Which customer attributes are the most influential drivers of churn across different business domains?

RQ3: Can predicted churn probabilities be translated into meaningful low-, medium-, and high-risk customer segments?

RQ4: Can a counterfactual retention strategy simulation estimate the potential reduction in churn risk associated with targeted managerial actions?

This study makes several contributions to customer churn prediction, explainable AI, and business-intelligence decision support. First, the study proposes an explainable multi-dataset churn management framework that integrates churn prediction, SHAP-based explanation, risk segmentation, and retention strategy simulation. By validating the framework on IBM Telco, Orange Telecom, and Banking Churn datasets, the study provides evidence of both cross-dataset and cross-industry applicability. Second, the study combines predictive performance evaluation with interpretability. Logistic Regression, Random

Forest, and XGBoost are compared using multiple classification metrics, and SHAP analysis is then applied to explain the key factors driving churn predictions. This combination improves both technical evaluation and managerial interpretability. Third, the study translates predicted churn probabilities into low-, medium-, and high-risk customer groups and validates these groups using actual churn rates. This step connects model outputs with practical customer prioritization and helps managers identify which customers should receive stronger retention attention. Fourth, the study extends churn prediction to action-oriented decision support by conducting counterfactual retention-strategy simulations. These simulations estimate the potential reduction in churn risk associated with managerial interventions such as contract upgrades, service recovery, support bundles, cross-selling, product portfolio adjustments, and pricing-related actions. Finally, the study contributes an explainable AI-driven intelligent decision-support framework that moves beyond black-box churn prediction by explaining why customers are at risk, validating their risk levels, and estimating which actions may reduce predicted churn probability.

The remainder of this paper is organized as follows. Section 2 reviews related work on customer segmentation, churn prediction, imbalanced classification, explainable AI, and counterfactual decision support. Section 3 presents the background and fundamental tools used in the study. Section 4 introduces the proposed explainable AI-driven churn prediction and retention framework. Section 5 describes the methodology, including datasets, preprocessing procedures, machine learning models, evaluation metrics, SHAP analysis, risk segmentation, and counterfactual simulation. Section 6 presents the experimental results and comparison with recent literature. Section 7 discusses the theoretical, managerial, and practical deployment implications. Finally, Section 8 concludes the paper by outlining its limitations and future research directions.

2. Related work

This section reviews the literature relevant to explainable customer churn prediction, intelligent decision support, and retention management. The review is organized into three streams. The first stream discusses customer segmentation and risk grouping as tools for customer relationship management and resource allocation. The second stream reviews machine learning-based churn prediction and imbalanced classification, with emphasis on intelligent systems developed for telecommunications, banking, and customer analytics. The third stream examines explainable artificial intelligence and counterfactual decision-support methods, which are essential for translating model predictions into managerial action. Together, these streams establish the need for an integrated framework that moves beyond model accuracy toward explainable, validated, and action-oriented churn management**.**

2.1. Customer segmentation and risk grouping

Customer segmentation is a central analytical process in customer relationship management. It divides customers into distinct groups according to shared characteristics, behavioral patterns, value levels, service usage, or predicted risk profiles. In churn management, segmentation is especially important because customers do not have the same probability of leaving, economic value, or response to retention actions. Instead of applying the same retention strategy to all customers, firms can use customer segmentation to identify low-,

medium-, and high-risk customers and allocate retention resources more efficiently.

Traditional segmentation approaches often rely on behavioral and transactional indicators. One widely used method is recency, frequency, and monetary value analysis, which groups customers according to how recently they purchased, how frequently they purchased, and how much they spent (Anitha & Patil, 2022; Ernawati et al., 2021a; Wu et al., 2020). RFM-based segmentation has been widely applied in e-commerce and customer analytics because it provides an intuitive and interpretable way to classify customers according to purchasing behavior (Rungruang et al., 2024; Zaghloul et al., 2025). Rungruang et al. (2024) proposed an RFM model for customer segmentation based on a hierarchical approach using formal concept analysis. Their study demonstrated that recency, frequency, and monetary value can generate meaningful customer groups for decision support. Similarly, Zaghloul et al. (2025) integrated RFM analysis, K-means clustering, and deep neural networks to predict customer churn in online retail, demonstrating that segmentation can support churn analytics when combined with predictive modeling.

Other studies have used clustering and multi-criteria decision-support methods to improve segmentation quality. K-means clustering is commonly used to partition customers into behaviorally similar groups based on their distances from cluster centroids (Rajput & Singh, 2023; Shirole et al., 2021). The quality of clustering is often evaluated using internal validation methods such as the silhouette score, which measures the separation and cohesion of clusters (Januzaj et al., 2023; Wang et al., 2017). Barrera et al. (2024) developed a multiple-criteria decision-support system for customer segmentation using a sorting outranking method. Their work shows that segmentation can be improved by combining customer characteristics with structured decision rules.

Although segmentation studies provide useful tools for classifying customers, several limitations remain. First, segmentation is often treated as an independent analytical stage rather than being integrated with predictive churn modeling. Second, many segmentation approaches classify customers but do not validate whether these groups correspond to actual churn behavior. Third, segmentation studies often stop at customer grouping and do not estimate how specific managerial interventions may reduce churn risk. These limitations suggest the need for a framework that connects predicted churn probabilities, validated risk groups, explainable AI, and retention strategy simulation**.**

2.2. Customer churn prediction and imbalanced classification

Customer churn prediction has been widely studied using statistical models, traditional machine learning, ensemble learning, deep learning, and hybrid intelligent systems. The main objective of churn prediction is to identify customers likely to discontinue using a service before churn occurs, enabling firms to implement proactive retention strategies. In telecommunications, banking, and online retail, churn prediction is particularly important because customer acquisition costs are often high and customer switching is relatively easy (Ahn et al., 2020; Ascarza, 2018; Lalwani et al., 2022).

Traditional machine learning models such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines have been widely applied in churn prediction because they are computationally efficient and relatively easy to implement (Baghla & Gupta,

2022; Xiahou & Harada, 2022). Logistic Regression is often used as a baseline classifier because it estimates the probability of churn in an interpretable manner. However, linear models may struggle to capture nonlinear relationships and interaction effects in complex customer behavior data. Tree-based and ensemble models address this limitation by capturing more flexible decision structures. Random Forest and boosting-based models have shown strong performance in structured churn datasets because they can model nonlinear feature interactions and reduce overfitting through ensemble learning (Zhang et al., 2022; Zhu et al., 2023).

A major challenge in churn prediction is class imbalance. In many churn datasets, non-churn customers represent the majority class, while churn customers form the minority class. This imbalance can cause models to achieve high accuracy while failing to identify actual churners correctly. Haddadi et al. (2024) examined customer churn prediction in imbalanced datasets using resampling methods and showed that class imbalance is a critical issue in churn classification. Rao et al. (2024) proposed a multi-strategy collaborative processing method for imbalanced customer churn classification. Their study further confirms that imbalance-aware modeling is necessary when evaluating churn prediction systems. Zhu et al. (2023) developed a bagging-based selective ensemble model for churn prediction on imbalanced data, demonstrating that ensemble approaches can improve prediction performance under class imbalance.

Recent research has also moved toward profit-driven, deep learning, and hybrid intelligent approaches. Jiang et al. (2023) proposed a profit-driven hybrid model based on cost minimization and return maximization, showing that churn prediction should be evaluated not only through statistical accuracy but also through business value. Wang et al. (2024) developed an FCLCNN-LSTM model for assessing telco customer churn risk, demonstrating that deep learning models can capture complex behavioral patterns in telecommunications data. Panimalar et al. (2025) proposed an intensified customer churn prediction model using a weighted multilayer perceptron and enhanced multipath backpropagation. Haridasan et al. (2025) used gazelle optimization and a conditional variational autoencoder for telecom user service recommendation based on churn analysis. These studies show that advanced intelligent systems are increasingly used to improve churn prediction performance.

However, the reviewed churn prediction studies remain limited in several ways. Many studies focus primarily on predictive accuracy, F1 Score, AUC-ROC, AUC-PR, or other classification metrics. While these metrics are important, they do not fully address the managerial problem of churn reduction. A high-performing model may identify likely churners, but managers still need to understand why customers are at risk and which interventions may reduce that risk. Moreover, many studies evaluate only a single dataset or industry, limiting the generalizability of their findings across different business contexts. Therefore, churn prediction research needs frameworks that combine robust model evaluation, explainable prediction, validated risk grouping, and actionable retention decision support.

2.3. Explainable AI and counterfactual retention decision support

Explainable artificial intelligence has become an important research direction because many

high-performing machine learning models operate as black boxes. In business decision support, interpretability is essential because managers must understand the reasons behind model predictions before using them to make decisions. In churn prediction, explainability is especially important because a predicted churn probability does not automatically indicate which customer attributes or service conditions to target with retention actions.

Abusitta et al. (2024) reviewed explainable AI techniques, challenges, and open issues, emphasizing that explainability improves transparency, trust, and accountability in intelligent systems. In customer churn prediction, explainability can help managers understand whether short tenure, month-to-month contracts, high service charges, repeated service calls, low account activity, product ownership, or other customer-specific factors drive churn risk. Among explainable AI techniques, SHAP has become widely used for its ability to provide both global and local explanations. Global SHAP explanations identify the most influential features across the dataset, while local explanations indicate how each feature contributes to the predicted churn probability for an individual customer.

Counterfactual explanation methods are also relevant for churn management because they support action-oriented decision-making. Balbino et al. (2023) proposed an agnostic method for counterfactual, selected, and social explanations for classification models. Counterfactual explanations are useful because they not only explain why a prediction occurs; they also show how a prediction may change if selected conditions are modified. In churn management, this logic can be used to estimate whether changing a customer's contract type, improving service support, modifying product ownership, or increasing customer activity may reduce predicted churn risk.

Despite the growing use of explainable AI, many churn prediction studies use interpretability only as a post-hoc feature-ranking tool. This limits its managerial usefulness because feature importance alone does not tell managers which retention strategy to implement. Similarly, many segmentation studies identify customer groups but do not estimate how specific interventions affect predicted churn risk. These limitations create a gap between explainable prediction and managerial action.

The present study addresses this gap by integrating SHAP-based interpretation with probability-based risk segmentation and counterfactual retention strategy simulation. SHAP analysis identifies the main drivers of churn across the IBM Telco, Orange Telecom, and Banking Churn datasets. Predicted probabilities are then used to classify customers into low-risk, medium-risk, and high-risk groups. Finally, counterfactual simulations estimate the potential reduction in churn risk associated with targeted managerial strategies, including contract upgrades, service recovery, support bundles, cross-selling, product portfolio adjustments, and customer activation. This approach extends churn prediction from a classification task to an explainable, action-oriented decision-support framework.

Overall, the literature shows important progress in churn prediction, imbalanced classification, segmentation, and explainable AI. However, most studies remain limited in three ways. First, they often evaluate a single dataset or business domain. Second, explainability is frequently treated as a post-hoc interpretation step rather than being

connected to validated risk segmentation. Third, few studies translate predicted churn risk into counterfactual retention-strategy simulations. The present study addresses these limitations by proposing an explainable multi-dataset churn management framework that integrates machine learning prediction, SHAP-based interpretation, probability-based risk segmentation, and action-oriented retention simulation.

3. Background and fundamental tools

The proposed framework relies on several computational tools commonly used in intelligent decision-support systems: Logistic Regression, Random Forest, XGBoost, SHAP explainability, probability-based risk segmentation, and counterfactual retention strategy simulation. These tools are not used only for algorithm comparison; rather, they form an integrated pipeline for designing, testing, and interpreting an explainable churn management system. Logistic Regression provides a transparent baseline; Random Forest and XGBoost provide nonlinear ensemble learning capabilities; SHAP explains model outputs; risk segmentation translates predicted probabilities into managerial priority groups; and counterfactual simulation estimates the potential effect of alternative retention actions. Together, these components support the development of a practical business-intelligence system for churn prediction and retention decision-making.

3.1. Logistic Regression

Logistic Regression is used as a transparent baseline classifier for binary churn prediction. It estimates the probability that a customer belongs to the churn class based on customer-level attributes. Because of its simplicity and interpretability, Logistic Regression provides a useful benchmark against more flexible machine learning models. However, it may be limited when churn risk depends on nonlinear relationships and interaction effects among behavioral, demographic, service, and financial variables.

3.2. Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive stability and reduce overfitting. In churn prediction, it is useful because it can capture nonlinear relationships and interaction effects among customer features. Prior churn studies have shown that tree-based ensemble methods can perform effectively in imbalanced and nonlinear churn classification settings (Zhu et al., 2023). In this study, Random Forest is used as a nonlinear ensemble benchmark.

3.3. XGBoost

Extreme Gradient Boosting, or XGBoost, is a gradient boosting algorithm that builds decision trees sequentially, with each new tree correcting errors from previous trees. It is widely used for structured tabular data because of its predictive strength, regularization mechanisms, and ability to capture complex feature interactions. In this study, XGBoost is used as the primary explainable model because it delivers competitive performance and can be interpreted with SHAP values.

3.4. SHAP explainability

Shapley Additive explanations (SHAP) are an explainable AI method that assigns each feature a contribution value, indicating how much it increases or decreases a model's prediction. In churn prediction, SHAP helps explain why a customer is classified as high risk by identifying the features that push the prediction toward churn. SHAP provides both global explanations, which identify overall

churn drivers, and local explanations, which explain individual customer predictions. This is important because explainable AI can improve transparency, trust, and managerial adoption of machine learning systems (Abusitta et al., 2024).

3.5. Risk segmentation

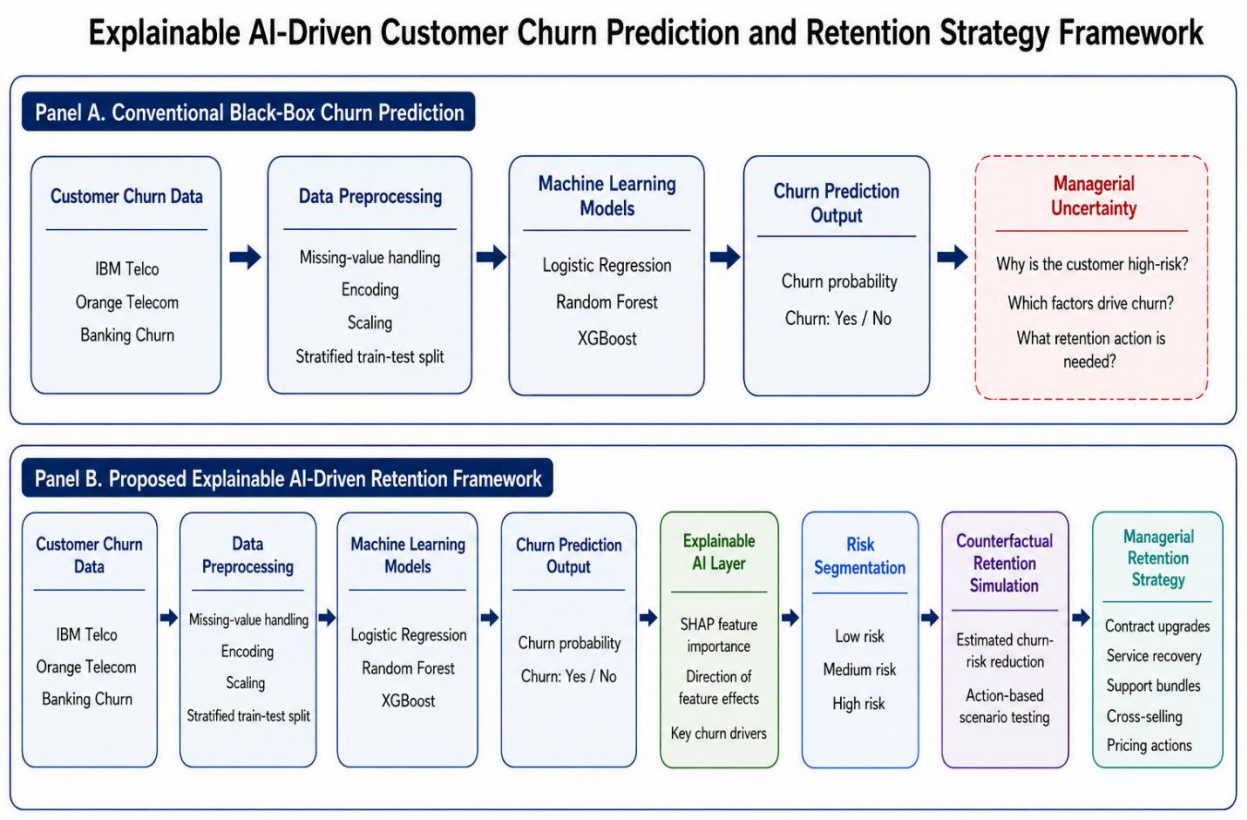
Risk segmentation divides customers into low-risk, medium-risk, and high-risk groups based on predicted churn probabilities. This converts continuous model scores into actionable customer categories. Such segmentation helps managers prioritize retention resources for customers at higher churn risk rather than applying costly interventions to all customers. Recent segmentation studies emphasize the value of customer grouping for decision support and targeted strategy design (Barrera et al., 2024; Rungruang et al., 2024). In this study, risk segmentation is validated by comparing actual churn rates across the predicted risk groups.

3.6. Counterfactual retention strategy simulation

Counterfactual retention strategy simulation estimates how predicted churn risk may change if selected customer attributes are modified to represent possible managerial actions. Counterfactual explanation methods are useful because they show how model predictions may change under alternative conditions (Balbino et al., 2023). In this study, simulations are used to evaluate potential interventions, including contract upgrades, support-service bundles, service recovery, product cross-selling, customer activation, and product portfolio simplification. These simulations are interpreted as model-based decision-support estimates rather than causal treatment effects.

4. The proposed approach

The proposed approach develops an explainable AI-driven churn management framework to support business intelligence and retention decision-making. The framework integrates machine learning prediction, SHAP-based explainability, probability-based risk segmentation, and counterfactual retention simulation into a unified intelligent system. Unlike conventional churn prediction models that primarily produce probability scores or binary churn labels, the proposed framework links predictive outputs to interpretable explanations and actionable managerial strategies. In this way, the study addresses the practical need for intelligent systems that not only identify high-risk customers but also help managers understand churn drivers, prioritize customers, and evaluate potential retention actions.



As illustrated in Figure 1, the framework is structured around two linked perspectives. Panel A represents conventional black-box churn prediction, where customer churn data are preprocessed, machine learning models are trained, and churn prediction outputs are generated. This conventional process can identify whether a customer is likely to churn. However, it often leaves managers uncertain about why the customer is high risk, which factors drive the prediction, and what retention action to take. Panel B presents the proposed explainable AI-driven retention framework, which extends the conventional prediction pipeline by adding SHAP-based explanation, risk segmentation, counterfactual retention simulation, and managerial strategy formulation.

The framework begins with three publicly available customer churn datasets: IBM Telco, Orange Telecom, and Banking Churn. These datasets represent different service contexts and enable both same-domain and cross-industry validation. IBM Telco and Orange Telecom provide validation in the telecommunications sector, while the Banking Churn dataset enables evaluation of the framework in a financial services context. This multi-dataset design enhances the generalizability of the proposed framework compared with studies based on a single dataset or a single industry.

The first stage is data preprocessing. Each dataset is cleaned and transformed before model training. Missing values are handled, categorical variables are encoded, numerical variables are scaled where required, and stratified train-test splits are applied to preserve the original churn distribution across the

training and test subsets. These steps are necessary because churn datasets are typically imbalanced and contain mixed data types, including demographic, service, usage, billing, and account-level variables.

The second stage is machine learning-based churn prediction. Three supervised models are trained and evaluated: Logistic Regression, Random Forest, and XGBoost. These models are selected because they represent different levels of complexity. Logistic Regression provides a transparent statistical baseline, Random Forest provides a nonlinear ensemble benchmark, and XGBoost provides a high-performing gradient boosting model suitable for SHAP-based interpretation. Model performance is evaluated using accuracy, precision, recall, F1-score, AUC-ROC, and AUC-PR. Using multiple evaluation metrics is important because churn prediction is an imbalanced classification problem, and accuracy alone may be misleading.

The third stage is SHAP-based explainability. After model evaluation, SHAP analysis is applied to the XGBoost models to identify the most important drivers of churn in each dataset. This stage explains how different features influence predicted churn probability. For example, in telecom datasets, churn risk may be influenced by tenure, contract type, internet service, monthly charges, technical support, and customer service calls. In the banking dataset, churn risk may be influenced by age, number of products, balance, activity status, credit score, geography, and salary. SHAP therefore provides interpretable insight into both common and domain-specific churn drivers.

The fourth stage is probability-based risk segmentation. Predicted churn probabilities are used to classify customers into low-, medium-, and high-risk groups. The validity of these groups is examined by comparing actual churn rates across the predicted risk categories. A reliable risk segmentation system should show a clear increase in actual churn rates from low-risk to high-risk customers. This stage transforms model outputs into practical customer prioritization groups that managers can use for retention planning.

The fifth stage is counterfactual retention strategy simulation. In this stage, selected customer attributes are modified to reflect potential managerial interventions, and the model estimates the resulting change in the predicted churn risk. The simulated actions include contract upgrades, service recovery, technical support bundles, cross-selling, product portfolio adjustment, customer activation, and pricing-related interventions. The purpose of this stage is not to claim causal effects, but to provide model-based estimates of how predicted churn risk may change under alternative customer conditions.

The framework's final output is a managerial retention strategy. The results of prediction, explanation, segmentation, and simulation are combined to support targeted retention decisions. Rather than only identifying high-risk customers, the framework helps managers understand why customers are at risk, which groups should be prioritized, and which interventions may reduce predicted churn probability. This makes the proposed framework suitable for business intelligence applications that require interpretability, customer prioritization, and actionability.

The proposed approach, therefore, consists of four integrated decision-support components: machine learning-based churn prediction, SHAP-based explanation, probability-based risk segmentation, and counterfactual retention

strategy simulation. By combining these components, the framework provides a more complete churn management system than prediction accuracy alone. It advances churn analytics from black-box classification to an explainable, actionable intelligent system that can support business decision-making in telecommunications and banking contexts.

5. Methodology

This section describes the datasets, preprocessing procedures, modeling strategy, evaluation metrics, SHAP-based explainability analysis, risk segmentation procedure, and counterfactual retention simulation. The same analytical pipeline was applied to all three datasets to ensure consistency and comparability across domains.

5.1. Data sources and dataset description

This study uses three publicly available customer churn datasets obtained from Kaggle: IBM Telco Customer Churn, Orange Telecom Churn, and Banking Customer Churn. These datasets were selected because they represent different service contexts and allow the proposed framework to be evaluated across both same-domain and cross-industry settings. IBM Telco and Orange Telecom provide evidence from the telecommunications sector, while the Banking Churn dataset allows validation in a financial services context.

The IBM Telco dataset contains 7,043 customer records and includes demographic information, account characteristics, service subscriptions, billing variables, contract type, tenure, monthly charges, total charges, and churn status. The Orange Telecom dataset contains 3,333 customer records and includes account length, international plan, voicemail plan, service usage variables, call-related variables, customer service calls, and churn status. The Banking Churn dataset contains 10,000 customer records and includes credit score, geography, gender, age, tenure, balance, number of products, credit card ownership, active membership status, estimated salary, and churn status.

For consistency across datasets, all target variables were converted into a common binary format, where 1 represents churn, and 0 represents non-churn. Non-predictive identifier variables were removed before modeling. The Orange Telecom data were provided as two subsets in the original repository; these were combined into a single dataset before applying the same preprocessing and stratified train-test split used for the other datasets.

Table 1. Dataset overview

Dataset	Domain	Records	Features	Churners	Non-churners	Churn rate
IBM Telco	Telecommunications	7,043	19	1,869	5,174	26.5%
Orange Telecom	Telecommunications	3,333	19	483	2,850	14.5%
Banking Churn	Banking	10,000	10	2,037	7,963	20.4%

Note. Predictive features used refer to the number of variables retained after removing non-predictive identifier variables and excluding the target variable

5.2. Exploratory data analysis

Exploratory Data Analysis (EDA) was conducted prior to model development to examine the dataset's structure, feature distributions, missing values, and target-class composition. This step was important because churn datasets are typically imbalanced, with non-churners representing the majority class. Understanding class distribution helped guide the choice of evaluation metrics and imbalance-handling procedures.

For the IBM Telco dataset, the exploratory analysis focused on tenure, monthly charges, total charges, contract type, internet service, payment method, and support-related services. These variables reflect the duration of the customer relationship, billing exposure, contractual commitment, and service configuration. For the Orange Telecom dataset, the analysis focused on account length, international plan, voicemail plan, usage minutes, usage charges, and customer service calls, which capture service usage intensity and customer interaction patterns. For the Banking Churn dataset, the analysis examined age, credit score, tenure, balance, number of products, active membership, geography, gender, and estimated salary, which reflect customer financial profile and relationship depth.

Both graphical and non-graphical procedures were used during EDA. Non-graphical procedures included descriptive statistics, missing-value checks, churn-rate calculations, and class distribution summaries. Graphical procedures included churn distribution plots, feature comparison plots, and numerical correlation heatmaps. The EDA results informed preprocessing decisions, model evaluation, and interpretation of SHAP-based churn drivers. The EDA figures are reported in the supplementary material to keep the main manuscript focused on model performance, explainability, risk segmentation, and retention simulation.

5.3. Data preprocessing

Data preprocessing was applied consistently across the three datasets to ensure that the data were clean, comparable, and suitable for machine learning. First, non-predictive identifier variables were removed. In the IBM Telco dataset, the customer ID variable was removed. In the Banking Churn dataset, variables such as Row Number, Customer ID, and Surname were excluded because they do not provide meaningful behavioral or managerial information for churn prediction.

Second, numerical variables stored as text were converted into numeric format. In the IBM Telco dataset, Total Charges was converted into a numeric variable because it may contain blank or invalid entries. Missing numerical values were treated using median imputation, while missing categorical values were treated using the most frequent category. This ensured that incomplete values did not prevent model estimation.

Third, the churn target was harmonized across datasets. In the telecommunications datasets, churn labels were encoded as 1 for churn and 0 for non-churn. In the banking dataset, the Exited variable was treated as the churn outcome, with 1 indicating that the customer left the bank and 0 indicating that the customer remained.

Fourth, categorical variables were one-hot encoded. This allowed categorical predictors such as contract type, internet service, payment method, gender, geography, international plan, and voicemail plan to be used in the machine learning models. Numerical variables were

standardized using StandardScaler, particularly to support Logistic Regression and to ensure that variables with larger numeric ranges did not dominate the modeling process.

Finally, each dataset was divided into training and testing subsets using an 80/20 stratified split. Stratification preserved the original churn and non-churn proportions in both subsets, which was important because all three datasets showed class imbalance.

5.4. Machine learning models and imbalance handling

Three supervised machine learning models were trained and compared: Logistic Regression, Random Forest, and XGBoost. These models were selected because they represent different levels of model complexity and interpretability. Logistic Regression served as a transparent baseline model. Random Forest was used as a nonlinear ensemble benchmark to capture interactions among customer attributes. XGBoost was used as a gradient boosting model because of its strong performance on structured tabular data and its compatibility with SHAP-based explainability.

The models were trained separately on each dataset using the same preprocessing pipeline. To mitigate the effects of class imbalance, Logistic Regression and Random Forest were estimated with class-weighting strategies. For XGBoost, an imbalance-adjusted positive-class weight was calculated from the ratio of non-churners to churners in the training data. This allowed the model to give greater attention to the minority churn class.

5.5. Probability estimation and evaluation metrics

The proposed framework uses predicted churn probabilities for risk segmentation and counterfactual simulation; therefore, probability estimation is an important part of the modeling pipeline. After each model was trained on the training set, predicted churn probabilities were generated for the held-out test set using the model's probability output. These predicted probabilities were then used for model evaluation, risk segmentation, and simulation of counterfactual retention strategies.

Model performance was evaluated using accuracy, precision, recall, F1-score, AUC-ROC, and AUC-PR. Accuracy measures the overall proportion of correct predictions. Precision measures the proportion of predicted churners who actually churned, while recall measures the proportion of actual churners correctly identified by the model. The F1-score combines precision and recall into a single balanced measure. AUC-ROC evaluates the model's ability to distinguish churners from non-churners across classification thresholds. AUC-PR evaluates the precision-recall trade-off and is particularly useful when the positive class is relatively rare. Because churn prediction is an imbalanced classification problem, the interpretation emphasizes F1-score, AUC-ROC, and AUC-PR rather than relying only on accuracy.

5.6. SHAP explainability

After model evaluation, SHAP analysis was applied to the XGBoost model for each dataset. SHAP values quantify how each feature contributes to the predicted probability of churn for a given observation. Positive SHAP values indicate that a feature increases the predicted probability of churn, while negative SHAP values indicate that a feature reduces the predicted probability of churn. Mean absolute SHAP values were used to rank the most important churn drivers in each dataset.

For IBM Telco, SHAP analysis focused on variables such as tenure, contract type, monthly charges, total charges, internet service, and support-related services. For Orange Telecom, SHAP analysis examined service usage variables, customer service calls, international plan, voicemail plan, and charge-related features. For Banking Churn, SHAP analysis examined age, number of products, balance, active membership, geography, gender, and estimated salary. The SHAP results provided the explanatory basis for identifying why customers were classified as high risk.

5.7. Risk segmentation

Predicted churn probabilities were used to classify customers into low-risk, medium-risk, and high-risk groups. This step translated continuous model scores into actionable customer-prioritization categories. In this study, customers were divided into three quantile-based risk groups based on their predicted churn probabilities. Customers in the lowest probability tertile were classified as low risk; those in the middle tertile as medium risk; and those in the highest probability tertile as high risk.

After assigning customers to risk groups, the actual churn rate was calculated within each group. A valid risk segmentation should show that actual churn rates increase from the low-risk group to the high-risk group. This validation step is important because it demonstrates whether the predicted probability scores are meaningful for managerial prioritization. If the high-risk group contains a substantially larger proportion of churners than the low-risk group, managers can use the segmentation to allocate retention resources more efficiently.

5.8. Counterfactual retention strategy simulation

A counterfactual retention strategy simulation was conducted to estimate how predicted churn risk may change under selected managerial interventions. For each dataset, intervention scenarios were designed based on SHAP-identified churn drivers and domain logic. This step extended the analysis from prediction and explanation to action-oriented decision support.

For IBM Telco, simulated strategies included upgrading month-to-month customers to one-year or two-year contracts and bundling support services such as Tech Support and Online Security. For Orange Telecom, simulated strategies included replacing customers' international plans with international plans and improving service recovery for customers with three or more service calls. For Banking Churn, simulated strategies included cross-selling a second product to one-product customers and simplifying product portfolios for customers with more than two products.

For each intervention, the XGBoost model first estimated the original churn probability for the target customers. Then the relevant customer attribute was modified to represent the counterfactual intervention scenario, and the model generated a new predicted churn probability. The difference between the original and counterfactual probabilities was interpreted as the predicted reduction in churn risk. Expected retained customers were estimated as:

Expected retained customers were estimated as:

$$\text{Expected Retained customers} = \text{Target Customers} \times \frac{\text{Predicted Risk Reduction}}{100}.$$

This value should be interpreted as a model-based decision-support estimate rather than a guaranteed number of retained customers.

5.9. Reproducibility and implementation

All experiments were implemented in Python using standard machine learning and explainable AI libraries. Scikit-learn was used for preprocessing, train-test splitting, scaling, one-hot encoding, Logistic Regression, Random Forest, and evaluation metrics. XGBoost was used for gradient boosting classification, while SHAP was used for model interpretation. Matplotlib and Seaborn were used for visualization. A fixed random seed was used to ensure reproducibility, and the same general pipeline was applied across all datasets.

6. Experiments and results

This section presents the empirical findings of the proposed explainable multi-dataset churn management framework. The results are organized into five parts: model performance comparison, SHAP-based explainability, risk segmentation validation, counterfactual retention strategy simulation, and comparison with recent literature.

6.1. Model performance comparison

Table 2 presents the holdout performance of Logistic Regression, Random Forest, and XGBoost across the three datasets.

Table 2. Performance comparison of machine learning models on the holdout test set

Dataset	Model	Accuracy	Precision	Recall	F1	AUC-ROC	AUC-PR
IBM Telco	Logistic Regression	0.738	0.504	0.783	0.614	0.842	0.634
IBM Telco	Random Forest	0.789	0.627	0.508	0.561	0.820	0.611
IBM Telco	XGBoost	0.759	0.536	0.695	0.605	0.832	0.635
Orange Telecom	Logistic Regression	0.784	0.376	0.732	0.497	0.826	0.463
Orange Telecom	Random Forest	0.925	1.000	0.485	0.653	0.929	0.872
Orange Telecom	XGBoost	0.960	0.927	0.784	0.849	0.946	0.889
Banking Churn	Logistic Regression	0.714	0.387	0.700	0.499	0.777	0.468
Banking Churn	Random Forest	0.859	0.766	0.442	0.561	0.851	0.682
Banking Churn	XGBoost	0.823	0.558	0.619	0.587	0.836	0.660

The results show that model performance varies across datasets and evaluation metrics. For the IBM Telco dataset, Logistic Regression achieved the highest AUC-ROC, whereas XGBoost achieved the highest AUC-PR. Random Forest produced the highest accuracy and precision but showed lower recall, indicating a more conservative churn classification pattern. For the Orange Telecom dataset, XGBoost achieved the strongest

overall performance, with the highest accuracy, recall, F1-score, AUC-ROC, and AUC-PR. For the Banking Churn dataset, Random Forest achieved the highest accuracy, precision, AUC-ROC, and AUC-PR, while XGBoost produced the highest F1-score. The Random Forest model achieved perfect precision on the Orange Telecom test split (1.000). However, its relatively low recall (0.485) indicates a conservative classification strategy that avoids false positives at the cost of missing many true churners.

To provide a clearer visual comparison, Fig. 2 reports the cross-dataset performance of the three models using AUC-ROC and F1-score. AUC-ROC captures the models' ability to discriminate between churners and non-churners, while F1-score provides a balanced measure of precision and recall, which is particularly important for imbalanced churn datasets.

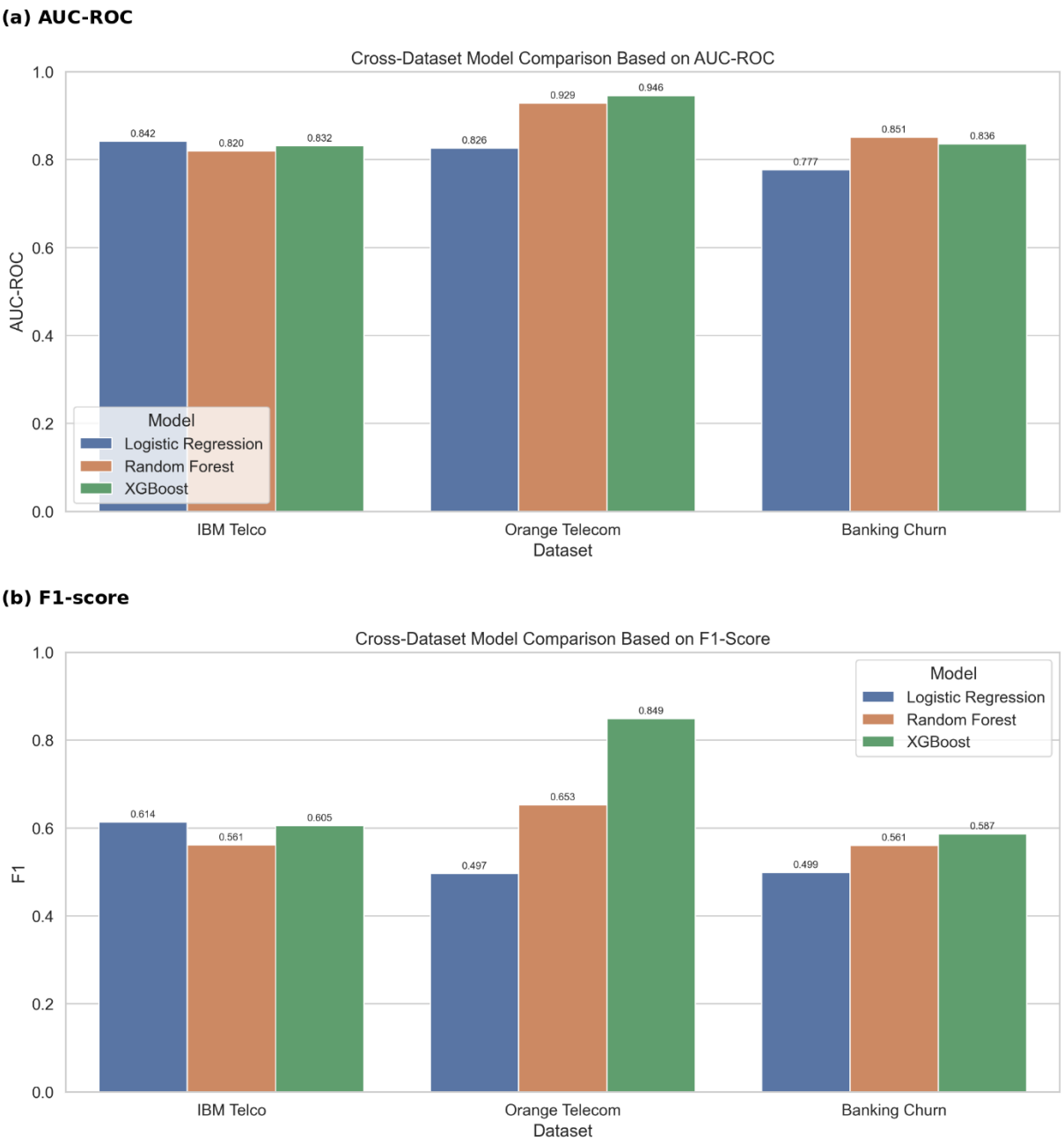


Fig. 2. Cross-dataset model comparison based on (a) AUC-ROC and (b) F1-score.

As shown in Fig. 2, XGBoost performs particularly well on the Orange Telecom dataset, while Random Forest performs strongly on the Banking Churn dataset. Logistic Regression remains competitive in IBM Telco, especially in terms of AUC-ROC and recall. These findings indicate that no single model dominates across all datasets and all metrics. However, tree-based ensemble approaches, particularly Random Forest and XGBoost, show strong performance on structured churn data. XGBoost was selected for the explainability stage because it delivers competitive predictive performance across

datasets and is well-suited for SHAP-based interpretation. This enables the analysis to move beyond model comparison and examine the customer-level drivers of churn risk.

6.2. SHAP explainability results

After evaluating model performance, SHAP analysis was applied to the XGBoost models to identify the most influential drivers of churn across the three datasets. SHAP values quantify the contribution of each feature to the model prediction and therefore help explain why specific customers are classified as high or low churn risk. Table 3 reports the five most important features for each dataset based on mean absolute SHAP values.

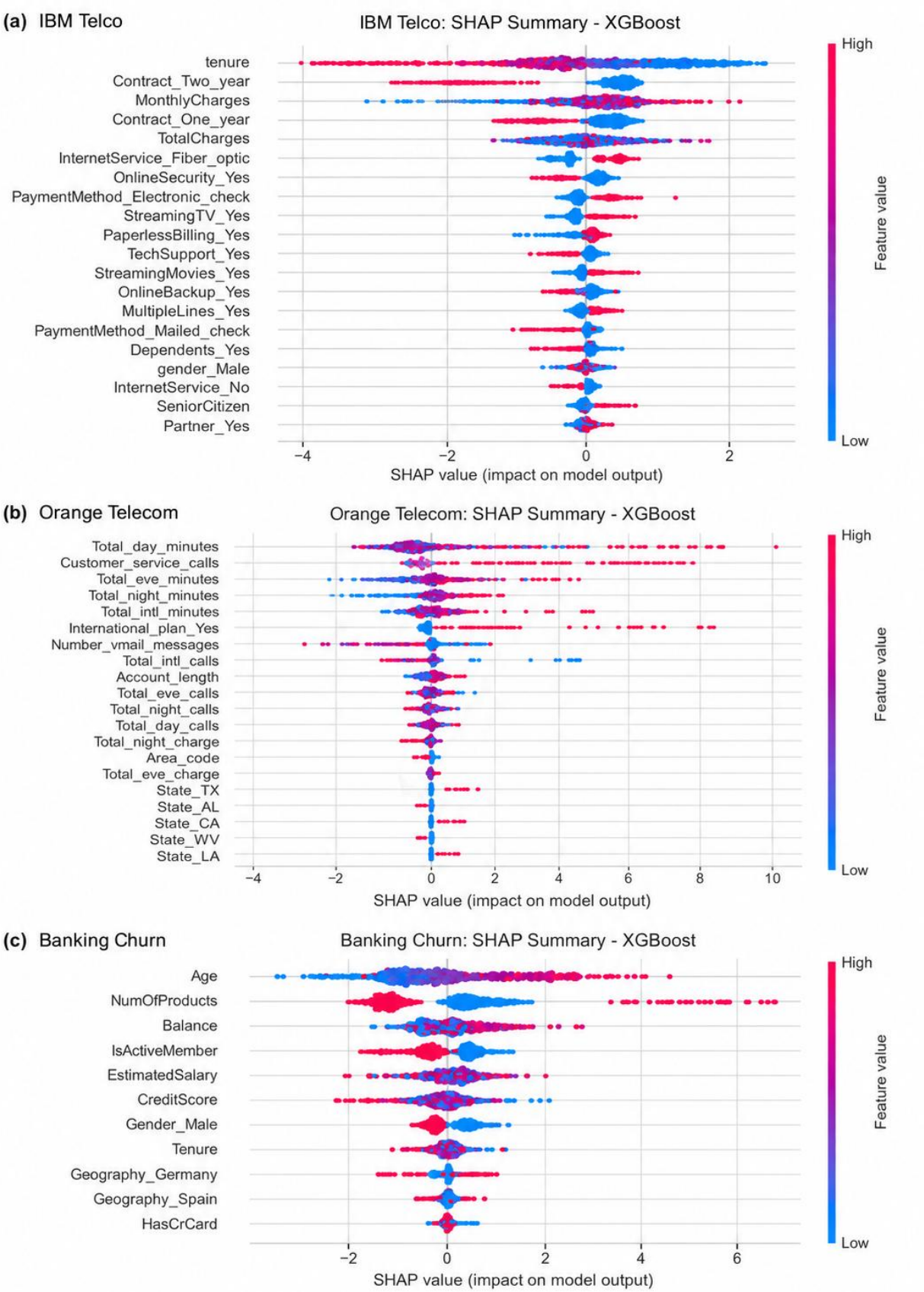
Table 3. Top five churn drivers based on mean absolute SHAP values

Rank	IBM Telco	Orange Telecom	Banking Churn
1	Tenure (0.994)	Total day minutes (1.224)	Age (1.062)
2	Contract two-year (0.806)	Customer service calls (0.796)	Number of products (0.991)
3	Monthly charges (0.523)	Total eve minutes (0.656)	Balance (0.487)
4	Contract one year (0.446)	Total night minutes (0.519)	Active Membership (0.484)
5	Total charges (0.393)	Total intl minutes (0.481)	Estimated salary (0.380)

The SHAP results reveal clear domain-specific patterns. In the IBM Telco dataset, tenure and contract-related variables are the dominant drivers of churn, indicating that customer relationship length and contractual commitment strongly shape churn risk. In the Orange Telecom dataset, usage intensity and service-related variables, particularly total day minutes and customer service calls, are the most influential predictors. In the Banking while the horizontal spread of SHAP values shows whether higher or lower feature values increase or decrease churn risk. These explanations strengthen the model's managerial Churn dataset, churn is mainly explained by demographic and product-relationship variables, especially age, number of products, balance, and active membership status.

The combined SHAP summary plots in Fig. 3 provide a more detailed view of both feature importance and direction of effect. For each dataset, features at the top of the plot have the greatest influence on the XGBoost prediction,

usefulness by showing not only which customers are likely to churn but also which customer characteristics drive the prediction.


(a) IBM Telco
IBM Telco: SHAP Summary - XGBoost
tenure
Contract_Two_year
MonthlyCharges
Contract_One_year
TotalCharges
InternetService_Fiber_optic
OnlineSecurity_Yes
PaymentMethod_Electronic_check
StreamingTV_Yes
PaperlessBilling_Yes
TechSupport_Yes
StreamingMovies_Yes
OnlineBackup_Yes
MultipleLines_Yes
PaymentMethod_Mailed_check
Dependents_Yes
gender_Male
InternetService_No
SeniorCitizen
Partner_Yes
−4
−2
0
2
SHAP value (impact on model output)
High
Feature value
Low
(b) Orange Telecom
Orange Telecom: SHAP Summary - XGBoost
Total_day_minutes
Customer_service_calls
Total_eve_minutes
Total_night_minutes
Total_intl_minutes
International_plan_Yes
Number_vmail_messages
Total_intl_calls
Account_length
Total_eve_calls
Total_night_calls
Total_day_calls
Total_night_charge
Area_code
Total_eve_charge
State_TX
State_AL
State_CA
State_WV
State_LA
−4
−2
0
2
4
6
8
10
SHAP value (impact on model output)
High
Feature value
Low
(c) Banking Churn
Banking Churn: SHAP Summary - XGBoost
Age
NumOfProducts
Balance
IsActiveMember
EstimatedSalary
CreditScore
Gender_Male
Tenure
Geography_Germany
Geography_Spain
HasCrCard
−2
0
2
4
6
SHAP value (impact on model output)
High
Feature value
Low

6.3. Risk segmentation validation

Customers were segmented into low-, medium-, and high-risk groups based on predicted churn probabilities. The validity of this segmentation was verified by comparing actual churn rates across the predicted risk groups.

Table 4. Actual Churn Rate by Predicted Risk Group

Risk Group	IBM Telco	Orange Telecom	Banking Churn
Low risk	0.045	0.014	0.040
Medium risk	0.207	0.018	0.101
High risk	0.545	0.404	0.469

As shown in Figure 4, there is a strong monotonic increase in actual churn rates from the low-risk to the high-risk segment in all datasets. This pattern validates the practical usefulness of probability-based risk segmentation. Managers can therefore prioritize retention resources toward the high-risk group, where the potential impact of interventions is highest, rather than applying uniform strategies across the entire customer base. The successful validation of risk groups also strengthens confidence in using the predicted probabilities for downstream counterfactual retention strategy simulation.

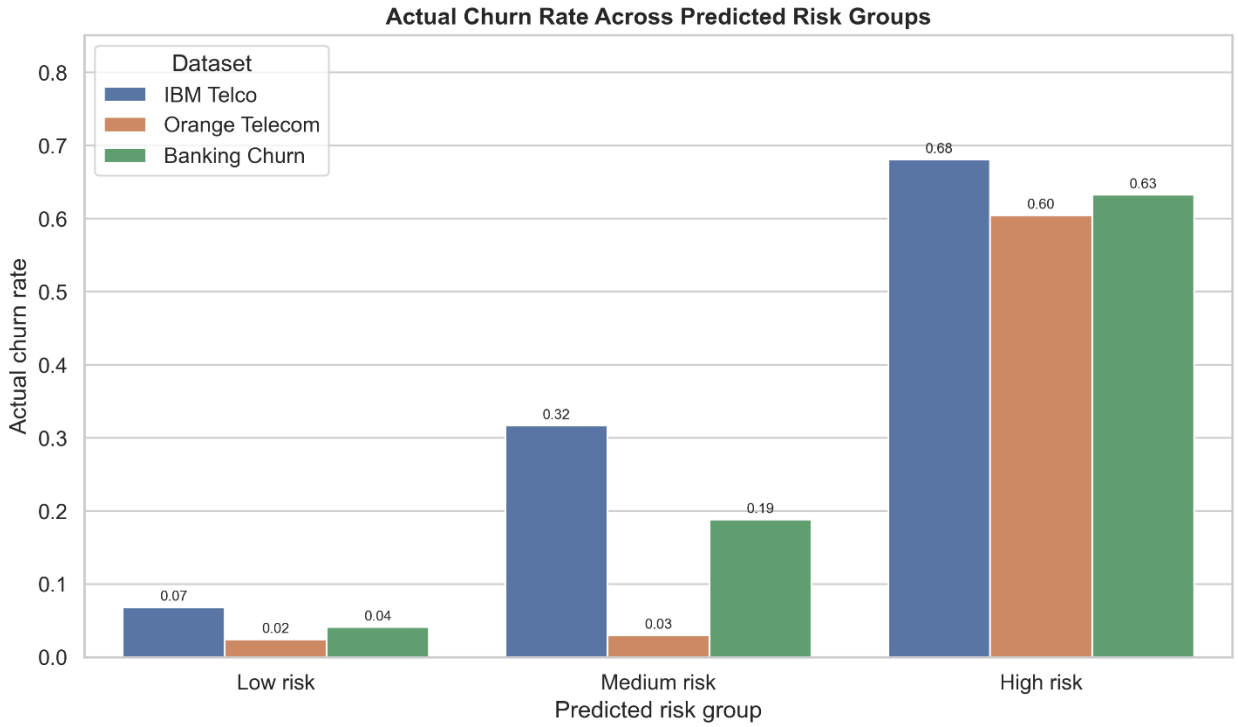


Figure 4. Actual Churn Rate Across Predicted Risk Groups.

As shown in Figure 4, there is a strong monotonic increase in actual churn rates from the low-risk to the high-risk segment in all datasets. This pattern validates the practical usefulness of probability-based risk segmentation. Managers can therefore prioritize retention resources toward the high-risk group, where the potential impact of interventions is highest, rather than applying uniform strategies across the entire customer base. The successful validation of risk groups also strengthens confidence in using the calibrated probabilities for downstream counterfactual retention strategy simulation.

6.4. Counterfactual retention strategy simulation

Counterfactual retention strategy simulations were performed to estimate the impact of targeted managerial interventions on predicted churn risk.

Table 5. Counterfactual Retention Strategy Simulation Results (Selected Strategies)

Dataset	Strategy	Target Customers	Risk Reduction (pp)	Expected Retained
IBM Telco	Upgrade month-to-month to two-year contract	773	51.76	400

Dataset	Strategy	Target Customers	Risk Reduction (pp)	Expected Retained
IBM Telco	Upgrade month-to-month to one-year contract	773	29.93	231
Banking Churn	Simplify product portfolio (>2 products)	69	66.19	46
Banking Churn	Cross-sell second product	1,013	17.71	179
Orange Telecom	Replace international plan	65	21.77	14
Orange Telecom	Improve service recovery (3+ calls)	145	18.54	27

The simulations demonstrate substantial potential impact. The strongest predicted reduction is observed in the Banking Churn dataset, where simplifying the product portfolio for customers with more than two products yields a 66.19 percentage-point reduction in churn risk. In the IBM Telco dataset, upgrading month-to-month customers to a two-year contract produces a substantial reduction of 51.76 percentage points. In Orange Telecom, replacing the international plan and improving service recovery for customers with multiple service calls also show meaningful predicted improvements. To provide a clearer visual overview, Figure 5 illustrates the predicted churn-risk reduction for the evaluated retention strategies across the three datasets.

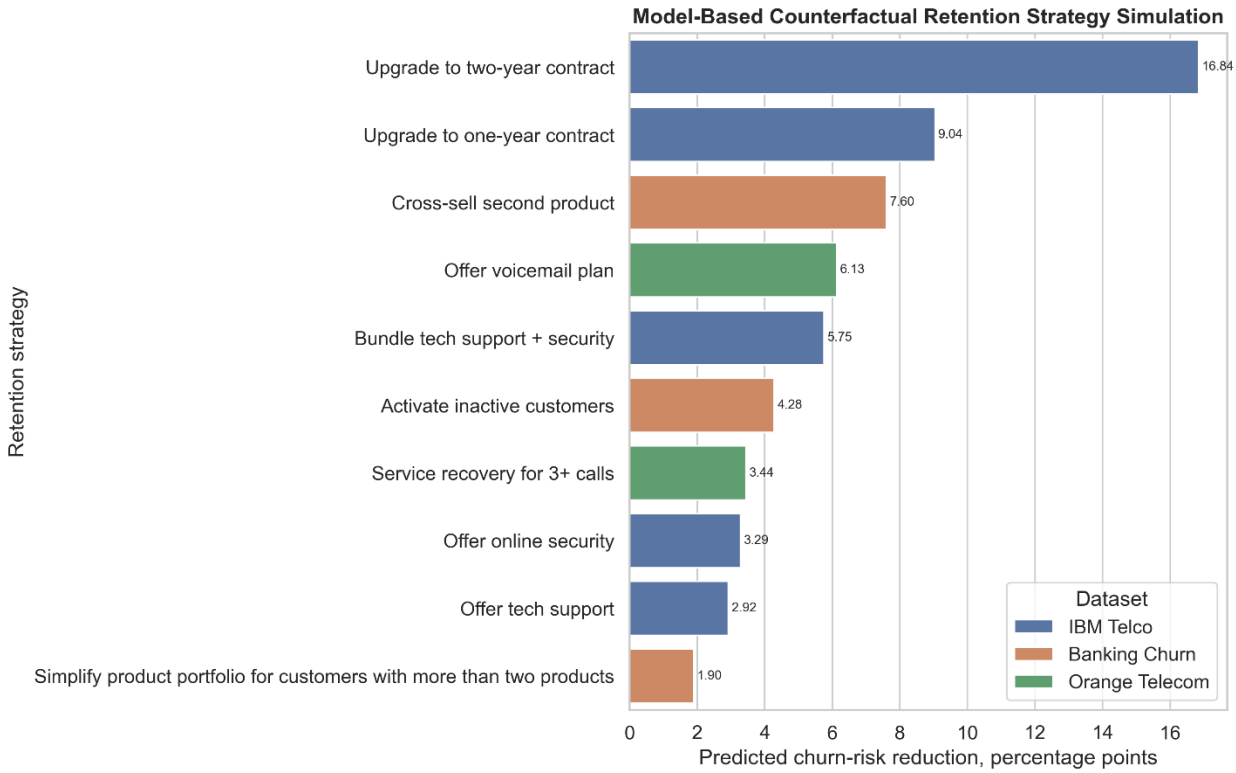


Figure 5. Model-Based Counterfactual Retention Strategy Simulation Across Datasets.

As shown in Figure 5, targeted interventions can produce considerable reductions in predicted churn risk, particularly for high-value or high-risk customer segments. These findings highlight the practical value of the proposed framework by translating model predictions and SHAP insights into quantifiable, actionable retention strategies. However, the results should be interpreted as model-based counterfactual estimates rather than guaranteed causal outcomes, as they are derived from observational data.

6.5. Comparison with recent literature

Table 6 compares the proposed study with recent studies published in Expert Systems with Applications. The comparison focuses on research objective, methodological orientation, empirical scope, and the extent to which each study connects prediction, explainability, segmentation, and managerial action. The purpose of this comparison is not only to benchmark predictive performance, but also to clarify how the present framework extends prior work by integrating multiple decision-support components into a single churn management pipeline.

Table 6. Comparison of the proposed framework with recent studies.

Study	Main Research Focus	Methodological Approach	Domain / Empirical Scope	Main Contribution	Difference from the Present Study
Zhu et al. (2023).	Churn prediction under class imbalance	Bagging-based selective ensemble model	Churn prediction datasets	Proposed a selective ensemble strategy for improving churn prediction on imbalanced data	Focuses mainly on ensemble performance; does not integrate SHAP, risk segmentation, and counterfactual retention simulation
Jiang et al. (2023).	Profit-driven churn prediction	Hybrid model based on cost minimization and return maximization	Customer churn context	Connects churn prediction with profit-oriented decision logic	Emphasizes profit-driven prediction but provides limited integration of explainable AI and validated risk grouping
Balbino et al. (2023).	Explainable AI and counterfactual explanations	Model-agnostic counterfactual, selected, and social explanations	Classification models	Provides actionable explanation logic for classification decisions	Focuses on explanation methodology generally rather than churn-specific multi-dataset retention strategy simulation

Study	Main Research Focus	Methodological Approach	Domain / Empirical Scope	Main Contribution	Difference from the Present Study
Rungruang et al. (2024).	Customer segmentation	RFM model with hierarchical formal concept analysis	Customer relationship management	Develops a hierarchical segmentation approach for customer grouping	Focuses on segmentation rather than linking churn probability, SHAP explanations, and retention simulations
Barrera et al. (2024).	Customer segmentation and decision support	Multiple-criteria sorting outranking method	Business-to-business customer segmentation	Provides a structured decision-support model for segmentation	Does not focus on churn prediction, SHAP interpretation, or counterfactual retention scenarios
Haddadi et al. (2024).	Churn prediction in imbalanced datasets	Comparative study of resampling methods and classifiers	Multiple churn datasets	Shows the importance of imbalance handling in churn prediction	Primarily evaluates resampling and classification performance; does not extend to explainability and managerial intervention simulation
Rao et al. (2024).	Imbalanced churn classification	Multi-strategy collaborative processing method	Customer churn classification	Proposes an improved method for imbalanced	Concentrates on classification improvement rather than explainable and

Study	Main Research Focus	Methodological Approach	Domain / Empirical Scope	Main Contribution	Difference from the Present Study
				churn classification	action-oriented churn management
Wang et al. (2024).	Telco churn risk assessment	FCLCNN-LSTM deep learning model	Telecommunications	Develops a deep learning model for telco churn risk assessment	Strong domain-specific predictive model, but limited cross-industry validation and managerial counterfactual simulation
Abusitta et al. (2024).	Explainable AI	Survey of XAI techniques, challenges, and open issues	General AI applications	Highlights the need for transparency, trust, and accountability in intelligent systems	Provides conceptual and methodological XAI guidance but does not develop a churn-specific empirical framework
Haridasan et al. (2025).	Telecom churn and service recommendation	Gazelle optimization and conditional variational autoencoder	Telecommunications	Links churn analysis with telecom service recommendation	Focuses on telecom recommendation and optimization; does not test a cross-dataset churn

Study	Main Research Focus	Methodological Approach	Domain / Empirical Scope	Main Contribution	Difference from the Present Study
					management framework
Panimalar et al. (2025).	Customer churn prediction	Weighted multilayer perceptron and enhanced multipath backpropagation	Telecommunications	Develops an intensified deep learning approach for churn prediction	Emphasizes predictive modeling; limited focus on SHAP-based explanation, risk validation, and intervention simulation
Zaghloul et al. (2025).	Customer retention through churn prediction	Hybrid RFM, K-means, and deep neural network framework	Online retail with validation datasets	Integrates segmentation and deep learning for churn prediction and retention	Closest to the present study, but focused on online retail and RFM/deep learning; the present study emphasizes SHAP, probability-based risk segmentation, counterfactual simulation, and cross-domain telecom–banking validation.

Study	Main Research Focus	Methodological Approach	Domain / Empirical Scope	Main Contribution	Difference from the Present Study
Present study	Explainable multi-dataset churn management	Logistic Regression, Random Forest, XGBoost, SHAP, probability-based risk segmentation, and counterfactual retention simulation	IBM Telco, Orange Telecom, and Banking Churn	Integrates prediction, explainability, risk validation, and actionable retention simulation across telecom and banking datasets	Provides a unified decision-support pipeline that moves beyond model accuracy toward interpretable and actionable churn management

The comparison shows that recent studies have made substantial progress in customer churn prediction, imbalanced classification, segmentation, explainable AI, and retention-focused analytics. However, most prior work emphasizes a single component of the churn management problem. Some studies focus mainly on improving predictive performance under class imbalance, while others emphasize segmentation, deep learning, or general explainability. In contrast, the present study integrates these components into a single framework. Specifically, the proposed approach combines cross-dataset model evaluation, SHAP-based explainability, probability-based risk segmentation, and counterfactual retention strategy simulation.

This integration is important because churn management is not only a classification problem. A model that predicts churn accurately may still be difficult to use if managers cannot understand the reasons behind the prediction or translate predicted probabilities into concrete actions. The proposed framework addresses this limitation by demonstrating how predictive outputs can be converted into explainable churn drivers, validated risk groups, and estimated effects of retention strategies. Therefore, the main contribution of this study lies not only in predictive performance, but also in linking machine learning outputs with business-intelligence decision support.

7. Discussion

The findings of this study provide several important insights for explainable customer churn prediction and retention management in service operations. First, the results confirm that model performance varies across datasets and evaluation metrics. Logistic Regression, Random Forest, and extreme gradient boosting

each exhibit distinct strengths depending on the data structure, churn prevalence, feature composition, and class imbalance. This finding reinforces the importance of evaluating churn prediction models using multiple metrics rather than relying only on accuracy. In imbalanced churn classification tasks, F1-score, area under the receiver operating characteristic curve, and area under the precision-recall curve provide a more informative assessment of model performance because they account for the difficulty of identifying minority-class churners (Haddadi et al., 2024; Rao et al., 2024; Saito & Rehmsmeier, 2015; Zhu et al., 2023).

Second, Shapley Additive Explanations analysis shows that the drivers of churn are strongly domain-specific. In the IBM Telco dataset, tenure and contract-related variables are the strongest drivers of churn risk, suggesting that customer relationship duration and contractual commitment are central to telecom retention. In the Orange Telecom dataset, usage intensity and customer service calls are more influential, indicating that service consumption patterns and customer interaction problems are important churn signals. In the Banking Churn dataset, age, number of products, balance, and active membership status are the most important factors, indicating that banking churn is shaped by customer profile and relationship depth. These findings highlight the value of context-aware explainability in churn analytics, as model interpretation must reflect the specific engineering and operational environment in which the prediction is applied (Abusitta et al., 2024; Lundberg & Lee, 2017; Wang et al., 2024).

Third, the risk segmentation results demonstrate that predicted churn probabilities can be used to define practical customer prioritization groups. Across all three datasets, actual churn rates increase substantially from low-risk to high-risk groups. This indicates that the model-generated probabilities are not only statistically useful but also operationally meaningful for resource allocation in service systems (Barrera et al., 2024; Rungruang et al., 2024; Zaghloul et al., 2025).

Fourth, the counterfactual retention simulations show how churn analytics can move from explanation to actionable engineering decisions. By modifying selected customer attributes to represent possible managerial interventions, the framework estimates how predicted churn risk may change under different scenarios. Contract upgrades, service recovery, support bundles, product cross-selling, and product portfolio simplification were evaluated as possible retention strategies. These simulations, while based on observational data and not causal proofs, provide useful model-based decision-support estimates that can help engineers and managers identify promising retention actions before conducting field experiments, pilot programs, or A/B tests (Ascarza, 2018; Balbino et al., 2023; Jiang et al., 2023).

7.1. Theoretical Implications

This study contributes to the literature on customer churn prediction and explainable decision support in several ways. First, it extends churn prediction research by moving beyond single-model or single-dataset evaluation. By using three publicly available datasets from telecommunications and banking, the present study demonstrates how churn prediction frameworks can be evaluated across different customer relationship structures and industry settings (Ahn et al., 2020; Haddadi et al., 2024; Zaghloul et al., 2025).

Second, the study contributes to explainable artificial intelligence research by embedding Shapley Additive Explanations analysis within a broader churn management framework. Instead of using explanations only as a post-hoc feature-ranking tool, this study connects them to risk segmentation and counterfactual retention simulation. This provides a more complete view of how explainability can support operational decision-making by linking prediction, interpretation, prioritization, and action (Abusitta et al., 2024; Balbino et al., 2023; Lundberg & Lee, 2017).

Third, the study contributes to customer segmentation research by validating predicted risk groups against actual churn behavior. This strengthens the connection between predictive modeling and practical segmentation for service operations (Barrera et al., 2024; Rungruang et al., 2024; Wu et al., 2020).

Fourth, the study contributes to intelligent decision-support systems research by integrating prediction, interpretation, prioritization, and intervention simulation. This supports a shift from purely predictive analytics toward prescriptive decision support in engineering applications (Jiang et al., 2023; Zaghloul et al., 2025).

7.2. Managerial and Engineering Implications

The findings offer several practical implications for managers and engineers in telecommunications, banking, and other service-oriented industries. First, model selection should consider multiple performance metrics rather than accuracy alone, especially in imbalanced classification settings common to churn problems. Metrics such as recall, F1-score, AUC-ROC, and AUC-PR provide a more balanced view of model performance, particularly when the business objective is to identify at-risk customers before they leave (Haddadi et al., 2024; Rao et al., 2024; Saito & Rehmsmeier, 2015).

Second, churn drivers differ significantly across domains. Telecom engineers should focus on contract type, tenure, billing characteristics, service usage, and customer service interactions. Banking operations teams should prioritize age, product ownership, account balance, and active membership. Retention strategies must therefore be tailored to domain-specific drivers rather than generic policies (Abusitta et al., 2024; Lundberg & Lee, 2017; Wang et al., 2024).

Third, probability-based risk segmentation enables efficient resource allocation in service operations. Low-risk customers can receive routine engagement, medium-risk customers soft interventions, and high-risk customers targeted actions. This approach helps avoid wasting limited operational resources on low-impact customers (Ascarza, 2018; Barrera et al., 2024).

Fourth, counterfactual simulations support evidence-based retention planning by estimating the potential impact of operational interventions such as contract upgrades or service bundles. These simulations serve as a low-cost screening tool before expensive real-world experiments (Balbino et al., 2023; Jiang et al., 2023; Zaghloul et al., 2025).

Finally, the framework improves collaboration between data science teams and operations engineers by providing interpretable outputs and actionable recommendations, thereby increasing trust and adoption of artificial intelligence systems in service operations (Abusitta et al., 2024; Balbino et al., 2023).

7.3. Practical Deployment Implications

For practical deployment, the proposed framework should be implemented as an intelligent decision-support system rather than a one-time prediction exercise. A practical deployment pipeline includes four stages: (1) regular updating and preprocessing of customer data, (2) generation of churn probabilities using the trained model, (3) computation of Shapley Additive Explanations for high-risk customers, and (4) counterfactual simulation to evaluate potential retention actions (Barrera et al., 2024; Jiang et al., 2023; Zaghloul et al., 2025).

However, counterfactual results based on observational data should be treated as decision-support estimates rather than causal guarantees. Selected interventions must be validated through A/B testing or pilot programs before large-scale rollout. Continuous monitoring of model performance, probability stability, and business outcomes is essential to handle concept drift in dynamic service environments (Ascarza, 2018; Haddadi et al., 2024; Rao et al., 2024).

7.4. Engineering and Systems Implications

From a systems engineering perspective, the proposed framework supports intelligent automation and decision-support in service operations, a key area in Industry 4.0 and cyber-physical customer management systems. It enables proactive resource allocation, better capacity planning (e.g., call center staffing and network resources), and performance monitoring of retention initiatives. By integrating explainable artificial intelligence with operational workflows, the framework helps transform raw customer data into actionable engineering insights, ultimately improving system reliability, profitability, and customer experience in service-oriented organizations.

The framework's modular design also facilitates integration with existing enterprise systems, including customer relationship management platforms, billing systems, and service operation dashboards. This interoperability is essential for real-world deployment in engineering environments where legacy systems and data silos often present implementation challenges.

8. Conclusion, Limitations, and Future Research

This study proposed an explainable multi-dataset churn management framework that integrates machine learning prediction, Shapley Additive Explanations-based interpretability, probability-based risk segmentation, and counterfactual retention-strategy simulation. The framework was evaluated using three publicly available datasets from telecommunications and banking domains: IBM Telco (7,043 records), Orange Telecom (3,333 records), and Banking Churn (10,000 records). Logistic Regression, Random Forest, and extreme gradient boosting were compared using accuracy, precision, recall, F1-score, AUC-ROC, and AUC-PR.

The results showed that tree-based ensemble models generally perform strongly on structured churn data, with extreme gradient boosting achieving the highest F1-score (0.849) and AUC-ROC (0.946) on the Orange Telecom dataset. In contrast, Random Forest performed competitively on the Banking Churn dataset (AUC-ROC = 0.851). Shapley Additive Explanations revealed domain-specific churn drivers: tenure and contract type in IBM Telco, usage intensity and customer service calls in Orange Telecom, and age, product ownership, and active membership in banking. Risk segmentation successfully translated model outputs into operationally meaningful customer

groups, with actual churn rates increasing from 4.5% (low-risk) to 54.5% (high-risk) in IBM Telco, and from 4.0% to 46.9% in the banking dataset. Counterfactual simulations estimated potential churn risk reduction under realistic interventions, with reductions reaching up to 66.19 percentage points in the banking dataset.

The main contribution of this study is the development of an integrated, explainable, and action-oriented churn management framework that bridges predictive modeling, interpretability, and prescriptive decision support for service operations engineering. The framework demonstrates practical value through: (1) cross-dataset and cross-industry validation using public data, (2) identification of domain-specific churn drivers, (3) validated risk segmentation for resource allocation, and (4) counterfactual simulation for intervention planning.

Limitations. The study has several limitations. First, the analysis is based on publicly available datasets, which may not fully capture the complexity of real organizational customer databases with richer behavioral histories and temporal dynamics. Second, the counterfactual simulations are based on model-estimated changes in predicted churn probability and should not be interpreted as causal treatment effects. Third, the study uses structured tabular datasets and does not incorporate unstructured data such as customer complaints, call transcripts, social media posts, or customer reviews. Fourth, although the study uses three datasets, further validation is needed across additional industries and real-time business environments.

Future Research. Future research can extend this study in several directions. First, validation using private organizational datasets with richer customer histories and actual intervention outcomes would strengthen the framework's practical applicability. Second, experimental or quasi-experimental designs could test whether simulated retention strategies produce real reductions in churn. Third, incorporating unstructured data through natural language processing could capture customer sentiment, complaint patterns, and signals of service dissatisfaction. Fourth, comparing additional explainable AI methods, such as LIME, counterfactual explanation algorithms, partial dependence plots, and accumulated local effects, would further enhance interpretability. Fifth, developing real-time churn-management dashboards that integrate model predictions, SHAP explanations, risk segmentation, and recommended retention actions would support operational decision-making. Finally, extending the framework to other engineering domains such as predictive maintenance, supply chain risk management, or equipment failure prediction would demonstrate broader applicability of the proposed approach.

Overall, the proposed framework provides a practical, explainable, and prescriptive approach to churn management in service operations. By integrating machine learning, explainable AI, risk segmentation, and counterfactual simulation, the study offers a pathway to transform churn prediction from a black-box classification task into an interpretable, actionable intelligent decision-support system for engineering applications.

Funding Declaration

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Abusitta, A., Li, M. Q., & Fung, B. C. M. (2024). Survey on explainable AI: Techniques, challenges, and open issues. Expert Systems with Applications, 255, 124710. https://doi.org/10.1016/j.eswa.2024.124710

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A survey on churn analysis in various business domains. IEEE Access, 8, 220816–220839. https://doi.org/10.1109/ACCESS.2020.3042657

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using the K-means algorithm. Journal of King Saud University - Computer and Information Sciences, 34(5), 1785–1792. https://doi.org/10.1016/j.jksuci.2019.12.011

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. Journal of Marketing Research, 55(1), 80–98. https://doi.org/10.1509/jmr.16.0163

Baghla, S., & Gupta, G. (2022). Performance evaluation of various classification techniques for customer churn prediction in e-commerce. Microprocessors and Microsystems, 94, 104680. https://doi.org/10.1016/j.micpro.2022.104680

Balbino, M. de S., Zárate, L. E., & Nobre, C. N. (2023). CSSE: An agnostic method of counterfactual, selected, and social explanations for classification models. Expert Systems with Applications, 228, 120373. https://doi.org/10.1016/j.eswa.2023.120373

Barrera, F., Segura, M., & Maroto, C. (2024). Multiple criteria decision support system for customer segmentation using a sorting outranking method. Expert Systems with Applications, 238, 122310. https://doi.org/10.1016/j.eswa.2023.122310

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2021). A review of data mining methods in RFM-based customer segmentation. Journal of Physics: Conference Series, 1869(1), 012085. https://doi.org/10.1088/1742-6596/1869/1/012085

Haddadi, S. J., Farshidvard, A., Silva, F. dos S., dos Reis, J. C., & Reis, M. S. (2024). Customer churn prediction in imbalanced datasets with resampling methods: A comparative study. Expert Systems with Applications, 246, 123086. https://doi.org/10.1016/j.eswa.2023.123086

Haridasan, V., Hariharanath, K., & Muthukumaran, K. (2025). Gazelle optimization and conditional variational autoencoder for telecom user service recommendation based on churn analysis. Expert Systems with Applications, 267, 126199. https://doi.org/10.1016/j.eswa.2024.126199

Januzaj, Y., Beqiri, E., & Luma, A. (2023). Determining the optimal number of clusters using silhouette score as a data mining technique. International Journal of Online and

Biomedical Engineering, 19(4), 174–182. https://doi.org/10.3991/ijoe.v19i04.37059

Jiang, P., Liu, Z., Zhang, L., & Wang, J. (2023). Hybrid model for profit-driven churn prediction based on cost minimization and return maximization. Expert Systems with Applications, 228, 120354. https://doi.org/10.1016/j.eswa.2023.120354

Kumar, A., & Singh, P. (2024). Explaining customer churn prediction in the telecom industry using tabular machine learning models. Engineering Applications of Artificial Intelligence, 133, 109512. https://doi.org/10.1016/j.engappai.2024.109512

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: A machine learning approach. Computing, 104(2), 271–294. https://doi.org/10.1007/s00607-021-00908-y

Li, H., & Chen, Y. (2024). GA optimization-based BRB AI reasoning algorithm for determining the factors affecting customer churn for operators. Engineering Applications of Artificial Intelligence, 133, 109562. https://doi.org/10.1016/j.engappai.2024.109562

Liu, Y., & Zhang, L. (2024). Enhancing e-commerce customer churn management with a profit- and AUC-focused prescriptive analytics approach. Engineering Applications of Artificial Intelligence, 133, 109318. https://doi.org/10.1016/j.engappai.2024.109318

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1705.07874

Panimalar, S. A., & Krishnakumar, A. (2023). Customer churn prediction model in cloud environment using DFE-WUNB: ANN deep feature extraction with Weight Updated Tuned Naïve Bayes classification with Block-Jacobi SVD dimensionality reduction. Engineering Applications of Artificial Intelligence, 123, 107015. https://doi.org/10.1016/j.engappai.2023.107015

Panimalar, S. A., Krishnakumar, A., & Senthil Kumar, S. (2025). Intensified customer churn prediction: Connectivity with weighted multi-layer perceptron and enhanced multipath back propagation. Engineering Applications of Artificial Intelligence, 265, 125993. https://doi.org/10.1016/j.engappai.2024.125993

Pasquadibisceglie, V., Appice, A., Ieva, G., & Malerba, D. (2023). TSUNAMI – An explainable PPM approach for customer churn prediction in evolving retail data environments. Engineering Applications of Artificial Intelligence, 120, 105904. https://doi.org/10.1016/j.engappai.2023.105904

Rajput, L., & Singh, S. N. (2023). Customer segmentation of e-commerce data using K-means clustering algorithm. In 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 658–664). https://doi.org/10.1109/Confluence56041.2023.10048834

Rao, C., Xu, Y., Xiao, X., Hu, F., & Goh, M. (2024). Imbalanced customer churn classification using a new multi-strategy collaborative processing method. Engineering Applications of Artificial Intelligence, 133,

109703. https://doi.org/10.1016/j.engappai.2024.109703

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. Expert Systems with Applications, 237, 121449. https://doi.org/10.1016/j.eswa.2023.121449

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Shirole, R., Salokhe, L., & Jadhav, S. (2021). Customer segmentation using RFM model and K-means clustering. International Journal of Scientific Research in Science and Technology, 591–597. https://doi.org/10.32628/IJSRST2183118

Wang, C., Rao, C., Hu, F., Xiao, X., & Goh, M. (2024). Risk assessment of customer churn in telco using FCLCNN-LSTM model. Expert Systems with Applications, 248, 123352. https://doi.org/10.1016/j.eswa.2024.123352

Wang, F., Franco-Penya, H.-H., Kelleher, J. D., Pugh, J., & Ross, R. (2017). An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. In Machine Learning and Data Mining in Pattern Recognition (pp. 291–305). https://doi.org/10.1007/978-3-319-62416-7_21

Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., & Xu, G. (2020). An empirical study on customer segmentation by purchase behaviors using an RFM model and K-means algorithm. Mathematical Problems in Engineering, 2020, 8884227. https://doi.org/10.1155/2020/8884227

Xiahou, X., & Harada, Y. (2022). B2C e-commerce customer churn prediction based on K-means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 458–475. https://doi.org/10.3390/jtaer17020024

Zaghloul, M., Barakat, S. I., & Rezk, A. (2025). Enhancing customer retention in online retail through churn prediction: A hybrid RFM, K-means, and deep neural network approach. Expert Systems with Applications, 290, 128465. https://doi.org/10.1016/j.eswa.2025.128465

Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. International Journal of Distributed Sensor Networks, 18(6), 155013292211069. https://doi.org/10.1177/15501329221106935

Zhang, Y., & Wang, X. (2025). Decoding industrial-scale battery manufacturing process through integration of causal graphs into explainable artificial intelligence. Engineering Applications of Artificial Intelligence, 159, 109428. https://doi.org/10.1016/j.engappai.2025.109428

Zhu, B., Qian, C., vanden Broucke, S., Xiao, J., & Li, Y. (2023). A bagging-based selective ensemble model for churn prediction on imbalanced data. Expert Systems with Applications, 227, 120223. https://doi.org/10.1016/j.eswa.2023.120223

Badole, S. (n.d.). Banking Customer Churn Prediction Dataset. Kaggle. Retrieved June 8, 2026.

Mnassri, B. (n.d.). Telecom Churn Dataset. Kaggle. Retrieved June 8, 2026.

Aqeel, M. (n.d.). WA_Fn-UseC_-Telco-Customer-Churn.csv. Kaggle. Retrieved June 8, 2026.