International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Evaluating Explainability Tradeoffs in Machine Learning-Based Retention Intelligence Systems

Wael Breich

Senior Data Analyst, Pinnacle Group, DIRECTV

Abstract

Machine learning systems are increasingly used in customer churn analytics, yet many high-performing predictive models suffer from limited interpretability. This study investigates explainability trade-offs in churn prediction using a telecommunications dataset containing approximately 3,150 customer records and a combination of behavioral, demographic, billing, and subscription-related attributes. Logistic regression, random forest, and XGBoost models were comparatively evaluated to analyze the relationship between predictive performance and interpretability.

To improve transparency within the nonlinear ensemble framework, SHapley Additive exPlanations (SHAP) were applied to the XGBoost model to generate both global and local behavioral explanations. The results show that XGBoost achieved the strongest predictive performance, while SHAP substantially improved interpretability by identifying the customer attributes most strongly associated with churn predictions. Engagement and subscription-related variables emerged as the most influential behavioral drivers.

The findings demonstrate that explainable artificial intelligence can improve the operational transparency and usability of high-performing churn prediction systems without substantially sacrificing predictive capability.

Keywords: Explainable Artificial Intelligence; SHAP; Machine Learning Interpretability; Customer Churn Analytics; XGBoost; Behavioral Analytics.

1. Introduction

The growing integration of machine learning into operational decision-making systems has created increasing demand for predictive models that are not only accurate but also interpretable. Across industries such as telecommunications, healthcare, finance, insurance, and subscription-based digital services, organizations increasingly rely on predictive analytics systems to support strategic and operational decisions involving risk assessment, resource allocation, customer retention, and intervention planning. However, as machine learning systems become more complex, explainability and transparency challenges have emerged as major barriers to organizational trust and operational deployment.

This issue is particularly important in customer churn analytics. Customer churn prediction represents one of the most commercially important applications of machine learning because of its direct relationship with customer lifetime value, recurring revenue stability, acquisition efficiency, and long-term growth. Organizations increasingly deploy predictive systems to identify subscribers at risk of cancellation and initiate proactive retention interventions before disengagement becomes irreversible [1], [2].

Recent advances in ensemble learning methods such as random forests, gradient boosting, and XGBoost have significantly improved churn prediction accuracy [3], [4]. Prior telecommunications research by Verbeke et al. [5] similarly demonstrated that ensemble and hybrid learning frameworks frequently outperform traditional statistical techniques in churn analytics environments. These algorithms are particularly effective in large behavioral datasets because they capture nonlinear relationships and interaction effects that are often missed by traditional statistical models. However, despite their predictive advantages, these ensemble systems frequently function as black box models whose internal decision processes remain difficult to interpret.

In many real-world business environments, predictive performance alone is insufficient. Stakeholders, retention analysts, and operational managers often require interpretable explanations capable of clarifying why a model predicts churn for a specific subscriber [6], [7]. Without transparency, organizations may hesitate to operationalize predictive systems, particularly when retention actions involve financial costs, customer experience implications, or strategic resource allocation.

This tension between predictive performance and interpretability has become a central concern within explainable artificial intelligence (XAI) research [8], [9]. Transparent models such as logistic regression offer naturally interpretable outputs through coefficients and additive relationships but frequently underperform when applied to nonlinear behavioral environments. Conversely, high-performing ensemble methods typically sacrifice interpretability in exchange for predictive flexibility.

The present study investigates explainability tradeoffs in machine learning-based churn analytics using a telecommunications dataset containing behavioral, billing, demographic, and subscription-related attributes.

Importantly, this study evaluates explainability tradeoffs rather than proposing a novel predictive architecture. The primary contribution lies in the interpretability analysis and comparative explainability framework rather than the development of a new machine learning algorithm. Unlike prior churn prediction studies focused primarily on predictive capability, this research positions churn prediction as an application domain for evaluating explainable machine learning systems.

This study extends prior engagement-centric churn research conducted by the author by shifting the analytical focus from predictive performance toward explainability, operational transparency, and interpretability tradeoffs in machine learning systems [10]. Earlier work demonstrated that engagement behavior alone provides substantial predictive power for churn prediction independent of demographic and contractual variables. The present study instead focuses on evaluating how different model classes represent behavioral relationships and how SHAP-based explanations improve interpretability in nonlinear ensemble learning environments.

Specifically, the study compares logistic regression, random forest, and XGBoost models while evaluating explainability across multiple dimensions. SHAP explanations are applied to the XGBoost framework to generate global feature attribution analysis, nonlinear behavioral interpretation, and local prediction explanations.

The contributions of this study are fourfold. First, the study evaluates interpretability tradeoffs between transparent linear models and nonlinear ensemble learning systems. Second, the study demonstrates how SHAP-based explanations restore operational transparency to black-box churn models. Third, the study identifies the behavioral mechanisms most strongly associated with churn predictions through explainable feature attribution analysis. Finally, the study contributes to the growing literature on interpretable

business analytics systems by presenting a reproducible explainability framework for customer retention environments.

The remainder of this paper is organized as follows. Section 2 reviews the literature on churn analytics, explainable artificial intelligence, and interpretable machine learning systems. Section 3 presents the explainability-oriented methodological framework. Section 4 presents the comparative explainability analysis and empirical findings. Section 5 discusses operational implications for explainable retention systems. Section 6 highlights limitations and future research directions. Section 7 concludes the study.

2. Literature Review

2.1 Machine Learning in Customer Churn Analytics

Customer churn analytics has become one of the most active application domains in predictive analytics because of its importance across subscription-based industries. Early churn prediction systems relied primarily on statistical methods such as logistic regression, survival analysis, and decision trees because of their simplicity, interpretability, and ease of operational implementation. These models frequently relied on demographic variables, billing records, contractual attributes, and aggregate usage measures.

As behavioral data availability increased, contemporary churn analytics shifted toward machine learning methods capable of modeling complex interaction patterns across large customer datasets [2]. Ensemble learning approaches such as random forests, gradient boosting, and XGBoost consistently demonstrated superior predictive performance compared to traditional statistical techniques [3], [4]. These improvements are especially pronounced in imbalanced churn environments characterized by heterogeneous user behavior and nonlinear engagement patterns.

Recent research increasingly emphasizes the importance of engagement-centric behavioral variables within churn analytics systems. Features related to interaction consistency, usage frequency, service reliability, complaint activity, and communication diversity frequently emerge as dominant churn predictors [7], [8]. Similar findings were reported by Amin et al. (2019), who demonstrated that behavioral usage patterns significantly outperform static customer descriptors in telecommunications churn environments. Behavioral engagement variables are particularly valuable because they provide continuously observable signals capable of capturing evolving customer dissatisfaction and disengagement [1], [11].

However, despite strong predictive performance improvements, the increasing complexity of machine learning systems introduces substantial transparency challenges. In many operational churn environments, retention teams require interpretable behavioral explanations rather than isolated churn probabilities. Consequently, explainability has become an increasingly important consideration within applied churn analytics research.

2.2 Explainable Artificial Intelligence

Explainable artificial intelligence emerged in response to growing concerns regarding transparency, trust, accountability, and interpretability in machine learning systems [7], [9]. Although modern machine learning algorithms frequently outperform traditional statistical models in predictive tasks, their internal decision structures are often difficult to interpret directly.

This challenge is particularly important in operational business systems where machine learning output influences real-world decisions involving customers, financial resources, healthcare interventions, or organizational strategy. In such environments, stakeholders increasingly require explanations capable of clarifying how predictions are generated and which variables contribute most strongly to model outputs.

The explainable AI literature generally distinguishes between inherently interpretable models and post-hoc explainability approaches. Interpretable models such as logistic regression provide transparent outputs through additive coefficients and linear relationships. However, these models may oversimplify complex nonlinear environments.

Conversely, post-hoc explainability approaches seek to interpret complex black-box systems after model training. These methods attempt to preserve predictive performance while improving transparency through feature attribution analysis, local explanations, or surrogate modeling techniques.

Within operational analytics environments, explainability is increasingly viewed not merely as a technical preference but as a practical deployment requirement [12]. Recent work by Carvalho et al. (2019) similarly argues that interpretable machine learning systems are essential for organizational trust, accountability, and real-world AI adoption. Organizations often hesitate to operationalize predictive systems that cannot provide actionable and trustworthy explanations.

2.3 SHAP and Feature Attribution Frameworks

Among contemporary explainability approaches, SHapley Additive exPlanations (SHAP) has become one of the most widely adopted interpretability frameworks because of its strong theoretical foundation and compatibility with ensemble learning systems [13], [14].

SHAP is derived from cooperative game theory and quantifies the marginal contribution of individual variables to model predictions by evaluating prediction changes across all possible feature combinations. The framework provides both global interpretability and local prediction explanations.

One major advantage of SHAP is additive consistency [13]. The theoretical SHAP framework introduced by Lundberg and Lee (2017) has become particularly influential within interpretable machine learning research because of its mathematical consistency and compatibility with tree-based ensemble systems. Feature attributions sum directly to the final prediction output, improving interpretability and explanation stability. This property allows analysts to evaluate how multiple variables interact to influence predictive outcomes.

SHAP has become increasingly important within churn analytics because it enables analysts to identify nonlinear behavioral thresholds, feature interactions, and asymmetric churn drivers that are often hidden within black-box ensemble systems. More recent explainability research by Molnar [15] further emphasizes the growing role of model-agnostic interpretability techniques within operational machine learning workflows. Prior research demonstrates that SHAP explanations frequently reveal abrupt churn probability increases among low-engagement subscribers and identify complaint-related events as disproportionately influential behavioral accelerators.

Despite growing interest in explainable churn analytics, relatively few studies explicitly position churn prediction as an explainability evaluation problem rather than purely a predictive modeling problem. Most existing studies continue to emphasize predictive performance while treating explainability as a supplementary analytical layer.

The present study contributes to the literature by reframing churn analytics as an explainability-oriented machine learning problem focused on evaluating interpretability tradeoffs between transparent and black-box predictive systems.

3. Explainability-Oriented Methodology

3.1 Comparative Explainability Framework

Rather than treating churn prediction solely as a classification problem, the present study evaluates churn

analytics as an explainability challenge involving tradeoffs between predictive performance and interpretability. Three machine learning approaches were comparatively evaluated:

- Logistic Regression
- Random Forest
- XGBoost Gradient Boosting

These models were selected because they represent different levels of interpretability and predictive flexibility.

Logistic regression serves as the transparent baseline model. The approach provides naturally interpretable coefficients and additive feature relationships that are easy for stakeholders to understand operationally.

Random forest and XGBoost represent increasingly complex nonlinear ensemble learning systems capable of modeling interaction effects and threshold behavior within behavioral datasets. However, these systems sacrifice transparency because their internal decision structures are not inherently interpretable.

The comparative framework therefore evaluates not only predictive performance but also:

1. Model transparency
2. Feature attribution interpretability
3. Nonlinear relationship discovery
4. Local prediction explainability
5. Operational usability

This framework enables direct evaluation of explainability tradeoffs across different machine learning paradigms.

3.2 Dataset and Behavioral Feature Representation

The analysis uses a publicly available telecommunications churn dataset consisting of approximately 3,150 anonymized subscriber records. Each observation corresponds to a unique subscriber and includes a binary churn indicator.

The dataset contains multiple engagement-centric behavioral variables aggregated across a twelve-month observation period. Unlike many churn prediction studies that combine demographic, contractual, and pricing variables, the present analysis primarily emphasizes behavioral engagement features while also incorporating demographic, billing, and subscription-related attributes.

The engagement variables were selected because they provide operationally actionable behavioral signals while minimizing interpretability complications associated with heterogeneous feature sets.

The variables were grouped into four behavioral categories:

1. Engagement frequency
2. Engagement intensity
3. Engagement breadth
4. Service reliability indicators

Engagement frequency variables capture interaction consistency and usage regularity. Engagement intensity variables capture aggregate behavioral activity levels. Engagement breadth variables capture interaction diversity and behavioral embeddedness. Service reliability variables capture operational friction events such as complaints and service failures.

All usage-based variables were normalized by subscriber tenure to ensure comparability across subscribers with different exposure durations.

3.3 Interpretability Evaluation Pipeline

The explainability evaluation pipeline consisted of three major stages.

First, predictive performance was evaluated using Area Under the ROC Curve (AUC), F1 score, precision, and recall. These metrics were included not as the primary research objective but rather as contextual indicators for evaluating explainability tradeoffs.

Second, interpretable and black-box models were comparatively analyzed to evaluate how different modeling paradigms represent behavioral churn relationships.

Third, SHAP-based explanations were applied to the XGBoost framework to evaluate:

- Global feature importance
- Nonlinear behavioral effects
- Feature interaction behavior
- Local prediction explanations
- Operational interpretability

The explainability analysis therefore extends beyond traditional feature importance interpretation and instead evaluates how machine learning systems communicate behavioral reasoning.

3.4 SHAP-Based Explainability Analysis

SHAP explanations were selected because of their additive consistency, scalability for ensemble learning systems, and ability to provide both local and global interpretability.

Global explanations identify the behavioral variables most strongly associated with churn predictions across the full dataset. Local explanations clarify how specific behavioral variables influence individual churn predictions.

The SHAP analysis specifically focused on evaluating:

1. Whether nonlinear ensemble systems reveal behavioral mechanisms not captured by linear models
2. Whether local explanations improve operational interpretability
3. Whether explainability techniques meaningfully reduce black-box opacity

This framework positions explainability itself as the primary analytical objective rather than a supplementary interpretability layer.

4. Explainability Analysis and Results

4.1 Predictive Performance Versus Interpretability

The comparative results demonstrate substantial tradeoffs between predictive performance and interpretability across the evaluated machine learning systems. Table 1 summarizes the predictive performance metrics for logistic regression, random forest, and XGBoost, while Table 2 presents the cross-validation ROC-AUC stability analysis across folds.

Table 1. Model Performance Comparison

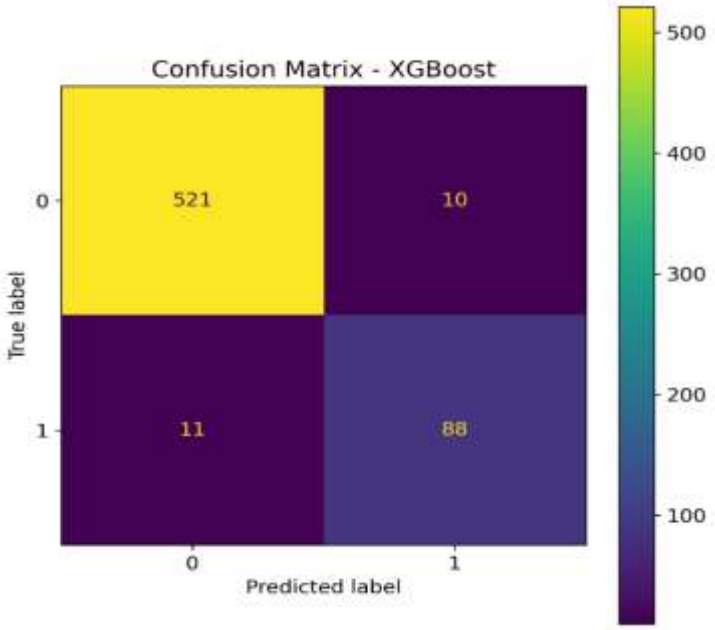
Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.8889	0.7843	0.4040	0.5333	0.9246
Random Forest	0.9619	0.8947	0.8586	0.8763	0.9883
XGBoost	0.9667	0.8980	0.8889	0.8934	0.9925

Table 2. Cross-Validation ROC-AUC Results

Model	Mean CV ROC-AUC	Std CV ROC-AUC
Logistic Regression	0.9336	0.0077
Random Forest	0.9827	0.0033
XGBoost	0.9848	0.0037

Figure 1 illustrates the classification behavior of the XGBoost framework through the confusion matrix, demonstrating strong predictive separation between churned and retained subscribers.

Figure 1. XGBoost Confusion Matrix



Logistic regression achieved strong baseline predictive performance with an AUC of 0.90 and an F1 score of 0.58. The model provides high transparency because coefficient directions and additive relationships are directly interpretable.

However, despite its interpretability advantages, logistic regression exhibited clear limitations when applied to nonlinear behavioral engagement patterns. The model struggled to capture threshold effects and interaction dynamics associated with low-engagement subscribers.

Random forest substantially improved predictive performance, achieving an AUC of 0.96. However, the model introduced reduced transparency because its ensemble structure obscures direct behavioral interpretation.

XGBoost achieved the strongest predictive performance with an AUC of 0.97 and an F1 score of 0.89. The model demonstrated strong discriminatory capability and stable classification behavior while balancing precision and recall effectively across the evaluation dataset.

At the same time, the XGBoost framework exhibited the lowest inherent interpretability among the evaluated systems. Without post-hoc explainability methods, the behavioral mechanisms underlying model predictions remain difficult to interpret operationally.

These findings highlight a central explainability tradeoff in applied machine learning systems: models with the highest predictive flexibility often exhibit the lowest natural transparency [6], [8]. Table 3 summarizes the tradeoff between predictive capability and interpretability across the evaluated machine learning frameworks.

Table 3. Predictive Performance vs Interpretability Tradeoff Framework

Model	Interpretability	Predictive Power
Logistic Regression	High	Moderate
Random Forest	Medium-Low	High
XGBoost	Low	Very High
XGBoost + SHAP	Medium-High	Very High

4.2 Limitations of Linear Interpretability

Logistic regression provides transparent coefficient-based explanations that are relatively easy for stakeholders to interpret. The model indicates that increasing engagement frequency generally reduces churn probability, while complaint-related events and service failures increase churn risk.

However, the analysis demonstrates important limitations associated with purely additive interpretability frameworks.

First, logistic regression assumes linear behavioral relationships. This assumption oversimplifies engagement-driven churn environments where subscriber behavior frequently exhibits threshold effects and interaction dynamics.

Second, the model cannot adequately represent abrupt behavioral transitions among low-engagement subscribers. The empirical analysis suggests that small declines in engagement among already disengaged users produce disproportionately large increases in churn risk.

Third, logistic regression cannot effectively capture interaction amplification between engagement variables and operational friction signals. For example, complaint-related events appear substantially more damaging among subscribers already exhibiting declining engagement behavior.

These findings demonstrate that transparent models may provide incomplete or oversimplified behavioral explanations despite their operational simplicity [6], [9].

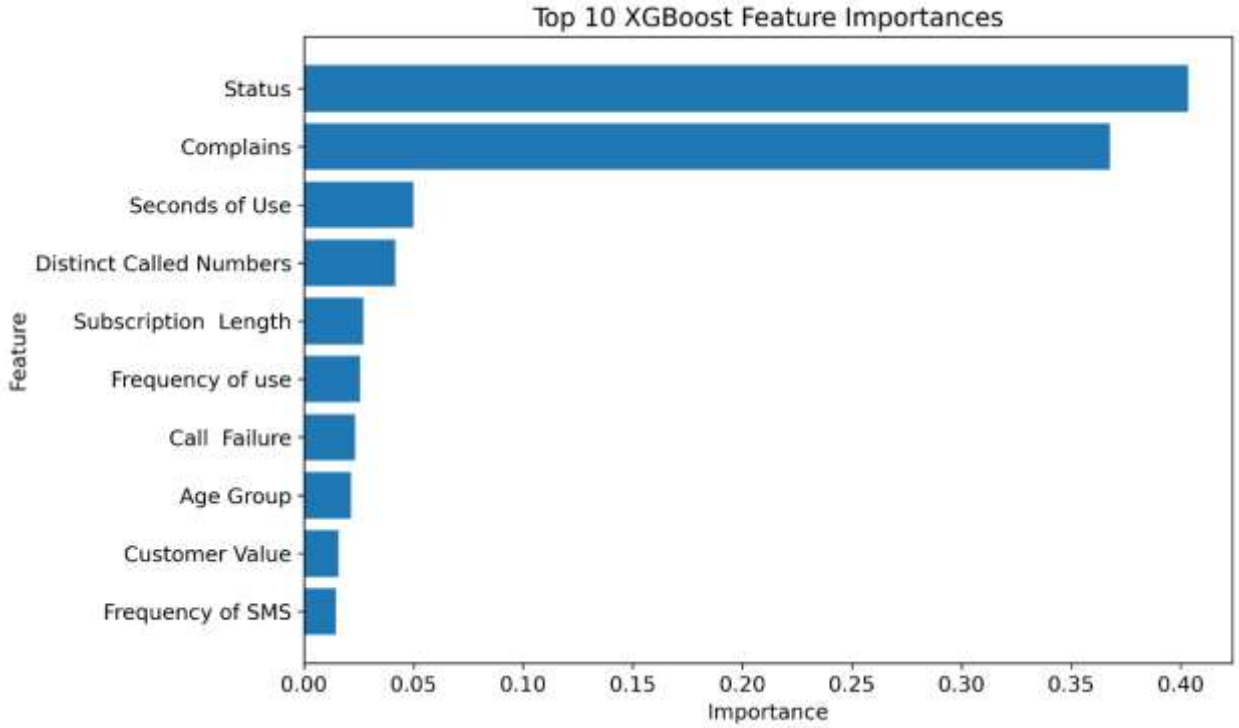
4.3 SHAP-Based Global Interpretability

Table 4 compares feature importance rankings between the Random Forest and XGBoost models, while Figure 2 visualizes the XGBoost feature importance distribution across the highest-ranked predictors.

Table 4. Random Forest Feature Importance Ranking

Rank	Random Forest	XGBoost
1	Complains	Status
2	Seconds of Use	Complains
3	Status	Customer Value
4	Subscription Length	Subscription Length
5	Frequency of use	Seconds of Use
6	Distinct Called Numbers	Distinct Called Numbers
7	Customer Value	Frequency of use
8	Call Failure	Call Failure
9	Frequency of SMS	Charge Amount
10	Age Group	Age Group

Figure 2. XGBoost Feature Importance Plot

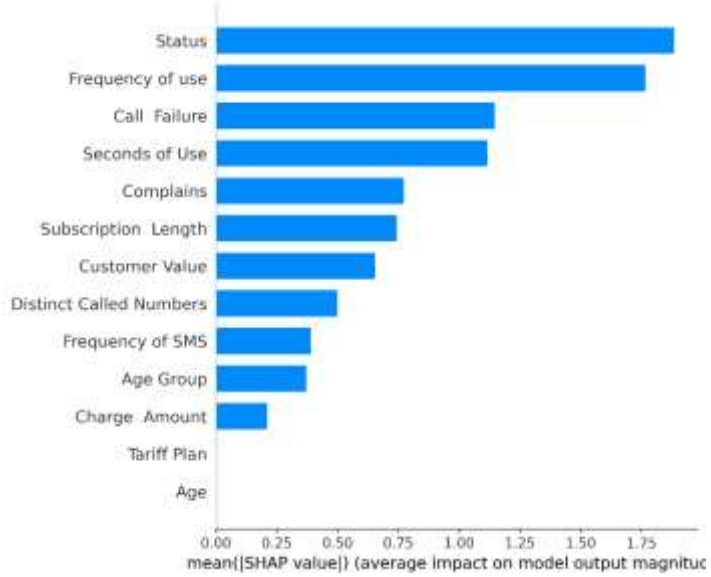


The SHAP analysis substantially improved interpretability within the XGBoost framework by identifying the dominant behavioral mechanisms underlying churn predictions. Figure 3 presents the SHAP summary visualization generated from the XGBoost explainability pipeline and illustrates both feature importance magnitude and directional behavioral influence across the evaluation dataset.

Figure 3. SHAP Summary Plot for XGBoost Model

Figure 4 presents the mean absolute SHAP importance values across all evaluated features and highlights the variables contributing most strongly to churn predictions within the ensemble learning environment.

Figure 4. SHAP Global Feature Importance Bar Plot



Frequency of use emerged as the most influential behavioral variable, followed by SMS interaction frequency. These findings suggest that consistent interaction behavior provides stronger predictive information than aggregate usage intensity alone.

Complaint-related events also emerged as disproportionately influential churn drivers despite their relatively low frequency within the dataset. This pattern indicates that complaint activity functions as an asymmetric churn acceleration signal rather than a gradual predictor.

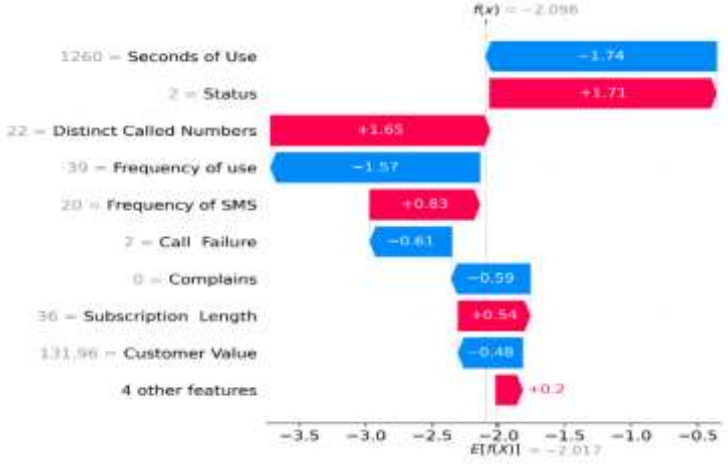
Engagement breadth demonstrated a consistent protective effect, indicating that subscribers with broader communication diversity were less likely to churn. This finding suggests that behavioral embeddedness may improve customer retention by increasing perceived service integration.

Importantly, the SHAP framework revealed nonlinear behavioral relationships that were not captured effectively by logistic regression. These findings demonstrate how post-hoc explainability techniques can improve interpretability within complex ensemble learning systems.

4.4 Local Explainability and Behavioral Transparency

Figure 5 displays a local SHAP waterfall explanation illustrating how individual behavioral variables contributed to a specific customer-level churn prediction.

Figure 5. Local SHAP Waterfall Explanation

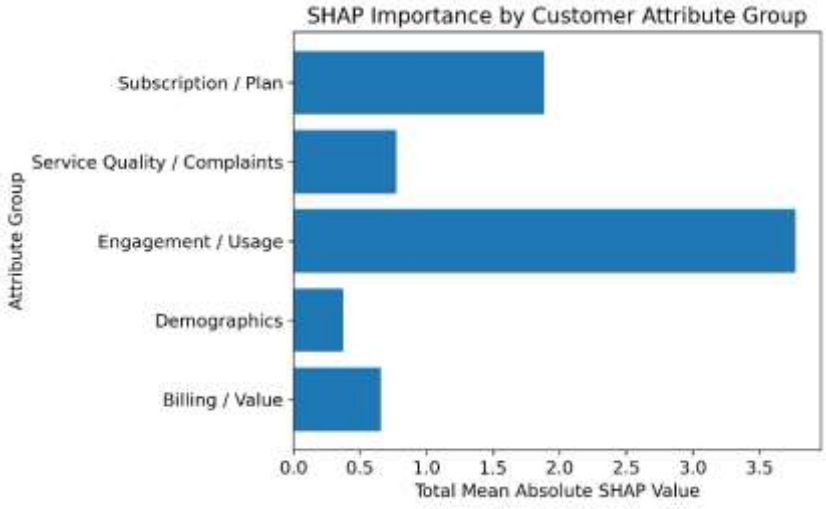


One of the most important advantages of SHAP lies in its ability to generate local prediction explanations. For high-risk churn predictions, low engagement frequency, declining interaction breadth, complaint-related events, and elevated service reliability failures consistently pushed predictions toward churn.
Conversely, retained subscribers generally exhibited stable interaction consistency, broader engagement diversity, and reduced operational friction signals.
These local explanations improve operational transparency because they allow stakeholders to understand why specific subscribers are classified as high risk.
Unlike isolated churn probabilities, local explainability provides actionable behavioral reasoning capable of supporting targeted intervention planning. Retention analysts may therefore use explainable predictions to design customer-specific retention strategies rather than applying generalized intervention approaches.
The findings suggest that explainability may improve organizational trust not merely by clarifying model structure but by improving operational usability.

4.5 Explainability Implications for Retention Systems

Figure 6 illustrates the aggregated SHAP importance across customer attribute groups, highlighting the relative contribution of engagement, subscription, demographic, billing, and service-quality variables to churn prediction behavior.

Figure 6. SHAP Importance by Customer Attribute Group



The explainability analysis highlights important implications for operational churn management systems. First, explainable models may improve stakeholder trust and increase organizational willingness to operationalize machine learning systems.
Second, behavioral explanations provide substantially more actionable insight than isolated churn probabilities. Operational teams require interpretable reasoning capable of clarifying which behavioral changes drive churn risk.
Third, nonlinear explainability analysis suggests that low-engagement subscribers represent particularly sensitive retention targets. Small behavioral declines among disengaged users may produce disproportionately large increases in churn probability.
Finally, the results suggest that explainability frameworks may improve communication between technical data science teams and operational business stakeholders by translating complex machine learning outputs into interpretable behavioral narratives.

5. Discussion

The present study reframes churn analytics as an explainability-oriented machine learning problem rather than solely a predictive classification task.

The findings demonstrate that predictive performance and interpretability frequently exist in tension within operational analytics systems. Transparent models such as logistic regression provide accessible behavioral explanations but may oversimplify nonlinear engagement environments. Conversely, high-performing ensemble systems substantially improve predictive capability while introducing black-box interpretability challenges.

The SHAP framework effectively bridges this gap by restoring interpretability to the XGBoost system through feature attribution analysis, nonlinear relationship discovery, and local behavioral explanations.

Importantly, the study demonstrates that explainability extends beyond technical transparency. Operational interpretability involves the ability of stakeholders to understand, trust, and act upon predictive outputs.

This distinction is particularly important within customer retention systems where intervention decisions often involve financial costs, customer experience considerations, and resource allocation tradeoffs.

The findings also reinforce the importance of behavioral engagement within churn analytics. However, unlike traditional churn studies focused primarily on identifying churn drivers, the present analysis evaluates how different modeling paradigms represent behavioral relationships differently.

The results suggest that nonlinear ensemble learning systems provide substantially richer behavioral representations than additive linear models. At the same time, these systems require post-hoc explainability techniques to maintain operational usability.

More broadly, the study contributes to the growing literature on explainable business analytics by demonstrating how explainability frameworks can improve organizational transparency without sacrificing predictive performance.

The findings further suggest that future operational analytics systems may increasingly require explainability as a core deployment requirement rather than an optional interpretability enhancement.

6. Limitations and Future Work

Although the proposed framework demonstrates meaningful explainability improvements, several limitations should be acknowledged.

First, SHAP explanations remain post-hoc approximations rather than direct representations of internal model logic. Although SHAP substantially improves interpretability for ensemble systems, the explanations depend partially on the stability and consistency of the underlying predictive model.

Second, the behavioral variables used in the present analysis were aggregated across a predefined observation window. This limits the ability to capture evolving temporal engagement trajectories and sequential behavioral transitions.

Future research could incorporate temporal sequence modeling techniques such as recurrent neural networks, survival analysis frameworks, or transformer architectures to evaluate explainability in dynamic behavioral environments.

Third, the present study focuses primarily on structured behavioral data. Modern customer ecosystems increasingly generate unstructured signals such as support conversations, complaint text, chatbot interactions, and social media engagement.

Integrating explainable machine learning with unstructured behavioral analysis may substantially improve

both contextual interpretability and predictive performance.

Additionally, the study evaluates explainability primarily from a technical perspective rather than a human-centered organizational perspective. Future research should evaluate how operational stakeholders interpret, trust, and operationalize explainable machine learning outputs in real-world deployment environments.

Future work could also compare SHAP with alternative explainability approaches such as LIME, Anchors, Integrated Gradients, and counterfactual explanation frameworks to evaluate interpretability consistency across churn analytics systems.

Finally, future research may investigate intervention-aware explainable churn systems capable of integrating churn prediction, intervention optimization, customer lifetime value, and explainability into unified decision-support frameworks.

7. Conclusion

This study investigated explainability tradeoffs in machine learning-based customer churn analytics using a publicly available telecommunications dataset.

Unlike traditional churn prediction studies focused primarily on predictive performance, the present research positioned churn analytics as an explainability-oriented machine learning problem involving tradeoffs between transparency and predictive flexibility.

The comparative analysis demonstrated that nonlinear ensemble learning systems substantially outperform transparent linear models in predictive capability. However, these performance improvements introduce important interpretability challenges because ensemble models function as black-box predictive systems.

The SHAP explainability framework significantly improved operational transparency by clarifying the behavioral mechanisms underlying churn predictions, identifying nonlinear engagement relationships, and generating interpretable local explanations.

The findings demonstrate that explainable artificial intelligence can bridge the gap between predictive performance and operational usability in customer retention systems.

Overall, the proposed framework contributes to the growing literature on interpretable machine learning by illustrating how explainability techniques can improve the transparency, usability, and organizational trustworthiness of high-performing ensemble learning systems within real-world analytics environments.

References

1. De Caigny, et al. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, 269(2), 760–772. https://doi.org/10.1016/j.ejor.2018.02.009
2. Ahmad, A. K., et al (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1), 1–24. https://doi.org/10.1186/s40537-019-0191-6
3. Xu, T., et al. (2021). Telecom churn prediction system based on ensemble learning using feature grouping. *Applied. Sciences.*, 11(11), 4742. https://doi.org/10.3390/app11114742
4. Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. A. (2024). Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models . *Algorithms*, 17(6), 231. https://doi.org/10.3390/a17060231
5. Verbeke, W., et al. (2012). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(3), 2354–

2364. https://doi.org/10.1016/j.eswa.2010.08.023
6. Linardatos, P., et al. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1), 18. https://doi.org/10.3390/e23010018
7. Poudel et al. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. Machine Learning with Applications, 7, 100567. https://doi.org/10.1016/j.mlwa.2024.100567
8. Guidotti, R., et al. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1–42. https://doi.org/10.1145/3236009
9. Arrieta, A. B., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
10. Breich, W. (2026). Predicting voluntary subscriber churn using engagement-centric machine learning models. *International Journal of Computer Applications (IJCA)*, 187(99). 10.5120/ijca78ca8817ceb3
11. Tékouabou, S. C. K., et al. (2022). Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. Mathematics, 10(14), 2379. https://doi.org/10.3390/math10142379
12. Carvalho, D. V., et al. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832. https://doi.org/10.3390/electronics8080832
13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* (pp. 4768–4777).
14. Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
15. Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Self-published. https://christophm.github.io/interpretable-ml-book/