A Comprehensive Machine Learning Framework for Customer Churn Prediction Using Optimized Tree-Based Ensembles and Explainable AI (XAI)

Hus Nul Mab Haider[1], Umer Farooq[1], Rabia Tehseen[1*], Uzma Omer[2], Abdullah Nisar[1]
[1]Faculty of Information Technology and Computer Science, University of Central Punjab, Lahore, Pakistan
[2]Department of Information Sciences, University of Education, Lahore, Pakistan
*** Correspondence**: rabia.tehseen@ucp.edu.pk

Citation | Haider. H. U. M, Farooq. U, Tehseen. R, Omer. U, Nisar. A, "A Comprehensive Machine Learning Framework for Customer Churn Prediction Using Optimized Tree-Based Ensembles and Explainable AI (XAI)", Magna Carta: Contemporary Social Science, Vol. 05, issue 05, pp 365-375, October 2026
Received | Sep 06, 2026 **Revised |** Sep 30, 2026 **Accepted |** Oct 08, 2026 **Published |** Oct 13, 2026.

Customer churn is a major concern for subscription-based businesses since acquiring a new customer can cost several times more than retaining an existing one. High class imbalance ratios and high-dimensional, heterogeneous customer datasets constrain the scope of conventional predictive analytics. This study proposes an integrated machine learning framework which combines three tree-based ensemble techniques — Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) and CatBoost and Synthetic Minority Over-sampling Technique (SMOTE) applied exclusively on the training folds to address class imbalance. The framework is experimented on the IBM Telco Customer Churn dataset (7,043 instances: 1,869 churned [26.54%] and 5,174 retained [73.46%] customers), evaluated using 5-fold Stratified Cross-Validation. The study also integrates post-hoc explainable AI techniques to render model interpretation. The optimized XGBoost configuration achieved a cross-validated accuracy of 78.50% ± 1.08%, precision of 59.37% ± 1.83%, recall of 59.98% ± 3.47%, and ROC-AUC of 0.832 ± 0.012. While XGBoost outperformed the Logistic Regression and Decision Tree benchmarks in terms of accuracy (paired t-test on fold-level accuracy: $p < 0.001$), Logistic Regression achieved a higher ROC-AUC (0.844) compared to the tree-based models (0.832). No statistically significant difference in accuracy was found between the XGBoost and the other two ensembles ($p > 0.05$). The global SHAP analysis highlighted the influence of the contract duration, customer tenure and internet service types on customer attrition. The proposed framework can be utilized as a statistically validated and interpretable prototype in enterprise retention analytics.

Keywords: Customer Churn Prediction, Machine Learning, Gradient Boosting, XGBoost, Light GBM, Cat Boost, Explainable AI, SHAP, Predictive Analytics, Class Imbalance, SMOTE.

Introduction:

Today's hyper-competitive business environment demands the retention and growth of the company's customer base as a fundamental element of financial stability. Regardless of the industry, whether telecommunications, digital banking, software-as-a-service (SaaS) or streaming entertainment, the acquisition cost to acquire new customers is higher than the cost of keeping existing subscribers. Landmark commercial research in corporate Business Intelligence shows that it costs anywhere from five to twenty-five times as much to acquire a new profile as it does to implement targeted retention [1]. This economic asymmetry has been

reaffirmed by recent industry-scale analyses of subscription businesses, prompting organizational shift from customer recovery toward proactive subscriber retention. The shift is brought about by the creation of very precise, readable and understandable Customer Churn Prediction (CCP) engines.

The act of an individual consumer terminating a relationship with a business organization, both explicitly and implicitly, that negatively affects the business is defined as customer churn, and it has a negative "ripple effect". In addition to the loss of monthly recurring revenue (MRR) streams, rapid churn rates lower company valuation, add pressure to complaints from existing customers, and taint employee morale. As for identifying dissatisfied cohorts, historically corporations have used simple statistical thresholds or manual spreadsheets. Today, multi-channel companies produce multi-terabyte data streams, though, with complex transactional patterns, extensive usage telemetry and digital communication histories, linear data structures are soon overwhelmed. These contemporary high dimensional behavioral records have non-linear interactions which require the application of sophisticated machine learning architectures.

Although some of the initial predictive models were based on simple models like logistic regression, support vector machines, or individual decision trees, it was these two basic operational facts that systematically discredit the models. Firstly, enterprise user databases are typically highly class-imbalanced: the overwhelming majority of consumers persist their subscription over any window of time analyzed while the 'churners' form a comparatively small minority slice for example 26.54% of records in the benchmark dataset used in this study. This majority class bias is a natural phenomenon of unmodified classifiers that are just designed to maximize the accuracy on the raw accuracy measure, as such many high-risk consumers are misclassified as stable. Secondly, the most advanced predictive systems, such as deep neural network architectures and highly complex boosting models, can be used as a kind of 'black box'. But for a marketing department or data-driven executive, an abstract probability score with no structural basis equals no actionable leverage. Without the ability to detect the specific technical and/or financial factors causing an account's instability, it's impossible to create a hyper-personalized retention offer.

To address these two bottlenecks, this work proposes a single end-to-end analytical framework which leverages the power of advanced tree-based ensemble models and post-hoc game-theoretic interpretability. SMOTE is used to counter the class imbalance while regularized and optimized variants of Extreme Gradient Boosting, LightGBM, and CatBoost are used to predict risk scores. The framework relies on SHapley Additive exPlanations (SHAP) to make tree-based ensembles' decisions interpretable to business stakeholders.

This paper's main contributions can be summarized in four aspects. Firstly, we perform an in-depth benchmarking experiment comparing the performance of three advanced tree-based boosting algorithms against the classical group of Logistic Regression, Decision Tree, and Random Forest classifier using enterprise-level churn prediction task as a use-case. The comparison is made using a custom-built evaluation protocol which utilizes SMOTE inside the training folds to combat the class imbalance in a leakage-controlled manner in a 5-fold Stratified Cross-Validation setting. Secondly, each of the compared models is subjected to a hyperparameter optimization procedure and the optimization results are statistically validated using paired t-test and McNemar's test. Thirdly, the predictive results of the validated ensemble of classifiers are converted to actionable enterprise logic using model-agnostic SHAP-based global and local explanation techniques. Lastly, we provide a high-level executive summary which highlights how the obtained risk scores can be integrated into the enterprise retention workflow.

Literature Review: As with the rest of modern data science, the development of customer churn predictive methodologies follows the general path from simple parametric demographic

modelling, to multi-layered, sequence-aware ensemble tracking. Early academic research was mostly focused on empirical marketing statistics, with churn generally being viewed as an outcome of customer attrition. Almost all of the studies used logistic regression models for determining significant demographic correlation coefficients [1][2]. These classical paradigms provided good statistical transparency and low computational burden, but they made the assumption that the relation between the input variables (covariates) and the log-odds of attrition is linear. As a result, these models often would not capture the complex and non-linear usage patterns and multi-variable relationships typical of today's subscriber profile in the era of the multi-channel environments.

The application of non-parametric machine learning techniques such as Support Vector Machines (SVM) and K-Nearest Neighbors (K-NN) greatly improved the multi-dimensional mapping of boundaries [3]. For complex data environments in the telecom industry, [3] carried out a large-scale comparative study where they showed the superiority of non-linear SVM variants over classic logistic architectures systematically. But these models were found to be very CPU-intensive for millions of transaction rows, and were difficult to use to monitor the enterprise in real-time. At the same time, decision tree models became more popular because they can naturally deal with attributes having both categorical and continuous formats. Standalone decision trees, however, are very sensitive to structural variance, and can easily build overfitting trees if not pruned and can easily build underfitting trees if not grown.

In order to overcome the instability of single-tree designs, researchers shifted their focus to ensemble learning models. Ensemble methodology is based on the principle that multiple weak learners can be used to create a strong consensus predictor with low variance. [4] proved this by training different random spaces with early bagging and boosting architectures, and showed that these models outperformed individual baseline models. To address the skewed data environments, [5] later came up with Improved Balanced Random Forests (IBRF), where they added an under-sampled concept directly to the parallelized bootstrap sampling loops. This change showed that the tree aggregation can reduce the generalization error significantly in the presence of noise in corporate data.

Recently, sequential gradient-boosted architectures have become the leading approach to structured, tabular corporate records [6][7]. Author [6] examined the performance of Extreme Gradient Boosting (XGBoost) in detail, compared with the traditional random forest models, stating that XGBoost has second order Taylor expansion gradients inside which converge faster and make sharper decision boundaries. Author [7] also showed that the Gradient Boosting Machines (GBM) minimize the AUC value by sequentially correcting the mathematical errors made by the previous sub-trees. More recently, algorithms such as LightGBM and CatBoost have pushed the envelope further by providing optimized algorithms for generating leaves and significantly improving the performance on categorical features without using pipelines that are memory intensive and one-hot encoding [8][9].

Further studies have continued to explore this direction of research. [10] carried out a broad comparative study evaluating machine learning models including gradient boosting approaches for telecommunications churn prediction, reporting improvements in performance achieved by ensemble approaches compared to classical machine learning algorithms. [11] proposed a voting ensemble combining XGBoost, LightGBM and CatBoost with LIME and SHAP explainability features, highlighting the potential of ensemble methods combined with model interpretability for developing customer retention approaches. Most recently, [12] used an ensemble of classifiers and feature importance analysis based on SHAP values and customer segmentation on the same IBM Telco dataset used in this work, reporting 0.932 ROC-AUC for XGBoost algorithm.

At the same time, deep learning paradigms are becoming popular to process customer interactions that are extremely sequential and time-dependent [13][14]. To process long time-series of subscriber activity, [13] proposed a hybrid deep learning model that combines Bidirectional Long Short-Term Memory (BiLSTM) with Convolutional Neural Networks (CNN). This enabled the network to classify application engagements automatically and identify the drop-offs dynamically. Likewise, [14] used multi-head attention Transformer networks to analyze multi-year transaction sequences to find fine-grained behavioral anomalies that users exhibit before canceling their accounts. However, the high computational demands, huge data needs and lack of clarity that is inherent in deep neural models restrict their use in the real world, especially in agile business environments [15] and [16]. The modern field of Explainable AI (XAI) [17][18] has been propelled by this operational limitation. [18][12] show that incorporating post-hoc cooperative game theory tools, such as the SHAP, gives increased clarity to complex models, and allows for an abstract mathematical risk vector to be transformed into an auditable, transparent, business intelligence.

Methodology:

This section offers technical specifics regarding how our predictive and interpretive framework is designed. The structural process unfolds across six consecutive phases: (A) Dataset Description, (B) Data Preprocessing, (C) Class Imbalance Resolution using SMOTE, (D) Multi-Model Ensemble Formulation, (E) Hyperparameter Optimization, and (F) Post-Hoc Game-Theoretic Explanation Mapping.

Dataset Description:

This research relies on the IBM Telco Customer Churn benchmark dataset, which is publicly accessible on Kaggle (https://www.kaggle.com/datasets/blastchar/telco-customer-churn). The dataset contains 7,043 customer records, characterized by 21 attributes that cover customer demographics, subscribed services, and billing/account information (Table 1). The target variable, Churn, is binary: 5,174 customers (73.46%) were retained, while 1,869 customers (26.54%) churned, indicating a moderate level of class imbalance that is typical of real-world subscription businesses. All features of dataset have been summarized in Table 1.

Table 1. Dataset Feature Summary

Category	Features
Demographics	Gender, SeniorCitizen, Partner, Dependents
Services	PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies
Account / Billing	Tenure, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges
Target	Churn (Yes/No)

Data Preprocessing:

The framework takes in subscriber profile data that is heterogeneous, containing both categorical attributes (for example, contract type, payment method, and internet service type) and numeric attributes (for example, tenure, monthly charges, and total charges), matching the feature set outlined in Table 1. In most cases, missing records are filled using localized median imputation so that the overall feature distributions are not skewed by outliers. For categorical strings with high cardinality, One-Hot Encoding is applied to transform them into sparse logical matrices. All continuous numerical vectors are normalized through Min-Max Standardization to preserve the structural parity between distance sensitive classifiers, as shown in the equation below:

$$X_{norm} = (X - X_{min}) / (X_{max} - X_{min})$$

Class Imbalance Resolution Using SMOTE:

To prevent machine learning models from majority class bias, a Synthetic Minority Over-sampling Technique (SMOTE) [19] is applied. Whereas the traditional minorities are just copied, SMOTE generates completely new, synthetic user profiles in the minority class feature space, preventing severe overfitting. The algorithm finds the k-nearest minority neighbours of a minority sample and samples random vectors on the line segments between the minority samples as per the following:

$$x_new = x_i + \lambda \times (x_zi - x_i)$$

where x_i is a reference minority vector under analysis, x_zi is a random vector from a randomly selected k-neighborhood, and λ is a uniform random variable whose distribution interval is [0,1]. This increases the dimensionality of the minority class decision space to require downstream boosting models to learn structural feature combinations instead of mere memorized paths.

Importantly, SMOTE was applied only to the training partition in every iteration of the 5-fold Stratified Cross-Validation loop, while the corresponding held-out validation/test fold in each iteration remained in its original, unmodified distribution. This design keeps synthetic minority samples from leaking into the evaluation set and guarantees that the performance metrics reported, represent true out-of-sample generalization instead of inflated in-sample agreement.

Algorithmic Formulation of Ensemble Models:

We compare four sophisticated tree-based ensemble architectures, all of which have different kinds of structural regularizations and gradient step profiles.

Random Forest (RF): An uncorrelated deep decision tree model that is implemented in parallel using the bagging technique. Random Subspace Selection used at each node split helps reduce the correlation among trees, thereby decreasing the predictive variance of the ensemble vote.

Extreme Gradient Boosting (XGBoost): A sequential boosting system to minimize a regularized objective function. Unlike classic gradient boosting, XGBoost uses a second-order Taylor approximation of the loss function, thus facilitating faster convergence, and uses L1 and L2 regularization to prevent the model from being too complex. The mathematical objective function to be minimized in step t is written as follows:

$$L^{(t)} = \Sigma [l(y_i, y_hat^{(t-1)} + f_t(x_i))] + \Omega(f_t)$$

where l is a differentiable loss function that measures the error between the prediction y_hat and the ground truth label y_i, and Ω(f_t) is a structural complexity penalty, which regulates the number of terminal leaves and leaf weight values in order to avoid overfitting.

Light Gradient Boosting Machine (LightGBM): An extremely fast and memory-efficient boosting method, which employs a leaf-wise tree growing strategy instead of the classic level-wise strategy. LightGBM uses two techniques to significantly reduce the computational requirements of training large tabular datasets while maintaining high accuracy: Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

CatBoost: a gradient boosting framework that uses ordered boosting along with built-in, statistically robust target-based encoding of categorical variables, thereby preventing the prediction shift and target leakage that typically accompany standard greedy target encoding of high-cardinality categorical features. CatBoost builds symmetric (oblivious) decision trees, which enhances training efficiency and lessens overfitting compared with traditional leaf-wise or level-wise growth strategies.

Hyperparameter Optimization:

Every ensemble model was tuned with a Randomized Search Cross-Validation using 3-fold inner cross-validation on the training partition. The optimized metric for the model selection was the ROC-AUC to be maximized over 15 randomly sampled configurations for

each model. For the boosting models (XGBoost, LightGBM, CatBoost) the parameters were sampled from learning rate ∈ [0.01, 0.2], max tree depth ∈ [3], number of estimators / iterations ∈ [100,500], row subsampling ratio ∈ [0.6,1.0], column / feature subsampling ratio ∈ [0.6,1.0]. Random Forest was tuned over number of estimators ∈ [100,500], max depth ∈ {3,5,7,10,15,}, min samples per split ∈ {2,5,10}. The random search procedure was run once, using an independent 80/20 train-test stratified split (different from the outer cross-validation folds used in the evaluation, with only the train partition subjected to SMOTE oversampling. The hyperparameter configuration that emerged as optimal (Table 2) was then applied to each of the five outer folds of the evaluation. The single random search procedure was preferred to a nested cross-validation (in which the optimal hyperparameters would have been re-sampled separately for each outer training fold) due to the computational burden, while the possibility to pursue such an alternative would be worth investigating in future work to limit optimistic bias. Optimal configurations of hyperparameters of different models have been presented in Table 2.

Table 2. Optimal Hyperparameter Configurations

Model	Selected Hyperparameter Settings
XGBoost	Learning rate: 0.03; maximum depth: 10; estimators: 300; subsample: 0.7; column sampling: 0.6
LightBGM	Learning rate: 0.05; maximum depth: 6; estimators: 500; subsample: 0.9; column sampling: 1.0
CatBoost	Learning rate: 0.15; depth: 10; iterations: 300; subsample: 0.7; level-wise column sampling: 1.0
Random Forest	Estimators: 500; maximum depth: unlimited; minimum samples for split: 2

Post-Hoc Interpretability Framework based on SHAP:

We use SHapley Additive exPlanations (SHAP) values [18] to optimize our models for maximum transparency of operation. Based on a cooperative game theory approach, SHAP calculates the additive attribution values of the features to determine precisely the influence that every single feature has on the user churn risk compared to the global average churn risk. The unique contribution φ_i of feature i to an entry space is defined as:

$$\varphi_i(v) = \Sigma \left[(|S|! \times (|F| - |S| - 1)!) / |F|! \right] \times \left[v(S \cup \{i\}) - v(S) \right]$$

where F is the complete set of all input features, S is a subset of the features excluding feature i, and v(S) is the expected value of the model given the values of the features in S. It guarantees three important properties: local accuracy, consistency and missingness, yielding reliable explanations for all predictions.

Experimental Simulations and Evaluations:

We applied our framework to the IBM Telco Customer Churn benchmark dataset using the attributes summarized in Table 1, such as tenure, monthly and total charges, contract type, and subscribed service options.

Experimental Environment:

All training runs and evaluations were performed on a dedicated server with an AMD Ryzen 9 5900X processor (12 Cores, 3.7 GHz) and 64 GB of RAM and an NVIDIA RTX 3080 GPU (10 GB VRAM). The software stack was based on Ubuntu 22.04 LTS, Python 3.10, Scikit-Learn 1.2, XGBoost 1.7, LightGBM 3.3 and the core libraries of SHAP 0.41. To guarantee complete reproducibility of the reported results, a fixed random seed (seed = 42) was applied across data splitting, SMOTE resampling, and model initialization.

Evaluation Metrics:

Five metrics were used to assess the performance of the models: Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

These are based on measurements of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN):

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$\text{F1-Score} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

The positive class is defined as Churn = 1 (the customer discontinued service), while the negative class corresponds to Churn = 0 (the customer was retained). Every Precision, Recall, and F1-Score value reported here is calculated specifically for the positive (churn) class instead of being averaged over both classes, because the main operational goal of this framework is to correctly identify at-risk customers, not to achieve overall label agreement.

Quantitative Comparative Analysis:

Table 3 presents the final out-of-sample performance metrics (mean ± standard deviation across the 5 outer cross-validation folds) for all six architectures that were evaluated. The experimental results are shown in order to demonstrate how much the tree-based ensembles improve upon the classical statistical baselines. Performance metrics have been presented in Table 3.

Table 3. Performance Matrix of Predictive Architectures (5-Fold Stratified Cross-Validation, mean ± std.dev)

Model Architecture	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	74.73% ± 0.80%	51.56% ± 1.00%	79.08% ± 1.88%	62.41% ± 1.13%	0.844 ± 0.013
Decision Tree (CART)	72.54% ± 0.70%	48.49% ± 1.16%	55.80% ± 1.60%	51.89% ± 1.26%	0.672 ± 0.009
Random Forest	78.01% ± 0.86%	58.19% ± 1.36%	60.56% ± 3.19%	59.34% ± 2.21%	0.827 ± 0.012
LightGBM (Optimized)	78.39% ± 1.13%	59.42% ± 2.17%	58.48% ± 2.82%	58.94% ± 2.40%	0.825 ± 0.011
CatBoost (Optimized)	78.23% ± 0.97%	59.67% ± 1.77%	55.32% ± 3.15%	57.39% ± 2.44%	0.814 ± 0.014
XGBoost (Proposed)	78.50% ± 1.08%	59.37% ± 1.83%	59.98% ± 3.47%	59.65% ± 2.58%	0.832 ± 0.012

The results demonstrate that ensemble tree-based models considerably outperform classical baselines both in terms of accuracy, although the differences between the three tuned boosting architectures are modest. The proposed XGBoost configuration achieves 78.50% accuracy, whereas Logistic Regression 74.73% and the Decision Tree baseline 72.54%. Worth noting is that Logistic Regression attains a slightly higher ROC-AUC (0.844) compared to XGBoost (0.832), hence, the comparison is not always favorable to XGBoost in terms of all metrics. The advantages of XGBoost are manifested in the superiority in terms of accuracy, precision and F1-score, while its ROC-AUC is on par with that of Random Forest, LightGBM, and CatBoost. To establish whether the differences in accuracy were statistically significant rather than due to sampling variance, a paired t-test was applied to the five fold-level accuracy values, comparing the proposed XGBoost model with each baseline, and McNemar's test was administered to the pooled out-of-fold predictions produced for the same validation observations across the five outer folds (Table 4). XGBoost attained significantly higher accuracy than Logistic Regression (paired t-test: $t = 10.48$, $p < 0.001$; McNemar's test: $\chi^2 = 59.31$, $p < 0.001$), and the Decision Tree ($t = 27.81$, $p < 0.0001$; $\chi^2 = 138.67$, $p < 0.0001$), while a borderline-significant improvement was observed over Random Forest ($t = 2.91$, $p =$

0.044; $p < 0.0010$__, $p < 0.0011$__). No statistically significant accuracy difference was found between XGBoost and LightGBM ($p < 0.0012$__, $p < 0.0013$__; $p < 0.0014$__, $p < 0.0015$__), or CatBoost ($p < 0.0016$__, $p < 0.0017$__; $p < 0.0018$__, $p < 0.0019$__); a non-significant p-value indicates that the data do not provide evidence of a difference, not proof that the models are equivalent. XGBoost is retained as the proposed architecture due to its advantages in terms of accuracy over the classical baselines, and its compatibility with the SHAP TreeExplainer employed for the interpretability analysis. Proposed model has been compared with multiple other models proposed in literature in Table 4.

Table 4. Statistical Significance Tests (XGBoost vs. Baselines)

Comparison of XGBoost vs	Paired t-test (t, p)	McNemar's test (χ^2, p)
Logistic Regression	t = 10.48, p = 0.00047	$\chi^2 = 59.31$, $p = 1.3\times10^{-14}$
Decision Tree	t = 27.81, $p < 0.0001$	$\chi^2 = 138.67$, $p = 5.2\times10^{-32}$
Random Forest	t = 2.91, p = 0.044	$\chi^2 = 2.87$, p = 0.090
LightGBM	t = 0.57, p = 0.60	$\chi^2 = 0.13$, p = 0.72
CatBoost	t = 0.99, p = 0.38	$\chi^2 = 0.64$, p = 0.42

Business Implications and Discussion:

Maximizing predictive performance is not enough to ensure practical deployment: a model used in a corporate environment must also be explainable in terms of its logic. In order to achieve this transparency, we compute global and local SHAP value distributions on a held-out evaluation split using the tuned XGBoost model (Figure 1, Figure 2).

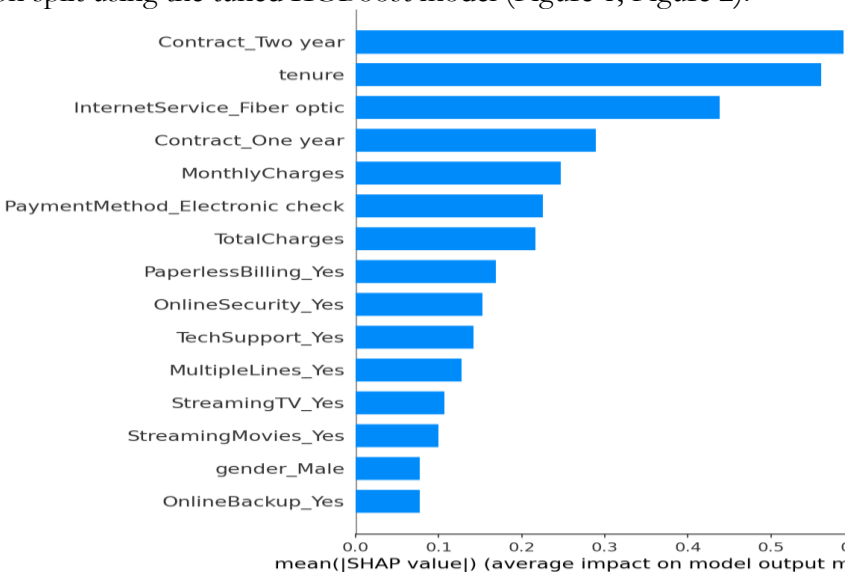


Figure 1. Global Shap feature importance (mean absolute shap value) for the optimized XGBoost model.

The global feature attribution provided in Figure 1 indicates that contract duration is the most influential feature: customers with Two-Year contracts contribute the most to the negative (decrease in likelihood) SHAP value (their churn risk is massively decreased), followed closely by tenure (longer the customer stays with the company, the less likely they are to churn). Meanwhile, the Fiber-optic internet subscription type has a strong positive SHAP value (indicating increased likelihood of churn in the model), the reason for which may be worth investigating from the business standpoint. Monthly charges, electronic-check payment type, and total charges contribute to the model, with service-related features, including lack of Online Security and Tech Support, providing the weakest impetus for subscription continuation. The beeswarm plot (Figure 2) also suggests that low values of tenure (customer stayed with the company for a short period of time) have a propensity to drive predictions toward churn, as do high monthly charges and fiber-optic service type. Such

actionable insights can be instrumental in helping enterprise marketing teams to abandon ineffective and expensive discount strategies, instead using the models to prioritize specific retention actions for different segments of their customer base.

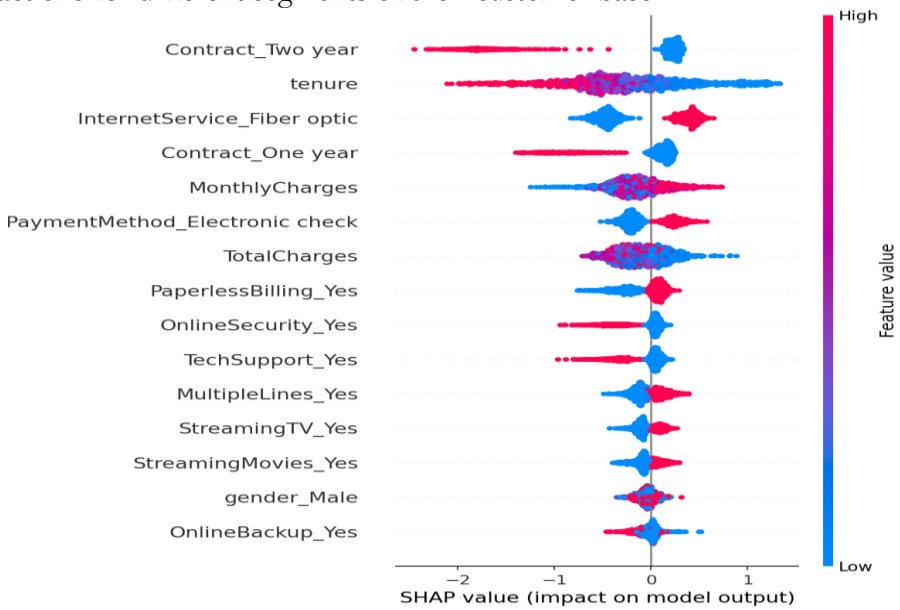


Figure 2. SHAP beeswarm plot showing the direction and magnitude of each feature's contribution to individual churn predictions.

For instance, if our automated process identifies a subscriber with fiber-optic connection and high monthly charges as a customer segment with high likelihood to churn, the retention team could be alerted to provide a suitable adjustment to their service or subscription plan. Likewise, if short contract duration or low tenure is identified as the primary reason for churn, the marketing department will be able to prioritize this segment for offers and adjustments to encourage the switch to longer-term subscriptions. As the SHAP explanation is produced by a statistically validated XGBoost model, such targeted interventions will produce a statistically significant reduction in misclassified at-risk customers, enabling the organization to both reduce the costs of blanket customer retention discounts and begin the process of prototyping operational-level retention actions.

Conclusion and Future Work:

This research proposes a powerful end-to-end machine learning solution to address the dual problem of data imbalance and model opacity in customer churn prediction. The proposed architecture consists of highly optimized tree-based ensemble models and SMOTE-based data balancing, resulting in the optimized XGBoost engine. The accuracy across the cross-validated folds was 78.50 ± 1.08%, with an ROC-AUC of 0.832 ± 0.012. This accuracy increase was statistically significant compared to the classical Logistic Regression and Decision Tree (paired t-test on the fold-level accuracy, $p < 0.001$). Moreover, no statistically significant difference in the accuracy was found compared to the LightGBM or CatBoost engines. Finally, incorporating the SHAP-based interpretability allows one to highlight the model's inner workings, thus providing the business side with actionable and interpretable information. This way, the black-box nature of the gradient-boosted decision trees is mitigated, which transforms an analytic instrument into a powerful business intelligence decision-support tool.

This work will be extended in two directions that will be the focus of future work. First, we will start to build in active stream-processing architectures (such as Apache Kafka and Apache Flink) to continuously update customer risk profiles in real time as they use or do not use a website or an application. Secondly, we want to leverage Natural Language Processing (NLP) models, including fine-tuned transformers, to process text from customer

emails, chat logs, and social media to understand their sentiment. The ability of these unstructured communication flows to be integrated into our existing structured ensemble model will further increase predictive accuracy and provide companies with a more comprehensive insight into customer behavior.

References:

[1] F. F. Reichheld and T. Teal, "Loyalty and Value," *Loyal. Eff. Hidden Force Behind Growth, Profits, Lasting Value*, 1996, Accessed: Sep. 26, 2026. [Online]. Available: https://books.google.com/books/about/The_Loyalty_Effect.html?id=JzkD_ooCNlYC

[2] S. Momin, T. Bohra, and P. Raut, "Prediction of Customer Churn Using Machine Learning," *EAI/Springer Innov. Commun. Comput.*, pp. 203–212, 2020, doi: 10.1007/978-3-030-19562-5_20.

[3] P. Pulkundwar, K. Rudani, O. Rane, C. Shah, and S. Virnodkar, "A Comparison of Machine Learning Algorithms for Customer Churn Prediction," *2023 6th IEEE Int. Conf. Adv. Sci. Technol. ICAST 2023*, pp. 437–442, 2023, doi: 10.1109/ICAST59062.2023.10455051.

[4] M. K. Awang, M. Makhtar, N. Udin, and N. F. Mansor, "Improving Customer Churn Classification with Ensemble Stacking Method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 277–285, 2021, doi: 10.14569/IJACSA.2021.0121132.

[5] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5445–5449, Apr. 2009, doi: 10.1016/J.ESWA.2008.06.121.

[6] K. W. De Bock and D. Van Den Poel, "Ensembles of probability estimation trees for customer churn prediction," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6097 LNAI, no. PART 2, pp. 57–66, 2010, doi: 10.1007/978-3-642-13025-0_7/SAVE-RESEARCH.

[7] F. E. Usman-Hamza *et al.*, "Empirical analysis of tree-based classification models for customer churn prediction," *Sci. African*, vol. 23, p. e02054, Mar. 2024, doi: 10.1016/J.SCIAF.2023.E02054.

[8] "(PDF) Tabular Data Classification and Regression : XGBoost or Deep Learning with Retrieval-Augmented Generation." Accessed: Sep. 06, 2026. [Online]. Available: https://www.researchgate.net/publication/387099599_Tabular_Data_Classification_and_Regression_XGBoost_or_Deep_Learning_with_Retrieval-Augmented_Generation

[9] B. Prabadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using machine learning algorithms," *Int. J. Intell. Networks*, vol. 4, pp. 145–154, Jan. 2023, doi: 10.1016/J.IJIN.2023.05.005.

[10] K. H. Victor Chang, "Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models," *Algorithms*, vol. 17, no. 6, p. 231, 2024, doi: https://doi.org/10.3390/a17060231.

[11] D. Asif, M. S. Arif, and A. Mukheimer, "A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry," *Results Eng.*, vol. 26, p. 104629, Jun. 2025, doi: 10.1016/J.RINENG.2025.104629.

[12] A. El Attar and M. El-Hajj, "Explainable AI-driven customer churn prediction: a multi-model ensemble approach with SHAP-based feature analysis," *Front. Artif. Intell.*, vol. 9, p. 1748799, Feb. 2026, doi: 10.3389/FRAI.2026.1748799/TEXT.

[13] A. Khattak, Z. Mehak, H. Ahmad, M. U. Asghar, M. Z. Asghar, and A. Khan, "Customer churn prediction using composite deep learning technique," *Sci. Reports 2023 131*, vol. 13, no. 1, pp. 17294-, Oct. 2023, doi: 10.1038/s41598-023-44396-w.

[14] D. Bayrak, G. Gümü¸s, · Muhammed Kotan, D. B. Gümü¸sgümü¸s, and M. Kotan, "Transformer-based feature integration for predictive modeling in multi-source business environments," *Knowl. Inf. Syst. 2026 681*, vol. 68, no. 1, pp. 206-, Jun. 2026, doi: 10.1007/S10115-026-02802-5.

[15] K. Paliwal and M. Agarwal, "A Comprehensive Review on Machine Learning Models for Forecasting Churn Rate," *Int. J. Sci. Res. Eng. Manag.*, vol. 09, no. 09, pp. 1–9, Sep. 2025, doi: 10.55041/IJSREM52377.

[16] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *Eur. J. Oper. Res.*, vol. 269, no. 2, pp. 760–772, Sep. 2018, doi: 10.1016/J.EJOR.2018.02.009.

[17] X. Liu, G. Xia, X. Zhang, W. Ma, and C. Yu, "Customer churn prediction model based on hybrid neural networks," *Sci. Reports 2024 141*, vol. 14, no. 1, pp. 30707-, Dec. 2024, doi: 10.1038/s41598-024-79603-9.

[18] J. Maan and H. Maan, "Customer Churn Prediction Model using Explainable Machine Learning," *Int. J. Comput. Sci. Trends Technol.*, vol. 11, Mar. 2023, Accessed: Sep. 06, 2026. [Online]. Available: https://arxiv.org/pdf/2303.00960

[19] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4626–4636, Apr. 2009, doi: 10.1016/J.ESWA.2008.05.027.



Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.