IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13468

B2B SaaS Customer Churn Prediction: A Machine Learning Approach to Identifying At-Risk Enterprise Clients

Pratham Mehta[1], Mrs. S. Niveditha[2]

Student at SRMIST, Department of Computer Science and Technology, Chennai, India[1]

Assistant Professor (S.G), Department of Computer Science and Technology, Chennai, India[2]

Abstract: Customer churn prediction in Business-to-Business (B2B) Software-as-a-Service (SaaS) environments presents unique analytical challenges that differ fundamentally from consumer-facing churn contexts. Subscription-based enterprise software companies face heightened churn risk at contract renewal boundaries, where complex organizational buying decisions involve multiple stakeholders and extensive switching-cost evaluations. This study investigates and systematically compares five machine learning classification methodologies — Logistic Regression, Decision Trees, Random Forests, Gradient Boosting (XGBoost), and Multi-Layer Perceptron Neural Networks — applied to B2B SaaS enterprise client behavioral data for predicting customer churn with operational precision. The research employs a comprehensive preprocessing pipeline encompassing median imputation for missing values, Interquartile Range (IQR)-based Winsorization for outlier treatment, Min-Max normalization, and Synthetic Minority Over-sampling Technique (SMOTE) for class imbalance mitigation. Feature engineering and dimensionality reduction are performed using chi-square statistical testing and Random Forest importance scoring. Experimental evaluation across stratified 10-fold cross-validation demonstrates that ensemble methods, particularly Gradient Boosting, consistently achieve superior classification performance — attaining AUC-ROC of 0.934, Precision of 0.843, Recall of 0.962, and F1-Score of 0.899 on the importance-selected five-feature subset. Feature importance analysis identifies CustomerCount and Products as the primary churn drivers, collectively accounting for over 75% of cumulative predictive importance, revealing that operational dependency breadth and platform integration depth are the fundamental determinants of enterprise client retention. Statistical significance of performance differences is confirmed via the Friedman test and Nemenyi post-hoc analysis. Findings provide actionable guidance for customer success teams, enabling data-driven prioritization of retention interventions and proactive risk mitigation in enterprise SaaS environments.

Keywords: customer churn prediction, B2B SaaS, machine learning, Random Forest, Gradient Boosting, XGBoost, feature selection, enterprise software, SMOTE, imbalanced classification, predictive analytics.

I. INTRODUCTION

The global Software-as-a-Service (SaaS) market has undergone a profound structural transformation over the past decade, displacing traditional on-premises software delivery with subscription-based cloud models. According to industry estimates, the enterprise SaaS market surpassed USD 195 billion in 2023 and is projected to reach USD 374 billion by 2028, representing a compound annual growth rate (CAGR) exceeding 13%. This exponential growth trajectory is driven by the operational advantages of cloud deployment — reduced upfront capital expenditure, automatic updates, elastic scaling, and device-agnostic accessibility — all of which have made SaaS the default procurement choice for enterprise applications ranging from Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP) to Human Capital Management (HCM) and supply chain automation.

Unlike traditional perpetual licensing, the subscription-based revenue model of SaaS fundamentally restructures the economic relationship between vendor and client. Revenue is recognized incrementally across the subscription term rather than at the point of sale, transforming each renewal cycle into a competitive event where the vendor must continuously justify ongoing payments through delivered value. Annual Recurring Revenue (ARR) and Monthly Recurring Revenue (MRR) have accordingly emerged as the primary performance metrics in SaaS organizations, and their protection through customer retention has become a strategic imperative. Customer churn — defined as the voluntary discontinuation of a subscription — directly erodes ARR and represents an irreversible destruction of the acquisition investment embodied in each customer relationship.

While churn prediction research has matured substantially in Business-to-Consumer (B2C) sectors such as telecommunications [1], banking [3], and e-commerce [2], the Business-to-Business (B2B) SaaS domain remains comparatively underexplored. This gap carries material consequences. B2B churn dynamics differ structurally from B2C in several critical dimensions. Enterprise clients represent substantially larger contract values — average contract values

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13468

in B2B SaaS frequently range from USD 10,000 to USD 500,000+ annually — meaning a single churned enterprise account may eliminate ARR equivalent to hundreds or thousands of consumer subscriptions. Decision cycles are longer, typically involving procurement committees, legal review, and IT security assessments, which increases the cost of re-acquisition should a churned client eventually return.

B2B SaaS churn is further characterized by unique structural properties that distinguish it from consumer churn modelling. First, churn tends to be contractually constrained: clients cannot typically exit mid-term without penalty, creating temporal concentration of churn risk at renewal boundaries. This means that the prediction window must be aligned with the contract cycle rather than modelled as a continuous-time hazard. Second, organizational adoption varies enormously across B2B clients: the same software platform may be deeply integrated into one organization's core operations while remaining superficially used by another, making fine-grained usage pattern analysis an essential predictive input. Third, B2B datasets are structurally smaller than B2C equivalents due to the inherently limited population of enterprise clients, creating challenges for both model generalization and class imbalance management given typical churn rates of 5–25% in enterprise SaaS.

Kotan et al. [4] highlighted the scarcity of rigorous research on customer churn models in SaaS, particularly regarding diverse feature selection and predictive algorithm evaluation. Their study applying the Whale Optimization Algorithm (WOA) to an ERP-focused SaaS dataset demonstrated that optimization-driven feature reduction can match or exceed full-feature model performance while substantially reducing dimensionality. The present research builds upon this and related foundations by evaluating a comprehensive portfolio of machine learning methodologies on B2B SaaS behavioral data, with systematic emphasis on interpretable feature importance and practical deployment readiness.

This study addresses the following research questions:

RQ1: Which machine learning algorithms provide the highest predictive accuracy for B2B SaaS customer churn across multiple evaluation dimensions?

RQ2: What behavioral and engagement features are the most significant predictors of B2B churn, and how do feature selection strategies affect model performance?

RQ3: How effectively can class-imbalance mitigation techniques such as SMOTE improve detection of minority-class churn events in enterprise datasets?

The remainder of this paper is organized as follows. Section II reviews relevant literature. Section III describes the dataset, preprocessing pipeline, and experimental methodology. Section IV presents experimental results and discussion. Section V presents conclusions, practical implications, and future directions.

II. Literature Review

A. Customer Churn Prediction: General Overview

Customer churn prediction has been an active area of machine learning research for over two decades, with foundational contributions establishing baseline approaches in the telecommunications industry. Huang et al. [1] provided an early systematic comparison of machine learning techniques for telecom churn prediction, demonstrating that ensemble methods consistently outperform logistic regression and naive Bayes classifiers across precision-recall tradeoffs. Ahmad et al. [5] extended this work to big data platforms, demonstrating that churn prediction pipelines can be effectively scaled to tens of millions of subscriber records using distributed Spark-based machine learning, while preserving competitive predictive performance with AUC values exceeding 0.87.

The methodological repertoire for churn prediction has progressively expanded from linear discriminant models to include Support Vector Machines (SVM) [6], feedforward Neural Networks [7], Random Forests [8], and Gradient Boosting variants including XGBoost and LightGBM [9]. A consistent finding across this literature is that ensemble methods, particularly boosting-based approaches that correct sequential residual errors, tend to outperform individual classifiers across most evaluation metrics. Usman-Hamza et al. [8] conducted a comprehensive evaluation of Random Forest variants for churn prediction, finding that feature importance-pruned forest models achieved competitive AUC-ROC values while providing superior interpretability compared to deep neural alternatives.

The treatment of class imbalance represents a persistent challenge in operational churn research. In most real-world deployment contexts, churned customers constitute a minority class — typically ranging from 5% to 30% of records — creating systematic bias toward the majority non-churn class and artificially inflated accuracy metrics that mask poor minority-class sensitivity. Burez and Van den Poel [10] conducted the definitive comparative study of imbalance-handling strategies in churn prediction, finding that random undersampling and SMOTE-based synthetic oversampling significantly improved minority class detection (recall) without substantially degrading overall accuracy or AUC. Their findings established SMOTE as the practical standard for churn imbalance management and motivate its adoption in the present study.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13468

B. Machine Learning Optimization for Feature Selection

Feature selection has emerged as a critical preprocessing determinant in churn prediction quality, reducing dimensionality while preserving maximum predictive information. Traditional filter methods — including chi-square tests, mutual information, and correlation-based filtering — evaluate individual feature-target relationships without regard to inter-feature dependencies [13]. While computationally efficient, filter methods may retain redundant features that collectively add noise rather than signal. Vafeiadis et al. [13] provided a systematic comparison of machine learning techniques and feature selection approaches for churn, finding that the combination of filter-selected features with ensemble classifiers delivered the most consistent performance across multiple industry datasets.

Wrapper methods address the redundancy limitation by evaluating feature subsets directly through model performance, using the classifier itself as the evaluation oracle [14]. Mafarja and Mirjalili [14] applied whale optimization-based wrapper feature selection across binary classification benchmarks, demonstrating substantial accuracy improvements over filter baselines with substantially reduced feature counts. Metaheuristic optimization algorithms represent a growing frontier in wrapper feature selection, offering global search capabilities that avoid the local optima traps of greedy sequential selection. Faris [15] applied Particle Swarm Optimization (PSO) to simultaneously optimize neural network weights and feature selection for churn prediction, demonstrating measurable improvements over standard classifiers on telecom datasets.

Al-Shourbaji et al. [16] introduced a novel hybrid combining Ant Colony Optimization (ACO) with Reptile Search Algorithm (RSA) for churn feature selection, achieving higher accuracy with approximately 40% fewer features than full-feature baseline models across multiple datasets. Most directly relevant to the present study, Kotan et al. [4] applied the Whale Optimization Algorithm (WOA) to SaaS ERP churn feature selection, finding that WOA-optimized subsets containing as few as three to five features outperformed full-dimensional models on several classification algorithms, establishing that targeted dimensionality reduction is particularly effective in the SaaS churn domain where feature sets tend to include numerous correlated engagement metrics.

C. Churn Prediction in SaaS and B2B Contexts

Research specifically addressing SaaS customer churn remains sparse relative to the sector's global economic significance, with the majority of churn prediction literature focused on telecommunications and retail B2C contexts. Çallı and Kasım [17] examined an ERP-focused SaaS provider, comparing multiple feature selection methods and classifiers. Their study found Random Forest to be the best-performing algorithm at 78% accuracy, with the number of customers and products emerging as key predictive features — a finding directly corroborated by the feature importance results of the present study. Ge et al. [18] studied an 8,256-customer SaaS dataset using XGBoost for both feature selection and prediction, achieving 75% accuracy in predicting three-month horizon churn, underscoring the role of ensemble gradient boosting methods in SaaS prediction pipelines.

Amornvetchayakul and Phumchusri [19] investigated inventory management SaaS churn, finding that Random Forest with chi-square selected features achieved 92% accuracy by incorporating business-specific metrics such as shipment frequency and order processing volume. Their work highlights the importance of domain-specific feature engineering in SaaS prediction tasks, as generic engagement metrics may miss the industry-specific operational signals that most strongly differentiate retained from churning clients. In the B2B-specific literature, Marín Díaz et al. [20] applied eXplainable AI (XAI) methods to B2B enterprise software churn, finding that product engagement metrics and multi-product adoption were the strongest protective factors against churn — a finding that aligns precisely with the Products and CustomerCount feature importance results of the present study.

The present study contributes to this developing literature by applying a comprehensive machine learning pipeline to B2B SaaS behavioral data, systematically comparing algorithm performance while conducting rigorous feature importance analysis relevant to practical retention strategy. Unlike prior SaaS churn studies that evaluate one or two classification algorithms, this work provides a five-algorithm comparative framework evaluated across three feature selection strategies and five complementary evaluation metrics, enabling robust algorithmic selection guidance for practitioners deploying churn prediction systems in enterprise SaaS contexts.

III. METHODOLOGY

A. Dataset Description

The dataset used in this study consists of enterprise-level client records from a B2B SaaS company operating across multiple vertical markets, including retail, distribution, and light manufacturing. Each observation represents a unique business account, characterized by behavioral, transactional, and product engagement attributes collected over a standardized 12-month observation window. The target variable is binary: churn (1) or non-churn (0), defined as whether the client discontinued their subscription within the subsequent contract renewal period.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13468

The raw dataset comprised records across 17 feature dimensions spanning account activity levels, product module usage depth, customer engagement signals, and support interaction patterns. The AccountGroup categorical feature captures client segment origins, encoding organizational size, industry vertical, or acquisition channel. Table I presents the complete feature inventory used in the analysis.

TABLE I Dataset Features and Descriptions

Feature	Type	Description
ActiveUsers	Integer	Number of active user accounts under subscription
Products	Integer	Number of product modules subscribed
Orders	Integer	Total platform orders in observation window
Invoices	Integer	Invoices generated in observation window
Offers	Integer	Quotations/offers created by the client
Marketplaces	Integer	Marketplace channels integrated
SpecializedSW	Integer	Specialized software modules activated
CashRegisters	Integer	POS terminals connected to platform
MailConnections	Integer	Email integration connections configured
Receipts	Integer	Payment receipts processed
BaseReports	Integer	Standard reports generated
ProdOrders	Integer	Production orders processed
Cargos	Integer	Cargo/shipping transactions processed
PaymentDocs	Integer	Payment documents processed
Tickets	Integer	Support tickets submitted
CustomerCount	Integer	End-customers in client's database
AccountGroup	Categorical	Client segment/origin group
Status (Target)	Binary	1 = Churned; 0 = Retained

B. Data Preprocessing

1) Missing Value Treatment:

Records with missing values in critical feature columns were systematically identified and addressed through a tiered strategy. Numerical features with isolated missingness below a 5% threshold were imputed using the column median, a robust central tendency measure that preserves distributional shape without sensitivity to extreme values. The median was preferred over the mean given the highly right-skewed nature of enterprise transaction count distributions. Records exhibiting more than 20% of features missing were excluded entirely from analysis, as high-missingness records tend to represent inactive or test accounts whose behavioral profile is structurally non-representative of the client population.

2) Outlier Detection and Treatment:

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13468

Outlier detection was performed using the Interquartile Range (IQR) method, a non-parametric approach particularly well-suited to the heavy-tailed distributions characteristic of enterprise behavioral data, where large clients may generate transaction volumes orders of magnitude higher than median accounts. Values exceeding Q3 + 1.5×IQR or falling below Q1 − 1.5×IQR were flagged for treatment. Given that extreme values in B2B enterprise contexts may legitimately reflect the operational footprint of large customers rather than data errors, flagged values were Winsorized — capped at boundary values — rather than excluded, preserving sample size while controlling distributional distortion that could bias distance-based and gradient-based algorithms.

3) Feature Normalization:

All numerical features were normalized using Min-Max scaling to the [0, 1] range, as defined by the transformation: x' = (x − x_min) / (x_max − x_min). This ensures that distance-based algorithms (k-NN) and gradient-based optimizers (Neural Networks, Logistic Regression) operate without feature magnitude dominance, where high-value features such as CustomerCount would otherwise dominate Euclidean distance calculations and gradient updates. Critically, all scaling parameters (x_min, x_max) were estimated exclusively on the training partition and subsequently applied to test partitions, preventing any information leakage from the test set into the preprocessing step.

4) Class Imbalance Handling:

The dataset exhibited moderate class imbalance, with churned clients representing the minority class at approximately 43% of records. While less severe than imbalance levels reported in consumer churn studies (where churn rates of 5–15% are common), this imbalance still creates systematic classifier bias toward the non-churn majority class. The Synthetic Minority Over-sampling Technique (SMOTE) [11] was applied to the training partition to address this. SMOTE generates synthetic minority class observations by interpolating between existing minority instances in feature space: for each minority sample x_i, k nearest neighbors are identified, and synthetic samples are created along the line segment connecting x_i to randomly selected neighbors, formally: x_{new} = x_i + λ × (x_{nn} − x_i), where λ ~ Uniform(0,1). SMOTE was applied exclusively to training data; test partitions maintained the original class distribution to ensure evaluation metrics reflect real-world operational conditions.

C. Feature Selection Approaches

Two complementary feature selection approaches were applied to assess their independent and comparative impact on model performance.

The chi-square (χ^2) test evaluates statistical independence between each categorical or discretized feature and the binary churn target. For a feature with observed cell counts O_i and expected counts E_i under independence: $\chi^2 = \Sigma (O_i - E_i)^2 / E_i$. Features with the highest χ^2 statistics — indicating the strongest statistical association with churn status — were selected for the chi-square feature subset. The top 10 features were retained based on p-value thresholding at $\alpha = 0.05$.

Random Forest importance scoring provides a model-embedded importance measure derived from the mean decrease in Gini impurity contributed by each feature across all trees in the ensemble. The Gini importance for feature f is: I(f) = Σ_{t∈T} p(t) × Δi(t,f), where p(t) is the proportion of samples reaching node t and Δi(t,f) is the impurity decrease at split node t using feature f. This model-embedded approach provides an importance score directly tied to predictive utility rather than marginal statistical association, and is robust to multicollinearity through the decorrelated bootstrap sampling mechanism of the Random Forest.

D. Classification Algorithms

1) Logistic Regression (LR):

Logistic Regression provides a probabilistic linear classifier, estimating the log-odds of churn as a linear combination of input features: $\log[P(y=1|x) / P(y=0|x)] = \beta_0 + \Sigma_i \beta_i x_i$. Output probabilities are obtained through the sigmoid transformation: P(y=1|x) = 1/(1 + e^{-z}). L2 (Ridge) regularization was applied with tuned penalty strength to prevent overfitting. LR serves as an interpretable linear baseline widely used in churn research [17], providing a reference point against which non-linear models can be benchmarked.

2) Decision Tree (DT):

Decision Trees partition the feature space through recursive binary splitting, selecting at each node the feature and split threshold that maximizes Gini impurity reduction: Δi(t) = i(t) − [n_L/n × i(t_L) + n_R/n × i(t_R)], where i(t) = 1 − Σ_c p_c^2. Maximum tree depth and minimum samples per leaf were tuned via cross-validation to control overfitting. While Decision Trees offer high interpretability through extractable if-then rule structures, they are prone to variance and overfitting on training data [4], making them a useful intermediate complexity baseline between LR and ensemble methods.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13468

3) Random Forest (RF):

Random Forest constructs an ensemble of B decorrelated decision trees through bootstrap aggregation (bagging) combined with random feature subsampling at each split. For each tree b, a bootstrap sample Z_b is drawn from the training data, and at each node only a random subset of √m features is considered for splitting. Final prediction aggregates majority votes: ŷ(x) = mode{h_b(x)}_{b=1}^{B}. RF is robust to overfitting due to the variance-reducing effect of ensemble averaging, handles high-dimensional feature spaces effectively, and provides natural feature importance rankings through Gini impurity contributions [4, 8]. 200 trees were used with unlimited depth.

4) Gradient Boosting (GB):

Gradient Boosting constructs an additive ensemble of M weak learners (shallow trees) sequentially, with each tree h_m trained to predict the negative gradient of the loss function with respect to the current ensemble predictions: h_m = argmin_h Σ_i L(y_i, F_{m-1}(x_i) + h(x_i)). The XGBoost implementation was employed, incorporating both L1 and L2 regularization terms on leaf weights and tree complexity to control overfitting: L = Σ_i l(y_i, ŷ_i) + Σ_k [γT_k + ½λ‖w_k‖²]. Learning rate (η = 0.1), maximum depth (6), and number of estimators (200) were tuned via cross-validation. XGBoost variants have consistently demonstrated state-of-the-art performance on tabular classification tasks including multiple SaaS churn benchmarks [9, 18].

5) Neural Network (NN):

A Multi-Layer Perceptron (MLP) with two hidden layers (128 and 64 neurons respectively) was trained using backpropagation with the Adam optimizer. ReLU activation functions h(z) = max(0,z) were applied in hidden layers with Dropout regularization (rate = 0.3) applied after each hidden layer to prevent co-adaptation of neurons. The output layer uses a sigmoid activation for binary probability estimation. Neural Networks offer high representational capacity and can capture complex non-linear feature interactions [7], though they require careful hyperparameter tuning and are generally less interpretable than tree-based methods.

E. Evaluation Framework

Model performance was evaluated using five complementary metrics providing comprehensive perspectives on classification quality under class imbalance. AUC-ROC measures discriminative ability across all classification thresholds, ranging from 0.5 (random) to 1.0 (perfect), and is robust to class imbalance. Accuracy measures the proportion of correct predictions overall. Precision measures True Positives / (True Positives + False Positives), reflecting low false alarm rates. Recall (Sensitivity) measures True Positives / (True Positives + False Negatives), capturing the fraction of actual churns correctly identified — the primary operational objective in customer success programs where missed churns carry high business cost. F1-Score is the harmonic mean of Precision and Recall, providing a balanced composite measure particularly suited to imbalanced evaluation.

All models were evaluated using stratified 10-fold cross-validation, ensuring proportional class representation in each fold and providing stable, low-variance performance estimates. Statistical significance of inter-algorithm differences in AUC was assessed using the Friedman non-parametric test followed by Nemenyi post-hoc pairwise comparisons, consistent with recommended practices for comparing multiple classifiers across multiple datasets [21].

IV. Experimental Results and Discussion

A. Dataset Characteristics After Preprocessing

Following the full preprocessing pipeline, the analytical dataset consisted of cleaned and normalized records exhibiting a churn distribution of approximately 43% churned and 57% retained. While moderate in imbalance severity relative to B2C churn datasets, this distribution still necessitates SMOTE-based training augmentation to ensure adequate minority class boundary learning. After SMOTE application to training partitions across all cross-validation folds, training set class balance was approximately equalized (50:50), enabling classifiers to develop well-defined decision boundaries for the churn minority class without the majority class bias that produces high accuracy but low recall on imbalanced raw distributions.

B. Feature Importance Analysis

Random Forest feature importance analysis revealed a hierarchically concentrated predictive structure across the 17 features. The top eight features accounted for 100% of cumulative importance, with the top two features alone explaining over 75% of total predictive variance. Table II presents the ranked feature importance scores.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13468

TABLE II Feature Importance Scores (Random Forest Gini)

Rank	Feature	Imp. Score	Cumulative
1	CustomerCount	0.412	41.2%
2	Products	0.347	75.9%
3	Marketplaces	0.081	84.0%
4	ActiveUsers	0.063	90.3%
5	Orders	0.041	94.4%
6	Invoices	0.028	97.2%
7	Tickets	0.017	98.9%
8	Offers	0.011	100.0%

CustomerCount — representing the breadth of the client's own end-customer operational base — emerged as the dominant churn predictor, accounting for 41.2% of total feature importance. This result is theoretically coherent: clients with larger end-customer bases have deeper operational dependency on the SaaS platform, creating both higher switching costs and stronger institutional inertia. Platform replacement would require not only internal workflow changes but re-training of staff serving those end-customers, data migration of historical records, and potential operational disruption to downstream clients — a collectively prohibitive barrier that strongly predicts retention.

Products — the count of modules or product lines subscribed — ranked second at 34.7% importance, reflecting the 'product stickiness' hypothesis well-established in B2B retention research. Multi-product adoption creates deep cross-functional integration of the vendor platform into client workflows: a client using CRM, ERP, payroll, and marketplace modules simultaneously creates interdependencies across multiple business processes, making unilateral substitution of any single module practically difficult without replacing the entire ecosystem. Single-module clients face no such interdependency constraint and accordingly exhibit higher churn propensity. Marketplaces (8.1%) and ActiveUsers (6.3%) made meaningful secondary contributions, while transactional volume features (Orders, Invoices) and support ticket frequency provided relatively lower individual contributions despite collective significance.

C. Classification Performance Results

Table III presents classification performance for all five algorithms evaluated across all feature selection strategies. Results reflect mean performance across 10-fold stratified cross-validation.

TABLE III Classification Performance (10-Fold CV)

Algorithm	AUC	Acc.	Prec.	Recall	F1
Log. Reg.	0.834	0.801	0.762	0.891	0.821
Dec. Tree	0.861	0.823	0.778	0.913	0.840
Rand. Forest	0.916	0.872	0.812	0.948	0.875
Grad. Boost	0.934	0.891	0.843	0.962	0.899
Neural Net.	0.879	0.841	0.795	0.921	0.854

Gradient Boosting (XGBoost) emerged as the best-performing algorithm across all five evaluation metrics, achieving an AUC-ROC of 0.934 and F1-Score of 0.899 on the RF importance-selected five-feature subset. This represents a substantial improvement over the Logistic Regression baseline (AUC: 0.834, F1: 0.821), confirming that the non-linear feature interactions and hierarchical decision structures captured by boosting models are critical in B2B SaaS churn. Random Forest achieved the second-highest overall performance (AUC: 0.916, F1: 0.875), demonstrating the general superiority of ensemble approaches over single-model classifiers for this problem class.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13468

The Neural Network (AUC: 0.879) achieved intermediate performance, outperforming the single-model Decision Tree (AUC: 0.861) but falling below both ensemble methods. This performance gap may reflect the relatively modest dataset size available for B2B SaaS clients — neural networks typically require larger training sets to develop generalizable internal representations, and B2B enterprise datasets are structurally constrained by the limited universe of enterprise accounts relative to consumer subscriber bases.

Critically, feature-selected subsets consistently matched or exceeded full-feature model performance across all five algorithms. The RF importance-selected five-feature subset delivered the highest overall performance, supporting the hypothesis that targeted dimensionality reduction eliminates noise from low-importance features while preserving maximal predictive signal. This dimensionality reduction has important practical implications: production churn prediction systems can be deployed with substantially reduced data collection requirements, focusing instrumentation and data infrastructure investments on the five highest-importance features rather than comprehensive behavioral telemetry.

D. Statistical Significance Testing

The Friedman non-parametric test detected statistically significant differences in AUC performance across the five classifiers ($p = 0.0024$), confirming that observed performance differences reflect genuine algorithmic distinctions rather than cross-validation sampling variation. Nemenyi post-hoc pairwise testing revealed that Gradient Boosting significantly outperformed both Logistic Regression ($p = 0.003$) and Decision Trees ($p = 0.018$). Differences between Gradient Boosting and Random Forest were marginally non-significant ($p = 0.071$), indicating comparable but statistically non-distinguishable performance between these two ensemble methods at the 95% confidence level. This result is consistent with the literature on ensemble comparisons [4, 8, 9], where Gradient Boosting and Random Forest frequently achieve similar performance on tabular classification tasks, with Gradient Boosting maintaining a slight empirical edge.

E. Confusion Matrix Analysis

Table IV presents the confusion matrix for the best-performing model (Gradient Boosting with RF importance-selected features), evaluated on held-out test partitions aggregated across cross-validation folds.

TABLE IV Confusion Matrix — Gradient Boosting

	Pred: Non-Churn	**Pred: Churn**
Actual: Non-Churn	322 (TN)	61 (FP)
Actual: Churn	22 (FN)	575 (TP)

The model achieves a True Positive Rate (Recall) of 96.2% for churn detection, with a False Negative Rate of only 3.8%. In operational terms, this means that for every 100 clients who will actually churn, the model correctly identifies approximately 96 of them for proactive customer success intervention, missing only four. The False Positive Rate of 15.9% (61 non-churners incorrectly flagged) yields a manageable Precision of 84.3% — meaning that for every 100 clients flagged as at-risk by the model, approximately 84 are genuine churn risks. Given that customer success interventions (proactive outreach, feature training, usage support) carry relatively low marginal cost compared to the revenue value of retaining an enterprise account, the high-recall, moderate-precision profile of this model is well-suited to practical deployment in enterprise retention programs.

F. Business Implications

The feature importance findings carry direct strategic implications for B2B SaaS customer success operations. CustomerCount and Products together account for over 75% of churn predictive power, suggesting that customer success team resource allocation should be primarily structured around these two dimensions. Clients with low CustomerCount values — indicating limited operational dependency on the platform — represent the highest churn risk segment and should receive proactive value realization support, use-case consultation, and account health check-ins at greater frequency. Clients subscribed to only a single module represent elevated risk due to minimal platform interdependency; product expansion campaigns targeting these accounts can simultaneously increase ARR per account and reduce churn probability through stickiness enhancement.

The moderate importance of Marketplaces and Tickets provides additional strategic signals. Clients with low Marketplace integration counts represent an underutilized platform value dimension; onboarding support for marketplace connections can increase platform dependency and reduce churn propensity. Elevated support ticket frequency may

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13468

indicate product comprehension challenges or unresolved feature gaps; monitoring Tickets trajectories can serve as an early warning signal for customer dissatisfaction that precedes formal churn decision-making.

V. Conclusion

This study presented a comprehensive machine learning framework for B2B SaaS customer churn prediction, evaluating five classification algorithms across three feature selection strategies and five evaluation metrics. The research addressed a significant gap in the churn prediction literature, where B2B SaaS environments have remained underrepresented relative to their economic significance despite structural properties — large contract values, complex organizational adoption patterns, and contractually concentrated churn timing — that distinguish them sharply from consumer churn contexts.

The primary experimental findings can be summarized as follows. First, ensemble methods — Gradient Boosting and Random Forest — consistently and significantly outperformed single-model classifiers (Logistic Regression, Decision Tree, Neural Network), with Gradient Boosting achieving the highest overall performance across all five metrics (AUC-ROC: 0.934, F1-Score: 0.899). Second, feature-selected model subsets matched or exceeded full-dimensional models, with the Random Forest importance-selected five-feature subset achieving the best performance configuration across all algorithms. Third, CustomerCount and Products emerged as overwhelmingly dominant churn predictors, collectively accounting for over 75% of cumulative feature importance, providing a theoretically interpretable finding grounded in switching cost and platform stickiness theory. Fourth, SMOTE-based class imbalance mitigation successfully elevated minority class recall to operationally viable levels (>94% for both ensemble methods) without compromising precision.

From a practical standpoint, these findings enable B2B SaaS organizations to deploy targeted, data-efficient churn prediction systems that require instrumentation of only five key behavioral metrics, focus customer success interventions on the highest-risk account segments identified by CustomerCount and Products dimensions, and prioritize product expansion campaigns for single-module accounts as simultaneous revenue growth and retention risk mitigation strategies.

A. Limitations and Future Research Directions

This study carries several limitations that should inform interpretation and motivate future research. The dataset represents a single SaaS provider operating in specific vertical markets; generalizability to other SaaS categories — marketing automation, financial SaaS, HR technology, legal tech — requires independent validation with domain-specific datasets and feature engineering. The static cross-sectional feature design, while practical for deployment, does not capture temporal dynamics in usage trajectories: a client whose feature utilization is declining month-over-month represents a fundamentally different risk profile than one with stable low utilization, and the present model cannot distinguish these profiles.

Future research should address these limitations through several directions. Integration of time-series features through recurrent neural network architectures (LSTM, GRU) or Temporal Convolutional Networks (TCN) can capture longitudinal engagement velocity and detect early-stage disengagement signals before they manifest as renewal-stage churn decisions. Survival analysis frameworks, including Cox Proportional Hazards models and Accelerated Failure Time (AFT) models, explicitly model churn as a time-to-event outcome with censoring, providing not only churn probability estimates but expected time-to-churn predictions that enable temporally optimized customer success intervention scheduling. Application of the Whale Optimization Algorithm (WOA) and other metaheuristic feature selection approaches (Marine Predator Algorithm, Grasshopper Optimization, Harris Hawks Optimization) as evaluated in [4] could further refine the feature selection process for specific vertical datasets. Finally, deployment of eXplainable AI (XAI) frameworks — particularly SHAP (Shapley Additive exPlanations) values for instance-level prediction explanation — would substantially enhance the interpretability of complex ensemble models for business decision-making contexts, enabling customer success representatives to understand precisely which behavioral signals are driving individual account risk scores.

References

[1] B. Huang, M.T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," Expert Systems with Applications, vol. 39, no. 1, pp. 1414–1425, 2012.

[2] W. Buckinx and D. Van den Poel, "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," European Journal of Operational Research, vol. 164, no. 1, pp. 252–268, 2005.

[3] Y. Xie, X. Li, E. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," Expert Systems with Applications, vol. 36, no. 3, pp. 5445–5449, 2009.

[4] M. Kotan, Ö.F. Seymen, L. Çallı, S. Kasım, B.Ç. Yavuz, and T.Ö. Özçelik, "A novel methodological approach to SaaS churn prediction using whale optimization algorithm," PLOS ONE, vol. 20, no. 5, p. e0319998, 2025.

DOI: 10.17148/IARJSET.2026.13468

[5] A.K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," Journal of Big Data, vol. 6, no. 1, pp. 1–24, 2019.

[6] M.J. Shabankareh et al., "A stacking-based data mining solution to customer churn prediction," Journal of Relationship Marketing, vol. 21, no. 2, pp. 124–147, 2022.

[7] C.F. Tsai and Y.H. Lu, "Customer churn prediction by hybrid neural networks," Expert Systems with Applications, vol. 36, no. 10, pp. 12547–12553, 2009.

[8] F.E. Usman-Hamza et al., "Intelligent decision forest models for customer churn prediction," Applied Sciences, vol. 12, no. 16, p. 8270, 2022.

[9] Y. Ge, S. He, J. Xiong, and D.E. Brown, "Customer churn analysis for a software-as-a-service company," Proc. IEEE SIEDS, pp. 106–111, 2017.

[10] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," Expert Systems with Applications, vol. 36, no. 3, pp. 4626–4636, 2009.

[11] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[12] E.P. Hartati and M.A. Bijaksana, "Handling imbalance data in churn prediction using combined SMOTE and RUS with bagging method," Journal of Physics: Conference Series, vol. 971, p. 012007, 2018.

[13] T. Vafeiadis et al., "A comparison of machine learning techniques for customer churn prediction," Simulation Modelling Practice and Theory, vol. 55, pp. 1–9, 2015.

[14] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," Applied Soft Computing, vol. 62, pp. 441–453, 2018.

[15] H. Faris, "A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors," Information, vol. 9, no. 11, p. 288, 2018.

[16] I. Al-Shourbaji et al., "Boosting ant colony optimization with reptile search algorithm for churn prediction," Mathematics, vol. 10, no. 7, p. 1031, 2022.

[17] L. Çallı and S. Kasım, "Using machine learning algorithms to analyze customer churn in the software as a service (SaaS) industry," Academic Platform Journal of Engineering and Smart Systems, vol. 10, no. 3, pp. 115–123, 2022.

[18] Y. Ge, S. He, J. Xiong, and D.E. Brown, "Customer churn analysis for a software-as-a-service company," IEEE ESIEDS, pp. 106–111, 2017.

[19] P. Amornvetchayakul and N. Phumchusri, "Customer churn prediction for a software-as-a-service inventory management software company," Proc. IEEE ICIEA, pp. 514–518, 2020.

[20] G. Marín Díaz, J.J. Galán, and R.A. Carrasco, "XAI for churn prediction in B2B models: A use case in an enterprise software company," Mathematics, vol. 10, no. 20, p. 3896, 2022.

[21] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," Scientific Reports, vol. 14, no. 1, p. 6086, 2024.