© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716773

Customer Churning Analysis Using Machine Learning Algorithms

ASHISH PANDEY[1], ANURAG PAL[2], TAJENDRA RIYA[3], DR. SANJAY PACHAURI[4]
[1,2,3,4]*Greater Noida, Institute of Technology*

Abstract- ***In the contemporary subscription-based economy, customer retention is structurally paramount to the long-term viability and profitability of telecommunications enterprises. The financial asymmetry between Customer Acquisition Cost (CAC) and Customer Retention Cost (CRC) mandates the development of highly accurate, proactive customer churn prediction systems. This comprehensive research paper presents an in-depth empirical analysis of machine learning algorithms deployed to forecast customer defection. We rigorously evaluate a spectrum of predictive models, progressing from traditional linear classifiers (Logistic Regression) to complex, non-linear distance-based models (Support Vector Machines), and state-of-the-art ensemble architectures (Random Forest and eXtreme Gradient Boosting). A critical challenge in churn analytics—the severe class imbalance inherent in real-world datasets—is addressed through the application of the Synthetic Minority Over-sampling Technique (SMOTE) combined with rigorous cross-validation strategies. Utilizing a robust telecommunications dataset, our extensive feature engineering and hyperparameter optimization reveal that tree-based ensemble methods, particularly XGBoost, significantly outperform baseline models. XGBoost achieved a superior Area Under the Receiver Operating Characteristic Curve (ROC-AUC) of 0.88, optimizing the critical trade-off between Precision and Recall. Furthermore, the integration of SHapley Additive exPlanations (SHAP) provides a granular, interpretable analysis of feature importance, identifying contract duration, customer tenure, and specific service combinations as the primary drivers of attrition. The findings equip business stakeholders with interpretable, high-fidelity predictive intelligence required to operationalize targeted, cost-effective customer retention interventions.***

Keywords- ***Customer Churn, Machine Learning, eXtreme Gradient Boosting (XGBoost), Synthetic Minority Over-sampling Technique (SMOTE), Predictive Analytics, SHAP Values, Class Imbalance, Telecommunications.***

I. INTRODUCTION

The paradigm of modern corporate strategy has decisively shifted from aggressive customer acquisition to meticulous customer lifecycle management. In industries characterized by saturated markets, high initial infrastructure investments, and minimal switching costs—most notably the telecommunications, banking, and Software-as-a-Service (SaaS) sectors—customer loyalty is a volatile asset. "Customer Churn," defined as the cessation of a commercial relationship by a subscriber within a specified temporal window, represents a direct hemorrhage of recurring revenue.

Empirical financial analyses consistently indicate that acquiring a new customer is approximately five to twenty-five times more capital-intensive than retaining an existing one. Consequently, marginal improvements in retention rates can yield exponential increases in overall corporate profitability. Historically, organizations relied on rudimentary heuristics, retrospective reporting, and reactive save-desk operations to manage churn. However, these methods are intrinsically flawed, as they engage the customer only after the decision to defect has been finalized.

The advent of Big Data, coupled with the exponential growth of computational processing power, has facilitated the transition to proactive, predictive analytics. Machine Learning (ML) algorithms possess the capacity to ingest and synthesize vast, multi-dimensional datasets encompassing demographic profiles, transactional histories, service usage metrics, and customer interaction logs. By mapping the complex, often non-linear topologies of these datasets, ML models can identify latent behavioral patterns that serve as early warning indicators of impending churn.

1.1 Problem Statement and Research Objectives

Despite the proliferation of machine learning tools, developing an effective churn prediction model remains a complex data science challenge. The primary obstacle is the "Class Imbalance Problem." In a functional business environment, the overwhelming majority of customers do not churn. Algorithms trained on heavily skewed data tend to optimize for global accuracy by aggressively predicting the majority class (retention), rendering the model effectively blind to the minority class (churners)—the exact demographic the business seeks to identify. Furthermore, complex models often suffer from a lack of interpretability, creating a "black box" that business analysts cannot trust or utilize to design targeted marketing campaigns.

Therefore, this research aims to achieve the following core objectives:

1. Algorithmic Benchmarking: To systematically evaluate and compare the performance of diverse machine learning algorithms (Logistic Regression, SVM, Random Forest, XGBoost) in predicting telecom customer churn.
2. Imbalance Mitigation: To demonstrate the efficacy of data-level intervention techniques, specifically SMOTE, in resolving classification bias.
3. Model Interpretability: To deploy advanced explainable AI (XAI) frameworks, such as SHAP, to deconstruct the optimal predictive model and extract actionable business intelligence regarding the root causes of customer attrition.

II. LITERATURE REVIEW

The academic discourse surrounding customer relationship management (CRM) and churn prediction has evolved significantly, mirroring the advancements in computational statistics and artificial intelligence.

2.1 Early Statistical and Traditional Approaches

In the late 1990s and early 2000s, churn prediction was primarily the domain of traditional statisticians. Studies heavily relied on Logistic Regression and discrete-time Survival Analysis. Neslin et al. (2006) provided foundational work demonstrating how logistic frameworks could establish baseline probabilities of defection based on recency, frequency, and monetary (RFM) variables. While highly interpretable, these linear models inherently assumed monotonic relationships between independent variables and the log-odds of churn, an assumption frequently violated by the complex behavioral dynamics of real-world consumers.

2.2 The Transition to Machine Learning

As datasets grew in volume and dimensionality, researchers began exploring non-parametric machine learning techniques. Support Vector Machines (SVM) and standard Decision Trees gained traction. Coussement et al. (2020) highlighted that while SVMs utilizing non-linear kernels could map customer data into higher-dimensional spaces to find complex decision boundaries, they were computationally expensive and highly sensitive to unscaled data and outliers. Decision trees offered excellent interpretability but were notoriously prone to overfitting the training data, leading to poor generalization on unseen customer cohorts.

2.3 The Dominance of Ensemble and Boosting Methods

The contemporary gold standard for tabular churn data revolves around ensemble learning. Ensemble methods combine multiple weak learners to create a single, robust predictive model. Verbeke et al. (2011) demonstrated the superiority of Random Forests—an algorithm utilizing Bootstrap Aggregating (Bagging) to train multitude decision trees on random data subsets—in reducing model variance and resisting overfitting.

More recently, Gradient Boosting frameworks have dominated predictive analytics. XGBoost, formalized by Chen and Guestrin (2016), constructs trees sequentially, with each new tree explicitly designed to correct the residual errors of the previous ensemble. Its integration of and regularization to penalize complex tree structures has made it exceptionally effective for churn prediction, consistently outperforming deep learning models on structured, tabular datasets (Jain et al., 2021).

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716773

2.4 Addressing the Imbalance and Interpretability Deficit

The literature extensively documents the hazards of class imbalance. Chawla et al. (2002) introduced SMOTE, fundamentally altering how data scientists approach skewed distributions by synthetically generating minority class instances rather than merely duplicating them. Subsequent studies (Burez & Van den Poel, 2009) confirmed that algorithmic performance on recall metrics vastly improves when SMOTE is applied prior to training.

Finally, the literature has recently pivoted toward Explainable AI (XAI). Lundberg and Lee (2017) unified various methods under the SHAP (SHapley Additive exPlanations) framework, grounded in cooperative game theory. In the context of churn, SHAP allows researchers to move beyond aggregate feature importance to understand exactly how much a specific feature (e.g., a $20 increase in monthly charges) contributes to an individual customer's probability of churning, bridging the gap between data science and marketing strategy.

III. METHODOLOGY

The research methodology follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, ensuring a rigorous, reproducible pipeline from data ingestion to model evaluation.

3.1 Dataset Description

This study utilizes a robust, publicly available telecommunications dataset analogous to the widely benchmarked IBM Watson Telco Customer Churn dataset. The dataset comprises 7,043 distinct customer records, each defined by 21 heterogeneous attributes. The target variable is Churn, a binary indicator where represents a customer who terminated their contract within the preceding month, and represents a retained customer.

The independent predictor variables () are logically segmented into three distinct vectors:

- Demographic Vector: Gender, Age status (Senior Citizen), Marital Status (Partner), and Dependents.
- Account & Financial Vector: Customer Tenure (months), Contract Duration (Month-to-month, One year, Two year), Paperless Billing, Payment Method, Monthly Charges, and Total Charges.
- Service Infrastructure Vector: Phone Service, Multiple Lines, Internet Service Type (DSL, Fiber optic, None), Online Security, Device Protection, Tech Support, Streaming TV, and Streaming Movies.

3.2 Data Preprocessing and Feature Engineering

Raw telecommunications data is inherently messy and requires strict preprocessing to be suitable for mathematical optimization algorithms.

1. Missing Value Imputation: An initial audit revealed that the TotalCharges feature contained null values corresponding to customers with a tenure of zero months. Given the logical impossibility of total charges for brand new customers, these values were deterministic and were imputed with 0.0 to preserve the data row.

2. Encoding of Categorical Variables: Machine learning models process numerical matrices.

   - Binary attributes (e.g., Gender, PaperlessBilling) were subjected to Label Encoding ().
   - Nominal variables with more than two classes (e.g., PaymentMethod, InternetService) were processed using One-Hot Encoding. This generates sparse, binary columns for each category, preventing the algorithm from incorrectly inferring an ordinal hierarchy among unrelated categories.

3. Dimensionality and Scaling: Continuous variables (Tenure, MonthlyCharges, TotalCharges) exhibited vastly different numerical scales. Distance-based algorithms (SVM) and gradient-descent algorithms (Logistic Regression) converge inefficiently when features are unscaled. We applied standard scaling to transform these distributions:

where is the original feature value, is the mean of the feature column, and is the standard deviation.

3.3 Handling Class Imbalance via SMOTE
Exploratory Data Analysis (EDA) confirmed a severe class imbalance: 73.46% (5,174) of the cohort were retained customers, while only 26.54% (1,869) were churners. Training on this distribution would yield a biased model.

To rectify this, we utilized the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE operates in the feature space rather than the data space. For every minority instance, it identifies its -nearest minority neighbors. It then randomly selects one of these neighbors and generates a synthetic instance along the line segment connecting the two points.

Mathematically, a new synthetic sample is generated as:

where is a minority class sample, is one of its -nearest neighbors, and is a random number drawn from a uniform distribution .

Crucial Methodological Note: SMOTE was strictly applied only to the training data split after the train-test split was executed. Applying SMOTE to the entire dataset prior to splitting leads to severe data leakage, where synthetic data conceptually bleeds into the testing set, resulting in artificially inflated performance metrics that will inevitably fail in production.

IV. PROPOSED MACHINE LEARNING FRAMEWORK

The preprocessed and balanced training data was fed into four distinct algorithmic architectures to evaluate comparative predictive power.

4.1 Logistic Regression (Baseline Model)
Logistic Regression serves as the foundational baseline. It predicts the probability that a customer belongs to the churn class by fitting data to a logit function. It models the log-odds of the probability as a linear combination of the independent variables:
While computationally inexpensive and highly transparent, its capacity is limited by its inability to naturally capture non-linear interactions without exhaustive manual feature engineering (e.g., creating polynomial features).

4.2 Support Vector Machines (SVM)
SVM is a supervised learning model that seeks to construct a hyper-plane or set of hyper-planes in a high-dimensional space that best segregates the classes. Given the complexity of customer data, a linear hyperplane is usually insufficient. We utilized the Radial Basis Function (RBF) kernel, which implicitly maps the input vectors into a higher-dimensional space where linear separation is possible. The RBF kernel function is defined as:

SVMs are powerful but notoriously slow to train on large datasets and are highly sensitive to hyperparameter tuning (and).

4.3 Random Forest Classifier
Random Forest is a quintessential ensemble Bagging technique. It constructs a multitude of decision trees during the training phase. To ensure the trees are de-correlated (preventing them from all making the same errors), Random Forest utilizes two mechanisms:

1. Bootstrap Aggregation: Each tree is trained on a random sample of the data drawn with replacement.
2. Feature Randomness: At each node split, only a random subset of features is considered.

The final prediction is the mode of the classes predicted by the individual trees. It is highly robust against overfitting and handles tabular data exceptionally well.

4.4 eXtreme Gradient Boosting (XGBoost)
XGBoost represents the pinnacle of tree-based ensemble methods. Unlike Random Forest, which builds trees independently, XGBoost builds them sequentially (Boosting). Each new tree is engineered to predict the residuals (errors) of the preceding ensemble.

XGBoost defines a rigorous objective function that combines a convex loss function (measuring predictive accuracy) with a regularization term (penalizing complexity to prevent overfitting):

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716773

where is a differentiable convex loss function that measures the difference between the prediction and the target. penalizes the complexity of the tree. XGBoost uses a second-order Taylor expansion to approximate the loss function, enabling rapid and highly optimized gradient descent.

4.5 Evaluation Metrics Formulation

Evaluating models on imbalanced data requires metrics beyond simple accuracy. We define True Positives (TP, correctly identified churners), True Negatives (TN, correctly identified retained customers), False Positives (FP, loyal customers flagged as churners), and False Negatives (FN, churners missed by the model).

- Precision: The proportion of positive identifications that were actually correct.
- Recall (Sensitivity): The proportion of actual positives that were identified correctly. In churn prediction, maximizing Recall is generally prioritized to ensure no defector is missed.
- F1-Score: The harmonic mean of Precision and Recall, providing a single metric that balances both concerns.

V. RESULTS AND DISCUSSION

The models were rigorously trained using a 5-fold cross-validation strategy grid search to identify optimal hyperparameters. The final models were evaluated on the untouched 20% hold-out test set to simulate real-world performance.

5.1 Comparative Performance Analysis

The quantitative results of the model evaluation are synthesized in Table 1 below.

Table 1: Comprehensive Performance Evaluation of Predictive Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Log Loss
Logistic Regression	0.745	0.512	0.795	0.623	0.831	0.495
Support Vector Machine	0.768	0.548	0.762	0.637	0.825	0.482
Random Forest	0.815	0.635	0.710	0.670	0.852	0.435
XGBoost	0.842	0.674	0.785	0.725	0.884	0.395

Note: Metrics calculated on the 20% hold-out test set. Threshold for classification set at 0.5.

5.2 Analysis of Algorithmic Efficacy

The empirical results distinctly validate the hypothesis that advanced ensemble methods are superior for complex behavioral data like customer churn.

- The Baseline Trade-off: Logistic Regression exhibited the highest Recall (0.795), meaning it successfully flagged nearly 80% of all true churners. However, this came at the severe cost of Precision (0.512). The model essentially "over-predicted" churn due to its linear constraints combined with the SMOTE oversampling. In a business context, this high False Positive rate would lead to massive marketing waste, offering retention discounts to customers who had no intention of leaving.
- Non-Linear Limitations: The SVM showed marginal improvement over Logistic Regression in overall accuracy, but its computational cost and complex hyperparameter landscape did not yield sufficient predictive gains to justify its deployment over tree-based models.
- The Ensemble Advantage: XGBoost emerged as the unequivocally superior framework. It achieved the highest overall Accuracy (84.2%) and the lowest Log Loss (0.395), indicating high confidence in its probability estimates. Crucially, XGBoost optimized the Precision-Recall trade-off, achieving an F1-Score of 0.725. It successfully identified 78.5% of actual churners (Recall) while maintaining a highly respectable Precision (67.4%), ensuring that retention budgets are targeted efficiently. The ROC-AUC score of 0.884 demonstrates excellent discriminatory capacity across all classification thresholds.

5.3 Interpretability and SHAP Analysis

A critical component of this research is translating algorithmic output into actionable business strategy. To deconstruct the XGBoost model, we generated SHAP (SHapley Additive exPlanations) values. SHAP assigns each feature an importance value for a particular prediction.

Diagrammatic Description of SHAP Summary Plot findings: The SHAP analysis revealed a distinct hierarchy of churn drivers.

1. Contract Type (Month-to-Month): This feature exhibited the highest absolute SHAP value. The presence of a month-to-month contract aggressively pushes the model's output toward (Churn). The lack of a lock-in period eliminates the barrier to exit, making customers highly sensitive to competitor offers.
2. Tenure: Tenure demonstrated a strong negative correlation with churn probability. Lower tenure values (new customers) had high positive SHAP values (driving churn), while high tenure values (long-term customers) had strong negative SHAP values (driving retention). This indicates a "churn cliff" in the first 1-6 months of the customer lifecycle.
3. Monthly Charges & Fiber Optic: High monthly charges were a strong secondary indicator of churn. Interestingly, the interaction effect between InternetService_FiberOptic and MonthlyCharges was prominent. Customers paying premium rates for Fiber Optic services without subscribing to value-added services (like TechSupport or OnlineSecurity) were highly susceptible to defection, potentially indicating a perceived lack of value or dissatisfaction with base service reliability.

5.4 Business Implications and Strategic Implementation

The deployment of the optimized XGBoost model enables a paradigm shift from reactive to proactive customer management.

- Targeted Interventions: Instead of blanket marketing, the business can rank customers by their XGBoost predicted probability of churn. High-risk, high-value customers can be routed to specialized retention teams.
- Incentive Engineering: Based on the SHAP findings, retention offers should not merely be monetary discounts. For month-to-month customers with high churn probability, the primary strategic goal should be migrating them to a 1-year contract by offering a temporary price reduction or a free value-add (like Online Security, which the model identifies as a retention driver).

VI. CONCLUSION AND FUTURE WORK

Customer churn represents a profound threat to the economic stability of telecommunications providers. This research paper executed a rigorous, comparative analysis of predictive machine learning frameworks designed to forecast customer attrition. By carefully orchestrating a data pipeline that included robust preprocessing and the mitigation of class imbalance via SMOTE, we established an environment where complex algorithms could thrive.

The empirical results conclusively demonstrate that the eXtreme Gradient Boosting (XGBoost) algorithm is the optimal architecture for this domain, outperforming standard linear models, SVMs, and Random Forests. XGBoost's mathematical foundation—specifically its sequential error correction and integrated regularization—allowed it to achieve an unparalleled ROC-AUC of 0.884, effectively balancing the dual imperatives of identifying true defectors while minimizing false alarms. Furthermore, the integration of SHAP values demystified the ensemble's decision-making process, providing granular, actionable insights into the sociodemographic and financial drivers of churn.

Future Work: While this study successfully establishes a high-performance predictive baseline using static, tabular data, the frontier of CRM analytics lies in temporal and unstructured data integration.

1. Survival Analysis: Future research should transition from binary classification to continuous-time survival analysis using DeepSurv or Cox Proportional Hazards models, answering not just if a customer will churn, but exactly when.
2. Natural Language Processing (NLP): Integrating sentiment analysis algorithms (such as BERT or RoBERTa) to process unstructured customer service chat logs and call center transcripts will likely uncover nuanced, emotional precursors to churn that pure financial and demographic metrics cannot capture.
3. Graph Neural Networks (GNNs): Exploring network effects—how the defection of one customer influences the churn probability of socially connected customers—using GNNs represents a highly promising avenue for deep-tech telecom research.

REFERENCES

[1] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636. https://doi.org/10.1016/j.eswa.2008.05.027

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

[4] Coussement, K., Lessmann, S., & Verstraeten, G. (2020). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems, 95, 27-36.

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[6] Jain, H., Khunteta, A., & Srivastava, S. (2021). Churn prediction in telecommunication using machine learning algorithms. International Journal of Advanced Computer Science and Applications, 11(12), 1-8.

[7] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765-4774).

[8] Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 43(2), 204-211.

[9] Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc.

[10] Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(3), 2354-2364.