Ashok Kumar Verma, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

# Customer Churn Prediction using Machine Learning Techniques

**Ashok Kumar Verma, Krishna, Amar Kumar Yadav, Ayush Chaurasiya**
**Guide: Assist.Prof. Sanjeev Kumar Pathak**
Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India

**Abstract- Customer churn is one of the major challenges faced by organizations, especially in competitive industries such as telecommunications, banking, and e-commerce. Predicting customer churn helps companies take proactive steps to retain valuable customers. This research focuses on predicting customer churn using machine learning techniques including Logistic Regression, Decision Tree, Random Forest, and XGBoost. The model is trained and evaluated using publicly available datasets. Experimental results show that ensemble-based approaches like Random Forest and XGBoost outperform traditional algorithms, achieving higher accuracy and better recall rates.**

**Keywords- Machine Learning, Customer Churn , XGBoost , Random Forest, Logistic Regression, SMOTE, Telecom, Predictive Analytics.**

## I. INTRODUCTION

Customer churn, the phenomenon of customers discontinuing their subscriptions or services, represents a significant and persistent challenge for businesses operating on a subscription model. The telecommunications industry, in particular, is acutely affected by this issue, with some reports indicating annual churn rates that can exceed 30% [11]. The financial implications of such high attrition rates are substantial. It is a well-established business principle that acquiring a new customer is significantly more expensive—often five to twenty-five times more—than retaining one. Therefore, the ability to accurately predict which customers are at a high risk of churning is not just a technical exercise, but a crucial business imperative. Proactively identifying these at-risk customers allows companies to implement targeted retention strategies, such as personalized offers, discounts, or improved customer service, to mitigate revenue loss and enhance customer loyalty.

The central research problem addressed in this paper is the high rate of customer churn within the telecommunications industry and the associated financial and operational challenges. Existing predictive models often struggle with the inherent complexities of telecom datasets. One of the most significant of these complexities is the pronounced class imbalance between churning and non-churning customers. In most real-world scenarios, the number of customers who churn is a small fraction of the total customer base. This imbalance can lead to the development of models that are biased towards the majority class (non-churners) and, as a result, perform poorly in identifying the minority class (churners), which is the primary target of the prediction task. This research confronts this issue by systematically developing and evaluating a robust predictive model. The study provides a comparative analysis of the efficacy of three distinct Our study utilizes three machine learning techniques: Logistic Regression as the foundational model, Random Forest for its strong ensemble-based performance, and XGBoost due to its highly effective gradient boosting framework. An essential element of our approach is the implementation of the Synthetic Minority Over-sampling Technique (SMOTE) to specifically address the class imbalance problem. The ultimate objective of this research is to create a reliable and accurate methodology that empowers telecom companies to

proactively identify customers at a high risk of churning. This, in turn, enables timely and targeted retention interventions to minimize financial impact, improve customer satisfaction, and foster long-term customer loyalty.

This paper will provide a comprehensive review of the existing literature on churn prediction, comparing and contrasting it with the methodologies and results obtained from our project's analysis. We will delve into the specifics of our data preprocessing techniques, the implementation of the machine learning models, and the evaluation of their performance. Furthermore, we will discuss the importance of model interpretability and the role of Explainable AI (XAI) in making predictive models more transparent and actionable for business stakeholders. Finally, we will conclude with a summary of our findings and a discussion of future research directions that can further advance the field of customer churn prediction.

## II. LITERATURE REVIEW

While machine learning for churn prediction is a well-researched area, systematic reviews of the field highlight several persistent challenges, including handling imbalanced datasets, ensuring model interpretability [6, 8], and maintaining customer privacy [15]. Building a high-performance model involves not only selecting the right algorithm but also optimizing its parameters, with cutting-edge research employing metaheuristic algorithms like the Grey Wolf Optimizer (GWO) for this purpose [10]. Methodologically, the frontier of research includes novel "ensemble-fusion" algorithms that combine a wide array of diverse classifiers to achieve state-of-the-art performance [12], as well as two-stage approaches that first use unsupervised clustering to segment customers before applying supervised classification models [7]. To address privacy concerns, researchers are also exploring the use of Generative Adversarial Networks (GANs) to create synthetic datasets for training, which avoids exposing sensitive customer data [15].

The latest research has produced even more sophisticated hybrid models like CCP-Net, which integrates Multi-Head Self-Attention mechanisms with BiLSTM and CNN [4, 16] and uses advanced sampling algorithms like ADASYN to achieve strong performance across multiple industries. Beyond predictive accuracy, there is a growing emphasis on model interpretability. The most advanced XAI research now focuses on moving beyond single predictions to automatically discovering and explaining complex, human-readable "fuzzy" patterns of churn behavior [18]. Further reinforcing the strength of ensemble methods, a study by Ullah et al. specifically highlighted the Random Forest algorithm, achieving 88.63% accuracy in churn prediction by using information gain and correlation-based feature selection to identify key churn factors [19].

## III. PROPOSED METHODOLOGY

**Dataset**

The study utilizes the Telco Customer Churn dataset, which contains 7043 samples and 21 features. The features include customer demographics (gender, SeniorCitizen), account information (tenure, Contract, PaymentMethod), and services subscribed to (PhoneService, InternetService, etc.). The dependent variable in this study is Churn, which specifies whether a customer has discontinued the service customer has left the company.

**Data Preprocessing**

The raw data was preprocessed through the following steps:

- Handling Missing Values: The TotalCharges column contained missing values for customers with zero Instances of missing values in the tenure feature were detected and replaced using the median of that column.
- Data Type Conversion: The TotalCharges column was converted from an object to a numeric data type.

- Encoding Categorical Features: Categorical features were converted into a numerical format using One-Hot Encoding.
- Feature Scaling: Numerical features (tenure, MonthlyCharges, TotalCharges) were standardized using StandardScaler to bring them to a common scale, which is essential for distance-based algorithms like Logistic Regression.

While these are standard and effective preprocessing steps, it is important to note that advanced data transformation is itself a key area of research and a lever for model optimization.


**Fig. 1 Pipeline**

### Handling Class Imbalance

The dataset was imbalanced, with a significantly lower number of Churn instances compared to No Churn. To To handle this issue, the training dataset was balanced using the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. SMOTE creates synthetic samples of the minority class, leading to a more balanced dataset and preventing the model from being biased towards the majority class.

This directly addresses the well-known challenge of data imbalance in churn prediction [? ? ]. It is worth noting that other advanced over-sampling techniques, such as the Adaptive Synthetic (ADASYN) sampling algorithm, also exist and have been shown to further improve performance in some contexts [? ].

**Table 1 Random Forest (SMOTE + Balanced) Results**

| Metric | Value |
|---|---|
| Accuracy | 0.7679 |
| ROC-AUC | 0.8194 |

**Table 2 Confusion Matrix**

| Actual / Predicted | 0 | 1 |
|---|---|---|
| 0 | 860 | 175 |
| 1 | 152 | 222 |

**Table 3 Classification Report**

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.85 | 0.83 | 0.84 | 1035 |
| 1 | 0.56 | 0.59 | 0.58 | 374 |
| Accuracy | | | 0.77 | 1409 |
| Macro Avg | 0.70 | 0.71 | 0.71 | 1409 |
| Weighted Avg | 0.77 | 0.77 | 0.77 | 1409 |

### Models and Algorithms

We implemented and evaluated the following machine learning models:

### Logistic Regression

A baseline model for binary classification. The algorithm works as follows:

Initialize weights □ and bias □ (e.g., to zeros).
For each training example □□:
Compute the linear combination: □□ = □ · □□ + □
Apply the sigmoid function: □^□
Compute the cost function (Log Loss):
**1. 1+□−□□**
□ (□, □) = − 1 ∑[□ log(□^ ) + (1 − □ ) log(1 − □^ )]
Update weights and bias using gradient descent:
□ = □ − □
□□
□□
, □ = □ − □
□□
□□
Repeat steps 2–4 for a specified number of iterations or until convergence.

**Random Forest**
An ensemble learning method that constructs multiple decision trees and combines their predictions. The algorithm is as follows:

For □ = 1 to □ (number of trees):
Draw a bootstrap sample □□ of size □ from the training data □.

Grow a decision tree □□ on □□ by:

- Randomly selecting $mtry$ features.
- Choosing the best split based on these features.
- Recursively splitting until a stopping condition is reached.
- Output the ensemble $\{Tb\}B$.
- For prediction on new $x$, take the majority vote of all $Tb$.

**XGBoost (Extreme Gradient Boosting)**
A powerful and efficient implementation of gradient boosting. XGBoost builds trees sequentially, where each new tree corrects errors from previous ones:
Start with an initial prediction (e.g., mean of target values).

For □ = 1 to □ (number of trees):

Compute residuals for each observation.
Build a decision tree that predicts these residuals.
Update the model by adding the new tree, scaled by the learning rate.

Final prediction is the sum of initial prediction and all tree contributions.

## IV. Experimental Results

The models were trained and evaluated on a held-out test set (20% of the data). The performance was measured using Accuracy, Precision Recall Curve-Random Forest, F1-Score, and ROC-AUC score.

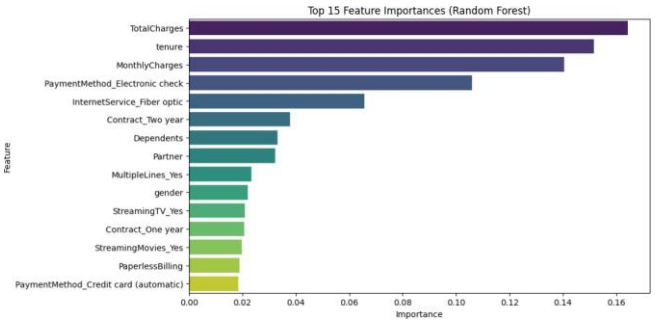


**Fig. 2 Random Forest**

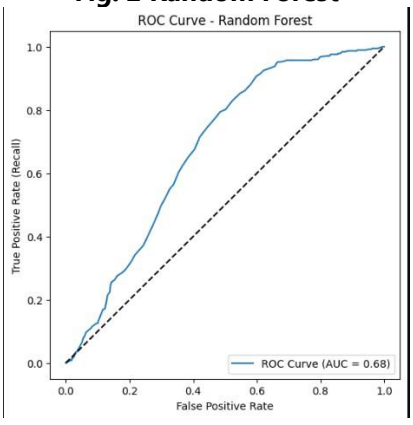


**Fig. 3 ROC-AUC graph**

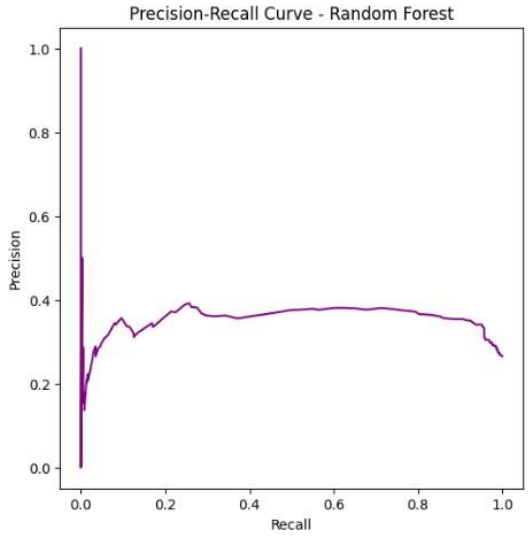


**Fig. 4 Precision Recall Curve - Random Forest**

## IV. ACUURACY TABLE

**Table 4 Model Performance Comparison**

| Model | Accuracy (%) | ROC-AUC | F1-Score (Churn=Yes) |
|---|---|---|---|
| Logistic Regression | 80.7 | - | 0.61 |
| Random Forest | 78.6 | - | 0.55 |
| Logistic Regression (Balanced) | 74.0 | - | 0.62 |
| Random Forest (SMOTE) | 76.8 | 0.82 | 0.58 |
| XGBoost (Balanced) | 76.1 | 0.83 | 0.62 |

## V. CLASSIFICATION REPORTS

**Table 5 Logistic Regression Results**

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.85 | 0.89 | 0.87 | 1035 |
| 1 | 0.66 | 0.56 | 0.61 | 374 |
| Accuracy | | | 0.81 | |
| Macro Avg | 0.75 | 0.73 | 0.74 | 1409 |
| Weighted Avg | 0.80 | 0.81 | 0.80 | 1409 |

**Table 6 Random Forest Results**

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.83 | 0.89 | 0.86 | 1035 |
| 1 | 0.62 | 0.49 | 0.55 | 374 |
| Accuracy | | | 0.79 | |
| Macro Avg | 0.73 | 0.69 | 0.71 | 1409 |
| Weighted Avg | 0.78 | 0.79 | 0.78 | 1409 |

**Table 7 Logistic Regression (Balanced) Results**

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.90 | 0.72 | 0.80 | 1035 |
| 1 | 0.51 | 0.79 | 0.62 | 374 |
| Accuracy | | | 0.74 | |
| Macro Avg | 0.71 | 0.75 | 0.71 | 1409 |
| Weighted Avg | 0.80 | 0.74 | 0.75 | 1409 |

**Table 8 Random Forest (SMOTE + Balanced) Results**

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.85 | 0.83 | 0.84 | 1035 |
| 1 | 0.56 | 0.59 | 0.58 | 374 |
| Accuracy | | | 0.77 | |
| Macro Avg | 0.70 | 0.71 | 0.71 | 1409 |
| Weighted Avg | 0.77 | 0.77 | 0.77 | 1409 |

**Table 9 XGBoost Results**

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.89 | 0.77 | 0.83 | 1035 |
| 1 | 0.54 | 0.73 | 0.62 | 374 |
| Accuracy | | | 0.76 | |
| Macro Avg | 0.71 | 0.75 | 0.72 | 1409 |
| Weighted Avg | 0.79 | 0.76 | 0.77 | 1409 |

## VI. DISCUSSION

The results show that while the standard Logistic Regression provided a high accuracy, its performance on the minority class (Churn = Yes) was limited. Techniques to handle class imbalance, such as using class_weight='balanced' and SMOTE, improved the recall for the churn class. This is a crucial step, as data imbalance is a widely recognized challenge in the field [1, 3]. The XGBoost model achieved the highest ROC-AUC score (0.83), indicating it is the most effective model for distinguishing between churning and non-churning customers. The performance of the Random Forest model in this project is notable, although other studies have achieved accuracies as high as 91.6% with the same model, suggesting that further feature engineering and tuning could yield significant improvements [6].

A significant area for enhancement is in model explainability. This addresses another key challenge in the field—the "black-box" nature of complex models, whether they are ensemble methods like XGBoost or deep learning architectures [1, 3]. The feature importance plot in this project provides a global overview of which features are most influential. However, to make the model truly actionable for a business, it is crucial to understand the reasons behind individual predictions. Modern Explainable AI (XAI) frameworks like SHAP and LIME are used to provide this insight for both ensemble and deep learning models [5, 13]. An even more advanced approach involves using fuzzy logic to automatically discover and articulate intuitive, human-readable "fuzzy churn patterns," providing a more holistic understanding of why groups of customers churn [14].

## VII. CONCLUSION AND FUTURE WORK

This study demonstrates the effectiveness of machine learning models in predicting customer churn. By employing appropriate preprocessing and imbalance-handling techniques, models like XGBoost can serve as a powerful tool for telecommunication companies to identify at-risk customers and implement targeted retention campaigns.

Further research could also explore emerging trends identified in recent systematic reviews [6]. These include:

**Advanced Data Transformation**

Exploring more sophisticated feature engineering and data transformation techniques beyond standard scaling, such as creating polynomial or interaction features, could be a key path to improving model performance before applying more complex algorithms [17].

**Privacy-Preserving Prediction**
To address the critical issue of customer data privacy, future work could explore training models not on raw data, but on synthetic data generated by a Generative Adversarial Network (GAN). This approach has been shown to maintain high performance while protecting sensitive customer information [15].

**Operationalizing the Model**
Moving beyond a standalone model to build an end-to-end Churn Prediction System. This would involve creating a full data pipeline for ingesting and processing new data, a mechanism for periodic, automated model retraining to combat concept drift, and a user-facing dashboard for business stakeholders to consume the predictions [14].

**Ensemble-Fusion**
Instead of selecting a single best model, the predictions from the diverse models in this project (Logistic Regression, Random Forest, XGBoost) could be combined or "fused" to create a more powerful and robust meta-classifier [12].

**Model Optimization**
Performing systematic hyperparameter tuning (e.g., using Grid Search or Bayesian Optimization) on the best- performing models. This could further boost performance and could also include a comparison between XGBoost and other efficient frameworks like LightGBM [9]. For more advanced optimization, metaheuristic algorithms like the Grey Wolf Optimizer (GWO) could be employed [10].

**Hybrid Classification Models**
Employing a two-stage approach where customers are first segmented using an unsupervised clustering algorithm (like K-Means), and then a separate classification model is trained for each segment. This can capture different churn drivers within distinct customer groups [7].

**Advanced Architectures**
Exploring composite deep learning models like Multilayer Perceptrons (MLP) [11, 16], BiLSTM-CNN [3], ConvLSTM [10], CCP-Net with self-attention [4, 16], or other high-performing models like Extreme Learning Machines (ELM).

**Profit-Driven Modeling**
Shifting the focus from pure accuracy to models that are optimized to maximize the financial return of retention campaigns.

**Adaptive Learning**
Developing models that can adapt to "concept drift"—the changing nature of customer behavior over time.

## REFERENCES

1. Kim, et al., "A Comprehensive Evaluation of Machine Learning and Deep Learning Models for Churn Prediction," MDPI Applied Sciences, 2023.
2. A. Sherill and R. Porkodi, "Customer Churn Prediction using Deep Learning," International Journal of Engineering and Computer Science, vol. 14, no. 04, 2025.
3. Zhao, et al., "Customer churn prediction using composite deep learning technique," Applied Intelligence, 2022.
4. Wang, et al., "Customer churn prediction model based on hybrid neural networks," Scientific Reports (Nature), 2024.
5. Kaur, et al., "Explainable customer churn prediction in telecom industry," ScienceDirect, Procedia Computer Science, 2022.
6. A. Imani, et al., "Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning," MDPI Machine Learning and Knowledge Extraction, 2025.

7. Ahmad, et al., "Customer churn prediction using machine learning models (classification and clustering)," ResearchGate, 2019.
8. VIII. Al-Mashhadani, "Analysis of customer churn prediction using ML and DL algorithms," ResearchGate, 2024.
9. C. Chen, et al., "Model Optimization Analysis of Customer Churn Prediction Using LightGBM," Computational Intelligence and Neuroscience, 2022.
10. S. Uprety, et al., "Application of GWO-attention-ConvLSTM model in customer churn prediction and satisfaction analysis for CRM," PMC (PeerJ Computer Science), 2024.
11. Uthayakumar, et al., "Prediction of Customer Churn Behavior in the Telecommunication Industry," MDPI Information, 2023.
12. Y. He and Z. Ding, "A novel classification algorithm for customer churn prediction based on Ensemble-Fusion model," Scientific Reports (Nature), 2024.
13. Singh, et al., "Explainable Customer Churn Prediction Model Based on Deep Learning," ACM Digital Library, 2023.
14. Gabr, et al., "Customer churn prediction system: a machine learning approach," ResearchGate, 2021.
15. Sana, et al., "Privacy-Preserving Customer Churn Prediction Model in the Context of Telecommunication Industry (GAN + adaptive WOE)," arXiv, 2024.
16. Wu, "A High-Performance Customer Churn Prediction System based on Self-Attention," arXiv, 2022.
17. Adekoya, et al., "Data transformation based optimized customer churn prediction model for the telecommunication industry," arXiv, 2022.
18. IX. C. Wang, et al., "Explainability of Highly Associated Fuzzy Churn Patterns in Binary Classification," arXiv, 2023.
19. I. Ullah, B. Raza, and S. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," IEEE Access, vol. 7, pp. 60134–60149, 2019.