Proceedings of the 3rd International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2025.DL27964

Machine Learning Strategies for Reducing Customer Churn in the Telecommunications Industry

Qunbo Meng

School of Arts and Sciences, Rutgers University, USA
Qm51@scarletmail.rutgers.edu

Abstract. Customer churn is a key issue in the telecom industry and has a significant impact on the revenue of related companies. For telecom companies, establishing an accurate customer churn prediction model is crucial for retaining existing customers. Therefore, it is necessary to develop reliable prediction models to effectively manage and reduce churn rates. This study explores how to use machine learning technology to predict customer churn based on various customer characteristics and usage patterns, and its application in accurately predicting customer churn based on demographic data, service usage behavior, and billing information. Through systematic analysis of model performance, the advantages and disadvantages of each method are highlighted, with particular emphasis on the superior predictive power of the integrated method when dealing with imbalanced data sets. This proves the potential of these methods in accurately predicting customer churn behavior, thereby providing relevant companies with methods to retain customers. In addition, the study explored the economic implications of the accuracy of attrition predictions and advocated the use of lower-cost methods to ensure the rational allocation of resources.

Keywords: Customer churn, Telecom industry, Machine learning, Logistic regression, Decision trees.

1. Introduction

Due to market demand and intense competition among different companies, customer churn has become a major challenge facing the telecommunications industry. Customer churn refers to the number of existing customers lost over a certain period due to various reasons. It enables companies to understand customer satisfaction and loyalty and identify potential changes in company profits. It directly impacts revenue streams and operational stability. Research consistently shows that the cost-effectiveness of retaining loyal customers may be several times higher than attracting new users, making preventing customer churn a top priority for telecommunications companies. In recent years, the availability of large-scale data such as demographic information and customer service feedback has provided new impetus for building predictive models. This data reveals patterns of customer dissatisfaction or behavioral changes, which may indicate customer churn risk.

Unlike traditional regression models, machine learning methods can capture nonlinear relationships and hidden correlations, thereby enhancing predictive capabilities. Techniques such as ensemble learning offer flexible and scalable solutions for companies to address challenges like data

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.DL27964

imbalance. Therefore, the growing importance of customer churn management necessitates the use of more advanced methods that integrate advanced data preprocessing, scalable modeling techniques, and interpretability.

Considering this evolving landscape, this study emphasizes the critical role of machine learning in addressing the dual challenges of predictive accuracy and actionable decision support in customer churn prediction.

In the following content, we will introduce some related work to give readers a specific understanding of the development in this field. In addition, we also explained some methods, including Data Preprocessing and Feature Engineering, Model Development and Training, Evaluation and Interpretability and Scalability.

2. Literature review

In recent decades, customer churn prediction research in the telecommunications industry has evolved from simple statistical models to advanced machine learning and hybrid methods. Early research relied heavily on logistic regression because it was easy to implement and interpret, helping telecommunications companies understand the relationship between customer churn probabilities [1].

However, the emergence of big data and increasingly complex behavioral patterns has exposed the limitations of such linear models in capturing nonlinear feature interactions. To address these shortcomings, Huang and Kechadi proposed a hybrid learning architecture that combines logistic regression, decision trees, and neural networks, leveraging the interpretability of linear models and the flexibility of nonlinear learning [1]. Their findings highlight the advantages of decision trees in identifying important customer churn levels while maintaining model accuracy and transparency. Building on this, Ahmad et al. incorporated ensemble algorithms such as random forests and gradient boosting into a big data platform, with results showing significantly better performance than the single model baseline and emphasizing the impact of engineering variables [2]. Subsequent developments introduced multi-stage and optimization-enhanced methods. Idris et al. combined filter-based feature selection with ensemble classifiers to achieve scalability and higher predictive accuracy on large telecommunications datasets [3]. Amin et al. advanced this approach by combining genetic algorithms with decision trees to optimize feature subsets to more effectively model complex nonlinear customer churn behavior [4]. Lalwani et al. expanded on these ideas, designing a six-stage process incorporating feature selection driven by a gravity search algorithm, significantly enhancing the performance of advanced learners such as AdaBoost and XGBoost [5].

A comparative analysis by Vafeiadis et al. confirmed that ensemble-based frameworks consistently outperform individual models [6], particularly in datasets with heterogeneity and class imbalance. In addition to traditional classifiers, hybrid neural network architectures have also garnered increasing attention. Tsai combined backpropagation networks with auxiliary architectures to better capture subtle changes in user behavior that traditional models might overlook [7]. De and Prabu et al. provided a more comprehensive review of recent advancements, highlighting areas that remain under-explored [8], such as leveraging interaction-level data and validating models across different geographic and demographic groups. Kernel-based learning has also led to methodological improvements; for example, Ly et al. applied kernel-optimized support vectors to enhance the stability of decision boundaries in complex feature spaces [9]. From a commercial application perspective, Saleh analyzed the drivers of customer churn in the Danish telecommunications industry, directly linking feature importance to actionable CRM strategies, thereby illustrating the practical necessity of interpretable outputs [10].

DOI: 10.54254/2753-8818/2025.DL27964

Overall, the literature reflects a shift toward hybrid and ensemble methods that incorporate various advanced techniques and are reinforced by complex feature engineering and optimization strategies.

3. Methodology

The methodological framework used in this study consists of four interrelated stages: data collection, data preprocessing, model building, and performance evaluation. The dataset used integrates customer demographic information, detailed service usage patterns, billing history, and interaction records, and appropriate measures have been taken to protect customer privacy.

3.1. Data preprocessing and feature engineering

Raw data often contains inconsistencies and missing values, which I addressed using statistical interpolation methods. To avoid excessive bias, I standardized continuous features and encoded categorical variables using a target encoding scheme.

To more effectively capture customer behavior, I constructed derived features such as service tenure trends, aggregated billing history, and call usage changes. To mitigate the category imbalance problem—an inherent challenge in customer churn prediction—I used the SMOTE method and a complementary under sampling strategy to maintain distribution balance.

3.2. Model development and training

I also used a hierarchical modeling approach, based on logistic regression because of its interpretability and ability to establish benchmark performance. Then, we applied decision trees to reveal rule-based structures and highlight key predictors of customer churn.

To further improve prediction accuracy, I used ensemble models—namely, random forests and GBM—to minimize variance and bias using multiple learners. I performed hyperparameter optimization using grid search and cross-validation, adjusting parameters such as learning rate, maximum depth, and regularization strength.

3.3. Evaluation and interpretability

Model performance evaluation includes not only traditional accuracy, but also recall, precision, F1 score, and AUC methods to ensure balanced evaluation even under imbalanced conditions. In addition, the model incorporates cost metrics to reflect the impact of the results, prioritizing the detection of high-value customer churners whose retention has a significant financial impact.

To ensure stability, k-fold cross-validation is applied across multiple data partitions. For ease of interpretation, the model uses the SHAP method to gain insight into how individual features affect customer churn prediction. This step bridges the gap between predictive performance and management decisions.

3.4. Scalability

Given the ever-increasing scale of telecommunications data, scalability has also been fully developed. The process is designed to be compatible with distributed computing environments, enabling real-time scoring and integration into customer relationship management systems. This ensures that the method's contributions are not only academic but also practical, making it suitable

DOI: 10.54254/2753-8818/2025.DL27964

for industry adoption. For example, it can trigger personalized retention campaigns or allocate resources to high-risk groups.

In this way, the methodological contributions go beyond theoretical exploration, demonstrating clear applicability in large-scale industry deployments while maintaining efficiency, accuracy, and operational relevance.

4. Conclusion

This study combines advanced data preprocessing techniques with high-performance machine learning algorithms to construct a comprehensive framework for predicting customer churn in the telecommunications industry.

The experimental results show that although logistic regression is easy to interpret, ensemble-based methods have better predictive accuracy, especially in solving class imbalance and capturing complex behavioral patterns. The use of methods such as SMOTE and feature targeting significantly improves the stability and generalization ability of the model. In addition, the evaluation process not only considers accuracy, but also comprehensively considers recall, precision, F1 score, and AUC to ensure a more balanced and practical evaluation of model effectiveness. Research consistently shows that the cost-effectiveness of retaining loyal customers may be several times higher than attracting new users, making preventing customer churn a top priority for telecommunications companies. In recent years, the availability of large-scale data such as demographic information and customer service feedback has provided new impetus for building predictive models. This data reveals patterns of customer dissatisfaction or behavioral changes, which may indicate customer churn risk.

These findings highlight the importance of combining technical performance with business priorities and emphasize the identification of high-value customers who are likely to churn. From a company perspective, this provides telecom operators with a practical method that directly supports targeted customer retention strategies, optimizes marketing investments, and improves customer satisfaction.

This proves the potential of these methods in accurately predicting customer churn behavior, thereby providing relevant companies with methods to retain customers. In addition, the study explored the economic implications of the accuracy of attrition predictions and advocated the use of lower-cost methods to ensure the rational allocation of resources

Future work will explore real-time data processing, the integration of advanced deep learning architectures, and explainable AI technologies to further improve the accuracy and interpretability of predictions, thereby enabling more adaptive and proactive customer churn management in the market environment.

References

[1] Huang, Y., & Kechadi, M.-T. (2012). An effective hybrid learning system for telecommunication churn prediction. Expert Systems with Applications, 39(12), 11792–11803.

[2] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 28.

[3] Idris, A., Rizwan, M., & Khan, A. (2018). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. Computers & Electrical Engineering, 67, 341–355.

[4] Amin, A., Anwar, A., Adnan, A., Nawaz, M., Howard, M., & Qadir, S. (2020). Customer churn prediction in the telecommunication sector using a rough set approach. Telecommunication Systems, 76, 565–579.

DOI: 10.54254/2753-8818/2025.DL27964

[5] Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: A machine learning approach. Computing, 104(8), 1781–1803.
[6] Vafeiadis, T., Diamantaras, K., Sarigiannidis, G., & Chatzisavvas, K. (2015). A comparison of machine learning techniques for customer churn prediction. Applied Artificial Intelligence, 29(5), 1–17.
[7] Tsai, C.-F., & Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. Expert Systems with Applications, 36(10), 12547–12553.
[8] De, S., & Prabu, P. (2022). Predicting customer churn: A systematic literature review. Journal of Discrete Mathematical Sciences and Cryptography, 25(7), 1965–1985.
[9] Ly, T. V., Nguyen, H. D., Nguyen, T. T., & Nguyen, M. T. (2022). Churn prediction in telecommunication industry using kernel support vector machines. PLOS ONE, 17(5), e0267935.
[10] Saleh, S. (2023). Customer retention and churn prediction in the Danish telecommunication industry. SN Applied Sciences, 5(1), 28.