Machine Learning-Based Prediction of Telecom Customer Churn: Comparative Model Analysis

Yixiang Zhang

University of Illinois Urbana-Champaign, Illinois, the United States

dithosw@gmail.com

Abstract. In the increasingly competitive telecommunications market, customer churn has become a key challenge affecting the profitability and sustainable growth of enterprises. This study aims to conduct an empirical analysis to compare the performance of various machine learning models in predicting customer churn in the telecommunications industry. In this study, Python and related machine learning libraries were utilized to build a prediction process, and strict performance evaluations were conducted on three models: Logistic Regression, Random Forest, and Extreme Gradient Boost. The evaluation indicators include accuracy rate, precision rate, recall rate, F1 score, and the area under the receiver operating characteristic curve (ROC-AUC). The research results show that the optimized XGBoost model exhibits the best performance in all evaluation indicators, demonstrating its outstanding ability to handle such classification problems. In addition, by analyzing the feature importance of the XGBoost model, this study identified the key drivers influencing customer churn, among which contract type, customer online duration, and monthly fee are the most significant predictors. These findings not only provide telecom operators with high-precision churn warning tools but also offer data-driven decision support for them to formulate precise and effective customer retention strategies.

Keywords: Customer Churn; Machine Learning; Python; Telecommunications; XGBoost.

1. Introduction

In today's highly saturated and competitive market environment, Customer Relationship Management (CRM) has become a core strategy for business survival and growth. Among all customer-related business metrics, customer churn is particularly critical [1]. The customer churn rate refers to the proportion of customers who cease using a company's products or services within a specific period. This metric directly impacts a company's revenue stability and long-term growth potential [2]. Extensive research and business practice demonstrate that retaining an existing customer costs significantly less than acquiring a new one. For instance, studies indicate that reducing the customer churn rate by 5% can increase corporate profits by 25% to 85% [3]. For the telecommunications industry, customer churn poses particularly severe challenges. Characterized by a subscription-based business model and near-saturated market penetration, competition among major operators primarily centers on existing customer bases [4]. Statistics indicate that the telecom industry's average annual customer churn rate ranks among the highest across major sectors, sometimes exceeding 30% [5]. Therefore, establishing a mechanism that accurately identifies potential churn customers and understands their reasons for leaving is not only a necessity for telecom operators to enhance profitability but also a strategic imperative to maintain core competitiveness [6].

Traditionally, customer retention strategies have been largely reactive, involving remedial actions only after customers submit cancellation requests—often too late. With the advancement of big data technology and artificial intelligence, machine learning has revolutionized customer churn management, propelling enterprises from passive response to proactive prediction [7]. By analyzing vast amounts of historical customer data—encompassing demographics, account details, service usage records, and consumption behavior—machine learning models can learn and identify complex, nonlinear patterns underlying customer churn [2]. This predictive capability enables companies to calculate each customer's probability of future churn. Based on this risk score, businesses can identify high-risk customer segments and implement precise, personalized retention interventions before they decide to leave—such as offering customized package deals, providing value-added services, or

initiating proactive customer care communications [8]. This data-driven proactive retention strategy not only significantly improves retention success rates but also optimizes marketing resource allocation. It concentrates limited budgets on customers most likely to churn and with the highest retention value, thereby maximizing the effectiveness of customer relationship management [3].

Existing literature employs various techniques ranging from traditional statistical models to advanced deep learning methods. For instance, Ahmad, Jafar, & Aljoumaa achieved an impressive AUC of 93.3% on a large telecom company's real-world big data platform [9]. This result was attained by constructing innovative feature engineering methods (such as social network analysis features) and leveraging advanced algorithms like XGBoost, demonstrating the high level of churn prediction achievable with rich data sources and robust computational platforms [9]. Concurrently, academia increasingly emphasizes model interpretability. A black-box model with high predictive accuracy but no decision logic explanation holds limited commercial value, as it fails to inform decision-makers about why customers churn, hindering effective intervention strategies. Addressing this challenge, De Caigny, Coussement, & De Bock proposed the Logit Leaf Model (LLM), a hybrid algorithm that ingeniously combines the strengths of decision trees (excelling at capturing variable interactions) and logistic regression (excelling at handling linear relationships) [10]. This approach aims to balance predictive performance with model interpretability. The structure of this paper is as follows: Section 2 details the dataset, data preprocessing methods, selected machine learning models, and evaluation protocols. Section 3 presents comprehensive experimental results, including comparative analysis of model performance, identification of key features, and in-depth exploration of their business implications. Section 4 summarizes the findings, discusses limitations, and outlines future research directions.

2. Data & Methods

2.1 Dataset and Preprocessing

The data source employed in this study is the Telco Customer Churn dataset, originally provided by IBM Analytics. The dataset features a clear structure, comprising 7,043 customer records (rows) and 21 features (columns) describing customer attributes [11]. The target variable, Churn, is a binary categorical variable representing whether a customer leaves the service. The preliminary analysis highlighted class imbalance, as about 73.5% of the customers were classified as non-churning customers and about 26.5% were classified as customers who quit [4], prompting the inclusion of multi-dimensional evaluation metrics. Table 1 summarizes several important features in the dataset, including data types and example values.

To ensure data quality and meet machine learning model input requirements, all operations were performed in a Python environment, primarily relying on Pandas and NumPy for data manipulation, and utilizing the Scikit-learn library for feature transformation [2]. The TotalCharges column was incorrectly identified as an object type. This occurred because the field contained empty strings for new customers with a tenure of 0. This study first used the pd.to_numeric function to force the column into a numeric type, setting unconvertible empty strings as missing values (NaN). Analysis revealed that 11 samples had their TotalCharges become missing, and all these samples had a tenure of 0 [8]. Given that these customers had not yet incurred total charges, filling these missing values with 0 was a logical approach. Beyond this, no other significant missing values required handling in the dataset [8]. Binary categorical features (e.g., Yes/No) were mapped to 1 and 0. For nominal variables with three or more categories (e.g., Contract, InternetService, PaymentMethod), one-hot encoding is applied. This method converts each category into a new binary feature column, preventing the model from erroneously assigning ordinal relationships between categories. Continuous numerical features like tenure, MonthlyCharges, and TotalCharges undergo standardization. Using Scikit-learn's StandardScaler, each feature's data is transformed into a distribution with a mean of 0 and a standard deviation of 1. This step is crucial for models sensitive to feature scaling (e.g., logistic regression and support vector machines), eliminating the influence of units and ensuring all features carry equal

weight during model training [12]. To objectively evaluate model generalization, the preprocessed dataset was randomly split into training and test sets at an 80:20 ratio. All models were trained exclusively on the training set, while final performance evaluation was conducted on the unseen test set to ensure impartial and reliable assessment [4].

Table 1. Dataset Feature Description.

Feature Name	Description	Data Type	Example Values
Gender	Customer Gender	Categorical	Male, Female
SeniorCitizen	Senior citizen (65+)	Binary	1, 0
Partner	Has a partner?	Binary	Yes, No
Dependents	Has dependents?	Binary	Yes, No
Tenure	Customer tenure (months)	Numeric	1, 34, 72
Contract	Contract Type	Categorical	Month-to-month, One year, Two-year
Payment Method	Payment Method	Categorical	e.h., Electronic check, Mailed check
Internet Service	Type of Internet Service	Categorical	DSL, Fiber optic, No
Online Security	Subscribes to online security services?	Binary	Yes, No, No internet service
TechSupport	Subscribe to technical support services?	Binary	Yes, No, No internet service
Monthly Charges	Monthly billing charges	Numeric	29.85, 56.95, 108.15
Total Charges	Total bill charges	Numeric	29.85, 1889.5, 8684.8
Churn	Churn status (target variable)	Binary	Yes, No

2.2 Machine Learning Models

This study selected three classification algorithms that are widely applied and representative in both academia and industry. Logistic regression serves as the baseline model for this comparative study. It is a classical generalized linear model that maps the output of linear regression to the interval (0, 1) via the sigmoid function, yielding the probability of a sample belonging to a specific class. Its advantages include simplicity, computational efficiency, fast training, and highly interpretable model coefficients that intuitively reflect the direction and strength of each feature's influence on prediction outcomes [4].

Random Forest is a representative algorithm of the Bagging concept in ensemble learning, renowned for its robust performance and resilience against overfitting [13]. It operates by constructing multiple decision trees using Bootstrap Aggregating (Bootstrap Sampling). During each tree's growth, a double randomness is introduced: not only are the training samples randomly drawn, but the features selected for node splitting are also randomly sampled from a subset of all features. The final prediction is determined by the majority vote of all decision trees. This process reduces the correlation between individual trees, effectively lowering model variance while enhancing overall

stability and accuracy. Additionally, Random Forest has the natural ability to output feature importance rankings. XGBoost is an efficient, flexible, and portable implementation of gradient boosting algorithms, ranking among the most powerful tools in contemporary machine learning. It consistently delivers outstanding performance in data science competitions and real-world applications. Unlike random forests that build trees in parallel, XGBoost employs a sequential approach, iteratively constructing a series of decision trees. Each new tree is trained to minimize the residuals (i.e., errors) from the previous round by optimizing along the negative gradient of the loss function, progressively correcting errors to achieve high prediction accuracy. XGBoost also incorporates several key optimizations, including regularization terms to prevent overfitting and support for parallel computation to enhance efficiency [14].

2.3 Evaluation Metrics and Protocol

Given the imbalanced nature of the dataset, relying solely on accuracy metrics may be misleading. For instance, a model predicting all customers as non-churn could achieve 73.5% accuracy, yet remain useless for identifying churners. Therefore, this study employs a comprehensive set of evaluation metrics that holistically reflect model performance on imbalanced data.

Confusion matrix: Serves as the foundation for all classification metrics, comprising four core components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
Accuracy: The proportion of correctly predicted samples out of the total samples.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

Precision: The proportion of customers who actually churn among all customers predicted by the model to churn. A high precision indicates less waste in retention costs.

$$\text{Precision}=\frac{TP}{TP+FP}$$

Recall: The proportion of all truly churned customers successfully identified by the model. A high recall rate indicates coverage of more potential churn customers.

$$\text{Recall}=\frac{TP}{TP+FN}$$

F1 Score: As the harmonic mean of precision and recall, it provides a single metric balancing both.

$$F_1=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

AUC (Area Under the ROC Curve): Assesses the model's skill in distinguishing positive from negative samples across all classification thresholds. Values can range from 0.5 to 1, with larger values indicating better discrimination. AUC is well-suited to evaluate performance on imbalanced datasets.

3. Results

3.1 Comparative Performance of Models

To fairly evaluate the predictive capabilities of logistic regression, random forest, and XGBoost, this study calculated the aforementioned performance metrics on the reserved 20% test set. All models utilized their final versions trained on the 80% training set. Detailed performance comparison results are shown in Table 2.

The comparison results in Table 2 reveal several key points:

- Benchmark Model Performance: The baseline logistic regression model achieved an accuracy of 80.4% and an AUC value of 0.843, indicating that even a simple linear model can capture some predictive signals on this dataset. This performance level aligns with reports of logistic regression on this dataset. Its recall (0.548) is relatively low, meaning the model missed nearly half of the actual churn customers.

- Superiority of Ensemble Models: Compared to logistic regression, two ensemble learning models—Random Forest and XGBoost—demonstrated stronger performance across multiple key metrics, particularly in AUC, a core indicator measuring overall model discriminative capability. XGBoost achieved the highest AUC value of 0.857 among the three models.
- XGBoost's Leading Performance: Between Random Forest and XGBoost, XGBoost outperformed in accuracy, precision, F1 score, and AUC. This aligns with XGBoost's established reputation as an advanced implementation of gradient boosting algorithms. While Random Forest reduces variance through bagging, XGBoost progressively minimizes bias via boosting. Its sequential learning approach allows it to focus more intently on correcting errors from previous rounds, achieving higher precision across many tasks.

Table 2. Machine Learning Model Performance Comparison.

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.804	0.651	0.548	0.595	0.843
Random Forest	0.795	0.635	0.579	0.606	0.849
XGBoost	0.812	0.673	0.585	0.626	0.857

3.2 Feature Importance Analysis of XGBoost

To investigate the primary factors driving customer churn, this study leveraged the built-in feature importance evaluation function within the top-performing XGBoost model. This function quantifies each feature's contribution to prediction outcomes by calculating the total number of times it was used for node splits or the total gain across all decision trees in the model. The following figure displays the top 15 influential features and their relative importance.

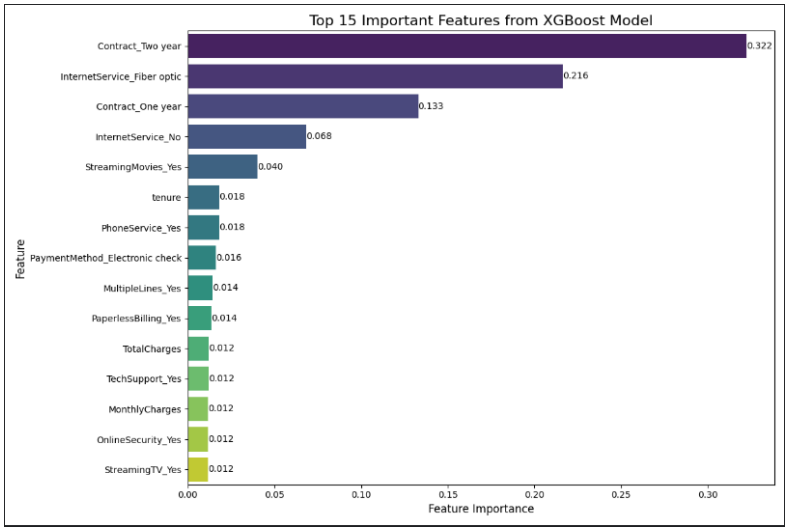

Figure 1. Top 15 Important Features Output by the XGBoost Model.

The analysis reveals that the most influential features for predicting customer churn are highly concentrated in contract details and service selections, rather than personal demographic characteristics. The top-ranked key features and their relative importance are as follows:

1. Contract Type: A two-year contract (importance score 0.322) is the most critical factor for stabilizing customer relationships. One-year contract (0.133). Customers on monthly payment plans constitute the group with the highest churn risk.
2. Internet Service Type: Whether using Fiber Optic (InternetService_Fiber optic) is the second most important predictive feature (0.216). This typically indicates that customers using fiber

optic services may exhibit higher churn tendencies due to price sensitivity or service stability concerns.

3. Tenure: Tenure ranked sixth in importance (0.018), indicating that relatively stable long-term customers have lower churn probability, while newer customers with shorter tenure are more prone to churn.
4. Payment Method: The importance score for Electronic check (PaymentMethod_Electronic check) is 0.016, indicating that customers using this payment method face a higher churn risk.
5. Total Charges & Monthly Charges: Both features scored 0.012 in importance. Customers with lower total charges may be more prone to churn as they haven't invested heavily in the service yet; similarly, customers with higher monthly charges also exhibit greater churn tendencies.
6. Value-added Services Subscriptions: Whether customers subscribe to value-added services like OnlineSecurity_Yes (0.012) and TechSupport_Yes (0.012) also aids in predicting churn. Customers not subscribing to these services typically exhibit higher churn rates.

These findings align with previous exploratory data analysis of this dataset. Notably, as shown in the chart, demographic characteristics like gender and relationship status did not appear among the top 15 most significant features. This strongly indicates that customer churn decisions are more influenced by their service contracts with the company, fees, and specific service usage experiences than by their personal identity background.

4. Discussion

The analysis of feature importance provides actionable insights for telecom operators seeking to reduce customer churn. This section translates the model's key findings into data-driven business strategies. A primary finding is that customers on month-to-month contracts are the most prone to churn due to a lack of long-term commitment and low switching costs. To address this, operators should design targeted marketing campaigns that encourage these customers to transition to one- or two-year contracts. Effective incentives could include first-month discounts, complimentary data allowances, or service upgrades to lock in customer relationships and significantly reduce churn risk. Similarly, new customers with a short tenure represent another high-risk segment, as the early stage is a critical period for building loyalty. It is recommended to implement a robust New Customer Onboarding Program. Proactively engaging with customers within the first three months via SMS, app notifications, or phone calls can ensure a seamless experience, address any initial queries, and reinforce the value of their subscription. The balance between price and perceived value is also crucial. Churn rates are elevated among customers with high-fee fiber optic plans who do not subscribe to value-added services, suggesting they feel the service value does not match the premium price. To counter this, operators should redesign their product portfolio by bundling services like online security and technical support with these premium plans. This approach not only increases customer stickiness but also better justifies the higher price point, enhancing overall perceived value. Finally, the payment method significantly influences retention, with electronic check users exhibiting a higher churn rate, likely due to payment failures or inconvenience. Operators should promote more stable and convenient automated payment methods, such as credit cards or bank auto-debit, by offering small bill discounts as an incentive. This can effectively reduce unintentional churn caused by payment friction. In summary, the model not only predicts who is likely to churn but also reveals the underlying reasons through feature importance analysis, providing telecom enterprises with a clear roadmap for precise and effective interventions.

5. Conclusion

This study systematically compared three machine learning models for telecom churn prediction and validated a reproducible workflow that transforms model outputs into actionable business guidance. The findings confirm that ensemble learning models significantly outperform traditional

linear models in this context. Specifically, the XGBoost model demonstrated the most robust performance, achieving the highest scores across all key metrics, including an Area Under the Curve (AUC) of 0.857. Through feature importance analysis of the XGBoost model, the study identified that the most critical churn drivers relate to contract and financial status—including contract type, customer tenure, and monthly fees—rather than demographic characteristics. However, the study has certain limitations that should be acknowledged. The findings are based on a single, public benchmark dataset, which may limit generalizability to specific commercial environments. Furthermore, the analysis used a static data snapshot, failing to capture the dynamic, time-series nature of customer behavior that may contain critical churn signals. The comparative scope was limited to three representative models, and no specialized techniques like SMOTE were applied at the data level to address the dataset's class imbalance. Future research should therefore aim to incorporate time-series data to capture evolving customer behaviors for more timely and accurate predictions. It is also recommended to construct a business-oriented evaluation framework that weighs technical accuracy against factors like customer value and retention costs to identify the model that delivers maximum business profit. Finally, enhancing the interpretability of high-performance models like XGBoost is of vital importance. Applying advanced interpretability tools can provide customized churn reasons for each high-risk customer, allowing for the formulation of more precise and targeted retention strategies.

References

[1] Kaggle. (2025). Customer churn dataset. https:// www.kaggle.com/ datasets/ muhammadshahidazeem/ customer-churn-dataset.

[2] Google Dataset Search. (2025). Telco customer churn dataset. https://toolbox. google. com/ datasetsearch/ search? query=Telco%20Customer%20Churn%20dataset%20-site%3Akaggle.com.

[3] Xu, T., Ma, Y., & Kim, K. (2021). Telecom churn prediction system based on ensemble learning using feature grouping. Applied Sciences, 11(11), 4742. https://doi.org/10.3390/app11114742.

[4] Wei, S. (2025). Comparative analysis of machine learning models for telecom customer churn prediction. Theoretical and Natural Science, 134, 24–30. https://www.ewadirect. com/ proceedings/ tns/ article/ view/26483.

[5] Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., & Benson, V. (2024). Prediction of customer churn behavior in the telecommunication industry using machine learning models. Algorithms, 17(6), 231. https://doi.org/10.3390/a17060231.

[6] Kaggle. (2025). Telco customer churn: Exploratory data analysis. https://www.kaggle.com/ code/ supratimhaldar/ telco-customer-churn-exploratory-data-analysis.

[7] Medium. (2019). Machine learning case study: Telco customer churn prediction. https://medium.com/ @ manureservations/machine-learning-case-study-telco-customer-churn-prediction-a5f228364945.

[8] Kaggle. (2025). Telco customer churn - EDA & predictions + SHAP. https://www.kaggle. com/ code/ jonaspalucibarbosa/telco-customer-churn-eda-predictions-shap.

[9] Ahmad, J., Jafar, R., & Aljoumaa, F. (2025). Customer churn prediction in telecom using machine learning and social network analysis in big data platform. ResearchGate. https://www. researchgate. net/ publication/386615188_Customer_churn_prediction_in_telecom_using_machine_learning_and_social_ network_analysis_in_big_data_platform.

[10] De Caigny, S., Coussement, K., & De Bock, K. (2025). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. ResearchGate. https://www.researchgate.net/publication/323130432_A_New_Hybrid_Classification_Algorithm_for_Customer_Churn_Prediction_Based_on_Logistic_Regression_and_Decision_Trees.

[11] Kaggle. (2025). Telco Customer Churn. https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

[12] Observable. (2023). End-to-end machine learning project: Telco customer churn / His CodeShip. https://observablehq.com/@ealecho/end-to-end-machine-learning-project-telco-customer-churn.

[13] Medium. (2020). Comprehensive report: Telecom customer churn analysis and recommendations. https://medium.com/@KingHenryMorgansDiary/comprehensive-report-telecom-customer-churn-analysis-and-recommendations-398eedaf3466.

[14] Ahmad, J., Jafar, R., & Aljoumaa, F. (2024). Customer churn prediction in telecom using machine learning and social network analysis in big data platform. ResearchGate. https://www. researchgate. net/ publication/386615188_Customer_churn_prediction_in_telecom_using_machine_learning_and_social_ network_analysis_in_big_data_platform.