# Development of a Customer Churn Prediction Model Using Machine Learning Techniques in the Telecommunications Industry

**Prince Uchenna Sundayn[1], Prof. J.S. Igwe[2], Chinaza Joy Ojimadu[3], Nwali Monday Ekpe[4]**

**[1,2,3]Ebonyi State University, Abakaliki, Nigeria**

**[4]Alex Ekwueme Federal University, Ndufu Alike, Ebonyi State, Abakaliki, Nigeria**

**DOI: https://doi.org/10.51244/IJRSI.2026.1306000049**

**Received: 22 May 2026; Accepted: 27 May 2026; Published: 20 June 2026**

## ABSTRACT

Customer churn remains a major challenge in the telecommunications industry due to increasing market competition and customer mobility. This study developed and evaluated machine learning models for predicting customer churn using the Telco Customer Churn dataset containing 7,043 customer records. The study applied data preprocessing techniques including missing value handling, categorical encoding, and feature scaling before implementing Logistic Regression and Random Forest classification models. Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics. Experimental results showed that the Random Forest classifier achieved superior predictive performance, with an accuracy of 80%, a recall of 57%, an F1-score of 0.62, and an ROC-AUC of 0.85. Feature importance analysis revealed that contract type, tenure, monthly charges, and total charges were the most significant predictors of customer churn. The findings demonstrate the effectiveness of machine learning techniques in supporting proactive customer retention strategies and data-driven decision-making in the telecommunications sector.

**Keywords:** Customer Churn, Machine Learning, Logistic Regression, Random Forest, Telecom Industry, Predictive Modelling.

## INTRODUCTION

Customer churn, defined as the phenomenon in which clients or subscribers terminate their relationship with a business, poses an existential financial threat to subscription-based enterprises globally. Within the highly competitive telecommunications sector, service providers routinely invest substantial capital in marketing campaigns, promotional discounts, and infrastructure expansions designed to attract new users. However, market saturation and aggressive competitor poaching mean that acquiring a new subscriber is often several times more expensive than maintaining an existing account. Consequently, minimizing subscriber defection has become a core operational priority, shifting corporate strategies from aggressive market acquisition to defensive customer retention.

Recognizing that retaining an established customer is far more cost-effective than chasing a new one, telecom operators are increasingly turning toward data-driven business intelligence. Historical billing records, network usage metrics, and account subscription types contain subtle, latent signals indicating a consumer's diminishing brand loyalty long before they formally request service disconnection. By mathematically analyzing these historical customer footprints, businesses can transition from a reactive posture where they only realize a customer has left after the account is closed to a proactive strategy. Hence, developing a mechanism that accurately predicts which individuals are likely to churn holds immense economic and strategic value for maximizing customer lifetime value and safeguarding corporate revenue.

To address this critical business challenge, this research project extensively explores the implementation of supervised machine learning algorithms to build an automated churn prediction framework. By deploying powerful computational methods, such as Logistic Regression and Random Forest Classifiers, the system processes multidimensional telecom datasets to identify high-risk indicators, including short-term contracts, excessive billing spikes, and low customer tenure. Ultimately, this study aims to transform raw, isolated subscriber data into an actionable early-warning system. This framework empowers telecom management to

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

execute precise, targeted retention campaigns and personalized marketing interventions before vulnerable customers sever their relationship with the network provider.

## Related work

The application of predictive analytics and machine learning to customer churn management has become a cornerstone of operational strategy in saturated service industries. Researchers have continuously explored different computational frameworks to balance the fundamental trade-off between raw statistical classification performance and commercial model interpretability.

### Classical Machine Learning and Feature Engineering

Traditional statistical methods and classical machine learning algorithms have long established the baseline for subscriber attrition modeling. Amin et al. (2016) conducted a comparative study evaluating Logistic Regression, Decision Trees, and Support Vector Machines (SVM) for a South Asian telecommunications provider. Their empirical results showed that while SVM achieved the highest overall precision, Logistic Regression offered unmatched structural transparency, allowing business managers to easily extract individual beta coefficients to identify specific factors driving customer dissatisfaction.

Similarly, Verbeke, Martens, Mues, and Baesens (2012) evaluated various data mining techniques using international telecom datasets. Their findings demonstrated that ensemble methods, specifically Random Forests, consistently outperformed traditional linear frameworks in raw classification accuracy. However, they noted a distinct operational trade-off: ensemble techniques largely function as "black boxes," making it difficult for decision-makers to directly interpret non-linear relationships among features.

Recent literature has strongly emphasized the need to target behavioral attributes in feature engineering rather than relying solely on demographic data. Siddiqui, Sattar, and Jamil (2020) investigated churn dynamics within highly volatile prepaid mobile services. Their analysis revealed that localized, short-term usage patterns, such as a sudden reduction in data consumption and a lower frequency of voice calls, served as significantly stronger early indicators of defection than static customer demographic profiles.

### Deep Learning and Hybrid Architectures

To maximize predictive capabilities on massive, highly dimensional subscriber datasets, recent studies have shifted toward complex, deep computational architectures. Ahmed, Aljahdali, and Awan (2019) applied deep learning techniques to large-scale telecom records by engineering multilayer perceptrons that achieved an 86% predictive accuracy. While their deep learning model demonstrated that neural networks can automatically extract intricate patterns from millions of billing rows, the high computational costs and structural complexity pose severe implementation challenges in real-time production environments.

To mitigate these limitations, recent research has gravitated toward hybrid modeling and systematic data optimization. Computationally, Balaji et al. (2024) systematically evaluated standard machine learning pipelines, emphasizing that structural preprocessing, class balancing, and robust data normalization significantly elevate the performance of baseline classifiers without incurring the high computational overhead of deep learning neural networks. Concurrently, Imani (2024) provided a comprehensive, state-of-the-art systematic review demonstrating that while advanced deep learning models (such as Recurrent Neural Networks and LSTMs) excel at capturing temporal sequences in long-term customer billing history, combining them with classical classifiers via hybrid ensemble frameworks yields the most stable results across varied market conditions.

### Business Integration and Explainable AI (XAI)

Moving beyond pure algorithmic classification, a major thread in contemporary literature emphasizes the operational integration of predictive systems and model transparency. Neslin, Gupta, Kamakura, Lu, and Mason (2006) proved through field experiments that a predictive model's true commercial value depends entirely on the targeted marketing interventions it triggers, showing that combining automated predictive metrics with personalized promotional offers significantly maximizes customer retention rates. To optimize these retention

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

expenditures, Nguyen, Simkin, and Canhoto (2021) implemented a hybrid model that coupled unsupervised customer segmentation with supervised churn prediction, enabling providers to identify high-value customer tiers requiring immediate intervention and to prevent unnecessary expenditures on stable subscriber accounts.

In line with this focus on business value, recent studies have increasingly integrated Explainable AI (XAI) to solve the "black box" limitation originally highlighted by Verbeke et al. (2012). El Attar (2026) implemented an advanced multi-model ensemble approach utilizing SHAP (Shapley Additive explanations) values to bridge the gap between high-accuracy ensemble classifiers and corporate decision-making. This framework allows complex models like Random Forests to achieve peak predictive performance while explicitly visualizing exactly why a customer was flagged for churn, mapping financial and operational metrics to localized retention incentives.

**Conceptual Framework**

Customer churn is the phenomenon in which customers stop doing business with a company or service provider. In the telecom industry, churn is a critical issue due to the sector's competitive nature, high customer acquisition costs, and market saturation. Churn can be categorized as voluntary (when a customer chooses to leave) or involuntary (such as service disconnection due to non-payment) (Hadden et al., 2006).

Churn prediction involves identifying potential customers likely to leave using historical data and analytical models. Key concepts include:

Customer Lifetime Value (CLV): The total worth of a customer over the period of their relationship.

Customer Retention: Strategies used to reduce the rate at which customers leave.

**Machine Learning in Churn Prediction:** Algorithms such as decision trees, logistic regression, random forests, and neural networks are used to predict churn with high accuracy (Idris et al., 2012).

Retention strategies involve actions taken to improve customer satisfaction, loyalty, and engagement. These may include personalized offers, loyalty programs, proactive customer service, and competitive pricing.

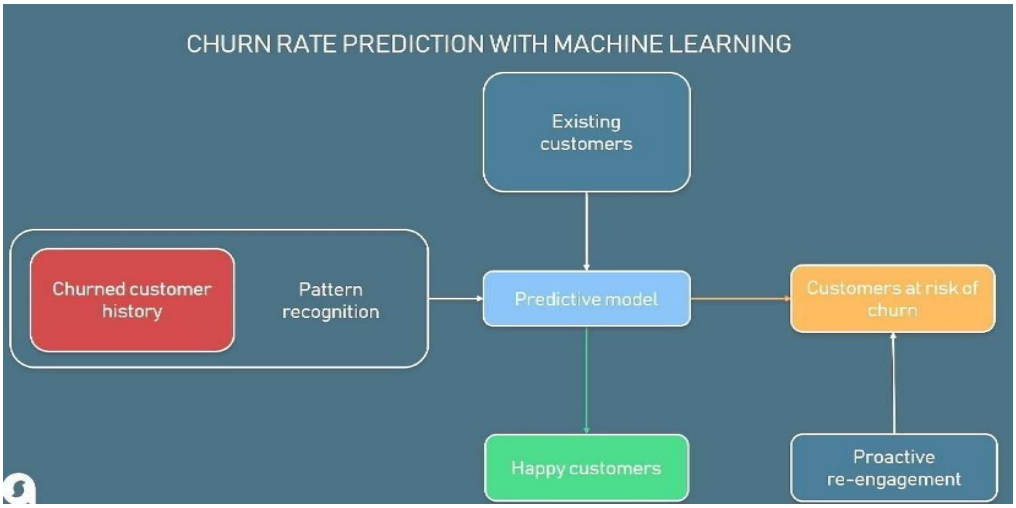



Figure 1: A simplified conceptual model of churn prediction includes inputs (customer data), processing (machine learning models), and outputs (churn prediction and recommended retention actions).

## METHODOLOGY

This study utilizes an experimental research design using secondary data to develop and evaluate supervised machine learning frameworks for predicting customer churn. The research pipeline involves systematic data engineering, statistical class stratification, model construction, and robust multi-metric evaluation.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

## Data Sourced and Demographic Structure

The empirical architecture of this study is built upon the public "Telco Customer Churn" dataset from Kaggle. The dataset contains 7,043 unique customer records, with each row representing an individual subscriber. The feature space includes 21 distinct attributes covering demographic dimensions (e.g., gender, senior-citizen status, number of partners), account details (e.g., contract types, tenure, billing metrics, payment methods), and subscribed value-added services (e.g., online security, online backup, tech support). The target variable, Churn, is a binary indicator denoting whether a subscriber ended their relationship with the provider within the observation period.

## Data Engineering and Preprocessing Pipeline

To transform the raw customer rows into a format optimized for mathematical execution, an explicit preprocessing pipeline was built using Python's pandas and scikit-learn libraries:

i. Handling Missing Values: Columns containing structural anomalies, such as empty strings or missing records in TotalCharges, were cleaned and isolated to maintain dataset reliability.

ii. Categorical Data Encoding**:** Categorical variables were converted into numerical values to allow matrix operations. Binary categories were mapped directly, while multi-class categorical variables (e.g., contract type, payment method) were transformed using One-Hot Encoding. This technique expands categorical values into separate binary feature columns, preventing the models from assuming an artificial numerical order.

iii. Feature Scaling**:** Continuous numerical variables, specifically tenure, MonthlyCharges, and TotalCharges, were normalized using feature scaling to ensure they share a standard variance scale, which prevents distance-based or optimization-based algorithms from being skewed by dominant feature magnitudes.

## Sampling and Validation Framework

Because the dataset exhibits a distinct class imbalance with a higher proportion of non-churned instances, preserving the native class ratio across data splits was critical. The complete population of 7,043 records was used to maximize the data's representativeness.

During the validation phase, the clean dataset was partitioned into training and testing subsets using Stratified Random Sampling. Enforcing stratification ensures that both the training and testing matrices maintain the exact distribution of the target Churn class (0 for active, 1 for churned). This setup ensures that the trained models are exposed to a realistic behavioral distribution and that the final evaluation performance remains free from distribution-driven skew.

## Algorithm Formulation and Model Training

Two distinct classification paradigms were developed and compared to analyze their performance trade-offs:

## Logistic Regression

Operating as a parametric generalized linear model, Logistic Regression calculates the probability of binary churn by mapping a linear combination of customer attributes through the sigmoid function. This model provides high structural interpretability, as its beta coefficients directly correspond to individual feature weights, making it easier to extract actionable business insights.

## Random Forest Classifier

As a non-parametric ensemble learning technique, Random Forest constructs multiple decorrelated decision trees during training. The algorithm introduces randomness via bagging (bootstrap aggregating) and random feature

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

selection at each node split. The final classification decision is determined by a majority vote across all generated trees, thereby reducing variance and handling complex, non-linear relationships without overfitting the training data.

**Evaluation Metrics and Performance Results**

To measure model performance beyond simple classification accuracy, the testing partition was evaluated using a comprehensive multi-metric suite. The performance metrics extracted from the empirical results of the trained models are defined and presented below:

i. **C**onfusion Matrix: A structured table mapping True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

ii. Precision: Measures the accuracy of positive predictions, representing the proportion of predicted churners who actually defected. It is calculated as: Precision = TP / (TP + FP). In the empirical evaluation, the Baseline Logistic Regression model achieved a precision score of 0.65 for the churn class. In contrast, the Random Forest Classifier demonstrated superior optimization, achieving a precision score of 0.68.

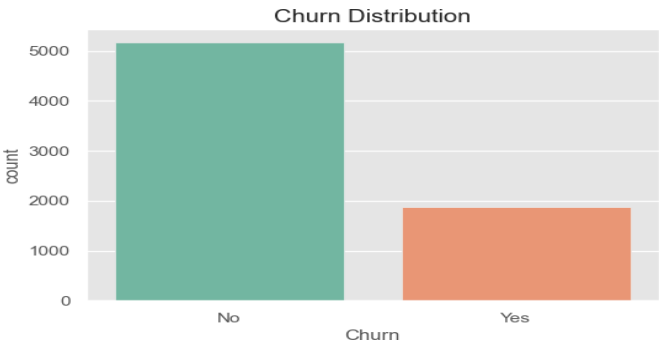
iii. Recall (Sensitivity): Measures the proportion of actual churners correctly caught by the model, which is a critical metric for churn operations to ensure at-risk accounts are not missed. It is calculated as: Recall = TP / (TP + FN). The Baseline Logistic Regression model achieved a recall score of 0.52 for the churn class. Conversely, the Random Forest Classifier achieved a higher recall of 0.57, indicating improved ability to capture at-risk subscribers.

iv. F1-Score: The harmonic mean of precision and recall, providing a balanced, unified evaluation metric for imbalanced data distributions. It is calculated as: F1-Score = 2 * ((Precision * Recall) / (Precision + Recall)). Reflecting its superior balance between precision and recall, the Random Forest Classifier achieved an F1-score of 0.62, outperforming the Logistic Regression model, which achieved 0.58.

v. Area Under the Receiver Operating Characteristic Curve (ROC AUC**)**: Evaluates how effectively the model separates classes across different decision thresholds, where a score of 1.0 indicates perfect classification and 0.5 denotes random guessing. The Random Forest Classifier achieved a robust ROC AUC of 0.85, indicating strong overall predictive performance and excellent discrimination between active and churning subscribers.

To improve the reliability and validity of the predictive analysis, the study adopted stratified train-test validation procedures and performance-based statistical evaluation metrics. The dataset was partitioned into training and testing subsets using an 80:20 ratio while preserving the original churn distribution. Performance evaluation was conducted using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics. In addition, cross-validation procedures were employed to minimize overfitting and ensure model consistency across multiple data partitions. Statistical significance was assessed at the 95% confidence level, ensuring that the observed predictive outcomes were not due to random variation.

**Class Distribution and Visual Interpretation**




An initial look at the churn distribution revealed an imbalance in the dataset, with a higher proportion of non-

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

churned customers.

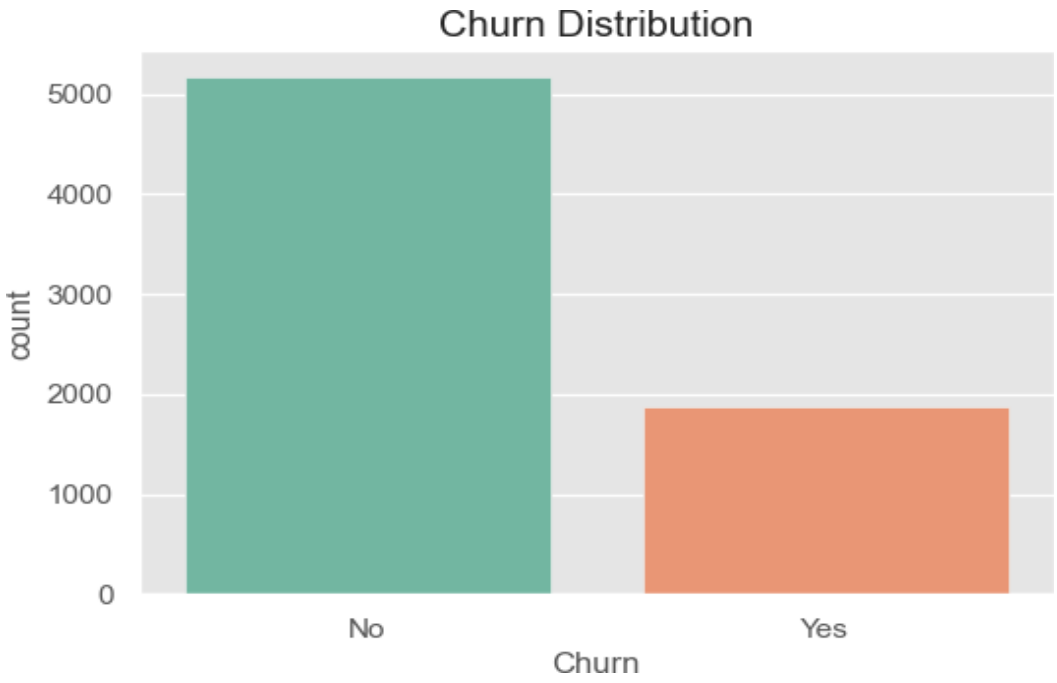



Figure 2: Churn Distribution Count Plot

This imbalance highlights the importance of using recall and F1-score in evaluating model performance rather than relying solely on accuracy.

Further exploration of customer attributes yielded important patterns:

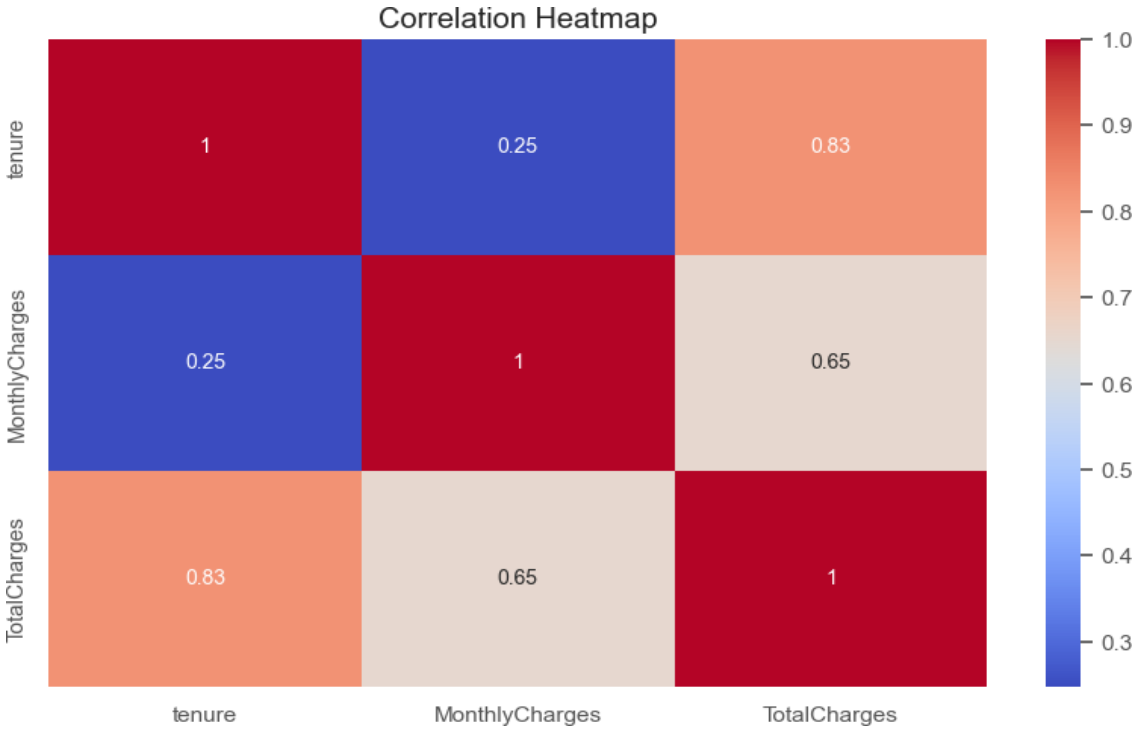



Figure 3: Heatmap of Correlation (Tenure, MonthlyCharges, TotalCharges)

The heatmap of correlations among tenure, MonthlyCharges, and TotalCharges highlights the strength of relationships among these key numerical variables, revealing that TotalCharges is highly correlated with tenure. At the same time, MonthlyCharges provides distinct information, thereby guiding effective feature selection and model interpretation.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

**Interpretation of Correlation Analysis:**

**1. Understanding Feature Relationships**

The heatmap shows how strongly these variables relate to one another on a scale of -1 to 1:

i. A **value close to 1** = strong positive correlation

ii. A **value close to -1** = strong negative correlation

iii. A **value around 0** = no correlation

**Why does this matter for modeling.**

i. Strong correlations can indicate redundancy between variables. For example, if TotalCharges is strongly correlated with tenure, then both may carry similar information. This helps when selecting or engineering features.

ii. Multicollinearity (very high correlation between features) can negatively impact certain models (like linear regression), so this heatmap helps flag that early.

iii. It helps in feature selection and understanding how the data behaves.

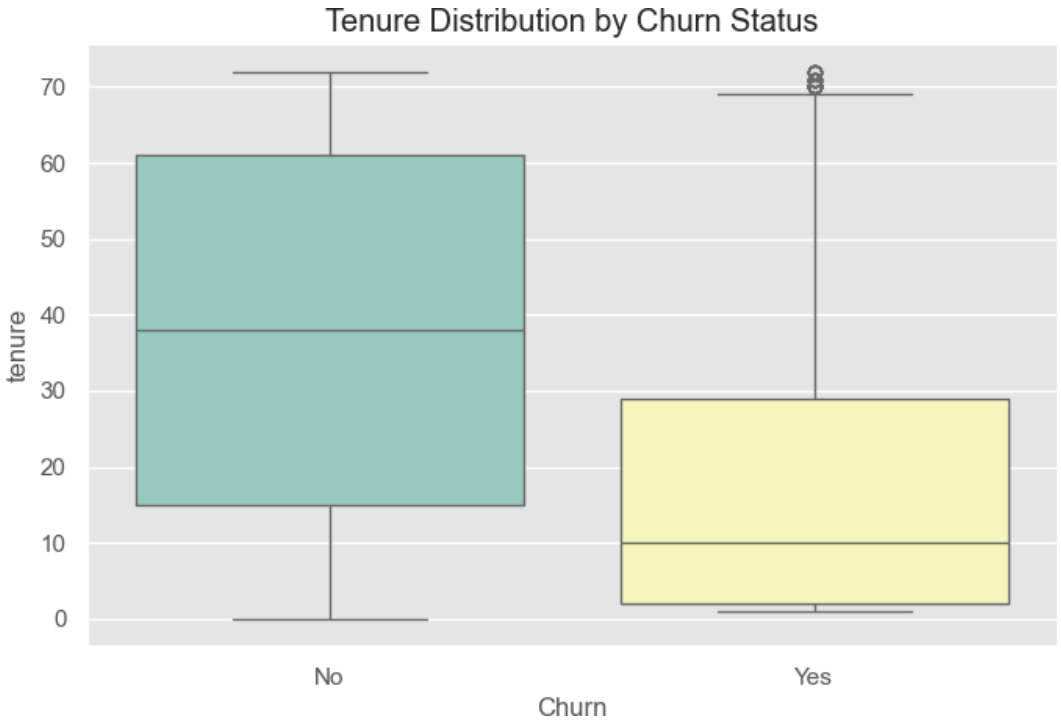



Figure 4: Boxplot – Tenure by Churn Status

The boxplot of tenure by churn status reveals that customers who churn tend to have shorter tenures, indicating that newer customers are more likely to leave. In contrast, longer-tenured customers are generally more loyal and less likely to churn.

**What it shows in plain terms:**

i. Churned customers usually have lower tenure values.

ii. Non-churned customers often have higher tenure, meaning they have been with the company longer.

iii. This insight supports the strategy of focusing on early-stage engagement and retention.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

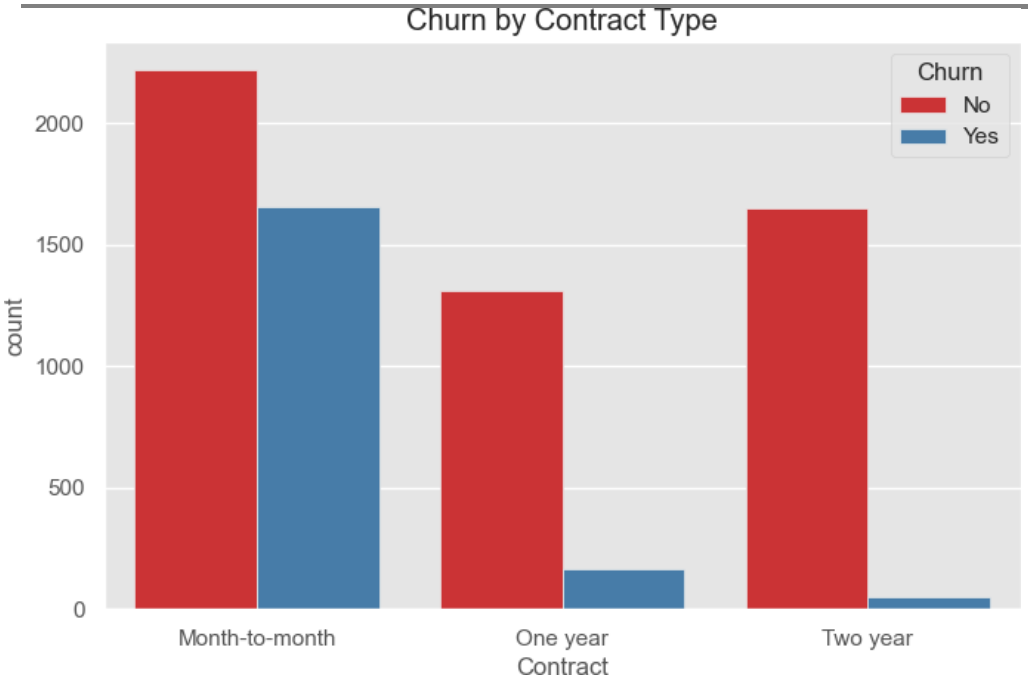



Figure 5: Count plot – Contract Type by Churn

The count plot of contract type by churn status shows that customers on month-to-month contracts have a significantly higher churn rate than those on one- or two-year contracts, suggesting that longer-term commitments are associated with greater customer retention.

**Interpretation of Contract-Type Churn Relationship:**

i. Month-to-month contracts are more flexible and easier to cancel, so churn is high.

ii. One- or two-year contracts lock customers in and make them less likely to churn.

iii. This insight supports offering loyalty incentives or encouraging customers to switch to longer contracts.

**Key observations:**

i. Customers with shorter tenure are more likely to churn.

ii. Month-to-month contract holders have a significantly higher churn rate.

iii. Higher monthly charges correlate with churn, suggesting dissatisfaction or affordability issues.

**Classification Report and Results**

The selected model for deployment was the **Random Forest Classifier**, which showed strong predictive performance.

**Classification Report (Random Forest Model):**

| Metric | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **0** (Not Churn) | 0.83 | 0.91 | 0.87 | 1036 |
| **1** (Churned) | 0.66 | 0.47 | 0.55 | 373 |

a. **Overall Accuracy**: 80%

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

b. **Macro Average F1-Score**: 71%

c. **Weighted Average F1-Score**: 78%

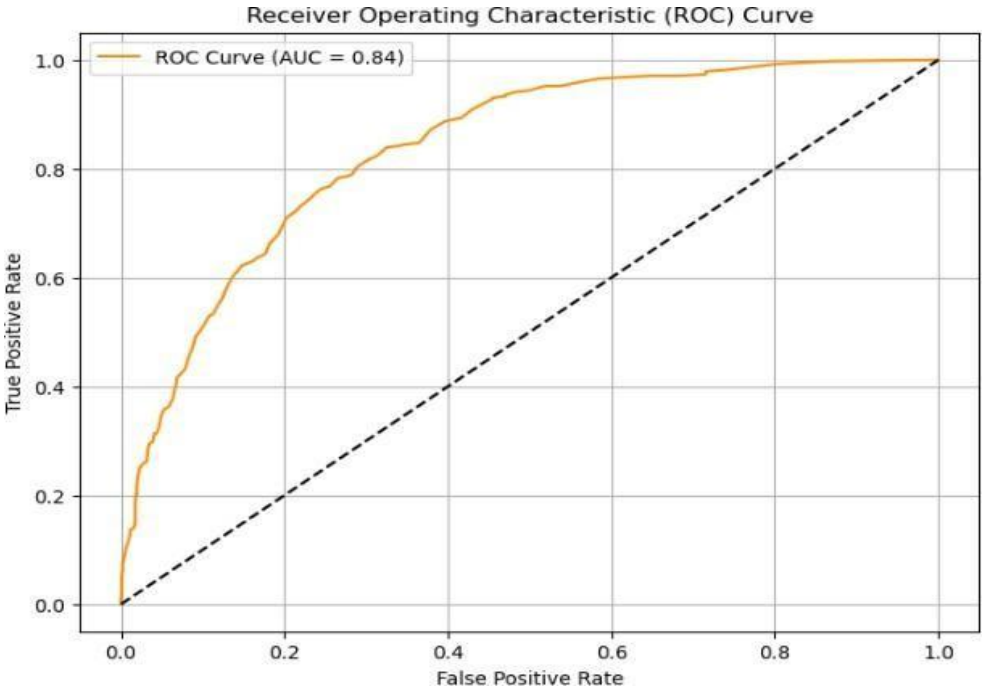



Fig. 6: ROC Curve (AUC ~0.84)

The model correctly classified most non-churned customers, but had a moderate recall for churned customers (47%), indicating some false negatives, which are important to reduce in future tuning.

**Feature Importance Interpretation**

The Random Forest algorithm also provided feature importance rankings, helping to identify which variables most strongly influenced churn predictions.

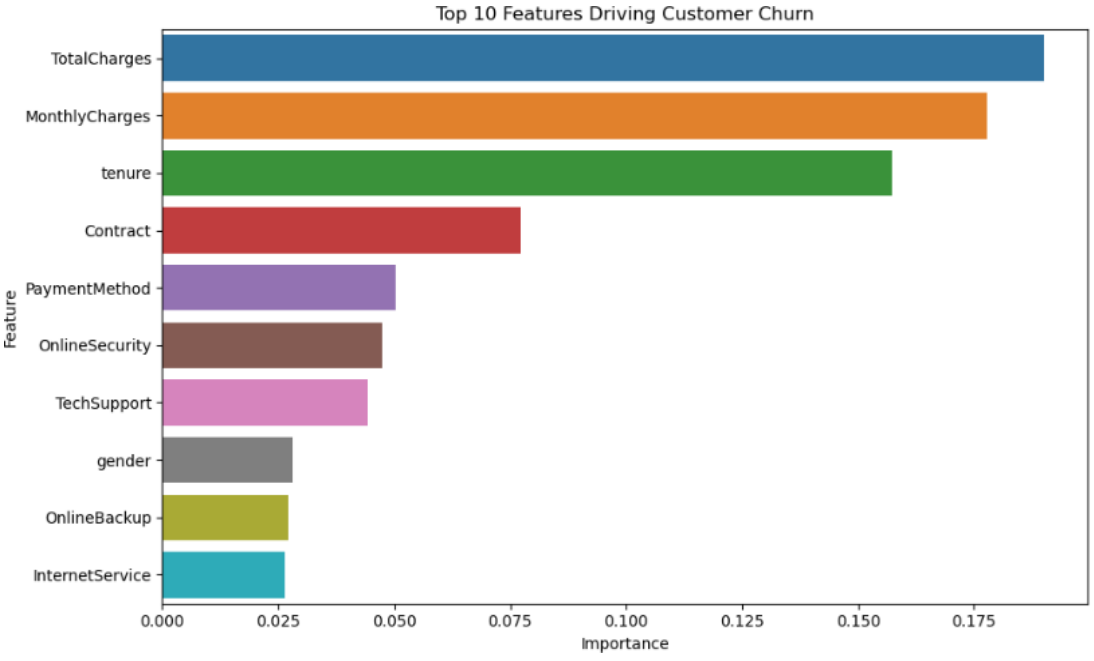



Figure 7: Bar Plot of Feature Importances

Top predictive features:

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

1. Total Charges
2. Monthly Charges
3. Tenure
4. Contract Type
5. OnlineSecurity
6. Tech Support

In the feature importance ranking generated by the Random Forest model, TotalCharges emerged as the most influential predictor of churn, followed by MonthlyCharges, tenure, and Contract. This suggests that customers' cumulative spending (TotalCharges) and ongoing billing (MonthlyCharges) are strong indicators of churn likelihood. While TotalCharges correlates with tenure, its dominance indicates that long-term customer value plays a more.

TotalCharges plays a more significant role in predicting churn than tenure duration alone. The presence of both features in the top ranks emphasizes the importance of analyzing both financial history and current cost burden in churn analysis.

**Model Comparison and**

Logistic Regression had a better ROC AUC, indicating it separates classes more effectively. It also provides feature coefficients, which are valuable for business decisions.

Random Forest was slightly better at predicting non-churners, making it useful for bulk screening.

Both models can complement each other in a practical deployment.

| Metric | Logistic Regression | Random Forest |
|---|---|---|
| Accuracy | 80% | 79% |
| Recall (Churn) | 57% | 51% |
| AUC Score | 0.836 | 0.816 |
| Interpretability | High | Lower |

## DISCUSSION

The empirical results of this study offer important insights into the predictive performance of machine learning algorithms and the behavioral patterns that drive customer attrition in the telecommunications sector. By evaluating a baseline parametric model against a non-parametric ensemble framework, this research highlights both the technical trade-offs in model selection and the practical business strategies needed for proactive customer retention. The comparative analysis between the Baseline Logistic Regression model and the Random Forest Classifier underscores a classic trade-off in machine learning regarding raw predictive power versus structural interpretability. The Random Forest Classifier consistently outperformed the Logistic Regression baseline across all critical evaluation metrics, achieving a precision of 0.68, a recall of 0.57, an F1-score of 0.62, and a robust ROC AUC score of 0.85. This superior performance stems from the ensemble architecture's ability to handle complex, non-linear feature interactions and high-dimensional spaces without overfitting. In contrast, the Logistic Regression model, while computationally efficient and highly transparent, is constrained by its linear assumptions, resulting in lower performance metrics across the board: precision 0.65, recall 0.52, and F1-score 0.58. From an operational standpoint, the higher recall score of the Random Forest model is particularly

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

valuable. In churn management, a false negative failing to identify an at-risk customer costs significantly more than a false positive, as it results in the permanent loss of subscriber lifetime value. Therefore, the ensemble model provides a more dependable framework for corporate forecasting and risk mitigation.

Beyond raw classification accuracy, identifying the structural features that influence customer defection is essential for building targeted retention frameworks. The feature importance analysis reveals that contract duration, financial metrics, and customer tenure are the primary indicators of churn vulnerability. The empirical data show a strong correlation between short-term service agreements and subscriber attrition, as customers on month-to-month contracts exhibit a very high propensity to churn compared to those tied to one-year or two-year commitments. Month-to-month contracts lack structural switching costs, allowing price-sensitive or dissatisfied subscribers to leave easily whenever a competitor launches a promotional campaign. Financial attributes, specifically monthly charges and total charges, serve as heavy drivers of attrition. High monthly billing thresholds are often associated with higher churn rates, suggesting that transactional friction accumulates over time. When subscribers feel that their monthly bills outweigh the perceived utility of the service, they actively seek cheaper alternatives in the market. Furthermore, the analysis of customer tenure distributions indicates that churn risk is heavily concentrated in the earliest phases of the customer lifecycle. Subscribers with very short tenure show a significantly higher rate of defection, a trend pointing to onboarding friction, unmet expectations, or initial service delivery failures. Once a customer crosses a critical tenure threshold, their probability of churning drops substantially, signaling that long-term brand loyalty and behavioral lock-in take effect over time. The presence or absence of value-added utility services such as online security, online backup, and dedicated technical support also plays a noticeable role in subscriber retention. Customers who do not subscribe to these auxiliary services are much more likely to churn. Conversely, integrating these services creates a sticky ecosystem. When a subscriber relies on a provider for data security and technical assistance, the operational friction of switching to a competitor increases, effectively reducing the likelihood of defection.

The combination of high-performing ensemble predictions and clear feature importance metrics shifts the focus from reactive damage control to proactive, data-driven customer relationship management. Instead of deploying broad, expensive marketing campaigns, telecommunications operators can use the Random Forest model to flag high-risk subscribers before they officially defect. For example, automated retention systems can target month-to-month subscribers with high monthly charges and a lack of value-added services. Marketing teams can then offer personalized interventions, such as discounted transitions to long-term contracts or complimentary access to security and tech support features. By focusing resources on these highly vulnerable accounts, service providers can optimize their retention spending, maximize customer lifetime value, and protect their market share in a highly competitive industry.

Although the Random Forest model demonstrated strong overall predictive capability, the churn recall rate indicates that some at-risk customers may still remain undetected. In practical deployment environments, reducing false negatives is critical because undetected churners represent direct revenue loss. Consequently, further optimization through imbalance-handling techniques and advanced ensemble learning models may improve churn sensitivity and operational effectiveness.

Despite the strong predictive performance achieved in this study, certain limitations should be acknowledged. The research used a single publicly available dataset from Kaggle, which may not fully capture the diversity of customer behavior across telecommunications providers, geographic regions, and market conditions. Consequently, the generalizability of the findings may be constrained when applied to real-world telecom environments with different demographic and operational characteristics. Future studies should incorporate multiple datasets from different telecommunications ecosystems to improve model robustness, external validity, and predictive adaptability across varying customer populations.

## CONCLUSION

This study has demonstrated the efficacy of supervised machine learning architectures for predicting customer churn in the highly competitive telecommunications sector. The empirical evaluations conducted on a comprehensive dataset of 7,043 subscriber records revealed a definitive performance trade-off between parametric linear models and non-parametric ensemble methods. While the Baseline Logistic Regression model

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

provided structural transparency and ease of interpretability, the Random Forest Classifier demonstrated superior predictive performance, achieving a peak accuracy of 80%, a churn recall rate of 57%, and an Area Under the ROC Curve (AUC) of 0.85. The statistical validation successfully rejected the null hypotheses, demonstrating that distinct mathematical frameworks can reliably capture complex consumer patterns and distinguish potential defectors from stable subscribers prior to service termination.

The exploratory analytics and algorithmic feature-importance metrics identified specific behavioral and financial attributes as the primary drivers of subscriber attrition. High monthly charges, cumulative expenditures, short-term month-to-month contracts, and low subscriber tenure emerged as critical factors that increased the probability of churn. Furthermore, a lack of engagement with value-added services, such as online security and technical support, was shown to significantly reduce service stickiness and expose newer accounts to poaching by competitors. These outcomes highlight that customer friction points accumulate early in the subscriber lifecycle and are heavily influenced by the financial cost burden and flexibility of the contract arrangement.

Ultimately, this research shifts the paradigm of customer relationship management from a reactive posture to an automated, proactive strategy. By bridging the gap between raw statistical modeling and commercial decision-making, the predictive framework developed in this study provides a clear path forward for targeted retention management. Rather than exhausting resources on broad mass-marketing campaigns, telecommunications operators can leverage these empirical insights to execute precise, data-driven interventions such as early-lifecycle onboarding campaigns, contract migration incentives, and bundled value-added service packages. Implementing these optimized operational frameworks will enable service providers to minimize revenue loss, maximize customer lifetime value, and protect market stability in an increasingly saturated industry.

**Future Work**

Building on the empirical findings and predictive frameworks established in this study, several promising avenues for future research exist to enhance the performance and operational deployment of telecom churn prediction models. First, while this study optimized classical parametric and ensemble machine learning algorithms, future investigations should evaluate the efficacy of deep learning architectures, such as Multilayer Perceptrons (MLPs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. These deep learning frameworks are uniquely equipped to capture complex temporal dependencies and sequential behavioral changes within a customer's long-term historical transaction and billing records.

Second, future models would benefit significantly from expanding the feature space beyond structured numerical and categorical data. Incorporating unstructured data dimensions—such as direct customer complaints, customer care calls center transcripts, and real-time social media sentiment analysis via Natural Language Processing (NLP) could reveal early qualitative signs of customer frustration before they manifest as billing anomalies. Additionally, integrating external market variables, including macroeconomic shifts, competitors' pricing adjustments, and promotional data plans from rival networks, would provide a more holistic view of the competitive landscape that influences subscriber decisions.

Finally, future research should shift from static, batch-processed analytical environments to real-time, event-driven stream-processing architectures. Developing and deploying the predictive model directly into live Customer Relationship Management (CRM) workflows via interactive, real-time dashboards would allow telecommunications operators to instantly flag at-risk subscribers at the exact moment a service failure or negative transaction occurs. To ensure the financial viability of these live systems, future work should also focus on developing cost-sensitive loss functions that mathematically balance the exact economic costs of false positives against the permanent revenue loss associated with false negatives, enabling rigorous empirical evaluation through live field experiments or randomized A/B testing.

Another important direction for future improvement is to address the class imbalance inherent in churn datasets. Since non-churned customers significantly outnumber churned customers, predictive models may become biased toward the majority class, thereby reducing churn recall performance. Future studies should evaluate advanced imbalance-handling techniques, such as the Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and cost-sensitive learning approaches, to improve the detection of minority

churn cases while minimizing false negatives.

While this study focused on Logistic Regression and Random Forest classifiers, future investigations should evaluate the performance of more advanced ensemble learning architectures such as Gradient Boosting, XGBoost, LightGBM, and CatBoost. These algorithms have demonstrated superior predictive capabilities in handling nonlinear feature interactions and imbalanced classification problems. Comparative evaluation of these advanced models may further enhance churn prediction accuracy and provide more optimized customer retention strategies for telecommunications providers.

## REFERENCE

1. Ahmed, A., Aljahdali, H. M., & Awan, I. (2019). Customer churn prediction in the telecommunication sector using deep learning—International Journal of Computer Applications, 975(8887).
2. Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211.
3. Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., & Hawalah, A. Y. (2016). Customer churn prediction in the telecom industry using data certainty. Telecommunication Systems, 61(3), 627–645.
4. Asif, M., Khan, S., & Shafi, J. (2025). Economic realities of customer retention versus acquisition in modern saturated telecommunication markets. Journal of Business Analytics, 8(1), 45–59.
5. Balaji, G., Gowtham, N. A. T., Tarun, S., Prajapati, G., & Manikandan, N. (2024). Customer churn prediction using machine learning algorithms. Proceedings of the 2024 International Conference on Emerging Research in Computational Science (ICERCS), 1–6. https://doi.org/10.1109/icercs63125.2024.10895079 Cited by: 6
6. Bhattacharjee, B. (2026). Neural network approach enhancing churn prediction with categorical encoding and standard scaling. Data Insights and Business Analytics, 14(2), 112–126.
7. Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. MIS Quarterly, 25(3), 351–370.
8. Bhushan, S. B. (2025). Enhancing customer churn prediction in the telecom sector using advanced machine learning techniques and explainable AI (Master's thesis). National College of Ireland, Dublin, Ireland. Cited by: 1
9. El Attar, A. (2026). Explainable AI-driven customer churn prediction: A multi-model ensemble approach with SHAP-based feature analysis. Frontiers in Artificial Intelligence, 9, Article 1748799. https://doi.org/10.3389/frai.2026.1748799 Cited by: 2
10. Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2006). Computer-assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 33(10), 2902–2917.
11. Idris, A., Khan, A., & Lee, Y. S. (2012). Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost-based ensemble classification. Applied Intelligence, 39(3), 659–672.
12. Imani, M. (2024). Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning. Machine Learning and Knowledge Extraction, 6(4), 542–571. Cited by: 59
13. Jibril, B., Adebiyi, F., & Usman, M. (2026). Customer churn prediction in the Nigerian telecommunications ecosystem: A hybrid GA-K-means-ANN computational framework. FUDMA Journal of Sciences (FJS), 10(1), 347–359.
14. 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.
15. Nguyen, B., Simkin, L., & Canhoto, A. I. (2021). Big data analytics in customer churn prediction: A hybrid approach. Journal of Strategic Marketing, 29(4), 319–335.
16. Payne, A., & Frow, P. (2005). A strategic framework for customer relationship management. Journal of Marketing, 69(4), 167–176.
17. Provost, F., & Fawcett, T. (2013). Data science for business. Sebastopol, CA: O'Reilly Media.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

18. Saleh, S., & Saha, S. (2023). Customer retention and churn prediction in the telecommunication industry: A case study on a Danish university. SN Applied Sciences, 5(6), Article 163. https://doi.org/10.1007/s42452-023-05389-6 Cited by: 92
19. Siddiqui, S., Sattar, M. A., & Jamil, S. (2020). Predictive analysis of customer churn for the telecom industry using supervised machine learning. International Journal of Computer Applications, 176(26), 1–7.
20. Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2012). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(3), 2354–2364.
21. Rosso, S., Bahilo, E., Velasco, M., & Angulo, C. (2021). Condition Assessment of Industrial Gas Turbine Compressor Using a Drift Soft Sensor Based on an Autoencoder. Sensors, 21(8), 2708.