International Journal of Management and Humanities (IJMH)
ISSN: 2394-0913 (Online), Volume-12 Issue-12, August 2026

Customer Churn Prediction Using Machine Learning Models: A Comparative Study Using R Software

Uppu Venkata Subbarao, Tedlapu Narayana Rao, Vantaku Bala, K.B Rajeswara Rao

Abstract: ***Predicting customer churn is essential for improving retention and supporting long-term business growth. In this study, we compared explainable machine learning models for predicting customer churn using the IBM Telco Customer Churn dataset in R. Our approach included data preprocessing, exploratory analysis, model development, performance evaluation, and further analysis. We developed and evaluated four classification algorithms: Logistic Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting, using an 80:20 train-test split. We assessed each model's accuracy, precision, recall, and F1-score. Logistic Regression achieved the best performance, with an accuracy of 82.30% on this dataset. Feature importance analysis indicated that contract type, customer tenure, monthly charges, total charges, and internet service were the key factors influencing churn. These results suggest that explainable machine learning offers both strong predictive performance and greater transparency. The R-based framework we present provides a practical, reproducible approach to support customer retention strategies and help managers make evidence-based decisions in customer relationship management.***

Keywords: Customer Churn; Explainable Machine Learning; Management Science; R programming; Logistic Regression; Random Forest; XGBoost; Decision Tree; Customer Analytics; Decision Support.

Nomenclature:
CRM: Customer Relationship Management
PDP: Partial Dependence Plots
EDA: Exploratory Data Analysis
LR: Logistic Regression
DT: Decision Tree
RF: Random Forest
TP: True Positives
TN: True Negatives
FP: False Positives
FN: False Negatives
RO: Receiver Operating Characteristic
AUC: Area Under the Curve

Manuscript received on 28 July 2026 | First Revised Manuscript received on 01 August 2026 | Second Revised Manuscript received on 10 August 2026 | Manuscript Accepted on 15 August 2026 | Manuscript published on 30 August 2026.

*Correspondence Author(s)

Dr. Uppu Venkata Subbarao*, Department of Basic Sciences, N.S Raju Institute of Technology, Autonomous, Visakhapatnam (Andhra Pradesh), India. Email ID: druvsrao.sh@nsrit.edu.in, ORCID ID: 0000-0003-2951-7399

Dr. Tedlapu Narayana Rao, Department of Management Studies, N.S Raju Institute of Technology, Autonomous, Visakhapatnam (Andhra Pradesh), India. Email ID: tnarayanarao@nsrit.edu.in, ORCID ID: 0000-0002-9441-1008

Dr. Vantaku Bala, Department of Management Studies, N.S Raju Institute of Technology, Autonomous, Visakhapatnam (Andhra Pradesh), India. Email ID: hodmba@nsrit.edu.in, ORCID ID: 0000-0001-5356-2347

K.B Rajeswara Rao, Department of Computer Science and Engineering Lendi Institute of Technology, Autonomous Vizianagaram (Andhra Pradesh), India. Email ID: rajeswararao.kb@lendi.edu.in, ORCID ID: 0009-0009-6330-657X

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open-access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

I. INTRODUCTION

In today's highly competitive digital economy, retaining customers has become a key priority for organizations. Rapid technological advancements, low switching costs, and the availability of numerous competing service providers have made it easier than ever for customers to change providers. As a result, customer churn the decision of a customer to discontinue a service has become a major concern because it directly affects customer lifetime value, profitability, and long-term business sustainability [15] [17]. The rapid growth of digital technologies has enabled organizations to collect large amounts of customer data from transactions, billing records, online interactions, and service usage. These data have created new opportunities for predictive analytics, allowing organisations to identify customers likely to churn, inform marketing strategies, and implement proactive retention initiatives based on data-driven insights [10]. Machine learning has become an increasingly popular approach to customer churn prediction because it can uncover complex patterns and relationships in customer data that traditional analytical methods may miss [20]. Algorithms such as Logistic Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost) have been widely adopted for their ability to produce accurate predictions and support business decision-making [20]. Despite their strong predictive performance, many machine learning models operate as black-box systems, making it difficult for managers to understand how predictions are generated. This lack of transparency often reduces trust in analytical models and limits their practical use in organisational decision-making [1]. Explainable Artificial Intelligence has emerged as an effective solution to this problem by making machine learning models more transparent and easier to interpret. By identifying the variables that contribute most to prediction outcomes, Artificial Intelligence helps managers better understand model behaviour and supports more informed, evidence-based decisions [3]. Although customer churn prediction has been widely studied, relatively few studies present a complete and reproducible analytical framework implemented entirely in R. To address this gap, the present study compares the performance of Logistic Regression, Decision Tree, Random Forest, and XGBoost using the IBM Telco Customer Churn dataset [12]. Model performance is evaluated using Accuracy, Precision, Recall, and F1-score, and feature importance analysis is used

© Copyright: All rights reserved.

to identify the variables that most strongly influence customer churn. The proposed R-based framework provides a practical, transparent, and reproducible approach for customer analytics and business decision support.

II. LITERATURE REVIEW

A. Customer Churn Prediction in Customer Relationship Management

Customer churn prediction is one of the most important applications of predictive analytics in Customer Relationship Management (CRM). Because retaining existing customers is generally more cost-effective than acquiring new ones, organisations increasingly use predictive models to identify customers likely to leave and implement targeted retention strategies [17]. Predictive analytics also supports customer segmentation, personalised marketing, efficient allocation of marketing resources, and customer lifetime value management, helping organisations strengthen customer relationships and improve business performance [18]. Recent studies have shown that factors such as customer tenure, contract type, monthly charges, payment method, and internet service play important roles in predicting customer churn. Understanding these factors enables organisations to identify high-risk customers and design more effective retention strategies [20]. Although predictive performance has improved considerably, ensuring that machine learning models remain understandable and useful to managers remains an important challenge. Organisations increasingly require predictive models that not only deliver accurate results but also provide clear explanations for their predictions [21].

B. Machine Learning Approaches for Customer Churn Prediction

Machine learning has significantly improved customer churn prediction by identifying complex patterns and relationships within customer data that conventional statistical techniques often fail to capture [20] [14]. Logistic Regression remains widely used for its simplicity, computational efficiency, and straightforward interpretation, making it a reliable baseline model for binary classification problems [11] [13]. Decision Tree models are popular because they generate simple decision rules that are easy for managers to understand and interpret [2]. Random Forest improves predictive performance by combining multiple decision trees into an ensemble model, resulting in greater stability, robustness, and more reliable estimates of variable importance [4]. Extreme Gradient Boosting (XGBoost) has become one of the most effective machine learning algorithms due to its ability to improve prediction accuracy while controlling model complexity through gradient boosting and regularisation [8] [22]. Since no single machine learning algorithm consistently performs best across all datasets, comparing multiple models remains essential for identifying the most suitable approach for customer churn prediction [20].

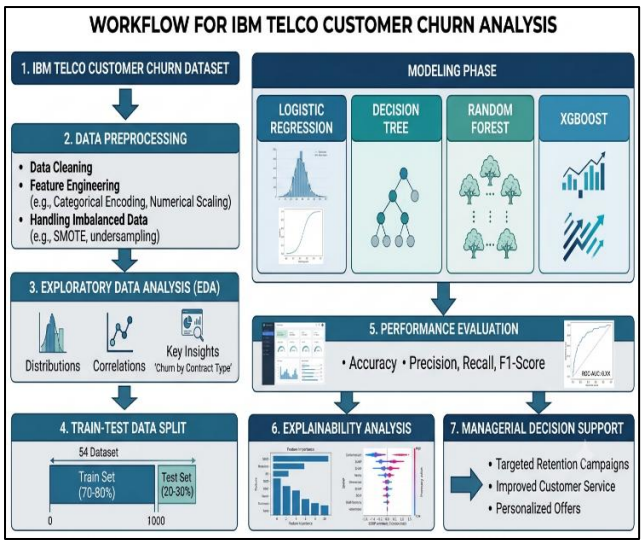
C. Explainable Artificial Intelligence in Business Decision Support

Although machine learning models often achieve excellent predictive performance, many operate as black-box systems, making it difficult for users to understand how predictions are produced [1]. Explainable Artificial Intelligence addresses this issue by improving model transparency and helping users understand the contribution of individual variables to prediction outcomes [3]. Techniques such as feature importance analysis, Local Interpretable Model-Agnostic Explanations, Partial Dependence Plots (PDPs), and SHAP provide valuable insights into model behaviour and make predictive models easier to interpret [16] [19]. In customer churn prediction, explainability enables organisations to identify the key drivers of customer attrition and develop more focused retention strategies based on actionable business insights [6] [7] [9]. By combining comparative machine learning with explainable analytics, the present study provides a framework that not only delivers accurate predictions but also supports transparent and informed managerial decision-making [20].

III. MATERIALS AND METHODS

A. Research Framework

To assess the performance of explainable machine learning models for customer churn prediction, this study uses a quantitative, comparative research approach. Within a reproducible R programming environment, the proposed analytical approach combines data preprocessing, exploratory data analysis, predictive modelling, model evaluation, and explainability analysis. To enhance managerial decision-making, the main goal is to evaluate the predictive performance of advanced machine learning models alongside traditional statistical models and identify the key factors influencing customer churn.



[Fig.1: Research Structure of the Study]

The IBM Telco Customer Churn dataset is the first step in the workflow. Data pretreatment and exploratory data analysis (EDA) come next. Stratified sampling is used to separate the processed dataset into training and testing sets. Accuracy, Precision, Recall, F1-score, and AUC-ROC are used to evaluate four machine learning models: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and XGBoost. The main factors driving customer

© Copyright: All rights reserved.

turnover are then identified through feature significance analysis, thereby enhancing the model's interpretability. The entire process is developed in R, offering a repeatable framework for decision support and customer churn prediction [5].

B. Dataset Description

The IBM Telco Customer Churn dataset, a publicly accessible benchmark dataset frequently utilised in machine learning and customer analytics research, is used in this work. There are 7,043 customer records and 21 variables in the dataset, including 20 predictor variables and one target variable (churn). The target variable indicates whether a customer has been retained ("No") or churned ("Yes"). Customer demographics, account information, subscribed services, and billing details are all described by the predictor variables. Tenure, contract type, internet service, payment method, monthly prices, and total charges are significant variables identified as important determinants of customer churn. As a result, the dataset is ideal for creating and evaluating machine learning models for predicting customer attrition.

C. Data Preprocessing

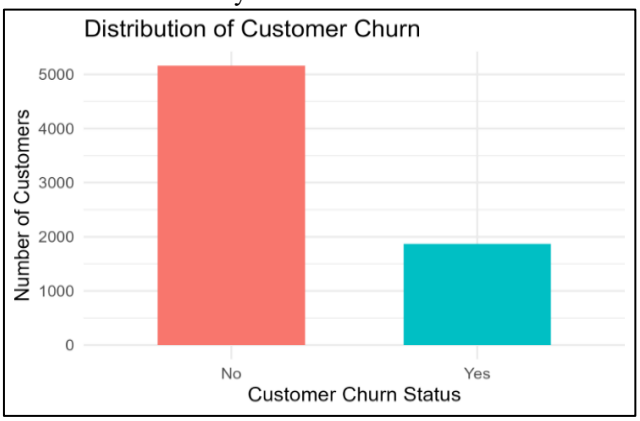
Before model building, data preprocessing was done to enhance data quality. After importing the IBM Telco Customer Churn dataset into R, it was checked for mismatched data types, duplicate records, and missing values. Because it lacked predictive value, the customer ID variable was eliminated. Total Charges' blank values were replaced with missing values, and the variable was then formatted numerically. The corresponding records were removed because missing values accounted for less than 1% of the dataset. There were no duplicate records found. Numerical variables were retained in their original form, whereas categorical variables were converted to factor variables. Descriptive statistics and visualisations were used to examine the cleaned dataset. Lastly, stratified random sampling was used to split the data into 80% for training and 20% for testing. To ensure a fair and repeatable comparison, the Logistic Regression, Decision Tree, Random Forest, and XGBoost models were developed using the identical preprocessed dataset.

D. Exploratory Data Analysis

To understand the characteristics of the cleaned dataset and identify patterns relevant to customer churn prediction, exploratory data analysis was conducted. The distributions of numerical and categorical variables were examined, and their interactions with the objective variable were investigated, using both statistical summaries and graphical visualisations in the R programming environment. Of the 7,032 customer records in the final dataset, 5,163 (73.42%) were non-churn customers and 1,869 (26.58%) were churn customers. The percentage of churn cases was adequate for supervised classification despite the dataset's moderate class imbalance. As a result, several metrics, such as accuracy, precision, recall, and F1-score, were used to assess the model's performance. According to descriptive statistics, the average monthly and total costs were 64.80 and 2,283.30, respectively, and the customer tenure spanned from 1 to 72 months, with an average of 32.42 months. Additionally, graphical analysis revealed that while gender showed minimal differences between churn and non-churn groups, month-to-month contracts, fibre-optic internet access, and electronic check payments were more prevalent among churned customers. The EDA insights informed the selection and interpretation of predictor variables for the ensuing machine learning models, validated the data preprocessing techniques, and provided a preliminary understanding of customer behaviour. The important customer attributes and churn-related trends are graphically analysed in the ensuing subsections

i) Distribution of Customer Churn

Before model development, the distribution of the target variable (Churn) was analysed to evaluate the dataset's class balance. 5,163 (73.42%) non-churn customers and 1,869 (26.58%) churn customers comprised the 7,032 customer records in the cleaned dataset, as seen in Figure 2. Although the dataset exhibits moderate class imbalance, the proportion of churn cases is sufficient for supervised classification. The observed churn rate underscores the importance of developing reliable predictive models to identify customers at risk of departing. Model performance was assessed using a variety of criteria, such as accuracy, precision, recall, and F1-score, to prevent false results resulting from class imbalance. These results confirm that the dataset is appropriate for comparative evaluation of the four machine learning models considered in this study.



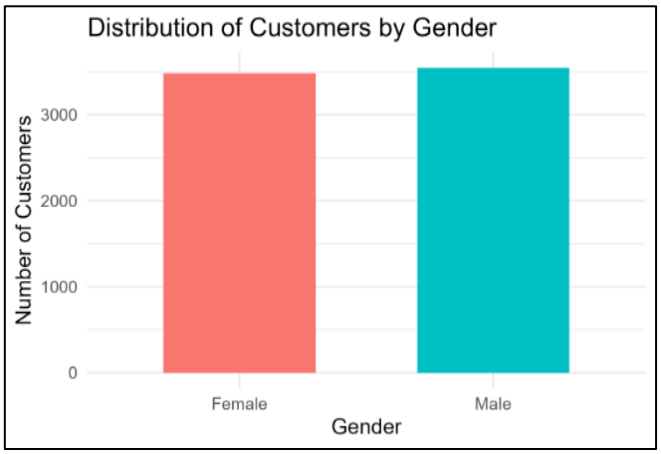
[Fig.2: Distribution of Customer Churn in the IBM Telco Customer Churn Data Set]

ii) Distribution of Customer Demographic Characteristics

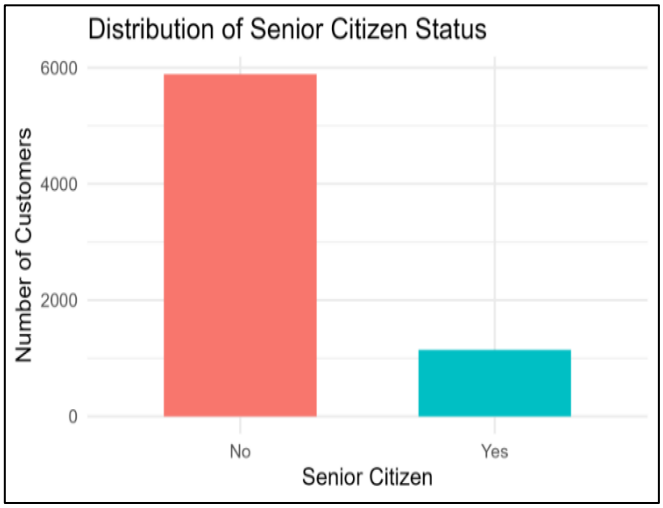
Customer demographic characteristics were examined to understand the customer base's composition before model development. Figure 3 shows that the cleaned dataset comprised 3,483 female (49.53%) and 3,549 male (50.47%) customers, indicating a nearly equal gender distribution and the absence of demographic bias. The Senior Citizen variable was also explored to examine age-related differences in the customer population. Together, these demographic variables provide useful background information for customer segmentation and are retained as predictor variables in the machine learning models. However, previous studies suggest that demographic characteristics generally have less influence on customer churn. than contractual, financial, Furthermore, service-related factors; overall, the

© Copyright: All rights reserved.

demographic analysis confirms that the dataset represents a balanced customer population, providing a suitable basis for subsequent predictive modelling and comparative evaluation of customer churn models.



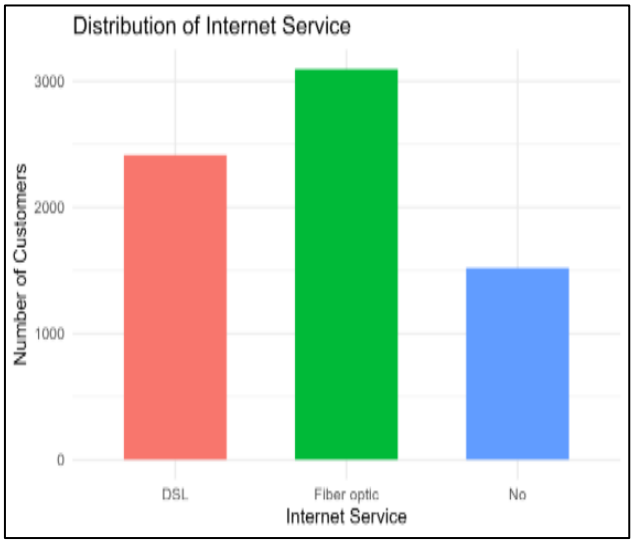
[Fig.3: Customer Distribution by Gender]



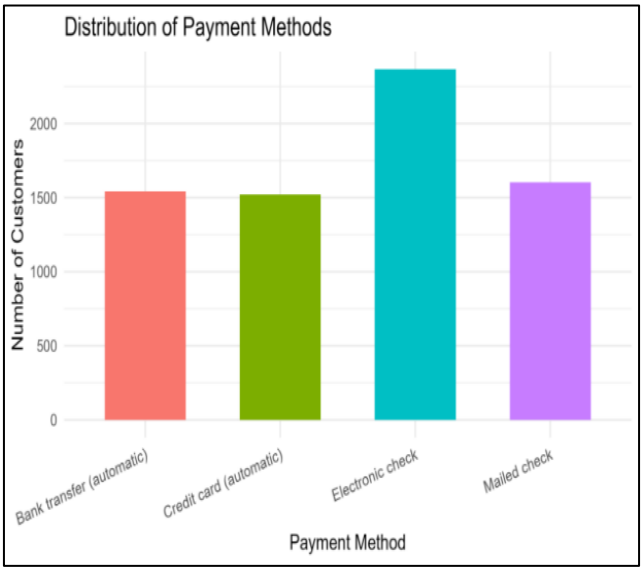
[Fig.4: Distribution of Customers by Senior Citizen Status]

iii) Distribution of Service-Related Characteristics

Service-related variables were explored because they are widely recognised as important predictors of customer churn. Figures 5 and 6 present the distributions of Internet Service and Payment Method**.** The analysis shows that month-to-month contracts represent the largest customer segment, followed by one-year and two-year contracts. Customers with short-term contracts generally have greater flexibility to switch service providers and are therefore more likely to churn. The majority of customers subscribed to either fibre optic or DS**L** internet services, while a smaller proportion did not use internet services. In addition, electronic check was the most frequently used payment method, followed by automatic bank transfer, credit card, and mailed check. These service-related characteristics provide valuable insights into customer behaviour and are expected to play an important role in churn prediction. Their predictive contribution is evaluated in subsequent machine learning models and further examined through feature-importance analysis.



[Fig.5: Distribution of Customers by Internet Service]

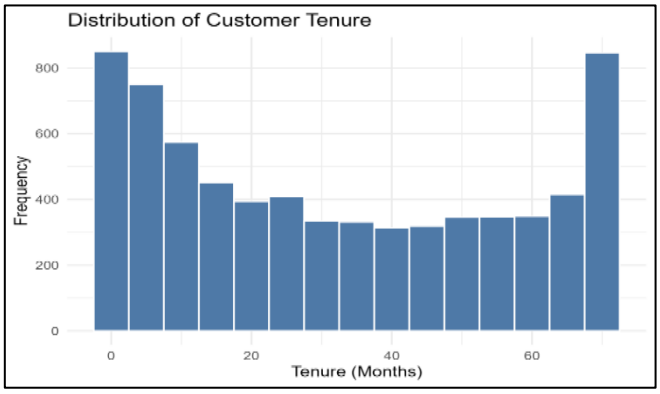


[Fig.6: Distribution of Customers by Payment Method]

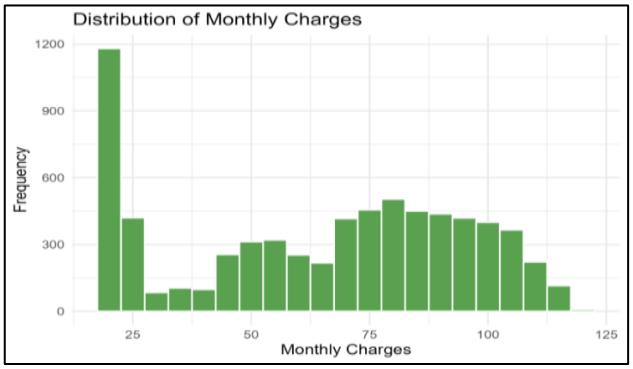
iv) Distribution of Numerical Variables

The numerical variables, namely tenure, Monthly Charges, and Total Charges, were analysed using descriptive statistics and graphical visualisations to gain insights into customer subscription duration and expenditure patterns. The distributions of these variables are shown in Figures 7 and 8. The client base included both new and long-term members, as evidenced by an average customer tenure of 32.42 months (range: 1 to 72 months). Due to variations in subscription services and price plans, monthly charges ranged from 18.25 to 118.75. The distribution of Total Charges, which ranged from 18.80 to 8,684.80, was slightly right-skewed, indicating that long-term clients incurred significantly higher service costs. The observed differences in billing features and subscription duration suggest that these factors provide useful information for differentiating between customers who have churned and those who have not. Accordingly, all numerical variables were retained for subsequent machine learning modelling and comparative evaluation.

© Copyright: All rights reserved.



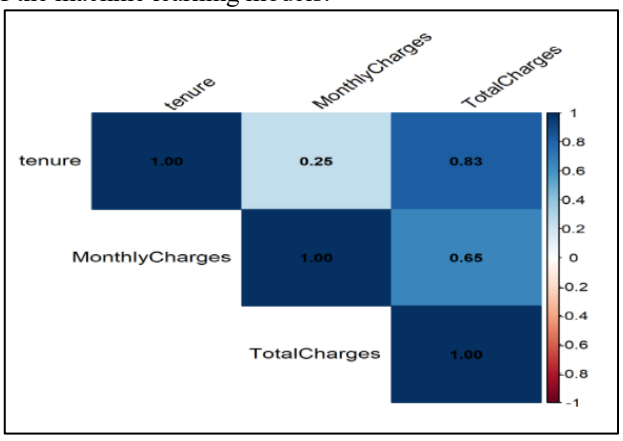
[Fig.7: Distribution of Customer Tenure]



[Fig.8: Distribution of Monthly Charges]

v) Correlation Analysis

Before developing the model, a correlation analysis was conducted to examine relationships among the numerical variables (tenure, monthly charges, and total charges). A correlation heat map was used to display the results of Pearson correlation coefficient calculations. Customers with longer subscription durations incurred higher service expenses, according to the data, which showed a clear positive association between tenure and total charges. Tenure and Monthly Charges had a somewhat weak link, while Monthly Charges and Total Charges showed a moderately positive correlation. These results imply that different facets of consumer behaviour are captured by subscription length and billing features. Overall, there is no evidence of significant multicollinearity among the numerical variables; the correlation analysis indicates that all variables were retained for subsequent model development, providing complementary information for the comparative evaluation of the machine learning models.



[Fig.9: Correlation Heatmap of Numerical Variables]

E. Machine Learning Model Development

To predict customer churn, this study used four supervised machine learning algorithms: Extreme Gradient Boosting, Random Forest, Decision Tree (DT), and Logistic Regression. These models were chosen because they represent widely used classification methods with varying levels of complexity and interpretability. Logistic Regression served as the statistical baseline, and the increasingly complex tree-based machine learning models were Decision Tree, Random Forest, and XGBoost. All models were created with an 80:20 stratified train–test split and the same preprocessed dataset to guarantee a fair comparison. While the testing data was set aside for performance assessment, the training data were used to train the models. The models were implemented using a standard machine learning package in the R programming environment. The predictive performance of each model was evaluated using multiple classification metrics, including accuracy, precision, recall, and F1 Score. The following subsections briefly describe the theoretical foundations and implementation of each machine learning algorithm used in this study.

i) Logistic Regression

Logistic Regression is a supervised binary classification algorithm that estimates the probability of customer churn using the logistic (sigmoid) function. The model is expressed as

$$p(Y = 1|X) = 1 \Big/ (1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)))$$

where (P(Y=1|X)) is the probability of customer churn, β_0 is the intercept, β are the regression coefficients, and (Xi) represent the predictor variables. Logistic Regression was implemented in **R** using the **glm ()** function with the **binomial** family.

ii) Decision Tree

Decision Tree is a supervised learning algorithm that classifies observations by recursively partitioning the dataset into homogeneous groups [23]. The CART algorithm was used with the Gini Impurity Index as the splitting criterion, defined as: $\text{Gini}=1-\sum(p_i)^2$ where p_i denotes the proportion of observations belonging to the i-th class, the model was implemented in R using the rpart package and evaluated on the test dataset.

iii) Random Forest

Random Forest (RF) is an ensemble learning algorithm that combines multiple decision trees built from bootstrap samples and random subsets of predictor variables. The final prediction is obtained by majority voting: $\hat{y}_i = Mode\{T_1(x), T_2(x), T_3(x) \ldots . T_B(x)\}$ where $\hat{y}_i$ is the predicted class, $T_i(x)$ are the predictions of individual trees, and B is the number of trees. The model was implemented using the Random Forest package in R.

iv) Extreme Gradient Boosting (XG Boost)

Extreme Gradient Boosting is a gradient boosting algorithm that sequentially builds decision trees to minimise prediction errors

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

while controlling model complexity through regularisation. Its objective function is: $\text{obj} = L + \Omega$ where L is the loss function and on and Ω is a regularisation term. The model was implemented using the xgboost package in R and evaluated on the test set.

F. Model Evaluation Metrics

The predictive performance of the four machine learning models was evaluated using the confusion matrix and the Receiver Operating Characteristic (ROC) curve. The same evaluation framework was applied to all models to ensure a fair comparison. The confusion matrix classifies predictions into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Based on these outcomes, the following metrics were calculated: Accuracy, Precision, and Recall. $\frac{TP}{TP+FN}$ The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC), with higher AUC values indicating better classification performance, were used to evaluate the model's discriminative performance further. Together, these metrics provide a comprehensive basis for comparing the predictive performance of the four machine learning models.

IV. RESULTS AND DISCUSSION

A. Descriptive Statistics

The cleaned dataset contained 7,032 customer records after removing missing values. Table-II summarises the descriptive statistics of the numerical variables. The average customer tenure was 32.42 months, while the mean monthly and total charges were 64.80 and 2,283.30, respectively. The variability in these variables indicates differences in customer subscription duration and service usage, making them important predictors of customer churn.

Table-I: Descriptive Statistics of Numerical Variables

Variable	N	Mean	Standard Deviation	Minimum	Maximum	Median	Skewness	Kurtosis
Tenure (months)	7,032	32.42	24.55	1	72	29	0.24	-1.39
Monthly Charges	7,032	64.8	30.09	18.25	118.75	70.35	-0.22	-1.26
Total Charges	7,032	2,283.30	2,266.77	18.8	8,684.80	1,397.47	0.96	-0.23

Source: *Computed by the Authors Using the IBM Telco Customer Churn Data Set*

Table-I: Summarizes the descriptive statistics of the numerical variables. The average customer tenure was 32.42 months, while the mean monthly and total charges were 64.80 and 2,283.30, respectively. The variation in total charges reflects differences in customer tenure and service utilization. Table-III presents the distribution of the categorical variables. Male and female customers were almost equally represented, while month-to-month contracts, fibre optic internet service, and electronic check were the most common categories. The target variable comprised 5,163 (73.4%) non-churn customers and 1,869 (26.6%) churn customers, indicating a moderate class imbalance that remained suitable for supervised machine learning.

Table-II: Distribution of Customer Characteristics

Variable	Category	Frequency	Percentage (%)
Gender	Female	3,483	49.53
	Male	3,549	50.47
Contract Type	Month-to-month	3,875	55.11
	One year	1,472	20.93
	Two years	1,685	23.96
Internet Service	DSL	2,416	34.36
	Fiber optic	3,096	44.03
	No Internet Service	1,520	21.62
Payment Method	Bank transfer (automatic)	1,542	21.93
	Credit card (automatic)	1,521	21.63
	Electronic check	2,365	33.63
	Mailed check	1,604	22.81
Customer Churn	No	5,163	73.42
	Yes	1,869	26.58

Source: *Computed by the authors using the IBM Telco Customer Churn data set*

The distribution of the main categorical variables is shown in Table-II. The dataset included nearly equal numbers of male and female customers. The most prevalent categories were electronic check payments (33.63%), fibre optic internet service (44.03%), and month-to-month contracts (55.11%). A moderate level of customer churn was also indicated by the 26.58% of users who had stopped using the service. Overall, the descriptive statistics indicate that the data set provides sufficient variation in contractual, financial, service, and demographic attributes to support the development and comparative assessment of the machine learning models. Additionally, the observed distribution shows that rather than a single dominant feature, a variety of demographic, subscription, and payment-related characteristics drive customer churn. Predictive modelling is more reliable when categories are balanced and well represented, as this reduces the likelihood of bias during model training. Additionally, the variety of contract kinds, online services, and payment options provides useful data for identifying consumer behaviour trends linked to churn. These descriptive insights establish a solid basis for feature selection, model development, and performance assessment. As a result, the dataset is well-suited for developing reliable and broadly applicable machine learning models to predict customer churn.

B. Comparative Performance of Machine Learning Models

© Copyright: All rights reserved.

Table-III: Comparative Performance of Machine Learning Models

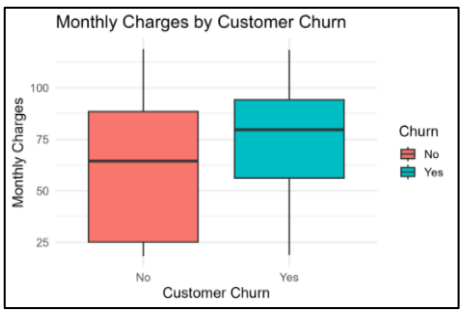
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	82.3	70.27	57.14	63.03
Decision Tree	81.75	69.48	47.53	56.45
Random Forest	81.89	68.36	48.05	56.44
XGBoost	81.47	69.03	51.1	58.73

Source: *Computed by the authors using the IBM Telco Customer Churn data set*

With an accuracy of 82.30%, precision of 70.27%, recall of 57.14%, and an F1-score of 63.03%, logistic regression performed best overall. The Decision Tree, Random Forest, and XGBoost models achieved similar results, but their recall and F1-scores were lower, suggesting reduced ability to identify customers who churned or to detect customer churn. Overall, the findings indicate that the best model for the IBM Telco Customer Churn dataset was Logistic Regression. The characteristics of this dataset favoured the more straightforward Logistic Regression model, despite the general expectation that ensemble approaches would achieve higher predictive accuracy. Instead of assuming that more sophisticated models will always perform better, these results highlight the importance of objectively evaluating multiple techniques.

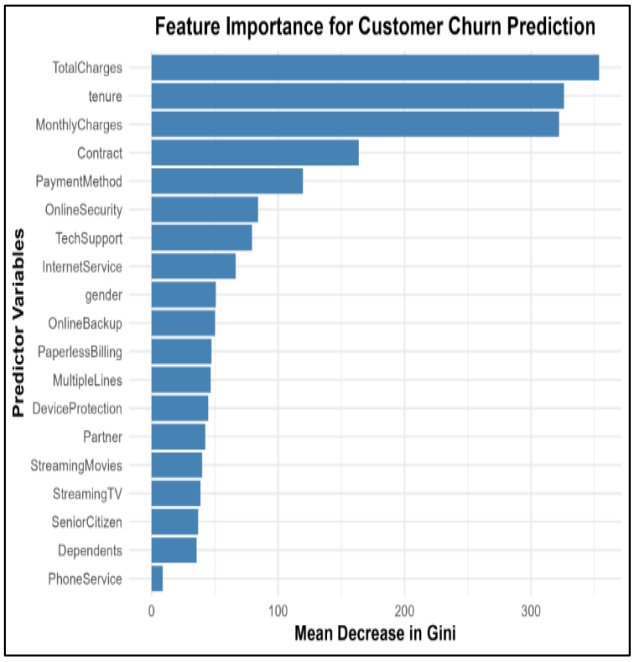
C. Explainable Machine Learning Analysis

For practical decision-making in customer relationship management, model interpretability is crucial, even while predictive accuracy is important. The most significant factors influencing customer churn, according to the feature importance analysis, are contract type, customer tenure, monthly charges, total charges, and internet service. Customers with month-to-month contracts, shorter tenure, and higher monthly charges are more likely to churn, according to the findings. Customers with longer tenure and long-term contracts, on the other hand, have a lower probability of churn, indicating greater customer loyalty. Churn prediction is also influenced by the type of internet service, indicating that service characteristics influence customer retention. By identifying consumer segments at increased risk of turnover, these studies give managers useful insights. Instead of using uniform marketing techniques, organisations can use this information to create specialised retention strategies, such as pricing plans, contract renewal incentives, and personalised offers. In general, implementing explainable machine learning into predictive modelling enhances model transparency and facilitates evidence-based decision-making for efficient customer churn management.



[Fig.10: Box Plot of Monthly Charges by Customer Churn]

The figure above shows that customers on month-to-month contracts have a much higher churn rate than those on one-year or two-year contracts. These findings highlight the importance of contractual commitment in retaining customers.



[Fig.11: Customer Churn Across Contract Type]

The relative importance of the predictor variables, as determined by the Random Forest model, is shown in Figure 11. The most significant factors that contributed to customer churn prediction were contract type, customer tenure, monthly charges, and total charges.

D. Managerial Implications

The findings show that by identifying customers at high risk of churn, predictive analytics can support proactive customer retention efforts. Logistic Regression achieved the highest overall performance, suggesting that accurate and interpretable churn prediction can be made without the need for extremely complex models. The primary causes of customer attrition were identified through feature importance analysis as contract type, tenure, monthly charges, total charges, and internet service. These results enable managers to implement targeted retention strategies, such as encouraging long-term contracts, offering personalised pricing or loyalty rewards, and enhancing service quality for high-risk customer segments. In addition to telecommunications, the proposed R-based framework provides a practical, reproducible decision-support tool for other customer-centric industries, including banking, insurance, healthcare, and e-commerce.

V. CONCLUSION

This study used the IBM Telco Customer Churn dataset in R to evaluate four machine learning models for customer churn prediction: Logistic Regression, Decision Tree, Random Forest, and XGBoost. Data preprocessing, exploratory data analysis, model development, performance evaluation, and feature importance

© Copyright: All rights reserved.

analysis were all part of a systematic approach. All four models performed satisfactorily; the results showed that Logistic Regression achieved the highest overall accuracy. Additionally, the analysis identified the most significant factors influencing customer churn as contract type, customer tenure, monthly price, total charges, and internet service. These results provide useful information for developing targeted customer retention plans. The proposed R-based framework supports data-driven decision-making and provides a practical, transparent, and reproducible method for predicting client attrition. However, the research is restricted to four supervised machine learning techniques and a single telecoms dataset. Future studies might compare other machine learning and deep learning approaches, validate the framework on larger, more varied datasets, and employ sophisticated explainability techniques such as SHAP and LIME. Overall, the results show that managerial decision-making can be enhanced and effective customer attrition prediction can be supported by combining explainable machine learning with predictive analytics.

DECLARATION STATEMENT

Some of the cited references are older and are noted explicitly as [12]. However, these works remain significant for the current study, as they are pioneering in their fields.

As the article's author, I must verify the accuracy of the following information after aggregating input from all authors.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted objectively and without external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate, with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible at: https://www.kaggle.com/datasets/blastchar/telco-customer-churn.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. DOI: 10.1109/ACCESS.2018.2870052
2. Alloghani, M., Al-Jumeily, D., Hussain, A., Aljaaf, A. J., Mustafina, J., & Petrov, E. (2020). Decision tree algorithms: A survey. In D. Al-Jumeily et al. (Eds.), Artificial Intelligence Review. Springer. https://www.academia.edu/97048115/
3. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. DOI:10.1016/j.inffus.2019.12.012
4. Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. Nature Methods, 15(4), 233–234. DOI:10.1038/nmeth.4642
5. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), Article 832. DOI: 10.3390/electronics8080832
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). DOI: 10.1145/2939672.2939785
7. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. (2024). xgboost: Extreme Gradient Boosting (R package Version 1.7.10.1). https://CRAN.R-project.org/
8. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, Article 6. DOI: 10.1186/s12864-019-6413-7
9. Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data—Evolution, challenges and research agenda. International Journal of Information Management, 48, 63–71. DOI: 10.1016/j.ijinfomgt.2019.01.021
10. Géron, A. (2023). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.
11. Hastie, T., Tibshirani, R., & Friedman, J. (2021). The elements of statistical learning: Data mining, inference, and prediction (2nd corrected ed.). Springer. DOI: 10.1007/978-0-387-84858-7
12. IBM. (2013). Telco customer churn [Data set]. Kaggle. https://www.kaggle.com. Work remains significant, see the declaration
13. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. DOI: 10.1007/978-1-0716-1418-1
14. Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice (2nd ed.). Morgan Kaufmann. DOI: 10.1016/C2018-0-01769-6
15. Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. Journal of Marketing, 80(6), 36–68. 10.1509/jm.15.0414
16. Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. Journal of Marketing, 80(6), 69–96. DOI: 10.1509/jm.15.0420
17. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/2017/
18. Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A review on machine learning methods for customer churn prediction and recommendations for business practitioners. IEEE Access, 12, 70434–70463. DOI: 10.1109/ACCESS.2024.3402092
19. Molnar, C. (2022). Interpretable machine learning (2nd ed.). Leanpub. https://christophm.github.io/interpretable-ml-book/
20. Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery, 9(3), e1301. DOI: 10.1002/widm.1301
21. R Core Team. (2025). R: A language and environment for statistical computing (Version 4.5.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/
22. Sarker, I. H. (2022). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 3, Article 160. DOI: 10.1007/s42979-021-00865-8
23. Therneau, T., & Atkinson, B. (2023). rpart: Recursive partitioning and regression trees (R package Version 4.1-21). https://CRAN.R-project.org/package=rpart

AUTHOR'S PROFILE

Dr. Uppu Venkata Subbarao is a Professor of Statistics in the Department of Basic Science and Humanities at NSRIT (Autonomous), Visakhapatnam, Andhra Pradesh, India. He earned his PhD in Statistics from Andhra University, Visakhapatnam. With over 30 years of experience in teaching, research, and academic administration, he has made significant contributions to higher education and statistical research. Prior to joining NSRIT, he served as an Assistant Professor at Mekelle University, Ethiopia (2010–2014); an Associate Professor at Adigrat University, Ethiopia (2014–2018); and an Associate Professor at Asmara University, Eritrea. (2019–2021). He has acted as a Convener and Resource Person at several national and international conferences and has actively contributed to academic and professional development initiatives. He has guided more than 50 research projects in Statistics, Data Science, and Machine Learning. His research interests include statistical modelling, applied statistics, data

© Copyright: All rights reserved.

analytics, and machine learning. He has published over 25 research papers in reputed national and international journals and serves as a reviewer for several scholarly journals. He possesses strong expertise in statistical software packages such as R, SPSS, and Stata, with extensive experience in statistical modelling and data analysis. He is a Life Member of professional bodies including ISTE, ISPS, and IAENG, and remains committed to advancing research, innovation, and excellence in statistical education. activities for the past 30 years. He is a statistics reviewer for several national and international journals and has published 25 articles in various journals.

Dr. T. Narayana Rao is working as Assistant Professor & In-charge HOD in the Department of Management Studies at NSRIT (Autonomous) College, Visakhapatnam, Andhra Pradesh, India. He was awarded his PhD in May 2026 for his research work on "Stress Management among Teachers". Dr Rao is a dedicated researcher, academician, and reviewer for several national and international management journals. He has been actively contributing to academics, research, and student development for the past several years. In recognition of his research contributions, he has received the Young Researcher Award twice. He is a member of INSc, TERA, IAENG and other professional bodies. He regularly participates as a resource person and convener in various workshops, seminars, and conferences, and continues to focus on research, academic excellence, and professional development in the field of management.

Dr. Vantaku Bala has been working as a Professor in the Department of Management Studies at NSRIT Autonomous College, Visakhapatnam. She currently serves as the Head of the MBA Department at NSRIT. She obtained her PhD in Management Studies from Andhra University and has successfully guided more than 50 research projects. Her areas of expertise include Human Resource Management, Marketing, and Finance. She has over two decades of experience in teaching, research, and academic administration. She has published several research papers in reputed national and international journals and has presented papers at various conferences. She actively participates in faculty development programs, workshops, and seminars to promote academic excellence. Her research interests include organisational behaviour, strategic human resource management, and sustainable business practices. She is committed to mentoring students and fostering innovation, leadership, and industry-oriented learning among management graduates.

K. B. Rajeswara Rao is an Assistant Professor in the Department of Computer Science and Engineering at LENDI (Autonomous), Andhra Pradesh. He holds an M. Tech in Computer Science and Engineering from JNTU-GV and an M.Sc. in Computer Science & Statistics from Andhra University. With over 16 years of combined experience in academia, industry, and professional training, he has developed strong expertise in teaching, research, and software technologies. His teaching interests include C Programming, C++, Python, Data Structures, Cloud Computing, Database Management Systems, Operating Systems, and Business Statistics. His research focuses on Artificial Intelligence, Machine Learning, Deep Learning, Computer Vision, Cybersecurity, Data Science, and AI-enabled healthcare applications. He has published research papers, authored a book chapter, and held patents in emerging technology domains. He actively participates in Faculty Development Programs and workshops on Artificial Intelligence, Machine Learning, Cybersecurity, Outcome-Based Education, and Research Methodology.

APPENDIX

A. Variables and Their Measurement Scales

Table-I: Study Variables and Their Measurement Scales

Variable	Description	Type
Customer ID	Unique customer identifier	Identifier
gender	Customer gender	Categorical
Senior Citizen	Senior citizen status (0/1)	Binary
Partner	Customer has a partner	Binary
Dependents	Customer has dependents	Binary
Tenure	Number of months with the company	Numerical
Phone Service	Subscription to phone service	Binary
Multiple Lines	Multiple phone lines	Categorical
Internet Service	Type of internet service	Categorical
Online Security	Online security service	Categorical
Online Backup	Online backup service	Categorical
Device Protection	Device protection plan	Categorical
TechSupport	Technical support service	Categorical
Streaming TV	Streaming TV subscription	Categorical
Streaming Movies	Streaming movie subscription	Categorical
Contract	Contract type	Categorical
Paperless Billing	Paperless billing status	Binary
Payment Method	Customer payment method	Categorical
Monthly Charges	Monthly service charges	Numerical
Total Charges	Total charges incurred	Numerical
Churn	Customer churn status (Target Variable)	Binary

Source: *IBM Telco Customer Churn Dataset collected from Kaggle*

B. Data Source:

The dataset used in this study was obtained from the Kaggle repository:
Dataset: Telco Customer Churn Dataset
Source: Kaggle
URL: BlastChar. *Telco Customer Churn*. Kaggle. Available at: https://www.kaggle.com/datasets/blastchar/telco-customer-churn.
The machine learning models described in this paper were developed and evaluated using this dataset, which includes customer demographics, service subscriptions, account, and churn-related factors.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Publisher /Journal and/or the editor(s). The Publisher /Journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

© Copyright: All rights reserved.