International Journal of Engineering Science Technologies
March-April 2026 10(2)
www.ijoest.com

Original Article

CUSTOMER CHURN PREDICTION IN TELECOM USING MACHINE LEARNING AND DATA MINING

Smita Pandey [1*], **Dr. Shashank Swami** [2]
[1] Research Scholar, Vikrant University Gwalior, Madhya Pradesh, India
[2] Professor, Department of Computer Science and Engineering Vikrant University, Gwalior Madhya Pradesh, India

ABSTRACT

Customer churn prediction is crucial in reducing the loss of customers and enhancing the retention strategies by telecom companies. This paper suggests a machine learning-based system in the case of the IBM Telco Customer Churn data to identify the probability of a customer switching the service. The methods of data preprocessing, including dealing with missing values, coding of categorical variables, log transformation, and feature scaling are used to improve the quality of data. Exploratory Data Analysis (EDA) will be performed to find some trends and factors that are important in churn. There are several supervised learning models, such as Decision Tree, Random Forest, and X GBoost that are implemented and evaluated. Random Oversampling is used to deal with the issue of class imbalance in order to enhance the model performance on minority class examples. Training and testing accuracy is used to evaluate the models with the ensemble models (Random Forest and XG Boost) performing well and generalizing better than the Decision Tree model. The findings show that the type of contract, technical support and payment method are important factors influencing the customer churn, which means that machine learning techniques are quite helpful in the customer retention strategies of the telecom industry.

Keywords: Customer Churn Prediction, Machine Learning, Random Forest, XG Boost, Decision Tree, Data Preprocessing, Exploratory Data Analysis, Class Imbalance, Oversampling, Telecom Analytics

INTRODUCTION

The telecommunication sector has experienced a high rate of development with the emergence of digital technology, as people no longer need to rely on the old method of communication via voice but rather the high data speeds. This change has raised competition among the service providers and has had a bigger impact on customer expectations Zhang and Zhang (2022). Customer retention has become a primary issue in this type of competitive atmosphere as the constant change of the users has a direct influence on the income and sustainability in the long-term Kumar and Mehta (2023).

Customer churn is the number of customers leaving services of a telecom company. It is mostly motivated by elements like prices, quality of services and presence of superior alternatives Sharma and Roy (2023). A high churn rate results in loss of revenue and higher cost of acquiring customers and therefore churn management is a major concern to telecom companies Verma and Raj (2023).

*Corresponding Author:
Email address: Smita Pandey (Smitapandey22@gmai.com)
Received: 15 February 2026; **Accepted:** 23 March 2026; **Published** 20 April 2026
DOI: 10.29121/IJOEST.v10.i2.2026.752
Page Number: 74-81
Journal Title: International Journal of Engineering Science Technologies
Journal Abbreviation: Int. J. Eng. Sci. Tech
Online ISSN: 2456-8651
Publisher: Granthaalayah Publications and Printers, India
Conflict of Interests: The authors declare that they have no competing interests.
Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
Authors' Contributions: Each author made an equal contribution to the conception and design of the study. All authors have reviewed and approved the final version of the manuscript for publication.
Transparency: The authors affirm that this manuscript presents an honest, accurate, and transparent account of the study. All essential aspects have been included, and any deviations from the original study plan have been clearly explained. The writing process strictly adhered to established ethical standards.

TELECOM CHALLENGES AND CHURN IMPACT

The shift to data-based services and the emergence of digital platforms have increased the level of competition in the telecom industry. The challenges encountered by companies include decreasing traditional revenue, high operational cost and saturation in the market. The above factors render customer retention hard to achieve, particularly where users have the freedom to change to competitors providing superior value Khan and Ahmed (2023).

Churn does not only decrease revenue but also market share and customer loyalty. It is cheaper to retain existing customers compared to acquiring new customers, a fact that underscores the need to have effective churn management strategies Zhao and Zhang (2023).

ROLE OF MACHINE LEARNING IN CHURN PREDICTION

Data mining and machine learning algorithms are critical in customer churn prediction. Predictive models can determine customers at risk of leaving by examining customer behavior, usage, and interactions with the services Li and Sun (2023).

Such insights would allow telecom companies to go out of their way, with personalized offers and better services to retain customers. Thus, churn prediction helps in enhancing customer satisfaction and improving business performance Brown et al. (2022).

RESEARCH OBJECTIVE

The key aim of the research is to create a machine learning-powered model to predict customer churn within the telecom industry and to measure its efficiency in enhancing customer retention policies.

LITERATURE REVIEW

A number of researches have examined customer churn prediction in the telecommunications industry with the help of various data-driven and machine learning methods. The next review is a summary of major contributions in this field.

Lee and Chen (2020) analyzed the impact of demographic factors on customer churn prediction. Their analysis pointed out that age, income and location are some of the attributes that play a significant role in churn behavior. They discovered that younger and low-income customers have a higher likelihood of switching their providers with the help of logistic regression and support vector machines. The paper has highlighted that demographic characteristics can be used to enable telecom companies to develop specific retention strategies.

Zhang et al. (2020) concentrated on the usage of techniques of deep learning to predict churn. They were able to show that neural networks and especially multi-layer perceptron models can be used to capture complex patterns in large datasets. They found that their results were better than traditional models but the method is more expensive in terms of the amount of computation required. The paper also emphasized the significance of optimizing hyperparameters in order to achieve the best possible performance

Patel and Zhao (2020) investigated how time series analysis can be used to predict customer churn. Using ARIMA models and historical customer data (usage patterns and payment history) they could have determined trends and seasonal behavior. Their results indicated that time-dependent analysis improves the accuracy of prediction and it can be used to complement machine learning models.

Singh and Sharma (2020) investigated data mining application in the customer behavioral trend. The study found out that low service usage, late payment and high complaints are good predictors of churn. They combined these behavioral variables with machine learning algorithms to come up with more precise prediction models and suggested early intervention interventions.

Kumar et al. (2021) have highlighted the significance of feature engineering when it comes to enhancing churn prediction models. Their findings indicated that the right choice of features, scaling and encoding are important to enhance model performance. They also stressed that the time aspect such as customer tenure and recent activity should be included in addition to increase predictive accuracy.

Garcia and Kim (2021) examined the possibility of incorporating social media data into churn prediction models. They discovered that any customer who posts a negative opinion on the internet has a higher probability of churning with the help of sentiment analysis. The research revealed that forecasting is enhanced by the integration of structured customer data and unstructured information on social media.

Nguyen and Tran (2021) paid attention to sentiment analysis through natural language processing techniques. Their analysis found that customer dissatisfaction and negative customer feedback are good predictors of churn. They also used sentiment analysis

and machine learning models together to yield improved accuracy and indicated that emotions of a customer are important to churn management.

RESEARCH METHODOLOGY

This section gives an overview of the general methodology and methodologies to formulate an effective customer churn prediction model. It outlines data, preprocessing methodology and machine learning algorithms used in the research. The methodology is aimed at converting the raw telecom data into valuable insights by performing a systematic data analysis and creating a model. Further, relevant methods are used to deal with the imbalance of data and enhance the precision of predictions, guaranteeing meaningful and effective outcomes.

RESEARCH DESIGN

This paper applies a quantitative research method predicting churn of customers in the telecommunications industry through the use of machine learning. The formulation of the problem can be defined as a binary classification problem, where customers can be classified into churned or retained. Supervised learning techniques are used, since the dataset includes labeled results (a churn variable).

DATASET DESCRIPTION

The dataset utilized in the research is the IBM Telco Customer Churn dataset which comprises of the customer data in terms of demographics, service usage and billing data. The data is in 7,043 records and 21 features, with each record corresponding to a single customer.

The variable of interest is the so-called Churn that implies whether the customer has ceased using the service. Such key attributes as: are present in the dataset.

- Demographic (gender, senior citizen, dependents)
- Features Service-related (internet service, online security, tech support)
- Information on account (tenure, type of contract, payment method, charges)

All these help in determination of patterns that are related to customer attrition.

DATA PREPROCESSING

Data preprocessing was done to guarantee the quality and appropriateness of data to machine learning models.

1) **Handling Missing Values:** The TotalCharges column was converted to a numeric one and missing values were dealt with accordingly to prevent disparity.
2) **Encoding:**
   - Binary variables were encoded using Label Encoding.
   - Multi-category features (contract type and payment method) were encoded with one-hot.
3) **Feature Transformation:** Log transformation was used on the numerical data such as MonthlyCharges and TotalCharges to eliminate skewness.
4) **Scaling:** Standardization was done on variables like tenure to put all variables on a similar scale.
5) **Data Splitting:** The data was split into training (70) and testing (30) parts to assess the models.

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis was done to get an idea of the structure and distribution of the data. The churn distribution and its association with features (gender, contract type and payment method) were analyzed using various visualization tools including count plots and pie charts.

EDA helped in identifying:

- lass imbalance in the churn variable
- Relationships between customer attributes and churn behavior
- Important features influencing customer retention

MACHINE LEARNING MODELS

Several machine learning algorithms were used to forecast customer churn:

- **Decision Tree:** This is a rule-based model to classify data by dividing data according to feature values.
- **Random Forest:** An ensemble type of learning method which is used to combine various decision trees to enhance accuracy and minimize overfitting.
- **XGBoost:** A gradient boosting type of algorithm with high performance and the capacity to learn complicated patterns within structured data.

The models have been chosen to compare performance on simple and advanced algorithms.

HANDLING IMBALANCED DATA

As the dataset had class imbalance, it used the Random Oversampling technique in order to balance the distribution of churn and non-churn classes. This enhances the model to accurately forecast instances of minorities classes.

Figure 1

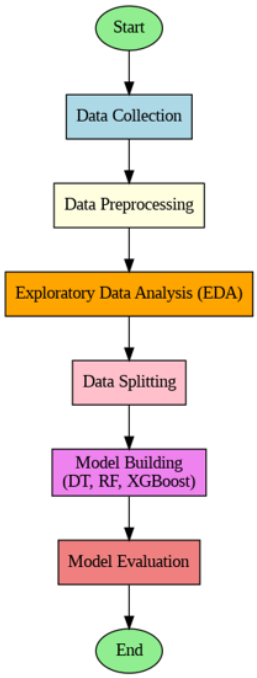

Figure 1 Flowchart of the Proposed Customer Churn Prediction Methodology

RESULTS AND DISCUSSION

This section includes the findings of the data that was implemented to predict customer churn using machine learning models. The main purpose is to review the performance of the models and also determine the major factors that affect customer attrition. Exploratory Data Analysis (EDA) is conducted to comprehend customer behavior, and then model evaluation is conducted based on performance measures of accuracy, precision, recall, and F1-score. The findings have valuable implications to enhance customer retention in the telecom industry.

EXPLORATORY DATA ANALYSIS

DISTRIBUTION OF CHURN

The data is imbalanced in terms of the classes with the non-churn customers by far surpassing the churn customers. Such imbalance is also worthy to note because it may affect the performance of the models when training.

Figure 2

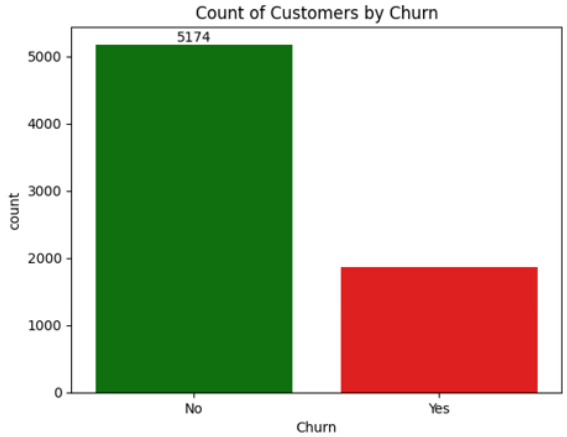

Figure 2 Distribution of Churned vs Non-Churned Customers

It is evident that most customers are under the non-churn category meaning that there is an imbalance in the distribution of classes.

CHURN BY CONTRACT TYPE

Contract type is one of the factors that determine customer retention and long term engagement. Varying contract periods indicate the level of commitment by customers.

Figure 3

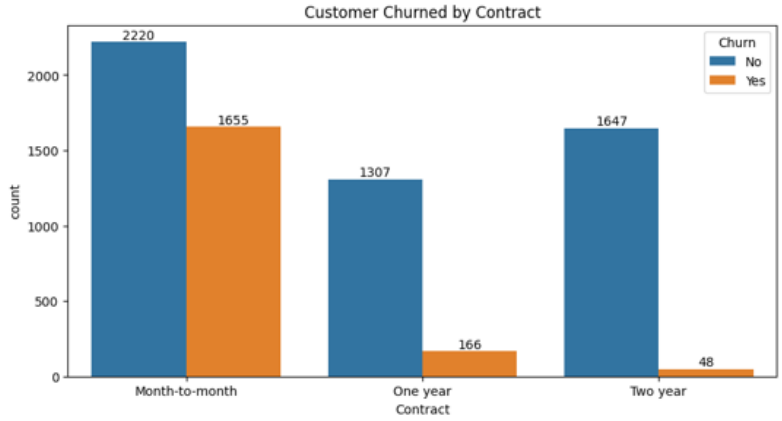

Figure 3 Customer Churn by Contract Type

The figure depicts that the churn is more among customers whose contracts are month to month and the churn is lower in the case of long term contracts.

CHURN BY TECH SUPPORT

Customer satisfaction and service experience are greatly influenced by technical support. Availability of support services can impact customer decisions to stay or leave.

Figure 4

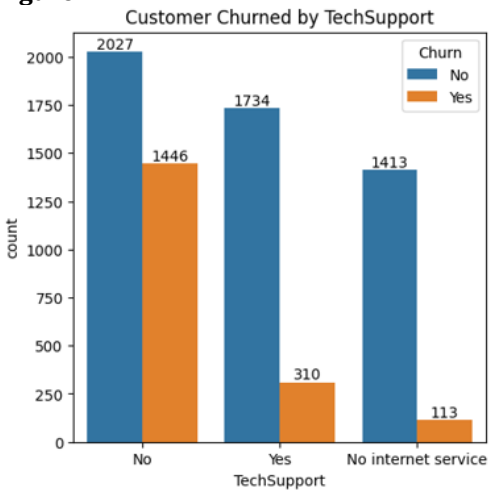

Figure 4 Customer Churn by Tech Support

The figure shows the churn rate of customers who do not get tech support is higher than the churn rate of customers who get the support services.

CHURN BY PAYMENT METHOD

Customer convenience, reliability and behavior of using a service are affected by modes of payment. Various approaches can have different impacts on customer retention.

Figure 5

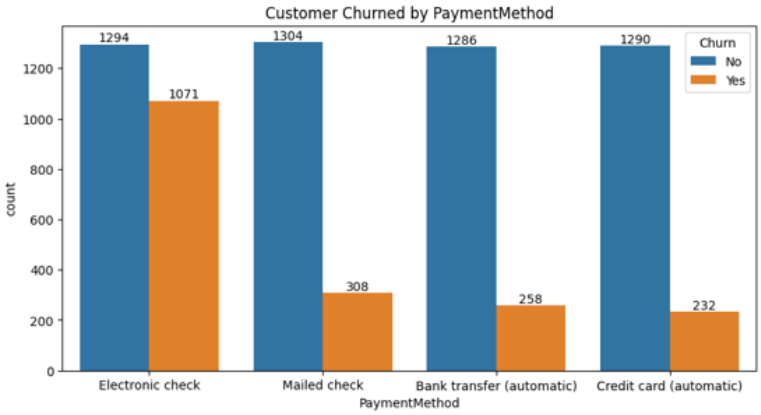

Figure 5 Customer Churn by Payment Method

The graph indicates that the number of customers on electronic check is more prone to churn as compared to the automatic payment method.

CHURN BY ONLINE SECURITY

Security services help in the customer trust and the perceived quality of the service. When customers feel safe with their data and services, they will be more inclined to remain.

Figure 6

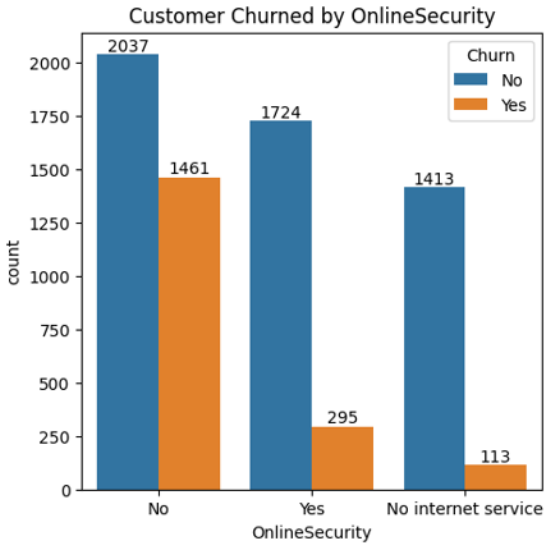

Figure 6 Customer Churn by Online Security

The number underscores the fact that customers who lack online security services are more likely to churn as opposed to the customers who utilize security services.

MODEL PERFORMANCE

DECISION TREE

Decision Tree Supervised learning algorithm: This algorithm is used to classify data with the help of a tree-like structure on the basis of feature splits. It is easy and straightforward to interpret yet can be overfitted, impacting its accuracy with new data.

RANDOM FOREST

Random Forest is an ensemble method, which is a combination of several decision trees to enhance accuracy. It minimizes overfitting and offers more accurate and consistent predictions as compared to one decision tree.

XGBOOST

XGBoost is a superior boosting algorithm which is designed to build up models in sequence to enhance performance. It is fast and can deal with complicated patterns and is extensively employed in making predictions of high accuracy.

MODEL COMPARISON

Table 1

Model	Training Accuracy	Testing Accuracy
Decision Tree	99.83%	73.12%
Random Forest	99.83%	79.55%
XGBoost	93.98%	79.51%

DISCUSSION

The findings demonstrate that the Decision Tree model is vulnerable to overfitting and has a weak generalization ability. Random Forest and XGBoost, in turn, have better performance and reliability.

Random Forest was found to have the best testing accuracy hence the best churn prediction model. XGBoost also worked well because of the regularization ability.

The outcomes of EDA show that the churn depends on the type of contract, technical support, the way of payment and online security. Customers on short term contract, less supported and less secure are more likely to change service providers.

KEY FINDINGS

- Customer churn is strongly influenced by service quality and contract type
- Class imbalance impacts model performance
- Random Forest provides the best predictive performance
- Technical support and security services play a crucial role in reducing churn

CONCLUSION

This paper used the IBM Telco Customer Churn dataset to show a machine learning-based method of predicting customer churn in the telecommunications industry. Different preprocessing methods were used to enhance the quality of data such as addressing missing values, encoding, feature transformation, and scaling. Exploratory Data Analysis (EDA) aided in the discovery of the noteworthy trends and the main elements affecting the customer churn, including the type of contract, technical support, payment method, and online security.

Several classification models, which included Decision Tree, random Forest, and XGBoost were applied and compared. The findings show that though the Decision Tree model had a high training accuracy, it had the disadvantage of overfitting and lower performance on test data. Conversely, the ensemble approaches like the Random Forest and XGBoost showed more generalization and predictive accuracy. Random Forest was the most appropriate model to use in this study due to its performance being the highest of all models.

Moreover, Random Oversampling was effective in correcting the problem of class imbalance and enhancing the effectiveness of the model in predicting churned customers. In sum, the results prove that machine learning methods could be effective in identifying the customers who may leave and helping telecom firms to build data-driven customer retention models.

ACKNOWLEDGMENTS

None.

REFERENCES

Brown, C., Wilson, E., and Taylor, D. (2022). The Impact of Customer Lifetime Value on Churn Prediction.

Garcia, M., and Kim, Y. (2021). Utilizing Social Media Data for Churn Prediction. Social Media Analytics Journal, 18(2), 45–63.

Khan, M., and Ahmed, R. (2023). Churn Prediction in Telecom: A Comparative Study of Classical Machine Learning Algorithms. International Journal of Data Mining and Applications, 18(3), 143–155.

Kumar, A., and Mehta, S. (2023). Improving Churn Prediction Accuracy with Hybrid Deep Learning Models. Journal of Machine Learning in Business, 16(2), 234–245.

Kumar, P., Singh, R., and Verma, K. (2021). Feature Engineering for Churn Prediction. Machine Learning Applications in Telecom, 29(1), 54–71.

Lee, H., and Chen, J. (2020). Influence of Customer Demographics on Churn Prediction. Telecom Analytics Journal, 19(3), 67–80.

Li, Z., and Sun, J. (2023). A Survey of Churn Prediction Models in Telecommunications. Journal of Network and Computer Applications, 15(2), 101–115.

Nguyen, T., and Tran, V. (2021). Sentiment Analysis for Predicting Customer Attrition. AI and Customer Insights, 22(1), 55–70.

Patel, S., and Zhao, L. (2020). Forecasting Attrition Utilizing Temporal Data. Time Series Analytics in Business, 16(1), 33–49.

Sharma, R., and Roy, S. (2023). The Significance of Customer Lifetime Value in Churn Analysis.

Singh, V., and Sharma, T. (2020). Analysis of Customer Behavior for Churn Prediction. Big Data and Customer Analytics, 14(3), 78–92.

Verma, S., and Raj, N. (2023). Predicting Telecom Churn Using Neural Network-Based Approaches. Computational Intelligence in Telecommunications, 22(7), 77–92.

Zhang, H., and Zhang, Y. (2022). Customer Churn Prediction in Telecom Using Random Forest and XGBoost models. Journal of Data Science Applications, 18(4), 202–213.

Zhang, W., Li, H., and Zhou, K. (2020). Churn Prediction Utilizing Neural Networks. Deep Learning for Business, 17(4), 98–115.

Zhao, Z., and Zhang, W. (2023). Customer Churn Prediction in Telecommunications Using Deep Learning and Ensemble Methods. Telecom Research Journal, 8(4), 220–234.