STM JOURNALS

Journal of
Computer Technology & Applications

ISSN: 2229-6964 (Online)
ISSN: 2347-7229 (Print)
Volume 16, Issue 2, 2025
May–August
DOI (Journal): 10.37591/JoCTA

https://journals.stmjournals.com/jocta

Review JoCTA

Prediction of Customer Churn Using Machine Learning Classification Models

Nikita Khandelwal[1,*], Vikas Sakalle[2]

Abstract

Customer churn prediction is a critical task in both the telecommunication and medical industries, where retaining customers or patients is essential for ensuring long-term profitability and maintaining high-quality service. To address this, a range of machine learning models—including logistic regression, decision trees, random forests, gradient boosting machines, and support vector machines—were employed to accurately forecast churn behavior. Prior to model training, the dataset underwent thorough preprocessing, which included handling missing values, normalizing numerical features, and encoding categorical variables to prepare the data for analysis. Class imbalance, a common challenge in churn datasets, was mitigated using SMOTE (synthetic minority over-sampling technique), resulting in a more balanced dataset for effective model learning. Feature selection techniques such as recursive feature elimination (RFE) and mutual information were utilized to pinpoint the most influential predictors of churn, enhancing model interpretability and performance. The models were evaluated using comprehensive metrics including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC), offering a multifaceted view of each model's effectiveness. Overall, the combination of robust preprocessing, balanced training data, and diverse evaluation metrics contributed to the development of reliable and generalizable churn prediction models.

Keywords: Customer churn, telecommunications, medical industry, machine learning, financial considerations, SMOTE (synthetic minority over-sampling technique)

INTRODUCTION

Customer churn, which refers to the loss of customers who discontinue using a service or product, presents a major challenge for businesses in multiple sectors. In the telecommunications industry, where competition is intense and the cost of acquiring new customers is steep, retaining current customers is essential for sustaining profitability and market position [1]. Likewise, in the healthcare sector, keeping patients engaged is critical not only for financial viability but also for ensuring consistent and quality care. Predicting customer churn allows businesses to proactively address issues that lead to dissatisfaction and disengagement, thereby implementing strategies to enhance customer loyalty and retention. The emergence of machine learning (ML) has created new opportunities for accurately forecasting churn by utilizing vast amounts of data and advanced algorithms [2].

***Author for Correspondence**
Nikita Khandelwal
E-mail: nikitakhandelwal0000@gmail.com

[1]Research Scholar, Department of Computer Science and Engineering, LNCT University, Bhopal, Madhya Pradesh, India
[2]Associate Professor, Department of Computer Science and Engineering, LNCT University, Bhopal, Madhya Pradesh, India

Received Date: February 04, 2025
Accepted Date: April 16, 2025
Published Date: April 24, 2025

Citation: Nikita Khandelwal, Vikas Sakalle. Prediction of Customer Churn Using Machine Learning Classification Models. Journal of Computer Technology & Applications. 2025; 16(2): 86–92p.

ML classification models, which are designed to categorize data into predefined classes, have proven to be particularly useful for churn prediction. These models can analyze complex patterns and relationships within data, making them ideal for identifying the factors that contribute to customer churn. The use of ML in churn prediction involves

several key steps: data collection, preprocessing, feature selection, model training, and evaluation. Every one of these steps is essential to guarantee that the models are both precise and dependable [3].

Data Collection

In both the telecommunication and medical industries, data is generated from various sources. For telecommunication companies, this includes call records, billing information, service usage patterns, customer demographics, and interaction logs [4].

Data Pre-processing

Data pre-processing is an essential process that involves cleaning and converting raw data into a format ready for analysis. This process includes addressing missing values, normalizing numerical features, and encoding categorical variables. In the telecommunication industry, for example, customer records may have missing billing information or incomplete service usage logs [5].

Feature Selection

Feature selection involves identifying the key variables that play a significant role in predicting churn. This step is vital as it helps reduce data complexity, enhances model performance, and prevents overfitting [6].

Model Training

After selecting the relevant features, the next step is to train the ML models. Various ML classification techniques, such as logistic regression, decision trees, random forests, gradient boosting machines, and support vector machines, can be used for churn prediction. Each algorithm has its own advantages and limitations, with the choice depending on the specific nature of the dataset and the problem being addressed [7].

Evaluation

Evaluating the model is a crucial step to determine the performance of the trained models. Metrics like accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) are used to assess the models' effectiveness [8].

Telecommunication Industry Application

In the telecommunications sector, churn prediction models can help implement focused retention strategies. For example, if a model identifies a group of customers at risk of leaving due to billing problems, the company can proactively engage with these customers by offering personalized deals or support [9].

Medical Industry Application

In the medical industry, predicting patient churn is essential for maintaining continuous and effective healthcare delivery. Patient churn can lead to disruptions in treatment, poorer health outcomes, and financial losses for healthcare providers. ML models can help identify patients who are at risk of discontinuing their care, allowing healthcare providers to take proactive measures [10].

The use of ML classification models for customer churn prediction in the telecommunication and medical industries holds great promise for improving customer retention and service quality [11].

BACKGROUND STUDY

Customer churn prediction is a critical aspect of maintaining business sustainability and growth, particularly in the telecommunication and medical industries. These sectors face unique challenges and dynamics that necessitate robust and reliable methods to anticipate and mitigate churn [12]. ML classification models offer powerful tools to address these challenges by analyzing complex datasets to identify patterns and predict customer behavior.

Similarly, in the medical industry, patient churn—when patients discontinue their treatment or switch healthcare providers—can lead to severe consequences. Continuity of care is crucial for effective medical treatment, and interruptions can result in poor health outcomes [13]. Additionally, patient churn can cause financial strain on healthcare providers due to the loss of revenue from ongoing treatments.

The initial step in applying ML for churn prediction is gathering data. Collecting thorough and high-quality data is crucial for training precise models. In the telecommunication industry, data is often collected from customer relationship management (CRM) systems, billing systems, and service usage logs. In the healthcare sector, electronic health records (EHRs) are a key source of data [14].

Assessing the performance of ML models is essential to verify their reliability and effectiveness. Metrics like accuracy, precision, recall, F1-score, and the AUC-ROC are used to evaluate the models. Precision and recall are particularly important in churn prediction, as they provide insights into the model's ability to correctly identify churners and non-churners.

Future directions in machine learning for churn prediction include developing more sophisticated algorithms that can handle large and complex datasets, integrating external data sources such as social media and customer reviews, and using real-time data for dynamic churn prediction [15].

LITERATURE REVIEW

Jajam et al. 2023: This research introduces the ensemble deep learning SBLSTM RNN-IGSA (Stacked Bidirectional Long Short-Term Memory Recurrent Neural Network - Improved Genetic Simulated Annealing) model, which focuses on optimizing arithmetic operations for predicting customer churn. This paper highlights advancements in ensemble deep learning techniques, demonstrating how combining various models can enhance prediction accuracy and efficiency [16].

Nagaraj et al., 2023: Focusing on the e commerce sector, this paper proposes a machine learning scheme for predicting customer churn based on customer behavior. The study reflects the nuances of online shopping behaviors and their implications for customer retention, highlighting the adaptability of machine learning techniques to various industries [17].

Mohamed and Al-Khalifa, 2023: This review paper examines various machine learning methods employed to predict churn in the telecommunications sector [18].

Thakkar et al., 2023: Concentrating on the telecommunications industry, this paper undertakes exploratory data analysis to uncover underlying patterns of customer churn. The authors delve into predictive models to offer solutions for the industry's churn problem, providing practical insights into data-driven strategies for reducing churn rates [9].

Şenyürek and Alp, 2023: This paper focuses on the application of ML methods specifically in the telecommunication sector for predicting churn. It provides insights into the effectiveness of various ML techniques in addressing the churn problem inherent to the industry [19].

Elgohary et al., 2023: This study introduces a smart evaluation approach for deep learning models using churn prediction as a case study. It provides insights into the practical application and performance metrics of deep learning in real-world scenarios, emphasizing the importance of smart evaluation techniques [20].

Naidu et al., 2022: Focusing on the telecom industry, this study employs the random forest algorithm to predict customer churn. Presented at an international conference on computational methods in systems and software, this research evaluates the capabilities of this machine learning technique in addressing retention challenges in telecom [21].

Pandithurai and Sriman, 2023: Investigating the telecom industry, this study employs a voting classifier ensemble method combined with supervised ML techniques. The aim is to create a robust prediction model that captures various facets of customer behaviors leading to churn, enhancing the reliability and accuracy of predictions [22].

Al-Shourbaji et al., 2022: The research aims to enhance churn prediction by boosting ant colony optimization using a reptile search algorithm. This innovative approach employs nature-inspired algorithms to address a pressing industry challenge, showcasing the potential of unconventional techniques in churn prediction [23].

RESEARCH GAP

Limited Algorithms

Many reviews focus predominantly on traditional ML models such as decision trees, support vector machines, and logistic regression. A valuable opportunity for comparative analysis across industries is frequently neglected in the literature. Applying similar churn prediction models to both the telecommunication and medical industries can reveal how different factors influence churn across these sectors.

Data Challenges

While much research emphasizes the development and performance of algorithms, there is often a lack of focus on the practical challenges related to data collection, cleaning, and preprocessing specific to different industries.

Ethical and Privacy Concerns

In industries handling sensitive data, particularly the medical sector, ethical and privacy considerations are paramount [24].

1. *Temporal dynamics:* Factors such as seasonal trends, market fluctuations, and external events (e.g., economic downturns or pandemics) can significantly impact churn rates. Incorporating temporal data into models can improve their predictive power and provide more timely and relevant insights.
2. *Integration with other systems:* The practical integration of churn prediction models with existing operational systems such as customer relationship management (CRM) tools in telecommunications or hospital management systems in healthcare is often overlooked. Addressing how these models can seamlessly interact with other systems to provide actionable insights can enhance their utility and effectiveness [25].
3. *Economic impact:* The direct and indirect economic implications of customer or patient churn are not always deeply explored. Understanding the financial impact of churn and how effective prediction and mitigation strategies can reduce costs is essential for justifying the investment in developing and deploying these models.
4. *Cultural and geographical variations:* Many reviews focus on studies conducted in specific regions, neglecting how cultural and geographical factors can influence churn behavior and prediction models. Exploring these variations can provide more generalizable models and insights that are applicable across different populations and markets.

RESEARCH METHODOLOGY

Logistic Regression

- Logistic regression is a statistical technique used to estimate the probability of a binary outcome based on one or more predictor variables. It employs the logistic function to describe the relationship between the dependent and independent variables.
- *Application:* In churn prediction, logistic regression can be used to estimate the likelihood of a customer churning based on factors such as usage patterns, service issues, and demographic information.

Decision Trees

- *Description:* Decision trees are tree-like models that make decisions based on splitting data into subsets using criteria like information gain or Gini impurity. Each node symbolizes a feature, each branch signifies a decision rule, and each leaf corresponds to a class label.
- *Application:* decision Trees can classify customers as churners or non-churners by analyzing their historical data, such as service usage and interaction history.

Random Forests

- *Description:* Random forests are an ensemble learning technique that builds several decision trees during training and returns the most common class (classification) from the individual trees. It helps in reducing overfitting and improving accuracy.
- *Application:* Random forests can handle large datasets with higher accuracy, making them suitable for churn prediction in both telecommunication and medical industries by combining multiple decision trees' predictions.

Gradient Boosting Machines

- *Description:* Gradient boosting machine (GBM) is an ensemble method that constructs models in a sequence, where each new model addresses the errors made by the previous ones. It utilizes a gradient descent algorithm to minimize the loss function.
- *Application:* GBM can be used for churn prediction by iteratively improving the model's accuracy through boosting, effectively handling complex data patterns.

K-Nearest Neighbors

- *Description:* K-nearest neighbors (KNN) is a non-parametric technique that categorizes data points according to the majority class of the K nearest neighbors. It is straightforward and efficient for small datasets.
- *Application:* KNN can classify customers as churners or non-churners by comparing their attributes to those of the nearest neighbors, identifying patterns in customer behavior.

Artificial Neural Networks

- *Description:* An artificial neural network (ANN) is made up of interconnected layers of nodes (neurons) that process input data using weights and activation functions to identify complex patterns. It is a robust method for capturing non-linear relationships.
- *Application:* ANNs can be used for churn prediction by learning from historical customer data and identifying complex patterns and relationships that indicate the likelihood of churn.

XGBoost (Extreme Gradient Boosting)

- *Description:* XGBoost is an enhanced, distributed gradient boosting library that utilizes a gradient boosting framework. It is highly efficient and scalable, often outperforming other algorithms in competitions.
- *Application:* XGBoost can be used for churn prediction by efficiently handling large datasets and complex data patterns, providing robust predictions with high accuracy.

ADVANTAGES

1. *Holistic overview:* Comprehensive reviews offer a broad perspective on the current state of churn prediction techniques, enabling stakeholders to understand the progression of the field and the methodologies currently in use. This holistic view helps in recognizing how various approaches have evolved over time and their relative effectiveness in different contexts.
2. *Cross-industry insights:* By examining the application of churn prediction models in both the telecommunication and medical industries, readers can gain valuable insights into the unique challenges and opportunities in each domain. This cross-industry analysis helps identify best practices that might be transferable between industries, enhancing the applicability and robustness of churn prediction techniques.

3. *Algorithm comparison:* Reviews provide a detailed analysis of which machine learning models perform best in specific scenarios. This comparison offers companies actionable insights into which algorithms they should consider implementing based on their unique requirements and the nature of their data. Recognizing the advantages and limitations of different models aids in making well-informed choices regarding model selection.
4. *Identification of best practices:* From data preprocessing to model evaluation, reviews highlight best practices in the churn prediction process identified across multiple studies. These best practices serve as guidelines for researchers and practitioners, helping them adopt the most effective methods to enhance model performance and reliability.
5. *Industry-specific challenges:* By differentiating between the telecommunication and medical industries, reviews highlight the unique challenges faced by each sector. This differentiation leads to better-targeted solutions that address specific issues pertinent to each industry, improving the overall effectiveness of churn prediction models.
6. *Promoting collaboration:* Showcasing interdisciplinary efforts between ML experts, telecommunication specialists, and medical professionals, reviews encourage further collaborations between these fields. Such collaborations foster innovation and the development of more robust and versatile churn prediction models.
7. *Enhancing trustworthiness:* A well-conducted review, especially if peer-reviewed, enhances the trustworthiness of the summarized findings. Businesses and researchers can base their decisions and further research on robust evidence, confident that the information is credible and thoroughly vetted.
8. *Stimulating innovation:* By presenting the state of the art in the field, reviews stimulate further innovation by challenging researchers and industry professionals to develop improved methodologies or tackle identified gaps. This stimulation leads to continuous advancements in churn prediction techniques, driving the field forward.

CONCLUSION

The high predictive performance of these models enables organizations to proactively implement targeted retention strategies, such as personalized marketing campaigns and enhanced patient follow-up programs. As a result, companies can tackle potential problems before they result in churn, thereby enhancing customer retention and promoting long-term loyalty.

The insights gained from this research highlight the potential of ML in driving data-driven decision-making processes, ultimately contributing to the growth and stability of companies in these competitive markets. The successful integration of these predictive models into operational workflows can provide a significant competitive advantage, ensuring that both telecommunication providers and healthcare institutions maintain a robust and satisfied customer base.

REFERENCES

1. Jafari MJ, Tarokh MJ, Soleimani P. An interpretable machine learning framework for customer churn prediction: a case study in the telecommunications industry. J Indus Eng Manage Stud. 2023; 10 (1): 141–157.
2. Sudharsan R, Ganesh EN. A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. Connection Sci. 2022; 34 (1): 1855–1876.
3. Oluwatoyin AM, Misra S, Wejin J, Gautam A, Behera RK, Ahuja R. Customer churn prediction in banking industry using Power BI. In: Singh PK, Wierzchoń ST, Tanwar S, Rodrigues JJPC, Ganzha M, editors. Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021. Singapore: Springer Nature; 2022. pp. 767–774..
4. Alshamari MA. Evaluating user satisfaction using deep-learning-based sentiment analysis for social media data in Saudi Arabia's telecommunication sector. Computers. 2023; 12 (9): 170.
5. AlShourbaji I, Helian N, Sun Y, Hussien AG, Abualigah L, Elnaim B. An efficient churn prediction model using gradient boosting machine and metaheuristic optimization. Sci Rep. 2023; 13(1): 14441.

6. Seymen OF, Ölmez E, Doğan O, Er O, Hiziroğlu K. Customer churn prediction using ordinary artificial neural network and convolutional neural network algorithms: a comparative performance assessment. Gazi Univ J Sci. 2023; 36 (2): 720–733.
7. Jeyaprakaash P, Sashirekha K. Accuracy measure of customer churn prediction in telecom industry using Adaboost over random forest algorithm. J Pharm Negative Results. 2022; 13 (4): 1486–1494.
8. Kim S, Lee H. Customer churn prediction in influencer commerce: an application of decision trees. Procedia Computer Sci. 2022; 199: 1332–1339.
9. Thakkar HK, Desai A, Ghosh S, Singh P, Sharma G. Clairvoyant: AdaBoost with cost-enabled cost-sensitive classifier for customer churn prediction. Comput Intell Neurosci. 2022; 2022 (1): 9028580.
10. Rahmaty M, Daneshvar A, Salahi F, Ebrahimi M, Chobar AP. Customer churn modeling via the grey wolf optimizer and ensemble neural networks. Discrete Dynam Nature Soc. 2022; 2022 (1): 9390768.
11. Wu Z, Jing L, Wu B, Jin L. A PCA-AdaBoost model for E-commerce customer churn prediction. Ann Oper Res. 2022. doi: 10.1007/s10479-022-04526-5.
12. Jeyaprakaash P, Sashirekha K. Accuracy measure of customer churn prediction in telecom industry using Adaboost over decision tree algorithm. J Pharm Negative Results. 2022; 13 (4): 1495–1503.
13. Wu X, Li P, Zhao M, Liu Y, Crespo RG, Herrera–Viedma E. Customer churn prediction for web browsers. Expert Syst Appl. 2022; 209: 118177.
14. Mozaffari F, Rahimi M, Yazdani H, Sohrabi B. Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data. Benchmarking Int J. 2023; 30 (10): 4140–4173.
15. Matuszelański K, Kopczewska K. Customer churn in retail e–commerce business: spatial and machine learning approach. J Theor Appl Electron Commerce Res. 2022; 17 (1): 165–198.
16. Jajam N, Challa NP, Prasanna KS, Deepthi CV. Arithmetic optimization with ensemble deep learning SBLSTM-RNN-IGSA model for customer churn prediction. IEEE Access. 2023; 11: 93111–93128.
17. Nagaraj P, Muneeswaran V, Dharanidharan A, Aakash M, Balananthanan K, Rajkumar C. E-commerce customer churn prediction scheme based on customer behaviour using machine learning. In: 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, January 23–25, 2023. pp. 1–6.
18. Mohamed FA, Al-Khalifa AK. A review of machine learning methods for predicting churn in the telecom sector. In: 2023 International Conference on Cyber Management and Engineering (CyMaEn), Bangkok, Thailand, January 26–27, 2023. pp. 164–170.
19. Şenyürek A, Alp S. Churn prediction in telecommunication sector with machine learning methods. Int J Data Mining Model Manage. 2023; 15 (2): 184–202.
20. Elgohary EM, Galal M, Mosa A, Elshabrawy GA. Smart evaluation for deep learning model: churn prediction as a product case study. Bull Electric Eng Informatics. 2023; 12 (2): 1219–1225.
21. Naidu RP, Oesch PA, van Dokkum P, Nelson EJ, Suess KA, Brammer G, Whitaker KE, Illingworth G, Bouwens R, Tacchella S, Matthee J. Two remarkably luminous galaxy candidates at z ≈ 10–12 revealed by JWST. Astrophys J Lett. 2022; 940 (1): L14.
22. Pandithurai O, Sriman B. Telecom churn prediction using voting classifier ensemble method and supervised machine learning techniques. ITM Web Conf. 2023; 56: 05012.
23. Al-Shourbaji I, Helian N, Sun Y, Alshathri S, Abd Elaziz M. Boosting ant colony optimization with reptile search algorithm for churn prediction. Mathematics. 2022; 10 (7): 1031.
24. Shastry C, Thangavel A. Telco big data analytics using open-source data pipeline: use cases, detailed use case implementation results and findings. Int J Innov Sci Res Technol. 2023; 7 (11): 2128–2138.
25. Almuqren L, Alrayes FS, Cristea AI. An empirical study on customer churn behaviours prediction using Arabic Twitter mining approach. Future Internet. 2021; 13 (7): 175.

