**International Journal of Advanced Reasoning and Computational Intelligence (IJARCI)**
Vol. 1, No. 1, July 2026, pp. xx~xx
ISSN: -, DOI: -


# Customer Churn Prediction in the Telecommunications Industry to Support Retention Strategies Using the XGBoost Algorithm

**Aida Fadhila[1], Halim Maulana[2]**
Faculty of Computer Science and Information Technology, Information Technology, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia
Email: [1]aidafadhila87@gmail.com, [2]thekazagami@gmail.com

**ABSTRACT**

The telecommunications industry faces significant challenges regarding customer churn, as the cost of acquiring new customers is substantially higher than retaining existing ones. This research aims to build an optimal customer churn prediction model using the XGBoost algorithm to support customer retention strategies. The methodology employed is the Cross-Industry Standard Process for Data Mining (CRISP-DM), utilizing a secondary dataset from Kaggle comprising 7,043 customer records. The data imbalance challenge was addressed using the Synthetic Minority Over-sampling Technique (SMOTE) on training data to enhance minority class detection. Evaluation results using Stratified 5-Fold Cross Validation demonstrate highly reliable model performance, achieving an accuracy and a recall value. This model was subsequently implemented into a web-based information system using the FastAPI framework and Neon Database. The system is capable of performing real-time customer risk classification (Low, Medium, High) and integrates Generative AI technology (Gemini API) to provide personalized and communicative retention strategy recommendations for customer service agents.

**Article Info**

***Article history:***

Received May, 30, 2026
Revised June 21, 2026
Accepted July 12, 2026

**Keyword : Customer Churn; FastAPI; Machine learning; SMOTE; Retention Strategy; XGBoost.**

***Corresponding Author:***
Name: Aida Fadhila
Department: Information Technology Study Program
Institution: Universitas Muhammadiyah Sumatera Utara
Address: Medan
Email : aidafadhila87@gmail.com



## 1. INTRODUCTION

Based on the annual reports of several global telecommunications operators, the mobile customer churn rate is in the double-digit range per year, with figures tending to be higher in the prepaid segment than in the postpaid segment [1]-[3]. Recent research and literature reviews indicate that customer churn is a major challenge in the telecommunications industry because it has a direct impact on decreasing recurring revenue and increasing the need to predict and retain customers through retention strategies [3]. Therefore, churn management is seen as a crucial strategic issue for the business sustainability of telecommunications companies because it is directly related to the company's long-term profitability and competitiveness.

In practice, telecommunications companies face challenges in early identification of customers who are likely to churn due to the high volume of data and the complexity of customer behavior [4]-[6]. This complexity is influenced by heterogeneous data characteristics, class imbalance, and non-linear relationships between variables, making them difficult to analyze using conventional linear models [7]-[9]. Recent studies have shown that customer churn patterns are formed from the interaction of various factors such as service usage, contract type, and consumption behavior, which influence each other. Therefore, a machine learning-based approach is needed to more accurately capture these complex relationship patterns [10]-[12].

The data used in this study is a global telecommunications customer dataset obtained from the Kaggle platform, which is also widely used in machine learning-based churn prediction research. Although the dataset is global in scale, the analyzed churn patterns remain relevant to the Indonesian telecommunications industry

ISSN: -

context because the characteristics of cellular services, prepaid and postpaid models, and data-based systems are similar to global industry practices [13]. Furthermore, industry reports indicate that increasing data service penetration and high levels of competition between operators are global phenomena that also occur in various developing countries, including Indonesia, so churn analysis based on global data remains contextually relevant [14].

As technology advances and customer data complexity increases, machine learning approaches are becoming increasingly relevant due to their ability to learn complex and non-linear patterns from historical data more effectively than conventional analysis methods. However, several previous studies have shown that conventional classification algorithms such as Support Vector Machine (SVM), Random Forest, and Logistic Regression still have variable churn prediction performance and are relatively inferior to more sophisticated boosting approaches. In a study of the application of SVM, Random Forest, and Logistic Regression to a telecommunications churn dataset, the highest accuracy results only reached around 79% for Logistic Regression and 76% for Random Forest, while SVM showed even lower accuracy at around 74% [15]-[17]. A comprehensive review also reported that Random Forest and SVM performed moderately with accuracy below 90% on various churn datasets, while simpler models like Logistic Regression tended to be less stable when applied to imbalanced data [18]-[20].

Based on these limitations, this study chose the XGBoost algorithm because of its ability to model complex and non-linear relationships through an iterative gradient boosting approach to minimize prediction errors [21]-[23]. In addition, XGBoost has a regularization mechanism to reduce the risk of overfitting, supports handling imbalanced data through class weight settings, and is able to manage missing values automatically, making it more adaptive to the characteristics of heterogeneous and dynamic churn data [24]-[26]. Based on these advantages, XGBoost was chosen in this study to handle imbalanced data and capture non-linear relationships accurately. This study not only builds a prediction model, but also implements it into a web-based system using FastAPI and Neon Database to provide automatic customer retention strategy recommendations and integrates through a data and model synchronization mechanism.

## 2. METHOD

### 2.1. Customer Churn in the Telecommunications Industry

Customer churn is defined as a condition where a customer decides to stop using a service or switch to another service provider within a specific time period. In this study, churn is represented as a binary target variable (Churn): Yes (customer unsubscribes) and No (customer remains subscribed). This variable is a key indicator of the success of customer retention strategies in the telecommunications industry [23].

In general, churn is divided into two main types: voluntary churn and involuntary churn. Voluntary churn occurs when customers consciously choose to unsubscribe due to dissatisfaction with price or service quality. Conversely, involuntary churn occurs due to administrative factors such as service termination due to unpaid payments. This study focuses on voluntary churn because this type can be predicted and prevented through accurate analysis of historical customer data [27]-[29].

### 2.2. XGBoost (Extreme Gradient Boosting) Algorithm

Gradient boosting is an ensemble learning technique that gradually builds a predictive model by combining several simple models (weak learners), typically decision trees, into a single, more robust model. The learning process is iterative, with each new model trained to correct the residual error generated by the previous model [30]. This approach uses gradient-based optimization principles, where the model gradually minimizes a loss function by adjusting predictions to the direction of the error gradient. In this way, gradient boosting is able to capture non-linear patterns and complex interactions between features that frequently appear in telecommunications customer data [31].

In the context of customer churn prediction, gradient boosting is considered effective because it can improve prediction accuracy on complex and heterogeneous data. Several studies have shown that gradient boosting-based algorithms, including XGBoost, consistently outperform single classification models in identifying customers at risk of churn [32].

### 2.3. Decision Threshold Optimization

When classifying churn risk, using a single threshold, such as 0.5, is often suboptimal for business needs. This research ensures that each threshold used (e.g., 0.7 for high-risk segmentation) has a quantitative basis and is not arbitrary. Threshold determination is based on a Receiver Operating Characteristic (ROC)

ISSN: -

Curve analysis using the Youden Index approach (J = Sensitivity + Specificity - 1) to maximize the model's discrimination ability, or based on a Precision-Recall Curve analysis to optimize the F1-Score. This threshold selection considers the trade-off between precision and recall to align with the business objective of customer retention; where failure to detect potential churners (False Negatives) has a greater financial impact than misdetecting loyal customers (False Positives).

## 3. RESULTS AND DISCUSSION

### 3.1. Dashboard View

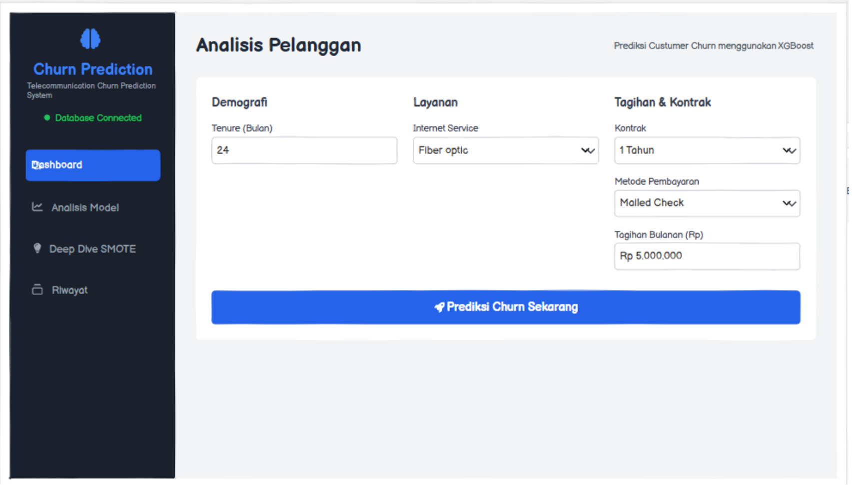


Figure 1. Dashboard View

The system's primary interface serves as a centralized hub for managing and analyzing customer data for churn prediction. It features a structured navigation panel (Dashboard, Model Analysis, Deep Dive SMOTE, and History) and a database connection status indicator. The main section contains a customer analysis form designed around XGBoost model variables to streamline data input and support data-driven retention strategies.

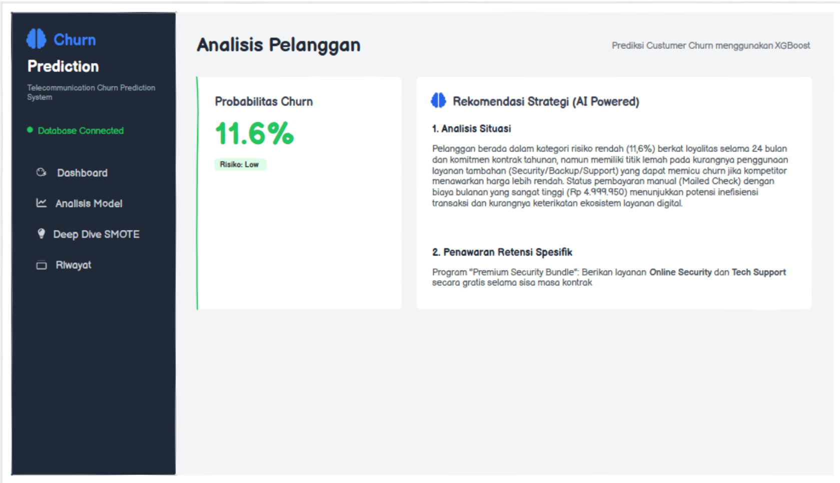


Figure 2 Churn Prediction Results

Utilizing the XGBoost algorithm, the system calculates and displays the churn probability for individual customers. This probability is derived from processed input variables that have undergone preprocessing and modeling based on training data, providing a foundational metric for making informed customer retention decisions.

ISSN: -

Figure 3 Model Analysis Interface

This page evaluates the performance of the XGBoost algorithm after training and testing phases. It displays key evaluation metrics- alongside a confusion matrix to analyze classification errors and a feature importance chart identifying the most influential predictors


Figure 4 Imbalance Detection & Preprocessing

This figure illustrates the initial stage of addressing data imbalance by showing class distribution before applying SMOTE. The preprocessing workflow includes feature-target separation, One-Hot Encoding for categorical variables, and numerical feature standardization to prepare the data for the model and prevent scaling bias.


Figure 5 SMOTE Mechanism

The diagram explains the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic data for the minority class rather than simple duplication. Based on the k-Nearest Neighbors (k-NN) approach $K$=5, the process selects a minority sample and creates new data points between it and its neighbors to enrich data variety and improve the model's ability to learn churn patterns.

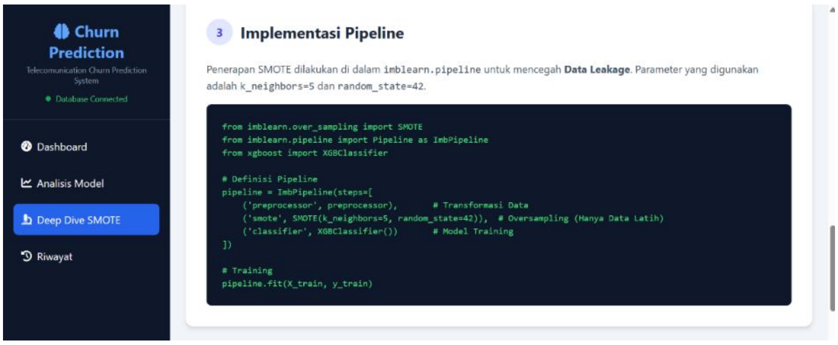


Figure 6 Pipeline Implementation

The system implements SMOTE within an imblearn.pipeline to prevent data leakage. The pipeline integrates three distinct stages: data preprocessing, SMOTE oversampling (applied only to training data), and XGBoost model training, ensuring a structured and consistent workflow.

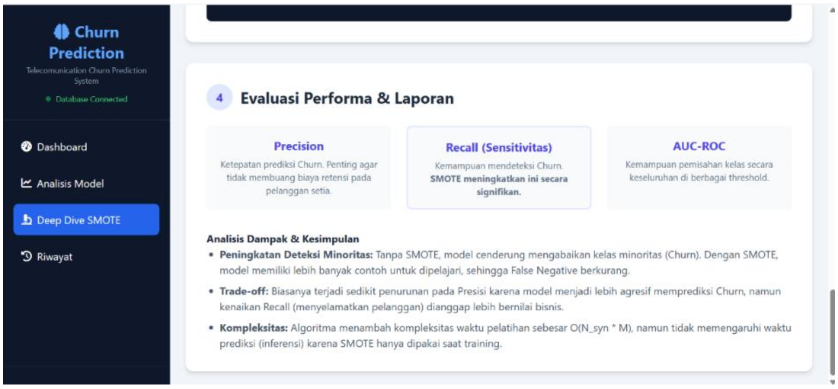


Figure 7 Performance Evaluation & Reporting

Post-SMOTE evaluation demonstrates a significant increase in the Recall metric, which is critical for identifying at-risk customers. Although a slight trade-off in Precision may occur, the enhanced ability to detect churn is considered more valuable for business retention strategies.

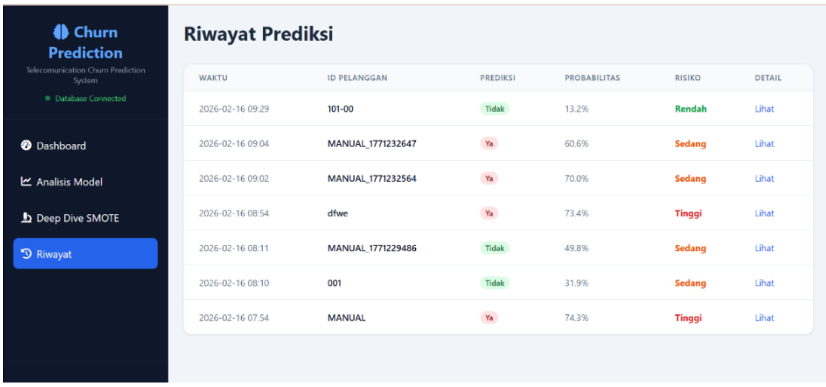


Figure 8 Prediction History Interface

The History page provides comprehensive documentation of all previous predictions, including timestamps, customer IDs, classification results, probability values, and risk levels. This feature ensures transparency and enables long-term tracking and evaluation of data-driven retention efforts.

ISSN: -

## 4. CONCLUSION

Several conclusions were drawn from the discussion on customer churn prediction in the telecommunications industry to support customer retention strategies using the XGBoost algorithm: An optimal customer churn prediction model was successfully built using the XGBoost algorithm, integrating the Synthetic Minority Oversampling Technique (SMOTE) method into the training data to address data imbalance. This approach has proven capable of enabling the model to recognize customer churn patterns without bias toward the majority class. The XGBoost model's performance demonstrated highly reliable results, with an accuracy value and a Recall. These evaluation results demonstrate that the model is not only accurate overall but also effective in minimizing false negatives in detecting customers at risk of churning. A web-based information system was successfully implemented using the FastAPI framework, capable of synchronously presenting prediction results and analyzing churn factors. The use of Neon Database as cloud storage enables automatic synchronization of customer data and model versions, ensuring that the predictions provided are relevant and consistent with the latest data entered into the system. 4. The prediction results and analysis of influencing factors (such as contract type and subscription period) have been successfully utilized as the basis for formulating retention strategies. Integration with Generative AI (Gemini API) in this system can provide specific follow-up strategy recommendations for each customer profile to increase their loyalty.

## REFERENCES


[1] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016(August 2016), 785–794. https://doi.org/10.1145/2939672.2939785

[2] Dhangar, K., & Anand, P. (2021). A REVIEW ON CUSTOMER CHURN PREDICTION USING MACHINE. 8(5), 193–201.

[3] Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. Technologies, 13(3), 1–40. https://doi.org/10.3390/technologies13030088

[4] Ichsan, A., Zulherry, A., Lubis, T.A., & Shahnaz, B.A.Z. (2025). Utilization of Mobile Applications to Speed Up The Search for Android-Based Index Places. IJATCoS: Indonesian Journal of Applied Technology, Computer and Science 2 (1)

[5] Sari, I.P., Batubara, I.H., & Al-Khowarizmi, A. (2021). Sensitivity Of Obtaining Errors In The Combination Of Fuzzy And Neural Networks For Conducting Student Assessment On E-Learning. International Journal of Economic, Technology and Social Sciences (Injects) 2 (1), 331-338

[6] Amada, E.P., & Zulherry, A. (2025). Klasterisasi Minat Dan Bakat Siswa Menggunakan Metode X-Means Berbasis Web: Studi Kasus SMA Negeri 1 Hamparan Perak. sudo Jurnal Teknik Informatika 4 (4), 276-283.

[7] Zulherry, A., Ramadhani, F., & Satria, A. (2024). Klasifikasi Data Tracer Study Dengan Pemanfaatan Data Mining Menggunakan Algoritma Support Vector Machine dan Neural Network. Portal Riset dan Inovasi Sistem Perangkat Lunak 2 (1), 45-54

[8] Zulherry, A. (2023). Decision making for network security with simple additive weighting method. Journal of Intelligent Decision Support System (IDSS), 6(3), 155-159.

[9] Zulherry, A., Gunawan, T. S., & Wanayumini, W. (2021). Analisis Hasil Pendukung Keputusan Mendapatkan Rumah Dinas Perusahaan Menggunakan Metode Analytical Hierarchy Process (AHP) dan Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). Jurnal Media Informatika Budidarma, 5(2), 695-704.

[10] Sari, I.P., Al-Khowarizmi, A., & Batubara, I.H. (2021). Cluster Analysis Using K-Means Algorithm and Fuzzy C-Means Clustering For Grouping Students' Abilities In Online Learning Process. Journal of Computer Science, Information Technology and Telecommunication Engineering 2 (1), 139-144

[11] Apdilah, D., & Sari, I.P. (2021). Optimization Of The Fuzzy C-Means Cluster Center For Credit Data Grouping Using Genetic Algorithms. Al'adzkiya International of Computer Science and Information Technology (AIoCSIT) Journal 2 (2), 156-163

[12] Peng, K., & Peng, Y. (2022). Research on Telecom Customer Churn Prediction Based on GA-XGBoost and SHAP. 107–120. https://doi.org/10.4236/jcc.2022.1011008

[13] Sam, G., Asuquo, P., & Stephen, B. (2024). Customer Churn Prediction using Machine learning Models. 26(2), 181–193. https://doi.org/10.9734/JERR/2024/v26i21081

[14] Schröer, C., Kruse, F., Marx, J., Kruse, F., & Marx, J. (2021). ScienceDirect ScienceDirect A Systematic Literature Review A Systematic Literature Review on Applying Process Model on Applying CRISP-DM Process Model. Procedia Computer Science, 181(2019), 526–534. https://doi.org/10.1016/j.procs.2021.01.199

[15] Shaikhsurab, M. A., & Magadum, P. (2017). Enhancing Customer Churn Prediction in Telecommunications : An Adaptive Ensemble Learning Approach Abstract :

[16] Wakhidah, L. N., Zyen, A. K., & Wahono, B. B. (2025). Evaluation of Telecommunication Customer Churn Classification with SMOTE Using Random Forest and XGBoost Algorithms. Journal of Applied Informatics and Computing, 9(1), 89–95. https://doi.org/10.30871/jaic.v9i1.8740

[17] Sari, I.P., Hariani, P.P., Al-Khowarizmi, A., Ramadhani, F., Sulaiman, O.K., Satria, A, & Manurung, A.A. (2024). CLUSTERING HIV/AIDS DISEASE USING K-MEANS CLUSTERING ALGORITHM. Proceeding International Seminar on Islamic Studies 5 (1), 1668-1676

[18] Basri, M., & Zulherry, A. (2025). Analysis of the Impact of Gambling and Online Loans in the Perspective of Informatics, Islam, and Kemuhammadiyahan. Ar-Rasyid: Jurnal Pendidikan Agama Islam, 5(1), 65-73.

[19] Asadel, A., & Zulherry, A. (2025). Detecting Zero-Width Characters Obfuscated in Phishing URLs using the XGBOOST Algorithm. Hanif Journal of Information Systems 3 (1), 43-53

[20] Zulherry, A., Sari, I.P., & Basri, M. (2025). Perancangan Aplikasi Monitoring Kehadiran Pegawai Menggunakan RFID. sudo Jurnal Teknik Informatika 4 (4), 378-384

[21] Sari, I.P., Ramadhani, F., Satria, A., & Sulaiman, O.K. Leukocoria Identification: A 5-Fold Cross Validation CNN and Adaboost Hybrid Approach. 2023 6th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 486-491

[22] Manurung, A.A., Nasution, M.D., & Sari, I.P. (2023). Implementation of Fuzzy K-Nearest Neighbor Method in Dengue Disease Classification. 2023 11th International Conference on Cyber and IT Service Management (CITSM), 1-4

[23] Asosiasi Penyelenggara Jasa Internet Indonesia. (2024). Laporan Survei Penetrasi Internet Indonesia 2024. https://apjii.or.id/survei

[24] Vodafone: Vodafone Group Plc. (2023). Annual Report 2023. https://investors.vodafone.com/reports-and-presentations/annual-reports/2023

[25] AT&T: AT&T Inc. (2023). 2023 Annual Report. https://investors.att.com/financial-reports/annual-reports/2023

[26] GSMA. (2023). The Mobile Economy Asia Pacific 2023. https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-economy/asiapacific/

[27] Bisht, P. S., & Anjaria, B. (2025). Customer Relationship Management (CRM): Evolution. ResearchGate. https://doi.org/10.13140/RG.2.2.22338.49609

[28] Sari, I.P., Ramadhani, F., Satria, A., & Apdilah, D. (2023). Implementasi Pengolahan Citra Digital dalam Pengenalan Wajah menggunakan Algoritma PCA dan Viola Jones. Hello World Jurnal Ilmu Komputer 2 (3), 146-157

[29] Sari, I.P., Al-Khowarizmi, A, Sulaiman, O.K., & Apdilah, D. (2023). Implementation of Data Classification Using K-Means Algorithm in Clustering Stunting Cases. Journal of Computer Science, Information Technology and Telecommunication Engineering 4 (2), 402-412

[30] Sulaiman, O.K & Batubara, I.H. (2021). Implementation Data Mining For Level Analysis Traffic Violation By Algorithm Association Rule. Al'adzkiya International of Computer Science and Information Technology (AIoCSIT) Journal 2 (2), 128-135

ISSN: -

[31] Hermawan, A., Jayanti, N. R., Tabaruk, Z., Triadi, F. L. Y., Saputra, A., & Syachrudin, M. R. H. (2024). Membangun Model Prediksi Churn Pelanggan yang Akurat: Studi Kasus tentang TELCO Company. Merkurius: Jurnal Riset Sistem Informasi dan Teknik Informatika, 2(6), 67–81. https://doi.org/10.61132/merkurius.v2i6.398

[32] P. H. C. Eilers and J. J. Goeman, "Enhancing scatterplots with smoothed densities," Bioinformatics, vol. 20, no. 5, pp. 623-628, March 2004. [Online]. Available: www.oxfordjournals.org. [Accessed Sept. 18, 2004].