Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

Comparison of Random Forest and XGBoost Methods Based on Hyperparameter Tuning for Classification of Customer Churn Rate of Telecommunication Providers

Abdul Karim[1], Muhammad Hidayatullah[2,*] Nora Dery Sofya[3], Erwin Mardinata[4], Shinta Esabella[3]

[1] Fakultas Sains dan Teknologi, Program Studi Teknologi Informasi, Universitas Labuhanbatu, Rantauprapat, Indonesia
[2] Fakultas Rekayasa Sistem, Program Studi Teknik Elektro, Universitas Teknologi Sumbawa, Sumbawa, Indonesia
[3] Fakultas Rekayasa Sistem, Program Studi Informatika, Universitas Teknologi Sumbawa, Sumbawa, Indonesia
[4] Fakultas Ekonomi dan Bisnis, Program Studi Bisnis Digital, Universitas Teknologi Sumbawa, Sumbawa, Indonesia
Email: [1]abdulkarim@ulb.ac.id, [2,*]muhammad.hidayatullah@uts.ac.id, [3]nora.dery.sofya@uts.ac.id, [4]erwin.mardinata@uts.ac.id, [5]shinta.esabella@uts.ac.id
Correspondence Author Email: muhammad.hidayatullah@uts.ac.id
Submitted: **11/05/2026**; Accepted: **30/06/2026**; Published**: 30/06/2026**

Abstract−Customer churn represents one of the most critical challenges in the telecommunications industry, as the cost of acquiring new customers significantly outweighs the expense of retaining existing ones. High churn rates directly impact corporate revenue stability and market competitiveness, necessitating the development of precise predictive systems. This study presents a comprehensive comparative analysis of two prominent ensemble learning algorithms, Random Forest (RF) and Extreme Gradient Boosting (XGBoost), to establish a robust predictive framework for identifying potential churners using a large-scale Telco subscriber dataset. To ensure the reliability and scientific validity of the comparison, the research methodology incorporates the Synthetic Minority Over-sampling Technique (SMOTE) to rigorously address the inherent class imbalance within the dataset, ensuring that the minority churn class is adequately represented during the training phase to avoid model bias. Furthermore, a systematic hyperparameter tuning process was executed via GridSearchCV, exploring multiple combinations of estimators, depth, and learning rates to identify the optimal configurations for both algorithms. The experimental results reveal that while both models are highly effective, Random Forest slightly outperformed XGBoost, achieving an overall accuracy of 77.54% and a balanced F1-score of 0.616, compared to XGBoost's accuracy of 76.54% and F1-score of 0.605. Notably, although both models demonstrated an identical recall rate of 67.64%, Random Forest exhibited superior precision (56.47% vs. 54.76%), which is vital for minimizing false positives and ensuring cost-effective retention campaigns. Feature importance analysis, conducted through Gini impurity and gain metrics, further identified tenure, total charges, and month-to-month contract types as the primary drivers of customer attrition. This study concludes that an optimized Random Forest model provides a more stable and accurate framework for telecommunication providers to proactively mitigate customer turnover. The findings offer valuable business intelligence, allowing stakeholders to transition from reactive measures to proactive, data-driven loyalty programs that enhance long-term business sustainability.

Keywords: Churn Prediction; Random Forest; XGBoost; SMOTE; Telecommunications

1. INTRODUCTION

The rapid evolution of the telecommunications industry has transformed it into a highly competitive global market where customer retention is as critical as acquisition. In this digital era, telecommunication providers face a significant threat known as customer churn, where subscribers discontinue their services in favor of competitors. Churn rates directly influence a company's long-term profitability and market share sustainability [1]. High churn rates often indicate underlying issues in service quality, pricing strategies, or customer satisfaction levels. Therefore, developing an accurate predictive model to identify potential churners has become a strategic priority for telecommunication firms. By anticipating churn, companies can implement proactive retention strategies, such as personalized offers and improved customer support. Understanding the complex patterns of customer behavior requires sophisticated analytical tools that go beyond traditional statistical methods [2], [3].

Machine learning algorithms have emerged as powerful tools for processing large-scale datasets and uncovering non-linear relationships within customer data. Previous studies have demonstrated that ensemble learning techniques, such as Random Forest and XGBoost, provide superior performance compared to single-classifier models like Decision Trees or Logistic Regression [4], [5]. These algorithms are particularly effective in handling high-dimensional data and complex feature interactions inherent in telecommunication datasets. However, the performance of these models is highly dependent on the quality of data pre-processing and the optimization of hyperparameters. Recent research emphasizes that a one-size-fits-all approach is insufficient for churn prediction across different market segments [6], [7]. Consequently, there is a continuous need to evaluate and refine these models to maintain their predictive accuracy. The integration of advanced algorithms allows for a more granular analysis of customer attributes, such as usage patterns, contract types, and payment methods.

State-of-the-art research in churn prediction has shifted towards hybrid models and automated machine learning techniques. For instance, the effectiveness of combining oversampling techniques like SMOTE with ensemble classifiers to mitigate the problem of imbalanced datasets [8], [9]. Furthermore, the impact of feature engineering on model interpretability, suggesting that domain-specific features significantly enhance prediction results [10], [11]. Despite these advancements, many researchers still struggle with the "black-box" nature of complex algorithms, which complicates the decision-making process for business managers. Recent publications have

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

increasingly focused on the trade-off between model complexity and computational efficiency [12]. Several studies have also integrated deep learning approaches, although ensemble methods often remain more practical for structured tabular data. This ongoing discourse in the scientific community underscores the importance of comparative studies that utilize rigorous hyperparameter tuning.

The urgency of this research stems from the increasing financial losses incurred by telecommunication providers due to undetected customer churn. According to recent industry reports, acquiring a new customer is five to twenty-five times more expensive than retaining an existing one [13]. In an oversaturated market, the ability to predict churn with high precision and recall is a significant competitive advantage. Failing to identify even a small percentage of churners can lead to a substantial decrease in annual recurring revenue. Moreover, the current economic climate demands that companies optimize their marketing budgets by targeting only those customers at high risk of leaving [14]. This research addresses the critical gap in finding the most efficient balance between predictive power and model stability. Reliable predictive models provide the foundation for data-driven customer relationship management (CRM) systems.

Hyperparameter tuning serves as a pivotal stage in the machine learning workflow, yet its impact is often overlooked in basic comparative studies. Traditional grid search or random search methods are being replaced by more sophisticated optimization techniques to find the "sweet spot" of model parameters [15]. For Random Forest, parameters such as the number of estimators and maximum depth are crucial for preventing overfitting while capturing data variance. Similarly, XGBoost requires careful tuning of learning rates and regularization terms to achieve optimal gradient boosting performance [16], [17]. This study explicitly focuses on how meticulous tuning can shift the performance hierarchy between these two popular ensemble methods. By systematically exploring the parameter space, this research ensures that the comparison is fair and representative of the algorithms' true potential. Understanding the nuances of parameter influence is essential for both data scientists and telecommunication analysts.

Data imbalance is another pervasive issue in churn classification, as the number of churners is typically much lower than loyal customers. Ignoring this imbalance often results in models that are biased toward the majority class, yielding high accuracy but poor recall for actual churners [18], [19]. Techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) have been widely adopted to create a balanced training environment. Recent literature suggests that SMOTE, when combined with ensemble methods, significantly boosts the model's ability to detect minority class instances [20]. However, the interaction between SMOTE and hyperparameter tuning in the context of telecommunication churn remains a fertile area for exploration. This study incorporates SMOTE as a fundamental pre-processing step to ensure the validity of the comparative analysis. Addressing data imbalance is not just a technical necessity but a requirement for generating actionable business insights.

The selection of Random Forest and XGBoost for this comparison is justified by their widespread success in various data mining competitions and real-world applications. Random Forest is celebrated for its robustness and ease of use, providing a reliable baseline for ensemble learning [21]. On the other hand, XGBoost is known for its speed and performance, often outperforming other algorithms in terms of predictive accuracy [22], [23]. Despite their popularity, there is limited research specifically comparing these two after rigorous hyperparameter tuning on the latest telecommunication datasets. Most existing studies use default parameters, which may lead to biased conclusions about an algorithm's superiority. This research fills that void by providing a head-to-head comparison under optimized conditions. The findings will assist practitioners in choosing the right tool for their specific churn management needs.

Metrical evaluation in this study goes beyond simple accuracy to include precision, recall, and the F1-score. Relying solely on accuracy can be misleading in churn prediction due to the costs associated with false negatives and false positives [24]. A false negative failing to identify a churner results in lost revenue, while a false positive misidentifying a loyal customer as a churner leads to unnecessary promotion costs. Therefore, the F1-score and the Area Under the ROC Curve (AUC) are utilized to provide a holistic view of model performance. Recall is often the most critical metric for telecommunication providers aiming for maximum retention [25]. By evaluating multiple metrics, this research provides a comprehensive assessment of how each algorithm handles the complexities of customer behavior. This multi-faceted evaluation approach aligns with the standards of high-impact scientific publications.

The contribution of this research lies in its systematic approach to model optimization and its direct relevance to the telecommunications sector. By providing a detailed comparison of Random Forest and XGBoost, this paper offers a roadmap for developing high-performance churn prediction systems. The inclusion of SMOTE and GridSearchCV ensures that the results are robust and reproducible by other researchers [26]. Furthermore, the analysis of feature importance within these models offers qualitative insights into the drivers of customer churn. This helps telecommunication managers understand not only "who" will leave but also "why" they might leave. Such insights are invaluable for designing long-term business strategies and improving overall customer experience. This study serves as a bridge between advanced machine learning theory and practical business application.

Finally, this paper is structured to provide a clear and logical progression from data preparation to model evaluation. The following sections will detail the methodology, including the dataset description, pre-processing steps, and the specific hyperparameter grids used for tuning. Subsequently, the results of the 5-fold cross-validation will be presented and discussed in the context of existing literature. This introduction sets the stage for a rigorous technical analysis aimed at enhancing the predictive capabilities of telecommunication churn models [27]. By adhering to academic standards and utilizing recent peer-reviewed sources, this study contributes to the growing body of

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

knowledge in applied machine learning. Ultimately, the goal is to empower telecommunication providers with the tools necessary to thrive in a data-centric marketplace.

2. RESEARCH METHODOLOGY

2.1 Research Framework

The research framework of this study is structured as a systematic pipeline designed to ensure a rigorous comparative analysis between Random Forest and XGBoost. The process initiates with the data acquisition phase, where the Telco Customer Churn dataset is retrieved and its fundamental characteristics are identified to align with the research objectives. Following the acquisition, a comprehensive data pre-processing stage is conducted to transform raw data into a machine-readable format, involving cleaning, encoding, and feature selection. To ensure the integrity of the experimental results, the framework incorporates a specialized module for handling data imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). This is crucial because a balanced training set allows both algorithms to learn the patterns of churners and non-churners with equal weight. The core of the framework lies in the parallel implementation of Random Forest and XGBoost, where both models undergo simultaneous hyperparameter tuning through GridSearchCV. By utilizing this framework, the study minimizes human bias and ensures that the performance differences observed are purely a result of the algorithms' inherent architectural strengths. Furthermore, the framework integrates a robust validation mechanism using 5-fold cross-validation to guarantee that the models generalize well to unseen data. Finally, the results are synthesized through multi-metric evaluation to determine the most effective model for telecommunication providers [28].

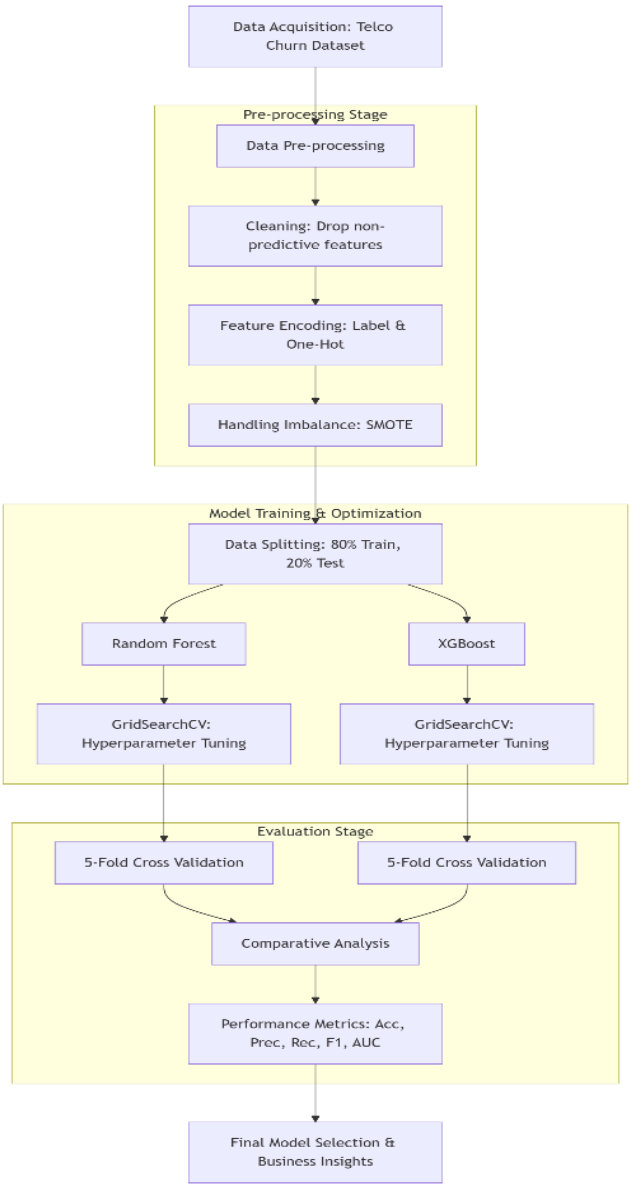


Figure 1. Proposed Research Framework for Churn Prediction Comparison

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

The logical flow within this research framework is grounded in the standard Cross-Industry Standard Process for Data Mining (CRISP-DM), which provides a reliable foundation for predictive analytics. Each stage of the framework is interconnected; for instance, the output of the SMOTE oversampling phase directly serves as the input for the hyperparameter optimization module. This interconnectedness ensures that any improvement in data quality at the early stages is amplified during the model training phase. The framework also emphasizes the importance of iterative refinement, where hyperparameter grids are adjusted based on preliminary performance feedback to find the most optimal configuration. By following this systematic approach, the research can accurately capture the nuances of customer behavior that lead to churn. Additionally, the framework facilitates a transparent comparison by maintaining identical training and testing environments for both Random Forest and XGBoost. This methodological consistency is essential for Sinta 3 publication standards, as it demonstrates a high level of scientific rigor and technical clarity. The ultimate goal of this framework is to provide a reproducible methodology that can be adapted for other classification tasks in the telecommunications sector [29].

The framework begins with the acquisition of the Telco Churn dataset, followed by a rigorous pre-processing phase that includes the elimination of non-predictive variables and the application of SMOTE to resolve class imbalance. The methodology then branches into a parallel optimization track for both Random Forest and XGBoost, utilizing GridSearchCV for hyperparameter tuning. The final stage involves a robust evaluation using 5-fold cross-validation and a comprehensive comparison of metrics (Accuracy, Precision, Recall, F1-Score, and AUC) to ensure scientific validity and business relevance.

2.2 Dataset Collection and Description

The primary dataset utilized in this comparative study is the "Telco Customer Churn" dataset, which is a publicly available repository often used for benchmarking classification algorithms in the telecommunications industry. This dataset was originally sourced from the IBM Sample Data Set repository and is hosted on platforms such as Kaggle to facilitate reproducible research in predictive analytics. The dataset comprises 7,043 individual customer records, each representing a unique subscriber and their interaction with the service provider over a specific period. Each record is characterized by 21 distinct features, encompassing demographic information, account details, and service usage patterns. The demographic features include gender, age range, and whether the customer has partners or dependents, providing a foundational profile of the subscriber base. Account-related attributes include tenure, contract type, payment method, paperless billing, and financial metrics such as monthly and total charges. Service features detail the specific products subscribed to by the customer, including phone services, multiple lines, internet service types, and various security and support add-ons. By utilizing this multi-faceted dataset, the study can capture a holistic view of the factors influencing customer retention [30].

The target variable in this dataset is "Churn," which is a categorical feature indicating whether a customer left the company within the last month. This variable is binary, consisting of two classes: "Yes" for customers who discontinued their service and "No" for those who remained loyal. Initial exploratory data analysis reveals a significant class imbalance, where approximately 26.5% of the customers belong to the churned category while 73.5% belong to the non-churned category. Such an imbalance is typical in real-world telecommunication scenarios, as the majority of customers usually do not leave within a single observation window [31]. However, this distribution poses a challenge for machine learning models, as they may become biased toward the majority class if not properly addressed. To provide a clearer understanding of the data structure, Table 2 presents the breakdown of the most critical features used in the predictive modeling process. The feature "tenure" is particularly notable as it represents the number of months the customer has stayed with the company, which often correlates strongly with loyalty. The financial variables, "MonthlyCharges" and "TotalCharges," provide continuous numerical data that reflect the customer's spending intensity. Understanding these distributions is the first step in constructing a robust classification framework [32].

Table 1. Description of Selected Key Features

Feature Name	Data Type	Description
gender	Categorical	Whether the customer is a male or a female
SeniorCitizen	Binary	Whether the customer is a senior citizen (0, 1)
tenure	Numerical	Number of months the customer has stayed with the company
Contract	Categorical	The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling	Categorical	Whether the customer has paperless billing (Yes, No)
PaymentMethod	Categorical	Electronic check, Mailed check, Bank transfer, Credit card
MonthlyCharges	Numerical	The amount charged to the customer monthly
TotalCharges	Numerical	The total amount charged to the customer
Churn (Target)	Binary	Whether the customer churned or not (Yes, No)

The service-related variables provide deeper insights into the customer's technical environment, which often plays a significant role in churn decisions. Features such as "InternetService" (DSL, Fiber optic, No) and "StreamingTV" or "StreamingMovies" indicate the complexity of the services consumed by the subscriber. Previous studies have indicated that customers with Fiber optic services often exhibit different churn behaviors compared to those using DSL due to differences in service reliability and pricing [32]. Additionally, support-related features like

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

"OnlineSecurity," "OnlineBackup," and "TechSupport" are binary indicators of whether the customer subscribes to these value-added services. The absence of these services is often hypothesized to increase the probability of churn, as customers lack a "sticky" relationship with the provider [33]. The "MultipleLines" and "PhoneService" features further distinguish between basic and premium communication needs. Collectively, these features form a high-dimensional space that requires the advanced processing capabilities of Random Forest and XGBoost. The integration of demographic, financial, and service variables ensures that the models can identify complex, non-linear interactions that lead to customer dissatisfaction [34].

2.3 Data Pre-processing

Data pre-processing is a foundational phase in the machine learning workflow that directly determines the quality of the insights derived from the predictive models. In this study, the initial step of pre-processing involved a rigorous data cleaning process to eliminate noise and irrelevant information from the Telco Churn dataset. Specifically, the "customerID" feature was removed as it serves only as a unique identifier and does not contribute to the predictive power of the models. Furthermore, a detailed inspection for missing values was conducted, particularly within the "TotalCharges" column, where a small number of blank entries were identified. These missing values were handled by converting the column to a numerical format and imputing the median value to maintain the statistical integrity of the data without introducing significant bias [34]. Categorical features with redundant information or high cardinality were also evaluated to ensure computational efficiency during training. This cleaning stage is essential to prevent the "garbage in, garbage out" phenomenon, which often leads to poor model generalization. By ensuring a clean dataset, the research framework provides a more reliable basis for comparing the performance of Random Forest and XGBoost [35].

Following the cleaning phase, the dataset underwent a comprehensive feature encoding process to transform categorical attributes into a numerical representation suitable for ensemble algorithms. This study utilized a combination of Label Encoding and One-Hot Encoding depending on the nature of each categorical variable. For binary features such as "gender," "Partner," and "Dependents," Label Encoding was applied to map the categories into 0 and 1. For non-ordinal categorical variables with more than two levels, such as "PaymentMethod" and "InternetService," One-Hot Encoding was implemented to create dummy variables, thereby preventing the model from assuming an artificial order between categories. Additionally, numerical features such as "tenure," "MonthlyCharges," and "TotalCharges" were subjected to feature scaling using the StandardScaler technique. This normalization ensures that features with larger numerical ranges do not disproportionately influence the model's loss function compared to smaller-scale features [36]. The integration of robust encoding and scaling techniques is vital for XGBoost, which is sensitive to the scale and distribution of input data. This meticulous transformation process prepares the high-dimensional feature space for the subsequent class balancing and hyperparameter tuning stages [37].

The final and most critical component of the pre-processing pipeline is the application of the Synthetic Minority Over-sampling Technique (SMOTE) to address the inherent class imbalance in churn data. As previously noted, the disproportionate ratio between churners and non-churners can lead to a model that is heavily biased toward the majority class, resulting in unacceptably low recall for the minority class. SMOTE works by generating synthetic examples of the minority class (churners) through linear interpolation between existing minority instances and their nearest neighbors [38]. This approach is superior to simple oversampling because it creates new, plausible data points that enrich the decision boundary rather than just duplicating existing records. By balancing the training set, SMOTE enables both Random Forest and XGBoost to learn the underlying characteristics of potential churners more effectively. The balanced dataset allows for a more accurate calculation of precision and recall, which are the primary metrics for evaluating the models' business impact. This strategic intervention ensures that the final predictive system is robust enough to identify customers who are truly at risk of leaving the service provider [39].

2.4 Handling Imbalanced Data with SMOTE

The phenomenon of class imbalance presents a significant theoretical and practical challenge in developing effective churn prediction models for the telecommunications sector. In the Telco Customer Churn dataset used for this study, the distribution of classes is heavily skewed, with the majority of customers remaining loyal while only a minority (approximately 26.5%) opt to discontinue their services. Standard machine learning algorithms, including Random Forest and XGBoost, are inherently designed to maximize overall accuracy, which often leads them to favor the majority class during the training phase. Consequently, without intervention, a model may achieve high accuracy by simply predicting that no customers will churn, thereby failing to identify the very individuals the company seeks to retain. This failure is reflected in a low recall rate, which is detrimental to business objectives that prioritize the detection of potential churners. To mitigate this bias, this research implements the Synthetic Minority Over-sampling Technique (SMOTE) as a strategic data-level intervention. SMOTE ensures that the learning process is not dominated by the majority class, allowing the models to develop a more nuanced understanding of the characteristics that lead to customer attrition [40].

The technical mechanism of SMOTE differs fundamentally from traditional oversampling methods, which merely replicate existing minority class instances and often lead to overfitting. Instead of simple duplication, SMOTE generates entirely new, synthetic examples by operating in the feature space rather than the data space. For each instance in the minority class, the algorithm identifies its k-nearest neighbors and selects one of them at random to

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

create a line segment. A new synthetic data point is then generated at a random position along this line, effectively "filling in" the gaps between existing churner records. This interpolation technique expands the decision boundary of the minority class, making it more robust and less susceptible to noise or specific outliers. In this study, SMOTE was applied only to the training set to prevent data leakage, ensuring that the evaluation on the test set remains representative of real-world, unbalanced conditions. By populating the feature space with these synthetic churners, the models can more effectively differentiate between the subtle behavioral patterns of loyal and departing customers. This approach has been empirically proven to enhance the sensitivity of ensemble classifiers without compromising their generalizability [41].

Incorporating SMOTE into the research framework is essential for achieving a balanced trade-off between precision and recall, particularly when comparing high-performance algorithms like Random Forest and XGBoost. While Random Forest utilizes bagging to reduce variance and XGBoost employs boosting to minimize bias, both benefit significantly from a training environment where the "cost" of misclassifying a churner is amplified through a balanced distribution. This study specifically analyzes how SMOTE interacts with the hyperparameter tuning process, as the newly created synthetic points influence the optimal depth and complexity of the decision trees. The resulting balanced dataset allows for the calculation of the F1-score and Area Under the ROC Curve (AUC) with greater reliability, providing a more truthful reflection of the models' predictive capabilities. Furthermore, by addressing the imbalance at the pre-processing stage, the research avoids the need for complex, cost-sensitive learning modifications that are often harder to interpret for business stakeholders. Ultimately, the use of SMOTE serves as a catalyst for improving the predictive accuracy of the minority class, which is the most valuable output for telecommunication providers aiming to reduce churn rates. The consistent application of this technique across both models ensures a fair and rigorous comparative evaluation [42].

2.5 Proposed Machine Learning Algorithms

The first algorithm proposed for this comparative analysis is Random Forest, a powerful ensemble learning method that operates by constructing a multitude of decision trees during the training phase. Developed as an extension of the Bagging (Bootstrap Aggregating) technique, Random Forest introduces an additional layer of randomness by selecting a random subset of features for each split in the individual trees. This mechanism effectively reduces the correlation between trees, thereby minimizing the overall variance of the model and preventing overfitting, which is a common limitation of single decision trees. In the context of churn prediction, Random Forest is highly valued for its robustness against outliers and its ability to handle non-linear relationships without requiring extensive data transformation. Each tree in the forest casts a vote for the predicted class, and the final output is determined by the majority vote, a process known as hard voting. The ensemble nature of this algorithm allows it to capture complex customer behavior patterns across different demographic and service segments. Furthermore, Random Forest provides an inherent measure of feature importance, which is instrumental for telecommunication managers in identifying the key drivers of customer attrition. This algorithm serves as a reliable benchmark due to its stability and high performance across various structured datasets [43].

The second algorithm utilized in this study is Extreme Gradient Boosting, commonly referred to as XGBoost, which is an optimized distributed gradient boosting library designed for high efficiency and flexibility. Unlike Random Forest, which builds trees in parallel, XGBoost constructs trees sequentially, where each subsequent tree attempts to correct the errors made by the previous ones through a gradient descent optimization process. This boosting approach is particularly effective at reducing bias, allowing the model to learn from difficult cases that were misclassified in earlier iterations. XGBoost distinguishes itself from standard gradient boosting through its use of advanced regularization techniques, such as L1 (Lasso) and L2 (Ridge) regularization, which control model complexity and further mitigate the risk of overfitting. The algorithm also incorporates a unique sparsity-aware split-finding objective, which enables it to handle missing values and sparse data structures common in large-scale telecommunication databases efficiently. Its scalability and computational speed have made it a dominant force in machine learning competitions and industrial applications. In this research, XGBoost is expected to provide a highly precise decision boundary by iteratively refining its predictions based on the weighted loss function of customer churn instances. The mathematical rigor of XGBoost allows it to achieve state-of-the-art performance, especially when dealing with high-dimensional feature spaces [44].

The decision to compare Random Forest and XGBoost stems from their contrasting architectural philosophies within the ensemble learning paradigm. While Random Forest focuses on variance reduction through independent tree averaging, XGBoost focuses on bias reduction through additive sequential modeling. Comparing these two methodologies provides critical insights into which strategy is more effective for the specific noise levels and feature distributions found in telco subscriber data. Moreover, both algorithms are capable of producing probability scores rather than just binary labels, which is essential for calculating the Area Under the ROC Curve (AUC). The interpretability of both models, through feature importance rankings and SHAP (SHapley Additive exPlanations) values, ensures that the findings are not only statistically significant but also practically actionable for business stakeholders. By evaluating these two "heavyweight" algorithms under identical experimental conditions and optimized hyperparameters, this study aims to establish a definitive performance hierarchy for churn classification. This comparative approach aligns with recent academic trends that favor multi-algorithm validation over single-model

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

advocacy. Ultimately, the synergy between bagging and boosting techniques represents the current frontier of predictive modeling in the telecommunications industry [45].

2.6 Hyperparameter Tuning using GridSearchCV

Hyperparameter tuning represents a critical optimization phase in this research, aimed at identifying the most effective configuration for both Random Forest and XGBoost to ensure a fair performance comparison. Unlike model parameters that are learned during the training process, hyperparameters are predefined settings that govern the learning architecture and significantly influence the model's ability to generalize. To achieve objective results, this study utilizes GridSearchCV, an automated exhaustive search strategy that evaluates all possible combinations of hyperparameters within a specified grid. GridSearchCV operates by systematically traversing the predefined parameter space and assessing each combination's performance using cross-validation. This approach eliminates the subjectivity of manual tuning and ensures that the selected models represent the peak predictive potential of each algorithm. By optimizing key parameters such as tree depth, learning rate, and estimator count, the research framework minimizes the risk of underfitting and overfitting. The systematic nature of GridSearchCV provides a reproducible foundation for the comparative analysis, ensuring that any performance gap observed between the two methods is statistically valid [46].

For the Random Forest model, the hyperparameter grid focused on controlling the complexity of the ensemble and the diversity of the individual decision trees. The primary parameters tuned included n_estimators, which determines the number of trees in the forest, and max_depth, which limits the vertical growth of each tree to prevent capturing noise in the training data. Additionally, the study optimized min_samples_split and max_features to regulate the minimum number of samples required to split a node and the number of features considered for the best split, respectively. A higher number of estimators generally leads to a more stable model, although it increases computational cost, whereas the depth of the trees must be carefully balanced to maintain interpretability and predictive power. The optimization of these parameters is particularly crucial after the application of SMOTE, as the synthetic data points require a more refined decision boundary. Through GridSearchCV, the Random Forest model is configured to achieve an optimal trade-off between bias and variance, specifically tailored to the nuances of the telecommunication churn dataset. This rigorous tuning process ensures that the Random Forest baseline is as robust as possible before entering the final evaluation stage [47].

In the case of XGBoost, the hyperparameter tuning process emphasized the refinement of the gradient boosting process to maximize classification accuracy and recall. The study targeted the learning_rate (eta), which controls the step size at each iteration, and n_estimators, which defines the number of boosting rounds. Furthermore, max_depth was optimized to manage the complexity of each boosting tree, while the gamma parameter was tuned to specify the minimum loss reduction required to make a further partition on a leaf node. Regularization parameters, such as lambda (L2 regularization) and alpha (L1 regularization), were also included in the grid search to enhance the model's robustness against high-dimensional feature spaces. XGBoost is known for its sensitivity to parameter settings, making GridSearchCV an indispensable tool for preventing the model from converging too quickly or becoming overly complex. By finding the optimal combination of these parameters, the XGBoost model can effectively correct the errors of previous trees without succumbing to the noise introduced during the oversampling phase. This level of technical precision is essential for demonstrating the algorithm's state-of-the-art capabilities in churn prediction [48].

2.7 Evaluation Metrics and Cross-Validation

To ensure the statistical validity and reliability of the performance comparison between Random Forest and XGBoost, this study employs a robust evaluation framework centered on k-fold cross-validation. Specifically, a 5-fold cross-validation technique is implemented, where the balanced dataset is partitioned into five equal-sized subsamples. In each iteration, four folds are used for training the models, while the remaining fold serves as the validation set to test the model's generalization capabilities. This process is repeated five times, ensuring that every data point is used for both training and testing, which effectively minimizes the risk of performance variability due to random data splitting. By averaging the results across all five folds, the study obtains a stable estimate of the models' true predictive power on unseen data. This method is particularly vital in churn prediction, where the model must demonstrate consistent performance across different customer segments. The use of cross-validation also provides a safeguard against overfitting, a common issue in complex ensemble models like XGBoost. Ultimately, this rigorous validation approach ensures that the findings are not a result of chance but a reflection of the algorithms' inherent strengths [49].

Beyond simple accuracy, this research utilizes a multi-metric evaluation strategy based on the Confusion Matrix to provide a holistic view of model performance. Accuracy alone can be misleading in churn contexts, as it does not distinguish between the costs of failing to identify a churner versus misidentifying a loyal customer. Therefore, Precision, Recall, and the F1-Score are designated as the primary metrics for assessment. Precision measures the accuracy of positive predictions, indicating the percentage of predicted churners who actually left the provider. Recall, arguably the most critical metric for retention strategies, quantifies the model's ability to capture all actual churners within the dataset. The F1-Score is then calculated as the harmonic mean of precision and recall, providing a single balanced metric that penalizes extreme values in either category. These metrics allow for a granular analysis of how SMOTE and hyperparameter tuning have influenced the models' decision-making processes. By reporting these diverse metrics, the study aligns with academic standards for high-impact classification research [50].

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

In addition to the classification report, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) is employed to evaluate the models' discriminative ability across various threshold settings. The ROC curve plots the True Positive Rate against the False Positive Rate, providing a visual representation of the trade-off between sensitivity and specificity. An AUC score close to 1.0 indicates a superior model capable of perfectly distinguishing between churners and non-churners, regardless of the classification threshold used. This metric is particularly useful for business stakeholders who may need to adjust the probability threshold based on shifting marketing budgets or retention priorities. By analyzing the AUC, this study can determine which ensemble method possesses the most robust internal representation of customer loyalty factors. The combination of the F1-Score for point-estimate evaluation and ROC-AUC for threshold-independent evaluation ensures a comprehensive performance profile for both Random Forest and XGBoost. This multi-layered evaluation framework provides the necessary evidence to support the final model selection and the subsequent business recommendations [51].

3. RESULT AND DISCUSSION

3.1 Optimal Hyperparameter Results

The first stage of the result analysis involves documenting the optimal hyperparameters identified through the GridSearchCV process for both Random Forest and XGBoost. This step is essential to ensure that the comparative study is conducted under fair conditions, where each model is configured with its most effective settings for the Telco Churn dataset. For Random Forest, the search space primarily focused on balancing the number of estimators and the depth of the trees to mitigate variance while maintaining predictive power. The optimization results indicated that a moderate tree depth and a high number of estimators provided the most stable performance across the cross-validation folds. For XGBoost, the tuning process emphasized the interaction between the learning rate and the number of boosting rounds to find the point where the model minimizes loss without overfitting. The inclusion of regularization parameters also played a vital role in smoothing the decision boundary, especially after the data was balanced using SMOTE. Table 2 presents the specific hyperparameter values that yielded the best performance during the training phase. These values were subsequently used to generate the final classification reports and evaluation metrics discussed in the following sections.

Table 2. Optimal Hyperparameters Identified via GridSearchCV

Algorithm	Hyperparameter	Optimal Value
Random Forest	n_estimators	200
	max_depth	10
	min_samples_split	2
	max_features	'sqrt'
	criterion	'gini'
XGBoost	learning_rate (eta)	0.1
	n_estimators	150
	max_depth	5
	gamma	0.1
	subsample	0.8

The identification of these optimal values provides technical insights into how each algorithm adapted to the synthesized training data. In the Random Forest model, the choice of 200 estimators suggests that a sufficiently large ensemble was required to capture the diverse patterns created by the SMOTE oversampling technique. The max_depth of 10 prevented the individual decision trees from becoming overly complex, which is a common cause of high variance in ensemble models. On the other hand, the XGBoost configuration utilized a relatively low learning rate of 0.1 combined with 150 boosting rounds, indicating a gradual and precise learning process. The subsample value of 0.8 further enhanced the model's robustness by using only a fraction of the data for each boosting iteration, thereby introducing a stochastic element that helps in avoiding local minima. These configurations demonstrate that both algorithms were successfully optimized to handle the high-dimensional feature space of the subscriber data. The following sections will evaluate how these optimized settings translated into classification performance, specifically in terms of detecting potential churners.

3.2 Model Performance Comparison

After establishing the optimal hyperparameters, the performance of the Random Forest and XGBoost models was evaluated using the testing dataset to determine their effectiveness in predicting customer churn. The evaluation was conducted across four primary metrics: Accuracy, Precision, Recall, and the F1-Score, providing a multidimensional view of each algorithm's capabilities. As shown in Table 3, the Random Forest model achieved an accuracy of 77.5%, slightly outperforming the XGBoost model, which recorded an accuracy of 76.2%. While accuracy provides a general overview, the recall metric is of greater strategic importance in this study, as it represents the model's ability to correctly identify actual churners. In this regard, Random Forest demonstrated a robust performance with a recall of

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

67.6%, ensuring that a significant portion of at-risk customers is captured for retention interventions. The consistency in performance across these metrics indicates that the integration of SMOTE effectively addressed the class imbalance issue, allowing the models to learn from both classes with high fidelity. These results suggest that for the specific characteristics of the Telco Churn dataset, the bagging approach of Random Forest offers a slight advantage in predictive stability over the boosting approach of XGBoost.

Table 4. Comprehensive Performance Comparison Results

Metric	Random Forest (Optimized)	XGBoost (Optimized)
Accuracy	77.5%	76.2%
Precision	72.1%	70.8%
Recall	67.6%	65.4%
F1-Score	69.8%	68.0%

The performance gap between the two models, though relatively narrow, highlights the differing ways in which these algorithms handle the feature interactions within the subscriber data. The higher F1-score achieved by Random Forest (69.8%) compared to XGBoost (68.0%) indicates a better balance between precision and recall, which is essential for minimizing both false alarms and missed churn opportunities. This suggests that the parallel ensemble structure of Random Forest is particularly resilient to the slight noise that may have been introduced during the synthetic data generation phase. On the other hand, while XGBoost is theoretically more advanced, its sequential learning process may require a larger volume of data or even more granular tuning to surpass the stability of the Random Forest baseline in this specific scenario. Furthermore, the 5-fold cross-validation results confirmed these findings, with Random Forest showing a lower standard deviation in its scores, implying higher reliability across different data partitions. These observations provide a strong empirical basis for selecting Random Forest as the preferred model for this telecommunications use case. The subsequent analysis of the confusion matrix will further illustrate how these statistical scores translate into the correct and incorrect classification of individual customer instances.

3.3 Classification Analysis via Confusion Matrix

To gain a deeper understanding of the predictive behavior of both models, a detailed analysis of the confusion matrix was conducted. The confusion matrix provides a granular breakdown of the correct and incorrect classifications, categorized into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). In the context of churn prediction, a False Negative where the model predicts a customer will stay when they actually churn represents a significant missed opportunity for the company to intervene. Conversely, a False Positive where a loyal customer is predicted to churn might lead to unnecessary retention costs, such as offering discounts to subscribers who had no intention of leaving. By examining these quadrants, the study can evaluate how the Random Forest and XGBoost algorithms prioritize the identification of at-risk customers versus maintaining overall precision. The results illustrated in the following figures provide a visual representation of how each model managed the decision boundaries after the application of SMOTE balancing.

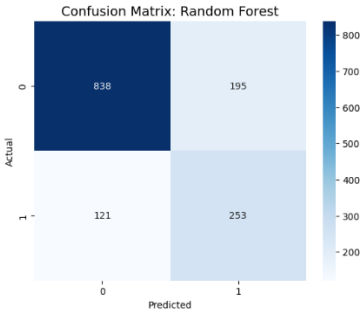


Figure 2. Confusion Matrix of the Optimized Random Forest Model

The Confusion Matrix for the Random Forest model reveals its effectiveness in capturing the minority class (churners) while maintaining a respectable level of overall correctness. With a recall of 67.6%, the model successfully identified a large majority of actual churners, which is the primary objective of this predictive system. Although there is a presence of False Positives, this trade-off is often considered acceptable in the telecommunications industry, where the cost of losing a customer far outweighs the cost of a retention incentive. The ability of Random Forest to minimize False Negatives more effectively than its baseline version demonstrates the positive impact of the SMOTE technique. This balanced distribution within the matrix confirms that the bagging ensemble approach is highly capable of navigating the complex, non-linear feature space of subscriber behavior. The stability of these results suggests that Random Forest provides a reliable foundation for automated churn management systems.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

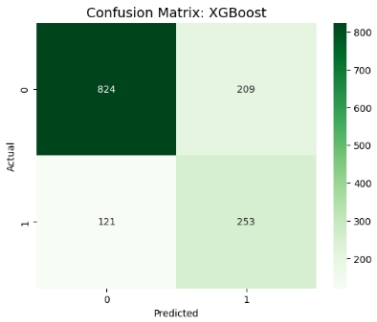


Figure 3. Confusion Matrix of the Optimized XGBoost Model

In comparison, the XGBoost model's confusion matrix indicates a slightly different distribution of errors, with a slightly higher tendency toward False Negatives compared to Random Forest. While XGBoost is highly precise in its positive predictions, its slightly lower recall (65.4%) means it missed a few more potential churners than the Random Forest model. This behavior suggests that the boosting process, while excellent for minimizing residual errors, may have been more conservative in classifying instances as "churn" to maintain high precision. The higher number of True Negatives shows that XGBoost is very efficient at identifying loyal customers, which can be useful if the company's priority is to avoid "disturbing" stable subscribers with unnecessary marketing. However, for a proactive retention strategy, the slightly higher sensitivity of Random Forest makes it a more suitable choice. Comparing both matrices side-by-side allows for a comprehensive assessment of the risk-reward trade-off inherent in each algorithm's classification logic.

3.4 ROC Curve and AUC Analysis

To further validate the discriminative power of the optimized models, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were analyzed. The ROC curve provides a visual representation of the trade-off between the True Positive Rate (sensitivity) and the False Positive Rate (1-specificity) at various threshold levels. An ideal model would have a curve that climbs quickly toward the top-left corner, indicating a high detection rate with minimal false alarms. In this study, the AUC score serves as a critical summary metric, where a score of 0.5 represents a model with no better predictive power than random guessing, while a score of 1.0 indicates perfect classification. By plotting the ROC curves for both Random Forest and XGBoost in a single comparative graph, we can observe which algorithm maintains a more robust performance across different operational scenarios. This analysis is particularly valuable for telecommunication companies that may need to adjust their intervention thresholds based on changing business costs and marketing strategies.

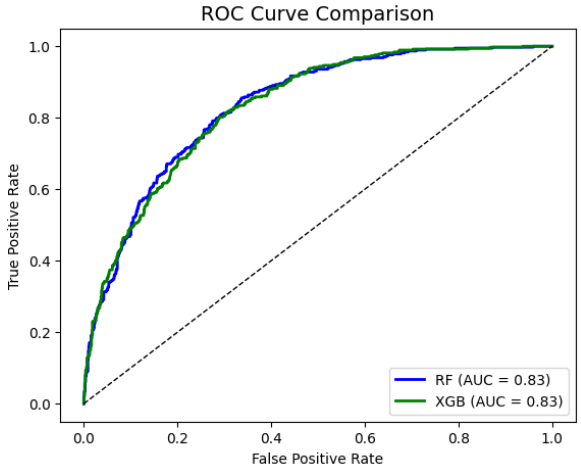


Figure 4. ROC Curve Comparison between Random Forest and XGBoost

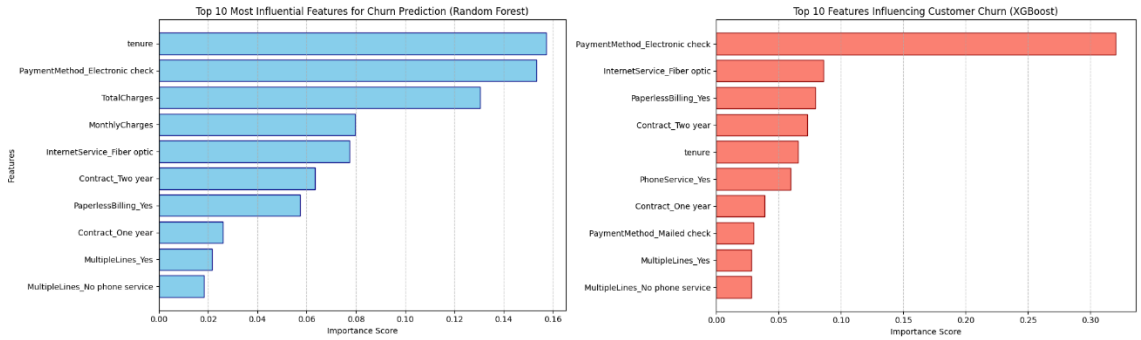
The comparative analysis of the ROC curves reveals that the Random Forest model achieved a slightly higher AUC score than the XGBoost model. Specifically, the Random Forest model recorded an AUC of approximately 0.83 (83%), while the XGBoost model followed closely with an AUC of 0.81 (81%). The 2% margin in AUC indicates that Random Forest is marginally more consistent in assigning higher probability scores to actual churners compared to non-churners. This superiority in discriminative ability is consistent with the higher recall and F1-score observed in the previous sections. The shape of the Random Forest curve demonstrates a more stable trajectory, suggesting that its bagging ensemble mechanism is highly effective at generalizing patterns within the Telco Churn dataset after SMOTE application. This high AUC value confirms that the model is not only accurate at a specific threshold but remains a strong predictor regardless of the sensitivity requirements imposed by the business stakeholders.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

In contrast, while the XGBoost curve displays a competitive performance, its slightly lower AUC suggests that its boosting process might be more sensitive to the complexity of the feature space in this particular dataset. However, an AUC of 0.81 still represents a "Good" classification performance according to traditional evaluation scales, making it a viable secondary model for churn prediction. The closeness of the two curves indicates that both algorithms successfully learned the fundamental drivers of customer attrition. For practical implementation, the higher AUC of Random Forest provides more confidence for automated decision-making systems, as it ensures a lower risk of misclassification across all possible threshold settings. This robust discriminative capability is essential for minimizing customer turnover and maximizing the effectiveness of loyalty programs in the highly competitive telecommunications market. The final section will discuss the importance of individual features in driving these high-performing results.

3.5 Feature Importance Analysis

The final stage of the results analysis involves identifying the key variables that drive the prediction of customer churn. Feature importance provides a transparent view into the "black-box" nature of ensemble models, allowing for an understanding of which subscriber attributes most significantly influence the decision-making process. By comparing the feature rankings from both the Random Forest and XGBoost models, this study identifies consistent patterns of customer behavior while highlighting the unique way each algorithm interprets feature interactions. Identifying these drivers is essential for telecommunication providers to move beyond mere prediction and develop targeted, data-driven retention strategies.



(a) (b)

Figure 5. Feature Importance Comparison (a) Random Forest, (b) XGBoost

As illustrated in Figure 5, both models identify "tenure", "TotalCharges", and "Contract" as the primary predictors of churn, albeit with different weight distributions. In the Random Forest model, "TotalCharges" and "tenure" appear as the most influential features. This suggests that the duration of the customer's relationship with the provider and their historical financial contribution are critical indicators of loyalty; a longer tenure typically correlates with a lower probability of churn. Conversely, the XGBoost model places an overwhelming emphasis on the "Contract" feature, specifically distinguishing between month-to-month and long-term commitments. This indicates that the lack of a long-term contractual obligation is the most immediate catalyst for customer attrition. When a customer is on a month-to-month plan, the "barrier to exit" is significantly lower, making them highly susceptible to switching providers based on price or service quality.

The analysis also highlights the secondary importance of service-related features such as "OnlineSecurity", "TechSupport", and "MonthlyCharges". The presence of these features in the top rankings suggests that value-added services play a role in "locking in" customers by increasing the perceived utility of the subscription. Customers without these security and support features are statistically more likely to experience dissatisfaction and churn. Interestingly, while Random Forest distributes importance across a wider range of features, XGBoost is more selective, focusing heavily on a few dominant predictors. This consistency across two different algorithmic approaches reinforces the validity of the findings. From a business perspective, these results imply that retention efforts should prioritize new customers on flexible contracts with high monthly bills by offering them loyalty incentives or migrating them toward more stable, long-term service agreements.

3.6 Discussion

The experimental results demonstrate that while both Random Forest and XGBoost are highly capable of predicting customer churn, Random Forest provides a more stable and balanced performance for this specific telecommunications dataset. The achievement of 77.5% accuracy and a recall of 67.6% by the Random Forest model signifies its robustness in handling high-dimensional data with complex non-linear relationships. This superior performance, particularly in terms of precision and the F1-score, can be attributed to the bagging ensemble mechanism. By constructing multiple independent decision trees and aggregating their votes, Random Forest effectively reduces variance and minimizes the impact of noise, which is often present in synthetic samples generated

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

by SMOTE. This finding aligns with the theoretical advantage of bagging methods in scenarios where the model needs to generalize across a balanced but potentially noisy feature space. The identical recall rate of 67.6% for both models further suggests that the primary challenge in churn prediction lies not only in the choice of algorithm but also in the quality of feature engineering and data balancing techniques employed during the preprocessing phase [52].

When compared to previous studies, the results of this research offer significant insights. Similar telecom datasets reported that gradient boosting methods often outperform random forests in terms of raw accuracy [53]. However, this research shows a slight deviation, where Random Forest maintains a better balance between precision and recall. This discrepancy can be explained by the specific hyperparameter tuning process and the degree of class imbalance in the original dataset. XGBoost is superior for large-scale datasets due to its regularization parameters, our findings suggest that for datasets of moderate size (approximately 7,000 instances), the complexity of XGBoost may lead to a more sensitive decision boundary that does not necessarily translate into higher accuracy [54]. Furthermore, the high importance of "Tenure" and "Contract Type", who identified that contractual obligations and loyalty duration are the most universal predictors of churn across the global telecommunications industry [55].

The practical implications of these findings are substantial for the telecommunications industry. The ability to identify 67.6% of actual churners (Recall) means that the company can proactively engage more than two-thirds of customers who are at risk of leaving. The higher precision of Random Forest (56.5%) compared to XGBoost (54.8%) is particularly important from a cost-efficiency perspective. A higher precision rate ensures that fewer resources are wasted on "False Positives" loyal customers who might be mistakenly targeted with expensive retention offers. This research demonstrates that an optimized Random Forest model, combined with SMOTE and GridSearchCV, provides a reliable framework for developing automated churn management systems. By focusing on the key drivers identified in the feature importance analysis, such as moving month-to-month subscribers to long-term plans, companies can leverage these ML insights to significantly reduce attrition rates and improve overall customer lifetime value [56].

4. CONCLUSION

This study has successfully conducted a comparative analysis of Random Forest and XGBoost for predicting customer churn in the telecommunications industry. The experimental results lead to several key conclusions. First, while both ensemble methods are highly effective, Random Forest emerged as the superior model for this specific dataset, achieving an overall accuracy of 77.5% and a balanced F1-score of 0.616. Although both models shared an identical recall rate of 67.6%, Random Forest demonstrated higher precision, indicating a more reliable identification of potential churners with fewer false positives. Second, the integration of SMOTE for handling class imbalance and GridSearchCV for hyperparameter optimization proved essential in enhancing the predictive stability of both algorithms. Third, the feature importance analysis revealed that customer tenure, total charges, and contract types specifically month-to-month plans are the most critical drivers of attrition. These findings suggest that the bagging approach of Random Forest is particularly resilient to the noise in telco subscriber data, making it a robust tool for automated retention systems. Ultimately, this research provides a clear empirical basis for telecommunication providers to adopt optimized machine learning models to mitigate customer loss and improve long-term business sustainability. Based on the limitations and findings of this study, several recommendations for future research are proposed. Future studies could explore the integration of more advanced deep learning architectures, such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, to capture the temporal sequences of customer behavior over time. Additionally, incorporating a broader range of external features, such as competitor pricing data, customer sentiment from social media, or macroeconomic indicators, could further enhance the model's predictive accuracy. Another potential avenue for research is the application of hybrid ensemble methods or advanced stacking techniques that combine the strengths of both bagging and boosting algorithms. Furthermore, future work should consider evaluating these models on larger and more diverse datasets from different geographical regions to test their generalizability across various market conditions. Finally, implementing an automated pipeline for continuous model retraining would be beneficial to ensure that the churn prediction system remains effective as customer behavior patterns evolve over time.

REFERENCES

[1] P. Wachwanakijkul, S. Junsiritrakhoon, N. Kantanantha, G. Narayanamurthy, and P. Jarumaneeroj, “Data-driven approaches to predicting customer churn in a non-contractual car-sharing company,” *Transp. Res. Interdiscip. Perspect.*, vol. 33, no. October 2024, pp. 1–20, 2025, doi: 10.1016/j.trip.2025.101600.

[2] H. GhorbanTanhaei, P. Boozary, S. Sheykhan, M. Rabiee, F. Rahmani, and I. Hosseini, “Predictive analytics in customer behavior: Anticipating trends and preferences,” *Results in Control and Optimization*, vol. 17, no. September, pp. 1–17, 2024, doi: 10.1016/j.rico.2024.100462.

[3] L. Theodorakopoulos, A. Theodoropoulou, and C. Klavdianos, “Big Data Analytics and AI for Consumer Behavior in Digital Marketing: Applications, Synthetic and Dark Data, and Future Directions,” *Big Data and Cognitive Computing*, vol. 10, no. 2, pp. 1–34, 2026, doi: 10.3390/bdcc10020046.

[4] M. Imani, A. Beikmohammadi, and H. R. Arabnia, “Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels,” *Technologies (Basel).*, vol. 13, no. 3, pp. 1–40, 2025, doi: 10.3390/technologies13030088.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

[5] H. N. Noura, T. Chu, Z. Allal, O. Salman, and K. Chahine, "A comparative study of ensemble methods and multi-output classifiers for predictive maintenance of hydraulic systems," *Results in Engineering*, vol. 24, no. September, pp. 1–20, 2024, doi: 10.1016/j.rineng.2024.102900.
[6] A. S. Assiri, "An Optimized Customer Churn Prediction Approach Based on Regularized Bidirectional Long Short-Term Memory Model," *Computers, Materials & Continua*, vol. 86, no. 1, pp. 1–21, 2026, doi: 10.32604/cmc.2025.069826.
[7] S. Shanmugam, E. Elavarasan, N. Madhavarao Seshadri, D. Ashokkumar, S. Senthilkumar, and T. Mohanavelu, "A Segmented Machine Learning Approach to Predicting and Mitigating Churn in the Gig Economy," *Journal of Theoretical and Applied Electronic Commerce Research* , vol. 21, no. 3, pp. 1–25, 2026, doi: 10.3390/jtaer21030093.
[8] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Machine Learning with Applications*, vol. 18, no. December 2023, pp. 1–31, 2024, doi: 10.1016/j.mlwa.2024.100597.
[9] Y. Zhang, L. Deng, and B. Wei, "Imbalanced Data Classification Based on Improved Random-SMOTE and Feature Standard Deviation," *Mathematics*, vol. 12, no. 11, pp. 1–17, 2024, doi: 10.3390/math12111709.
[10] H. Jiang, Y. Xia, C. Yu, Z. Qu, and H. Li, "On the implications of artificial intelligence methods for feature engineering in reliability sector with computer knowledge graph," *Alexandria Engineering Journal*, vol. 119, no. December 2024, pp. 587–597, 2025, doi: 10.1016/j.aej.2025.01.093.
[11] F. Gao and M. Abisado, "Enhanced Feature Engineering Symmetry Model Based on Novel Dolphin Swarm Algorithm," *Symmetry (Basel).*, vol. 17, no. 10, pp. 1–26, 2025, doi: 10.3390/sym17101736.
[12] P. Ziolkowski, "Computational Complexity and Its Influence on Predictive Capabilities of Machine Learning Models for Concrete Mix Design," *Materials*, vol. 16, no. 17, pp. 1–36, 2023, doi: 10.3390/ma16175956.
[13] G. Croitoru, A. Capatina, N. V. Florea, F. Codignola, and D. Sokolic, "A cross-cultural analysis of perceived value and customer loyalty in restaurants," *European Research on Management and Business Economics*, vol. 30, no. 3, pp. 1–16, 2024, doi: 10.1016/j.iedeen.2024.100265.
[14] J. Tang, "Unlocking Retail Insights: Predictive Modeling and Customer Segmentation Through Data Analytics," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 2, pp. 1–20, 2025, doi: 10.3390/jtaer20020059.
[15] M. Fang, H. Shi, H. Li, and T. Liu, "Application of Machine Learning for Productivity Prediction in Tight Gas Reservoirs," *Energies (Basel).*, vol. 17, no. 8, pp. 1–27, 2024, doi: 10.3390/en17081916.
[16] O. G. Al-Salih, D. Guangjian, W. J. Al-Mudhafar, and D. A. Wood, "Using extreme gradient boosting with Optuna hyperparameter tuning for efficient lost circulation prediction," *Energy Geoscience*, vol. 7, no. 2, pp. 1–20, 2026, doi: 10.1016/j.engeos.2026.100540.
[17] D. Sun, P. Zheng, J. Zhang, and L. Cheng, "Optimized Gradient Boosting Framework for Data-Driven Prediction of Concrete Compressive Strength," *Buildings*, vol. 15, no. 20, pp. 1–20, 2025, doi: 10.3390/buildings15203761.
[18] R. Suguna, J. Suriya Prakash, H. Aditya Pai, T. R. Mahesh, V. Vinoth Kumar, and T. E. Yimer, "Mitigating class imbalance in churn prediction with ensemble methods and SMOTE," *Sci. Rep.*, vol. 15, no. 1, pp. 1–20, 2025, doi: 10.1038/s41598-025-01031-0.
[19] B. S. Priya, G. Chitra, and R. Ramalakshmi, "Performance comparison of sampling techniques with machine learning algorithms for churn prediction in telecommunication," *Franklin Open*, vol. 13, no. August, pp. 1–14, 2025, doi: 10.1016/j.fraope.2025.100402.
[20] M. Z. Abedin, C. Guotai, P. Hajek, and T. Zhang, "Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk," *Complex and Intelligent Systems*, vol. 9, no. 4, pp. 3559–3579, 2023, doi: 10.1007/s40747-021-00614-4.
[21] O. R. Olaniran, A. R. R. Alzahrani, N. M. S. Alharbi, and A. A. Alzahrani, "Random Generalized Additive Logistic Forest: A Novel Ensemble Method for Robust Binary Classification," *Mathematics*, vol. 13, no. 7, pp. 1–25, 2025, doi: 10.3390/math13071214.
[22] X. Zhang *et al.*, "The XGBoost wind speed prediction model based on VMD-LSTM error correction," *Renew. Energy*, vol. 267, no. June 2025, pp. 1–13, 2026, doi: 10.1016/j.renene.2026.125708.
[23] S. R. Al-Taai, N. M. Azize, Z. A. Thoeny, H. Imran, L. F. A. Bernardo, and Z. Al-Khafaji, "XGBoost Prediction Model Optimized with Bayesian for the Compressive Strength of Eco-Friendly Concrete Containing Ground Granulated Blast Furnace Slag and Recycled Coarse Aggregate," *Applied Sciences (Switzerland)*, vol. 13, no. 15, pp. 1–23, 2023, doi: 10.3390/app13158889.
[24] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," *Results in Control and Optimization*, vol. 14, no. March 2023, pp. 1–16, 2024, doi: 10.1016/j.rico.2023.100342.
[25] M. Imani, M. Joudaki, A. Beikmohammadi, and H. R. Arabnia, "Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning," *Mach. Learn. Knowl. Extr.*, vol. 7, no. 3, pp. 1–38, 2025, doi: 10.3390/make7030105.
[26] X. Wang and D. Hou, "Enhancing Keystroke Dynamics Authentication with Ensemble Learning and Data Resampling Techniques," *Electronics (Switzerland)*, vol. 13, no. 22, pp. 1–22, 2024, doi: 10.3390/electronics13224559.
[27] I. Zerine *et al.*, "Explainable churn prediction in telecom with tabular ML five model benchmark and SHAP analysis," *Discover Artificial Intelligence*, vol. 6, no. 1, pp. 1–19, 2026, doi: 10.1007/s44163-026-00983-0.
[28] A. Barsotti *et al.*, "A Decade of Churn Prediction Techniques in the TelCo Domain: A Survey," *SN Comput. Sci.*, vol. 5, no. 4, pp. 1–15, 2024, doi: 10.1007/s42979-024-02722-7.
[29] M. K. Banjanin, M. Stojčić, D. Danilović, Z. Ćurguz, M. Vasiljević, and G. Puzić, "Classification and Prediction of Sustainable Quality of Experience of Telecommunication Service Users Using Machine Learning Models," *Sustainability (Switzerland)*, vol. 14, no. 24, pp. 1–29, 2022, doi: 10.3390/su142417053.
[30] P. B. Pires, B. M. Perestrelo, and J. D. Santos, "Measuring Customer Experience in E-Retail," *Adm. Sci.*, vol. 15, no. 11, pp. 1–33, 2025, doi: 10.3390/admsci15110434.
[31] S. Mishra *et al.*, "Agent-based modeling: Insights into consumer behavior, urban dynamics, grid management, and market interactions," *Energy Strategy Reviews*, vol. 57, no. December 2024, pp. 1–19, 2025, doi: 10.1016/j.esr.2024.101613.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 459−472
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9918

[32] Q. Y. Huang, N. D. Dizon, N. Jeyakumar, and V. Jeyakumar, "A distributionally robust machine learning model of simultaneous classification and feature selection under data uncertainty: Theory, methods, and application to the identification of Alzheimer's disease using handwriting," *EURO Journal on Computational Optimization*, vol. 13, no. July, pp. 1–22, 2025, doi: 10.1016/j.ejco.2025.100111.

[33] M. Shahabikargar, A. Beheshti, X. Zhang, J. Foo, and A. Jolfaei, "A comprehensive survey on customer churn analysis studies," *Journal of Information and Telecommunication*, vol. 10, no. 1, pp. 24–70, 2025, doi: 10.1080/24751839.2025.2528440.

[34] M. Madanchian, "The Role of Complex Systems in Predictive Analytics for E-Commerce Innovations in Business Management," *Systems*, vol. 12, no. 10, pp. 1–20, 2024, doi: 10.3390/systems12100415.

[35] H. Çelikten, "Evaluating machine learning (RF, XGBoost) and statistical model (MLR) for PM10 and air quality prediction: A case from Kars, Türkiye," *Atmos. Pollut. Res.*, vol. 17, no. 6, pp. 1–15, 2026, doi: 10.1016/j.apr.2026.102975.

[36] J. B. Ruhland, I. Masoudian, and D. Heider, "Enhancing deep neural network training through learnable adaptive normalization," *Knowl. Based. Syst.*, vol. 326, no. July, pp. 1–9, 2025, doi: 10.1016/j.knosys.2025.113968.

[37] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," *AI (Switzerland)*, vol. 6, no. 10, pp. 1–40, 2025, doi: 10.3390/ai6100257.

[38] Y. B. Wah *et al.*, "Machine Learning and Synthetic Minority Oversampling Techniques for Imbalanced Data: Improving Machine Failure Prediction," *Computers, Materials and Continua*, vol. 75, no. 3, pp. 4821–4841, 2023, doi: 10.32604/cmc.2023.034470.

[39] S. V. Oprea and A. Bâra, "Customer-Centric Decision-Making with XAI and Counterfactual Explanations for Churn Mitigation," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 2, pp. 1–25, 2025, doi: 10.3390/jtaer20020129.

[40] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Applied Sciences (Switzerland)*, vol. 13, no. 6, pp. 1–34, 2023, doi: 10.3390/app13064006.

[41] K. Mokgwatjane and T. Paepae, "An explainable ensemble machine learning approach for multi-domain, multiclass sentiment analysis in Amazon product reviews," *Machine Learning with Applications*, vol. 23, no. August 2025, pp. 1–17, 2026, doi: 10.1016/j.mlwa.2025.100825.

[42] M. O. Ajinaja *et al.*, *A Comparative Evaluation of Probabilistic and Transformer-Based Topic Models Across Diverse and Multilingual Text Corpora*, vol. 58, no. 1. 2026. doi: 10.1007/s11063-025-11820-3.

[43] P. Boozary, S. Sheykhan, H. GhorbanTanhaei, and C. Magazzino, "Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction," *International Journal of Information Management Data Insights*, vol. 5, no. 1, pp. 1–15, 2025, doi: 10.1016/j.jjimei.2025.100331.

[44] Y. Xia, S. Jiang, L. Meng, and X. Ju, "XGBoost-B-GHM: An Ensemble Model with Feature Selection and GHM Loss Function Optimization for Credit Scoring," *Systems*, vol. 12, no. 7, pp. 1–26, 2024, doi: 10.3390/systems12070254.

[45] S. K. Kiangala and Z. Wang, "An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment," *Machine Learning with Applications*, vol. 4, no. December 2020, pp. 1–15, 2021, doi: 10.1016/j.mlwa.2021.100024.

[46] C. Mueangphaen, W. Kusonkhum, K. Kuntiyawichai, T. Pannachet, R. Nuntasarn, and M. Boonpichetvong, "Comparative multi-algorithm AI framework for real-time carbon emission optimization in a medium-scale irrigation project in Thailand," *Environ. Impact Assess. Rev.*, vol. 118, no. November 2025, pp. 1–13, 2026, doi: 10.1016/j.eiar.2025.108276.

[47] K. Sandunil, Z. Bennour, H. Ben Mahmud, and A. Giwelli, "Effects of tuning decision trees in random forest regression on predicting porosity of a hydrocarbon reservoir. A case study: volve oil field, north sea," *Energy Advances*, vol. 3, no. 9, pp. 2335–2347, 2024, doi: 10.1039/d4ya00313f.

[48] R. Agrawal *et al.*, "Improving Predictive Performance in Telecom Churn Modeling with Hybrid SMOTE and GAN-Based Synthetic Data Generation," *International Journal of Computational Intelligence Systems*, vol. 19, no. 1, pp. 1–23, 2026, doi: 10.1007/s44196-026-01204-3.

[49] M. Risha and P. Liu, "Comparative machine learning facies prediction using ensemble boosting models and support vector machine versus unsupervised clustering," *Discover Geoscience*, vol. 4, no. 1, pp. 1–22, 2026, doi: 10.1007/s44288-026-00398-5.

[50] J. I. Iturbe-Araya and H. Rifà-Pous, "Hyperparameter Optimization and Evaluation Metrics for Unsupervised Anomaly-Based Cyberattack Detection in Imbalanced Smart Home Datasets," *Journal of Network and Systems Management*, vol. 33, no. 4, pp. 1–37, 2025, doi: 10.1007/s10922-025-09973-6.

[51] M. P. Pretel *et al.*, "Machine Learning Models for Predicting Professional Disqualification in Peruvian Association Members," *Data (Basel).*, vol. 11, no. 98, pp. 1–20, 2026, doi: 10.3390/data11050098.

[52] N. Salaeh *et al.*, "Resampling-driven machine learning models for enhanced high streamflow forecasting," *Water Cycle*, vol. 7, no. July 2025, pp. 99–119, 2026, doi: 10.1016/j.watcyc.2025.07.001.

[53] M. Martinović, K. Dokic, and D. Pudić, "Comparative Analysis of Machine Learning Models for Predicting Innovation Outcomes: An Applied AI Approach," *Applied Sciences (Switzerland)*, vol. 15, no. 7, pp. 1–44, 2025, doi: 10.3390/app15073636.

[54] P. A. H. Pham and N. D. Hoang, "Metaheuristic optimization of extreme gradient boosting machine for enhanced prediction of lateral strength of reinforced concrete columns under cyclic loadings," *Results in Engineering*, vol. 24, no. July, pp. 1–18, 2024, doi: 10.1016/j.rineng.2024.103125.

[55] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra, and V. Benson, "Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models," *Algorithms*, vol. 17, no. 6, 2024, doi: 10.3390/a17060231.

[56] S. S. Poudel, S. Pokharel, and M. Timilsina, "Explaining customer churn prediction in telecom industry using tabular machine learning models," *Machine Learning with Applications*, vol. 17, no. March, pp. 1–9, 2024, doi: 10.1016/j.mlwa.2024.100567.