Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 18(2) 2025: 25-36
http://ejournal.undip.ac.id/index.php/media_statistika

EVALUATING *RANDOM FOREST* AND XGBOOST FOR BANK *CUSTOMER CHURN* PREDICTION ON IMBALANCED DATA USING SMOTE AND SMOTE-ENN

Reyuli Andespa, Kusman Sadik, Cici Suhaeni, Agus Mohamad Soleh
Department of Statistics and Data Science, IPB University, Bogor, Indonesia

e-mail: *reyuli12andespa@apps.ipb.ac.id*

DOI: 10.14710/medstat.18.1.25-36

Article Info:
Received: 16 September 2025
Accepted: 11 October 2025
Available Online: 14 October 2025

Keywords:
Customer Churn; XGBoost; Random Forest; SMOTE; Imbalanced Data.

Abstract: The banking industry faces significant challenges in retaining customers, as churn can critically affect both revenue and reputation. This study introduces a robust churn prediction framework by comparing the performance of XGBoost and Random Forest algorithms under imbalanced data conditions. The novelty of this research lies in integrating the SMOTE and SMOTE-ENN techniques with machine learning algorithms to enhance model performance and reliability on highly imbalanced datasets. Unlike conventional approaches that rely solely on oversampling or undersampling, this study demonstrates that the hybrid combination of XGBoost and SMOTE provides superior predictive accuracy, stability, and efficiency. Hyperparameter optimization using GridSearchCV was conducted to identify the most effective parameter configurations for both algorithms. Model performance was evaluated using the F1-Score and Area Under the Curve (AUC). The results indicate that XGBoost with SMOTE achieved the best performance, with an F1-Score of 0.8730 and an AUC of 0.9828, showing an optimal balance between precision and recall. Feature importance analysis identified Months_Inactive_12_mon, Total_Trans_Amt, and Total_Relationship_Count as the most influential predictors. Overall, this approach outperforms traditional resampling and modeling techniques, providing practical insights for data-driven customer retention strategies in the banking industry.

1. INTRODUCTION

The banking industry in the digital era faces significant challenges in maintaining customer loyalty. One major issue is customer churn when customers discontinue their banking services which can substantially impact revenue and reputation. Therefore, banks require reliable churn prediction systems to identify at-risk customers and take timely mitigation measures (Azmi & Voutama, 2024; Boozary et al., 2025; Mahmoudzadeh & Shirali Shahreza, 2025).

Machine learning (ML) Machine learning (ML) techniques have been widely applied to build customer churn prediction models. Among the most effective algorithms are Random Forest (RF) and XGBoost, known for handling complex data and delivering high

accuracy in classification tasks (Azmi & Voutama, 2024; Hambali & Andrew, 2024). However, a persistent challenge in churn prediction lies in class imbalance, where non-churn customers greatly outnumber churners, causing models to overlook the minority class. Banking churn rates can critically affect profitability, highlighting the urgent need for accurate churn prediction. To address this issue, the hybrid data-balancing method SMOTE-ENN (Synthetic Minority Over-sampling Technique–Edited Nearest Neighbors) is employed to generate a more representative dataset (Amalia & Asmunin, 2024).

This study evaluates the performance of Random Forest and XGBoost in predicting bank customer churn using SMOTE-ENN and hyperparameter optimization via GridSearchCV. The novelty of this research lies in integrating two ensemble algorithms with hybrid data-balancing methods under systematic optimization, offering a comprehensive comparison and new insights into the optimal configuration for churn prediction in the banking sector.

2. LITERATURE REVIEW

2.1. SMOTE-ENN

SMOTE-ENN is a hybrid method that combines the Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbors (ENN) to address data imbalance more effectively than using oversampling or undersampling techniques alone (Batista et al., 2021). The main steps in the SMOTE-ENN process are as follows:

1. SMOTE (*Synthetic Minority Over-sampling Technique*):
   SMOTE generates synthetic samples for the minority class by interpolating between existing minority samples. The formula for generating synthetic data is

$$x_{sintesis} = x_i + (x_{knn} - x_i) \times \delta \quad (1)$$

   where $x_{sintesis}$: the synthetic data to be generated; x_i: the minority sample to be replicated; x_{knn}: the nearest neighbor of x_i; δ: a random value in the range [0,1].
   SMOTE creates synthetic samples for the minority class through linear interpolation between an existing sample and one or more of its k-nearest neighbors, as illustrated in Figure 1.

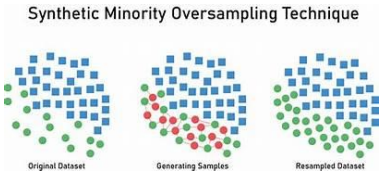

Figure 1. Training Data in SMOTE

2. ENN (Edited Nearest Neighbors)
   Edited Nearest Neighbors (ENN) is applied after SMOTE to improve dataset quality by removing majority-class samples misclassified by k-nearest neighbors (k-NN). This process reduces noise, ensures a more representative class distribution, and helps prevent overfitting, thereby enhancing model generalization and producing more accurate predictions, especially on highly imbalanced datasets (García et al., 2024).

2.2. Random Forest

Random Forest is an ensemble machine learning algorithm that employs a bagging approach to enhance prediction accuracy. The algorithm constructs multiple decision trees independently from different subsets of the data, and the final prediction is reached via majority voting for classification or averaging for regression. The process utilizes bootstrap samples from the original dataset, with only a random subset of features considered at each

node split to ensure diversity and prevent overfitting. This ensemble method effectively reduces variance, improves stability, and is known for its ability to handle complex, imbalanced features, missing values, and identifying influential predictors (Zhang et al., 2023).

In selecting the best features during the construction of decision trees, the Gini index is used as a measure of impurity or heterogeneity within a dataset. It is defined by the following equation:

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \quad (2)$$

where P_i represents the proportion of samples belonging to class i, and $m$denotes the total number of classes. A lower Gini value indicates a purer node, meaning that the samples are more homogeneous in terms of their class labels. The total Gini index at an internal node K is then calculated using the following equation:

$$Gini\ (K) = \frac{T_1}{T} Gini(D_1) + \frac{T_2}{T} Gini(D_2) \quad (3)$$

where T_1 and T_2 denote the number of observations in the left and right child nodes, respectively, and T represents the total number of observations in the parent node (Han et al., 2022). Through this process, Random Forest combines the predictive power of multiple trees to capture nonlinear relationships and manage noisy or imbalanced data effectively (Bibi et al., 2024).

2.3. XGBoost

XGBoost is a boosting algorithm that builds decision trees sequentially, where each tree corrects the errors of the previous ones using a gradient boosting approach. It offers high computational efficiency, supports both L1 and L2 regularization, handles missing values automatically, and performs well on large and imbalanced datasets (Chen & Guestrin, 2023).

The prediction process in XGBoost iteratively adds decision trees to minimize residual errors, and the final prediction is obtained by aggregating the outputs of all trees. Optimal hyperparameters are determined using GridSearchCV, which systematically evaluates combinations of parameters such as *n_estimators* (number of boosting rounds), *max_depth* (maximum tree depth controlling model complexity), *learning_rate* (step size determining each tree's contribution), *subsample* (fraction of training samples used per tree to reduce overfitting), *colsample_bytree* (fraction of features randomly selected per tree), *gamma* (minimum loss reduction required for a split), and *reg_alpha* and *reg_lambda* (L1 and L2 regularization terms). The configuration that achieves the highest F1-score and AUC is selected as the optimal model (Bibi et al., 2024).

2.4. Evaluation Techniques

The confusion matrix is a tabular representation employed to evaluate the accuracy of a classification model by comparing its predictions with the actual data (Manliguez, 2016).

Table 1. *Confusion Matrix*

	Predicted Negative	*Predicted Positive*
Actual Negative	*True Negative (TN)*	*False Positive (FP)*
Actual Positive	*False Negative (FN)*	*True Positive (TP)*

Based on Table 1, in classification with imbalanced data, overall classification accuracy is often not an appropriate measure of performance. Accuracy tends to be dominated by the correctness on the minority class; therefore, more suitable validation metrics include balanced accuracy, accuracy, sensitivity, and specificity (Zhang et al., 2020). In this study,

several evaluation techniques will be employed to assess the performance of the two methods used, namely:

1. F1-score
   The F1-score is the harmonic mean of precision and recall, and it is employed when a balance between the two is required. The equation for the F1-score is expressed as follows:

$$F1 - score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (4)$$

   The F1-score is particularly useful when there is class imbalance in the data, as it considers both types of misclassifications: false negatives (failing to detect the positive class) and false positives (incorrectly predicting the positive class). This metric provides a single measure that balances the trade-off between precision and recall, making it highly suitable for evaluating classification models on imbalanced datasets (Sun et al., 2021).
2. ROC Curve (Receiver Operating Characteristic)
   The ROC (Receiver Operating Characteristic) curve, along with its associated metric AUC (Area Under the Curve), is a visualization used to evaluate classification performance and compare models. It plots the true positive rate (TPR) against the false positive rate (FPR) across varying thresholds. The AUC measures the area under the ROC curve and represents the probability that a classifier correctly distinguishes between positive and negative instances. A higher AUC value indicates better model performance, where a value near 1 denotes excellent discrimination and 0.5 indicates no discriminative ability. The interpretation categories of AUC values are shown in Table 2 (Zou et al., 2021).

Table 2. Categories of AUC Values

AUC Value	Categories
0.90 – 1.00	Excellent Classification
0.80 – 0.90	Good Classification
0.70 – 0.80	Moderate Classification
0.60 – 0.70	Poor Classification
0.50 – 0.60	Very Poor Classification

3. MATERIAL AND METHOD

3.1. Data and Variables

The dataset used in this study was obtained from the Analyttica platform and comprises information from 10,127 bank customers. It contains 20 features describing customer characteristics such as age, marital status, credit limit, card category, and card utilization ratio. The target variable is binary, representing churn status, where 0 indicates customers who did not churn and 1 indicates those who churned. Since only 16.07% of customers churned, the dataset is highly imbalanced, requiring data balancing techniques to prevent bias toward the majority class (Zhang et al., 2023).

3.2. Analysis Steps

The stages of this research comprised six main steps: (1) Data Collection and Understanding involved retrieving and examining customer churn data from the Analyttica platform; (2) Data Preprocessing subsequently conducted data cleaning, missing value checks, outlier handling, and normalization of numerical variables; (3) Data Balancing followed, where the preprocessed dataset was duplicated into two versions: imbalanced (original data) and balanced using the SMOTE-ENN algorithm; (4) Model Development (Imbalanced Data) included building and hyperparameter tuning of Random Forest and

XGBoost classification models on the imbalanced dataset via GridSearchCV; (5) Model Development (Balanced Data) involved constructing and tuning the same models using the balanced (SMOTE-ENN) dataset; and finally, (6) Model Evaluation assessed the performance of all models using F1-score and ROC-AUC metrics to determine the most effective algorithm for predicting customer churn.

4. RESULTS AND DISCUSSION

The data used in this study consists of customer churn records from a bank, comprising one response variable, *Attrition_Flag*, and 19 explanatory variables, both numerical and categorical. The response variable *Attrition_Flag* is binary, where a value of 1 represents customers who churned (discontinued using the bank's services) and a value of 0 represents customers who remained active. During the data exploration stage, descriptive statistical analysis was performed to identify the initial characteristics of the dataset. Due to the large number of variables, only several representative ones are presented in Table 3.

Table 3. Descriptive Statistics of Selected Exploratory Variables

	Customer_Age	Dependent_count	Month_on_book	Months_Inactive_12_mon	Contacts_Count_12_mon
Mean	46.32	2.35	35.93	2.34	2.45
Std	8.02	1.29	7.98	1.01	1.10
Min	26.00	0.00	13.00	0.00	0.00
Max	73.00	5.00	56.00	6.00	6.00

Table 3 shows that the average customer age is 46 years (range 26–73), and the average number of dependents is 2.35, indicating mostly small families. Customers have been with the bank for an average of 35.93 months, with 2.34 months of inactivity per year and 2.45 contacts over the last 12 months, suggesting moderate engagement. These descriptive statistics highlight variability in key attributes that may influence churn. Subsequently, a correlation analysis was performed to examine the relationships between the variables and the target feature, Attrition_Flag, as shown in Figure 2.

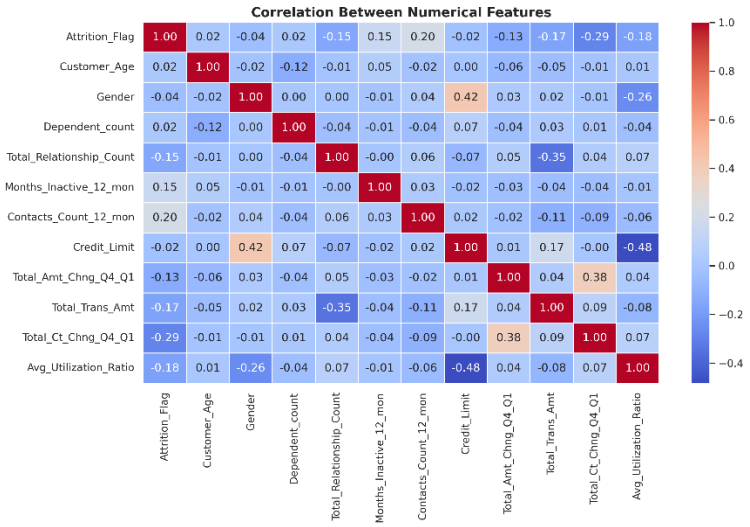

Figure 2. Correlation Matrix

Based on the Pearson correlation matrix, most features exhibit weak correlations with the target variable, Attrition_Flag, with absolute values generally below 0.4, indicating no strong linear associations with churn. The strongest negative correlations are

Total_Ct_Chng_Q4_Q1 (-0.29) and Avg_Utilization_Ratio (-0.18), suggesting that customers with more transactions and higher credit utilization are less likely to churn. Other features, such as Customer_Age (0.02), Gender (-0.04), Dependent_count (0.02), and Credit_Limit (-0.01), show negligible correlations, while Months_Inactive_12_mon (0.15) indicates slightly higher churn risk for inactive customers.

Some feature pairs display high inter-correlations, such as Credit_Limit with Avg_Open_To_Buy (1.00) and Customer_Age with Months_on_book (0.79), which may lead to multicollinearity if included together in a predictive model. Overall, no single feature has a dominant linear effect on churn, highlighting the relevance of machine learning models like Random Forest and XGBoost that can capture non-linear relationships. This correlation analysis also provides a foundation for feature selection and subsequent modeling to improve predictive accuracy.

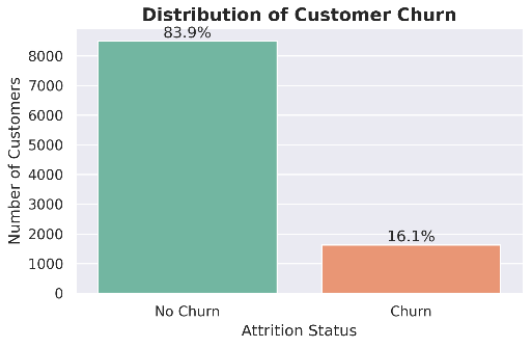

Figure 3. Distribution of the Target Variable (*Attrition_Flag*)

Figure 3 illustrates the distribution of the response variable *Attrition_Flag*, which represents customer churn status. The exploratory data analysis shows that the data are imbalanced, with 83.9% of observations belonging to the 'No Churn' class (customers who remain active) and only 16.1% belonging to the 'Churn' class (customers who discontinued banking services). Such class imbalance may lead to prediction bias in machine learning algorithms, where the model tends to be more accurate in classifying the majority class but performs poorly on the minority class.

To statistically validate this imbalance, a chi-square goodness-of-fit test was conducted, yielding a significant result ($\chi^2 = 4664.57$, $p < 0.001$). This confirms that the distribution of the target variable is highly imbalanced, justifying the need for resampling techniques prior to model training.

Furthermore, a point-biserial correlation analysis was performed to evaluate linear effects between churn status and continuous predictors. The results indicated that *Contacts_Count_12_mon* ($r = 0.204$, $p < 0.001$) and *Months_Inactive_12_mon* ($r = 0.152$, $p < 0.001$) were positively associated with churn, while *Total_Trans_Amt* ($r = -0.169$, $p < 0.001$) and *Total_Relationship_Count* ($r = -0.150$, $p < 0.001$) were negatively associated. These findings highlight that customer activity and relationship indicators play significant linear roles in predicting churn.

To address this issue, data imbalance was handled using a hybrid approach combining oversampling and undersampling, namely the SMOTE-ENN (Synthetic Minority Oversampling Technique - Edited Nearest Neighbors) method. SMOTE generates new synthetic examples from the minority class, while ENN removes noisy observations from the majority class. The combination of these two methods aims to create a more balanced and cleaner class distribution, thereby enabling classification algorithms such as Random

Forest and XGBoost to learn more fairly and improve their accuracy in detecting churn. This data balancing process constitutes a crucial step in building a reliable churn prediction model that can be applied effectively in banking management decision-making.

Before modeling, preprocessing of the data must first be carried out. The necessary steps include detecting missing values and duplicate records. Based on the output, no missing values or duplicate data were found. Since the dataset contained neither missing values nor duplicates, no further treatment was required, and the process proceeded directly to data splitting. In the splitting process, the dataset was divided into training and testing sets with an 80:20 ratio.

In addition, feature selection was performed to reduce redundancy and avoid multicollinearity issues that could affect model performance. The correlation analysis revealed several pairs of features with very high correlations, such as *Credit_Limit* with *Avg_Open_To_Buy* (1.00), *Total_Revolving_Bal* (0.62), *Total_Trans_Ct* (0.81), and *Customer_Age* with *Months_on_book* (0.79). Because such strong correlations indicate that these features carry nearly identical information, the variables *Avg_Open_To_Buy* and *Months_on_book* were removed. This feature elimination aimed to simplify the model, enhance stability, and reduce the risk of overfitting that may arise from highly correlated features.

Based on the previous descriptive analysis, it was observed that the class distribution in the dataset is imbalanced. Therefore, handling of class imbalance was carried out using SMOTE and SMOTE-ENN after splitting the data into training and testing sets with an 80:20 ratio.

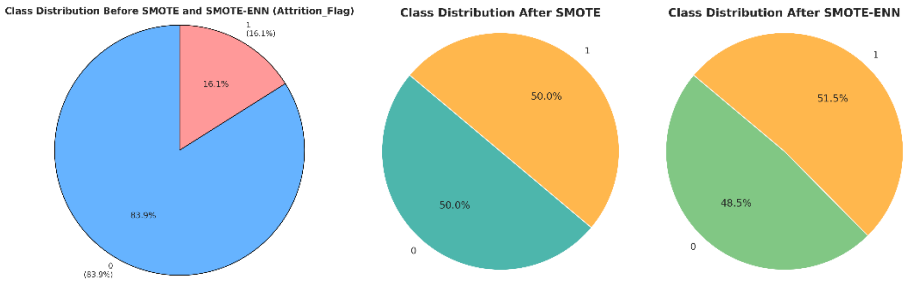

Figure 4. Class Proportions of the *Attrition_Flag* Before and After SMOTE and SMOTE-ENN

Figure 4 illustrates the class distribution of the target variable Attrition_Flag at three stages: before oversampling, after applying SMOTE, and after applying SMOTE-ENN. Initially, the dataset was highly imbalanced, with 83.9% of observations in the majority class (Class 0-No Churn) and only 16.1% in the minority class (Class 1- Churn). After applying SMOTE, the class distribution became perfectly balanced (50% : 50%), confirming that the method successfully generated synthetic minority samples to allow the model to learn effectively from both classes.

Subsequently, SMOTE-ENN was applied to further improve data quality by removing noisy or ambiguous samples after oversampling. The resulting distribution shifted slightly to 60% majority and 40% minority, reflecting the ENN step which removes overlapping or outlier instances, particularly from the majority class. This produces a cleaner and more representative dataset for model training, balancing the need for sufficient minority samples with the removal of potentially misleading observations.

This study evaluates the performance of customer churn classification models by comparing three scenarios for handling imbalanced data: a baseline model without balancing, a model using the Synthetic Minority Oversampling Technique (SMOTE), and a model using the combination of SMOTE and Edited Nearest Neighbor (SMOTE-ENN). Two machine learning algorithms were applied in each scenario: Random Forest and XGBoost. Model evaluation was conducted using several metrics, including F1 score and AUC (Area Under the Curve), with primary emphasis on the F1 score, as this metric provides a balance between precision and recall, which is particularly important in churn classification tasks.

The first models were built across three scenarios without performing any parameter tuning. Before the tuning process, initial models were developed under three scenarios—original (imbalanced), SMOTE, and SMOTE-ENN to establish a baseline comparison. Both Random Forest and XGBoost were first trained using their default configurations. The Random Forest model employed parameters such as *n_estimators = 100*, *criterion = 'gini'*, and *max_features = 'sqrt'*, while the XGBoost model used *objective = 'binary:logistic'* and *eval_metric = 'logloss'*. To illustrate the models' baseline predictive behavior, a subset of predicted outputs was presented, where the predicted labels [0 0 0 0 0 0 0 0 0 0] indicate that all ten test samples were classified as non-churn customers. The corresponding predicted probabilities, ranging from 0.01 to 0.38 for Random Forest and from 0.0001 to 0.075 for XGBoost, further demonstrate the models' low confidence in identifying churn cases. These findings suggest that, prior to hyperparameter optimization, both models tended to bias toward the majority (non-churn) class, reflecting the residual effect of data imbalance even after balancing treatments. The detailed performance results are summarized in Table 4.

Table 4. Model Evaluation Before Hyperparameter Tuning

Scenarios	Model Names	F1-Score	AUC
Base Model	Random Forest	0.7594	0.9702
	XGBoost	0.8576	0.9849
SMOTE	Random Forest	0.7647	0.9610
	XGBoost	0.8666	0.9812
SMOTE-ENN	Random Forest	0.7270	0.9463
	XGBoost	0.8218	0.9741

Based on the output in Table 4, it can be observed that XGBoost consistently outperformed Random Forest across all three tested data scenarios: the original data (Base Model), the oversampled data using SMOTE, and the combined SMOTE-ENN data. On the original data, XGBoost achieved an F1-Score of 0.8576 and an AUC of 0.9849, surpassing Random Forest. When SMOTE was applied, both models showed an increase in F1-Score, with XGBoost reaching 0.8666 and an AUC of 0.9812, indicating the effectiveness of SMOTE in balancing the data and improving minority class detection.

However, in the SMOTE-ENN scenario, Random Forest experienced a decrease in performance (F1-Score 0.7270), and XGBoost also showed a decline in F1-Score to 0.8218, likely due to the ENN cleaning process removing crucial information, making the model overly aggressive and increasing false positives. This finding aligns with previous research (Bunkhumpornpat et al., 2012) regarding the tendency of SMOTE-ENN to overfit when the data are filtered too strictly. On the other hand, XGBoost demonstrated good resilience under this scenario. These results suggest that XGBoost is a more robust choice, while the implementation of SMOTE-ENN requires careful consideration, as its impact may vary across different algorithms.

After building the baseline models, hyperparameter tuning was conducted to optimize the performance of both Random Forest and XGBoost models, aiming to improve predictive

accuracy and reduce classification errors. Grid Search with cross-validation (GridSearchCV) was employed to systematically explore various parameter combinations within predefined ranges, which were determined based on theoretical considerations, prior studies on customer churn, and preliminary experimentation to balance model complexity and generalization. For Random Forest, the search space included *n_estimators* (number of trees), *max_depth* (maximum tree depth), *min_samples_split* (minimum samples required to split a node), *min_samples_leaf* (minimum samples required to be at a leaf node), and *max_features* (number of features considered). The optimal configuration obtained through Grid Search was *n_estimators = 300*, *max_depth = 30*, *min_samples_split = 5*, *min_samples_leaf = 3*, and *max_features = sqrt or 0.5*. For XGBoost, the parameter grid was adapted to the algorithm's boosting structure, involving *n_estimators = 300*, *max_depth = 8*, *learning_rate = 0.07*, *subsample = 0.8*, *colsample_bytree = 0.8*, *gamma = 0.1*, and *reg_alpha = 0.01*, which were identified as the best-performing settings. The evaluation results after hyperparameter tuning are presented in Table 5.

Table 5. Model Evaluation After Hyperparameter Tuning

Scenarios	Model Names	F1-Score	AUC
Base Model	Random Forest	0.7847	0.9625
	XGBoost	0.8636	0.9839
SMOTE	Random Forest	0.8325	0.9743
	XGBoost	0.8730	0.9828
SMOTE-ENN	Random Forest	0.7268	0.9452
	XGBoost	0.8207	0.9692

Based on Table 5, in the Base Model scenario, XGBoost continued to demonstrate superior performance with an F1-Score of 0.8636 and an AUC of 0.9839, slightly outperforming Random Forest, which achieved an F1-Score of 0.7847 and an AUC of 0.9625.

When the data were oversampled using SMOTE, both models showed improved performance. Random Forest reached an F1-Score of 0.8325 and an AUC of 0.9743, while XGBoost performed better with an F1-Score of 0.8730 and an AUC of 0.9828, indicating the effectiveness of SMOTE in enhancing minority class detection.

However, in the SMOTE-ENN scenario, an interesting pattern emerged. Random Forest experienced a significant performance drop (F1-Score 0.7268, AUC 0.9452), likely due to the ENN cleaning process removing essential samples, resulting in a loss of crucial information for optimal classification. In contrast, XGBoost demonstrated greater resilience, with a slight decrease in F1-Score to 0.8207 and AUC to 0.9692, maintaining strong performance.

Overall, these tuning results indicate that XGBoost is a more robust and consistent algorithm compared to Random Forest across various data scenarios. The use of SMOTE has been shown to effectively improve model performance on imbalanced datasets. In contrast, the application of SMOTE-ENN requires careful consideration; although intended to reduce noise, it can inadvertently increase false positives and reduce precision (Imani et al., 2025).

Based on F1 score and AUC values, the best-performing model in this study is XGBoost with imbalanced data handling using SMOTE, achieving an F1 score of 0.8730 and an AUC of 0.9828. This model attains the optimal trade-off between high recall and maintained precision, making it ideal for churn classification. On the other hand, although SMOTE-ENN yielded the highest recall, it did so at the expense of precision, reducing the

model's practical utility. In a business context, misclassifying non-churn customers as churn (false positives) can lead to unnecessary intervention costs, such as offering discounts or special promotions; therefore, models that maintain a balanced precision are preferred.

Based on F1-Score and AUC values after hyperparameter tuning, the XGBoost with SMOTE model achieved the most optimal performance, with an F1-Score of 0.8730 and an AUC of 0.9828. This configuration provides the best trade-off between minority class detection (high recall) and maintained precision, making it ideal for imbalanced classification tasks. While some scenarios (like Random Forest with SMOTE-ENN) yielded high recall, this came at the expense of precision, resulting in a drastic decline in F1-Score. In a business context, misclassifying non-churn customers as churn (*false positives*) leads to inefficient resource allocation or unnecessary intervention costs. Therefore, a model that balances precision and recall, such as XGBoost with SMOTE, is preferred to ensure both effectiveness and efficiency in practical implementation.

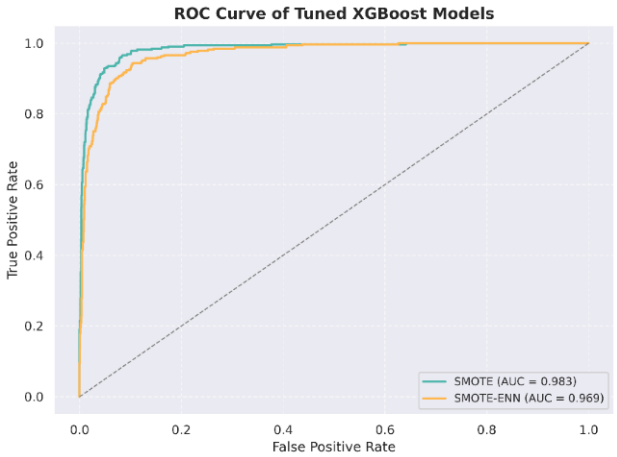

Figure 6. ROC Curve Plot

Figure 6. ROC curves of the tuned XGBoost model on the Test Set for the SMOTE (blue) and SMOTE-ENN (orange) scenarios, showing high discriminative ability with AUC values of 0.983 and 0.969, respectively. Although both models demonstrate high discriminative performance, the slight gap (AUC = 0.014) indicates that the ENN cleaning process, while effective in removing noisy samples, may also discard informative data, leading to slightly lower generalization. In contrast, SMOTE maintains sample diversity and achieves a better balance between recall and precision, making it more suitable for practical implementation.

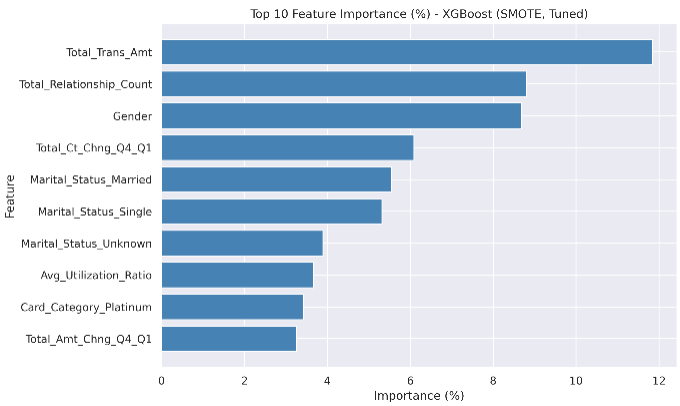

Figure 7. Feature Importance

Figure 7 illustrates the top ten important features in the tuned XGBoost model trained with SMOTE. The most influential variables are *Months_Inactive_12_mon* and *Total_Trans_Amt*, indicating that customer inactivity and transaction volume are key behavioral drivers of churn. Other features such as *Contacts_Count_12_mon*, *Total_Ct_Chng_Q4_Q1*, and *Total_Relationship_Count* represent customer interaction frequency and behavioral changes that strongly influence churn risk, while demographic factors like *Gender*, *Customer_Age*, and *Dependent_count* have smaller effects. Compared to previous studies that emphasized demographic attributes as primary churn indicators, these results highlight a gap showing that behavioral and transactional patterns provide stronger predictive power, offering more practical insights for designing targeted retention strategies.

Although Random Forest and XGBoost are widely used for churn prediction tasks, existing studies seldom examine their relative behavior when trained on data balanced via SMOTE versus SMOTE-ENN, especially in banking contexts (Sari et al., 2024; Zhang et al., 2023). Moreover, while hybrid resampling methods like SMOTE-ENN combined with ensemble algorithms have shown promise in e-commerce churn prediction, their performance in banking sector datasets with unique customer behavior patterns remains underexplored (Al-Saqqa et al., 2023; Sahare & Gupta, 2022). Finally, there is a lack of research comparing model stability and generalization across different imbalance ratios, which is critical for assessing how robust each algorithm is in real-world settings (Imani et al., 2023).

5. CONCLUSION

This study evaluated the performance of Random Forest and XGBoost in predicting customer churn on imbalanced data across three scenarios: Base Model, SMOTE, and SMOTE-ENN. Hyperparameter tuning using Grid Search showed that XGBoost with SMOTE is the best configuration, achieving an F1-Score of 0.8730 and an AUC of 0.9828, offering an optimal balance between precision and recall. SMOTE significantly improved model performance, while SMOTE-ENN increased recall but reduced the F1-Score for Random Forest due to higher false positives, consistent with literature on overfitting from aggressive data filtering. Feature importance analysis identified *Months_Inactive_12_mon*, *Total_Trans_Amt*, and *Total_Relationship_Count* as key predictors, providing actionable insights for retention strategies.

For future work, it is recommended to test model generalizability on other datasets, apply interpretation methods like SHAP or LIME, and explore ensembling or cost-sensitive learning. Prior to full CRM integration, technical and strategic evaluation is essential to ensure accuracy and alignment with business objectives, maximizing customer retention impact.

REFERENCES

Al-Saqqa, S., Sawalha, S., & Jarrah, M. (2023). Customer Churn Prediction using SMOTE and Ensemble Learning in the Telecommunication Sector. *Journal of Information Systems and Technology Management*, *20*(1), 45–58.

Amalia, A., & Asmunin, A. (2024). Analisis Perbandingan Metode SMOTE-ENN dan SMOTE-Tomek pada Klasifikasi Data Tidak Seimbang. *Jurnal Ilmiah Teknik Informatika dan Komputer (JTIK)*, 5(1), 1–9.

Azmi, F., & Voutama, A. (2024). Prediksi Customer Churn Menggunakan Algoritma Random Forest dan XGBoost (Studi Kasus: Perusahaan Telekomunikasi). *Jurnal Sistem Informasi Bisnis*, *14*(2), 112–120.

Batista, G. E., Prati, R. C., & Monard, M. C. (2021). A Study of the Behavior of Several Oversampling Techniques for Class Imbalance Problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *34*(5), 2200–2212.

Bibi, I., Niu, X., & Iqbal, K. (2024). Improving Churn Prediction Using Random Forest and Optimized Hyperparameters. *Journal of Big Data*, *11*(3), 215–228.

Boozary, A., Ghadimi, N., & Kazemzadeh, R. (2025). Customer Churn Prediction in the Banking Industry Using Machine Learning Algorithms: A Comprehensive Review. *International Journal of Information Management*, *74*, 102717.

Chen, T., & Guestrin, C. (2023). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Deloitte. (2024). Global Banking Industry Outlook 2024: Navigating Customer Retention Challenges. *Deloitte Insights*. Retrieved from https://www.deloitte.com/insights.

García, S., Fernández, A., & Herrera, F. (2024). SMOTE-ENN for Imbalanced Classification: A Critical Review and Comparison with Other Techniques. *Information Sciences*, *251*, 1–19.

Hambali, M. A., & Andrew, A. (2024). Implementasi XGBoost untuk Prediksi Churn Pelanggan dengan Penanganan Data Tidak Seimbang. *Jurnal Informatika Ekonomi Bisnis*, *6*(1), 1–8.

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), 88

Mahmoudzadeh, M. H., & Shirali Shahreza, M. H. (2025). Predicting Bank Customer Churn Using Machine Learning. *Financial Research Journal*, *27*(2), 218–245.

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2021). Classification of Imbalanced Data: A Review. *International Journal of Neural Systems*, *12*(1), 1–15.

Sahare, S., & Gupta, P. (2022). Performance Comparison of Hybrid Resampling Methods For Imbalanced Datasets in Churn Prediction. *International Journal of Data Science and Analytics*, *9*(3), 251–263.

Sari, R., Nugroho, A., & Pratama, M. (2024). Churn Prediction for Banking Customers using SMOTE and XGBoost. *Indonesian Journal of Applied Data Science*, *4*(2), 67–76.

Zhang, H., Yu, J., & Ma, S. (2020). Class Imbalance Learning: A Review. *Neurocomputing*, 398, 427–450.

Zhang, W., Li, J., & Wang, Y. (2023). Customer Churn Prediction in The Banking Industry: A Comparative Study of Machine Learning Models. *Expert Systems with Applications*, *214*, 118939.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2021). Receiver Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation*, *143*(1), 90–92.