Indonesian Journal of Electrical Engineering and Computer Science
Vol. 41, No. 2, February 2026, pp. 655~665
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v41.i2.pp655-665

Stable and accurate customer churn prediction: comparative analysis of eight classification algorithms

Vincent Alexander Haris[1], Muhammad Ilyas Arsyad[1], Nathanael Septhian Adi Nugraha[1], Yasi Dani[1], Maria Artanta Ginting[2]
[1]Department of Computer Science, School of Computer Science, Bina Nusantara University, Bandung Campus, Jakarta, Indonesia
[2]Research Center for Computing, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Article Info

Article history:

Received Nov 11, 2025
Revised Dec 11, 2025
Accepted Jan 11, 2026

Keywords:

Customer churn prediction
Machine learning
Classification algorithm
Evaluation metric
Performance evaluation

ABSTRACT

Predicting customer churn is a challenging problem in many subscription-based industries, though it is considered more cost-effective than acquiring new customers. In this research, customer churn is predicted using a public dataset from an internet service provider, with 72,274 instances and 55% churn rate. The main contribution is to provide a comprehensive comparison of the stability and performance of eight classification algorithms in customer churn prediction using a large-scale public dataset. The research process includes data collection, data preprocessing, feature engineering, and model evaluation. The metrics evaluation presents test accuracy, accuracy gap, precision, recall, F1-Score, and ROC AUC, with stratified K-Fold cross-validation. Since the proportion of churn and non-churn in the dataset is relatively balanced, the F1-score is considered as the primary evaluation metric, as it provides a balanced assessment of precision and recall for both classes. The results show that CatBoost and XGBoost are the most effective models that achieve high F1-scores of 94.97% and 94.92%, respectively.

Corresponding Author:

Maria Artanta Ginting
Research Center for Computing, National Research and Innovation Agency (BRIN)
Indonesia
Email: mari059@brin.go.id

1. INTRODUCTION

In the competitive business environment, a company must be able to retain customers to sustain revenue and market share. Retaining existing customers is generally more economical than acquiring new ones. In business, churn is a number expressing attrition of customers or subscribers. Churn rate is a measure that defines the proportion of individuals or items leaving a group or system in a period of time. A high rate of churn in a company can impact revenue and market position. In the present day, digitalization in industry allows companies to collect customer historical data. This valuable information can be used for one or multiple corporate objectives. A company can analyze consumer patterns and forecast potential churn using data analytics and machine learning [1]–[3]. With the aid of machine learning, we can categorize users based on their chances of churning. The findings can be extremely beneficial for organizations to develop preventative retention interventions. Churn prediction has been studied and implemented in various sectors, such as in banking [4], [5], telecommunications [6]–[10], and E-commerce [11], [12].

In the banking sector, [4] developed a churn prediction model using K-nearest neighbor (KNN), support vector machine (SVM), decision tree (DT), and random forest (RF) on a banking dataset. They reported that RF with oversampling gave the highest precision and accuracy. Other models, such as logistic regression (LR) and Naive Bayes, were also explored by [13] to predict customer churn in banking sector.

The results showed that Naive Bayes provided more reliable probabilistic predictions. A similar line of work was carried out in [14], [15] where the authors compared and assessed which algorithms are the most effective in churn prediction. In the credit card domain, studies in [16]–[18] proposed a more comprehensive churn prediction framework incorporating feature selection and multiple machine learning classifiers.

In the telecommunications sector, trained and evaluated several machine learning models with a gravitational search algorithm for feature selection [19]. A dataset from an Iranian mobile company was used by [20] to predict customer churn by implementing data mining and machine learning classification techniques. Meanwhile, a new approach based on a distance factor was proposed by [8] to classify churn and non-churn customers from the Telecommunication Industry (TCI). Furthermore, predicted customer churn using deep learning and compared the results with traditional machine learning algorithms [21].

In another sector, [22] discussed a customer churn prediction model for B2C E-commerce businesses. Additionally, various machine learning techniques were used by [23] and [24] to predict customer churn based in Brazilian E-commerce dataset. In [25], a process using a hybrid SVM classification approach to forecast E-commerce customer attrition was provided a hybrid recommendation strategy for targeted retention initiatives.

Numerous studies have assessed individual algorithms for churn prediction. However, algorithm performance varies depending on dataset characteristics and feature selection. Therefore, comparative analysis is necessary to identify the most effective algorithm for accurate churn prediction for a certain dataset. This study offers a focused comparison of eight classification algorithms using a churn dataset from an internet service provider. Unlike many benchmark-driven studies, we pay particular attention to the stability of each model and its training–testing performance gap. Highlighting this stability gap is important because it provides a clearer picture of how reliable a model will be in real-world deployment, yet this aspect has rarely been examined in previous churn prediction research.

A conceptual framework diagram is presented in Figure 1. The dataset from an internet service provider, which contains information about customers' service usage, contract attributes, and billing behavior, trains the predictive models. This study provides a comprehensive comparison of eight classification algorithms—both single and ensemble models—using a real-world ISP churn dataset. In a single classifier, LR, and linear SVM were used as linear models, Naïve Bayes as a probabilistic model, and a DT as a single tree-based model. Advanced ensemble methods include bagging (e.g., RF) and boosting (e.g., AdaBoost, XGBoost, and CatBoost). The study offers a novel focus on model stability and the train–test performance gap, providing insights rarely examined in previous churn prediction research. Analysis on discriminative power and model stability will be conducted. Finally, the goal is to give the business clear next steps.

This paper is organized as follows. In the next section, research methodology is described, including data identification, data preprocessing, modeling, and evaluation. The findings are discussed in section 3, including the overall model performance comparison, the analysis of discriminative power and model stability, mitigating overfitting in tree-based models, and the priority of F1-score in churn prediction. Finally, the conclusion is presented in section 4.

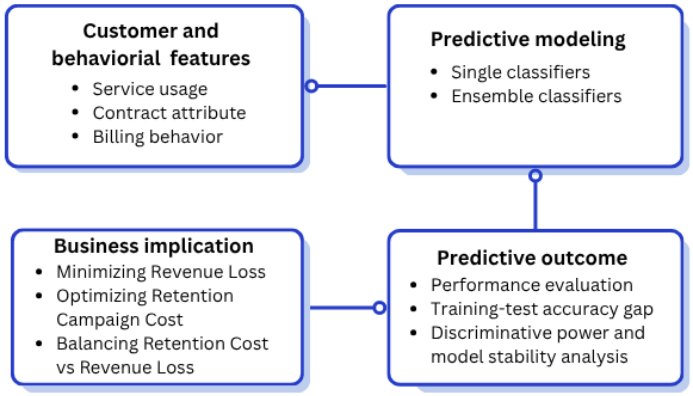

Figure 1. Conceptual framework diagram

2. METHOD

The research process is illustrated in Figure 2 and explained in the following subsections.

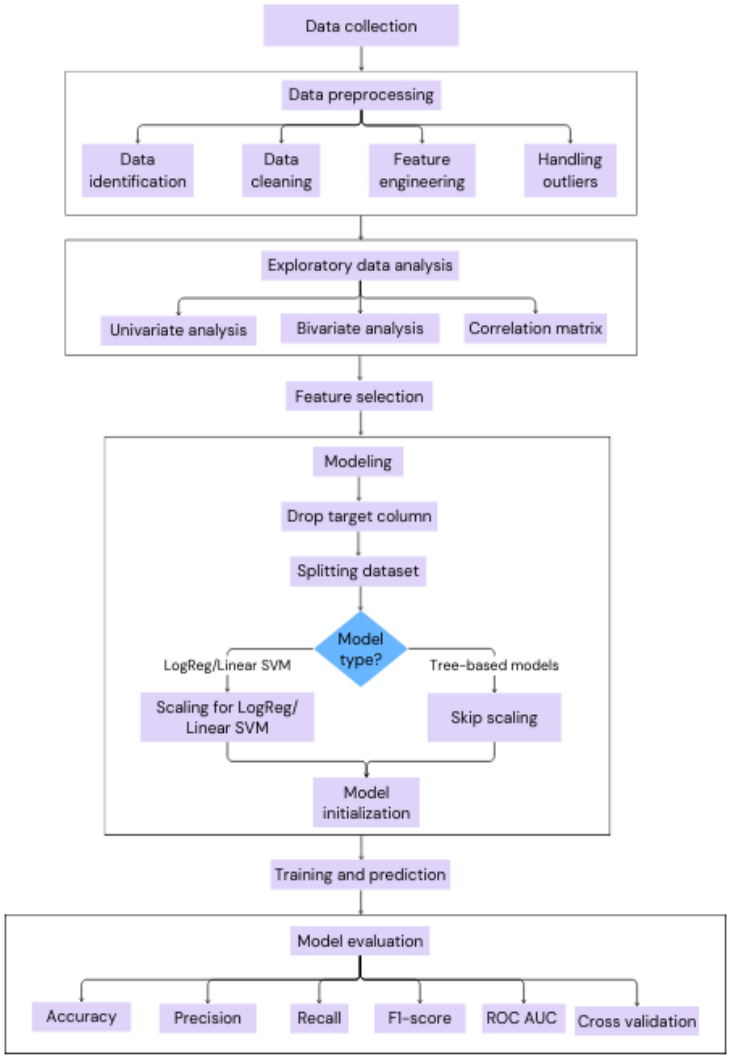

Figure 2. Research flowchart

2.1. Data identification

In this research, a public internet service dataset was used, obtained from the Kaggle platform provided by Kunt (2022). The dataset contains historical customer information from an internet service provider. There are 72,274 rows of customer data with 11 feature columns, all numerical. The dataset is relatively balanced with a churn rate of 55%, which is somewhat unusual since churn cases are typically much fewer than non-churn cases. The feature columns are described in Table 1.

Table 1. Feature columns and their description

Feature columns	Description
Id	Subscriber identifier.
is_tv_subscriber	Binary indicator of whether the customer has a TV subscription.
is_movie_package_subscriber	Binary indicator of whether the customer has a cinema/movie package subscription.
subscription_age	Subscription duration in months.
bill_avg	Average monthly bill over the last three months.
reamining_contract	Remaining contract duration, where null indicates no contract; customers with active contracts must use the service until the contract ends or pay a penalty if terminated early.
service_failure_count	Total customer calls reporting service issues to the call center in the last three months.
download_avg	Average download usage in GB over the last three months.
upload_avg	Average upload usage in GB over the last three months.
download_over_limit	Number of times the customer exceeded the download limit in the last nine months.
churn	A binary indicator where 1 denotes that the customer unsubscribed and 0 denotes that the customer remained subscribed.

2.2. Data preprocessing

This stage consists of data cleaning, dropping irrelevant columns, outlier handling, feature engineering, feature selection, and data scaling. To make sure that the evaluation results are free of data leakage, each step was trained exclusively on the training dataset. Each step was implemented using scikit-learn pipeline that assures the test dataset remains isolated and does not influence the data transformation.

a. Data cleaning
The first step is data cleaning, which identifies missing values in the dataset. The missing values were found in some feature columns, such as *remaining_contract*, *download_avg*, and *upload_avg*. Missing values were handled based on their proportion and potential impact. The *download_avg* and *upload_avg* columns had a very small percentage of missing values, that is 0.53%. The corresponding rows were deleted to maintain data quality without significant information loss. In contrast, there are 29.85% missing values in the *remaining_contract* column. In this case, the missing value was imputed with the value zero since this column represents customers who are not bound by a contract.

b. Dropping irrelevant columns
In this step, the *id* column was removed from the dataset. It contains the customer's unique identifier which is not strongly related to customer churn. This feature does not provide helpful information; otherwise, it may mislead the model during training. This column was deleted so that the dataset focuses on features that matter for customer behaviour, and the model can find meaningful patterns.

c. Outlier handling
Outliers were handled to prevent the model from being skewed by extreme values. Here, we used the Interquartile Range (IQR) method so that the outliers in the data were capped to limit their impact. In the dataset, the extreme values were found in some features, i.e., *bill_avg*, *subscription_age*, *remaining_contract*, *download_avg*, and *upload_avg*.

d. Feature engineering
In this step, a new binary feature was created, namely *is_contract*. This feature was created based on *remaining_contract* column. The value was set to 1 if the value in *remaining_contract* is greater than 0, and 0 otherwise. This transformation highlights whether a customer has a contractual commitment, providing a clear binary signal for the model.

e. Feature selection
The correlation matrix was computed and showed that the *service_failure_count* feature has a very low correlation with the target variable churn (coefficient of 0.02). This feature was considered as noise that has less power on churn prediction. Removing this feature simplifies the model and prevents it from learning spurious patterns.

f. Data scaling
This step is applied only to models that are sensitive to scale, like LR and linear SVM. StandardScaler from Scikit-learn was applied to these models. Other tree-based and ensemble models like DT, RF, AdaBoost, XGBoost, and CatBoost do not require scaling since their internal splitting mechanisms are not affected by differences in feature scales.

2.3. Modeling and evaluation metrics

a. Model selection
There are eight classification algorithms implemented in this comparative study, including probabilistic models (Naïve Bayes), linear models (LR, Linear SVM), single tree-based models (DT), and advanced ensemble methods, such as bagging (RF) and boosting (AdaBoost, XGBoost, CatBoost).

b. Training and testing
The dataset was split into a training set (80%) and a testing set (20%). Stratified K-Fold cross-validation with five folds was implemented to get a more robust evaluation. The stratification method keeps the same imbalance of the churn class in every fold, which would give a more reliable and fair estimation of the model performance.

c. Performance metrics
Model performance was evaluated using a set of standard metrics as follows.

- Accuracy: the ratio of the correctly predicted instances to all instances, formulated by,

$$\text{accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{total instances}}$$

- Precision: the ratio of true positive instances to all predicted positives, formulated by,

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

- Recall: the ratio of true positive instances to all actual positives; the metric is crucial for minimizing false negatives, formulated by,

$$\text{recall} = \frac{\text{true positives}}{\text{true positives+false negatives}}$$

- F1-score: the harmonic average of precision and recall, reflecting a balanced measure between the two metrics, formulated by,

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision+recall}}$$

- ROC: stands for receiver operating characteristic, illustrated by a curve that represents the trade-off between recall and specificity across various thresholds, where x-axis represents false positive rate and y-axis represents true positive rate.
- AUC: area under the curve of ROC that summarize the performance of a classifier across all possible thresholds. The value range is from 0 to 1, where 1 denotes perfect classification and 0.5 corresponds to random guessing.

3. RESULTS AND DISCUSSION

This section presents the performance comparison from the eight classification models, analysis of discriminative power and model stability, description of mitigating overfitting on tree-based models, and discussion about the priority of F1-score in this churn prediction.

3.1. Overall model performance comparison

We evaluated eight classification models on the test dataset. The performance of each model was measured in several metrics: train and test accuracy, accuracy gap, precision, recall, F1-score, and ROC/AUC, as shown in Table 2. The overall results show that models with gradient boosting and ensemble methods performed better than simpler linear and probabilistic models. CatBoost achieved the highest test accuracy of 94.44% and the highest F1-score of 94.97%. In recall, XGBoost performed best with score of 94.16%. Other tree-based models, like the tuned DT and RF, also show good performance with accuracies above 93.9%. The linear SVM and Naïve Bayes models gave decent but lower results.

Table 2. Comprehensive model performance comparison.

Model	Test accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC AUC	Accuracy Gap (%)	Accuracy gap notes
Naive Bayes	92,1	94,44	91,19	92,78	0,944	0,03	Lowest discriminative power
LR	92,52	95,49	90,86	93,12	0,955	0,15	Moderate performance
Linear SVM	92,09	95,44	90,1	92,69	0,954	0,15	Similar to LR
DT	93,95	95,7	93,33	94,5	0,964	0,02	Very stable after tuning
RF	93,9	95,95	93,4	94,46	0,970	0,1	Strong ensemble baseline
AdaBoost	93,46	95,14	93,02	94,07	0,961	0,12	Good but slightly weaker
XGBoost	94,38	95,69	94,16	94,92	0,981	1,01	Best recall, slight overfitting
CatBoost	94,44	95,81	94,14	94,97	0,981	0,5	Best overall balance

In Table 2, accuracy gap is also presented, showing the difference between training and test accuracy. A smaller gap indicates the model is better at generalization with new data. Based on this measure, the tuned DTs show the smallest gap at 0.02%. This result shows that the model generalized very well. In contrast, XGBoost had the largest gap, 1.01%, which suggests it might be slightly overfitting.

Figure 3 and 4 provide a visual comparison of test accuracy, precision, recall, and F1-score for each model. In the left panel of Figure 3, the test accuracy of each model is presented. It shows that CatBoost and XGBoost had the highest accuracy values of 94.44% and 94.38%, respectively. DT and RF followed with scores of 93.97% and 93.8%. In contrast, linear SVM, Naive Bayes, and LR show lower accuracy though it is still above 92%. These results point to the strong generalization capabilities of ensemble-based methods. In the right panel of Figure 3, comparison of precision for each model is shown. As depicted, CatBoost and XGBoost also achieve the highest precision scores of 95.81% and 95.69%. These indicate they have strong capability in producing fewer false positives. Other models like LR, linear SVM, DT, and RF maintain competitive precision scores around 95.5%, while Naive Bayes remains the least precise model with score of 94.44%.

Figure 4 presents the comparison of recall and F1-score of each model. In the left panel, the recall comparison is displayed, highlighting how well each model identifies actual churners. XGBoost has the highest recall score of 94.16%, followed by CatBoost (94.14%) and RF (93.43%). These models exhibit stronger sensitivity than linear classifiers, where LR and linear SVM produced recall scores of 90.86% and 90.10%. Recall is a particularly important metric in most churn prediction problems, since cases where the model fails to detect a customer who is actually about to churn can result in greater losses than cases where the model incorrectly marks a customer who has not churned as churned.

The comparison of F1-score is presented in the right panel of Figure 4. These scores balance the precision and recall values. As depicted, CatBoost again had the highest F1-score of 94.97%, followed closely by XGBoost (94.92%). The consistently superior results provided by the ensemble models emphasize their robustness in dealing with non-linear relationships and complex interactions between features. Meanwhile, Naïve Bayes and linear SVM obtain the lowest F1-scores (92.78% and 92.69%), confirming the earlier observation that the models with simple decision boundaries are not able to capture non-linear patterns as well as ensemble models.

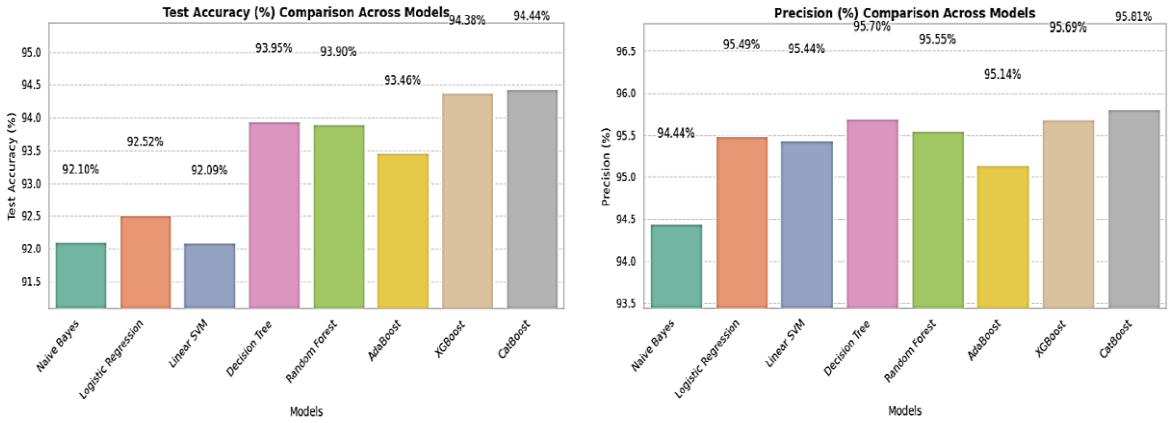

Figure 3. Bar chart comparing the accuracy (left) and precision (right) of each model

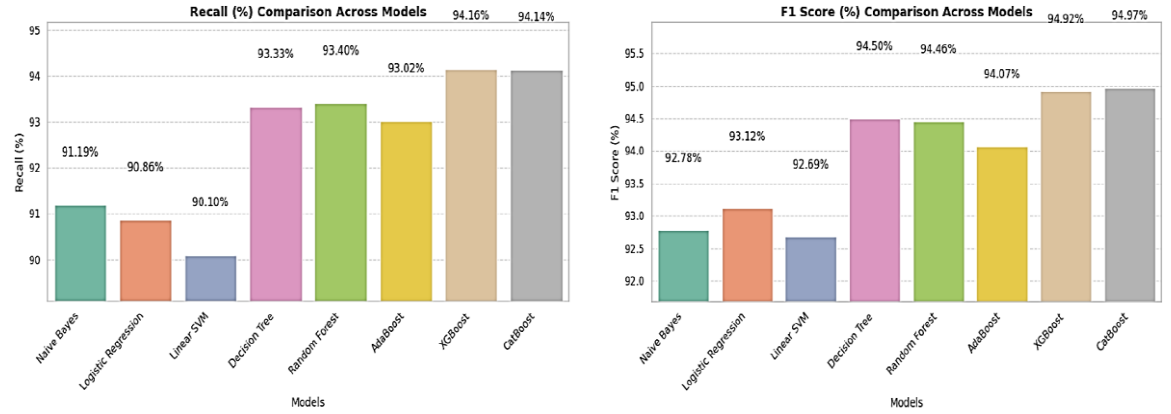

Figure 4. Bar chart comparing the recall (left) and F1-score (right) of each model

3.2. Analysis of discriminative power and model stability

The discriminative power of each model was assessed to evaluate its ability to effectively distinguish churners from non-churners. The discriminative power is analyzed based on the ROC curve and the AUC calculation. From Table 2, it is shown that CatBoost and XGBoost obtained the highest AUC score, 0.9812 and 0.9811 respectively. These indicate the models can clearly distinguish between customers who will churn and those who will not. Notably, these results match the results from other metrics. Tree-based ensemble methods, including RF and AdaBoost, also show strong discriminative power (0.9705 and 0.9612). Linear models such as LR and linear SVM show moderately lower ROC AUC scores (0.9553 and 0.9544),

suggesting a more limited ability to capture complex nonlinear patterns in the data. Naïve Bayes presents the lowest discriminative power with ROC AUC score of 0.9436, which may be due to the independence assumption not being in line with the observed feature relationships in the dataset.

Furthermore, the model stability was evaluated by implementing stratified 5-fold cross-validation on the training dataset. Table 3 presents the mean performance and standard deviation (SD) for each model across the five folds. CatBoost achieves the highest mean ROC AUC of 0.9804 (±0.0005), and XGBoost follows closely with 0.9798 (±0.0006). These cross-validation results support our earlier findings as well. Across all models, the SD remained relatively low, indicating stable performance across all folds, and no model experienced high variance during evaluation. This confirms the reliability of the stratified 5-fold cross-validation procedure in providing robust performance estimates.

Table 3. Stratified 5-fold cross-validation results

Model	Mean accuracy (± SD) (%)	Mean F1-score (± SD) (%)	Mean recall (± SD) (%)	Mean ROC AUC (± SD)
Naïve Bayes	92.09 (± 0.14)	92.79 (± 0.12)	91.39 (± 0.18)	0.9435 (± 0.0006)
LR	92.40 (± 0.13)	93.03 (± 0.13)	91.05 (± 0.28)	0.9554 (± 0.0006)
Linear SVM	91.96 (± 0.12)	92.59 (± 0.12)	90.23 (± 0.26)	0.9544 (± 0.0006)
DT	93.82 (± 0.07)	94.39 (± 0.06)	93.38 (± 0.21)	0.9646 (± 0.0010)
RF	93.75 (± 0.09)	94.34 (± 0.08)	93.51 (± 0.21)	0.9698 (± 0.0010)
AdaBoost	93.37 (± 0.13)	93.99 (± 0.12)	93.21 (± 0.22)	0.9615 (± 0.0015)
XGBoost	94.23 (± 0.06)	94.79 (± 0.06)	94.21 (± 0.22)	0.9798 (± 0.0006)
CatBoost	94.30 (± 0.13)	94.85 (± 0.13)	94.16 (± 0.32)	0.9804 ± 0.0005)

3.3. Mitigating overfitting in tree-based models

In the DT and RF models, we encountered signs of overfitting. To address this, tuning experiments were conducted on a single parameter, *max_depth*. This parameter denotes the maximum depth of the DT. If the tree depth is too large, the model might learn overly specific patterns, including noise. This condition may lead to model overfitting, where performance on training data is very high, but on testing data, it drops drastically. By limiting the maximum depth through parameter tuning, the tree is less likely to memorize noise and still maintain good generalization to new data. This tuning process finds the optimal *max_depth* value that balances model complexity and generalization ability, thereby reducing the risk of overfitting and improving accuracy on the test data.

Table 4 presents the performance evaluation before parameter tuning, indicating overfitting. With default parameters, the accuracy gaps for the DT and the RF are 8.88% and 5.75% respectively. The *max_depth* parameter was tuned by testing over values 1 to 20. As in Figure 5, it was found that *max_depth*=7 gave the best balance between model complexity and generalization. With this setting, the training and testing accuracy curves nearly matched. The accuracy gap reduced to 0.02% for the DT and 0.1% for the RF as presented in Table 2.

Table 4. Evaluation before parameter tuning

Model	Train accuracy (%)	Test accuracy (%)	Accuracy Gap (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
DT	99.96	91.08	8.88	92.23	91.71	91.97	0.909957
RF	99.95	94.2	5.75	95.95	93.53	94.73	0.942858

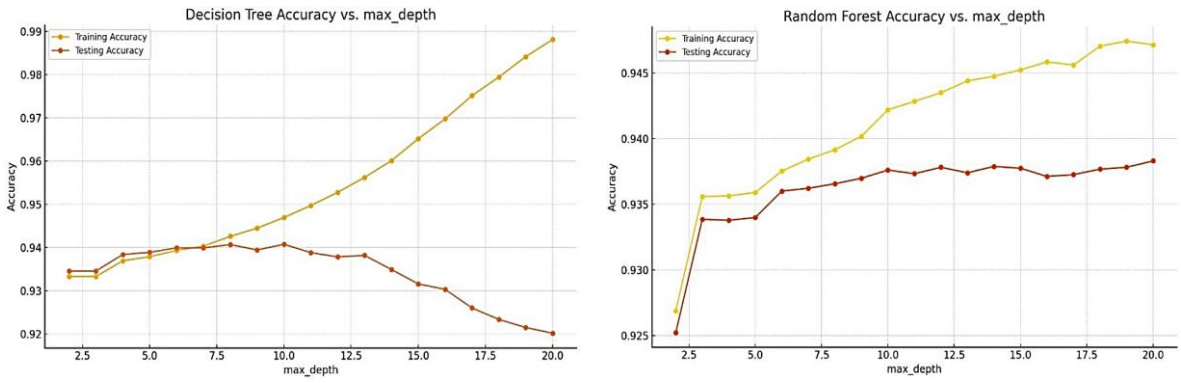

Figure 5. *max_depth* parameter tuning on the DT and the RF

3.4. The priority metric and business implication

In the churn prediction problem, it is common to have an imbalanced dataset where the number of churn customers is much smaller than the non-churn customers. This condition makes recall a critical metric to ensure that as many true churners as possible are correctly identified. However, our study presents a different case, with a churn rate of 55% the dataset is considered to be relatively balanced. Thus, one particularly important metric in this issue is the F1-score as it provides the assessment of the trade-off between precision and recall. A balanced dataset reduces the dominance of majority class, allowing the model to learn both classes more effectively. The F1-score becomes particularly suitable for evaluating predictive performance because it reflects both the cost of incorrectly identifying a customer who will churn (false negatives) and the potential expense of unnecessary retention actions towards loyal customers (false positives). Additionally, supporting metrics like accuracy, precision, recall, ROC-AUC are still examined to provide a comprehensive understanding of model behavior.

As noted in Subsection 3.1., CatBoost and XGBoost have the strongest performances with 94.97 and 94.92 F1 scores, respectively. This indicates that both models perform exceptionally well in maintaining a tradeoff between precision (out of the predicted churners, how many are actually churners) and recall (out of all actual churners, how many are predicted as churners). More specifically, both models are able to really pinpoint true churners, and they do not suffer from a high false positive rate, making them very reliable. This allows the company to make more informed business decisions to implement retention campaigns as well as to determine how to best use their available resources.

Because the dataset used in this study is relatively balanced, the findings also suggest that these models are well-suited to real world customer segments with balanced churn rates. Thus, they become relevant models for numerous subscription-based services, telecommunications providers, financial services, and other sectors where churn is in the moderate range. Also, the stability shown in the cross-validation folds means that the chosen model is consistent and could be confidently deployed in production environments. It would be possible for businesses to use the model results to design customer relationship management systems that support more advanced strategies, such as creating early warning dashboards, automated retention triggers, or campaign strategies based on customer tiers. If this is combined with relevant financial data (i.e., customer lifetime value (CLV), cost of retention offers, revenue at risk) the business would be able to quantify the impact of churn predictions and prioritize efforts on customer segments that would make the greatest financial impact.

4. CONCLUSION

This study successfully demonstrates the effective implementation of machine learning for high-accuracy customer churn prediction. The performances of eight classification algorithms were compared. Gradient boosting models, specifically CatBoost and XGBoost, performed best. Both models did well on metrics like accuracy (94,44% and 94,38%, respectively), F1-score (94,97 and 94,92), and ROC/AUC (0,98121 and 0,98108), indicating strong capability in distinguishing high-risk customers with minimal classification error. Based on the relatively balanced dataset characteristics, the F1-score is the most important metric since it evaluates model performance by combining both precision and recall into a harmonic measure. The quantitative results provide a more comprehensive reflection of how well the model distinguishes between the two classes, while maintaining a trade-off between correctly predicting the churns and avoiding unnecessary misclassifications. The performance consistency between training and testing also indicates model reliability since the model stability has been examined by executing the 5-fold cross-validation. This implies that the model is suitable for real-world deployment, like integration into customer relationship management to generate periodic churn-risk alert and to support more efficient retention strategies.

There are some limitations in this study that could be addressed in future research. First, a more comprehensive hyperparameter optimization could be explored to enhance performance. Second, future research should integrate profit or cost-sensitive analysis to measure the financial benefits of improvements in recall, as well as the economically optimal decision thresholds. Additionally, incorporating explainable AI (XAI) methods would allow practitioners to interpret predictions at both the global and individual levels.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

The specific contributions of each author to this research are outlined in the following table, following the CRediT (Contributor Roles Taxonomy) framework.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Vincent Alexander Haris		✓	✓			✓	✓	✓	✓		✓			
Muhammad Ilyas Arsyad		✓	✓	✓	✓	✓	✓		✓					
Nathanael Septhian Adi Nugraha		✓	✓			✓	✓	✓	✓		✓			
Yasi Dani	✓	✓		✓						✓		✓		
Maria Artanta Ginting	✓	✓		✓	✓		✓		✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The dataset used in this study is publicly available on Kaggle at https://www.kaggle.com/datasets/mehmetsabrikunt/internet-service-churn. The source code of the experiment is available on GitHub https://github.com/rsyd03/churn.git.

REFERENCES

[1] L. Geiler, S. Affeldt, and M. Nadif, "A survey on machine learning methods for churn prediction," *International Journal of Data Science and Analytics*, vol. 14, no. 3, pp. 217–242, Sep. 2022, doi: 10.1007/s41060-022-00312-5.

[2] A. Manzoor, M. A. Qureshi, E. Kidney, and L. Longo, "A review on machine learning methods for customer churn prediction and recommendations for business practitioners," *IEEE Access*, vol. 12, pp. 70434–70463, 2024, doi: 10.1109/ACCESS.2024.3402092.

[3] M. Imani, M. Joudaki, A. Beikmohammadi, and H. Arabnia, "Customer churn prediction: a systematic review of recent advances, trends, and challenges in machine learning and deep learning," *Machine Learning and Knowledge Extraction*, vol. 7, no. 3, p. 105, Sep. 2025, doi: 10.3390/make7030105.

[4] M. Bogaert and L. Delaere, "Ensemble methods in customer churn prediction: a comparative analysis of the state-of-the-art," *Mathematics*, vol. 11, no. 5, p. 1137, Feb. 2023, doi: 10.3390/math11051137.

[5] M. Rahman and V. Kumar, "Machine learning based customer churn prediction in banking," in *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, Nov. 2020, pp. 1196–1201, doi: 10.1109/ICECA49313.2020.9297529.

[6] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," *8th International Conference on Digital Information Management, ICDIM 2013*, pp. 131–136, 2013, doi: 10.1109/ICDIM.2013.6693977.

[7] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, Jun. 2015, doi: 10.1016/j.simpat.2015.03.003.

[8] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in the telecommunication industry using data certainty," *Journal of Business Research*, vol. 94, pp. 290–301, Jan. 2019, doi: 10.1016/j.jbusres.2018.03.003.

[9] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data*, vol. 6, no. 1, p. 28, Dec. 2019, doi: 10.1186/s40537-019-0191-6.

[10] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[11] M. Pondel *et al.*, "Deep learning for customer churn prediction in e-commerce decision support," *Business Information Systems*, vol. 1, no. July, pp. 3–12, 2021, doi: 10.52825/bis.v1i.42.

[12] Z. Wu, L. Jing, B. Wu, and L. Jin, "A PCA-AdaBoost model for E-commerce customer churn prediction," *Annals of Operations Research*, vol. 350, no. 2, pp. 537–554, Jul. 2025, doi: 10.1007/s10479-022-04526-5.

[13] V. Agarwal, S. Taware, S. A. Yadav, D. Gangodkar, A. L. N. Rao, and V. K. Srivastav, "Customer - Churn prediction using machine learning," in *Proceedings of International Conference on Technological Advancements in Computational Sciences, ICTACS 2022*, Oct. 2022, pp. 893–899, doi: 10.1109/ICTACS56270.2022.9988187.

[14] H. Tran, N. Le, and V. H. Nguyen, "Customer churn prediction in the banking sector using machine learning-based classification models," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 18, pp. 87–105, 2023, doi: 10.28945/5086.

[15] I. Kaur and J. Kaur, "Customer churn analysis and prediction in banking industry using machine learning," in *PDGC 2020 - 2020 6th International Conference on Parallel, Distributed and Grid Computing*, Nov. 2020, pp. 434–437, doi: 10.1109/PDGC50313.2020.9315761.
[16] D. AL-Najjar, N. Al-Rousan, and H. AL-Najjar, "Machine learning to develop credit card customer churn prediction," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 4, pp. 1529–1542, Nov. 2022, doi: 10.3390/jtaer17040077.
[17] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15273–15285, Nov. 2011, doi: 10.1016/j.eswa.2011.06.028.
[18] R. Rajamohamed and J. Manokaran, "Improved credit card churn prediction based on rough clustering and supervised learning techniques," *Cluster Computing*, vol. 21, no. 1, pp. 65–77, Mar. 2018, doi: 10.1007/s10586-017-0933-1.
[19] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, no. 2, pp. 271–294, 2022, doi: 10.1007/s00607-021-00908-y.
[20] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, "Improved churn prediction in telecommunication industry using data mining techniques," *Applied Soft Computing*, vol. 24, pp. 994–1012, Nov. 2014, doi: 10.1016/j.asoc.2014.08.041.
[21] S. Saha, C. Saha, M. M. Haque, M. G. R. Alam, and A. Talukder, "ChurnNet: deep learning enhanced customer churn prediction in telecommunication industry," *IEEE Access*, vol. 12, no. 1, pp. 4471–4484, 2024, doi: 10.1109/ACCESS.2024.3349950.
[22] X. Xiahou and Y. Harada, "B2C E-commerce customer churn prediction based on K-means and SVM," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 2, pp. 458–475, Apr. 2022, doi: 10.3390/jtaer17020024.
[23] S. Baghla and G. Gupta, "Performance evaluation of various classification techniques for customer churn prediction in E-commerce," *Microprocessors and Microsystems*, vol. 94, p. 104680, Oct. 2022, doi: 10.1016/j.micpro.2022.104680.
[24] K. Matuszelański and K. Kopczewska, "Customer churn in retail E-commerce business: spatial and machine learning approach," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 1, pp. 165–198, Jan. 2022, doi: 10.3390/jtaer17010009.
[25] S. J, C. Gangadhar, R. K. Arora, P. N. Renjith, J. Bamini, and Y. devidas Chincholkar, "E-commerce customer churn prevention using machine learning-based business intelligence strategy," *Measurement: Sensors*, vol. 27, p. 100728, Jun. 2023, doi: 10.1016/j.measen.2023.100728.

BIOGRAPHIES OF AUTHORS

Vincent Alexander Haris is was born in Bandung City on December 2, 2003. He completed his bachelor's degree in Computer Science at Bina Nusantara University in 2025. He interned at Xylo Solusi Indonesia from 2024 to 2025. He can be contacted at email: vincent.haris@binus.ac.id.

Muhammad Ilyas Arsyad is was born in Serang City on August 3, 2002. He completed his bachelor's degree in Computer Science at Bina Nusantara University in 2025. He interned as a web designer at Telkom Indonesia from 2024 to 2025. He can be contacted at email: muhammad.arsyad006@binus.ac.id.

Nathanael Septhian Adi Nugraha is was born in Bandung City on September 17, 2003. He completed his Bachelor's degree in Computer Science at Bina Nusantara University in 2025. He interned as a programmer at Mixtra Inti Tekindo from 2024 to 2025. He can be contacted at email: nathanael.nugraha@binus.ac.id.

Yasi Dani received her bachelor's, master's, and doctoral's degrees in Mathematics from Institut Teknologi Bandung, Bandung, Indonesia, in 2010, 2015, and 2025 respectively. Now, she joins as a lecturer in Computer Sciences at University of Bina Nusantara. Her research interests include outlier or anomaly detection, machine learning, applied mathematics and statistics. She can be contacted at email: yasi.dani@binus.ac.id.

Maria Artanta Ginting received her undergraduate, graduate, and doctoral studies in mathematics from Bandung Institute of Technology in Indonesia. Currently, she is a researcher in Research Center for Computing, National Research and Innovation Agency (BRIN), Indonesia. Her research focuses on applied and computational mathematics. She is also interested in exploring the applications of machine learning methods to data-driven problem solving. She can be contacted at email: mari059@brin.go.id.