CUSTOMER CHURN PREDICTION IN MOBILE BANKING APPLICATION BASED ON RANDOM FOREST

DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ ỨNG DỤNG NGÂN HÀNG DI ĐỘNG DỰA TRÊN MÔ HÌNH RỪNG NGẪU NHIÊN

Ngày nhận bài: 29/09/2025

Ngày nhận bản sửa: 27/11/2025

Ngày chấp nhận đăng: 13/01/2026

Le Dinh Hac, Nguyen Hoang Chung, Nguyen Thi Hang✉

ABSTRACT

The study was conducted to evaluate the performance of Random Forest in predicting customers' tendency to abandon mobile banking apps and to explore the input features that contribute significantly to identifying customers' intention to abandon. Qualitative, quantitative, and bibliometric methods were used in the study. The study used simulated data to train the Random Forest model with 5 input features, including (1) Login Frequency, (2) Balance Checks, (3) Transfer Transaction, (4) Online Savings, and (5) Bill Payments. Additionally, the study compared the performance of Random Forest with that of other supervised machine learning models, including Gradient Boosting, Logistic Regression, and SVM. The results show that Random Forest achieved the highest predictive performance, with up to 99.5% accuracy. Two characteristics were identified as strong indicators: "Login Frequency" and "Balance Checks." From the research results, the application of supervised machine learning models in early identification of customers who are likely to leave the bank and perform periodic identification to ensure accuracy when customers change their consumer behavior, and business strategies need to prioritize focusing on customer groups that tend to leave, in addition, banks need to strengthen cooperation with businesses to create incentives to stimulate user demand.

Keywords: Mobile Banking Application; Predicting Customer Churn; Random Forest; Supervised Machine Learning.

TÓM TẮT

Nghiên cứu được thực hiện nhằm đánh giá hiệu suất của mô hình Rừng ngẫu nhiên trong việc dự đoán xu hướng khách hàng rời bỏ ứng dụng ngân hàng số, đồng thời khám phá các đặc trưng đầu vào có đóng góp quan trọng trong việc nhận diện ý định rời bỏ của khách hàng. Phương pháp định tính, định lượng, và phân tích trắc lượng thư mục được sử dụng trong nghiên cứu. Dữ liệu mô phỏng được sử dụng để huấn huyện mô hình Random Forest với 5 đặc trưng đầu vào bao gồm (1) Tần suất đăng nhập, (2) Truy vấn số dư, (3) Giao dịch chuyển khoản, (4) Gửi tiết kiệm online, (5) Thanh toán hóa đơn định kỳ. Ngoài ra, nghiên cứu còn so sánh hiệu suất của Random Forest với các mô hình học máy có giám sát khác, bao gồm Gradient Boosting, Logistic Regression và SVM. Kết quả cho thấy Random Forest đạt hiệu suất dự đoán cao nhất, với độ chính xác lên đến 99,5%. Hai đặc trưng là chỉ báo mạnh gồm "Tần suất đăng nhập" và "Truy vấn số dư". Từ kết quả nghiên cứu, tác giả đề xuất một số giải pháp nhằm nâng cao hiệu quả quản lý và duy trì sự gắn bó của khách hàng với ứng dụng ngân hàng số trong bối cảnh kinh doanh trên nền tảng công nghệ số phát triển mạnh mẽ.

Từ khóa: Ứng dụng ngân hàng di động; Dự đoán khách hàng rời bỏ; Rừng ngẫu nhiên; Học máy có giám sát.

1. Introduction

Faced with the strong global digital transformation, all sectors of the economy are shifting from traditional to digital business, leveraging digital technology platforms across production, distribution, marketing, and sales. In the United Nations' sustainable development goals, digitalization plays an extremely important role. If businesses do not carry out digital transformation, economic and environmental challenges will be difficult to solve (Bican and Brem, 2020). Not only businesses but also banks are quickly implementing digital transformation and investing heavily in technology to improve their operations. One of the most obvious improvements of banks is the modernization of services through the development of mobile banking applications that gradually replace traditional banking services. With just a phone and an Internet connection, users can use the Mobile Banking Application to perform financial transactions such as money transfers, bill payments, balance inquiries, online savings deposits, and so on. In fact, when the Mobile Banking Application is deployed, it enhances customer experience, helping banks reduce operating costs, save time, and human resources. In addition, the increasing trend toward cashless payments has driven the increasing use of mobile banking applications. According to the forecast of Market.us Scoop (2025), by the end of the year, there will be 3.8 billion users globally using mobile banking applications. Although the number of digital banking users is increasing rapidly, banks are also facing the challenge of users abandoning the application after a period of use. Several customers register but do not activate their accounts, or some customers have used the application but have not logged in for a period of time. According to statistics from Longe (2024), the abandonment rate is 75% after the first day of use, 89.3% after 1 week, 94.4% after 1 month, and 71% after 3 months. This not only affects the revenue and efficiency of exploiting bank customers but also poses a risk of cybercriminals taking over these inactive accounts to commit financial fraud. Given this situation, predicting customers' abandonment behavior to develop a customer retention strategy has become an urgent requirement for banks. In the modern technological environment, the traditional survey method to grasp the psychology and needs of customers seems to be no longer effective because the bank is in a passive position. A feasible solution to this problem is to build a customer behavior prediction system based on user behavior history. In the context of the world economy rapidly shifting towards digital business platforms, user history data is considered a strategic asset for all businesses. When this data is used effectively, it will become a key factor in helping businesses improve their competitive advantage. Banks can rely on customers' behavioral history when using the Mobile Banking App, use supervised machine learning to predict future customer behavior, and identify early those likely to leave, thereby helping banks proactively approach and implement reasonable retention policies to retain existing customers and attract potential customers. Among supervised

Le Dinh Hac, Nguyen Hoang Chung, Ho Chi Minh University of Banking
Nguyen Thi Hang, Thu Dau Mot University
✉Email: nguyenthihang@tdmu.edu.vn

machine learning algorithms, Random Forest is proposed for its high applicability, fast processing of large datasets, and no requirement for in-depth technical knowledge (Zakariah, 2014). A study by Aburbeian and Ashqar (2023) demonstrates that the enhanced Random Forest model delivers outstanding performance in detecting credit card fraud under highly imbalanced data conditions. This firms the model's effectiveness and reliability in high-risk financial environments. Furthermore, Random Forest possesses a notable strength in identifying the most influential predictive features (Aysan et al., 2024), which is especially valuable in forecasting tasks where highlighting high-impact variables helps clarify the key drivers of risk. Hence, the study expects to successfully build a Random Forest-based model to predict the likelihood of abandoning mobile banking apps, thereby suggesting solutions to improve the user experience of mobile banking, helping to increase the rate of customers sticking with mobile banking apps, and contributing to promoting business activities on the bank's technology platform to develop strongly and sustainably.

2. Theoretical Backgrounds and Methodology

2.1. Theoretical backgrounds

2.1.1. Predicting Customer Churn

Predicting Customer Churn involves using technical methods to estimate the likelihood of leaving the service in the near future (Lalwani et al., 2022). Based on the customer's service usage behavior, predicting the likelihood of customers sticking with or leaving the business is completely possible. Traditionally, businesses often conduct surveys to assess customer satisfaction and identify customer needs. On that basis, businesses identify which customers are likely to leave. With the development of technology, the survey method has evolved from paper surveys to Google Forms to surveys within the apps customers use. However, this method does not really bring accurate results because not all customers are willing to take the survey. Some customers do it reluctantly, some feel it is a waste of time, refuse, and some just randomly select without reading the content. Therefore, understanding customer needs and classifying customers according to service usage needs has not achieved high accuracy.

Currently, artificial intelligence is widely used by businesses in their operations as well as their business strategies, including predicting user behavior and thereby developing appropriate strategies to improve customer experience. In the field of information technology, factors such as transmission speed and payment methods strongly affect users' behavior of leaving and returning to use the service (Phua et al., 2012). Cheng et al. (2019) found that customers abandon bank credit cards in Taiwan due to ineffective marketing campaigns; signs of customer churn include reduced consumption or no transactions during the period. According to Wang et al. (2022), information needs influence people's churn behavior. When users are not provided with clear, complete service information, their likelihood of churning is quite high. Not only is the group of loyal customers important to the development of the business, but the group of customers who are likely to churn is also extremely important because it takes a lot of time and money for businesses to find customers. Hence, they need to retain customers to exploit the potential and expand the customer base, thereby expanding the scale of the business. Therefore, predicting customer behavior to identify the risk of customer churn and develop response plans is extremely important.

2.1.2. Random Forest

Random Forest is a supervised machine learning algorithm used for object prediction and classification. The Random Subspace Method was first studied by Ho (1995). In Ho (1995)'s study, many submodels were randomly built from a set of features, ensuring high performance while avoiding overfitting. Breiman (2001) had built a complete Random Forest model by adding important techniques. The outstanding feature of Breiman's Random Forest model compared to Ho is the combination of two sources of randomness in data and randomness in features, and at the same time, evaluating the importance of features, thereby the accuracy in object recognition and classification is much higher than the original version. The Random Forest model is described in an intuitive and easy-to-understand image in Zhang (2022)'s study:

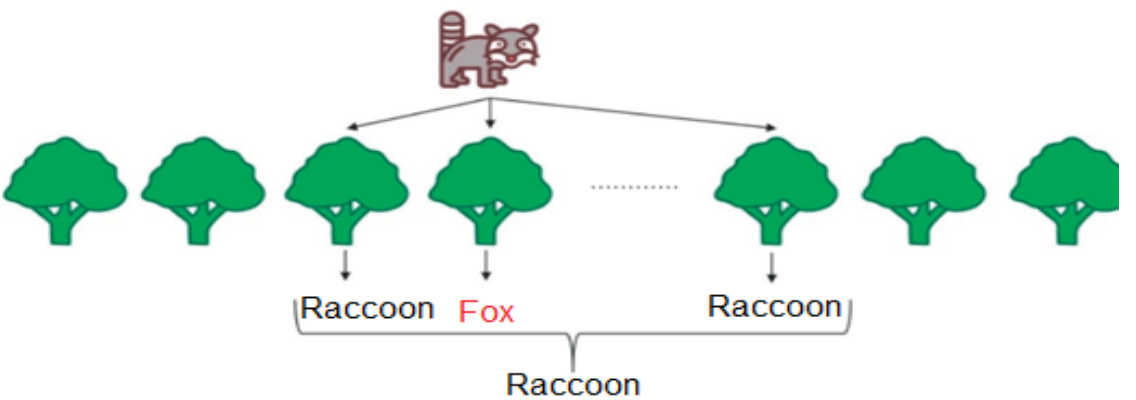

Figure 1. Random forest model

In Zhang's study, Random Forest is defined as "Random Forest is a combined classifier method" and illustrated in Figure 1 as the classification between Raccoon and Fox in a mixed group.

Recently, there have been quite a few studies using Random Forest for classification problems in finance and banking, such as Makariou et al. (2021) predicting market bond price differences, Carcillo et al. (2021) detecting credit card fraud, Al-Najjar et al. (2022) predicting credit card delinquency, Han et al. (2024) predicting credit risk of small and medium enterprises, and Aysan et al. (2024) analyzing risks in banking operations. In addition, Random Forest has also been studied for application in other fields, such as Magidi et al. (2021) classifying irrigated land in agriculture, Balla et al. (2021) predicting productivity of garment workers, Alariyibi et al. (2023) predicting the risk of heart disease in the medical field, Balabied et al. (2023) predicting student learning outcomes in education, and Deng (2025) predicting future tourist traffic. All studies show that Random Forest has high predictive performance, accurate results, and practical application significance. Studies recommend applying Random Forest in practice for object classification, predicting future trends, and thereby building effective response strategies.

2.2. Methodology

2.2.1. Bibliometric approach to identifying research trends

In the context of global digitalization, predicting the likelihood of customers abandoning banking services, especially the Mobile Banking App, is increasingly attracting the attention of researchers and financial institutions.

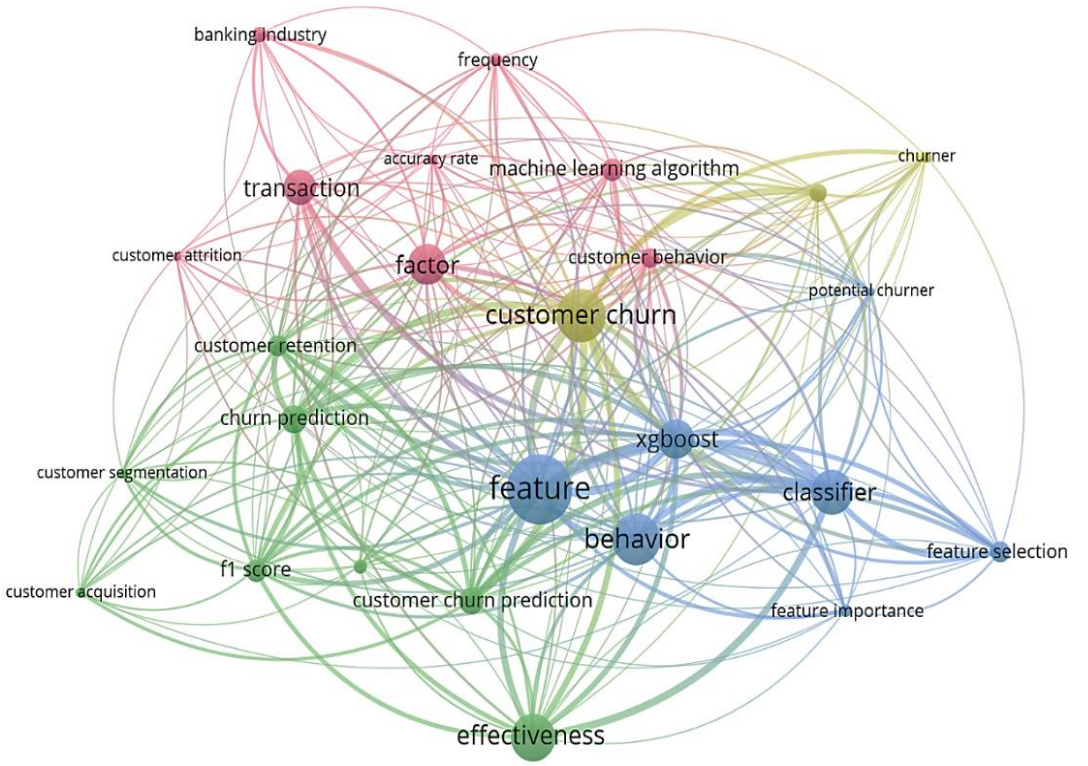

Figure 2. Bibliometric map of keywords associated with customer churn prediction

Based on the keyword co-occurrence map generated using VOSviewer in Figure 2, it is evident that the term "Customer Churn" stands out as a central research focus. It is strongly associated with other key terms such as "feature", "behavior", "effectiveness", and "transaction".

Cluster 1 serves as the central theme, characterized by the presence of key terms such as "Customer Churn," "Churn Prediction Model," "Customer Behavior," and "Machine Learning Algorithm". Studies in this cluster primarily explore the relationship between user behavior and the likelihood of service discontinuation, aiming to develop predictive models based on behavioral characteristics.

Cluster 2 focuses on transactional behavior data and the factors that influence customer churn. Prominent keywords in this cluster include "Transaction", "Customer Behavior", "Factor", and "Frequency". Research in this group often utilizes historical customer transaction data, particularly analyzing usage frequency, account fluctuations, and levels of interaction with mobile banking applications to detect early warning signs of churn.

Cluster 3 emerges as the most prominent, comprising keywords related to model development and machine learning algorithms, such as "Feature", "Classifier", and "Behavior". Within this cluster, the keyword "Feature" plays a central role, highlighting the importance of selecting relevant variables for predictive modeling. Studies in this group primarily focus on identifying key features that influence customer churn and integrating them into classification models.

The final cluster includes terms such as "Effectiveness", "Customer Retention", "Customer Satisfaction", and "Churn Prediction". Research in this area tends to emphasize the strategic application of churn prediction to improve customer relationship management and enhance user experience in digital banking to increase customer loyalty.

Based on the keyword structure, "Feature"—a core term in the third and most dominant cluster—emerges as the most significant node in the co-occurrence network. This reflects a prevailing research trend in the field of churn prediction: a strong focus on analyzing customer behavioral features.

2.2.2. Data simulation based on user behavior characteristics

The study employed qualitative methods to review previous studies and identify user behavior characteristics, from which simulated data were built for use in the Random Forest training process. This approach ensured compliance with legal and ethical requirements regarding AI usage and customer data privacy. Breiman (2001) emphasized that simulated data, when retaining the statistical properties and decision margins of the original phenomenon, is widely accepted for developing and evaluating ensemble algorithms such as Random Forest. Similarly, Meldrum et al. (2025) asserted that when real data are limited by privacy concerns, simulated data can be generated based on theoretical rules or empirical evidence to reflect real-world features and relationships. Therefore, this study constructs a simulated dataset based on the principle of empirically grounded, rule-based simulation, in which each behavioral feature is simulated according to a reasonable statistical distribution, and adds random noise to reflect the variability and imperfection of real data, ensuring that the Random Forest model is trained on both reasonable and diverse data, reflecting the actual behavioral trends of mobile banking users. Building on this approach, the simulated dataset incorporates behavioral features that have been shown in previous studies to play a crucial role for predicting users' intention to abandon mobile banking services.

According to recent studies, historical user behavior characteristics play an important role in predicting users' intention to abandon the service, including login frequency, number of balance checks, money transfer transaction frequency, online savings service usage, and recurring bill payment. These characteristics directly reflect the level of user interaction and attachment to the application, specifically:

(1) Login Frequency: Login frequency is a basic characteristic that shows the attachment or abandonment of using the Mobile Banking App. If customers frequently log in to the app, it indicates high demand; conversely, if login frequency is low or rare, it is a sign that customers will abandon the service. This characteristic is found in studies by Kaya et al. (2018), Mitchell (2020), Adekunle et al. (2023), Hasan et al. (2024), and Boozary et al. (2025).

(2) Balance Checks: The number of times the account balance is checked shows the customer's interest and control over spending. If the frequency of checking the balance gradually or suddenly decreases, it suggests a tendency to abandon the use of the app. Research by Zhan (2024), Boozary et al. (2025) has shown this.

(3) Transfer Transactions: The frequency of transfers shows the customer's need to use the app. The high level of transfers shows that the customer is attached to the app and vice versa, according to research by Zhan (2024), Hasan et al. (2024), and Adekunle et al. (2023)

(4) Online Savings: Online savings is a feature that can clearly show whether

customers are attached or intend to leave the app because the characteristics of this service are the deposit term and the interest that customers receive. If customers use this service regularly, the probability of customers leaving the app is very low (Isson, 2018; Zhan, 2024; Boozary et al., 2025).

(5) Bill Payments: Paying bills such as electricity and water periodically is an essential service of all banks. Bill payments have a fixed cycle, usually monthly or quarterly. If the payment frequency suddenly decreases or if the user no longer makes payments, it is a clear sign of leaving the app, according to the research of authors Hasan et al. (2024) and Boozary et al. (2025).

The number of samples used in the study was determined according to the proposal by Silvey and Liu (2024). The appropriate number of training samples for the classification model was 200 per feature. Five user behavior features were used as input factors, yielding a total of 1,000 samples.

2.2.3. Model construction using the Random Forest algorithm

The quantitative method was used to build a model to predict the likelihood of abandoning mobile Banking apps based on the Random Forest algorithm. The Python programming language and the Scikit-learn library were used to train the model on the Google Colab platform. The Random Forest model was built in the following order:

(1) Divide the data set into a Training Data set of 80% and a Test Data set of 20%

(2) Set parameters

(3) Train "h" sub-models on Training Data set

(4) Vote "h" sub-models that have been trained into the Random Forest model

(5) Use Random Forest to predict on the Test Data set

(6) Evaluate the performance of Random Forest through the indicators Accuracy, Precision, Recall, and F1-Score.

Inheriting the research of Breiman (2001), the research model was proposed as follows:

Subsidiary model:

$h_k(X) = h(X,\Theta k)$

In which:

X: the input feature vector

Θk: a random parameter

Ensemble model:

$$\hat{y}(X) = argmax_c(\sum_{k=1}^{t} I(h_k(X) = c))$$

In which:

I(·): indicator function

c ∈{0,1}: variable representing label (0: churn, 1: Not-Churn)

3. Results and Discussion

3.1. Results

Ten subtrees were trained on the training dataset using the features at the root node, and two features were randomly selected for training to ensure objectivity and diversity of the models.

Table 1 shows that, although randomly selected, it can be seen that the 3 features, Login Frequency, Balance Checks, and Bill Payments, are prioritized to be the features at the root node. At the branch nodes, all 5 features are used evenly in 10 models. The depth of all trees = 5 is enough to separate the samples without overfitting. Leaf nodes are the final separation level of each tree, for models with depth = 5, these indices are suitable.

Table 1. Summary of 10 Sub-Models

Sub-Model	Root Feature	Leaf Nodes	Tree Depth	Used Features
Tree 1	Login_Frequency	11	5	Login_Frequency, Bill_Payments
Tree 2	Balance_Checks	21	5	Transfers, Balance_Checks
Tree 3	Balance_Checks	17	5	Balance_Checks, Bill_Payments
Tree 4	Login_Frequency	11	5	Login_Frequency, Bill_Payments
Tree 5	Bill_Payments	19	5	Bill_Payments, Online_Savings
Tree 6	Login_Frequency	11	5	Login_Frequency, Bill_Payments
Tree 7	Login_Frequency	12	5	Login_Frequency, Balance_Checks
Tree 8	Login_Frequency	12	5	Balance_Checks, Login_Frequency
Tree 9	Balance_Checks	17	5	Bill_Payments, Balance_Checks
Tree 10	Bill_Payments	19	5	Bill_Payments, Online_Savings

Figure 3 illustrates the working mechanism of the Random Forest model. Ten submodels are trained independently on different training datasets, and then they are combined to produce the final model used to predict on the test dataset. This mechanism helps the model exploit the strength of many sub-trees, thereby improving accuracy in predicting objects. The performance of the model will be evaluated based on the prediction results on the Test Data set. Figure 3 illustrates the working mechanism of the Random Forest model. Ten submodels are trained independently on different training datasets, and then they are combined to produce the final model used to predict on the test dataset. This mechanism helps the model exploit the strength of many sub-trees, thereby improving accuracy in predicting objects. The performance of the model will be evaluated based on the prediction results on the Test Data set.

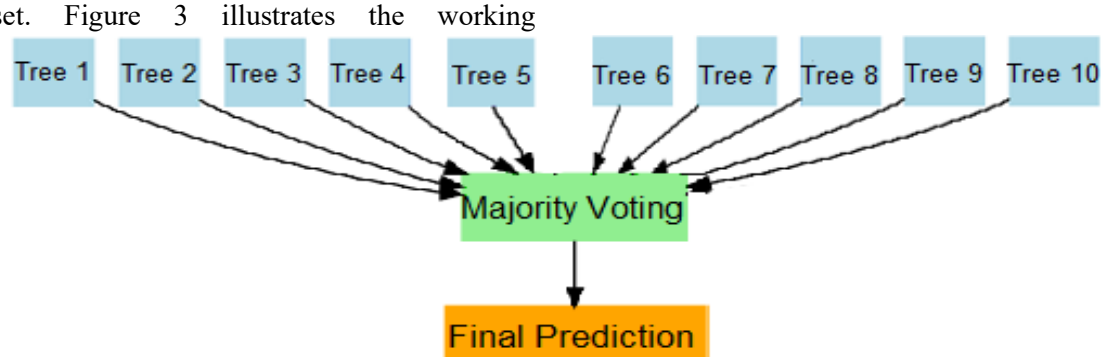

Figure 3. Random Forest Ensemble Structure with 10 Decision Trees

Figure 4 shows the level of error of the Random Forest model when predicting on the Test Data set. According to the actual data, there are 84 Not-Churn customers and 116 Churn customers. The matrix shows that the model correctly predicted 84 Not-Churn

customers and 115 Churn customers. That is, in reality, there are 116 Churn customers, but the model predicted 115 Churn customers; 1 Churn customer was incorrectly predicted as Not-Churn. But this level of error prediction is insignificant compared to the overall.

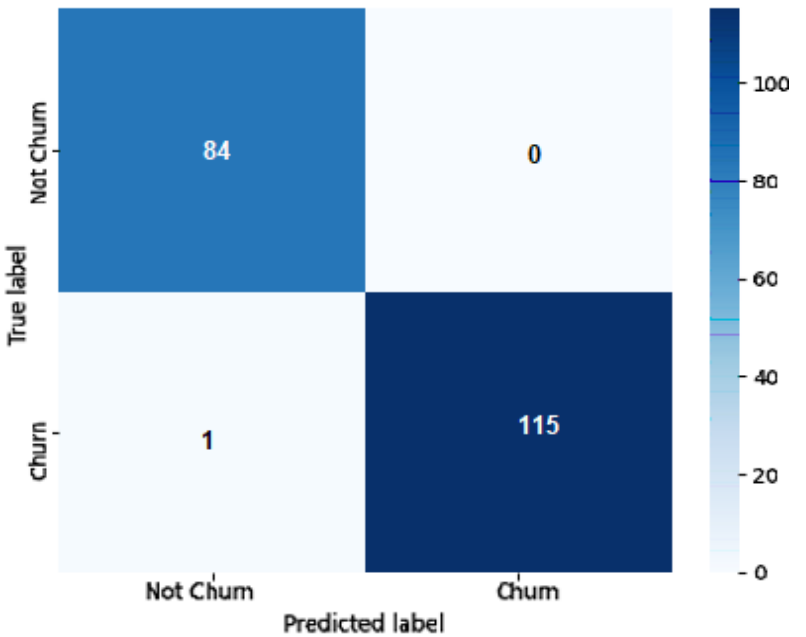

Figure 4. Confusion Matrix

Table 2 shows the model's prediction performance for each label. Specifically, the Not-Churn label has 84 actual samples, and Recall = 1.00 indicates that the model correctly predicts 84 samples as Not-Churn. However, Precision = 0.99 because there is 1 Churn sample, but the model mistakenly predicts it as Not-Churn. The F1-score of 0.99 indicates that the Not-Churn group prediction performance is very high. For the Churn label with 116 actual samples, the Recall is 0.99, as the model correctly predicts 115 samples as churn. Precision = 1 means that all 115 samples predicted as churn are actually in the Churn group; an F1-score of 1 indicates that the model is highly effective at predicting the Churn group.

Table 2. Classification Report

	precision	**recall**	**f1-Score**	**support**
Not Churn	0.99	1.00	0.99	84
Churn	1.00	0.99	1.00	116

Table 3 shows the comparative predictive performance of four supervised machine learning models: Random Forest, Gradient Boosting, Logistic Regression, and SVM. All Precision, Recall, and F1-Score are above 95% for all models, demonstrating good classification performance across the board. Random Forest and Gradient Boosting achieve near-optimal predictive performance, while Logistic Regression and SVM show slightly lower metrics.

Table 3. Comparative Performance Metrics of Machine Learning Models

	Model	Accuracy (%)	Recall (%)	F1 Score (%)
0	Random Forest	99.50	99.14	99.57
1	Logistic Regression	97.52	96.47	97.18
2	SVM	96.73	95.88	96.40
3	Gradient Boosting	99.03	98.72	98.91

3.2. Discussion

From Table 1, it can be seen that during training, each of the 10 trees randomly selected a subset of two features from the five available and chose the feature that provided the best split as the root node. As a result, no feature was explicitly prioritized, but inherently stronger features tended to appear more frequently at the root. This approach follows the principle described by Ho (1995), in which the algorithm introduces inherent randomization by selecting a subset of features as a subspace for each tree and then determining the optimal split within this subset. Consistent with this mechanism, three features, including Login Frequency, Balance Checks, and Bill Payment, appeared at the root node. This shows that these are factors that have a significant impact on the ability to identify churn customers. Among them, Login Frequency was selected as the Root feature 5 times, showing the strongest impact. This finding is consistent with Monetizely (2025)'s research, which states that a decrease in login frequency often occurs several weeks to several months before customers churn the service. The Balance Checks feature, which appears 3 times in the root node, is the Second-most-influential feature for detecting customer churn. From this result, it can be concluded that the Login Frequency and Balance Checks features are two strong predictors to identify the possibility of customer churn.

In Table 1, considering the structure of 10 sub-models trained with depth = 5, the maximum separation level in each tree is 25 = 32; if exceeding this threshold, the model is prone to overfit. The training results show that the final separation level of each tree is in the range of 11-21, which is appropriate and ensures a good balance between the learning ability of the tree, both avoiding overfitting and ensuring the capture of the trend of the data. With 800 samples with 5 features used for training, the tree depth = 5 and the separation level does not exceed 32, showing a reasonable model structure, not too simple nor too complex, so the synthesized model voted from 10 sub-models will achieve high performance and reliable prediction results (Figure 3).

The Random Forest model is synthesized from 10 trained sub-trees, so there is no overfitting. This has also been confirmed by Breiman (2001) that Random Forest, as a tree-based ensemble method, improves accuracy by combining multiple Decision Trees. This is also the outstanding strength of Random Forest compared to other supervised machine learning methods. In 200 Test samples, there was only 1 false negative sample, meaning "Churn", but the model predicted "Not-Churn", and there was no false positive error, meaning 100% of Not-Churn objects were predicted correctly (Table 2). Incorrect prediction will cause the bank to lose the opportunity to retain customers; however, in this study, the level of incorrect prediction of the model is insignificant. The prediction results in Table 3 show that the Random Forest performance is almost optimal with accuracy =

99.5%, Recall = 99.14%, and F1-score = 99.57%. According to the recommendation of Alwash et al. (2025), simulation data with high distribution fidelity helps to evaluate the performance of supervised machine learning models more stably. In the study, the simulation data in the study was built according to the principle of strictly controlling behavioral features based on practical grounds, while adding only a moderate amount of Gaussian noise to reflect natural fluctuations without blurring the classification boundaries (Breiman, 2001; Meldrum et al., 2025). Thanks to the good control of noise levels and clear classification boundaries, the Random Forest model achieved 99.5% accuracy on the test set without overfitting, providing a scientific basis for banks to invest in customer behavioral data collection and cleaning systems to maximize predictive efficiency.

In addition, the comparison in Table 3, conducted on the same simulated dataset, reveals that Logistic Regression and SVM have noticeably lower performance than Random Forest. Gradient Boosting performs close to Random Forest, yet Random Forest still demonstrates superior predictive ability. These results confirm that Random Forest has strong and superior predictive performance among supervised machine learning models. These findings are consistent with previous studies, such as Manorom et al. (2024), Yang (2024), and Salunke et al. (2025), which reported that Random Forest consistently outperforms other supervised models, such as Logistic Regression, SVM, and Decision Trees, in prediction tasks based on historical data.

A limitation of this study is that the simulated dataset is relatively small compared to real-world mobile banking data. While the dataset was carefully designed to capture key behavioral patterns and relationships, its limited size may limit the model's ability to generalize to larger, more diverse user populations. Therefore, future studies should consider using larger-scale datasets to validate the findings and ensure that the model's predictive performance remains robust under real-world conditions. Additionally, prediction performance will decrease if there is a data shift. This happens when customers change their service usage behavior, such as paying via e-wallet or e-commerce platform to enjoy incentives or change seasonal behavior, such as during holidays, the frequency of login and money transfer will often be higher than usual.

4. Conclusion and Policy Implications

The study uses supervised machine learning, specifically the Random Forest model, to predict the rate of customers leaving mobile banking apps in order to build customer retention strategies, attract potential customers, and thereby optimize the bank's business efficiency. The results show that the prediction performance of Random Forest is very high, with an accuracy of up to 99.5%, along with the discovery of two key features that are strong enough to predict the possibility of leaving the app, which are Login Frequency and Balance Checks. The results of the study make a significant contribution to business activities on digital technology platforms, not only in the banking industry but also across other sectors of the economy.

From the research results, some policy implications are proposed as follows:

First, applying supervised machine learning, especially the Random Forest model or exploiting other models suitable for the characteristics of each bank, with priority given to two important input features, Login Frequency and Balance-Check, to early forecast the trend of customer service usage based on historical user data instead of conducting traditional surveys. However, the

forecasting process needs to be performed periodically and the model retrained to ensure forecasting performance when user behavior changes. This application can be widely applied to fields in the economy.

Second, based on the analysis of user behavior history for both Churn and Not-Churn groups, upgrade popular services, and build new service packages to enhance customer experience. In addition, banks should collaborate with other business units to offer preferential combos when customers use partners' services, which will encourage customers to use the service more and stay with the bank longer.

Third, prioritize developing a customer care strategy focused on customers with a high probability of leaving, because once customers have abandoned, the probability of returning is extremely low. When customers are loyal, it not only provides a stable source of profit for the bank but can also help the bank expand its customer base through referrals from friends and relatives. Therefore, retaining customers who are on the verge of leaving is extremely important.

REFERENCES

Aburbeian, A.M., Ashqar, H.I. (2023). Credit card fraud detection using enhanced random forest classifier for imbalanced data. In: Daimi, K., Al Sadoon, A. (eds) *Proceedings of the 2023 International Conference on Advances in Computing Research* (ACR'23). ACR 2023. Lecture Notes in Networks and Systems, vol 700. Springer, Cham. https://doi.org/10.1007/978-3-031-33743-7_48

Adekunle, B. I., Chukwuma-Eke, E. C., Balogun, E. D., & Ogunsola, K. O. (2023). Improving customer retention through machine learning: A predictive approach to churn prevention and engagement strategies. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *9*(4), 507-523. https://doi.org/10.32628/IJSRCSEIT

Alariyibi, A., El-Jarai, M., & Maatuk, A. (2023). Evaluating the accuracy of classification algorithms for detecting heart disease risk. *arXiv preprint arXiv:2312.04595*. https://doi.org/10.48550/arXiv.2312.04595

Al-Najjar, D., Al-Rousan, N., & Al-Najjar, H. (2022). Machine learning to develop credit card customer churn prediction. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(4), 1529-1542.

Alwash, M., Al Hajj, G. S., Grytten, I., & Sandve, G. K. (2025). Meta simulation approach for evaluating machine learning method selection in data limited settings. *Scientific Reports*, *15*(1), 40766. https://doi.org/10.1038/s41598-025-24627-y

Aysan, A. F., Ciftler, B. S., & Unal, I. M. (2024). Predictive power of random forests in analyzing risk management in Islamic banking. *Journal of Risk and Financial Management*, *17*(3), 104. https://doi.org/10.3390/jrfm17030104

Balabied, S. A. A., & Eid, H. F. (2023). Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. *PeerJ Computer Science*, *9*, e1708. https://doi.org/10.7717/peerj-cs.1708

Balla, I., Rahayu, S., & Purnama, J. J. (2021). Garment employee productivity prediction using random forest. *Jurnal Techno Nusa Mandiri*, *18*(1), 49-54. https://doi.org/10.33480/techno.v18i1.2210

Bican, P. M., & Brem, A. (2020). Digital business model, digital transformation, digital entrepreneurship: Is there a sustainable "digital"?. *Sustainability*, *12*(13), 5239. https://doi.org/10.3390/su12135239

Boozary, P., Sheykhan, S., GhorbanTanhaei, H., & Magazzino, C. (2025). Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction. *International Journal of Information Management Data Insights*, *5*(1), 100331. https://doi.org/10.1016/j.jjimei.2025.100331

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, *557*, 317-331. https://doi.org/10.1016/j.ins.2019.05.042

Cheng, L. C., Wu, C. C., & Chen, C. Y. (2019). Behavior analysis of customer churn for a customer relationship system: An empirical case study. *Journal of Global Information Management (JGIM)*, *27*(1), 111-127. https://doi.org/10.4018/JGIM.2019010106

Deng, W. (2025). Prediction model of tourist traffic data based on random forest algorithm. In: Al-Turjman, F. (eds) *Smart Infrastructures in the IoT Era*. Sustainable Civil Infrastructures. Springer, Cham. https://doi.org/10.1007/978-3-031-72509-8_67

Han, L., Bo, Q., Wei, G., & Pan, Y. (2024). Research on SMEs credit risk prediction based on decision tree and random forest. In D. Gong, Y. Ma, X. Fu, J. Zhang, & X. Shang (Eds.), *Proceedings of the 13th International Conference on Logistics, Informatics and Service Sciences (LISS 2023)* (pp. 366-378). Springer. https://doi.org/10.1007/978-981-97-4045-1_29

Hasan, M. S., Siam, M. A., Ahad, M. A., Hossain, M. N., Ridoy, M. H., Rabbi, M. N. S., ... & Jakir, T. (2024). Predictive analytics for customer retention: Machine learning models to analyze and mitigate churn in e-commerce platforms. *Journal of Business and Management Studies*, *6*(4), 304-320. https://doi.org/10.32996/jbms.2024.6.4.22

Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278-282). IEEE.

Isson, J. P. (2018). *Unstructured data analytics: how to improve customer acquisition, customer retention, and fraud detection and prevention*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119378846

Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., & Bozkaya, B. (2018). Behavioral attributes and financial churn prediction. *EPJ Data Science*, *7*(1), 41. https://doi.org/10.1140/epjds/s13688-018-0165-5

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2), 271-294. https://doi.org/10.1007/s00607-021-00908-y

Longe, T. (2024, December 22). *Mobile app retention benchmarks by industries 2025*. UXCam. https://uxcam.com/blog/mobile-app-retention-benchmarks

Magidi, J., Nhamo, L., Mpandeli, S., & Mabhaudhi, T. (2021). Application of the random forest classifier to map irrigated areas using google earth engine. *Remote Sensing*, *13*(5), 876. https://doi.org/10.3390/rs13050876

Makariou, D., Barrieu, P., & Chen, Y. (2021). A random forest based approach for predicting spreads in the primary catastrophe bond market. *Insurance: Mathematics and Economics*, *101*, 140-162. https://doi.org/10.1016/j.insmatheco.2021.07.003

Manorom, P., Detthamrong, U., & Chansanam, W. (2024). Comparative assessment of fraudulent financial transactions using the machine learning algorithms decision tree, logistic regression, naïve bayes, k-nearest neighbor, and random forest. *Engineering, Technology & Applied Science Research*, *14*(4), 15676-15680. https://doi.org/10.48084/etasr.7774

Market.us Scoop. (2025). *Online banking statistics 2025 by finance, transactions, growth*. https://scoop.market.us/online-banking-statistics/

Meldrum, J., Suleiman, B., Rabhi, F., & Alibasa, M. J. (2025). New money: A systematic review of synthetic data generation for finance. *arXiv preprint arXiv:2510.26076*. https://doi.org/10.48550/arXiv.2510.26076

Mitchell, W. D. (2020). *Proactive predictive analytics within the customer lifecycle to Prevent Customer Churn* [Doctoral dissertation, Northcentral University].

Monetizely. (2025, June 22). *Understanding user engagement: How to track login frequency and session patterns*. https://www.getmonetizely.com/articles/understanding-user-engagement-how-to-track-login-frequency-and-session-patterns

Phua, C., Cao, H., Gomes, J. B., & Nguyen, M. N. (2012). Predicting near-future churners and win-backs in the telecommunications industry. *arXiv preprint arXiv:1210.6891*. https://doi.org/10.48550/arXiv.1210.6891

Salunke, Y., Phalke, S., Madavi, M., Kumre, P., Bobhate, G., Madavi, M. D., & Kumre, P. D. (2025). Fraud detection: a hybrid approach with logistic regression, decision tree, and random forest. *Cureus Journal Of Computer Science*, *2*(1), es44389-024-02350-5. https://doi.org/10.7759/s44389-024-02350-5

Silvey, S., & Liu, J. (2024). Sample size requirements for popular classification algorithms in tabular clinical data: empirical study. *Journal of Medical Internet Research*, *26*, e60231. https://doi.org/10.2196/60231

Wang, M., Hua, Y., Sun, H. L., Chen, Y., & Jiang, L. (2022). User churn behavior model of rural public digital cultural services: an empirical study in China. *Library Hi Tech*, *40*(5), 1267-1288. https://doi.org/10.1108/LHT-09-2020-0243

Yang, C. (2024). Machine learning algorithms based prediction for customer churn in banks. *Highlights in Business, Economics and Management*, *40*, 352-358. https://doi.org/10.54097/0svjfz52

Zakariah, M. (2014). Classification of large datasets using Random Forest Algorithm in various applications: Survey. *International Journal of Engineering and Innovative Technology (IJJEIT)*, *4*(3), 189-198.

Zhan, Y. (2024). Prediction and Feature Importance Investigation for Bank Churn Based on Machine Learning. *Highlights in Business, Economics and Management*, *40*, 409-415. https://doi.org/10.54097/4jkbx429

Zhang, Q. (2022). Financial data anomaly detection method based on decision tree and random forest algorithm. *Journal of Mathematics*, *2022*(1), 9135117. https://doi.org/10.1155/2022/9135117