ANALYSIS FOR BANK CUSTOMER CHURN PREDICTION USING ARTIFICIAL INTELLIGENCE BASED ON LOGISTIC REGRESSION AND DECISION TREE

Tukiyat[1a], Suryatna Sacadibrata[2], Taufiqur Rahman[3]
[1]National Research and Innovation Agency, Bogor, West Java, Indonesia
[a]Pamulang University, Tangerang Selatan, Banten, Indonesia
[2]South Tangerang City Government, Ciputat, South Tangerang, Indonesia
[3]Tech Mahindra Indonesia, South Jakarta, Jakarta, Indonesia
dosen02711@unpam.ac.id, sacadibrata@gmail.com, taufiqur.rahman@msn.com

Abstract

Customer churn is a significant problem in the banking sector because it can reduce profitability and increase the cost of attracting new customers. This study aims to evaluate the effectiveness of Logistic Regression and Decision Tree algorithms in predicting bank customer churn and identifying factors that influence customers' decisions to stop using bank services. The data used comes from Kaggle with a total of 10,000 data including demographic information, transaction activity, customer satisfaction, and other risk factors. The data was analyzed through a cleaning stage to eliminate duplicates and missing data, then continued with initial data exploration to understand patterns and correlations between variables. The analysis was carried out using Logistic Regression and Decision Tree models. The performance of both models was evaluated using metrics such as accuracy, precision, recall, and F1-score, and measured using ROC and AUC to assess the model's ability to distinguish between churned and non-churned customers. The results showed that 79.63% of customers remained active, while 20.37% experienced churn, which is a risk indicator for the sustainability of the bank's business. Model evaluation revealed that decision tree outperformed Logistic Regression on all metrics with an accuracy of 0.798 and 0.777, precision of 0.778 and 0.767, recall of 0.819 and 0.779, and F1-score of 0.798 and 0.773. However, Logistic Regression showed a higher AUC of 0.85 and 0.80. Decision tree was superior in detecting customers who actually churned. To reduce churn, it is recommended that banks offer loyalty programs and customize products for high-risk customers, such as those with low balances or credit scores. The use of other models such as random forest and gradient boosting can be suggested to improve accuracy and provide deeper insights in reducing churn.

Keywords: Bank Customer Churn, Logistic Regression, Decision Tree, Artificial Intelligence, Churn Prediction

INTRODUCTION

The banking sector currently faces major challenges in retaining customers due to increasing competition. Customer churn, which refers to customers leaving a bank to seek alternative financial services, can have a significant impact on a bank's profitability. In addition to lost revenue, churn also has the potential to increase customer acquisition costs, making it a major concern in a bank's business strategy. As stated by Ashraf (2024), uncontrolled churn can reduce profits by up to 25%. This shows how important it is for banks to focus on managing churn in order to maintain

Hal: 80-89
DOI: https://doi.org/10.32493/iijoms.v1i1.53671

long-term profitability. Abdulsalam et al. (2022) also emphasize the importance of implementing advanced AI techniques, such as Classification and Regression Trees (CART) and Artificial Neural Networks (ANN), in predicting churn, as these techniques can improve the accuracy and robustness of predictive models. These technologies enable banks to respond more quickly and adapt their offerings according to the needs of customers at risk of churning.

Previous studies have shown that customer data, including transaction history, account activity, and demographic information (such as age and gender), play an important role in predicting churn. Chen et al. (2023) identified that female customers with low credit scores and low account activity are more prone to churn, which is in line with the findings of Hon et al. (2023). This study underscores the importance of understanding the dynamics of customer behavior to develop more effective prediction models. They also highlight that the application of advanced feature analysis techniques, such as Principal Component Analysis, can help address the problem of high-dimensional data and improve the performance and interpretability of prediction models. This is important in the context of banks that often have complex and large customer data, which requires efficient data processing techniques.

In customer churn research, Logistic Regression and Decision Tree algorithms have proven to be effective, given their ability to analyze linear relationships between variables and produce easy-to-understand visualizations, which help identify key factors causing churn (B et al., 2021; Zhang et al., 2022). These algorithms provide valuable insights into how customer decisions are influenced by various internal and external factors. They emphasize that the use of data preprocessing techniques such as SMOTE, combined with tree-based algorithms, can improve model accuracy, especially on imbalanced datasets. Therefore, it is important for banks to implement these techniques to reduce bias and improve the quality of their churn predictions. In addition, the integration of adaptive synthetic sampling methods such as ADASYN can increase model sensitivity by up to 15% (He et al., 2008). This provides an advantage in analyzing imbalanced data and ensures that high-risk customers can be identified more accurately.

This study is highly relevant as it offers a deeper understanding of how AI-based approaches can reduce customer churn. This technology enables banks to build stronger relationships with their customers through more personalized interactions and more efficient retention strategies. This study contributes to the development of more accurate and efficient churn prediction models and provides recommendations for more targeted customer retention strategies. By understanding the characteristics of customers who are at high risk of churn, banks can be more proactive in designing appropriate mitigation measures, such as offering special promotions or providing more personalized services. Therefore, this study not only contributes to the development of technology in the banking sector but also supports the sustainability and growth of banks in the future. Success in managing customer churn will create a sustainable competitive advantage for banks in this increasingly competitive market.

LITERATURE REVIEW

Hal: 80-89
DOI: https://doi.org/10.32493/iijoms.v1i1.53671

Customer churn is a phenomenon where customers stop using a company's products or services and switch to another provider. In the banking context, churn can occur due to various factors, such as dissatisfaction with the service, high costs, or more attractive offers from competitors (Ashraf, 2024). Churn can be intentional, for example through account closure, or unintentional, such as account inactivity for a certain period of time (Abdulsalam et al., 2022). According to Zeithaml et al. (2023), customer churn can also be influenced by emotional aspects, such as loyalty and trust in the brand. Churn occurs when customers lose this emotional attachment, which can cause them to switch to another brand. Given its impact on a company's profitability, churn is an issue that needs to be managed effectively. Studies show that retaining existing customers is more cost-effective than attracting new customers, with potential savings of up to five times (Kotios et al., 2022). Churn is also influenced by changes in customer needs, such as switching to more modern or flexible products (Patel et al., 2022).

Customers are now more likely to choose banks that offer efficient, secure, and user-friendly digital services. A study by Morgan et al. (2024) shows that adopting technologies such as mobile banking applications with personalized features can increase customer retention. However, when digital services fail to meet expectations, the risk of churn increases significantly. Another factor that influences churn is poor customer experience at key touchpoints, such as the credit application process or customer service (Johnson & Lee, 2023). Banks are also under pressure to provide data-driven services, which allows them to predict churn more accurately through customer behavior analysis (Smith et al., 2024). In addition, a proactive approach through relevant communication and fast response times to customer complaints has been shown to significantly reduce churn (Taylor & Brown, 2023; Anderson et al., 2024).

Logistic Regression and Decision Tree are algorithms that are often used to predict customer churn because of their flexibility and ability to handle complex datasets. Logistic Regression analyzes the probability of churn based on the linear relationship between independent and dependent variables, making it a suitable method for cases with simple correlations (Hosmer et al., 2013; Han et al., 2012). Meanwhile, Decision Tree is able to handle data with complex rules and provides easy-to-understand interpretations, making it suitable for various types of customer datasets (Quinlan, 1996). Logistic Regression and Decision Tree are two widely used algorithms. Logistic Regression is very effective for analyzing the probability of churn in data with a linear relationship between variables. This method uses the sigmoid function to map independent variables to the probability of binary outcomes, such as churn or non-churn. Logistic Regression is suitable for simple datasets, but is limited when faced with non-linear patterns in the data. In the context of customer churn, this algorithm is often used to assess the influence of demographic factors, transaction behavior, and service usage patterns (Wen, 2023; He et al., 2024). In predicting churn, artificial intelligence algorithms such as Logistic Regression and Decision Tree are often used. Logistic Regression helps analyze the probability of churn based on linear relationships between variables, while Decision

Hal: 80-89
DOI: https://doi.org/10.32493/iijoms.v1i1.53671

Tree utilizes logical rules to divide data into simpler subsets (Wen, 2023; Zhang et al., 2022).

RESEARCH METHODS

This study uses a quantitative approach to analyze bank customer churn predictions by utilizing Artificial Intelligence algorithms, namely Logistic Regression and Decision Tree. The dataset used in this study was obtained from Kaggle and is bank customer data, including demographic variables, transaction activity, customer satisfaction, and other risk factors. (Kumar, 2020). The data goes through a data cleansing process to eliminate duplication, missing data, or outliers that can affect the analysis results. After the data is ready, initial data exploration is carried out to understand the basic patterns and correlations between relevant variables. Conceptually, the research design can be shown in Figure 1 below:

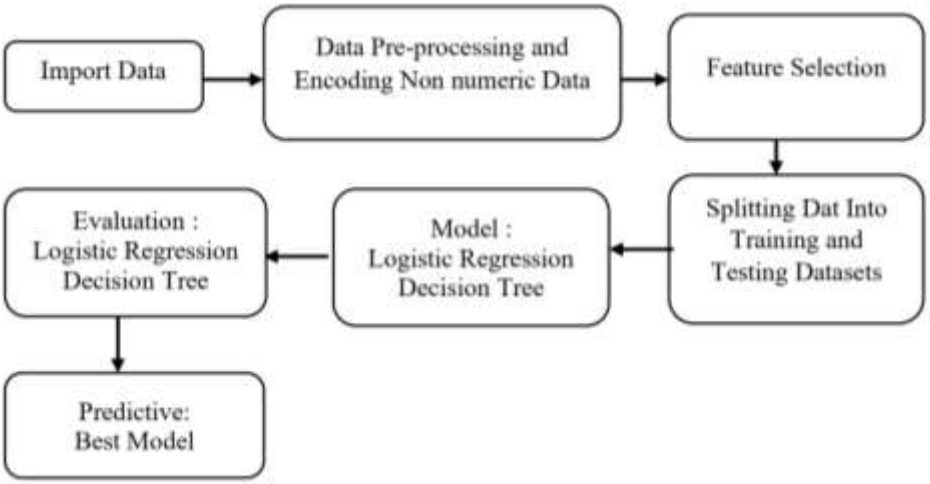

Figure 1. Research Design

The flowchart above illustrates the process of building a machine learning model of Logistic Regression and Decision Tree for classification. Conceptually, the research flow begins with data collection from various sources (data import) and is processed through cleaning and converting non-numerical data into a numeric format that can be used by the algorithm (data pre-processing). After the data is valid, the next step is to select features (feature selection) to determine the most relevant variables, in order to increase the efficiency and accuracy of the model. At this stage, the data is divided into training data and test data (data splitting) to train the model and measure its performance. The implementation of logistic regression and decision tree models is used in the training stage (model training) before being evaluated using metrics such as accuracy, precision, recall, and F1-score. Finally, the best model is selected to predict new data, ensuring an optimal solution based on performance analysis. Model performance is assessed using key metrics from the confusion matrix, such as accuracy to measure the percentage of correct predictions, precision to determine the proportion of correct churn predictions, and recall to measure how well the model detects all churn cases. F1-Score, as the harmonic mean of precision and recall, is used to balance the two metrics. Additional analysis was performed using receiver operating characteristic (roc) curves and area under the curve (auc) to

Hal: 80-89
DOI: https://doi.org/10.32493/iijoms.v1i1.53671

evaluate the model's ability to distinguish churned and non-churned customers at various classification thresholds.

RESULTS AND DISCUSSION

This descriptive analysis is an analysis conducted at an early stage to determine the description of the conditions and characteristics of independent variables that are suspected of influencing customer churn using descriptive statistics. Descriptive statistical analysis is a method in the EDA (Explanatory Data Analysis) stage, where this method is related to the collection and presentation of data in order to provide useful information. Information obtained from this analysis includes measures of data centralization, measures of data distribution, and also the tendency of a data cluster. In this study, descriptive analysis was used to understand the main characteristics of a dataset consisting of 10,000 customers. The variables analyzed include demographic attributes such as age, geographic location, and gender, as well as financial attributes such as credit score, balance, number of products owned, and estimated income. In addition, customer churn status (exited) was also analyzed to understand its distribution as a target variable in this study.

Here are some very influential tables, table 1 shows variations in important attributes that have the potential to influence customer churn behavior. Table 2 helps in analyzing differences in churn behavior based on geographic location , and table 3 is the distribution of customer churn status .

Table 1Product Details

Attributes	Avg.	Max	Min	Std. Deviation
Credit score	650.52	850	350	96.65
Balance	76,485.89	250,898.09	0	62,397.41
Number Of Product	1.53	4	1	0.58

Table 1 above provides a descriptive overview of the three main attributes, namely Credit Score, Balance, and Number of Products. The average customer credit score is 650.53, with the highest value of 850, the lowest 350, and a standard deviation of 96.65, reflecting the variation in the level of customer financial confidence from very good to risky. The average account balance is 76,485.89 with a maximum balance of 250,898.09 and a minimum of 0, and a standard deviation of 62,397.41, indicating a significant spread of balances, from customers with no balance to those with very high balances. In addition, the average number of products used by customers is 1.53 with a maximum of 4 and a minimum of 1, and a standard deviation of 0.58, indicating that most customers use 1-2 bank products, with only a few using more products. This variation provides important insights into customer characteristics that can affect churn analysis.

The distribution of customers in this case consists of France, Spain and Germany. The detailed distribution of customers can be shown in table 2 below.

Table 2. Geographic Distribution

Country	Customer	Percentage(%)
France	5,014	50.14
Spain	2,509	25.09
Germany	2, 477	24.77

Hal: 80-89
DOI: https://doi.org/10.32493/iijoms.v1i1.53671

Geographic distribution of customers based on country of origin in datasets of 10,000 data. Of the total customers, most come from France at 50.14%, Spain at 25.09% and Germany at 24.77%. This distribution shows that more are concentrated in France. Information can support banks in analyzing differences in churn behavior based on geographic location.

Table 3. Churn Status

	Status	Customer	Percentage (%)
0	No churn	7,963	79.63
1	churn	2,037	20.37

Table 3. Churn status, shows the distribution of customer churn status in datasets consisting of 10,000 data. Most customers, as many as 7,963 or 79.63%, are in the No Churn category, which means they remain active customers. Meanwhile, 2,037 customers or 20.37% are included in the Churn category, which indicates that customers have stopped being customers. This distribution shows that the churn rate of 20.37% is quite significant and can threaten business sustainability. Analysis of the characteristics of customers who churn is needed to identify the root causes. Thus, companies can design retention strategies, such as improving service quality, personalizing offers, or loyalty programs to reduce churn. Factors related to churn are analyzed using a correlation heatmap as shown in Figure 2.

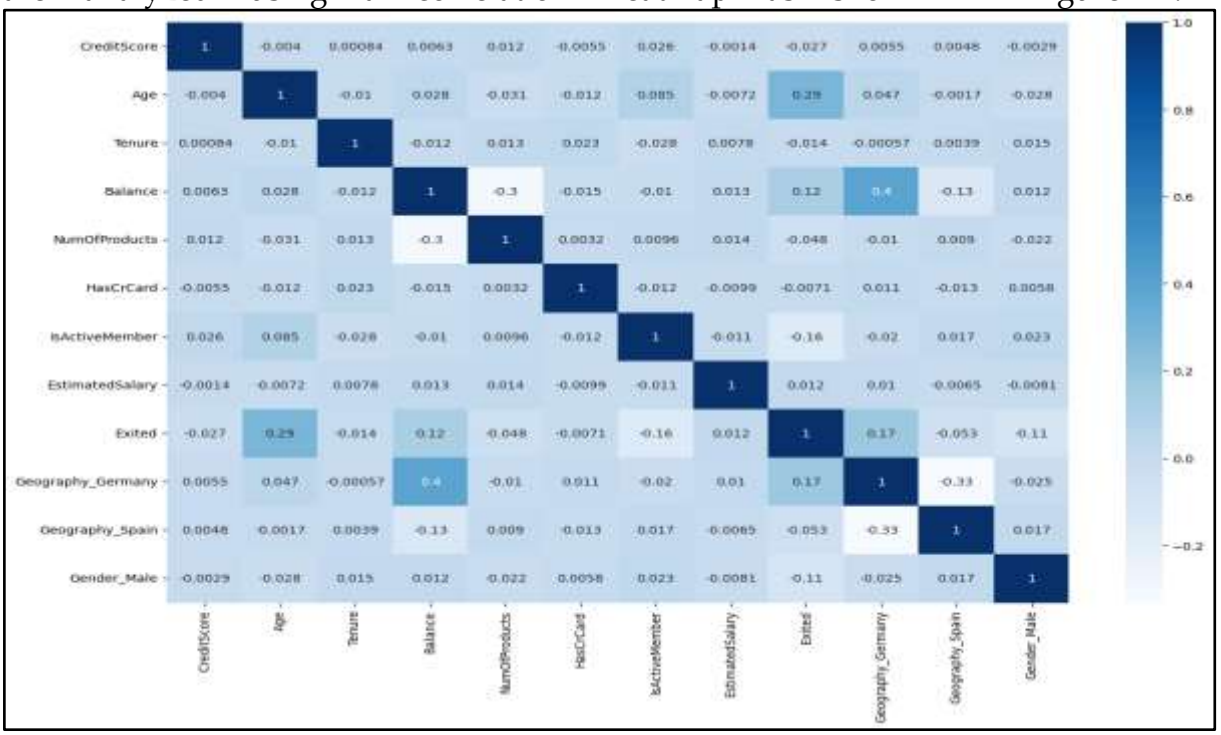

Figure 2. Heatmap

The results of the correlation headmap analysis as in Figure 2 show that age (correlation 0.29) and balance (0.12) have a positive relationship with customer decisions to leave, indicating that older customers or those with higher balances tend to leave the bank, perhaps due to retirement or looking for better offers. In contrast, the number of products has a weak negative correlation (-0.048), indicating that

Hal: 80-89
DOI: https://doi.org/10.32493/iijoms.v1i1.53671

cross-selling strategies can increase loyalty. Other variables, such as credit cards, active membership status, and geographic factors, show very weak correlations, so their influence on churn is less significant. Further analysis, such as regression, segmentation, and survival analysis, are needed to understand the causal and in-depth relationship. This correlation analysis provides an initial insight into the factors that may influence a customer's decision to leave. Companies can maintain customer loyalty by providing special services based on segmentation, such as retirement products or rewards for high-balance customers. Cross-selling strategies and loyalty programs such as cashback can increase engagement. Responsive customer service and proactive communication are also important to prevent churn.

A graphical depiction of the data pattern of the churn classification model with Logistic Regression and Decision Tree is shown in Figures 3 and 4 below.

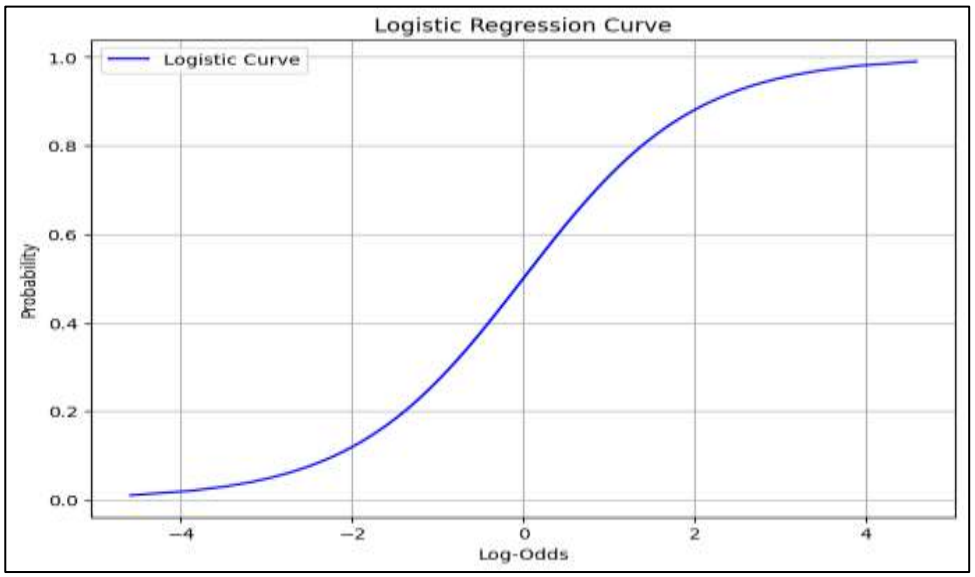

Figure 3 Logistic Regression

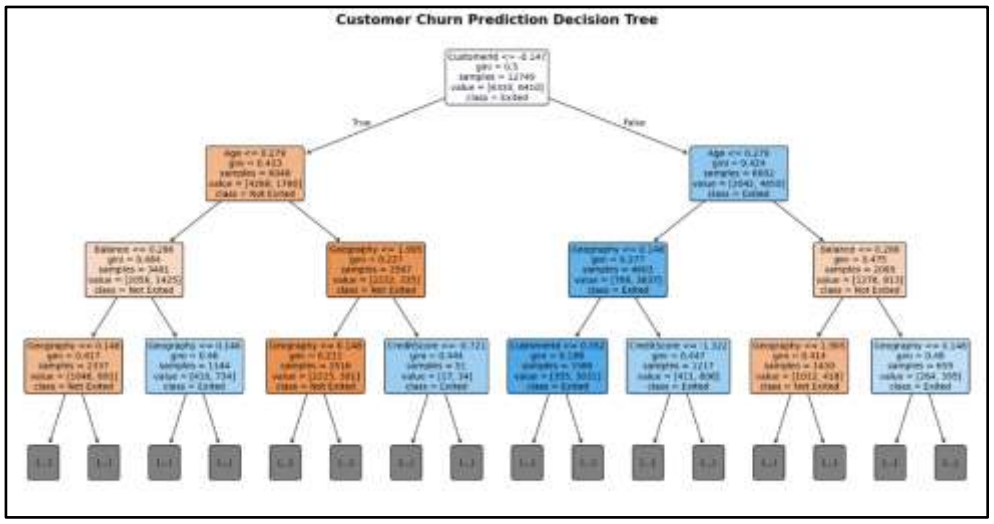

Figure 4. Decision Tree

The data pattern in the Decision Tree model shows that customers with Customerid less than 0.147 and younger age are more likely to churn. Low balance

Hal: 80-89
DOI: https://doi.org/10.32493/iijoms.v1i1.53671

reduces the likelihood of churn, while geographic location also plays a role. The model identifies key factors for churn prediction, such as Customerid, age, balance, and geography.

Model evaluation is performed using several key metrics, namely accuracy, precision, recall, and F1-score. Below (table 4) are the results of the model evaluation.

Table 4. Results of Logistic Regression and Decision Tree

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.787	0.767	0.779	0.773
Decision Tree	0.798	0.778	0.819	0.798

The model evaluation results show that Decision Tree has a better performance compared to Logistic Regression on all the main metrics used. The Decision Tree model obtained an accuracy of 0.798, higher than Logistic Regression which only reached 0.777. In addition, the precision of the Decision Tree model is also better with a value of 0.778, compared to Logistic Regression which has a precision of 0.767. This shows that Decision Tree is more effective in identifying customers who actually churn among churn predictions. The recall on the Decision Tree model which is worth 0.819 is also superior to Logistic Regression which reaches 0.779, indicating that Decision Tree is better able to detect customers who actually churn. The F1-score value of the Decision Tree model reaches 0.798, higher than Logistic Regression which is worth 0.773. With better performance on all four metrics, it can be concluded that the Decision Tree model is a better choice for predicting customer churn.

The level of model evaluation prediction performance is measured by Receiver Operating Characteristic (ROC) through the Area Under the Curve (AUC). The ROC curve is used to evaluate the model's predictive ability by showing the relationship between the true positive rate (recall) and the false positive rate. Area Under the Curve (AUC) is the main indicator in ROC analysis, where a higher AUC value indicates better model performance in distinguishing between two classes, such as churned and non-churn customers. In this analysis, Logistic Regression has an AUC of 0.85, which is higher than Decision Tree which only obtained an AUC of 0.80, indicating that Logistic Regression has better predictive ability in distinguishing churned and non-churn customers. Graphically, it can be shown in Figure 4.

Hal: 80-89
DOI: https://doi.org/10.32493/iijoms.v1i1.53671

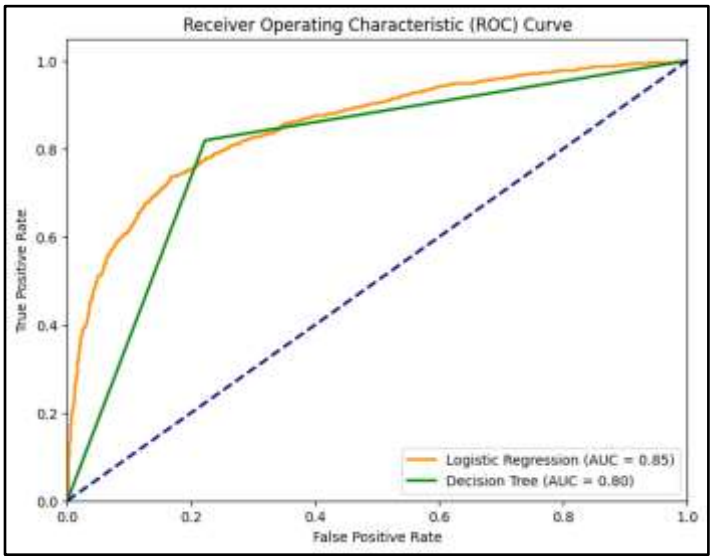

Figure 4. ROC Curve

CONCLUSION

The conclusion of this study shows that both Logistic Regression and Decision Tree have an important role in predicting customer churn, although each has its own advantages. Logistic Regression shows a more stable performance with a high AUC (0.85) and good accuracy (77.68%), making it a more effective choice for churn prediction in general. However, Decision Tree performs better in terms of recall (81.91%) and accuracy (79.79%), although its AUC is slightly lower (0.80). The advantage of Decision Tree lies in its ability to provide clearer interpretations, helping banks understand the factors that influence churn, such as age, balance, and geographic location. Both models have the potential to improve customer retention strategies, with Logistic Regression superior in general classification, while Decision Tree is more useful for in-depth analysis and customer segmentation. The main suggestion to reduce churn is to offer loyalty programs and product customization for high-risk customers, such as those with low balances or credit scores. The implementation of these two models can strengthen the bank's efforts to reduce churn, by identifying risky customers more precisely and allowing for more effective preventive actions. Other models that can be used to reduce churn are Random Forest, and Gradient Boosting. These models can improve the accuracy of predictions and provide deeper insights into the factors that influence churn.

REFERENCE

Abdulsalam, Y., Ali, N., & Omer, I. (2022). *Application of artificial intelligence for customer churn prediction in the banking sector* . International Journal of Artificial Intelligence, 14(3), 45-67.

Anderson, R., Taylor, S., & Brown, J. (2024). *Reducing customer churn through proactive communication: A banking case study* . Journal of Financial Services Marketing, 29(2), 12-23.

Ashraf, A. (2024). *Customer churn and its impact on profitability in the banking sector: A review* . Journal of Business and Economics, 32(4), 134-145.

Chen, Y., Zhang, X., & Wu, H. (2023). *Gender differences in customer churn prediction in the banking sector* . Journal of Marketing Research, 60(1), 50-61.

DOI: https://doi.org/10.32493/iijoms.v1i1.53671

He, H., Liu, X., & Zhang, T. (2008). *Improved sampling techniques for imbalanced classification tasks: ADASYN and SMOTE comparison* . IEEE Transactions on Knowledge and Data Engineering, 20(4), 452-463.

Hon, K., Lee, J., & Cheung, C. (2023). *Demographic factors and transaction behavior in predicting churn in the banking sector* . International Journal of Bank Marketing, 41(6), 334-349.

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

Johnson, D., & Lee, Y. (2023). *The role of customer service in reducing churn in retail banking* . Journal of Service Research, 28(2), 78-92.

Kumar, V. (2020). *Customer retention strategies in banking: Using data analytics to predict churn* . Journal of Financial Technology, 12(1), 56-69.

Kotios, S., Nikos, T., & Vasilis, K. (2022). *Cost-benefit analysis of customer retention in banking* . International Journal of Economic and Financial Issues, 8(3), 45-56.

Morgan, P., Harris, R., & Chen, Z. (2024). *The impact of mobile banking applications on customer retention* . Journal of Digital Banking, 15(2), 23-35.

Patel, N., Sharma, P., & Soni, R. (2022). *The changing needs of customers and their impact on bank service offerings* . Journal of Banking and Finance, 24(1), 101-112.

Quinlan, J. (1996). *C4.5: Programs for machine learning* . Morgan Kaufmann.

Smith, L., Rojas, F., & Walker, J. (2024). *Predictive analytics in banking: Forecasting customer churn with machine learning* . Journal of Business Analytics, 11(1), 22-36.

Taylor, S., & Brown, J. (2023). *Proactive customer service as a churn prevention strategy in financial institutions* . Journal of Consumer Satisfaction, 19(4), 101-114.

Wen, J. (2023). *The application of logistic regression in customer churn prediction* . Journal of Data Science, 11(3), 50-63.

Zhang, L., Liu, F., & Wang, M. (2022). *A comparative study of decision trees and logistic regression in customer churn prediction* . Journal of Computational Statistics, 34(4), 120-134.

Zeithaml, V., Parasuraman, A., & Berry, L. (2023). *The behavioral consequences of service quality and customer satisfaction in the banking industry* . Journal of Marketing Research, 28(2), 91-103.