ISSN 2085-4552

A Comparative Study : Predicting Customer Churn in Banking Using Logistic Regression & Random Forest

Mhd Basri[1], Muhammad Iqbal Pradipta[2], Krisna Aditya[3]
[1,2,3] Faculty of Computer Science and Information Technology,
Muhammadiyah University of North Sumatra, Medan, Indonesia
[1]mhd.basri@umsu.ac.id , [2]muhammad.iqbalpradipta@umsu.ac.id, [3]ozwellespencer49@gmail.com

Accepted 20 May 2025
Approved 13 June 2025

***Abstract*— Customer churn prediction is crucial in the banking industry, where retaining existing customers is often more cost-effective than acquiring new ones. Understanding the factors that lead customers to leave can help banks implement proactive retention strategies, improve customer satisfaction, and maintain long-term profitability. This research explores the prediction of bank customer churn using machine learning techniques. The dataset used includes various customer features such as demographics, transaction history, and interactions with the bank. After performing exploratory data analysis (EDA) and pre-processing, two machine learning models were applied: Logistic Regression and Random Forest. Logistic Regression was chosen for its simplicity and interpretability, while Random Forest was selected for its ability to handle complex, non-linear relationships in the data. The EDA results showed that factors such as number of transactions, total transaction value, and credit utilization rate were correlated with the likelihood of churn. Pre-processing included handling categorical data, removing irrelevant features, and dividing the data into training and testing sets. The Logistic Regression model achieved 84% accuracy on training data and 83.9% on testing data, but showed poor performance in terms of recall and F1-score for the "Attracted Customer" class. In contrast, the Random Forest model showed excellent performance with 100% accuracy on both datasets, as well as perfect precision, recall, and F1-score values for both classes. In conclusion, the Random Forest model was selected as the best model to predict bank customer churn. These findings can help banks identify customers at risk of churn and develop effective retention strategies.**

Index Terms*— *Customer churn; Attrition; Banking Industry; Machine learning; Logistic Regression; Random Forest.

I. INTRODUCTION

Customer attrition, also known as customer churn, is the phenomenon where customers terminate their relationship with a business or organization [1]. Customer churn has emerged as a pressing challenge in sectors such as banking, telecommunications, music media, and insurance [2]. Customer attrition in the banking industry occurs when consumers stop using the products and services offered by the bank for some time, and then end their relationship with the bank. Therefore, customer retention has become very important in today's highly competitive banking market [1]. The rate at which a company loses clients within a certain period of time is called the churn rate, also known as the customer attrition rate [3]. Although many banks strive to attract and retain customers, customers are often dissatisfied and choose to switch. Factors that can lead to customer loss can include poor service quality, fees that are perceived to be too high, or a lack of innovation in products and services. In addition, having a strong customer base helps attract new consumers by building trust and gaining referrals from existing customers. These factors make reducing customer attrition an important step for banks to take.

The impact of customer churn can be very significant for banks Losing a customer means losing potentially valuable revenue. In addition, the cost of attracting new customers is often higher than retaining existing ones. To find customers who may be lost, customer churn prediction techniques can be used. Then, the results can change the marketing strategy [4]. If churn is not managed well, it can also damage the bank's reputation, making potential customers hesitant to join. Therefore, companies must accurately predict customer churn and take appropriate actions. A high churn rate can create a negative perception in the community. When many customers leave a bank, it can affect the trust of other customers and make them consider switching banks. This domino effect can be very detrimental to the bank in the long run. With the advancement of technology, researchers are starting to look for solutions to this problem.

Customer churn prediction has become an important research focus in various industries, especially in banking, where customer retention directly affects revenue streams and satisfaction levels.

For example, Lalwani et al. [5] conducted churn prediction research in the telecommunications industry using a dataset consisting of approximately 7,000 customer records with 21 features, including numerical and categorical attributes such as tenure, monthly fee, contract type, and internet service usage. The dataset was pre-processed through missing value handling, categorical coding, and balancing using resampling techniques, thus enabling the application of machine learning models for accurate churn classification. Traditional methods like Logistic Regression have shown promising results, achieving accuracy rates of around 80.45%

Verma et al. study conducted research on churn prediction for savings bank customers using transaction and demographic data obtained from a national bank data warehouse in India [6]. The dataset included customers aged 21-50 years featuring 47 variables selected from 66 available variables, such as age, income, occupation, and transaction behavior. The Random Forest algorithm outperformed other models including GLM, ANN, Decision Tree, and XG-Boost, achieving 78% accuracy on test data and 79% on validation data, with an AUC of 0.844. This study highlights Random Forest as a reliable model for practical application in customer retention strategies.

The study presented by Wagh in predicting customer churn used the Telco Customer Churn dataset obtained from Kaggle, which included 7,043 customer records with 21 attributes (16 categorical and 5 numerical). The dataset labeled 26.53% of customers as churners. The Random Forest algorithm was used for classification, and outperformed the other models tested. Before up-sampling, the model achieved 98.91% accuracy, with 99% precision and recall. After applying SMOTE and ENN techniques to address class imbalance, the model performance further improved to reach 99.09% accuracy [7]. The results of this study conclude that Random Forest offers a robust solution for churn prediction in the telecommunications sector and can be effectively integrated with retention strategies such as survival analysis and Cox proportional hazard modeling.

Another example, research conducted by Prabadevi by testing several machine learning algorithms such as K-Nearest Neighbor (KNN), Logistic Regression (LR), Random Forest (RF) and also Stochastic Gradient Booster (SGB) shows that the Linear Regression algorithm is still superior to the Random Forest algorithm with Classification Score Linear Regression getting Training Score 0.797, Test Performance 0.782, and AUROC 0.826 [10]. 797, Test Performance of 0.782 and AUROC of 0.826. while the Random forest algorithm gets a Training Score of 0.803, Test Performance of 0.787 and AUROC of 0.829. meaning that the Random Forest model is claimed to be better than Linear Regression [8].

Similarly, Dhangar also tested with the Logistic Regression algorithm, GaussianNB, SVM and Random Forest. the results show that SVM and Random forest get the advantage to predict Customer Churn with the Area under the curve (AUC) score of SVM 0.921 and Random forest 0.945. on Logistic Regression Getting AUC of 0.908. shows that SVM and Random forest are better at predicting Customer churn [9].

Advanced techniques, including hybrid approaches such as BiLSTM-CNN models, have emerged as powerful tools for capturing sequential patterns in churn prediction [10]. The banking sector has particularly benefited from these developments, with studies showing that proper handling of class imbalance through under-sampling techniques can significantly improve model performance [6]. Recent research has also explored the effectiveness of ensemble methods and hybrid algorithms, with some studies indicating that traditional approaches like Logistic Regression can outperform more complex models in certain scenarios [5], [9].

In an effort to address customer churn in the banking sector, deep learning approaches offer a promising solution with its ability to comprehensively analyze complex patterns from customer data. Bank churn prediction is made to find out how likely clients are to switch from one bank to another [11]. In this research, the aim is to analyze the bank's data and predict which users are likely to stop using the bank's services and turn into paying customers. Analyzing this data will help banks identify trends and try to retain customers who are on the verge of attrition.

II. METHODOLOGY

The machine learning project used data-driven methodologies in an organized manner to tackle the specified challenge. The first phase was gathering and prepping the data, which involved cleaning it after obtaining it from a reliable source to guarantee its quality and dependability. This procedure involved resolving any discrepancies, standardizing the data, and managing missing numbers. After preprocessing, the dataset's distribution, correlations, and possible patterns were examined using exploratory data analysis (EDA), which influenced the modeling approaches selected.

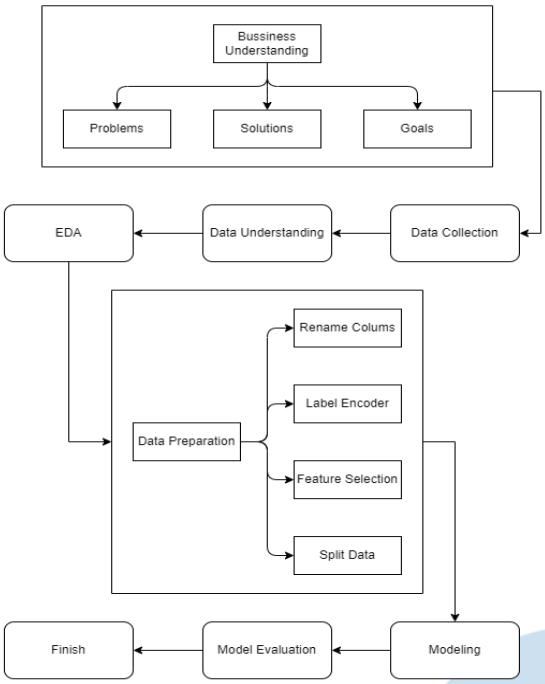

Fig. 1 Research Methods

A. *Bussiness Understanding*

This research focuses on the problem of customer attrition (churn) in the banking industry. Attrition occurs when customers stop using banking products and services, which negatively impacts the bank's revenue. In a highly competitive industry, customer retention is important to maintain business stability and reduce the cost of acquiring new customers. By having historical data from customers, we can use machine learning to predict customers who are at risk of quitting.

The data pre-processing steps that need to be performed to make the data ready to be used for training the model include various steps, such as data cleaning, handling missing values, normalization, and dividing the data into subsets suitable for training and testing.

The solution that can be done is to create a deep learning model and then evaluate the model using the Classification Report by looking at important metrics such as precision, recall, F1-score, and support, which are very helpful in assessing the performance of the model in classifying data and also creating heatmap graphs to analyze the correlation between existing features. Pre-processing the data properly so that it can be used, creating deep learning models to predict or identify customers who are at risk of churn and also knowing what features are most influential in customers who churn or not.

B. *Data Collection*

In the data collection stage, a dataset titled “Bank Customer Churn Prediction” sourced from the Hugging Face Datasets repository was used, and originally published through Analyttica's business analytics platform TreasureHunt. The platform is known for providing real datasets that are used to solve practical business problems. The dataset consists of 10,000 customer records from a private sector bank focused on credit card services. Each customer data includes 14 attributes, including: Customer ID, credit score, location, gender, age, tenure, account balance, number of products used, credit card ownership, active membership status, estimated income, and churn indicators. These attributes provide a comprehensive view of each customer's profile, enabling the construction of predictive models to identify customers at risk of churning. One important characteristic of this dataset is the presence of class imbalance, where only about 16.07% of customers are recorded to have churned. This poses a challenge in the model training process as it requires an appropriate strategy to accurately predict the minority class.

C. *Data Understanding*

The dataset that has been downloaded through Hugging face, then the dataset will be saved to google drive. In statistical analysis, this is very important as it enables proper handling of data, helps in model selection, and improves model performance and accuracy [12]. The dataset is accessed by mounting by specifying the file path to be accessed using Google Colab, then reading the file and adjusting it to pandas Dataframe and finally displaying information from the dataset which consists of 10,127 rows and 23 columns. In the context of customer churn prediction analysis in the banking industry, this stage is very important to identify patterns, trends, and relationships between variables that can affect churn.

D. *Exploratory Data Analysis*

Exploratory data analysis (EDA) process shows that the majority of customers are in the age range of 40-52 years, with some outliers whose age is above 70 years old as shown in the following figure:

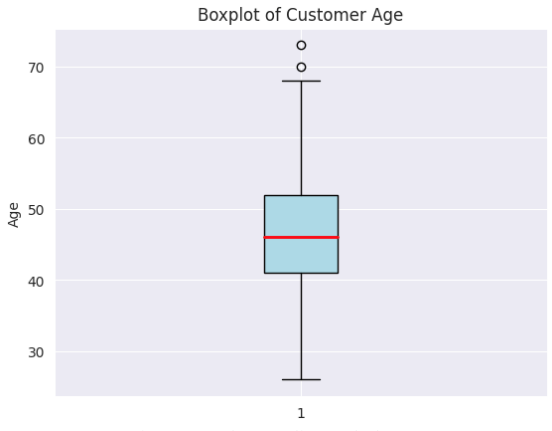

Fig. 2 Boxplot Credit_card_dataset

The figure above shows a boxplot of the age distribution of customers in the dataset used. The red line inside this visualization box shows that the median age of customers is about 46 years old. Most customers are between 35 and 55 years old, or the first and third quartiles. Of the few points above the age of 70, values outside this range are considered outliers. An initial overview of the age distribution of customers is given by this boxplot. This is important to analyze further as age is one of the factors that influence the likelihood of customer churn. In addition, the relatively symmetrical distribution indicates that the age variable does not suffer from significant skewness.

E. Data Preparation

In this section, we will explain in detail about the data cleaning and preprocessing processes carried out to prepare the dataset for analysis and modeling. This process is very important because good data quality will directly contribute to the accuracy and effectiveness of the machine learning models built.

To make the dataset easier to understand and use, column names that were too long were changed to shorter ones. To rename a column, a function with the parameter inplace=True is used, which indicates that the change is applied directly to the DataFrame. To ensure that the change is successful, the renamed column name is displayed.

The use of Label Encoder aims to convert categorical data into numerical format. This function allows categorical fields in a dataset to be processed by machine learning models by converting the category values into numbers, so that the model can learn patterns in the data. Label coding improves regression accuracy by enabling binary classification algorithms, improving error correction, and supporting end-to-end training across multiple regression tasks [13] . Feature importance analysis is essential to understand which input variables have the most significant impact on model predictions [14]. Removing columns that are irrelevant or unrelated to the model makes it easier for the model to process data so that it can focus on relevant features and improve its performance. By removing unrelated features, the model becomes simpler and more efficient.

The purpose of Split data is to divide the dataset into two parts training set and testing set, where there are independent features or X to predict the target and Y as the dependent variable or target label. in this case 20% of the total dataset will be used as testing data while 80% will be used as training data. Details of the data pre-processing used in this study can be seen in Table 1.

Table 1 Data Preprocessing Steps

Step	Description	Explanation
Column Renaming	Rename columns that are too long and change them to shorter versions.	use the rename() function with inplace = True to apply the changes directly to the DataFrame.
Label Encoding	converts categorical variables into a numerical format so that machine learning algorithms can process them effectively.	Each categorical feature is transformed using LabelEncoder().
Feature Removal	Irrelevant and insignificant columns were removed to improve model performance and reduce noise in the data.	This deletion uses the credit.card.drop() function
Data Splitting	The dataset was split into a training set and a testing set using an 80:20 ratio.	X represents the independent features, and y represents the dependent target variable Attrition_Flag. A fixed random_state ensures reproducibility.

F. Model Selection

In the Modeling stage, it discusses the use of machine learning algorithms to solve the churn prediction problem faced in the project. In this case, two models are used, namely Logistic Regression and Random Forest, to classify whether a customer will churn or not. Logistic Regression is a model of the Generalized Linear Model (GLM) where the response variable is a binary number (0 or 1) and follows a binomial distribution [15]. This model works by using a linear combination of input features to predict the probability that the data belongs to a particular category. It is easy to understand because it has a linear coefficient for each feature, which indicates how strongly each feature affects the prediction. This model is also fast and requires less computational resources. If there is a linear relationship between the features and the target variable, logistic regression works well.

Random forest is a machine learning algorithm that uses an ensemble of decision trees to make predictions [16]. Random Forest helps reduce overfitting and often produces more accurate predictions than single tree models. When compared to a single decision tree, this model is more stable against overfitting. Random Forest has complex interactions between features and can handle non-linear data. It can provide metrics to show which aspects affect the prediction the most.

G. Model Evaluation

This research uses a classification-type deep learning model which means that if it is close to 100% accuracy, the performance is good, while if it is below 75%, the performance is poor. The number of metrics that indicate different aspects of classification model performance is very important to consider when developing and evaluating models [13]. The following evaluation metrics and confusion matrix are results obtained from the best-performing model, which in this case is Random Forest :

- **Accuracy**

Accuracy is a measure of how many predictions are correct compared to all predictions made by the model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Based on the confusion matrix results, the model produces True Positive (TP) = 1699, True Negative (TN) = 327, and there are no False Positive (FP) or False Negative (FN). Then, the accuracy of the model is :

$$\frac{1669+327}{1699+327+0+0} = \frac{2026}{2026} = 100\%$$

- **Confusion Matrix**

Confusion matrix describes the results of model prediction against test data with the following details:

Table. 2 Confusion Matrix

	Negative Prediction (0)	Positive Prediction (1)
Actual Negative (0)	327	0
Actual Positive (1)	0	1699

- **Precision**

Precision is a measure of how many true positive predictions (True Positives, TP) are correct compared to all positive predictions made by the model (including False Positives, FP).

$$Precision = \frac{TP}{TP+FN} \quad (2)$$

With TP = 1699 and FP = 0, the precision is:

$$\frac{1699}{1699+0} = 1.0$$

- **Recall**

One of the evaluation metrics in classification is recall, which measures how many positive examples the model managed to find out of all the truly positive data.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Since FN = 0, recall:

$$\frac{1699}{1699+0} = 1.0$$

- **F1 – Score**

F1-Score, a metric that combines Precision and Recall in a single value, provides a more complete picture of model performance, especially in cases where Precision and Recall are not balanced.

$$F1-Score = \frac{2 \times Precision \times Recall}{Precision \times Recall} \quad (4)$$

Since Precision and Recall are both 1.0, then:

$$F1 = 2 \times \frac{1.0 \times 1.0}{1.0 \times 1.0} = 1.0$$

III. RESULT

In the methods section of this study after collecting and preparing the data, which involves cleaning the data after obtaining it from reliable sources to ensure its quality and reliability. This procedure involves resolving any discrepancies, standardizing data, and managing missing numbers. After preprocessing, the distribution, correlation, and possible patterns of the data set are examined using exploratory data analysis (EDA), which influences the modeling approach chosen. After the modeling stage, the algorithms used are evaluated and compared using performance metrics such as accuracy, precision, recall, and F1-score, to ensure the selection of the most effective model. Hyperparameter tuning is performed to further optimize model performance, using techniques such as grid search or random search.

The methodology also emphasizes model validation through techniques such as cross-validation, which helps assess the model's resistance to overfitting. Finally, the best-performing model is deployed and tested with unseen data to ensure its generalizability. Throughout the project, tools such as Python, Scikit-learn, and Matplotlib were used for implementation and visualization purposes.

Figure 3 below shows the distribution of the length of time a customer (in months) has been registered with a bank. The histogram shows the distribution of the original data, while the Kernel Density Estimation (KDE) adds a smooth curve that estimates the probability distribution of the data. This combination of histogram and KDE is useful for understanding how the data is distributed in more detail.

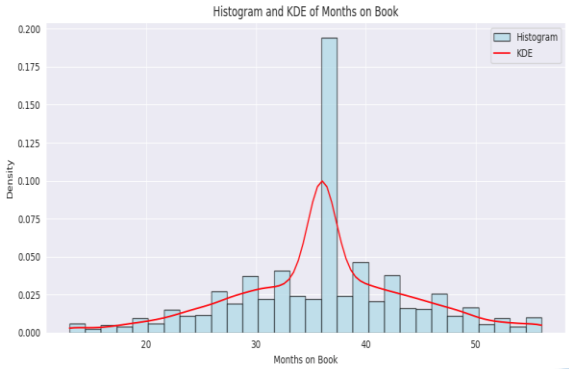

Fig. 3 Months_of_Book Relations

A. *Data Distribution*

The graph shows a distribution that is slightly skewed to the right. This means that there are fewer customers who have been customers for longer, and most customers have a subscription time of less than 50 months. There are some customers who have longer subscription times (more than 50 months), but the number is very small compared to the group around 36 months.

B. *Business Interpretation*

Banks may want to focus on customers with subscription times around 36 months as this is the largest group. This could be a potential target for increasing engagement or offering new products. Customers who have been subscribed for longer (more than 50 months) are a smaller group that may require special attention, especially if there are concerns about churn or customer loyalty.

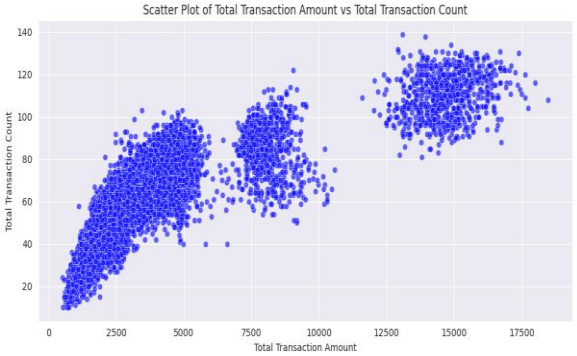

Fig. 4 Total Amount and Count Relationship in Segmentation

Customer Behavior Segmentation can be divided into two, namely High Frequency, Low Value Customers The group on the right shows customers who are actively transacting but with a small transaction value. They may be users who make routine or daily purchases, such as daily shopping. Low Frequency, High Value Customers The group on the left shows customers who rarely transact but when they do, the transaction value is high. They may be users who purchase expensive items regularly as shown in figure 4.

Visualization of correlation between Multivariate features using heatmap it will remove some features that are not needed, shown in Figure 5:

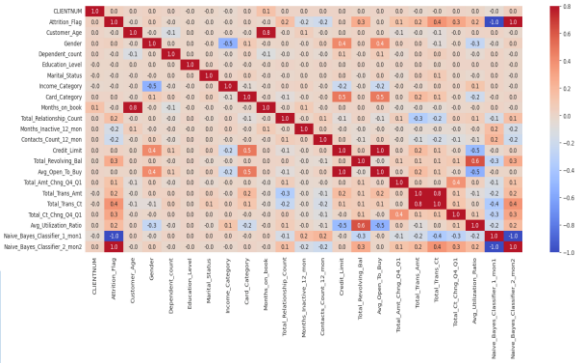

Fig. 5 Overall relationship between variables.

Total_Trans_Ct (Total Number of Transactions) There is a correlation of 0.4 between Attrition_Flag and Total_Trans_Ct, which shows a moderate positive relationship, indicating that the more transactions a customer makes, the less likely they are to churn and the more active customers tend to stay. Total_Trans_Amt (Total Amount of All Transactions) There is a correlation of 0.2 with Attrition_Flag, which indicates a positive relationship. Customers who spend more money through transactions are less likely to leave the service. Total_Relationship_Count has a correlation of 0.2 with Attrition_Flag, which indicates that the more interaction or relationship a customer has with the company (e.g. through products, services, or other contacts), the less likely they will churn.

The data analysis conducted on existing and attracted customers based on gender and marital status provides important insights into customer behavior and preferences. The following is the business interpretation generated from the following figure 6:

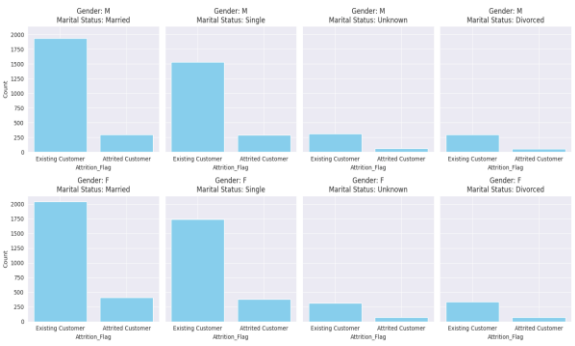

Fig. 6 Relationship between Attrition_Flag and marital status

Churn rate based on Marriage status

It can be seen that both married men and women have lower churn (attrited) rates than customers who are single or whose marital status is unknown. This suggests that married customers tend to be more stable in their relationship with the company, which can be interpreted as higher loyalty. Specifically in the "Married" category, there is a very high proportion of existing customers compared to attracted customers. This indicates that marital status plays a positive role in customer retention. While "Single" customers also show a similar pattern, their loyalty level is slightly lower.

The effect of marital status on churn

Interestingly, customers with "Unknown" and "Divorced" marital statuses tend to have higher churn rates, almost equal to the originality rate of existing customers in a given category. The results suggest that marital status uncertainty can be an important factor in customer retention. When customers' marital status changes, such as divorce, they may experience emotional and financial instability, which may affect their decision to stay in touch with the company.

Opportunities for customer segmentation

From the analysis conducted, it can be concluded that married customers show higher loyalty compared to other customer categories. This creates an opportunity for companies to better segment based on marital status. By understanding that married customers are more likely to remain loyal, companies can design more relevant products and services, specifically aimed at this segment.

The analysis of the distribution of existing and attrited (churned) customers by revenue category in Figure 7, provides valuable insights into the customer profile and strategies that the company can adopt. The following discussion outlines the key findings as well as the opportunities and challenges that arise from the data.

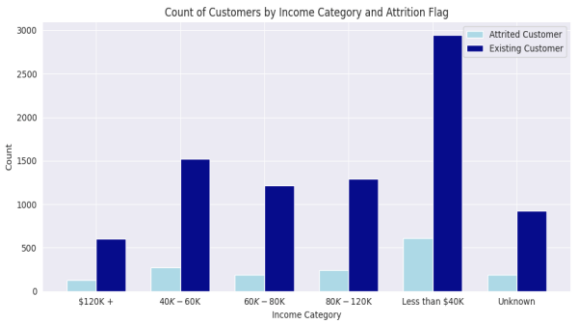

Fig. 7 income_category_with_attrition

Existing Customer Distribution

The data shows that customers with an income of less than 40K have the highest proportion in the existing customer category, much more than the other income categories. This indicates that the company has a strong customer base in the low-income segment, which may be a result of offering affordable products and services. The income category between 40K to 60K also shows a significant number of existing customers, although it is still below the low-income segment.

On the other, customers in the 120K+ and 80K to 120K income categories appear to be quite few. This indicates that the company has not been successful in attracting customers in the high-income segment, which could be an indication of an opportunity to improve marketing or products geared towards this segment.

Distribution of Attrited Customers (Churn)

Although the low-income segment has a high number of customers, the churn rate remains a concern, and companies need to develop strategies to retain customers in this category. Churned customers also tend to come from the "Unknown" category and the less than 40k income category. In contrast, higher income classes, such as 120k+ and 80k to 120k, show very low churn rates. This may indicate higher stability among high-income customers, or perhaps the company has not actively explored this market.

Opportunities in the Low-income Segment

The company can make cheap products and implement a strong loyalty program since the majority of its customers are from the low-income segment. This method has the potential to increase retention of existing customers and simultaneously attract new customers from comparable industries. Offering attractive discounts, referral programs, and manufacturing products that meet consumer needs and preferences are some of the strategies that can be implemented.

Challenges in the High-income Segment

The company struggles to attract customers from the high-income segment which has a very small existing and interested customer base. Developing premium products and services tailored to the needs of this segment can help expand penetration in this market. To discover what can attract high-income customers, more research needs to be done on their behavior and preferences.

Category "Unknown"

The presence of a customer in the "Unknown" category indicates that there is no information available about the customer's revenue. Identifying and collecting information about the revenue of these customers can help build better retention and acquisition strategies. Companies can improve the accuracy of their marketing and make offers that better suit the segment in question by collecting more complete information about their customers.

To make the analysis clearer and more concise, the following table summarizes the correlation between customer attributes and churn rates, along with corresponding insights and business implications.

Table. 3 Marital Status vs Churn

Marital Status	Churn Rate	Bussiness Insight
Married	Low	High loyalty, good retention opportunity
Single	Moderate	Slightly less loyal than married
Divorced / Unknown	High Churn	Potential emotional/financial instability

Table. 4 Income Category vs Churn

Income Category	Churn Rate	Bussiness Insight
< 40K	High existing, high churn	Price-sensitive, needs loyalty programs.
40K – 60K	Moderate existing, moderate churn	Growing segment, needs targeted offers.
80K – 120K / 120K+	Low churn, Low existing	Untapped premium market, opportunity to expand.
Unknown	Moderate Churn	Lack of data, needs better profiling.

Then we can also see in the card category where users are mostly in the Blue card type both in terms of churn and non churn which is found in Figure 8 :

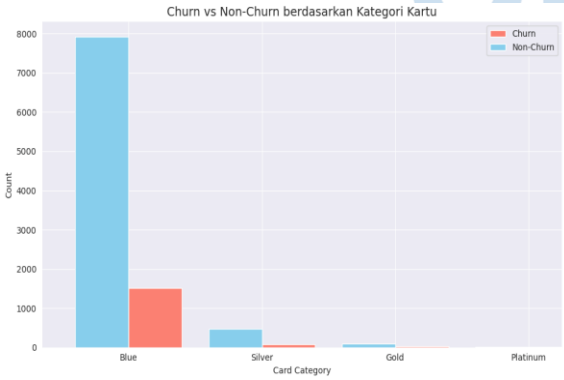

Fig. 8 Card Category

Comparison between churn and non-churn customers for various card categories. From the data shown, the Blue card has the highest number of customers with around 7,800 non-churn customers and around 1,500 churn customers. The Silver category comes in second with a much smaller number of about 500 non-churn customers and very few churn customers. Meanwhile, the Gold category has the least number of customers with only a few hundred non-churn customers and almost no churn. For the Platinum category, there is no significant customer data visible in the graph.

Interestingly, the data shows that the higher the card category, the lower the churn rate, which may indicate that higher category cardholders tend to be more loyal to the service. The Blue card, despite having the largest customer base, also shows a relatively higher churn rate compared to other categories.

IV. DISCUSSION

The performance analysis of the classification models used in this project is presented in this report. The three main metrics analyzed are Accuracy, Precision, and Recall. These metrics provide an in-depth look at the model's ability to correctly identify positive and negative classes.

A. Logistic Regression

In the precision section, it can be seen that Attirted Customer out of 327 data that is predicted by the prediction model, 100% is predicted by Attirted Customer. Existing Customer from 1699 data that is predicted by the prediction model, 84% is predicted by Existing Customer. In the Recall section, In the Recall section, Attirted Customer from 0% predicted Attirted Customer, 0% resulted correct. Existing Customer from 100% predicted Existing Customer, 100% resulted correct, but for Recall and F1 Score it is not good, so from the results it can be concluded that the model created is Not Good Fit. From both categories it can be seen that the resulting model performs poorly, especially in the Attirted Customer label Conclusion From the table results we can see that the model gets good performance with Accuracy on training data 0.8410072830514751 Accuracy Score 0.839091806515301 the full explanation is in table 5.

Table. 5 Classification Report Logistic Regression

	Precision	Recall	F1-Score	Support
Attrited Customer	1.00	0.00	0.01	327
Existing Customer	0.84	1.00	0.91	1699
Accuracy			0.84	2026
Macro avg	0.92	0.50	0.46	2026
Weighted Avg	0.86	0.84	0.77	2026

B. Random Forest

Based on the prediction results against the six labels listed in table 6, the model performed well with

all metrics exceeding 70% and testing accuracy of 90%, in accordance with the subsequent analysis, which concentrated on Precision, Recall, and F1 scores for the two main categories of customers: 'Interested Customers' and 'Existing Customers'. In the precision section of 327 data predicted as Attrited Customer, the model successfully predicted 100% as Attrited, this shows that there are no errors in determining attrited customers. Of the 1699 data predicted as Existing Customer, the model also successfully predicted one hundred percent with the same accuracy. This indicates that all customers identified as existing customers are existing customers.

In the Recall section the model managed to make 100% of all predictions made for the targeted customers, indicating that every targeted customer actually fell into the category, demonstrating the model's ability to find all positive cases. In addition, the model also successfully identified 100% of the predicted customers as existing customers, and all predictions in this category also proved to be correct, demonstrating the model's exceptional accuracy. F1-Score metric results show that the model performs well. This score, which combines Precision and Recall, gives an overall picture of the balance between these two metrics, and since all predictions are correct, this score will also show an excellent value.

Based on the analysis conducted, the model evaluation results are as follows Training Data Accuracy gets 1.0 (100%) while on Score Accuracy gets 1.0 (100%) From these results, it can be concluded that the tested model has excellent performance and high quality, and all metrics have perfect results. Therefore, the model created by the Random Forest algorithm is good fit.

Table. 6 Classification Report Random Forest

	Precision	Recall	F1-Score	Support
Attrited Customer	1.00	1.00	1.00	327
Existing Customer	1.00	1.00	1.00	1699
Accuracy			1.00	2026
Macro avg	1.00	1.00	1.00	2026
Weighted Avg	1.00	1.00	1.00	2026

However, this research has some limitations, including the datasets used may not be fully representative of real-world scenarios, as they are pre-processed and collected from specific sources. The memory and computation time requirements of the random forest model can be quite large especially if a large number of datasets are used to store the training data. Future research can overcome these limitations by studying larger datasets, improving model parameters, or integrating more machine learning algorithms for a more thorough analysis.

V. CONCLUSIONS

From Figure 5, it can be concluded that the influential features to create a model are the features Total_Trans_Ct (Total Number of Transactions):, Total_Trans_Amt (Total Amount of All Transactions):, Total_Relationship_Count (Total Number of Relationships with the Company):, Avg_Utilization_Ratio (Average Credit Usage):, Total_Revolving_Bal (Total Revolving Balance):, Months_Inactive_12_mon (Inactive Months in the Last 12 Months):, Contacts_Count_12_mon (Number of Contacts with the Company in the Last 12 Months):

From the Logistic Regression results we can see that the model performs well in train with 84% accuracy and 83% validation. From the test data, we can create a classification report in table 3. From table 3 we can see the score of the model against the 3 metrics generated by the classification report, judging from the results it can be concluded that the model created does not have a good value especially in the Recall and F1 Score sections.

On the other hand, from the Random Forest results, we can see that the model gets good performance in training with 100% accuracy and 100% validation, this good performance is also proven by evaluating the model using the prepared test data. From the test data, a classification report can be made in table 4. From table 4 we can see the score of the model against the 3 metrics generated by the classification report, judging from the results it can be concluded that the model made is Good Fit. So that the model used is the result of Random Forest.

REFERENCES

[1] Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. Data Science and Management, 7(1), 7–16. Https://doi.org/10.1016/j.dsm.2023.09.002

[2] Zhang, H., & Zhang, W. (2024). Application of GWO-attention-convlstm model in customer churn prediction and satisfaction analysis in customer relationship management. Heliyon, 10(17). Https://doi.org/10.1016/j.heliyon.2024.e37229

[3] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1). Https://doi.org/10.1186/s40537-019-0191-6

[4] Xiahou, X., & Harada, Y. (2022). B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 458–475. Https://doi.org/10.3390/jtaer17020024

[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

[6] Verma, P. (2020). Churn prediction for savings bank customers: A machine learning approach. Journal of Statistics

Applications and Probability, 9(3), 535–547. Https://doi.org/10.18576/JSAP/090310

[7] Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization, 14. Https://doi.org/10.1016/j.rico.2023.100342

[8] Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. International Journal of Intelligent Networks, 4, 145–154. https://doi.org/10.1016/j.ijin.2023.05.005

[9] Dhangar, K. (2021). A review on customer churn prediction using machine learning approach. In novateur publications international journal of innovations in engineering research and technology (Vol. 8, Issue 5).

[10] Khattak, A., Mehak, Z., Ahmad, H., Asghar, M. U., Asghar, M. Z., & Khan, A. (2023). Customer churn prediction using composite deep learning technique. Scientific Reports, 13(1). Https://doi.org/10.1038/s41598-023-44396-w

[11] 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. Https://doi.org/10.3390/jtaer17040077

[12] Afolabi, S., Ajadi, N., Jimoh, A., & Adenekan, I. (2025). Predicting diabetes using supervised machine learning algorithms on E-health records. Informatics and Health, 2(1), 9–16. Https://doi.org/10.1016/j.infoh.2024.12.002

[13] Shah, D., Xue, Z. Y., & Aamodt, T. M. (2022). Label Encoding for Regression Networks. Http://arxiv.org/abs/2212.01927

[14] Hamzat, A. K., Salman, U. T., Murad, M. S., Altay, O., Bahceci, E., Asmatulu, E., Bakir, M., & Asmatulu, R. (2025). Development of robust machine learning models for predicting flexural strengths of fiber-reinforced polymeric composites. Hybrid Advances, 8. Https://doi.org/10.1016/j.hybadv.2025.100385

[15] Setiawan, E., Azis Suprayogi, M., Kurnia, A., Setiawan, E., Suprayogi, M. A., & Kurnia, A. (2025). BAREKENG: Journal of Mathematics and Its Applications a comparison of logistic regression, mixed logistic regression, and geographically weighted logistic regression on public health development in java a comparison of logistic regression, mixed logistic regression and …. Issue 1 j. Math. & app, 19(1), 129–0140. Https://doi.org/10.30598/barekengvol19iss1pp0129-0140

[16] Chen, P., Xiong, H., Cao, J., Cui, M., Hou, J., & Guo, Z. (2025). Predicting postoperative adhesive small bowel obstruction in infants under 3 months with intestinal malrotation: a random forest approach. *Jornal de Pediatria*. Https://doi.org/10.1016/j.jped.2024.11.011