EARLY WARNING SYSTEM FOR DEBT GROUP MIGRATION: THE CASE OF ONE COMMERCIAL BANK IN VIETNAM

Quoc Hung NGUYEN[a)], Hoang Viet TRINH[b)], Truong Viet PHUONG[c)], Truong Thi Minh LY[d)]

University of Economics Ho Chi Minh City, Ho Chi Minh City, VIETNAM
[a)]e-mail: hungngq@ueh.edu.vn, ORCID: 0000-0002-6363-0160
[b)]e-mail: viet.th@vnp.edu.vn, ORCID: 0009-0000-4454-5000
[c)]e-mail: truong@ueh.edu.vn, ORCID: 0000-0001-8330-2714
[d)]e-mail: truongly@ueh.edu.vn, ORCID: 0000-0003-0025-7626

Abstract: This study utilizes machine learning models, including Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest, in the early warning system for debt group migration in a Vietnamese commercial bank. In predicting customers' overdue debt migration (B Score), the RF model achieves the highest accuracy of 81.84%. However, if the priority is to reduce Type I errors, SVM performs better with a recall of 91.48%, although the accuracy drops to 46.62%. When predicting customers' debt group improvement (C Score), SVM proves to be the optimal model in terms of both accuracy and criteria based on Type II errors, with an accuracy of 71.6% and precision of 62.3%. When applied to new datasets, the evaluation criteria decrease, but SVM remains the most optimal model for both B Score and C Score. Additionally, the research results demonstrate that tuning the model parameters leads to a significant improvement in accuracy compared to the default parameters.

Keywords: machine learning models, debt group migration, B score, C score, model parameters tuning.

JEL Classification: E50, E51, G33, G21, G24.

1 Introduction

Credit is one of the most crucial activities for commercial banks to generate the highest profits. Through credit operations, banks provide loans to customers, enabling them to engage in business activities, make purchases, invest, and consume. However, like any financial activity, credit also entails risks. Credit risk is defined as a situation where customers fail to fulfill or lack the ability to fulfill their debt repayment obligations as agreed upon in the contract or agreement with the bank (Bielecki and Rutkowski, 2013; Bluhm, et al., 2016). It can occur when customers default, when customers face financial difficulties, or due to external factors such as changes in government policies, economic recessions, natural disasters, or wars. In such cases, banks incur financial losses and experience negative impacts on their operations.

To ensure the safety and enhance the profitability of customer deposits, banks must invest heavily in researching and developing specialized departments to effectively manage credit risks. These departments are responsible for examining and assessing the repayment capacity of customers, identifying influencing factors, proposing appropriate solutions to minimize risks, and implementing preventive measures in case risks occur. Accurately predicting credit risks plays a significant role in preventing default situations and potential economic losses (Djeundje, et al., 2021; Dahooie, et al., 2021).

In the credit risk management, early warning systems for risks play a crucial role in minimizing losses for financial institutions. These systems provide information on changing trends in the credit status of customers and companies. The use of these warning systems helps financial institutions and investors assess and manage credit risks with greater accuracy and efficiency. For instance, the early warning system for credit risks in Chinese manufacturing enterprises achieves an accuracy rate of up to 87.29% (Wang and Zhang, 2023), and the early warning system in South Korea can explain the default risk of mortgage loans based on macroeconomic factors (Kwon and Park, 2023).

One of the most widely used early warning systems is the credit rating transition alert system developed by Standard & Poor's (S&P). This system relies on financial, business, and other factors to provide forecasts on

the potential changes in the credit rating of a customer or corporation. These forecasts are regularly updated to assist financial institutions and investors in rapidly and effectively assessing and managing credit risks. In addition to S&P, financial institutions and investors also utilize other credit rating transition alert systems such as Moody's and Fitch Ratings. These systems also rely on financial, business, and other information to provide forecasts on the potential changes in the credit rating of a customer. The credit rating transition alert systems are developed to aid financial institutions in accurately and efficiently assessing and managing credit risks. Specifically, these systems support the early detection and warning of potential credit rating transitions, enabling financial institutions to make informed and timely decisions in credit risk management. Research studies by Kim and Sohn (2008), Figlewski, et al. (2012), Forster, et al. (2016), and Slapnik, et al. (2021) all highlight the role of credit rating transition alert systems.

In Vietnam, Circular No. 11/2021/TT-NHNN (Circular 11) issued by the State Bank of Vietnam (SBV) plays a crucial role in risk management and ensures the safety and stability of operations for credit institutions and foreign bank branches in Vietnam. This Circular is aimed at regulating, updating, and enhancing the regulatory framework for asset classification, provisioning for risks, and the use of provisions to handle risks in the activities of credit institutions. Regarding asset classification, this Circular has provided guidelines for classifying the assets of credit institutions into corresponding groups based on different risk levels, ranging from non-risk assets to high-risk assets. This helps credit institutions assess the risk level of their assets and implement appropriate measures to manage and mitigate risks.

In this article, we use three machine learning models, namely Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF), to predict the migration of debt groups based on Circular 11 issued by the SBV. To achieve improved results, we also adjust the parameters of these machine learning models instead of relying on default settings, ensuring higher accuracy in their operation. This tuning process contributes to enhancing prediction capability and minimizing errors. Additionally, we compare the performance of the machine learning models with Logistic Regression, a fundamental classification model. This comparison allows us to evaluate the flexibility and effectiveness of the employed machine learning models.

2 Literature review

2.1 Types of credit quality assessment

Credit quality assessment is an important aspect in the core business of a bank, as it allows banks to assess the repayment ability of borrowers and determine their loan repayment capacity. Credit scores are a digital representation of a borrower's credit history and help lenders assess the risk level of borrowers. There are different types of credit scores, each providing specific information to bank managers about the borrower's credit history and behavior. This study will discuss three typical types of credit scores: Application Score (A Score – credit score for new customers), Behavioral Score (B Score – credit score regarding the repayment behavior of existing customers), and Collection Score (C Score – credit score regarding the debt recovery ability of defaulted customers).

The first type of credit score is the A Score, which is based on information provided by the borrower on the credit application. This score considers key factors such as the borrower's income, employment history, credit history, and debt-to-income ratio. Banks use this score to assess the borrower's repayment ability and decide whether to approve the credit application. The A Score is crucial in evaluating the borrower's initial debt repayment capability.

The second type of credit score is the B Score, which reflects the repayment behavior of customers and is also known as the performance credit score. This score is based on the borrower's credit behavior during the credit relationship with the bank, such as payment history, credit utilization, length of credit usage, and types of credit used. The B Score provides banks with information about the borrower's credit habits and their ability to repay debts on time. This score is essential in assessing the borrower's credit risk over an extended period.

The third type of credit score is the C Score, which is used to assess the debt recovery ability of borrowers and is typically applied to customers showing signs of default.

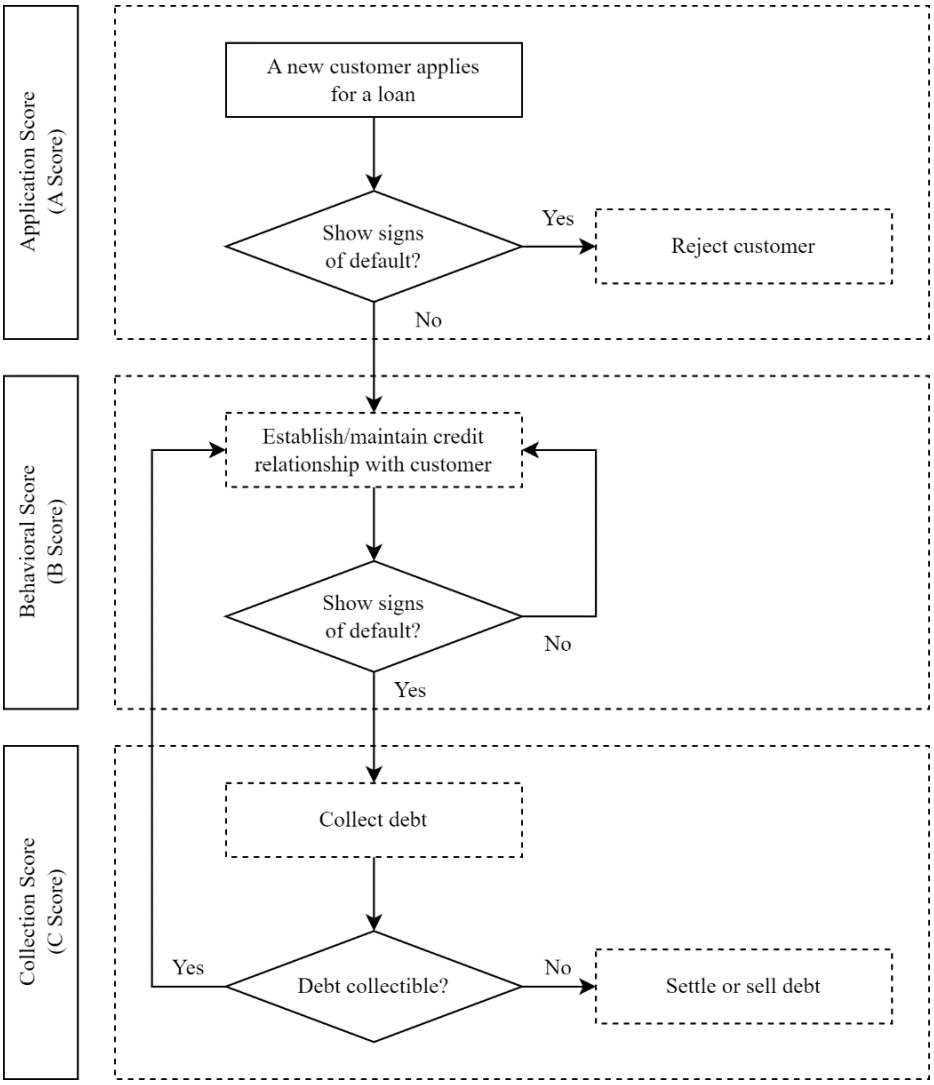

Figure 1. The lifecycle of a customer's credit relationship with a bank.
(*Source:* Author's own research)

The C Score is based on monitoring overdue or non-payment debts and the collection activities related to resolving these debts. It allows the bank to evaluate the borrower's ability to handle overdue debts and bad debts and decide whether to extend credit to the borrower. The collection credit score is an important indicator to assess the borrower's ability to manage and resolve debts as well as their ability to recover to a good credit status.

All three types of credit scores can reflect the lifecycle of a customer's credit relationship with the bank. According to Figure 1, a new customer submitting a credit application will be scored based on the A Score. If approved, the customer is granted credit and begins the repayment process with the bank. This process is monitored by the B Score. If the customer defaults, the bank will initiate collection efforts and assess the debt recovery ability through the C Score.

Within the scope of this study, an early warning system for the shift in the debt group has a close relationship with the B Score and C Score. The early warning system needs to provide signals indicating existing customers at risk of default in the future based on observable credit behaviors at present, thereby enabling the bank to proactively deal with potential losses and implement provisions for bad debts. Furthermore, for customers who have defaulted or have overdue debts, the system also helps identify customers from whom the bank can recover debts and restore the good debt group, thus providing timely support measures for customers.

2.2 Definition of credit impairment and credit improvement

According to Circular 11, customer loans are classified into risk categories as follows:

- Standard debt (Group 1 debt): Debt within the term and assessed as having the ability to recover both principal and interest on time; debt overdue for less than 10 days but assessed as being able to recover the overdue portion, with the remaining portion recoverable on time.
- Attention-required debt (Group 2 debt): Debt overdue from 10 to 90 days; debt with the first adjustment of repayment term.
- Substandard debt (Group 3 debt): Debt overdue from 91 to 180 days; debt with the first extension; debt waived or reduced in interest due to exceeding the repayment capacity under the credit contract.
- Doubtful debt (Group 4 debt): Debt overdue from 181 to 360 days; debt restructured with the first extension overdue for less than 90 days; debt restructured with the second extension.
- Potentially irrecoverable debt (Group 5 debt): Debt overdue for more than 360 days; debt restructured with the first extension overdue for 90 days or more; debt restructured with the second extension and overdue; debt restructured with the third extension or more.

The debt groups from 1 to 5 represent a decreasing order of credit quality, and based on the factors of overdue days, there can be a phenomenon of debt groups potentially shifting to better or worse groups. For example, if a customer fails to repay the debt within 3 months (> 90 days), it will be classified as Group 3 debt. However, if the customer makes full payment of principal and interest for one month, the remaining overdue debt will be below 90 days and will be classified as Group 2 debt.

Based on these foundations, we proposed the following criteria to develop an early warning system:

- The evaluation period for customer debt group transitions compared to the reporting date is 12 months.
- Debt showing signs of overdue will be classified as Group 2 or higher.

Regarding credit quality improvement, we have not found specific criteria for precise evaluation. However, most banks will recognize customer credit quality improvement based on the following two factors:

- The minimum period from the customer's default to returning to a non-default status, typically 3–6 months.
- The minimum period the customer maintains a non-default status after returning to it, typically 3–6 months.

Therefore, if a customer meets both of the above criteria within a total period of 6–12 months, the bank will recognize the improvement in the customer's credit quality.

2.3 Early warning models for credit risk across the world

To achieve the objective of forecasting or warning with the smallest possible error, early warning models for credit risk need to be built on solid statistical and economic theories. Various studies have proposed multiple methods for constructing high-reliability early warning models. However, there are two most common methods being applied in many countries: the signal method and the quantitative method.

2.3.1 The signal method

The signal method is one of the popular approaches to building early warning systems. This method assumes that there is a correlation between indicators and credit risk. When the indicators surpass the warning threshold, signals are emitted to indicate the potential risks of default, interest rate risk, or other credit risks. The process of selecting indicators can be based on economic theories, insights from managers, or past experiences through analyzing previous crises. According to several studies (Reihart, et al., 2010; Ionela, 2014; Ha, 2019; Kwon, 2023), there are three main groups of indicators that often appear in early warning models and produce the best forecasting results: customer financial indicators, customer behavioral indicators, and external environment indicators.

Financial indicators such as cash flow, collateral assets, and profitability play a crucial role as reliable

tools for banks to warn about the risk of default from customers. However, conventional financial indicators often have fixed update times with a low frequency, and thus, their early warning nature may not be as effective as behavioral indicators. The mechanism of recording behavioral data will continuously track the customer's credit relationship, such as timely debt repayment, early debt payment, or cash flows in and out of the customer's account. Both financial and behavioral indicators reflect the characteristics of the customer. However, the difference lies in the data recording method, where financial indicators are usually provided by the customer or independent third parties (such as auditing firms, appraisers), while behavioral indicators are continuously recorded by the bank from the customer's activities.

In terms of the purpose of these two types of indicators, financial indicators are better used for customer segmentation and risk optimization of credit portfolios, while behavioral indicators are used to quickly identify anomalies in customer behavior and provide early warnings of potential risks.

External environment indicators also contribute to credit risk warning. However, these warnings are of a systemic nature. A negative signal in external environment indicators can affect a large number of customers rather than specific individuals or groups. External environment indicators can be examined at different levels ranging from industry-specific indicators, regional indicators, to macroeconomic indicators at the national level, such as economic growth, real estate prices, inflation, interbank interest rates, and even globally influential measures like oil prices, gold prices, and the growth rates of major economies worldwide.

2.3.2 The quantitative method

Quantitative method, similar to the indicator method, assumes that the relationship between indicators and credit risk can be quantified. There are two approaches to the quantitative method:

- Regression method: This method is used when credit risk is a continuous variable. For example, credit risk can be measured by the ratio of non-performing loans, the ratio of defaulting customers, etc. The basic model in this method is Linear Regression.
- Classification method: This method is used when credit risk is a discrete variable. For example, it can be the default or non-default status of customers, customer rating groups, etc. The basic models in this method are Logistic Regression and Ordinal Logistic Regression.

With the rapid emergence of data, there have been significant improvements in these methods. Notable advancements include tree-based models such as Decision Trees, Random Forest, XGBoost, CatBoost, as well as deep learning models.

2.3.3 Evaluation of an early warning model

To achieve optimization, the alert threshold is set to minimize the weighted average value of both Type I and Type II errors. In each stage, if the observed results of the indicators exceed the threshold and fall into the danger zone, a warning signal will be issued (Reihart, et al., 2010). The alert signals for each indicator are divided into four types as follows.

According to Table 1, there are a total of 4 types of alerts as follows:

- A: Alert signal is issued and the event occurs. This is a true alert.
- B: Alert signal is issued but the event does not occur. This is a false alert, belonging to Type II error (also known as false positive): when the alert system predicts that the event will occur but in reality, but it does not.

Table 1. Matrix for different types of alerts
(*Source:* Reihart, et al., 2010)

-	The event occurred	The event did not occur
There is a warning signal	A	B
There is no warning signal	C	D

- C: No alert signal is issued but the event occurs. This is a missed alert, belonging to Type I error (also known as false negative): when the system does not issue any alert signal, but the event occurs.
- D: No alert signal is issued and the event does not occur. This is a true non-alert.

In the signal method, the trade-off between the two types of errors can be observed when setting the alert threshold. If the threshold is set too low to issue alert signals, it will increase the Type II error, leading to unnecessary decisions and costs. However, if the threshold is set too strict, it can result in events occurring without being alerted in advance, causing damages due to the lack of timely coping measures.

In the quantification method, an equivalent matrix is also evaluated as a confusion matrix for correctly or incorrectly predicting the event. In this case, Recall criterion is related to Type II error, and Precision criterion is related to Type I error (Fawcett, 2006).

2.4 Machine learning models

2.4.1 Logistics Regression (LG)

This is the most fundamental model. To gain a better understanding of logistic regression, we can start with a linear regression model in the following form:

$$y = w^T x$$

where y represents the dependent variable, also known as the target variable; x represents the independent variables; and w^T is a matrix of weights corresponding to each independent variable, also known as regression coefficients. Linear regression models are typically used when y is a continuous variable with a defined range $(-\infty; +\infty)$.

However, in the context of this study, the target variable is a binary variable with two values, 0 and 1. If we still apply the linear regression model (Gujarati and Porter, 2009) in this case, the predicted outcome y (denoted as $\hat{y}$) will fall into two scenarios:

- If $\hat{y} \in (0; 1)$, it can be interpreted as the probability of y = 1. This idea corresponds to the Linear Probability Model (LPM).
- If $\hat{y} < 0$ or $\hat{y} > 1$, it cannot be interpreted as a probability. If we constrain the values of $\hat{y}$ to the range $(0; 1)$, it would result in losing the interpretation of the slope coefficients in the linear regression model.

To address this issue, the component $w^T x$ will be passed through an activation function to transform its value range. At this point, the model can be rewritten as follows:

$$z = f(w^T x)$$

The notation z replaces y because y is a binary variable with only two values, while z is a variable with a different value range depending on the type of activation function.

If the activation function transforms w^T x from the value range $(-\infty; +\infty)$ to (0;1), then the model above becomes the logistic regression model. The commonly used activation function is the Sigmoid function (specifically, the Logit function), and the logistic regression model is written as follows:

$$z = \frac{1}{1 + e^{-w^T x}}$$

When the predicted outcome from the model, $\hat{z}$, falls within the range $(0; 1)$, it represents a probability. The probability threshold is typically chosen as 0.5, so if $\hat{z} \geq 0.5$, then $\hat{y} = 1$, and if $\hat{z} < 0.5$, then $\hat{y} = 0$.

2.4.2 Support Vector Machine (SVM)

The SVM model is constructed by creating a plane (or hyperplane) to separate the data in a way that optimizes the distance between data points and the separating plane. In general, the distance of a point with coordinates x_0 to the hyperplane with the equation $w^T x + b = 0$ is determined as follows:

$$\frac{|w^T x_0 + b|}{\|w\|_2}$$

The objective of SVM is to find the values of *w* and *b* such that the distance from the separating plane to the nearest data point is maximized. Therefore:

$$(w, b) = \arg\max \left\{ \min \frac{y_n (w^T x_n + b)}{\|w\|_2} \right\} =$$

$$= \arg\max \left\{ \frac{1}{\|w\|_2} \min y_n (w^T x_n + b) \right\}$$

The optimization problem can be simplified as follows (Bishop, 2006):

$$(w, b) = \arg\min \frac{1}{2}\|w\|_2^2$$

$$\text{subject to: } 1 - y_n(w^T x_n + b) \leq 0, \forall n = 1,2, \dots, N$$

where $y_n = 1$ or $y_n = -1$ depending on the binary label of data point *n*.

In practice, data are often difficult to perfectly separate linearly. As a result, there will be the occurrence of noisy data points. The optimization problem of SVM is then rewritten as follows:

$$(w, b) = \arg\min \frac{1}{2}\|w\|_2^2 + C\sum_{n=1}^{N} \varepsilon_n$$

$$\text{subject to: } 1 - \varepsilon_n - y_n(w^T x_n + b) \leq 0, \forall n = 1,2, \dots, N$$

$$-\varepsilon_n \leq 0, \forall n = 1,2, \dots, N$$

where *C* is the weight that determines the trade-off between optimizing the distance and handling noise. The ε_n of a data point reflects the level of noise:

- If the data point is far from the separating plane and correctly classified, then $\varepsilon_n = 0$.
- If the data point is close to the separating plane but still correctly classified, then $0 < \varepsilon_n < 1$.
- If the data point is misclassified, then $\varepsilon_n > 1$.

Another issue is that the data may not be linearly separable. However, it is possible to use certain activation functions to map the data to a different feature space where it can be linearly separable. These activation functions are called kernel functions. Common kernel functions are defined as follows:

- Linear kernel: $k(x, z) = x^T z$
- Polynomial kernel: $k(x, z) = (r + \gamma x^T z)^d$
- Radial Basic Function (RBF) kernel: $k(x, z) = e^{-\gamma\|x-z\|_2^2}, \gamma > 0$
- Sigmoid (tanh) kernel: $k(x, z) = \tanh(\gamma x^T z + r)$

For these kernel functions, *r* and γ are parameters that can be tuned. In the case of the Polynomial kernel, an additional parameter d represents the degree of the polynomial function (*d* is a natural number). It is worth noting that the Linear kernel does not have any adjustable parameters.

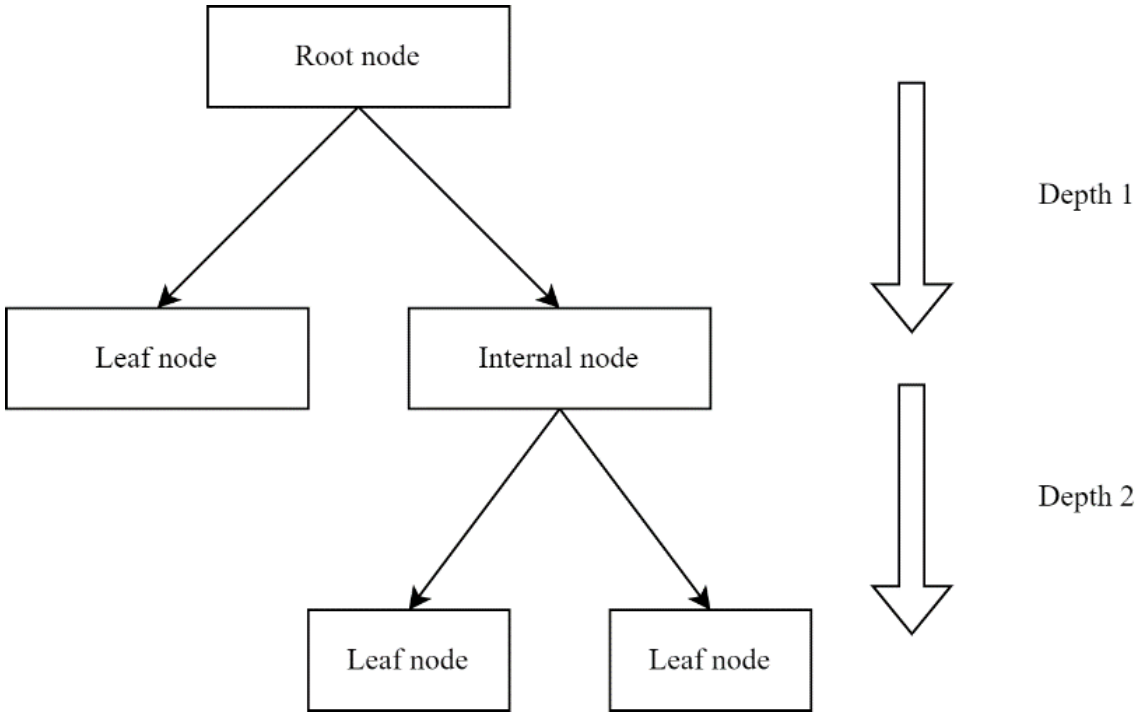

Figure 2. A decision tree classifier
(*Source:* Author's illustration)

2.4.3 Decision Tree (DT)

Decision tree is a classification model that uses variables to divide data into smaller samples. These samples are further divided in a similar manner until the labels of the data (target variable) can be predicted. Some related concepts in DT, as shown in Figure 2, include:

- Root node: It is the first node in the DT, where the data are not yet divided.
- Internal node: These are intermediate nodes in the DT. They are the result of the division at the previous node and serve as data for further division by a specific variable.
- Leaf node: It is the final node that is no longer divided by any variable. At the leaf node, the DT needs to make a prediction for the label of the data.
- Depth: Each division in the DT forms a depth in the tree, indicating the level of divisions.

At each internal node, the DT decides whether to continue dividing the data and if so, which variable to choose. Generally, the division process stops if the node contains data points with the same label (pure node). However, this may result in a very deep tree or the inability to find a variable for division. Therefore, limiting the depth and the number of leaf nodes in the DT is crucial.

The next issue is determining which variable to use for division. In DT, there are two basic criteria used to evaluate variables: Entropy and Gini (Rokach and Maimon, 2008; Shalev-Shwartz and Ben-David, 2014).

- Entropy at any given node is calculated as follows:

$$\text{Entropy} = -\sum_{i} p_i \times \log_2 (p_i)$$

Entropy has a minimum value of 0, indicating a pure node where all data points have the same label. The larger the entropy, the more diverse the distribution of labels in the node.

- Gini at any given node is calculated as follows:

$$\text{Gini} = 1 - \sum_{i} {p_i}^2$$

Gini also has a minimum value of 0. Similar to entropy, when Gini = 0, it indicates a pure node, and as Gini increases, the distribution of labels in the node becomes more diverse.

To select the variable for division, the Entropy (or Gini) is computed for the initial node before the division. Then, the Entropy (or Gini) is calculated for each node after the division. The information gain of a variable is determined as follows:

$$IG = E_{Initial} - \sum_{n} w_n E_n$$

where $E_{Initial}$ is the Entropy (or Gini) at the initial node, E_n is the Entropy (or Gini) of node *n* after the division, and w_n is the weight of the number of observations in node *n* relative to the number of observations at the initial node. Each variable will have a different information gain, and the variable with the highest information gain is chosen for the division.

2.4.4 Random Forest (RF)

Random forest is a model that consists of multiple decision tree (DT). On a dataset, RF performs random sampling with replacement to create n data samples, each with a smaller or equal size compared to the original sample. For each obtained data sample, a DT model is built and used to predict the entire original dataset. RF includes n constructed DTs and utilizes a voting mechanism to determine the label for each data point. Thus, the essence of RF lies in multiple DTs, where RF has the additional task of labeling through each DT and aggregating the results to obtain the final label assignment.

Due to the execution of multiple DTs, the number of DTs should be carefully considered when implementing RF to ensure computational efficiency, especially for large datasets. This is an important parameter of RF. Other tuning parameters for RF are similar to those of DT, such as the maximum number of leaf nodes and the depth of the tree. RF also has adjustments regarding the sample selection for conducting the DTs.

2.4.5 Parameters tuning

In general, each model has its own advantages to achieve the goal of classifying data. Fine-tuning the models is necessary to maximize the benefits of each model. Table 2 summarizes the important parameters that need to be adjusted to ensure the optimal performance of the model.

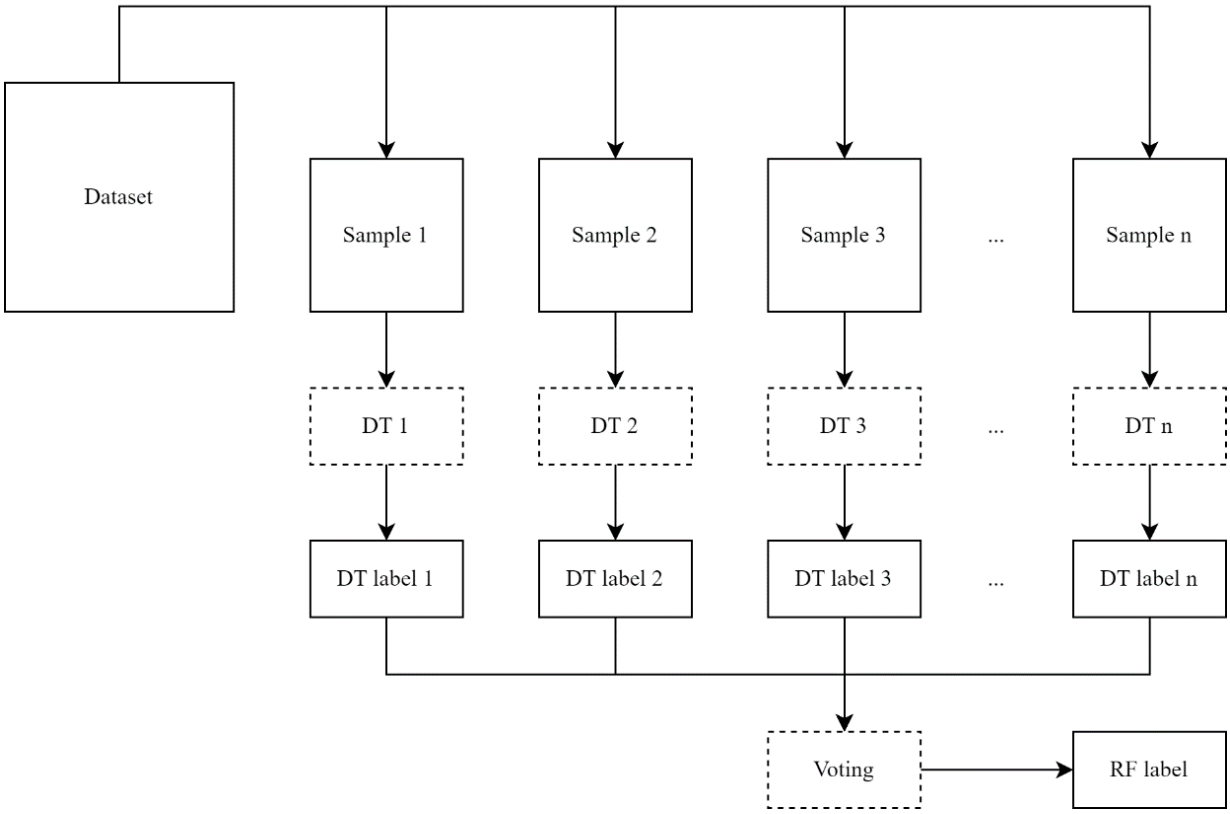

Figure 3. A random forest classifier
(*Source:* Author's illustration)

Table 2. Summary of parameters of models
(*Source:* Author's compilation)

Model	Parameter	Description
LG	None	The baseline model is a linear regression model combined with the sigmoid (logit) activation function, so no tuning is required
SVM	Kernel function	The activation function used to transform data into a different feature space for linear separation includes Linear, Polynomial, Sigmoid, and RBF
	C	The coefficient for balancing the weight between distance and noise
	d	The degree parameter when using the Polynomial kernel, which takes a natural number value
	γ	The gamma parameter for Polynomial, Sigmoid, and RBF kernels, which takes a non-negative value
	r	The intercept for the Polynomial and Sigmoid kernels
DT	Depth	It is necessary to limit the depth of the DT to avoid overfitting and reduce computational cost
	Number of leaf nodes	It is necessary to limit the number of leaf nodes of the DT to avoid overfitting and reduce computational cost
RF	Depth	It is necessary to limit the depth of each DT to avoid overfitting and reduce computational cost
	Number of leaf nodes	It is necessary to limit the number of leaf nodes of each DT to avoid overfitting and reduce computational cost
	Number of DTs	The number of DTs in Random Forest Classifier (RF) needs to be considered for computational cost when the number is too high

3 The proposed model

The three-stage architecture of the proposed model is shown in Figure 4.

The credit database used in this article consists of data related to customer transaction behavior on their payment and savings deposit accounts at a commercial bank in Vietnam. Additionally, information on collateral assets and early repayment behavior is also recorded as input for the early warning model. From this database, credit information and customer behavior data are preprocessed and divided into two datasets: (1) historical data used for model building and (2) current data used for assessing the early warning system. Among them, the historical data are further divided into two sets: the training set and the evaluation set. The training set is used to estimate the model, while the evaluation set is used to fine-tune the model. Subsequently, the optimized model is selected based on evaluation criteria considering the importance of two types of alert errors. Then, the selected model is deployed on the current dataset. After a period of time, the early warning system will continue to be evaluated for its accuracy in detecting shifts in debt groups among customers in the current dataset.

Based on the theoretical basis of credit impairment and improvement definition and considering the practical situation, the binary target variable is created to build the model based on the current debt group and the highest debt group within the next 12 months. Subsequently, the data are filtered and divided into two following sets:

- The dataset of customers with the current debt group as group 1 (good debt) will have a target variable value of 1 if the highest debt group within the next 12 months is greater than 1. These customers are referred to as B Score.
- The dataset of customers with the current debt group as groups 2 to 5 (delinquent and bad debt) will have a target variable value of 1 if the highest debt group within the next 12 months is smaller than the current debt group. These customers are referred to as C Score.

This study will not reflect the idea of the A Score because the research objective does not target new customers but only focuses on existing customers.

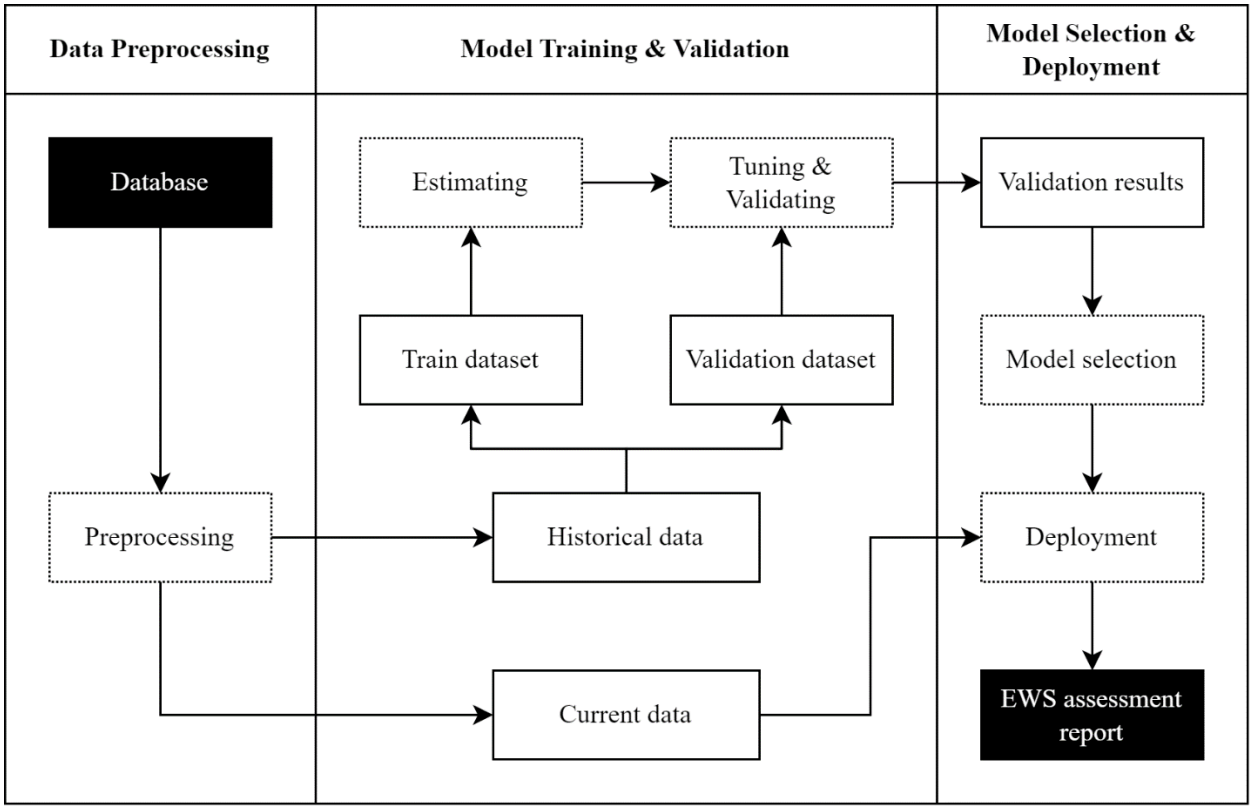

Figure 4. A random forest classifier
(*Source:* Author's own research)

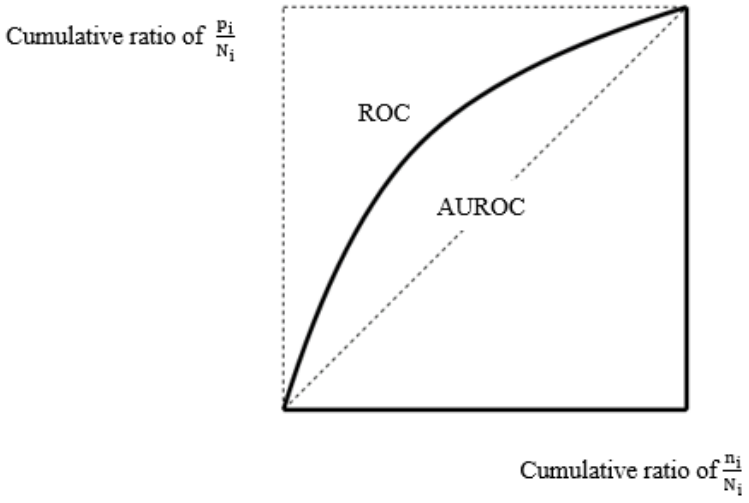

Figure 5. ROC and AUROC
(*Source:* Author's illustration)

3.1 Data preprocessing

During the data preprocessing phase, features are created to enhance the model's interpretability. We employ three transformation techniques as follows:

- Differential transformation: This technique captures the changes compared to the previous period. The time frames for change are 1, 2, and 3 months.
- Ratio transformation: It represents the percentage ratio instead of absolute numbers. Ratios are calculated relative to the outstanding loan balance.
- Logarithmic transformation: This transformation is used to reshape the distribution of the data into a bell-shaped curve.

With the transformed variables, we proceed to perform binning for each variable and calculate the Weight of Evidence (WoE) and Information Value (IV) values (Greiff, 1999; Siddiqi, 2006) as follows:

$$\mathrm{WoE_i} = \ln\left(\frac{\frac{n_i}{N_i}}{\frac{p_i}{N_i}}\right)$$

$$\mathrm{IV} = \sum_{1}^{k}\left[\mathrm{WoE_i} \times \left(\frac{n_i}{N_i} - \frac{p_i}{N_i}\right)\right]$$

where *i* denotes the bin, *k* represents the total number of bins, n_i indicates the number of observations with target variable equal to 0 in bin *i*, p_i represents the number of observations with target variable equal to 1, and N_i represents the total number of observations in bin *i*.

In addition, the GINI criterion is also determined based on the area under the receiver operating characteristics (AUROC) using the following formula:

$$\mathrm{GINI} = 2 \times \mathrm{AUROC} - 1$$

where AUROC is the area under the ROC curve illustrated in Figure 5.

The author conducted a Delphi survey with over 300 DX experts in SMEs in Vietnam using the Foresight approach to assess the features, challenges, and business vision of DX. The study also forecasts the future demand for DX in SMEs and the most important activities for implementing it. To calculate the sample size and determine the reliability level, the author used Yamane's (1967) sampling technique as a reference. The findings of the analysis were presented in detail in the study "Building a Digital Transformation Strategy for SMEs in Vietnam – Approaching Foresight" (Tuan, et al., 2023).

Most businesses need government assistance and tools during the DX process. In terms of support tools, most businesses have not used a comprehensive support system and desire one with the following capabilities.

The result of each bin corresponds to a WoE value. Assigning this WoE value to the corresponding bins in the original dataset will yield a series of WoE variables. These WoE variables also play a significant role in predicting the target variable of the models.

IV and GINI are used to measure the predictive power of features. IV and GINI are directly proportional to predictive ability. Since there are many additional

variables created, including all of them in the model can increase computational cost. Therefore, it is necessary to filter out variables with a good predictive power. We perform variable filtering based on the IV criterion (Siddiqi, 2006) with the following thresholds:

- <0.02: Not recommended for use in the model.
- [0.02, 0.1): Weak predictive power can be used in the model.
- [0.1, 0.3): Moderate predictive power can be used in the model.
- [0.3, 0.5): Strong predictive power can be used in the model.
- ≥0.5: Suspected variable for determining the target variable.

Therefore, this study will choose the IV threshold of ≥0.02 to limit the number of variables used in model estimation.

3.2 Model training and validation

The data used for model estimation are historical data divided into a training set and an evaluation set. The data splitting is performed using stratified sampling based on the target variable labels, with a ratio of 70% for training and 30% for evaluation. The models are estimated using the Python libraries of Scikit-learn. The model tuning parameters are described in Table 3.

Another important parameter that needs to be declared in all models is *class_weight* = 'balanced'. This is an important parameter to give higher weight to the target variable = 1 labels that are correctly predicted, as there is an imbalance in the data.

Model tuning will be based on optimizing the accuracy of predictions, specifically, creating a confusion matrix and calculating relevant metrics. According to Fawcett (2006), a confusion matrix reflects the true and false similarities between the predicted results and the actual data. Therefore, a confusion matrix is a 2×2 matrix with columns representing the number of model-predicted observations and rows representing the number of actual observations, as follows in Table 4.

Table 3. Model tuning parameters in Scikit-learn
(*Source:* Author's own research)

Model	Parameter	Parameters in Scikit-learn	Range of values for tuning
LG	None	None	None
SVM	Kernel function	*kernel:* accepts a value from 'linear', 'poly', 'rbf', 'sigmoid'. The default value is 'rbf'	'linear', 'poly', 'rbf', 'sigmoid'
	C	*C:* data type is float; the default value is 1	0.01, 0.1, 1, 10
	d	*degree:* data type is integer; the default value is 3	2, 3, 4, 5 (this is applicable only when *kernel* is set to 'poly')
	γ	*Gamma:* accepts a value from 'scale', 'auto'. The default value is 'scale'. It can also be specified as a non-negative float	0.01, 0.1, 1, 10 (not applicable when kernel is set to 'linear')
DT	Depth	*max_depth*: data type is integer or none. The default value is none, which means the tree is expanded until the maximum depth is reached	The range from 2 to 21 (with a step size of 2) and none
	Number of leaf nodes	*max_leaf_nodes:* data type is integer or None. The default value is none, which means an unlimited number of leaf nodes will be developed, regardless of *max_depth*	The range from 2 to 21 (with a step size of 2) and none
RF	Depth	Similar to DT	Similar to DT
	Number of leaf nodes	Similar to DT	Similar to DT
	Number of DTs	*n_estimators*, data type is integer. The default value is 100	10, 50, 100

Table 4. Confusion matrix
(*Source:* Author's illustration)

Target variable	Predicted: 1	Predicted: 0	Total
Actual: 1	TP: True positives	FN: False negatives	P
Actual: 0	FP: False positives	TN: True negatives	N
Total	$\widehat{P}$	$\widehat{N}$	P + N

It is important to note that both FP and FN reflect the model's errors, but the significance of these errors depends on the specific nature of what the model is predicting. In the specific case of this research, the emphasis on importance will focus on the following two errors:

- FN in the early warning models for transitioning to delinquent debt. Here, customers who are at risk of becoming delinquent without being alerted in advance can cause unforeseen losses for the bank. Evaluating the model should pay more attention to FN. If we consider FP in these models, they only increase caution for the bank, and their impact is not as severe as FN.
- FP in the improvement warning models for debt groups. Here, customers who still maintain delinquent debt but are given improvement warnings can lead to disadvantages for the bank when formulating policies for inefficiently operating customers. If we consider FN in these models, they also increase caution for the bank but do not cause as serious consequences.

The matrix can be transformed into a ratio (probability) form by dividing each element by the sample size (P + N) to obtain the following ratios: TPR, FNR, FPR, and TNR, in which R stands for Ratio. In addition, there are several criteria computed from the confusion matrix that are valuable for evaluating the model:

- Accuracy: A general measure of the overall accuracy of the model's predictions (both alerts and non-alerts).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$$

- Precision: A measure of the accuracy among the alerted objects.

$$\text{Precision} = \frac{\text{TP}}{\widehat{\text{P}}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall: A measure of the responsiveness of the alerts compared to the actual occurrences.

$$\text{Recall} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- F_1 score: The harmonic mean of Precision and Recall.

$$F_1 = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} =$$

$$= 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} =$$

$$= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

- F_β score: A more generalized form of the F_1 score that allows for different weights to be assigned to Precision and Recall.

$$F_\beta = (1 + \beta^2)\frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} =$$

$$= \frac{(1 + \beta^2) \times \text{TP}}{(1 + \beta^2) \times \text{TP} + \beta^2 \times \text{FP} + \text{FN}}$$

where:
$\beta = 1$: No weight is assigned, similar to F_1.
$\beta > 1$: Recall is more important than Precision.
$\beta < 1$: Precision is more important than Recall.

Powers (2011) pointed out that F_1 score is not a good measure, especially when dealing with imbalanced data, as it disregards True Negatives (TN) results. Hand and Christen (2017) also criticized the overemphasis on Precision and Recall when using F1 score. In reality, different types of misclassifications have different consequences. Chicco, et al. (2021) proposed using the Matthews Correlation Coefficient (MCC) to evaluate the model, with the formula:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

Therefore, to ensure effective model adjustment, the study uses the MCC criterion. However, due to the nature of credit risk, the adjustment and selection of the optimal model in operations should also consider the following criteria:

- For models alerting shifts to the overdue group: F_β with a higher weight on Recall than Precision ($\beta > 1$). The value of β is typically chosen as 2. In this study, it will be referred to as F-Recall.
- For models alerting shifts to the non-overdue group: F_β with a higher weight on Precision than Recall ($\beta < 1$). The value of β is typically chosen as 0.5. In this study, it will be referred to as F-Precision.

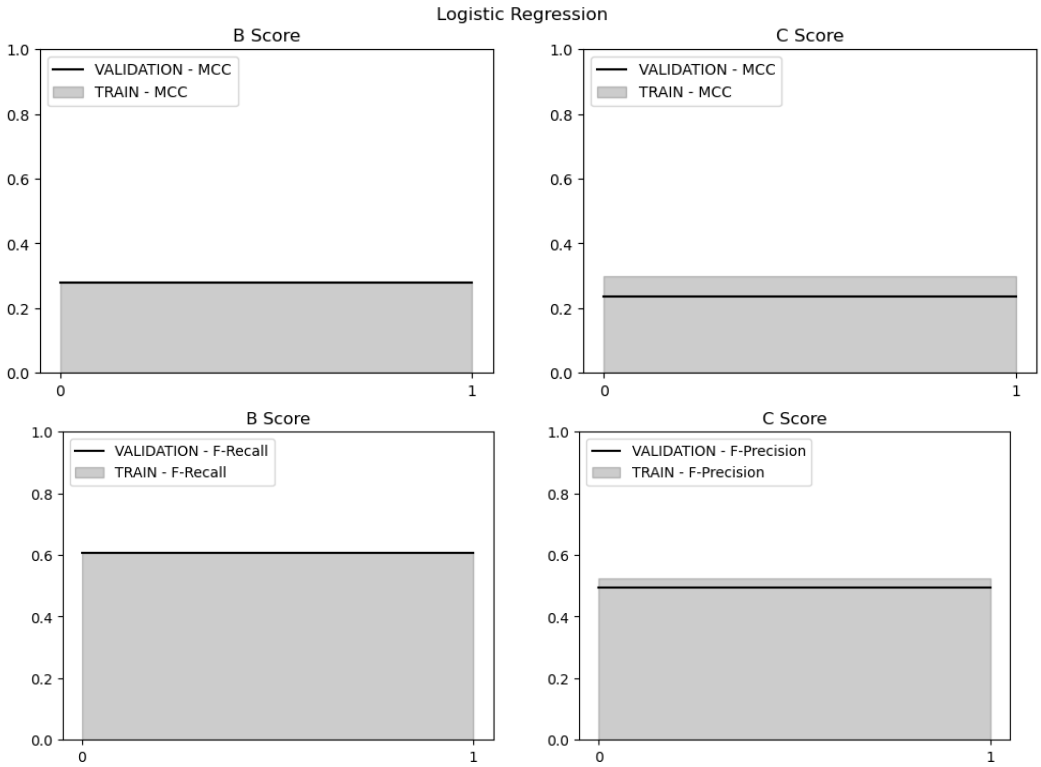

Figure 6. Logistic regression results
(*Source:* Author's own calculation)

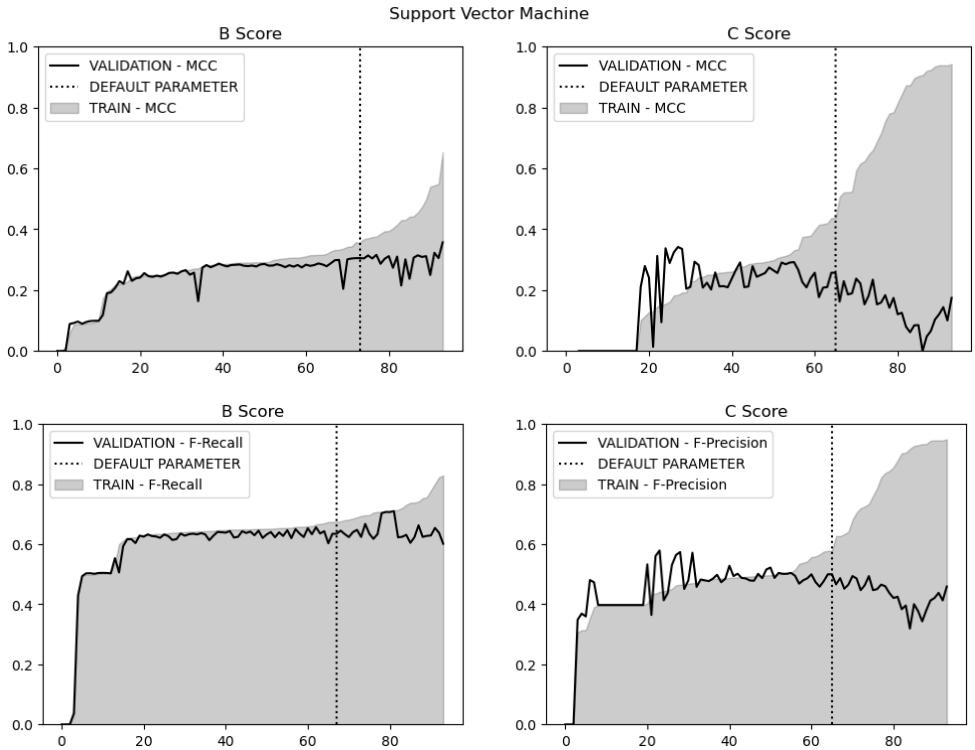

Figure 7. Support vector machine results
(*Source:* Author's own calculation)

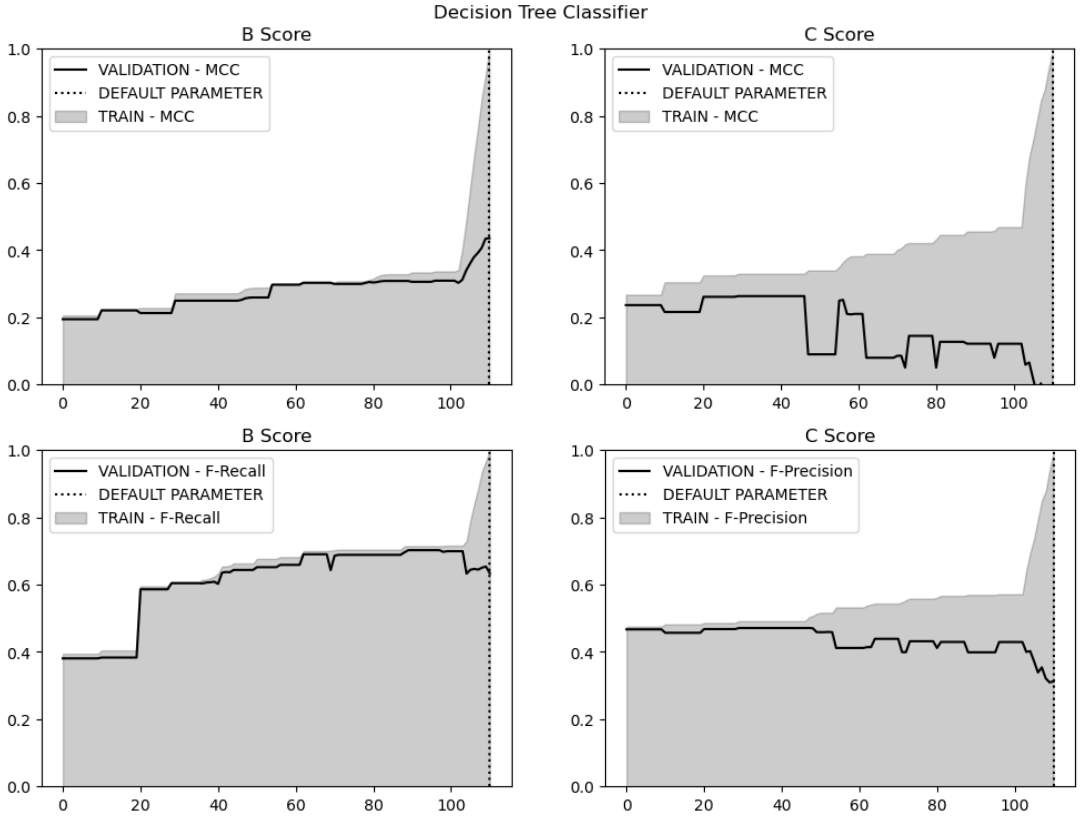

Figure 8. Decision tree results.
(*Source:* Author's own calculation)

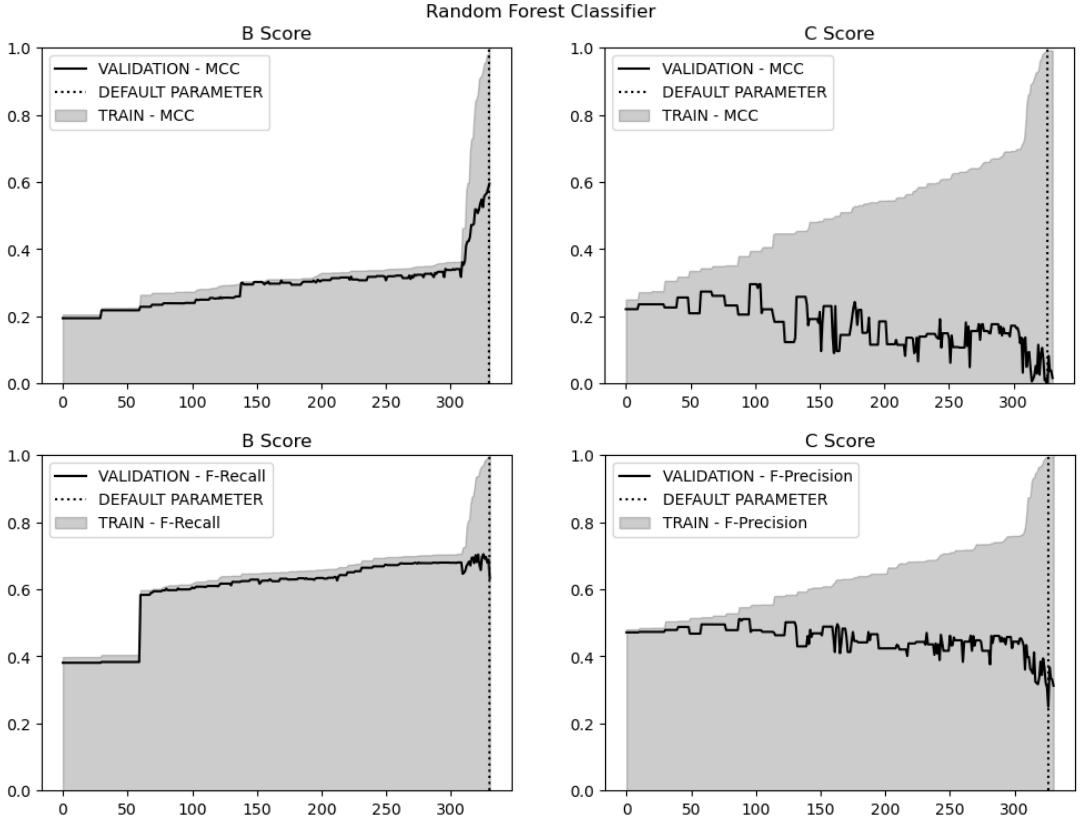

Figure 9. Decision tree results
(*Source:* Author's own calculation)

3.1 Model selection and deployment

At this phase, a comprehensive evaluation of the model will be conducted, considering both evaluation criteria: MCC and criteria relevant to credit risk management. Specifically, for the B Score customer group, the adjustment will be based on F-Recall, while for the C Score customer group, the adjustment will be based on F-Precision. Additionally, the comparison, analysis, and selection process should consider the relationship between these criteria across the datasets. The final selected model must ensure stability in both the training and test sets, while optimizing the necessary magnitude of the important criteria

During the implementation of an early warning system, besides the reported warning signs, the actual shift in the customer's debt group needs to be reviewed to determine if it aligns accurately with the early warnings. This allows for an evaluation of the system's effectiveness in issuing timely warnings. This study also performs simulations to assess the operational capability of the early warning system on a completely new dataset compared to the datasets used to build the model.

4 The proposed model

4.1 Model tuning and selection

The cases of each model during the tuning process will be sorted in ascending order according to the relevant criteria on the training set, and then, the performance trend of the model on the evaluation set will be assessed. The results are as follows Figure 6.

According to Figure 7, it can be observed that the LR, when used for both B Score and C Score datasets, does not exhibit significant overfitting as the MCC criterion does not show a significant difference between the training and evaluation sets. However, the effectiveness of the model remains relatively modest. Similarly, there is no difference in the F-Recall criterion for the B Score and F-Precision criterion for the C Score between the two datasets. In this study, the LR model has only one case and does not utilize any other parameters for tuning.

For SVM, although it has higher computational costs, it achieves better results compared to LG based on the highest MCC attained in both B Score and C Score datasets. Overfitting occurs with SVM, especially in the C Score dataset, where the MCC on the evaluation set tends to be contrary to the training set. In terms of trends, F-Recall and F-Precision follow a similar pattern to MCC. However, in terms of magnitude compared to LGR, there is not much improvement, fluctuating around 60% (B Score) and 50% (C Score).

Note: In each graph, the model cases are arranged in increasing order of the evaluation criteria on the training set. Therefore, the order of the model tuning cases will differ in the graphs. A case that achieves the highest MCC value may not necessarily have the highest F-Recall or F-Precision value.

Furthermore, it can be observed that in SVM, if default parameters are used, the evaluation criteria on the validation set are not optimal.

For DT models, in the case of B Score, the MCC on the evaluation set shows a significant improvement through different parameter cases. In most cases, overfitting does not occur, except for around 10% of the cases where the MCC on the training set increases faster than on the evaluation set. This is somewhat contrary to the F-Recall criterion, where approximately 10% of the cases on the training set shows a rapid increase but a slight decrease on the evaluation set. Note that these 10% of cases are not completely overlapping.

In the case of C Score, both MCC and F-Precision tend to decrease significantly on the evaluation set despite a rapid increase on the training set. Figure 19 clearly demonstrates the inverse relationship between the training and evaluation sets for both MCC and F-Precision criteria. Therefore, choosing an optimal case for DT requires a trade-off between accuracy on the training and evaluation sets.

In the DT model, it can be observed that when using default parameters, the results on the training set are the best. However, these results are not consistent with the evaluation set. Except for tuning based on MCC for B Score customers, the default parameters yield the best results on both the training and evaluation sets.

The results in RF and DT are quite consistent with each other in terms of the trends of evaluation criteria for both B Score and C Score. In the case of B Score, the highest MCC achieved on the evaluation set is significantly higher in RF compared to DT. However, when

considering F-Recall, the difference between the two models is not significant.

Similarly to the other models, overfitting in RF is not a major issue on the B Score, but it becomes more pronounced on the C Score as the accuracy on the training set decreases the accuracy on the evaluation set. When using the default parameters for the RF models, the results are also similar to the DT models.

In addition to evaluating the adjusted model cases based on the training and testing sets, it is necessary to assess the relationship between MCC and F-Recall, as well as F-Precision, to ensure the selection of appropriate criteria.

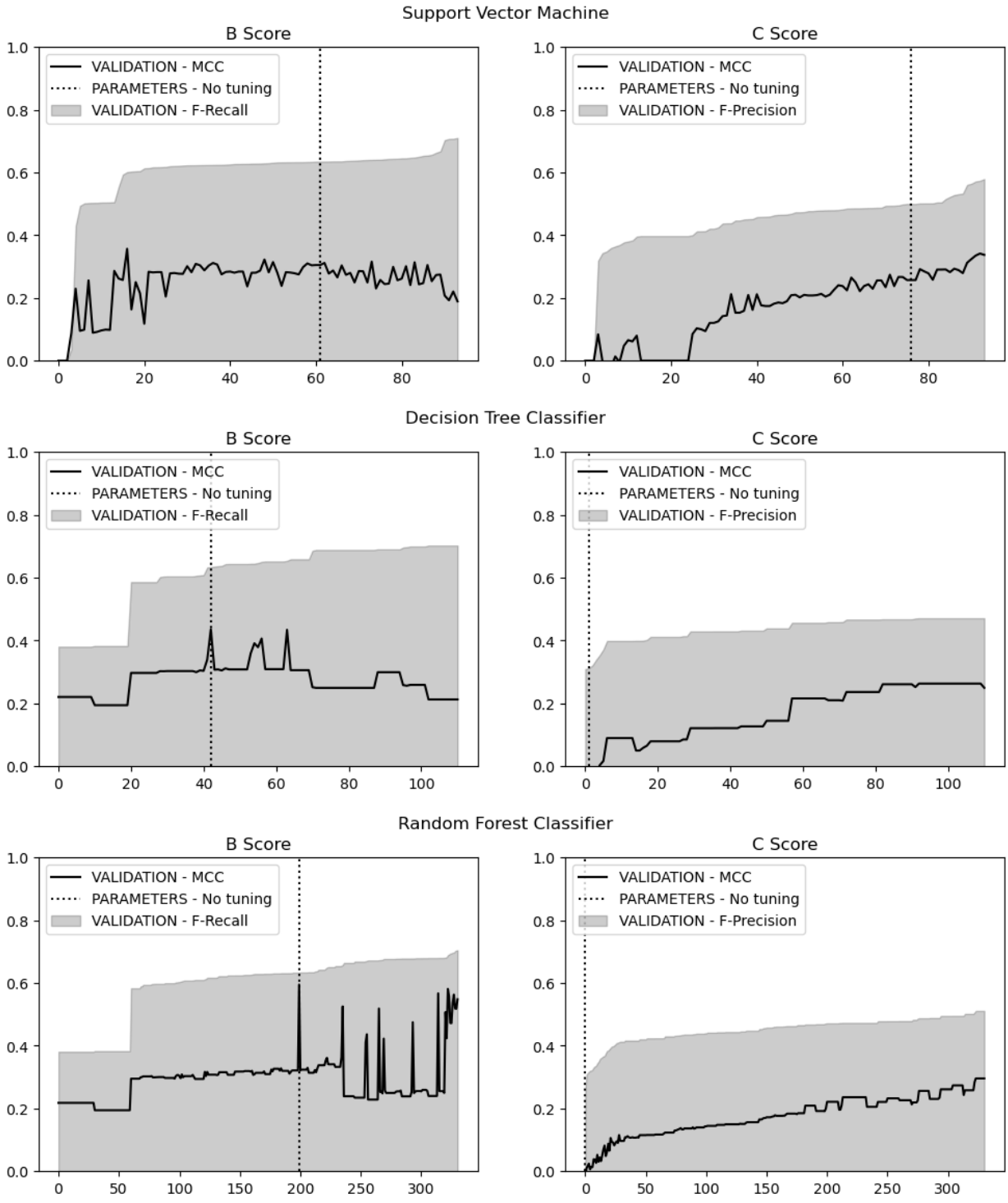

Figure 10. MCC versus F-Recall and F-Precision
(*Source:* Author's own calculation)

According to Figure 10, all criteria are represented on the evaluation set. When arranging the cases in increasing order of MCC, there is a clear positive correlation between F-Precision and the C Score dataset. However, on the B Score dataset, there is no clear relationship. This leads to the following consequences:

- On the B Score dataset, if a model is selected based on MCC, it is not necessarily guaranteed to effectively address the errors in credit risk management.
- On the C Score dataset, selecting a model based on either MCC or F-Precision ensures that the chosen model will be optimal or close to optimal.

In the case of DT and RF models, using the default parameters would be the best option when evaluating the models based on the MCC criterion.

According to the results in Table 5, for the B Score dataset, the choice of evaluation criterion has an impact on the selected model. Specifically, if MCC is used, RF should be chosen, and if F-Recall is used, SVM is the best model.

The use of MCC demonstrates consistency in both the training and evaluation sets. The RF model shows the best results across all evaluation criteria, with an overall accuracy of 81.84%. However, considering the importance of minimizing the error in identifying customer risk, if SVM is chosen based on the F-Recall index, the overall accuracy drops to 46.62%. The RF alert system correctly identifies only 67.45% of customers who actually become delinquent, while SVM achieves 91.48%, meaning that 32.55% (with RF) or only 8.52% (with SVM) of at-risk customers are overlooked. Additionally, from the perspective of alert effectiveness, 75.26% of customers receive accurate delinquency alerts, indicating that 24.74% of customers receive false alarms. In SVM, the rate of false alarms is more severe, with only 37.52% of customers receiving accurate alerts.

A comprehensive view of the selected models based on the F-Recall criterion reveals that RF has an F-Recall of 70.46%, which is not a significant improvement compared to SVM's 71.04%. However, RF has relatively balanced Recall and Precision rates (70.78% and 69.20%, respectively), while the best-performing SVM model has a significant disparity (91.48% Recall and 37.52% Precision). Therefore, RF based on the F-Recall criterion is also a promising model. It should be noted that RF based on MCC and RF based on F-Recall will have different parameter sets.

In the dataset of customers who have become delinquent (C Score), the results in Table 6 are inconsistent across different evaluation criteria. It can be observed that tree-based models (DT and RF) yield higher results on the training set compared to other models. However, when considering the evaluation set, each model has different effectiveness based on different evaluation criteria.

Returning to the model selection for C Score, if MCC or F-Precision is used, the best model would be SVM, but with different parameter sets. Considering the importance of minimizing errors in identifying customers for credit improvement, it is advisable to prioritize evaluation criteria that emphasize the magnitude of Precision. In other words, incorrectly identifying a customer who can benefit from credit improvement can have more severe consequences than accurately predicting how many customers can actually improve their credit situation. Incorrect identification could lead the bank to implement support policies for the wrong customer segments.

Therefore, the optimal model for customers in the C Score dataset could be SVM based on MCC with a Precision of 58.44% and an F-Precision of 57.40%. Alternatively, SVM based on F-Precision could also be considered with corresponding figures of 62.3% and 57.93%.

In summary, the model selection for early warning system can be as follows, Table 7.

4.2 Early warning system deployment

The results of the model evaluation during deployment period indicate the following:

Firstly, for the B Score dataset, when fine-tuning the model using the MCC criterion, the performance of the RF model is not particularly good when used with new data due to significantly lower Recall and F-Recall criteria compared to the evaluation set during the fine-tuning process. However, when using the models obtained from fine-tuning based on F-Recall, such as SVM and RF, there is a noticeable improvement in results, as shown in Table 8.

Table 5. The results of the evaluation criteria for the B Score model
(*Source:* Author's own calculation)

Dataset	Criteria (%)	Tuned by MCC				Tuned by F-Recall			
		LR	SVM	DT	RF	LR	SVM	DT	RF
Train	Accuracy	64.62	83.22	96.29	**99.06**	64.62	47.05	50.31	**96.11**
	Recall	65.19	86.98	98.20	**99.97**	65.19	91.53	90.05	**99.10**
	Precision	47.59	69.86	91.29	**97.27**	47.59	37.73	39.19	**90.15**
	F1	55.02	77.48	94.62	**98.60**	55.02	53.44	54.61	**94.41**
	F-Recall	60.70	82.91	96.74	**99.42**	60.70	71.22	71.49	**97.17**
	F-Precision	50.31	72.72	92.59	**97.80**	50.31	42.76	44.18	**91.81**
	MCC	27.92	65.34	91.94	**97.92**	27.92	19.60	22.82	**91.68**
Validation	Accuracy	64.66	70.58	74.09	**81.84**	64.66	46.62	50.41	**79.85**
	Recall	64.98	61.58	66.95	**67.45**	64.98	**91.48**	87.83	70.78
	Precision	47.62	55.08	59.79	**75.26**	47.62	37.52	39.02	**69.20**
	F1	54.96	58.15	63.17	**71.14**	54.96	53.22	54.04	**69.98**
	F-Recall	60.56	60.16	65.39	**68.88**	60.56	**71.04**	70.26	70.46
	F-Precision	50.31	56.27	61.10	**73.56**	50.31	42.54	43.90	**69.51**
	MCC	27.89	35.71	43.45	**58.13**	27.89	18.94	21.29	**54.83**

Table 6. The results of the evaluation criteria for the C Score model
(*Source:* Author's own calculation)

Dataset	Criteria (%)	Tuned by MCC				Tuned by F-Precision			
		LR	SVM	DT	RF	LR	SVM	DT	RF
Train	Accuracy	65.02	63.60	56.71	**65.72**	65.02	63.60	58.66	**65.72**
	Recall	67.69	44.62	**93.33**	85.64	67.69	36.92	**92.82**	85.64
	Precision	49.44	47.03	43.96	**50.15**	49.44	46.45	45.14	**50.15**
	F1	57.14	45.79	59.77	**63.26**	57.14	41.14	60.74	**63.26**
	F-Recall	63.04	45.08	**76.21**	75.02	63.04	38.50	**76.63**	75.02
	F-Precision	52.26	46.52	49.16	**54.68**	52.26	44.17	50.31	**54.68**
	MCC	29.80	18.44	33.02	**39.49**	29.80	15.51	35.05	**39.49**
Validation	Accuracy	**63.79**	70.78	53.91	62.55	63.79	**71.60**	54.73	62.55
	Recall	57.14	53.57	**89.29**	75.00	57.14	45.24	**85.71**	75.00
	Precision	48.00	**58.44**	42.13	47.37	48.00	**62.30**	42.35	47.37
	F1	52.17	55.90	57.25	**58.06**	52.17	52.41	56.69	**58.06**
	F-Recall	55.05	54.48	**72.96**	67.16	55.05	47.86	**71.15**	67.16
	F-Precision	49.59	**57.40**	47.11	51.14	49.59	**57.93**	47.12	51.14
	MCC	23.62	**34.19**	26.33	29.60	23.62	**33.75**	24.98	29.60

Table 7. The results of the evaluation criteria for the C Score model
(*Source:* Author's own calculation)

Customer	Model	Selection	Parameters

B Score	Best	RF tuned by MCC	n_estimators = 100;
			max_depth = 20;
			max_leaf_node = None

		SVM tuned by F-Recall	kernel = 'sigmoid'; C = 0.1; gamma = 0.01

	Second best	RF tuned by F-Recall	n_estimators = 100;
			max_depth = 16;
			max_leaf_node = None

C Score	Best	SVM tuned by MCC	kernel = 'poly'; degree = 4; C = 0.01 gamma = 0.1
		SVM tuned by F-Precision	kernel = 'poly'; degree = 4; C = 0.1 gamma = 0.01

Table 8. Early warning system deployment for B Score customers
(*Source:* Author's own calculation)

Criteria (%)	Tuned by MCC	Tuned by F-Recall	
	RF (best)	SVM (best)	RF (second best)
Accuracy	81.84 → **78.67**	46.62 → **36.86**	79.85 → **76.56**
Recall	67.45 → **29.01**	91.48 → **91.53**	70.78 → **38.14**
Precision	75.26 → **42.84**	37.52 → **22.44**	69.20 → **39.37**
F-Recall	68.88 → **31.01**	71.04 → **56.65**	70.46 → **38.38**

Table 9. Early warning system deployment for C Score customers
(*Source:* Author's own calculation)

Criteria (%)	Tuned by MCC	Tuned by F-Precision
	SVM (best)	SVM (best)
Accuracy	70.78 → **61.94**	71.60 → **65.54**
Recall	53.57 → **54.48**	45.24 → **50.00**
Precision (*)	58.44 → **40.33**	62.30 → **43.79**
F-Precision (*)	57.40 → **42.54**	57.93 → **44.91**

Overall, the effectiveness of real-world operations did not reach the level observed during the model evaluation in the tuning process. Most evaluation metrics decreased, but it can be observed that with the current data, tuning should be based on risk management criteria such as Recall or F-Recall, rather than a general accuracy indicator like MCC, as the decrease in important metrics is less significant. Additionally, the RF model did not perform as stably as the SVM model based on the important criteria. It is important to note that although SVM has a high Recall of 91.53%, its Precision is only 22.44%, which indicates an overly cautious approach as many customers are predicted to be at risk of default, which may be relatively safer than using RF, which only identifies 38.15% of customers who will actually default.

Secondly, in the C Score customer group, the evaluation criteria also decreased in the test set, but the decrease was not as significant as in the B Score group. Once again, it can be seen that tuning based on the F-Precision criterion yields better results than MCC. Furthermore, for the B Score group, SVM models performed the most stable although the overall accuracy level is not high.

Table 9 shows that both important model evaluation criteria decreased in the test set, but the decrease was not as substantial as in the B Score customer group. Despite the decrease, tuning based on F-Precision still achieved a higher level of 44.91% compared to 40.33% for MCC.

4.3 Discussion

This study provides evidence that various factors related to customer deposit behavior, and past loan characteristics can predict the likelihood of customer credit migration within one year. This was achieved through the development of two models: B Score, which predicts the likelihood of a customer transitioning to a delinquent group, and C Score, which predicts the potential for a customer in the delinquent group to improve their credit status.

The study applied four basic machine learning models: Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest. These models were tuned by adjusting their parameters to improve prediction results. These tuned parameters have been demonstrated to be better than the default parameters.

Among these models, LR served as the baseline model, commonly used in banks due to its simplicity and interpretability. However, the other models showed significant improvement in accuracy after tuning. Specifically, LR achieved accuracies of 64.66% and 63.79% for B Score and C Score, respectively, while the tuned models achieved higher accuracies of 81.84% and 71.60%. This suggests that there are

alternative algorithm choices that can be applied in the early warning system, replacing LR.

This study also investigated which evaluation criteria should be used to obtain an optimal model. Typically, models are tuned to maximize overall accuracy. However, this study highlights the importance of selecting appropriate criteria aligned with the significance of early warning alerts in credit risk management. Specifically, for B Score models, F-Recall should be chosen, while F-Precision should be prioritized for C Score models. These criteria emphasize caution in avoiding false alarms in credit risk management. Applying more general criteria, such as MCC, which considers all types of prediction errors, may lead to serious consequences in bank credit operations.

Besides, for the dataset used in this study and the applied tuning methods, the SVM model yielded the best results compared to the other models. Firstly, in the B Score customer set, although RFC achieved an impressive accuracy of 81.84% when tuned with MCC, it fell short in terms of criteria that prioritize caution in banking. RFC only achieved an F-Recall of 67.45%, which was lower than SVM's 71.04%. Another tuned RFC model with F-Recall achieved only 70.46%. Secondly, in the C Score customer set, SVM also had the advantage in both tuning methods using MCC and F-Precision, with F-Precision values of 57.40% and 57.93%, respectively. Although these two numbers are close, in practical applications, there will be more noticeable differences.

Finally, the study also implemented an alert system on the test dataset, which was entirely new data not used in training the models and was not a randomly sampled subset of the model-building dataset. These new data were generated after the model was constructed, specifically after the cutoff date of December 2021. After the actual credit migration of customers, the study compared the results with the model's predictions. The significant issue to acknowledge is that all evaluation criteria for the models decreased. In the B Score set, the accuracy criterion decreased to 78.67% and 76.56% for the two RFC models from the initial values of 81.84% and 79.85%, and more significantly, SVM dropped drastically from 46.62% to 36.86%. However, when considering cautious criteria, SVM's Recall remained stable at 91.48%, and F-Recall reached 71.04%, higher than both RFC models at 68.88% and 70.46%. For the C Score set, the two optimized SVM models also experienced a decrease in Precision and F-Precision values. As mentioned earlier, although these SVM models had approximate prediction values, the model with a slightly higher F-Precision would perform better on the test dataset. Thus, even though the models did not perform optimally on the test dataset, the cautious approach was still ensured to avoid losses for the bank.

5 Conclusion

This study presents a methodology for applying machine learning models to identify customers at risk of defaulting on their loans as well as customers who have the potential to improve their debt status. The machine learning models have achieved certain levels of effectiveness in prediction. Additionally, the study demonstrates different approaches to fine-tuning the machine learning models to optimize accuracy and minimize the cost of errors.

Due to limitations in data availability, this study only utilized data related to customer deposit behavior. Furthermore, the variables pertaining to deposit behavior were relatively rudimentary and did not provide substantial information for developing an early warning system. Therefore, future development could focus on researching data and variables that could be useful in constructing an early warning system for customer debt transitions.

Moreover, this study utilized four basic machine learning algorithms to construct the alert system, which may have limitations in terms of accuracy. Nowadays, there are many advanced tree-based models with improved algorithms such as CatBoost, XGBoost, and deep learning models that have proven to be effective in identifying customers at risk of default. Therefore, the next development direction could involve applying one of these advanced algorithms, combined with necessary refinements, to improve the research outcomes.

6 Acknowledgement

This research is funded by University of Economics Ho Chi Minh City, Vietnam (UEH).

7 References

[1] Bielecki, T.R., Rutkowski, M., 2013. Credit Risk: Modeling, Valuation and Hedging. Berlin, Heidelberg: Springer-Verlag.

[2] Bishop, C.M., 2006. Pattern recognition and Machine Learning. Springer.

[3] Bluhm, C., Overbeck, L., and Wagner, C., 2016. Introduction to Credit Risk Modeling. Florida: CRC Press.

[4] Chicco, D., Tötsch, N., and Jurman, G., 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, Vol.14(13), pp.1-22.

[5] Cramer, J.S., 2002. The Origins of Logistic Regression. Tinbergen Institute Working Paper, No.119/4.

[6] Dahooie, J.H., Hajiagha, S.H., Farazmehr, S., Zavadskas, E.K., and Antucheviciene, J., 2021. A novel dynamic credit risk evaluation method using data envelopment analysis with common weights and combination of multi-attribute decision-making methods. *Computers & Operations Research*, No. 129, 105223.

[7] Djeundje, V.B., Crook, J., Calabrese, R., and Hamid, M., 2021. Enhancing credit scoring with alternative data. *Expert Systems with Applications*, No. 163, 113766.

[8] Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27, pp.861-874.

[9] Figlewski, S., Frydman, H., and Liang, W., 2012. Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Economics & Finance*, No. 21(I), pp.87-105.

[10] Forster, J.J., Buzzacchi, M., Sudjianto, A., and Nagao, R., 2016. Modelling credit grade migration in large portfolios using cumulative t-link transition models. *European Journal of Operational Research*, Vol. 254(3), pp.977-984.

[11] Greiff, W.R., 1999. Maximum Entropy, Weight of Evidence and Information Retrieval. PhD Thesis, University of Massachusetts Amherst.

[12] Gujarati, D.N., Porter, D.C., 2009. Basic Econometrics. McGraw-Hill/Irwin.

[13] Ha, D.T., 2019. Early risk warning for credit provision to customers and related individuals is a critical concern for commercial banks in Vietnam. *Banking Science & Education Journal*.

[14] Hand, D., Christen, P., 2017. A note on using the F-measure for evaluating record linkage algorithms. Statistics and Computing, Vol. 28(3), pp.539-547.

[15] Ionela, S.A., 2014. Early Warning Systems - Anticipation's Factors of Banking Crises. *Procedia Economics and Finance*, No.10, pp.158-166.

[16] Kim, Y., Sohn, S.Y., 2008. Random effects model for credit rating transitions. European *Journal of Operational Research*, Vol. 184(2), pp.561-573.

[17] Kwon, Y., Park, S.Y., 2023. Modeling an early warning system for household debt risk in Korea: A simple deep learning approach. *Journal of Asian Economics*, No.84, 101574.

[18] Markov, A., Seleznyova, Z., and Lapshin, V., 2022. Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science*, No.8, pp.180-201.

[19] Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. Massachusetts Institute of Technology.

[20] Powers, D.M., 2011. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, Vol. 2(1), pp.37-63.

[21] Reinhart, C., Goldstein, M., and Kaminsky, G., 2010. Methodology for an Early Warning: The Signals Approach. University of Maryland, College Park, Department of Economics.

[22] Rokach, L., Maimon, O., 2008. Data Mining with Decision Trees: Theory and Applications. World Scientific Publishing.

[23] Shalev-Shwartz, S., Ben-David, S., 2014. 18 - Decision Trees. In Understanding Machine Learning. Cambridge University Press.

[24] Siddiqi, N., 2006. Credit Risk Scorecards, Developing and Implementing Intelligent Credit Scoring. Hoboken, NJ: John Wiley & Sons, Inc.

[25] Slapnik, U., Lončarski, I., 2021. On the information content of sovereign credit rating reports: Improving the predictability of rating transitions. *Journal of International Financial Markets, Institutions and Money*, No. 73, 101344.

[26] State Bank of Vietnam., 2021. Circular No. 11/2021/TT-NHNN: Prescribing classification of assets, amounts and methods of setting up risk provisions and use of provisions for control and management of risks arising from operations of credit institutions and foreign bank branches.

[27] Wang, L., Zhang, W., 2023. A qualitatively analyzable two-stage ensemble model based on machine learning for credit risk early warning: Evidence from Chinese manufacturing companies. *Information Processing & Management*, No. 60, 103267.