VIETNAM NATIONAL UNIVERSITY

HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

GRADUATION THESIS

STABILITY CLASSIFICATION BY CAMEL USING MACHINE LEARNING - A CASE OF VIETNAM COMMERCIAL BANKS

Supervisor:	**NGUYEN ANH PHONG**
Student:	**GIAP HOANG LONG**
Student ID:	**K194141728**
Class:	**K19414C**

Ho Chi Minh City, April 2023

VIETNAM NATIONAL UNIVERSITY

HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

GRADUATION THESIS

STABILITY CLASSIFICATION BY CAMEL USING MACHINE LEARNING - A CASE OF VIETNAM COMMERCIAL BANKS

Supervisor:	**NGUYEN ANH PHONG**
Student:	**GIAP HOANG LONG**
Student ID:	**K194141728**
Class:	**K19414C**

Ho Chi Minh City, April 2023

COMMENTS OF THE LECTURER:

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

....

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This study proposes a machine learning-based approach to evaluate the stability of commercial banks in Vietnam. The approach uses the CAMELS framework and key financial indicators to build forecasting models and classify banks into those with good or not good financial health. The study analyzed data from 45 Vietnamese banks between 2002 and 2021 using multiple machine learning algorithms, including Logistics Regression, Support Vector Machine, and AdaBoost, to identify the key drivers of banks' financial soundness.

The results show that the approach can forecast the stability of Vietnam's commercial banks with high accuracy, classifying banks into good and not good financial health. Three factors, including Non-Interest Expense, Cumulative Gaps Over Total Assets and Profits Before Tax, were found to be important for the classification of bank health.

The study provides a practical application of machine learning techniques in assessing stability in Vietnam banks, contributing to the fields of finance and machine learning. The findings could assist investors, depositors, and companies without knowledge or understanding of banking and finance to identify banks in good financial health and avoid risks associated with substandard banks.

Keywords: stability, CAMEL, machine learning, classification

1. INTRODUCTION

According to many studies, Huang et al. (2012) have shown bank failures due to poor management have effects on the economy as a whole, deposit holders, investors, and workers. In addition, Tung et al. (2004) have found bank failure can have negative financial effects and cause adverse effects such as large underwriting costs because of the failed bank and a drop in investor and depositor confidence. The study of Balgova et al. (2016) has shown that growth ignored by the economy due to an excess of bad debt can exceed two percentage points annually. Therefore, establishing an early warning system and prompt central bank intervention that can safeguard investors, depositors, and the economy is crucial.

Along with other countries throughout the world, in Vietnam as well, there are many previous studies that have examined the instances of bank failure there. Quang (2015) has found 13 banks that experienced financial crises between 2005 and 2013, the author of Determinants of Banking Crisis also discovers that bad debt and quality of assets are the main contributors to the banking crisis. Asset quality is declining, and private banks are more likely to experience a catastrophe than state-owned banks. Vuong et al. (2013) have proved the widespread financial crisis increased funding costs and reduced consumer trust in the economy.

There are now, however, not enough studies to recommend a rating system that is certain and can distinguish between banks with weak financial standing and banks with strong ones. Small depositors and those with limited banking and financial understanding may find it challenging to determine which banks are conducting themselves in a healthy manner. Therefore, further study is required to develop research models that will allow managers, investors, and the bulk of the bank's primary customers to approach corporate health more naturally.

After learning from previous studies, one of the approaches to evaluate the financial soundness of the banking system proposed by the IMF and World Bank (2005) is the CAMELS method based on the related finance indicators. CAMELS has proven to be one of the effective tools for regulators to observe and supervise the banking system

Le (2017) (cited in Gilbert, Meyer & Vaughan 2000; Hays, De Lurgio & Gilbert 2009). The CAMELS approach is additionally believed to be one of the primary methods for evaluating bank performance (Le, 2017) (cited in Derviz & Podpiera 2008; Evans et al. 2000; Kumar et al. 2012). Therefore, the CAMELS technique will be used in this study to evaluate the stability of Vietnam's commercial banks.

This study adds to the existing knowledge in various aspects. It makes a significant contribution to the literature by combining the CAMELS method with machine learning algorithms to assess the financial health status of banks. By integrating new technologies such as machine learning into traditional methods of determining a bank's stability, the study aims to produce more accurate and timely results based on the bank's financial ratios. In addition, at present given the limited number of studies predicting financial health in Vietnam, this study will refer to international research and provide new insights. The research is particularly relevant for depositors, investors, and corporate organizations who seek to identify banks with stable financial situations to invest in or use their services, especially those who do not have much knowledge of the financial sector and also corporate organizations.

Nowadays, with the development and application of algorithms and technology, the financial health of banks can be predicted by machine learning techniques, which is a typical classification problem. The majority of previous studies prioritized forecasting bank failure. In order to help depositors and retail investors determine which banks pose a significant risk, our study's primary goal was to categorize banks according to their financial soundness from the viewpoint of each individual depositor. Because there is currently no reliable framework for establishing this classification, therefore, in our analysis, we may divide banks into two categories: healthy banks and unhealthy banks. The model will help depositors in making wiser choices about where to store their funds by using the aforementioned classification.

In this research, 45 Vietnamese banks are selected and collected to be a dataset from Le et al. (2022) using annual and financial reports between the years 2002 and 2021. The information is based on the organizational structure of the banks, which includes foreign vs. domestic, state vs. private ownership, and commercial vs. policy banks. Following the raw data processing step, which labels the bank's stability as excellent or bad using the CAMELS technique, the variables that are truly significant for the bank are transformed and chosen to forecast then create machine learning models and finally assess the performance of the models.

The structure of this research paper is organized in the following manner: The definitions and literature review are presented in Section 2. The methodology is explained in Section 3. Section 4 describes the data that was utilized in this study. The research result is presented and discussed in Section 5. Section 6 is the conclusion and summary of the findings. References are in Section 7.

2. LITERATURE REVIEW

2.1 Stability of Bank

Stability is a sophisticated significant topic in the field of banking that is utilized to assess the financial health of any financial institution as a tool to indicate its capacity to use it effectively. Csikosova et al. (2019) have shown the financial status of the organization, as well as its ability to generate profits with little risk. Financial stability demonstrates the capacity to strike a balance between various environmental factors and all business stakeholders. It displays the business's financial well-being in terms of profitability, liquidity, financing, asset utilization, and market value. The company's financial statements serve as the primary source of information about its financial health demonstrated by Ross et al. (2008).

Bank stability research, bank with poor financial health can lead to bankruptcy, which has a negative impact on creditors, employees, investors, suppliers, people consumers, and local communities. It is a necessary part of running a business. When a company is unable to meet its debt obligations, it declares itself insolvent proved by Karim et al (2021).

Previous studies have focused much on bankruptcy prediction. Topics on bank financial soundness are required because there is a gap in previous research in this area. Additionally, there aren't enough studies that classify banks based on their financial stability status from the perspective of retail depositors, which is especially essential to create the conditions for retail investors to make an informed decision while selecting a commercial bank to deposit their money.

2.2 CAMELS

CAMELS is a method that is used to analyze performance of the banks and is a well-established method for assessing the financial soundness of banks. It was generated by regulatory authorities in the United States in the 1970s. The FDIC, which stands for the Federal Deposit Insurance Corporation, in the United States developed a system of ratings that ranks banks according to stability standards using CAMELS data shown by Affes & Kaffel (2019).

CAMELS is an acronym for the six factors that are used to evaluate the financial condition of banks: Capital adequacy, Asset quality, Management, Earnings, Liquidity, and Sensitivity to market risk. It is a regulatory rating system used by banking supervisors to assess the financial stability status of banks and determine whether they are operating in a safe and sound manner. The CAMELS approach provides a standardized framework for evaluating the overall health of banks and is widely used in many countries around the world.

The CAMELS framework will be applied during this study to assess the stability of commercial banks in Vietnam, Le (2017) has examined the financial stability of Vietnamese banking institutions between the years 2008 and 2013 using data envelopment analysis and the CAMELS approach. This study will also serve as a framework for future research on the six components of the CAMELS score model.

The components of the CAMELS approach include capital adequacy, asset quality, management quality, earnings ability, liquidity, and sensitivity to market risk.

Capital adequacy (C) ratio is used to evaluate the banking system's financial health because it shows how well this sector can withstand potential losses brought on by both internal and external factors, or even both. According to the study Le (2017), C can be calculated by the equity-to-total-assets ratio (ETA). ETA represents the proportion of total assets that are funded by a bank's shareholders. As a result, a bank is safer the higher the ratio. Therefore, this is one of the most crucial CAMELS framework indicators for ensuring a bank's soundness.

Asset quality (A) is a measure of a bank's strength that is closely related to capital adequacy because insolvency risk is accompanied by asset depreciation. According to

Le (2017), this study uses an indicator of loan quality is the ratio of non-performing loans (NPL).

Management quality (M) reflects the bank's management's ability to control operating costs. Comparable to the conventional method of representing financial ratios (M). The Cost-Income Ratio (CIR), based on Le (2017) prior study, is the indicator employed in this study. It is challenging to assess all of these (M) components using financial indicators due to their characteristics. It is difficult to find financial indicators that characterize this M factor and most indicators will not fully reflect the management thinking as well as the strategic and managerial situation of banks.

Earnings ability (E) can be measured by Return on Assets (ROA). ROA, as measured by the ratio of profit before tax to total assets. The higher this indicator is, the more effective a bank's performance is.

liquidity (L) of a bank measures its capacity to withstand shocks to cash flows and unforeseen withdrawals from depositors. This study uses the ratio of liquid assets to total assets (LTA). This ratio shows the proportion of liquid assets to total assets. In order to cover unforeseen depositor withdrawals, a bank may experience liquidity issues if it advances a higher volume of loans.

Sensitivity to market risks (S) reflects the way in which the market prices (the interest rates, the exchange rates, and the equity prices) impact the bank’s earnings and capital negatively. S is measured by the ratio of the difference between rate-sensitive assets and rate-sensitive liabilities to total assets. Rate-sensitive assets comprise dues from financial institutions and total loans whereas rate-sensitive liabilities include interbank liabilities and other liabilities. Accordingly, the higher gap means that a bank becomes less exposed to the risk of losses arising from changes in market prices because the value of sensitive assets is still able to cover the value of sensitive abilities.

To make sure that there are not too many variations in the phenomenon or indicators, these indicators are often calculated over a set period of time. However, the author will compute each year and utilize the C, A, M, E, and L ratios in this study with the exception of the S component because the S element does not have any documents or studies evaluating it with CAMEL that the author sees fit to score for the financial soundness of banks

2.3 Machine learning and Algorithms

Machine learning is a subfield of artificial intelligence that involves the development of algorithms and statistical models that enable computers to learn from data and make predictions or decisions without being explicitly programmed Giuffrè et al. (2023). Classification is a common problem in machine learning where the goal is to assign a label or category to a given input based on its features. Logistic Regression, Support Vector Machine (SVM), and AdaBoost have commonly used algorithms for classification tasks.

Logistic regression (LR) is a popular statistical method that models the relationship between a dependent variable and one or more independent variables. It is often used for binary classification tasks, where the goal is to predict one of two possible outcomes. Logistic regression works by fitting a logistic function to the data, which maps the input features to a probability of belonging to one of the two classes.

Support vector machine (SVM) is a popular machine learning algorithm that can be used for both linear and nonlinear classification tasks. The goal of SVM is to find the hyperplane that best separates the input data into different classes. SVM achieves this by finding the maximum margin hyperplane, which is the hyperplane that maximizes the distance between the closest data points from each class.

AdaBoost (Adaptive Boosting) is a boosting algorithm that combines multiple weak classifiers to form a strong classifier. In Adaboost, each weak classifier is trained on a subset of the input data, and the algorithm assigns higher weights to the misclassified data points. By iteratively adjusting the weights and combining the weak classifiers, Adaboost is able to improve the overall classification accuracy, even when dealing with complex and noisy data. The AdaBoost algorithm of Freund and Schapire was the first practical boosting algorithm and remains one of the most widely used and studied, with applications in numerous fields published by Schapire (2013). This will also be a new point in the research to apply Adaboost in predicting the classification problem of this bank.

2.4 Background and Empirical Research

Studies on predicting binary factors on the financial status of banks have been implemented both in Vietnam and throughout the world. However, most forecasts utilize conventional statistical models, and the scope of the subject is fairly limited. The prediction of bank collapse or bank failure brought on by crises has been the subject of studies (Thomson, 1991; Persons, 1999; Quang, 2015; Gaul et al, 2019; Affes & Kaffel, 2019). With a proper classification rate of 96.22%, the study by Affes & Kaffel (2019) employs a logit model to classify. In addition, the study employs the CAMELS framework's financial indicators and the early warning system of bank troubles to examine bank collapse. Gaul, Lewis, Jonathan Jones, and Pinar Uysal's (2019) study, which employed a logit model to estimate high-risk CAMELS ratings within a year, demonstrates that the findings may be used for forecasting. Most models base the predictor variable, which is a banking crisis or bank failure, on the analysis of financial indicators of the CAMELS framework. Due to this, the research will also make reference to the CAMELS set of financial indicators used in the Vietnamese market as provided by Le (2017) in order to assess the bank's financial health. Additionally, Desta (2016) (cited in Rozzani, Rahman, Babar, and Zeb, 2011) has shown the CAMELs Component and CAMEL Rating system as sources for the author to calculate and classify the type of banks for train and test data.

Machine learning has become a powerful tool for creating and resolving challenging problems where statistics have certain limits as a result of technological advancements. Machine learning is being used more and more often in the banking industry, leading to several ground-breaking research projects. Forecasting bank's financial health is not very useful, as was previously stated. However, there are a lot of themes and different ways to estimate the risk of bankruptcy or the danger of bank collapse. Viswanathan et al. (2020) reviewed several studies using algorithms to assess financial health. First is Decision trees (DTs) and Random Forest from Halteh et al., (2018). The study then used a more differentiated algorithm that is CART and multivariate adaptive regression splines (MARS) used by Affes & Kaffel (2019). Besides, Ekinci & Erdal (2016) Both hybrid ensemble learning (bagging and multi-boosting) and ensemble learning models (random subspaces) have been applied.

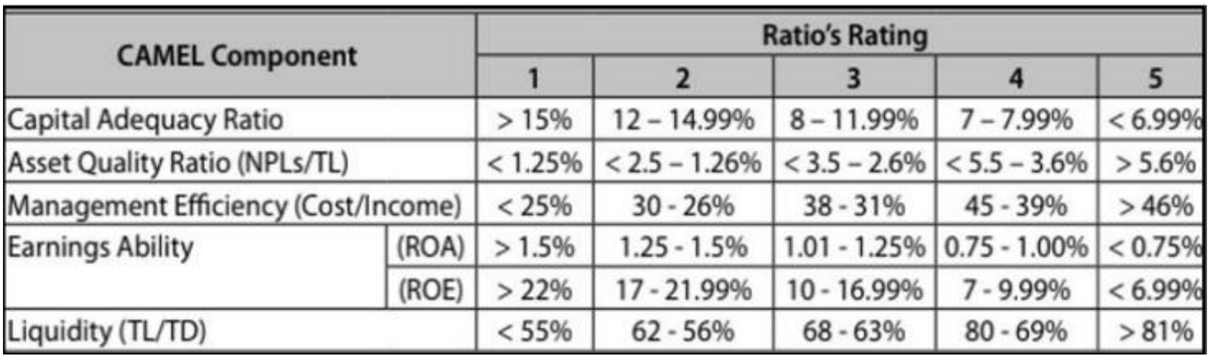
In recent times, there are not many studies that use machine learning to forecast the financial health of Vietnamese banks. However, since the CAMELS index's approach to evaluating financial health is comparable to its studies on bank failure, it is also conceivable to make use of other research articles in this study. discusses a bank collapse or catastrophe. Machine learning was utilized by Meitei et al. (2022) to detect weak banks in India. The study included five high-predictive-accuracy algorithms: Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Random Forest, and Average Neural Networks. Based on 12-year data from 2005 to 2017. Viswanathan et al. (2020) using machine learning algorithms classified 44 Indian banks into various financial health groups. Applying a machine learning approach to traditional financial ratios Le & Viviani (2018) have predicted the collapse of an innovative bank. The results show that artificial neural networks and k-nearest neighbor methods are the most accurate.

The goal of this study is to use methods from earlier studies to assess the variables that have a significant impact on the financial stability of Vietnamese banks. Applying the new algorithm and comparing it to other well-known algorithms will allow you to determine whether the model is truly applicable to Vietnamese banking data and will aid in the identification of health status. Financial stability of the Bank more quickly and easily.

3. METHODOLOGY

3.1 Target Variable Defination

The CAMELS framework is a well-established method for assessing the financial health of banks. In this framework, financial data is calculated for each of the six key factors: Capital Adequacy, Asset Quality, Management, Earnings, Liquidity, and Sensitivity to Market Risk. In previous studies, these factors are then assigned scores on a scale of 1 to 5, with 1 representing the strongest score and 5 representing the weakest score. Following this trend, Le (2017) has shown a framework will be calculated by each factor, and finnaly scored on a scale of 1 to 5 based on the following criteria in the CAMEL Component table of Desta (2016):



CAMEL Component		Ratio's Rating				
		1	2	3	4	5
Capital Adequacy Ratio		> 15%	12 – 14.99%	8 – 11.99%	7 – 7.99%	< 6.99%
Asset Quality Ratio (NPLs/TL)		< 1.25%	< 2.5 – 1.26%	< 3.5 – 2.6%	< 5.5 – 3.6%	> 5.6%
Management Efficiency (Cost/Income)		< 25%	30 - 26%	38 - 31%	45 - 39%	> 46%
Earnings Ability	(ROA)	> 1.5%	1.25 - 1.5%	1.01 - 1.25%	0.75 - 1.00%	< 0.75%
	(ROE)	> 22%	17 - 21.99%	10 - 16.99%	7 - 9.99%	< 6.99%
Liquidity (TL/TD)		< 55%	62 - 56%	68 - 63%	80 - 69%	> 81%

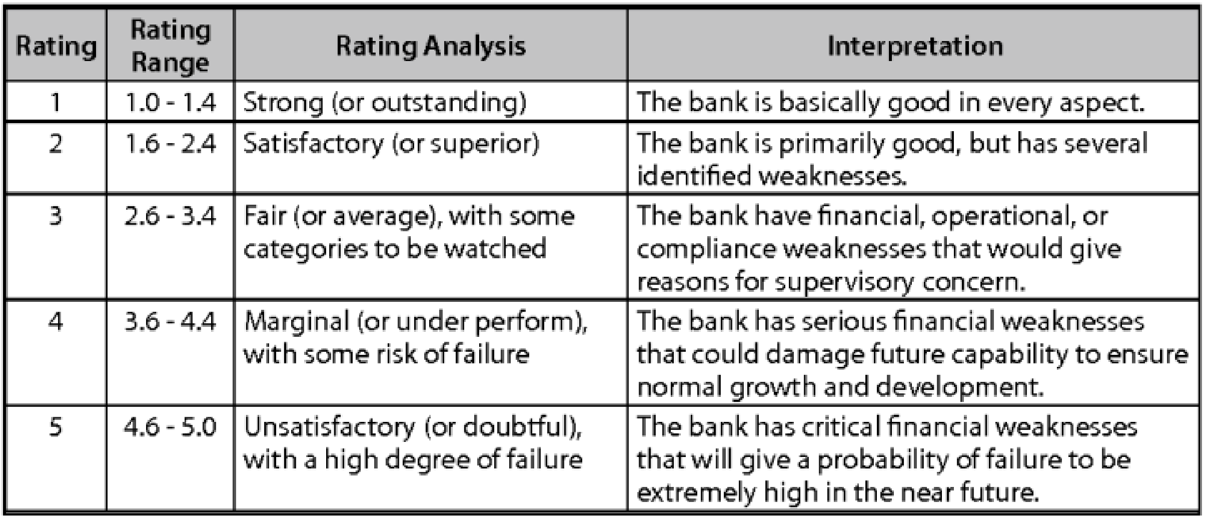
Figure 1: CAMELS Component

Source: Desta (2016)

Figure 1 shows the method to score each indicator based on (Rozzani and Rahman, 2013) and (Babar and Zeb, 2011) cited by (Desta, 2016). After scoring each factor in the model, the author calculated the final score of each bank for each year to determine the health status of the bank at different times. From there, it helps to identify categorical variables more conveniently. ***Overall CAMEL rating score*** formula:

$$\boxed{Target = \frac{Captital\ adequacy\ score\ +\ Asset\ quality\ score\ +\ Management\ score\ +\ Earning\ score\ +\ Liquidity\ score}{5}}$$

In which the component variables have been explained as above. After obtaining the value of the target variable, Overall CAMEL rating score, we will rely on the measurement table below to accurately distinguish current risks, and predict potential risks in the future. Determine the categorical variable by comparing the bank's final CAMEL score with Figure 2.



Rating	Rating Range	Rating Analysis	Interpretation
1	1.0 - 1.4	Strong (or outstanding)	The bank is basically good in every aspect.
2	1.6 - 2.4	Satisfactory (or superior)	The bank is primarily good, but has several identified weaknesses.
3	2.6 - 3.4	Fair (or average), with some categories to be watched	The bank have financial, operational, or compliance weaknesses that would give reasons for supervisory concern.
4	3.6 - 4.4	Marginal (or under perform), with some risk of failure	The bank has serious financial weaknesses that could damage future capability to ensure normal growth and development.
5	4.6 - 5.0	Unsatisfactory (or doubtful), with a high degree of failure	The bank has critical financial weaknesses that will give a probability of failure to be extremely high in the near future.

Figure 2: CAMELS Scoring

Source: Desta (2016)

Banks with a Camel overall score less than 3.4 will be considered to be in good stability and will be labeled as being in the “0” class which ranges from very good to satisfactory however there are several scoring factors. In contrast, banks with a final CAMEL score greater than or equal to 3.4 will be labeled “1” as banks with not good to very weak financial health status and high potential risk in the future.

Target Variable: “0” - *Healthy Bank and “1”* - *Unhealthy Bank*

3.2 Quantitative methods

This study presents a quantitative approach for forecasting the financial health status of Vietnam’s commercial banks using Machine Learning models. The study uses three algorithms, namely Logistics Regression, Support Vector Machine, and AdaBoost, to analyze financial data and classify banks based on their financial health. The analysis focuses on identifying the variables that primarily affect the classification ability of the model. To ensure the accuracy of the analysis, the input data is explored, cleaned, normalized, and transformed before building the models. The efficiency of the models is then evaluated using metrics from confusion matrix.

4. DATA

4.1 Data Source

The research data used in this study includes significant data about the activities of 45 Vietnamese banks over the period of 2002–2021, including deposits, loans, assets, and labor productivity. The dataset is the first comprehensive collection of data on the division between state-owned and private, as well as foreign-owned and domestic-owned, banks in Vietnam. It has a total of 643 observations covering each bank year.

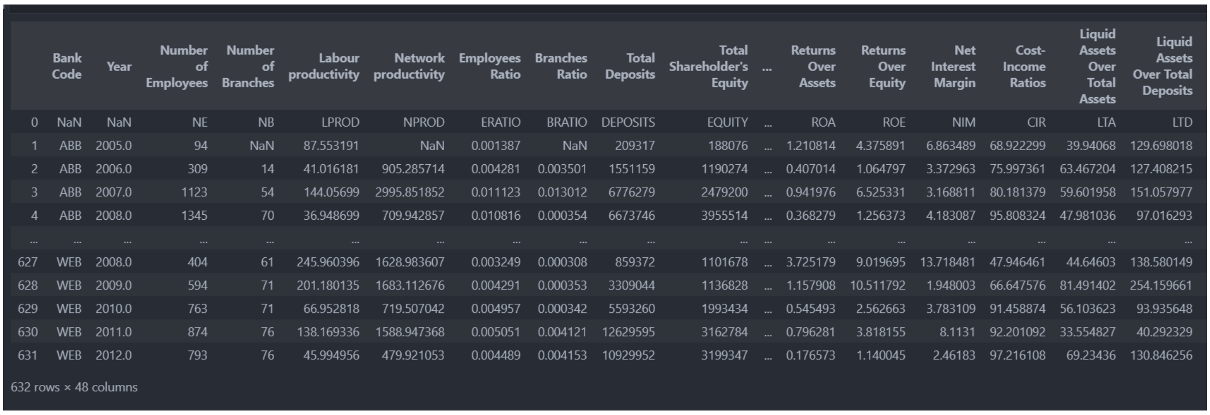
The dataset was created, verified, and published in the Harvard Dataverse by Le et al. (2022). The authors used a variety of sources, including annual reports, financial statements, and other publicly available data, to collect and validate the data to create the dataset. As a result, a special set of variables and indicators were produced, which accurately reflect the development and performance of the Vietnamese banking industry over time from a number of angles.

The dataset can be helpful for financial analysts, academics, and educators who are interested in banking efficiency and performance, risk and profit management, machine learning, and other relevant topics. It may also encourage more study of the Vietnam banking industry and aid in the development of industry-specific policies and strategies that are based on empirical research.

Le et al. (2022) description of the dataset's methodology, variables, and data quality control measures is very thorough. The dataset's characteristics and restrictions are detailed in their work, which interested parties are welcome to consult. Through the Harvard Dataverse website, the dataset is freely accessible to the general public.

4.2 Data Cleaning

In this study, the raw data consists of 632 rows and 48 columns.



	Bank Code	Year	Number of Employees	Number of Branches	Labour productivity	Network productivity	Employees Ratio	Branches Ratio	Total Deposits	Total Shareholder's Equity	...	Returns Over Assets	Returns Over Equity	Net Interest Margin	Cost-Income Ratios	Liquid Assets Over Total Assets	Liquid Assets Over Total Deposits
0	NaN	NaN	NE	NB	LPROD	NPROD	ERATIO	BRATIO	DEPOSITS	EQUITY	...	ROA	ROE	NIM	CIR	LTA	LTD
1	ABB	2005.0	94	NaN	87.553191	NaN	0.001387	NaN	209317	188076	...	1.210814	4.375891	6.863489	68.922299	39.94068	129.698018
2	ABB	2006.0	309	14	41.016181	905.285714	0.004281	0.003501	1551159	1190274	...	0.407014	1.064797	3.372963	75.997361	63.467204	127.408215
3	ABB	2007.0	1123	54	144.05699	2995.851852	0.011123	0.013012	6776279	2479200	...	0.941976	6.525331	3.168811	80.181379	59.601958	151.057977
4	ABB	2008.0	1345	70	36.948699	709.942857	0.010816	0.000354	6673746	3955514	...	0.368279	1.256373	4.183087	95.808324	47.981036	97.016293
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
627	WEB	2008.0	404	61	245.960396	1628.983607	0.003249	0.000308	859372	1101678	...	3.725179	9.019695	13.718481	47.946461	44.64603	138.580149
628	WEB	2009.0	594	71	201.180135	1683.112676	0.004291	0.000353	3309044	1136828	...	1.157908	10.511792	1.948003	66.647576	81.491402	254.159661
629	WEB	2010.0	763	71	66.952818	719.507042	0.004957	0.000342	5593260	1993434	...	0.545493	2.562663	3.783109	91.458874	56.103623	93.935648
630	WEB	2011.0	874	76	138.169336	1588.947368	0.005051	0.004121	12629595	3162784	...	0.796281	3.818155	8.1131	92.201092	33.554827	40.292329
631	WEB	2012.0	793	76	45.994956	479.921053	0.004489	0.004153	10929952	3199347	...	0.176573	1.140045	2.46183	97.216108	69.23436	130.846256

632 rows × 48 columns

Figure 3: Raw data

Source: Author

The first step in data cleaning was to remove any blank and useless columns that do not provide any useful information for the analysis. Next, the "Year" variable was converted to a DateTime format. The remaining variables were converted to numerical format to allow for further analysis and modeling.

Another important step in data cleaning was to identify columns with a high percentage of missing values and drop them from the dataset. In this case, columns with a missing value percentage greater than 9% were dropped to ensure that the remaining data was representative and not biased by missing values.

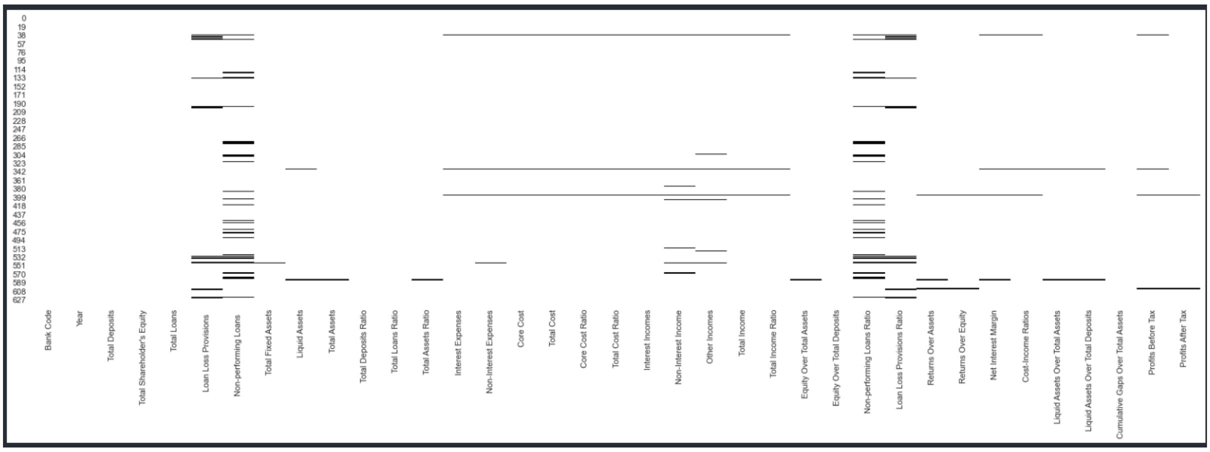


Figure 4: Missing values from data

Source: Author

Overall, these data cleaning steps were necessary to ensure that the dataset was of high quality and suitable for the subsequent analysis and modeling of financial stability forecasting for commercial banks in Vietnam using machine learning techniques.

4.3 Data Preprocessing

Before conducting any analysis or modeling, the data underwent a preprocessing phase to ensure its quality and suitability for the study. The following steps were taken in the data preprocessing phase:

4.3.1 Impute missing values

The missing values in the data were imputed using KNN imputation, a method that imputes missing values based on the values of the nearest neighbors. This helped ensure that the data remained as complete as possible. KNN imputation is a technique for filling in missing values in datasets, which has several advantages that make it suitable for financial data. This method preserves the original distribution of the data, uses available information, does not require assumptions, can handle both continuous and categorical variables, and reduces bias. These advantages make KNN imputation a reliable and flexible technique for filling in missing values in financial data, ensuring the accuracy and reliability of financial analyses and predictions.

4.3.2 Dealing with outliers

To deal with outliers in the dataset, a method was used to identify and alter them using a lower bound and an upper bound. This method involves setting a range of values, typically calculated as a multiple of the standard deviation, above and below the mean of the data. Any data point outside of these bounds was replaced with the value of the bound. This approach helps ensure that the data remains as representative as possible by minimizing the influence of extreme values that could skew the overall analysis. The lower bound and upper bound approach is a common technique used in data analysis to handle outliers and is often used in financial data analysis to ensure that the analysis is not skewed by extreme values that could affect investment decisions. However, it is important to note that altering outliers using this method can also result in the loss of potentially valuable information, so it is essential to carefully consider the potential impact on the analysis before using this approach.

4.4 Feature Selection

For the variables in the dataset used to calculate the target variable for the forecast will be discarded and not reused. Financial ratios that are similar in meaning and represent the CAMEL indices will also be removed to avoid high correlation and duplication in the data set. Moreover, the valuable financial index variables used to calculate the used indexes are also not suitable for the problem, so they are also removed. After removal, the remaining data set of 7 variables is used to run the model. These variables will be described in detail in the next section.

Table 1: List information of independent features

Variable	**Code**	**Note**
Non-Interest Expenses	NIE	
Profits Before Tax	PBT	
Total Deposits Ratio	DEPORATIO	= DEPOSITS/Total deposits of all available banks
Loan Loss Provisions	LLP	
Cumulative Gaps Over Total Assets	GTA	
Loan Loss Provisions Ratio	LLPRATIO	=LLP/LOANS
Target	Target	0 is Healthy Bank 1 is Unhealthy Bank

Source: Author

4.5 Variable

Table 2 : The statistical description of input features

	Loan Loss Provisions	Total Deposits Ratio	Non-Interest Expenses	Loan Loss Provisions Ratio	Cumulative Gaps Over Total Assets	Profits Before Tax
Count	614	614	614	614	614	614
Mean	835879.9	0.021288	1717864	1.19433	29.09488	1434938
Std	1015923	0.022047	2141522	0.640076	29.0491	1798033
Min	-674646	3.58E-05	269	-0.25367	-48.3686	-2876656
Median	343053.5	0.01195	628174	1.118466	33.54446	567442
Max	2887292	0.068268	6223097	2.632933	86.7251	5363616

Source: Author

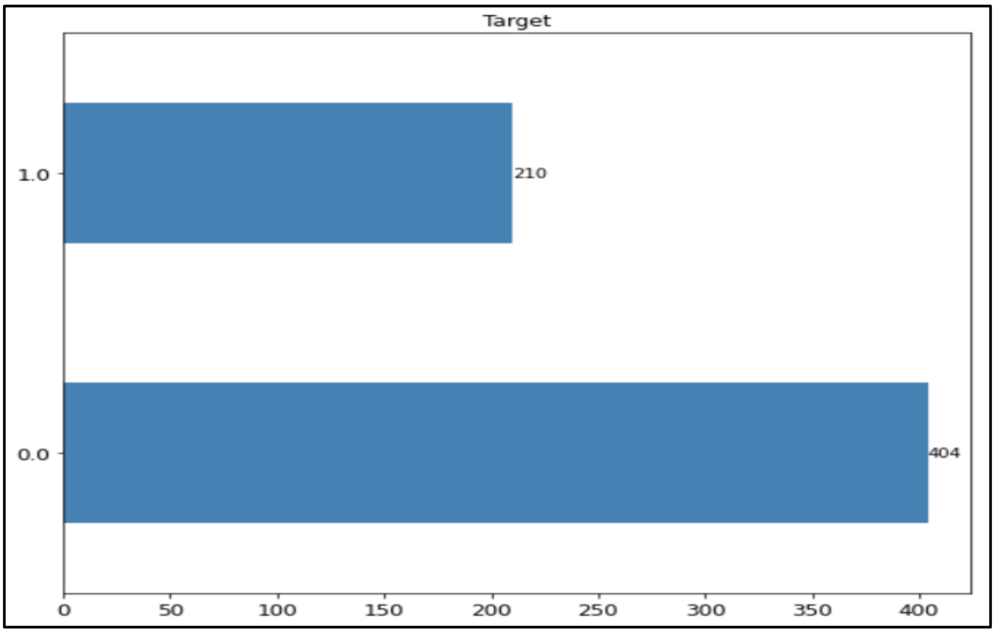


Figure 5 : Number of two classes of Target Variable

Source: Author

Statistics describe the variables used to run the predictive model. Figure 5 shows the number of observations of the target variable. Of which, 404 commercial banks in

Vietnam are classified as having good financial health and 210 banks with poor financial health.

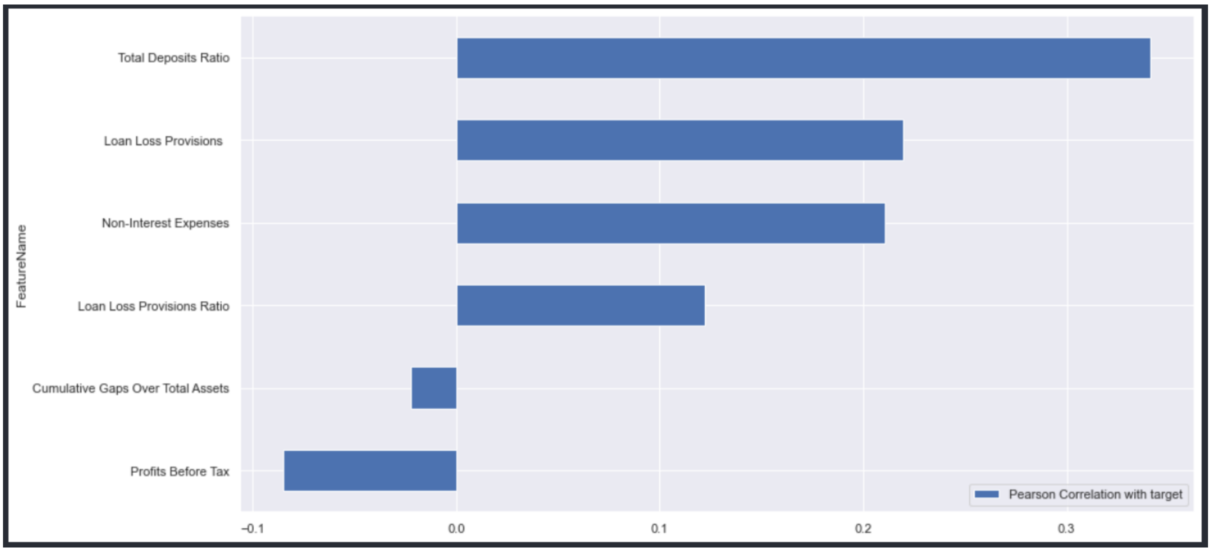


Figure 6 : Pearson Correlation with target variable

Source: Author

Figure 6 shows the correlation index of the independent variables with the target variable. 4 variables are positively correlated and 2 variables are negatively correlated. It is worth mentioning that all variables have a low correlation with the target variable.

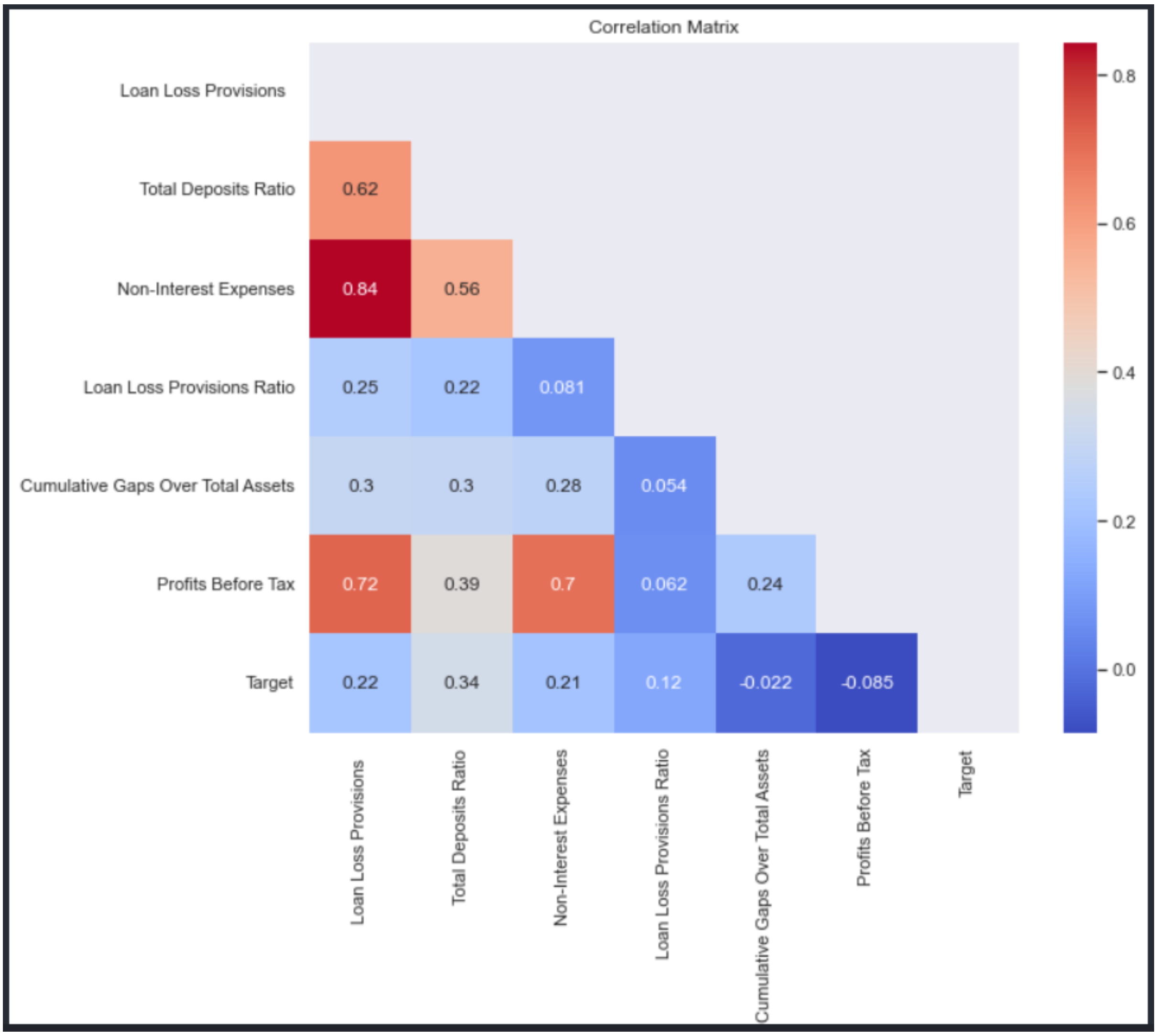


Figure 7 : Correlation matrix between all variables

Source: Author

Checking the level of correlation between variables can see that all variables have an acceptable and low correlation. Some cases have strong correlations from 0.7 or more, but very few and can be used to build predictive models.

4.6 Train-Test set

The complete dataset after processing has 614 rows and 7 columns. Because of limited data, the study will divide the train and test sets by the ratio of 80% for train and 20% for test. The train set includes 123 data rows, including 83 variables that are banks with stable health and 40 banks with unstable status.

5. RESEARCH RESULT

5.1 Feature Importance

Feature selection is a crucial process in machine learning, aimed at identifying and selecting the most impactful features from a dataset. It reduces complexity and dimensionality of the data, improves model performance, and minimizes overfitting risks.

One method of feature selection is using feature importance from DecisionTreeClassifier, which calculates and quantifies the contribution of each feature in the dataset to decreasing tree impurity (measured by Gini index or entropy). Features with higher importance scores are deemed more relevant to the target variable and can be selected for the model. This approach is popular and effective due to its simplicity, insight into feature significance, and ability to remove irrelevant features, resulting in improved accuracy and interpretability.

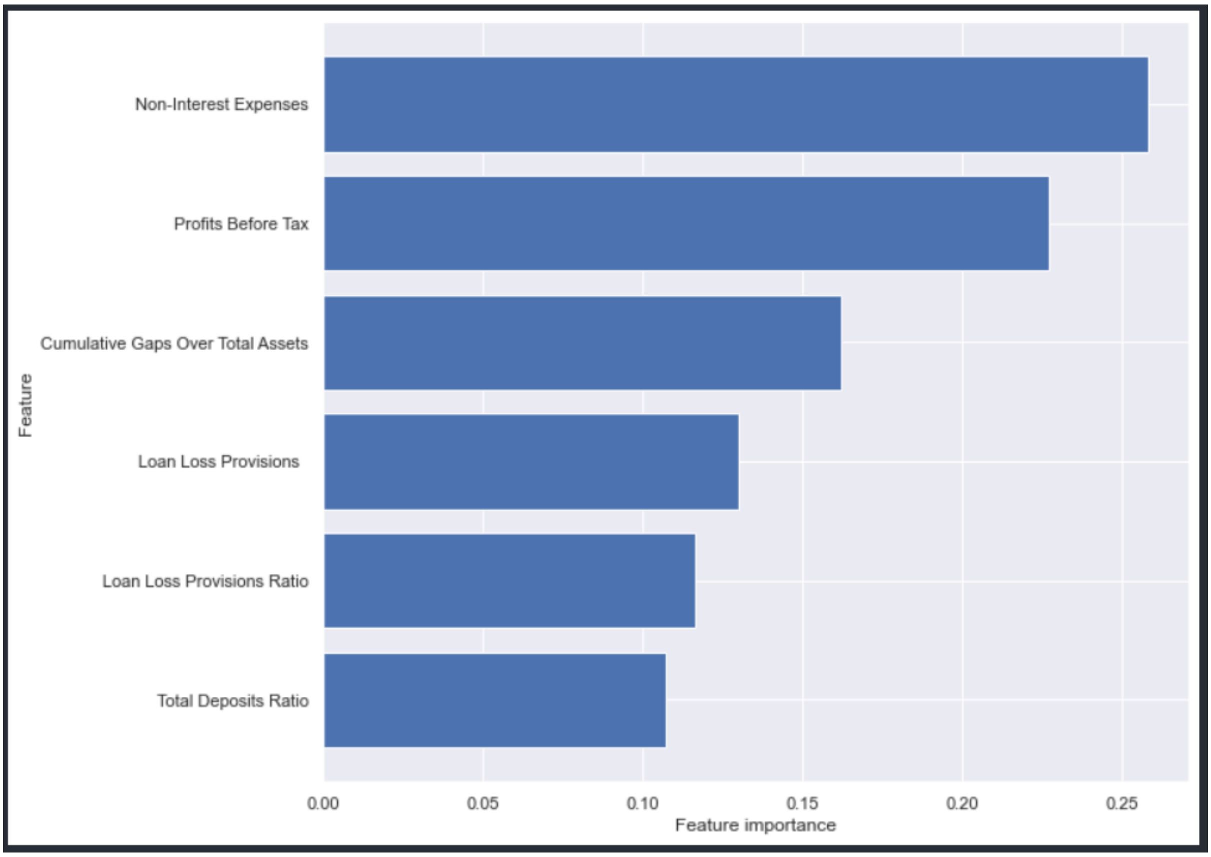


Figure 8 : Feature Importance of variables

Source: Author

From this step, the study finds three financial indicator variables that have the most impact on forecasting models are Non-Interest Expense, Profits Before Tax and Cumulative Gaps Over Total Assets. Besides, Total Deposit Ratio is the worst contributed ratio when classifiying.

5.2 Logistic regression

Table 3 : Classification Report of Logistic Regression

	precision	recall	f1-score	support
0.0	0.81	0.78	0.80	83
1.0	0.58	0.62	0.60	40
accuracy			0.73	123
macro avg	0.70	0.70	0.70	123
weighted avg	0.74	0.73	0.73	123

Source: Author

The model shows that a good class 0 prediction can correctly and accurately recognize this class. However, the metrics of class 1 are only at a good level, the model gives the ability to predict unstable banks with an accuracy of 58% on the predictions, but the accuracy compared to the number of observations actually 62%.

5.3 Support Vector Machines

Table 4 : Classification Report of Support Vector Machines

	precision	recall	f1-score	support
0.0	0.85	0.89	0.87	83
1.0	0.75	0.68	0.71	40
accuracy			0.82	123
macro avg	0.80	0.78	0.79	123
weighted avg	0.82	0.82	0.82	123

The SVM model exhibits a commendable accuracy of 82% when employed to predict the financial soundness of banks. Notably, it performs exceptionally well in predicting class 0, achieving an impressive accuracy rate of 89% in comparison to real-world data, while also achieving a favorable accuracy of 85% in comparison to forecasted values. Similarly to Logistic Regression, the recall index of layer 1 in the SVM model is 68%. However, it is noteworthy that the precision demonstrates significant improvement, contributing to an overall enhanced accuracy of the SVM model in both classes.

5.4 AdaBoost

Table 5 : Classification Report of Adaboost

	precision	recall	f1-score	support
0.0	0.79	0.83	0.81	83
1.0	0.61	0.55	0.58	40
accuracy			0.74	123
macro avg	0.70	0.69	0.70	123
weighted avg	0.73	0.74	0.74	123

Source: Author

The Adaboost model achieved an accuracy of 0.74 lower than SVM. In terms of precision, the model has a precision of 0.79 for class 0 and 0.61 for class 1, indicating that when the model predicts a bank as financially healthy (class 0), it is correct 79% of the time, and when it predicts a bank as not financially healthy (class 1), it is correct 61% of the time. The recall value for class 0 is 0.83, indicating that the model correctly identifies 83% of the financially healthy banks, and the recall value for class 1 is 0.55, indicating that the model correctly identifies 55% of the financially unhealthy banks. Overall, the model is still able to recognize well in both classes, however, the biggest limitation of the model is the accuracy compared to the reality of class 1, the class that needs to be predicted accurately is very limited.

5.5 Model evaluation

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model. It shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for different threshold values. The Area Under the Curve (AUC) is a numerical measure of the overall performance of the model, with a perfect classifier having an AUC of 1.0.

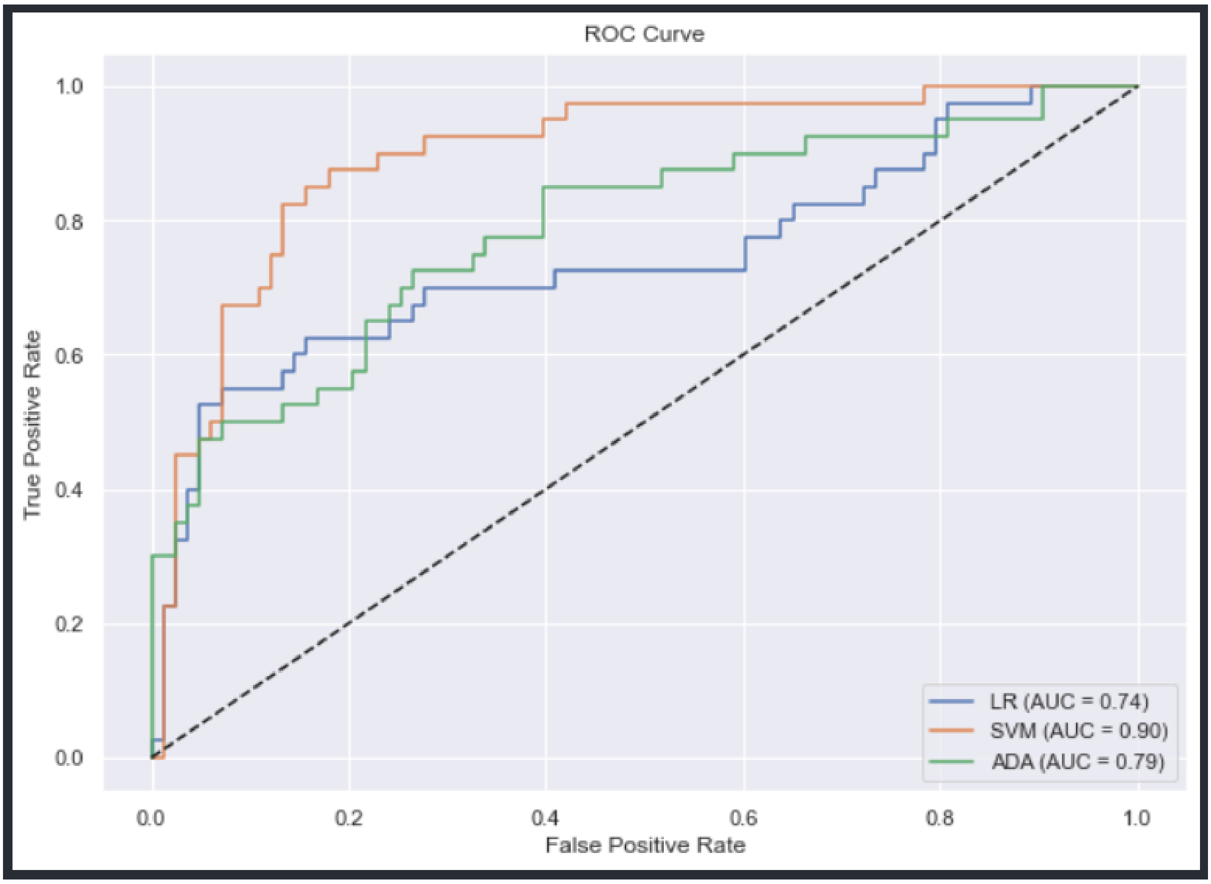


Figure 9 : Receiver Operating Characteristic and Area Under the Curve of all models

Source: Author

The models were evaluated using the receiver operating characteristic curve (ROC) and area under the curve (AUC) metrics. The results were highly promising and this showed that the model was doing well. All three models achieved good AUC scores, with Logistics Regression scoring 0.74, Support Vector Machine scoring 0.90, and Adaboost scoring 0.79. These scores indicate that the models were highly effective at predicting the financial stability of the banks in our dataset, with SVM emerging as the most accurate of the three.

The ROC curves for all three models were plotted on the same graph to visualize their performance. The curves show that all three models have a high TPR for a relatively low FPR. The closer the ROC curve is to the top-left corner, the better the model's performance. From the ROC curves, it is evident that the Adaboost model outperformed the other two models, followed by the Support Vector Machine and Logistics Regression models.

The models have a strong ability to distinguish between the two classes, can effectively differentiate between good and not good stability in Vietnam commercial banks banks and make accurate predictions. Especially, the higher the AUC, the better the A SVM model's performance, and a value of 0.90 indicates that the model has a very good discriminatory ability. Overall, the results of this study can be used to inform stakeholders in the banking sector and contribute to the ability to identify the health status of banks in Vietnam.

5.6 Interpretation of the results and implications for stability in Vietnam commercial banks

Based on the results of the four models, we can see that Logistic Regression and AdaBoost have the highest accuracy with 73% and 74%, while SVM has an accuracy of 82%. The precision, recall, and f1-score for all models are above 50%, indicating that they have good performance in predicting the financial soundness of Vietnam commercial banks. The ROC and AUC scores of the 3 models showed the result is from 0.79 to 0.90. With the SVM algorithm for the highest and lowest scores is Logistic Regresion.

In terms of implications for stability in Vietnam banks, these models can help identify potential risks and opportunities for improvement. By using the models to analyze financial data, banks can make more informed decisions about their operations, such as adjusting lending practices, investment strategies, or risk management practices. Additionally, the models can help regulators monitor the health of the banking sector and take appropriate actions to mitigate risks to the economy.

6. CONCLUSION

6.1 Summary of the main points

Based on our analysis of the stability of Vietnam commercial banks using four different machine learning models, we have found that AdaBoost and SVM have the highest accuracy in predicting the financial soundness of banks. SVM also performed well, achieving biggest AUC score.

The results suggest that SVM may be the most effective model for predicting financial soundness and stability in Vietnam commercial banks. However, the Logistic regression and AdaBoost models also show promising results, and they may be useful in situations where computational resources are limited or where simpler models are preferred.

The study finds Non-Interest Expense, Cumulative Gaps Over Total Assets and Profits Before Tax. Besides are three of the most crucial variables for predicting. The suggestion when using models that consider and pay close attention to these important metrics because they really mean something for banks.

Our findings suggest that machine learning models can be used as effective tools to predict the financial health of Vietnamese banks with high accuracy. This can be particularly useful for stakeholders in the banking industry, such as investors, regulators, and people with no prior understanding of banking and finance, to monitor and manage the financial risks associated with banks.

In conclusion, our study highlights the potential of machine learning models to classify and predict the stability of banks, and we recommend that stakeholders in the banking industry consider incorporating these models into their risk management practices. In order to get the most complete picture, it is also necessary to compare this method with other ones used to assess the bank's stability.

6.2 Limitations of the study

However, it will take a long time to confirm with a developing financial market like Vietnam because the current data is quite sparse and the level of financial distress in

Vietnamese banks is still low to ensure that the accuracy of the model in relation to reality. The calculated ratio of CAMEL is not assured to be representative for the bank industry in Vietnam.

For the Management (M) component, need more data about financial ratios can proxy for this or find another way to measure this factor more exactly. For the Sensitivity to Market Risk factor in CAMELS, there is still a need for studies incorporating this factor in the scoring process to apply to future studies and needs to be more harmonized to identify which of the 6 factors has the most weight in scoring the financial soundness of banks.

In addition, there is no truly absolute and better outcome than stability forecasting models. More data as well as better data processing methods are needed to make models less flawed because banks are an important pillar in the economy, if errors occur, it will greatly affect users and the development of the economy.

6.3 Future work

In the future, this study needs to add many other models to check the stability status of banks as a basis to compare the effectiveness of the current model. Provide a second system of evaluation criteria to test the authenticity and accuracy of the input dependent variable. Besides, applying advanced techniques in machine learning to improve the quality of the model. Collect more data in annual and financial reports, also data from the market to analyze the effects of them on the stability of Vietnam commercial banks. Research more about the use of CAMEL for Vietnam and must find a better way to adapt well in the reality.

REFERENCES

[1] Affes, Z., & Hentati-Kaffel, R. (2019). Predicting US banks bankruptcy: logit versus Canonical Discriminant analysis. Computational Economics, 54, 199-244.

[2] Affes, Z., & Hentati-Kaffel, R. (2019). Forecast bankruptcy using a blend of clustering and MARS model: Case of US banks. Annals of Operations Research, 281(1-2), 27-64.

[3] Balgova, M., Nies, M., & Plekhanov, A. (2016). The economic impact of reducing non-performing loans (Working Paper No. 193). European Bank for Reconstruction and Development.

[4] Csikosova, A., Janoskova, M., & Culkova, K. (2019). Limitation of financial health prediction in companies from post-communist countries. Journal of Risk and Financial Management, 12(1), 15.

[5] Derviz, A., & Podpiera, J. (2008). Predicting bank CAMELS and S&P ratings: the case of the Czech Republic. Emerging Markets Finance and Trade, 44(1), 117-130.

[6] Desta, T. S. (2016). Financial performance of “The best African banks”: A comparative analysis through CAMEL rating. Journal of accounting and management, 6(1), 1-20.

[7] Ekinci, A., & Erdal, H. İ. (2017). Forecasting bank failure: Base learners, ensembles and hybrid ensembles. Computational Economics, 49(4), 677-686.

Fachrudin, K. A. (2021). Insolvency and financial health prediction model for the listed companies on the Indonesia Stock Exchange. Jurnal Akuntansi dan Auditing Indonesia, 24-32.

[8] Gaul, L., Jones, J., & Uysal, P. (2019). Forecasting High-Risk Composite CAMELS Ratings.

[9] Gilbert, R. A., Meyer, A. P., & Vaughan, M. D. (2000). The role of a CAMEL downgrade model in bank surveillance. Federal Reserve Bank of St. Louis Working Paper Series, (2000-021)

[10] Giuffrè, M., Moretti, R., & Tiribelli, C. (2023). Gut Microbes Meet Machine Learning: The Next Step towards Advancing Our Understanding of the Gut Microbiome in Health and Disease. International Journal of Molecular Sciences, 24(6), 5229.

[11] Halteh, K., Kumar, K., & Gepp, A. (2018). Financial distress prediction of Islamic banks using tree-based stochastic techniques. Managerial Finance.

[12] Hays, F. H., De Lurgio, S. A., & Gilbert, A. H. (2009). Efficiency ratios and community bank performance. Journal of Finance and Accountancy, 1(1), 1-15.

[13] Hiển, N. T. (2017). Ứng dụng các mô hình tiêu chuẩn nhằm phân tích tình hình tài chính Vietinbank.

[14] Huang, D. T., Chang, B., & Liu, Z. C. (2012). Bank failure prediction models: For the developing and developed countries. Quality & Quantity, 46(2), 553–558.

IMF and World Bank 2005, Financial sector assessment: A handbook, Interntional Monetary Fund and the World Bank.

[15] Karim, M. R., Shetu, S. A., & Razia, S. (2021). COVID-19, liquidity and financial health: empirical evidence from South Asian economy. Asian Journal of Economics and Banking, 5(3), 307-323.

[16] Kumar, MA, Harsha, GS, Anand, S & Dhruva, NR 2012, 'Analyzing soundness in Indian banking: A CAMEL approach', Research Journal of Management Sciences, vol. 1, p. 1171.

[17] Le, H. H., & Viviani, J. L. (2018). Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. Research in International Business and Finance, 44, 16-25.

[18] Le, T. (2017). Financial soundness of Vietnamese commercial banks: A CAMELS approach. Available at SSRN 3068529.

[19] Le, T. D., Ho, T. H., Ngo, T., Nguyen, D. T., & Tran, S. H. (2022). A Dataset for the Vietnamese Banking System (2002–2021). Data, 7(9), 120.

[20] Luong Duy Quang, (2015), *The research of Determinants of Banking Crisis*

Meitei, A. J., Arora, P., Mohapatra, B. B., & Arora, H. (2022). Identification of Weak Banks Using Machine Learning Techniques: Evidence from the Indian Banking Sector. Global Business Review, 09721509221113631.

[21] Nguyen, A. H., Nguyen, H. T., & Pham, H. T. (2020). Applying the CAMEL model to assess performance of commercial banks: empirical evidence from Vietnam. Banks and Bank Systems, 15(2), 177.

[22] Persons, O. (1999). Using financial information to differentiate failed vs. surviving finance companies in Thailand: an implication for emerging economies. Multinational Finance Journal, 3(2), 127-145.

[23] Quang, L. D. (2015). Determinants of banking crisis: The case of Vietnam. HO CHI MINH CITY OPEN UNIVERSITY JOURNAL OF SCIENCE-ECONOMICS AND BUSINESS ADMINISTRATION, 5(2), 64-81.

[24] Schapire, R. E. (2013). Explaining adaboost. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, 37-52.

[25] Siddique, M. M. (2022). Banking performance of First Security Islami Bank Limited, measured by CAMELS rating system.

[26] Thomson, J. B. (1991). Predicting bank failures in the 1980s. Federal Reserve Bank of Cleveland Economic Review, 27(1), 9-20.

[27] Tung, W. L., Quek, C., & Cheng, P. (2004). GenSo-EWS: A novel neural-fuzzy based early warning system for predicting bank failures. Neural Networks, 17(4), 567–587.

[28] Viswanathan, P. K., Srinivasan, S., & Hariharan, N. (2020). Predicting financial health of banks for investor guidance using machine learning algorithms. Journal of Emerging Market Finance, 19(2), 226-261.

[29] Vuong, Q. H., Napier, N. K., & Tran, T. D. (2013). A categorical data analysis on relationships between culture, creativity and business stage: the case of Vietnam. International Journal of Transitions and Innovation Systems, 3(1), 4-24.