VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

FINAL REPORT

SCIENTIFIC RESEARCH TOPIC STUDENTS

2022-2023

Topic:

THE IMPACT OF MARGIN LENDING ON PREDICTING DIFFICULTIES OF VIETNAMESE ENTERPRISES

Scientific field: Economic field

Specialization: Finance - Banking

Group members:

NO	NAME	ID	Majors	Mission	Phone	Email
1.	Đặng Đức Duy	K205030798	Finance – Banking	Leader	0982789264	duydd20503@st.uel.edu.vn
2.	Phạm Thị Ngọc Thanh	K204141930	Finance – Banking	Member	0855730384	thanhptn20414c@st.uel.edu.vn
3.	Hứa Hoàn Châu	K204141909	Finance – Banking	Member	085599282	chauhh20414c@ss.uel.edu.vn
4.	Nguyễn Mai Phương	K204141926	Finance – Banking	Member	0987488678	phuongnm20414c@st.uel.edu.vn
5.	Huỳnh Thị Hồng Nhung	K204141925	Finance – Banking	Member	0947864250	nhunghth20414@st.uel.edu.vn

VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

FINAL REPORT

SCIENTIFIC RESEARCH TOPIC STUDENTS

2022-2023

Topic:

THE IMPACT OF MARGIN LENDING ON PREDICTING DIFFICULTIES OF VIETNAMESE ENTERPRISES

Đại diện nhóm nghiên cứu	**Giảng viên hướng dẫn**	**Chủ tịch Hội đồng**
(Ký, họ tên)	*(Ký, họ tên)*	*(Ký, họ tên)*

Lãnh đạo Khoa/Bộ môn/Trung tâm

(Ký, họ tên)

ABSTRACT

This study uses Machine Learning models to predict and evaluate profits and also identify factors that influence organizational profitability. The research was based on actual data from the financial statements of 512 companies listed on the Vietnamese stock exchange from 2010 to 2020. Recognizing the limits of past studies' techniques, the team proposes to perform profit forecasting research using two ways: by year and by industry, which yields two important results. First, the model's profit or loss performance is fairly strong, particularly for the 2012 and 2020 models, as well as the healthcare model. Second, in the data fields, ROA, Net profit margin, ROE, and DSO are qualities that have a significant impact on profit forecasting. Furthermore, unlike the model decomposed by industry, the model decomposed by years does not account for the effect of macro variables.

TABLE OF CONTENT

LIST OF TABLES

LIST OF PICTURES

LIST OF ACRONYMS

ANB	Augmented naive Bayes
ANN	Artificial neural network
AUC	Area Under The Curve
DM	Data mining
GBDT	Gradient boosting decision trees
GBM	Gradient boosting machines
HNX	Hanoi Stock Exchange
HOSE	Ho Chi Minh City Stock Exchange
KNN	K-nearest neighbors
ML	Machine learning
NAV	Net asset value
NPAT	Net profit after tax
RF	Random forest
RMNL	Random MultiNomial Logit
SMOTE	Synthetic Minority Over-sampling
SVM	Support Vector Machine
UPCom	Unlisted Public Company Market
ANB	Augmented naive Bayes
ANN	Artificial neural network
AUC	Area Under The Curve
DM	Data mining
GBDT	Gradient boosting decision trees
GBM	Gradient boosting machines

HNX	Hanoi Stock Exchange

CHAPTER 1: INTRODUCTION

1.1. Background and justification for the research project

In fact, in Vietnam or in any other country, financial difficulties are always a potential and core problem that every company must always put on top to closely monitor. The prediction contributes significantly to the process of changing business strategies as well as stabilizing economic recovery. Realizing that important role, the research team has focused on researching and researching in recent years. Topic: "*The impact of margin lending on predicting financial difficulties of Vietnamese enterprises*". Up to now, Vietnam has had a lot of research work on this issue, but in general, it still follows the path of no breakthrough innovation. The traditional approach is considered effective, but it takes a lot of time and effort. Instead, the Machine Learning approach proposed by the team in the current context of continuous digital transformation is to overcome the above disadvantages. The target variable in the article is whether the company falls into one of the financial difficulties leading to a trading warning, delisting on the stock exchange or not at a specified time based on the following factors... In addition, this study also contributes to complement previous studies through outstanding features such as: analyzing data by stock exchange and by year, applying legal documents.

1.2. Study overview

Margin loan eligibility is one of the important factors for assessing a company's financial position. Because in order to be granted margin, the company's shares often have to meet strict criteria such as being listed on the market for 6 months or more, profitable business results according to the consolidated audited financial statements. for the most recent year and the latest quarterly financial reporting period. The consolidated financial statements are reviewed, audited, fully accepted by a reputable auditing company and fully disclosed in accordance with regulations. Shares are not subject to warning, control, special control, trading suspension or delisting according to relevant regulations on securities listing. Therefore, it is possible to rely on the strict conditions of margin lending to make forecasts about the stable or difficult financial situation of Vietnamese companies listed on the stock exchange.

1.2.1. Domestic studies

In Vietnam, there are many research papers on the financial difficulties of companies listed on the Vietnamese stock market. Specifically, the following articles that our group has synthesized:

The study on the effect of cash flow on financial distress of the authors (Do Thi Van Trang, 2022). The authors used Bayesian approach and variables such as operating cash flow, cash

flow from investing activities, cash flow from financial activities, age of the enterprise, and size of the business. Research results showed that enterprises have increased operating cash flow, investment cash flow, and financial operating cash flow, reducing the risk of financial distress of enterprises.

In the research paper of the author Pham Thi Hong Van in 2018, the author used the Binary Logistic Regression model to measure the financial distress of listed companies on the stock market. The variables included in the model were current ratio (CA/CL), operating cash flow ratio (OCF/TA), debt ratio (Debt/TA), and retained profit ratio (RE). /TA), size of DN (Size). The research results showed that the current ratio and debt ratio had a positive impact on financial distress, a negative relationship between firm size and financial distress. Also, operating ratio and profit ratio did not show any dramatic relationship with financial distress.

1.2.2. Foreign studies

Besides, there are also international studies on the factors affecting the profitability of enterprises:

Financial Distress Prediction Through Cash Flow Ratios Analysis (Amrizah Kamaluddin, 2019). This study aims to use cash flow ratio to predict financial distress in industrial and consumer goods manufacturing companies in Malaysia through liquidity ratios, liquidity ratios, and liquidity ratios. The efficiency ratios and profitability ratios used in this study are taken from the cash flow statement and the Altman Z-score is used as the dependent variable. The results show that the solvency ratio has a negative impact on the profitability ratio and the financial distress of the enterprise. Other ratios, namely CFFO+INT/INT and CFFO/FA show an insignificant negative relationship. This indicates that efficiency ratio had no relationship with financial distress whilst solvency ratio has a mixed relationship on financial distress. Thus, it can be concluded that solvency and profitability cash flow ratio have a relationship with the financial distress. As for the control variable, the result shows that there is no relationship between the firms' size with financial distress.

Financial Distress Prediction for Small and Medium Enterprises Using Machine Learning Techniques (Aidas Malakauskas, Ausrine Lakstutiene, 2021) The study used data from 12,000 small and medium-sized companies and Machine Learning algorithms such as Logistic Regression, Artificial Neural Networks and Random with 23 independent variables to predict financial distress of enterprises. To improve the model, we added a time element instead of just one-year financial index. In addition, the authors have shown that using Random Forest is the best method for classifying static periods. In general, multi-period classification is a superior approach to a static period predictor, which suggests that time factors have reduced uncertainty by adjusting for time factors. adjusted for periods of risk.

Using neural networks and data mining techniques for the financial distress prediction model (Wei-Sen Chen, Yin-Kuan Du, 2009). The study was conducted on a sample of 68 listed companies with 34 companies facing financial difficulties and 34 normal companies in the same industry. Besides using 37 ratios, the authors also use the artificial neural network (ANN) and data mining (DM). As a result, the more factor analysis is used, the lower the accuracy obtained by ANN and DM methods. The closer we get to the actual financial distress, the better the accuracy we get, with an accuracy percentage of 82.14% in the two seasons prior to the financial distress, developing a financial difficulty prediction model, ANN method achieves better prediction accuracy than DM clustering method.

Finaical distress prediction in Indonesia companies: Finding an alternative model (Anggraini Dew, Mulya Hadri, 2017). The study is based on data of companies in Indonesia from 2006 to 2015. The study also uses dependent variables as financial indicators such as Working Capital to Total Assets; Current Ratio; Book value of equity to total liability; Total Debt to Total Assets; EBIT to Current Liabilities; and Institutional Ownership. In addition to the above variables, the authors also use a macro variable to represent the conditions faced by companies in Indonesia. As a result, businesses can use the model to assess the company's financial position, contributing to preventing financial difficulties that businesses encounter.

1.3. Objectives of the study

This study aims to identify and evaluate the factors affecting margin lending of companies listed on the Vietnam stock market, contributing to determining the companies' financial difficulties. Forecasting corporate financial distress is increasingly important due to having a significant impact on the lending decisions and profitability of financial institutions.

By using Machine Learning applications, the study finds out the factors affecting the margin lending situation at enterprises at a certain time, thereby predicting the financial difficulties of enterprises. Besides, the application of Machine Learning to research helps to process large amounts of data with high diversity, maximizing prediction accuracy. Finally, the research is carried out on current data of enterprises and industries in order to be relevant to Vietnam's economic conditions and provide important information for domestic and foreign managers and investors.

1.4. Research scope

Scope of space: Enterprises listed on Vietnam stock exchange. (HNX& HOSE).

Nowadays, regulations only allow stocks listed on HOSE and HNX to be traded on margin. The shares on UPCom floor have not been margined. Therefore, the research team only chooses on HNX and HOSE and excludes businesses on UPCom.

Time scope: 3-year research period from 2019 to 2021.

1.5. Research Methods

Data collection method:

Qualitative methods: calculating and integrating elements impacting the financial difficulties and seeking for new factors. This method requires the researchers deeply understanding the nature of the research issue, specifically, the target variable and features.

Quantitative methods: Financial difficulties forecasting for enterprises based on Machine Learning models using historical figures and scale them to the same rule. From there, the direction of future profit movement may be determined

Empirical research method: After collecting a quantitative data set, researchers can analyze, build the model and print the results, then verify the results by measuring the experimental probability of the study based on the real data.

Theoretical research method: researchers collect information through articles, documents, and other scientific studies to find and select the basic concepts and ideas that are the basis for the theory of the topic, form scientific hypotheses, predict the properties of research objects, and build the model

1.6. Expected results

If applied and researched effectively, the research will bring businesses more accurate financial difficulty warnings. At the same time, the study examines related uncertainties and factors affecting margin lending status of public companies such as economic data, subjective factors of enterprises, thereby supporting managers and investors to make the right investment decisions, thereby minimizing risks for the company.

1.7. New point of research

The study assesses the difficult situation of enterprises based on the status of margin lending, specifically Decision 87/QD-UBCK in 2017 (amended by Article 1 of Decision 1205/QD-UBCK in 2017), which will be suitable for the economic context of Vietnam.

Year-to-year and exchange analysis can assist in determining whether there is a substantial change in the elements that influence a company's profitability indicators from one year to the next. If yes, what caused the shift (from objective market mechanism variables) so that the analyst can capture the market's major trend throughout time in the future.

CHAPTER 2: THEORETICAL BASIS AND REALITY OF THESIS

2.1. Theoretical basis

2.1.1. Machine learning

Up to now, Machine Learning is a term that is not too unfamiliar to us. According to Tom Mitchell in his book Machine Learning: “Machine learning (ML) is a field of inquiry devoted to understanding and building methods that "learn" – that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence.” Or in the same book, he also states:

“The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, negative with experience E.”

Machine learning algorithms build a model based on the data set to make predictions or decisions. The data set used in the model is divided into two sets: training set and testing set. The training set is the initial data used to train the model. After building the model, the test set is used to evaluate how well the algorithm has been trained with the trained data set.

Machine Learning algorithms are usually classified into two groups. One is based on the learning method, the other is based on the function of the algorithm. Grouping based on learning methods includes algorithms such as: Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, Reinforcement Learning. With functional classification, there are algorithms such as: Linear Regression, Logistic Regression, Decision Tree, Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Random Fores, etc

2.1.2. Reasons for choosing Machine learning

Machine learning and statistical techniques are now used in pattern recognition, knowledge discovery, and data mining. Despite the fact that the figure below depicts them as almost exclusive, the two fields are converging, and they share the same goal: learning from data. These methods concentrate on gaining knowledge or insights from data. However, their methods are influenced by their inherent cultural differences. Machine learning models are built to make the most accurate predictions possible. Statistical models are intended to make inferences about the relationships between variables. Statistics has long been used to analyze data and draw conclusions. The sample space, the family of events, and the probability measure are the three most common components of these models. Statistical modeling techniques are most commonly used on low-dimensional data sets. Regression is

a classic example, in which one or more variables are used to determine the effect of each explanatory variable on the independent variable.

On the contrary, because there are few to no assumptions, Machine Learning modeling tolerance is much higher than statistics. Machine learning is used to predict outcomes, and its performance is measured by how well it generalizes to new data that it has not yet learned. The algorithm analyzes the data, identifies patterns, and forecasts on the new data set. It is most commonly used on high-dimensional data sets; the more data you have, the more accurate your prediction.

Statistics enables researchers to answer scientific questions about the causal impact of a particular variable on a desired outcome. Analysts use statistics to assess the impact of a redistributive policy on the distribution of wealth across a country's population, for example. Companies, on the other hand, may use Machine Learning to predict a customer's ability to repay a loan.

In this study, our aim is to predict which enterprises are in financial difficulties. Therefore, using Machine Learning will be a more optimal method and also a newer method than previous studies using regression statistics to predict results.

2.1.3. Financial difficulty

Financial failure occurs when a firm suffers chronic and serious losses or when the firm becomes insolvent with liabilities that are disproportionate to its assets, or there is a shortage of assets such as cash and other assets, leading to the risk of not being able to meet its payment obligations, which could lead the business is forced to close or go bankrupt at the request of its creditors.(Vu Thi Loan, 2017) The company is in financial difficulties facing the cash flow problems or cash shortage in their operation when they are unable to generate sufficient cash to supersede the current obligation (Outcheva, 2007). Common causes and symptoms of financial failure include lack of financial literacy, failure to plan capital, poor debt management, inadequate protection against unforeseen events, and difficulty in compliance discipline to operate properly in the financial markets. The common assumption underlying bankruptcy prediction is that a company's financial statements appropriately reflect the above characteristics. In this study, we consider and assess the financial difficulty of the company through the status of margin lending at the company.

According to Clause 10, Article 2 of *Circular 120/2020/TT-BTC*, Margin trading at a securities company (hereinafter referred to as margin trading) is a transaction to buy securities using borrowed money from a securities company, in which securities obtained from this transaction and other securities traded on margin by investors are used as security for the above loan. Pursuant to Article 3 of the Regulation on guiding securities margin trading issued by the State Securities Commission together with *Decision 87/QD-UBCK in*

2017 (amended by Article 1 of Decision 1205/QD-UBCK in 2017) Regulations on securities not eligible for margin trading are as follows:

1. Securities with listing period of less than 06 months from the first trading date to the time of consideration and selection for margin trading. In the case of securities moving to the listing floor, the listing time is calculated as the total time of listing at two Stock Exchanges;

2. Securities listed under warning, under control, under special control, suspended from trading, subject to delisting in accordance with relevant regulations on securities listing;

3. Securities of the issuer whose annual financial statements are audited or semi-annual financial statements are reviewed or audited with opinions that are not fully accepted by the auditing organization;

4. The listing organization is late in disclosing information about its audited annual financial statements and reviewed semi-annual financial statements for more than 5 working days from the date of expiry of information disclosure or the expiration of the time limit for public disclosure. disclose information according to regulations;

5. The Stock Exchange receives a report or information disclosure from a listed company or the Stock Exchange has information on:

- Decision of the person competent to sanction administrative violations of the listed company for acts of tax evasion or tax fraud;

- The decision of the person competent to sanction administrative violations of the listed company for the act of failing to comply with the conclusion on enforcement of the tax administrative decision;

- The decision to prosecute the accused of the agency conducting the proceedings against the listed company.

6. Business results of listed organizations with losses at the review period and/or accumulated losses based on the most recent audited financial statements or the most recently reviewed or approved semi-annual financial statements. audit. In case the listed organization is the parent company, the business results are based on the consolidated financial statements; In case the listed organization is a public investment fund with at least one month's net asset value (NAV) calculated per unit of fund certificates less than the par value based on the net asset value change report monthly for 3 consecutive months up to the selected time for margin trading.

If the company falls into one of the following 6 cases, it could be considered to be in financial difficulty.

2.2. Algorithm

2.2.1. Logistic

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category.

Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$Logit(pi) = \frac{1}{1 + \exp(-pi)}$$

$$\ln\left(\frac{pi}{1 - pi}\right) = Beta_0 + Beta_1 * X_1 + \cdots + B_k * K_k$$

2.2.2. Native Bayes

Naive Bayes is the simplest form of Bayesian network, in which all attributes are independent given the value of the class variable. This is called conditional independence. It is obvious that the conditional independence assumption is rarely true in most real-world applications. A straightforward approach to overcome the limitation of naive Bayes is to extend its structure to represent explicitly the dependencies among attributes. An augmented naive Bayesian network, or simply augmented naive Bayes (ANB), is an extended naive Bayes, in which the class node directly points to all attribute nodes, and there exist links among attribute nodes.

A classifier is a function that assigns a class label to an example. From the probability perspective, according to Bayes Rule, the probability of an example E = (x1, x2,..., xn) being class c is

E is classified as the class C = + if and only if fb(E) = p(C = +|E) p(C = −|E) ≥1

where fb(E) is called a Bayesian classifier

$$p(c|E) = \frac{p(E|\, c)p(c)}{p(E)}$$

2.2.3. K-Nearest Neighbors

KNN (K-Nearest Neighbors) algorithm is a machine learning algorithm that classifies and predicts data. It is a non-parametric method. In the KNN algorithm, "K" represents the number of nearest neighbors of a data point in the data set. KNN works by comparing the distance between the new data point and the known data points in the training set. These known data points are called "nearest neighbours". The algorithm will choose the nearest neighbors and based on them to predict the label for the new data point. KNN algorithms are widely used in many fields such as image processing, stock market prediction, text classification and many other applications.

2.2.4. Decision tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

2.2.5. XG boost

XGBoost is a robust machine-learning algorithm that can help understand data and make better decisions.

XGBoost is an implementation of gradient-boosting decision trees. It has been used by data scientists and researchers worldwide to optimize their machine-learning models.

XGBoost offers regularization, which allows you to control overfitting by introducing L1/L2 penalties on the weights and biases of each tree. This feature is not available in many other implementations of gradient boosting.

XGBoost is used for these two reasons: execution speed and model performance.

Execution speed is crucial because it's essential to working with large datasets. When you use XGBoost, there are no restrictions on the size of your dataset, so you can work with datasets that are larger than what would be possible with other algorithms.

Model performance is also essential because it allows you to create models that can perform better than other models. XGBoost has been compared to different algorithms such as random forest (RF), gradient boosting machines (GBM), and gradient boosting decision

trees (GBDT). These comparisons show that XGBoost outperforms these other algorithms in execution speed and model performance.

2.2.6. Random forest

Random Forests (Breiman, 2001)

AUC, also known as Area Under Curve, is a metric commonly used to evaluate machine learning models, which is the area below the ROC curve. The AUC helps the classifier to distinguish between classes. The higher the Area Under Curve, the better the positive and negative discrimination performance of the model.

The classifier is considered perfect when AUC = 1. And AUC=0 if the algorithm only makes random guesses.

Random Forests (Breiman, 2001), i.e. a classifier combining a forest of decision trees grown on random input vectors and splitting nodes on a random subset of features, have been introduced for the classification of binary and multiclass outputs. The majority of papers employs Random Forests for the prediction of a binary target and adduces further proof of Random Forests' high accuracy (Buckinx and Van den Poel, 2005, Lunetta et al., 2004, Schwender et al., 2004).

Each random forest is comprised of multiple decision trees that work together as an ensemble to produce one prediction.

Random forest algorithms can produce acceptable predictions even if individual trees in the forest have incomplete data. Statistically, increasing the number of trees in the ensemble will correspondingly increase the precision of the outcome.

Given Random Forests' robustness and competence for analyzing large feature spaces and MNLs weakness in the latter, why not applying the Random Forests approach to MNL, i.e. building a forest of MNLs, to unite the best of both worlds? To this end, this paper proposes a new method, the Random MultiNomial Logit (RMNL), a Random Forest of MultiNomial Logits. We explore the potential of RMNL and compare it with the traditional MNL with human expert feature selection. Our new innovative RMNL method is demonstrated on a cross-sell case within the home-appliances industry.

2.2.7. SVM

A new learning machine for two-group classification problems is the support-vector network. The following concept is conceptually implemented by the machine: input vectors are non-linearly mapped to a very high dimensional feature space. A linear decision surface is built in this feature space. The decision surface's unique properties ensure the learning machine's high generalization ability. The support-vector network concept was previously

implemented for the limited case where the training data could be separated without errors. This result is extended to non-separable training data in this paper. It is demonstrated that support-vector networks with polynomial input transformations have a high generalization ability. We also compare the performance of the support-vector network to that of various classical learning algorithms that participated in an Optical Character Recognition benchmark study.

The support-vector network implements the following idea: it maps the input vectors into some high dimensional feature space Z through some non-linear mapping chosen a priori. In this space a linear decision surface is constructed with special properties that ensure high generalization ability of the network

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier detection.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where the number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoiding over-fitting in choosing Kernel functions and regularization terms is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

2.2.8. Catboost

CatBoost is a gradient boosting machine learning algorithm developed by Yandex. It focuses on processing features with categorical values, called target variables, and optimizing the model's prediction accuracy on the dataset.

Some of the features of the CatBoost algorithm include:

- Automatically process data with categorical and numerical values. Automatically handle missing values in data.
- Use a new evaluation method to determine weights for rare and common samples.

- Improve the performance of models on large data sets using structural division and processing large amounts of features. Customizable to suit different machine learning problems.

2.2.9. Artificial Neural Network

ANN (Artificial Neural Network) algorithm is a type of machine learning model based on the structure of the nervous system in the human brain. It is used to solve classification, prediction, and recognition problems in various fields such as computer vision, natural language processing, and medical applications. ANN algorithms are often used to train models, where input data is fed to the neural network and the algorithm adjusts the weights of connections between neurons to optimize the output. This process is called neural network training and requires a large amount of data to achieve high accuracy.

CHAPTER 3: PREPROCESSING DATA AND BUILDING MODELS

3.1. Data resources

The data used in the report is collected from companies listed on 2 stock exchanges (HOSE & HNX) from 2019 to 2021 (3 years). Initial data is 2166 financial statements of the companies mentioned above (in 11 industries according to ICB-level 1). Because the financial reporting characteristics of the financial and banking sector are different from other industry groups, we decide to delete businesses in the above two industries. The final data remains 1768 enterprises in 9 industries (classified by ICB - level 1). These businesses operate in 9 industries such as: ['Industrials', 'Petroleum', ' Utilities', 'Consumer Services', 'Consumers', 'Basic Materials', 'Technology', 'Pharmaceuticals and Healthcare', 'Telecommunications'].

From the financial statements of companies, we using the Financial Indices Calculation Methodology indicator set provided by FiinPro to calculate the index. Index groups:

- Profitability/performance ratios: (EBT-EBIT)/EBT; perc Gross profit margin; perc EBIT; Asset turnover ratio; perc Net profit margin ratio; Operating profit margin; perc EBITDA, EBIT, Profit before tax %, The equity turnover ratio
- Efficiency/Activity Ratios: perc ROE; perc ROA, perc ROIC, Average time to collect money from customers, Receivable Turnover ratio, The equity turnover ratio.
- Financial leverage ratios: Long-term short-term loan/Equity capital; Cash ratio; Quick ratio, The goodwill to assets ratio, Current liabilities / Total assets, Current payout ratio, Long-term loan / Equity capital, Long-term loan / Total assets, Long-term short-term loan/Total assets, Short-term liabilities / Equity capital, Total liabilities/Equity capital, Total liabilities/Total assets, Total assets / Equity capital, EBITDA / (Short-term debt + Interest),
- Scoring ratios: EBIT/Total Assets; Working Capital / Total Assets; (Profit after tax - CFO)/Revenue, Retained Earnings / Total Assets, Sales/Total Assets.

A total of 35 metrics are calculated and considered as features for the model.

3.2. Preprocessing data

3.2.1. Data Cleaning

Initial data has a total of 3.78% missing values, focusing on: threshold above 40%, such as Long-term Loans / Total Assets (42.02), Long-term Loans / Equity (42.02); above the threshold of 20% with the ratio of intangible assets / Total assets (25.28); Above the threshold of 10%, there is Short-Term Loans / Equity (17.14) and Short-Term Loans / Total

Assets (17.14) and less than 1% has Gross Profit Margin % (0.06). To ensure that the data is not +Infinitive or -Infinitive, we replace the missing values with 10^-3 (equivalent to 1 million VND). Then the data with missing value above 1% will be filled fully, the feature perc Gross profit margin will be filled by KNN method (with k = 10).

To detect outliers, the study uses two methods, checking the effect of trimming outliers and checking the effect of winsorizing outliers, specifically checking the effect of trimming results in no outliers and checking the effect of winsorizing outliers results in outliers. Trimming outliers is used to remove extreme values from a dataset that may have a disproportionate effect on statistical analysis. Winsorizing involves replacing extreme values in a dataset with less extreme values to minimize the effect of outliers on the results.

After performing a hypothesis test to determine if there is a significant difference between the results obtained before and after winsorizing. This will give you a more formal way to evaluate the impact of winsorizing on the statistical analysis. After carefully assessing the impact of winsorizing on the data. Overall, the effect of winsorizing outliers will depend on the nature of the data and the goals of the analysis. Applying to this study, winorizing can improve the accuracy of the analysis by reducing the influence of outliers

3.2.2. Data transforming

To identify a model with useful features for predicting the financial difficulties of Vietnamese companies, we conducted base classifier selection as explained below. Firstly, dataset consisted of 35 variables, and we used the Mutual Information method and the feature importance model to identify variables that appeared frequently. This allowed us to select 15 variables.

Using Mutual Information, developing a method to select the best 15 features from the preliminary dataset of 35 features. The selected features are as follows: (EBT-EBIT)/EBT, EBIT%, ROA %, ROE%, Cash Ratio, EBIT/Total Assets, Operating Profit Margin Sales, Gross Profit Margin %, Net Profit Margin %, Quick Ratio, Working Capital/ Total Assets, EBITDA Ratio %, (EAT - CFO)/ Revenue, Asset turnover ratio, Long-term short-term loans / Equity capital.

Importance feature: total max 9 models/methods (Pearson, Chi-2, RFE, Logistics, Random Forest, LightGBM, CatBoost, XGB, Decision Tree) we use features where it has total greater than 6. But by default Financial significance to match the topic and improve the performance of the model, there are variables such as: EBIT/Total Assets, Customer Receivables Turnover Ratio, ROA%, EBIT % Ratio, Long-term Loans / Total assets, EBIT, EBITDA, NPAT - CFO)/Revenue, Net profit margin %, Profit before tax %, Operating profit margin, Quick ratio, Payment ratio cash, Working capital/ Total Assets, (EBT - EBIT)/EBT, ROE%.

From the two methods, this study defined 15 variables that will be used for the problem along with 9 models suitable for classification. In general, the minority class (grade 1 - DN in financial difficulty) is the class of interest and we aim to achieve the best in this class. If the unbalanced data is not preprocessed, then this will degrade the performance of the classifier model. Most of the predictions will correspond to the majority class and treat the minority class features as noise in the data and ignore them. This will lead to high bias in the model. Therefore, we use the data resampling method in order to deal with the unbalanced data set.

Before we jump into training the model, there's the problem with imbalanced data. Previously we saw that this dataset is highly skewed toward no margin. We need to figure out a way to balance the dataset before proceed. The challenge of working with imbalanced information is that most machine learning models will ignore, and in turn have poor performance on, the minority class, although in this case it is the minority class, predicting 'no margin' that is most important and valuable for us. In this case, oversampling is preferred to the undersampling technique - special data gain oversampling technique (SMOTE - applied only to the training dataset in order to properly tune our algorithm) themselves on the data. The test data is unchanged so that it accurately represents the original data.). The reason is, in the sampling below, removing versions from the data that may carry some important information. The data is unbalanced (class 0:1506 and class 1:262). Performing the models on the unequal real data set, all 9 classification models are not able to generalize well about the minority class compared to the majority class. Obviously, in such skewed class distribution cases, the accuracy data would be biased and not preferred. As a result, most of the negative class samples were correctly classified. Therefore, there are fewer FPs than more FNs. After oversampling, the test data clearly increased in the test data. Accuracy is reduced, but by much Recall, satisfying the goal of any binary classification problem. In addition, the AUC-ROC and F1-score models remain more or less the same and also the results of this study.

3.3. Variables

The study uses 16 observed variables including 15 feature variables (independent variables) and 1 target variable (dependent variable). The independent variables used are:

(EBT-EBIT)/EBT, EBIT%, ROA %, ROE%, Cash Ratio, EBIT/Total Assets, Operating Profit Margin Sales, Gross Profit Margin %, Net Profit Margin %, Quick Ratio, Working Capital/ Total Assets, EBITDA Ratio %, (EAT - CFO)/ Revenue, Asset turnover ratio, Long-term loans + short-term loans / Equity capital.

3.4. Target variable

The target variable is evaluated against the criteria to determine which securities are not eligible for margin trading. If the enterprise belongs to one of the six cases above, the target variable is 1, and vice versa is 0 if the enterprise does not fall into the above cases.

Based on Total Descriptive Statistics of 1768 observations, the default variable - the business is not margin trading has mean value of 0.148% and standard deviation of 0.355. This shows that up to 14.8% of businesses are prohibited from lending on margin.

3.5. Drawing correlation matrix of variables

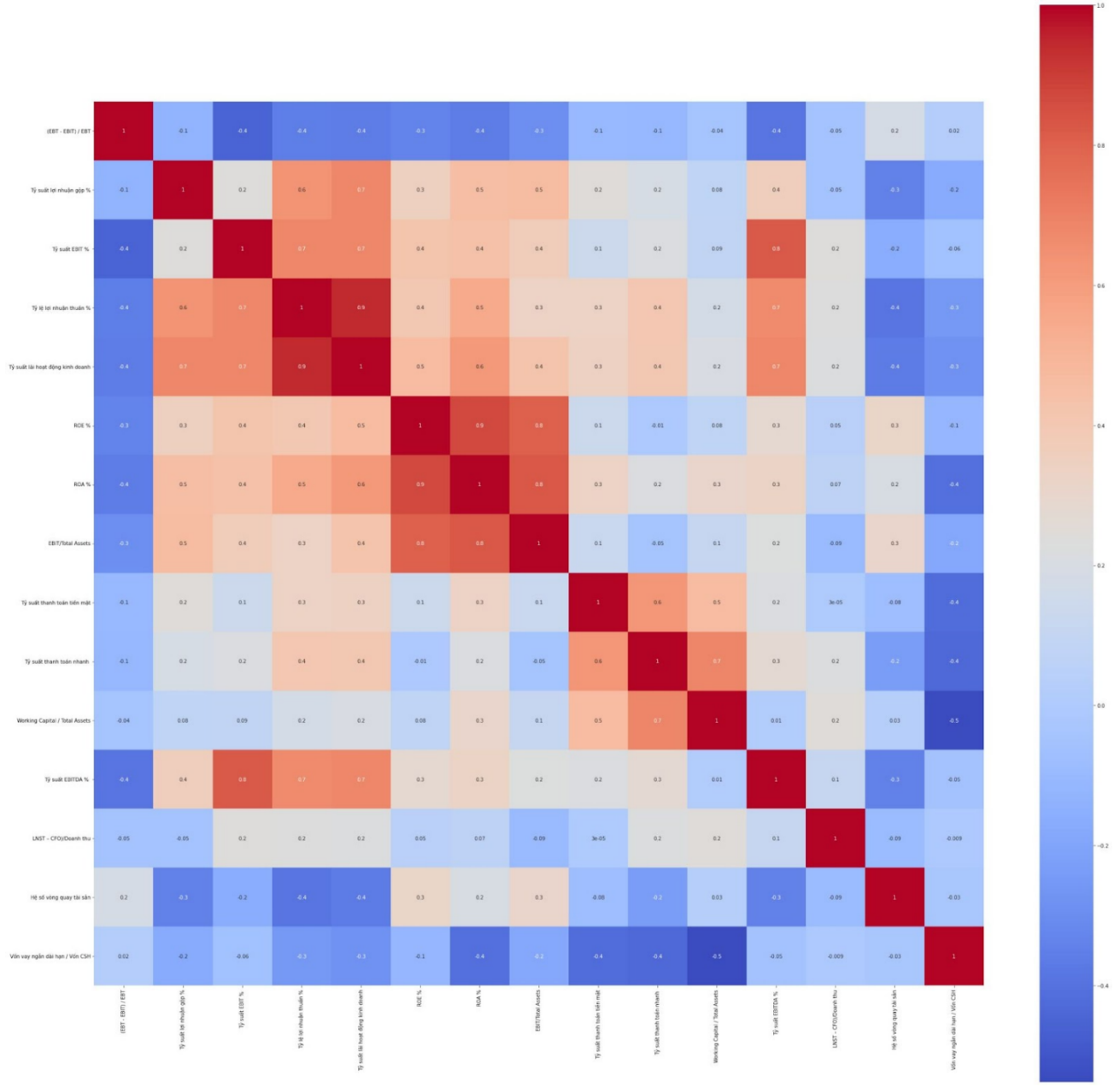


Figure 3.5: Heatmap of Correlation matrix of variables

After extracting 17 significant variables for the problem, we conducted a correlation analysis between the variables and found 2 variables: Retained Earnings/Total Assets; % of profit before tax has a high correlation (corr = 1) compared to 2 variables already in the model: %ROA; % percentage of net profit. ROE% and ROA%; Net profit ratio and operating profit margin are variables that are positively correlated with each other and are above 0.9.

3.6. Building models and visualizing results

3.6.1. Building models

As mentioned earlier, we will ignore the accuracy metric to evaluate the classifier's performance on this unbalanced dataset. Here, our team wants to know which companies are likely not to be allowed to deposit in the coming months. Thereby, focusing on metrics such as accuracy, withdrawal, F1 score to understand the performance of the classifiers in order to accurately determine the probability that the company is not allowed to borrow on margin.

We build models on python3 programming language combined with open source software packages.

To construct the machine learning model, divide the data into two parts, a training set and a testing set with a ratio of 70:30. The training set is used to train the machine learning model, and the testing set is used to test the model. We test machine learning algorithms to find the most optimal algorithm: Logistic Regression, Naive Bayes, KNN, Decision Tree, Random Forest, SVM, XGBoost, CatBoost, ANN.

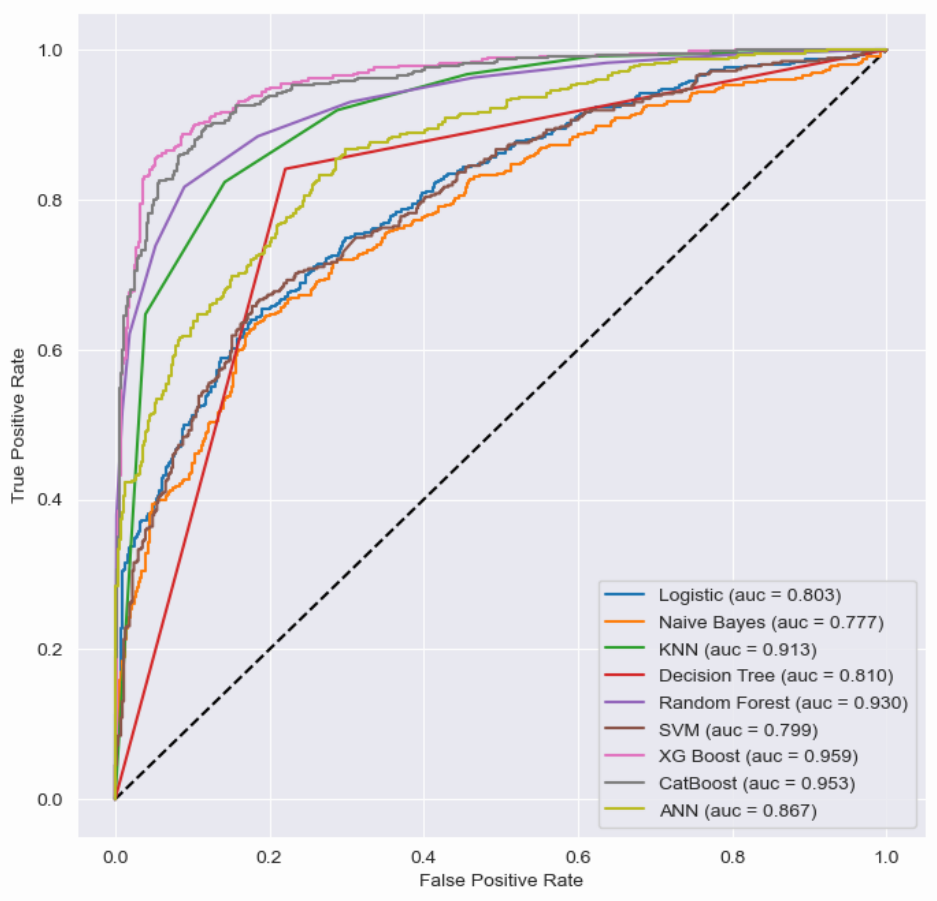


Figure 3.6.1: Chart of ROC curve in the total model

“The level of forecasting is aimed at businesses not having financial difficulties” shows that the forecasting ability is not correct according to the problem requirements of the model, including both financial difficulties and non-financial difficulties. Accordingly, the model that predicts the least will be the best model. Next, "the degree of accuracy of forecasting enterprises in financial difficulty" is more important than "the degree of forecasting so that enterprises do not face financial difficulties".

Table 3.6.1. Performance metric on test set by nine classification model

Algorithm	Test set						
	AUC-ROC	Precision		Recall		F1_score	
	Score	1	0	1	0	1	0
Logistic Regression	81%	69%	75%	72%	72%	70%	73%
Naive Bayes	79%	64%	75%	75%	64%	69%	69%
KNN	93%	77%	**93%**	**93%**	77%	85%	84%
Decision Tree	83%	79%	87%	86%	80%	82%	83%
Random Forest	94%	88%	86%	83%	90%	85%	88%
SVM	81%	69%	75%	72%	72%	71%	74%
XG Boost	96%	88%	**91%**	**90%**	89%	89%	90%
CatBoost	97%	89%	**90%**	**89%**	90%	89%	90%
ANN	98%	78%	79%	75%	72%	76%	81%

Source: Group research

In terms of financial significance, the problem is only interested in precision class 0 and Recall class 1. The score shows that the KNN model (k = 5, p = 2) is the best model for predicting margin lending with 93%.

Research model:

$$Target = \beta_0 + \beta_1 \frac{(\text{EBT} - \text{EBIT})}{\text{EBT}} + \beta_2\ \text{EBIT} + \beta_3\ \text{ROA} + \beta_4\ \text{ROE} + \beta_5\ \text{Cash ratio} + \beta_6 \frac{\text{EBIT}}{\text{Total Assets}} + \beta_7\ \text{Operating profit margin} + \beta_8\ \text{Gross profit margin} + \beta_9\ \text{Net profit ratio} + \beta_{10}\ \text{Quick ratio} + \beta_{11} \frac{\text{Working Capital}}{\text{Total Assets}} + \beta_{12}\text{EBITDA} + \beta_{13} \frac{\text{Profit after tax} - \text{CFO}}{\text{Revenue}} + \beta_{14}\ \text{Asset turnover ratio} + \beta_{15} \frac{\text{Long} - \text{term loans} + \text{short} - \text{term loans}}{\text{Equity capital}} + \varepsilon it$$

In there:

- Target variable: Using 1 figure is whether the enterprise is allowed to lend margin or not as the condition to create binary values (0 and 1). Information is published every quarter on the website of 2 stock exchanges: HOSE & HNX.

- Dependent variables:

+ Profitability/performance ratios: (EBT-EBIT)/EBT; Gross profit margin %; EBIT%; Asset turnover ratio; Net profit ratio %; Operating profit margin; EBITDA %

+ Efficiency/Activity Ratios: ROE %; ROA %

+ Financial leverage ratios: Long-term short-term loan/Equity capital; Cash ratio; Quick ratio

+ Scoring ratios: EBIT/Total Assets; Working Capital / Total Assets;(Profit after tax - CFO)/Revenue

The model above includes data of businesses in the same industry in the dataset from 2019-2021. In addition, we disaggregate the data by year and by stock exchange to compare the differences with variables like the one modeled above - the performance metric difference is just too big. Specifically, to decompose the model by year, we formed 11 models corresponding to 3 years from 2019 to 2021. According to the stock exchange, we divide into 2 exchanges, HNX and HOSE.

3.6.2. Results and evaluation of models

First, the analysis report to evaluate the reliability of the model The predictive KNN model gives good results with a high recall value of 93% and y = 1, showing that the accuracy of the obtained points is high, showing that the ratio of choosing the right businesses that are banned from margin trading is as high as 93% and with a precision rate of 93% in class 0, it can be seen that we will not be restricted when predicting the actual businesses that are

allowed to trade. escrow translation. Performance metrics with the remaining models such as XG Boost or Cat Boost respectively are close to or greater than 90%. The three models above accurately predict the ability of businesses to be free from financial difficulties compared to reality. Based on the above experimental results, the KNN model; XGBoost; Cat Boost reported the highest detection accuracy of financial distress (93%; 90%; 89%) of the 9 models built by this study to correctly identify businesses with financial distress.

Our test sample spans from 2019 to 2021. In general, the sample overall AUC ratings are in the range of 78% - 96%, with the highest being in the CatBoost and XGBoost models (see Appendix X).

Table.3.6.2. Accurate forecast results according to 3 algorithms KNN, XG Boost and Cat Boost

Algorithm	Obs	Test set								
		Test - set obs	Accuracy	AUC score	Precision		Recall		F1-score	
					1	0	1	0	1	0
KNN	3012	994	84%	93%	77%	93%	93%	77%	85%	84%
XG Boost			89%	96%	89%	91%	90%	89%	89%	90%
Cat Boost			90%	97%	89%	90%	89%	90%	89%	90%

Source: Group reseacher

Table.3.6.2. Accurate forecast results through 3 years according to 3 algorithms KNN, XG Boost and Cat Boost

Algorithm	Year	Obs	Test set									
			Test - set obs	Accuracy	AUC score	Precision		Recall		F1-score	
						1	0	1	0	1	0
KNN	2019	1032	341	77%	89%	72%	86%	90%	63%	80%	73%
XG Boost				87%	94%	88%	87%	88%	87%	88%	87%
Cat Boost				87%	95%	88%	86&	86%	87%	87%	86%
KNN	2020	978	323	82%	91%	79%	88%	91%	73%	84%	80%
XG Boost				87%	95%	90%	85%	85%	90%	88%	87%
CatBoost				89%	97%	92%	86%	86%	92%	89%	89%
KNN	2021	1002	331	86%	94%	85%	87%	89%	82%	87%	85%
XG Boost				90%	97%	91%	90%	91%	90%	91%	90%
CatBoost				92%	98%	93%	90%	91%	93%	92%	92%

Source: Group Research

Among the 35 indicators included in the forecasting model, the research results have a high consensus on the importance of the influence of the attributes on the predictability of enterprises experiencing financial difficulties based on the status of margin lending. of the enterprise. In Appendix x, there are 10 indicators that have the greatest influence on enterprises experiencing financial difficulties out of the 15 variables that bring the best results, respectively: ROE%, ROA%, Net profit ratio %, Operating Profit Margin, (EBT – EBIT) / EBT, EBIT / Total Assets, EBIT%, Gross Profit Margin %, Short Term Loans / Equity, (Profit after tax - CFO) / Revenue . The result of this study is a new point for

determining whether the enterprise is eligible for escrow deposit or not, thereby assessing whether the enterprise is facing financial difficulties or not.

Next, there is a distribution of performance metrics on 2 stock exchanges, HOSE and HNX using the 3 models and 15 financial indicators mentioned above .The results on HOSE are better than on HNX, specifically as follows:

Table.3.6.2: Performance metric on nine models decomposed by exchanges HOSE & HNX.

Algorithm	Exchange	Obs	Test set									
			Test - set obs	Accuracy	AUC score	Precision		Recall		F1-score	
						1	0	1	0	1	0
KNN	HOSE	1548	511	86%	93%	80%	95%	96%	75%	87%	84%
XG Boost				92%	97%	89%	95%	95%	89%	92%	92%
CatBoost				93%	98%	92%	94%	95%	91%	93%	93%

Algorithm	Exchange	Obs	Test set									
			Test - set obs	Accuracy	AUC score	Precision		Recall		F1-score	
						1	0	1	0	1	0
KNN	HNX	1464	484	80%	88%	77%	85%	89%	70%	82%	77%
XG Boost				84%	92%	84%	84%	86%	81%	85%	83%
CatBoost				86%	93%	87%	86%	88%	85%	87%	85%

Source: Group research

Our test sample spans from 2019 to 2021. In general, the sample overall AUC ratings are in the range of 78% - 96%, with the highest being in the CatBoost and XGBoost models (Appendix X).

CHAPTER 4: DISCUSSION AND CONCLUSION

4.1. Conclusion

We investigate whether machine learning (ML) methods can produce out-of-sample returns forecasts that outperform classical forecasts. We are motivated to use Machine Learning because the documentation and training show that traditional regression methods cannot produce out-of-sample forecasts that outperform random walk predictions, and Logistic Regression, K-Nearest Neighbor, Decision Tree, SVM (RBF Kernel), ANN and especially XG Boost and Catboost. Since they focus on predicting and optimizing prediction accuracy, ML methods have certain advantages over regression methods in generating out-of-sample predictions.

With a data sample of 1768 observations in the period of 2019 - 2021, the research results show that the model is built based on two motivational factors including financial stability: 15 variables, 3 models have ability to predict the situation of enterprises banned from margin trading leading to financial difficulties. In particular, the KNN model has the best predictive ability, this model has the ability to predict accurately over 95% when using all 15 variables mentioned above. This is followed by the XG Boost and Cat Boost models at 90% and 89%, respectively. The research results have helped to answer the research question, the important factors affecting the business being banned from margin trading have been identified with high predictive accuracy.

In the first phase, this study deploys Mutual information and feature importance to screen for important variables. A total of 15 variables that can affect the detection of margin trading prohibited enterprises are selected (including financial variables), in order to improve the detection accuracy of the models. Among the nine financial distress prediction models established by this study, three of them have precision class 0 over 90% and three of them have class 1 recall over 89%.

Furthermore, we assess the results obtained from the study by year of transactions. Stretching from 2019 to 2021, the sample overall AUC ratings are in the 78%-96% range, with the highest in two models such as CatBoost and XGBoost received from the year-to-year model review, which is changing. about 71% and 81% and there is no significant change across each exchange, specifically the KNN model on HNX and HOSE is 88% and 93% respectively, these figures are for CatBoost and XGBoost models are 92% and 97%, 93% and 98%, respectively. In addition, the significance of macro features in the model year is zero year over year. For different people and different purposes, we can choose different prototyping ways. When researchers require highly accurate predictions for investment and risk taking.

4.2. Discussion and future study

This new study in Vietnam opens up a new research method on the financial status of listed companies compared to previous studies when using machine learning to find out the

factors affecting the profitability of listed companies. allowed margin and found 15 variables that are significant financial indicators for the model.

Examples of such advantages include insensitivity to econometric issues such as multicollinearity, better handling of nonlinearity in the data than traditional regression based methods, and discovery of the functional form that best fits the data. We implement the ML method using a large sample of Vietnamese firms with valid required data over the period 2019-2021 and generate out-of sample predictions of directional changes (increases or decreases) in more than 15 variables.

In addition, the study assesses the financial difficulties of enterprises through the status of margin lending, specifically based on the margin lending criteria of article 1 of Decision 1205/QD-UBCK in 2017.

Based on the discussion above, this study provides models with variables that provide a rigorous and efficient predictability of firms in financial distress. This is also a guide for other researchers or practitioners. It can also be considered for future studies using other data mining techniques to predict business financial distress.

The weakness of the study is that it has not analyzed the impact of each industry on margin lending. It also has not analyzed whether or not the non-margin lending status of enterprises in one industry group significantly affects the margin lending status of enterprises in other industries. Besides, whether there is a differential influence from the input variables on the model performance.

Financial stress is not a surefire cause of bankruptcy, but a business that goes bankrupt is certainly going to experience financial stress. Therefore, forecasting the possibility of financial stress of enterprises has become a necessity, an issue that increasingly attracts the attention of investors, creditors and managers.

Enterprises that do not meet the margin requirements show that the business has had an inefficient period of operation. However, not only reflecting business results in an accounting period, behind the loss number are also separate stories about business characteristics, management, administration, accounting methods, the macro situation, etc. But for an enterprise banned from margin lending, it may come from the fact that the enterprise is dishonest in disclosing financial statements, or has acts of tax evasion. Based on the research results, we propose some recommendations:

From an investor's perspective, loss is the clearest warning signal about financial difficulties, as well as risks that investors face. If investors invest in businesses during a period of loss, the stock market price drops, investors will face a lot of risks. Therefore, forecasting financial difficulties will help investors make the right decisions in which companies to invest in. In addition, investors should invest in good businesses with positive business results and high growth, because the stocks of these companies often increase in price sustainably. In forecasting and making the right investment decisions, investors need to take more measures. The research article contributes to the basis for investors to easily realize their goal of optimizing their investment portfolio.

Predictions of a company's financial distress can also have a significant impact on bank lending decisions, such as loan interest rates, loan duration and periodic repayment amounts, Loan-to-Price Ratio, etc. collateral value. This helps the bank to minimize risk.

Forecasting financial distress based on margin lending serves to alert investors to potential risks in their investments. If a margin lending company is experiencing financial difficulty, there can be a greater risk that its loans may depreciate or become uncollectible. This can lead to a decrease in the company's stock price and reduce the value of investors' investments in that company.

As such, when making forecasts about the financial distress of margin lenders, investors can use this information to adjust their portfolios and reduce their exposure to risk. invest. They may also use this information to make decisions about whether to continue to hold or sell their existing investments in such margin lending companies.

REFERENCES

Aidas Malakauskas, Ausrine Lakstutiene (2021). *Financial Distress Prediction For Small and Medium Enterprises Using Machine Learning Techniques*. Inzinerine Ekonomika-Engineering Economics

Baltas, G., & Doyle, P. (2001). *Random utility models in marketing research: a survey*. Journal of Business Research, 51(2), 115-125.

Chen, W. S., & Du, Y. K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. Expert systems with applications, 36(2), 4075-4086.

Cortes, C., & Vapnik, V. (1995). *Support-vector networks.* Machine learning, 20, 273-297.

Clause 10, Article 2 of Circular 120/2020/TT-BTC.

Do T.V.T & Phan T.D & Dinh H.L (2022). *Dong tien va kiet que tai chinh cua cac doanh nghiep niem yet tren thi truong chung khoan Viet Nam: tiep can theo phuong phap Bayes*. Tap chi Kinh te & Phat trien

FiinPro Financial Indicators

Norhafizah Ishak & Nor Farizal Mohammed (2019). *Financial Distress Prediction Through Cash Flow Ratios Analysis*

Outecheva, N. (2007). *Corporate financial distress: An empirical analysis of distress risk. Doctoral dissertation,* University of St. Gallen.

Pham T.H.V (2018). *Do luong kha nang kiet que tai chinh cua cong ty niem yet tren thi truong chung khoan.*

Simplilearn. (2023*). What is xgboost? an introduction to XGBoost algorithm in Machine Learning: Simplilearn. Simplilearn.com.* Retrieved from https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article

The image section – university of copenhagen. Retrieved from http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdf

Support Vector Machines. scikit. Retrieved from https://scikit-learn.org/stable/modules/svm.html

Nhung Thuat Toan Hoc may thong dung nhat ban nen biet. ITguru.vn Blog. (2021). Retrieved from https://itguru.vn/blog/cac-thuat-toan-hoc-may-ban-nen-biet-vao-nam-2021/

Prinzie, A., & Van den Poel, D. (2008). *Random forests for multiclass classification: Random multinomial logit.* Expert systems with Applications, 34(3), 1721-1732.

sklearn.ensemble.RandomForestClassifier. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

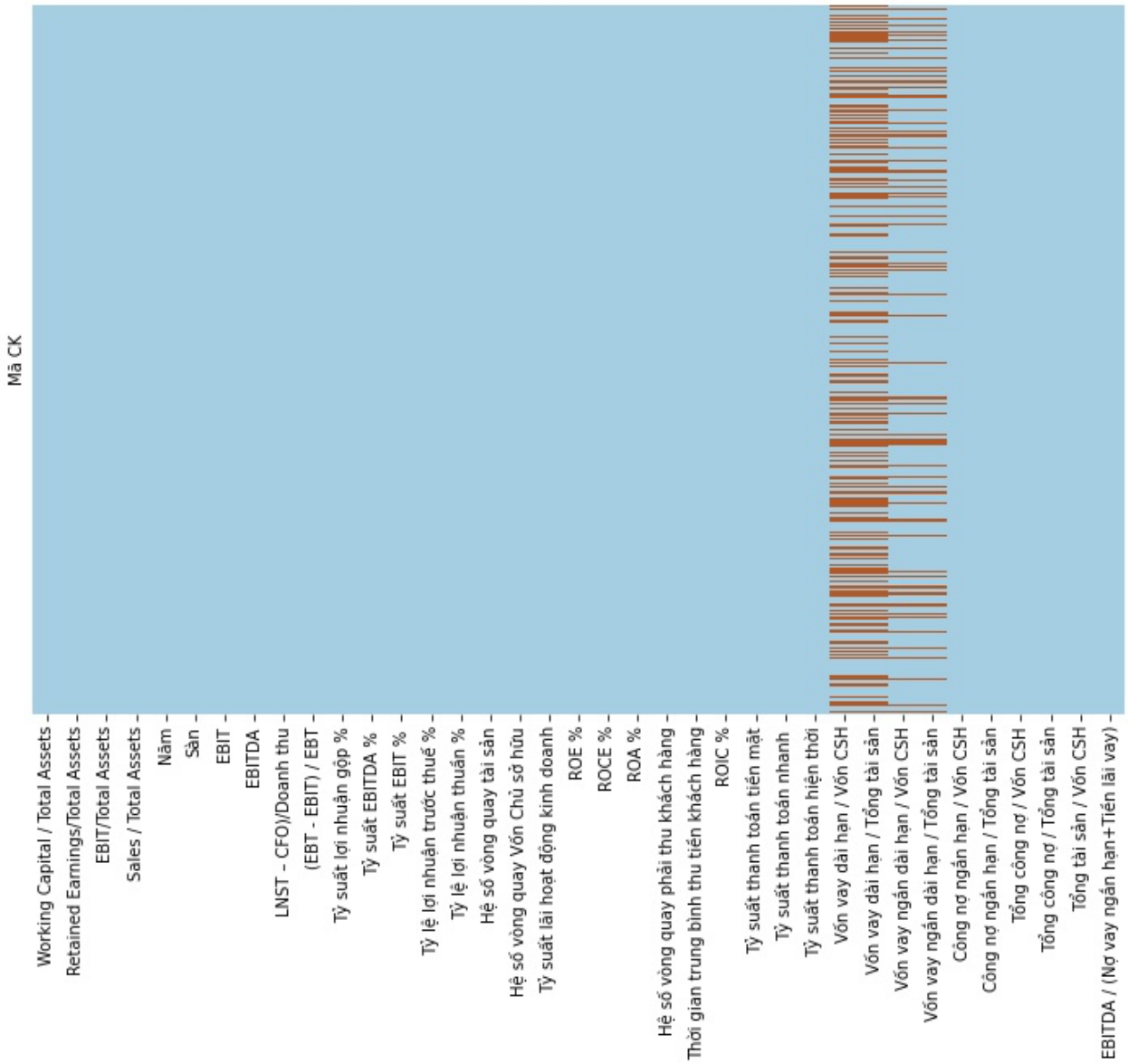
The optimality of naive bayes. Retrived from http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf

Wei-Sen Chen, Yin-Kuan Du (2009). *Using Neural Networks and Data Mining Techniques For The Financial Distress Prediction Model*

Thuat toan artificial neural network - learning ann. w3seo. (2022). Retrieved from https://websitehcm.com/thuat-toan-artificial-neural-network-tim-hieu-cach-learning-ann/

APPENDICES

Appendix. Chart of missing values' distribution in raw dataset



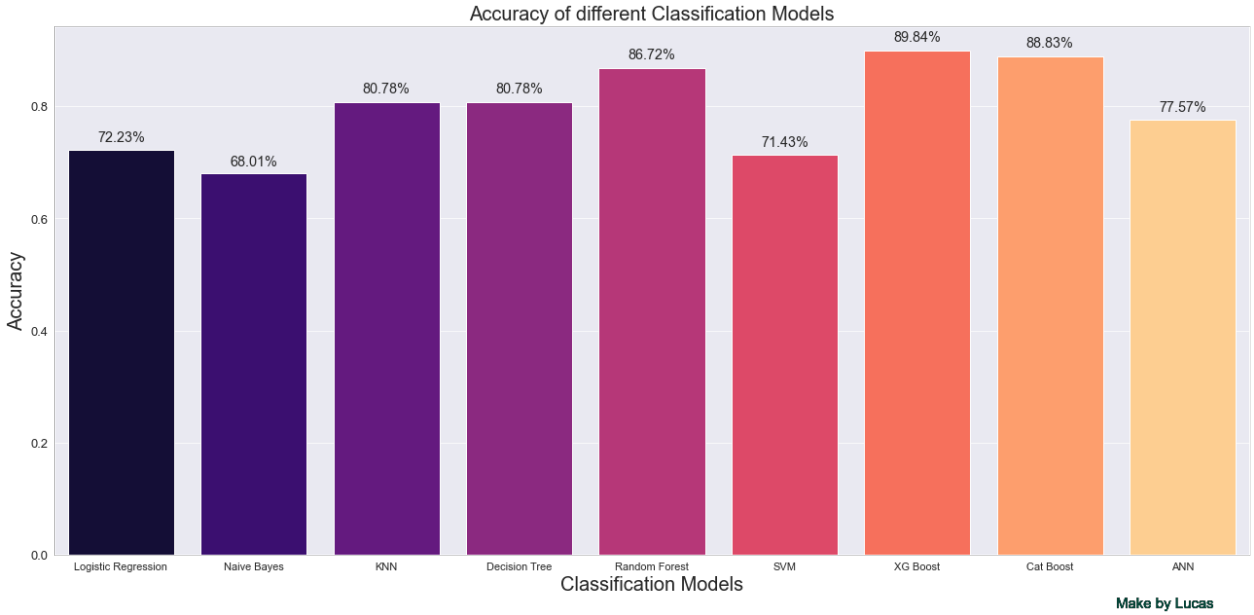
Appendix. Descriptive statistics measures of the base classifiers on dataset.

	count	mean	std	min	25%	50%	75%	max
default	890.0	0.130337	0.336863	0.000000e+00	0.000000	0.000000	0.000000	1.000000
EBIT	890.0	-32.060065	1886.235813	-1.588373e+04	-87.286339	-9.510003	57.001425	33612.447942
EBITDA	890.0	362.331674	2332.335614	-1.178488e+04	-14.497971	39.386194	246.104304	45765.480533
LNST – CFO)/Doanh thu	890.0	-0.290095	5.963761	-1.737445e+02	-0.074097	-0.006123	0.057464	3.189041
(EBT - EBIT) / EBT	890.0	2.632206	9.791101	-2.115873e+01	0.399945	1.047861	2.159014	170.333749
Tỷ suất lợi nhuận gộp %	890.0	0.193032	0.428695	-6.872605e+00	0.083458	0.155835	0.261289	9.384849
Tỷ suất EBITDA %	890.0	-0.173660	5.539628	-1.182521e+02	-0.016724	0.042221	0.162859	6.585149
Tỷ suất EBIT %	890.0	-0.276709	5.456786	-1.174968e+02	-0.060182	-0.012452	0.046769	6.575142
Tỷ lệ lợi nhuận trước thuế %	890.0	-0.185288	6.001667	-1.337451e+02	0.026721	0.067076	0.141191	6.860757
Tỷ lệ lợi nhuận thuần %	890.0	-0.169952	4.994609	-1.141053e+02	0.018573	0.049490	0.099749	5.942086
Hệ số vòng quay tài sản	890.0	1.029973	0.922869	-1.678898e-03	0.481513	0.841349	1.237268	8.166394
Hệ số vòng quay Vốn Chủ sở hữu	890.0	2.687040	6.210178	-2.074902e+00	0.808800	1.528834	2.937258	144.353958
Tỷ suất lãi hoạt động kinh doanh	890.0	-0.195590	6.033787	-1.351348e+02	0.023655	0.065096	0.138124	7.059860
ROE %	890.0	0.078849	0.864771	-2.533141e+01	0.049779	0.099992	0.163484	1.586803
ROCE %	890.0	-0.068085	0.467784	-1.138153e+01	-0.088253	-0.014879	0.049399	0.577383
ROA %	890.0	0.059414	0.077165	-4.672669e-01	0.020034	0.047295	0.087112	0.467896
Hệ số vòng quay phải thu khách hàng	890.0	7.779748	11.906274	-7.706655e-03	2.322700	4.561546	8.360415	127.630395
Thời gian trung bình thu tiền khách hàng	890.0	144.812265	3102.476209	-5.345245e+04	43.521728	79.781784	155.077227	55732.564670
ROIC %	890.0	-0.019006	0.363418	-2.024701e+00	-0.066500	-0.010818	0.042635	8.935139
Tỷ suất thanh toán tiền mặt	890.0	0.452970	1.181119	4.131720e-04	0.073980	0.195972	0.468779	25.381438
Tỷ suất thanh toán nhanh	890.0	1.985070	3.884709	7.291532e-02	0.691303	1.100340	1.840875	62.702583
Tỷ suất thanh toán hiện thời	890.0	2.514313	4.065046	9.712876e-02	1.130530	1.510169	2.368557	62.702583
Vốn vay dài hạn / Vốn CSH	890.0	0.259925	1.368772	-5.743858e-01	0.000004	0.045253	0.247060	38.963037
Vốn vay dài hạn / Tổng tài sản	890.0	0.080647	0.125323	1.588022e-08	0.000002	0.021204	0.114942	0.676780
Vốn vay ngắn dài hạn / Vốn CSH	890.0	0.731002	2.362893	-8.672885e-01	0.125951	0.438123	0.988666	66.385082
Vốn vay ngắn dài hạn / Tổng tài sản	890.0	0.238133	0.179170	1.234657e-07	0.082702	0.220453	0.380142	0.748336
Công nợ ngắn hạn / Vốn CSH	890.0	1.129334	3.425288	-4.777170e+00	0.288544	0.621391	1.266061	78.583043
Công nợ ngắn hạn / Tổng tài sản	890.0	0.357758	0.204022	2.673529e-03	0.184475	0.335332	0.502325	1.246841
Tổng công nợ / Vốn CSH	890.0	1.464683	4.598482	-4.831419e+00	0.440221	0.867854	1.700742	119.290872
Tổng công nợ / Tổng tài sản	890.0	0.465116	0.211763	2.673529e-03	0.308071	0.465719	0.630807	1.294471
Tổng tài sản / Vốn CSH	890.0	2.464707	4.598498	-3.831419e+00	1.440221	1.867854	2.700742	120.290872
EBITDA / (Nợ vay ngắn hạn+Tiền lãi vay)	890.0	0.382002	1.799618	-7.132443e+00	-0.041843	0.079845	0.422087	27.708397
Working Capital / Total Assets	890.0	0.212760	0.216123	-5.659188e-01	0.060949	0.185982	0.343807	0.902852
Retained Earnings/Total Assets	890.0	0.049098	0.065774	-4.180711e-01	0.016639	0.040863	0.078396	0.378283
EBIT/Total Assets	890.0	0.070084	0.083521	-3.699270e-01	0.026804	0.059297	0.102013	0.491633
Sales / Total Assets	890.0	1.029973	0.922869	-1.678898e-03	0.481513	0.841349	1.237268	8.166394

Appendix. Feature selection by importance on 9 methods

	Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	LightGBM	CatBoost	XGB	DTR	Total
1	ROE %	True	True	True	True	True	True	True	True	True	9
2	(EBT - EBIT) / EBT	True	True	True	True	True	True	True	True	True	9
3	Working Capital / Total Assets	True	True	True	True	False	True	True	False	True	7
4	Tỷ suất thanh toán tiền mặt	True	True	True	True	False	True	True	False	True	7
5	Tỷ suất thanh toán nhanh	True	True	True	True	False	True	True	False	True	7
6	Tỷ suất lãi hoạt động kinh doanh	True	True	True	True	True	False	True	True	False	7
7	Tỷ lệ lợi nhuận trước thuế %	True	True	True	False	True	True	True	True	False	7
8	Tỷ lệ lợi nhuận thuần %	True	True	True	True	True	True	True	False	False	7
9	LNST – CFO)/Doanh thu	True	True	True	True	False	True	True	False	True	7
10	EBITDA	True	True	True	True	False	True	True	True	False	7
11	EBIT	True	True	True	True	False	True	True	False	True	7
12	Vốn vay dài hạn / Tổng tài sản	True	True	True	True	False	False	True	False	True	6
13	Tỷ suất EBIT %	True	True	True	True	False	True	False	False	True	6
14	ROA %	True	True	True	False	True	False	True	True	False	6
15	Hệ số vòng quay phải thu khách hàng	True	True	True	True	False	True	True	False	False	6
16	EBIT/Total Assets	True	True	True	True	False	True	True	False	False	6
17	Vốn vay ngắn dài hạn / Vốn CSH	True	True	True	True	False	True	False	False	False	5
18	Tỷ suất lợi nhuận gộp %	True	True	True	False	False	True	True	False	False	5
19	Thời gian trung bình thu tiền khách hàng	True	True	True	True	False	False	True	False	False	5
20	Vốn vay dài hạn / Vốn CSH	True	True	True	False	False	False	True	False	False	4
21	Tỷ suất thanh toán hiện thời	True	True	True	True	False	False	False	False	False	4
22	Tổng công nợ / Vốn CSH	True	True	True	False	False	False	False	True	False	4
23	Sales / Total Assets	True	True	True	True	False	False	False	False	False	4
24	Retained Earnings/Total Assets	True	True	True	False	True	False	False	False	False	4
25	ROIC %	True	True	True	True	False	False	False	False	False	4
26	ROCE %	True	True	True	True	False	False	False	False	False	4
27	Hệ số vòng quay tài sản	True	True	True	True	False	False	False	False	False	4
28	Hệ số vòng quay Vốn Chủ sở hữu	True	True	True	True	False	False	False	False	False	4
29	EBITDA / (Nợ vay ngắn hạn+Tiền lãi vay)	True	True	True	True	False	False	False	False	False	4
30	Công nợ ngắn hạn / Vốn CSH	True	True	True	False	False	True	False	False	False	4
31	Công nợ ngắn hạn / Tổng tài sản	True	True	True	True	False	False	False	False	False	4
32	Vốn vay ngắn dài hạn / Tổng tài sản	True	True	True	False	False	False	False	False	False	3
33	Tỷ suất EBITDA %	True	True	True	False	False	False	False	False	False	3
34	Tổng tài sản / Vốn CSH	True	True	True	False	False	False	False	False	False	3
35	Tổng công nợ / Tổng tài sản	True	True	True	False	False	False	False	False	False	3

Appendix. Feature selection by importance on 9 methods



Appendix. Table of 15 independent variables

No.	Variable	Formula
01	$\frac{EBT - EBIT}{EBT}$	$\frac{EBT - EBIT}{EBT}$ In which, net profit before tax is accumulated for 4 consecutive years (If there are financial statements year) or 4 quarters (If no annual financial statements are available), EBIT is based on quarterly data or year to count
02	EBIT	$\frac{EBIT}{Net\ Sales}$ In which EBIT is calculated by quarter or year and Net Revenue by year or slip in the last 4 quarters
03	ROA	$\frac{\text{Parent company's NPAT}}{\text{Average Total Assets}}$ In which the parent company's NPAT by year or slip in the last 4 quarters, Average Total assets for 2 consecutive years (If annual financial statements are available) or Average Total assets falling in the last 4 quarters (If no annual financial statements are available).
04	ROE	$\frac{\text{Parent company's EAT}}{\text{Average Equity}}$ In which NPAT of parent company by year or by last 4 quarters, Average of Owner's Equity for 2 consecutive years (If annual financial statements are available) or Average of Owner's Equity in last 4 quarters (If no annual financial statements are available).
05	Cash ratio	$\frac{\text{Cash and cash equivalents} + \text{Trading securities}}{\text{Current debt}}$ In which Cash and cash equivalents, trading securities, and short-term liabilities are based on quarterly or annual data.
06	$\frac{EBIT}{Total\ Assets}$	(Gross profit + Selling expenses + General and administrative expenses + Profit/loss from joint ventures (before 2015) + Profit/loss from joint ventures)/ Total assets

		In which the targets are taken according to the quarterly or annual data to be calculated
07	Long-term + short-term loans / Equity capital.	Long-term + Short-term debt / Equity = (Short-term borrowings and leases + Long-term loans and leases) / Equity In which short-term loans and finance leases, Loans and finance leases long-term, Equity based on quarterly or annual data to calculate
08	Asset turnover ratio	$\text{Asset turnover ratio} = \frac{\text{Net sales}}{\text{Average Total assets}}$ In which: Net revenue by year or slip for 4 consecutive quarters; Average Total assets 2 years in a row (If annual financial statements are available) or Average Total assets slip 4 quarters (If no annual financial statements are available).
09	$\frac{\text{EAT} - \text{CFO}}{\text{Revenue}}$	$\frac{EAT - CFO}{Revenue} = \frac{\text{EAT of parent company} - \text{NCF from Business activities}}{Net\ revenue}$ In which, EAT of the parent company, Net cash flow from production and business activities, Net Revenue accumulated for 4 consecutive years (If annual financial statements are available) or slipped 4 quarters (If there is no annual financial statement).
10	EBITDA	$EBITDA\ Ratio = \frac{EBITDA}{Net\ Sales}$ In which EBITDA is calculated by quarter or year and Net Revenue by year or slip in the last 4 quarters.
11	$\frac{Working\ Capital}{Total\ Assets}$	$\frac{Working\ Capital}{Total\ Assets} = \frac{Current\ Assets - Currents\ Liabilities}{Total\ Assets}$ In which the targets are taken according to the quarterly or annual data to be calculated.
12	Quick ratio	$Quick\ Ratio = \frac{\text{Current Assets} - \text{Net Inventory}}{Current\ Liabilities}$

		In which Current Assets, Net Inventory, Current Liabilities are taken on a quarterly or annual basis.
13	Net profit ratio	$$Net\ profit\ ratio = \frac{\text{Parent company's NPAT}}{\text{Net sales}}$$ In which the parent company's NPAT and net revenue by year or by slip in the last 4 quarters best.
14	Gross profit margin	$$\text{Gross profit margin} = \frac{Gross\ profit}{Net\ revenue}$$ In which Gross profit and Net revenue are taken by year or the sum of the last 4 quarter1.
15	Operating profit rate	Operating profit rate $$= \frac{\text{Profit (loss) from business activities by year either 4 quarters}}{\text{Yearly Net Sales or 4 Quarters}}$$

Appendix. Table of 15 independent variables's meaning

No.	Variable	Meaning
01	$\frac{EBT - EBIT}{EBT}$	This index tells investors the structure of profit before tax and interest of the business or the ability to pay interest on loans by the income before tax.
02	EBIT	The EBIT ratio shows profitability before interest and tax expenses.
03	ROA	ROA shows how much profit a company generates based on the assets it has.
04	ROE	ROE indicates the ability to bring profit to shareholders per dollar spent. ROE can be used to estimate the growth rate of a business.,
05	Cash ratio	Cash ratio shows the ability of the business to pay short-term debt in cash and cash equivalents
06	$\frac{EBIT}{Total\ Assets}$	Variable for the effective variable of using assets of the business to generate profits. This indicator can used to determine which businesses are reporting the most efficient use of their assets relative to their earnings.
07	Long-term + short-term loans / Equity capital.	The short-term debt-to-equity ratio is a solvency measure that shows the ratio of debt a company uses to finance its assets, compared to the amount of equity used for the same. one goal.
08	Asset turnover ratio	Asset turnover ratio helps investors know the efficiency of converting assets into revenue of the business.
09	$\frac{\text{EAT} - \text{CFO}}{\text{Revenue}}$	
10	EBITDA	The EBITDA ratio indicates profitability that eliminates the effects of capital structure and depreciation policies. Therefore, this ratio can be used to compare different industries or businesses in different industries.

11	$\frac{Working\ Capital}{Total\ Assets}$	This variable shows the ability of the business to pay short-term debts. This index also shows the liquidity of the business because it also shows the percentage of liquid assets compared to the total assets of the company.
12	Quick ratio	Quick ratio measures the ability of a business to quickly liquidate assets to meet short-term liabilities.
13	Net profit ratio	Net profit ratio helps investors assess the ability to convert profit from sales after deducting operating and other expenses.
14	Gross profit margin	Gross profit margin indicates the profit after deducting the costs to obtain the goods and services, excluding selling costs and related general and administrative expenses.
15	Operating profit rate	Operating profit margin shows how much profit a business's operations contribute