VIETNAM NATIONAL UNIVERSITY

HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

GRADUATION THESIS

FORECASTING PROFITABILITY OF LISTED COMPANIES IN VIETNAM WITH MACHINE LEARNING

Lecturer :	**PHAM CHI KHOA**
Student :	**VO THUY UYEN NHI**
Student ID :	**K194141736**
Class :	**K19414C**

Ho Chi Minh City, May 2023

VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

GRADUATION THESIS

FORECASTING PROFITABILITY OF LISTED COMPANIES IN VIETNAM WITH MACHINE LEARNING

Lecturer : **PHAM CHI KHOA**
Student : **VO THUY UYEN NHI**
Student ID : **K194141736**
Class : **K19414C**

Ho Chi Minh City, May 2023

TABLE OF CONTENTS

LIST OF TABLES

LIST OF PICTURES

ABSTRACT

These days, data science and big data are growing rapidly, and they can offer a range of options to advance the study of finance and commercial strategy. Over the past two decades, the field of computer science known as machine learning has been increasingly popular. Using Machine Learning in the profit modelling of listed organizations is an effective method, such as Regression, Random Forest, Random Walk, Classification, Neural Network Models or deep neural networks,...when compared to traditional forecasting approaches. Following this trend, a model that evaluated and predicted the financial performance is created, which is expressed through profitability, of listed companies using a Random Forest approach based on machine learning. And 512 listed companies in Vietnam and their actual historical data from 2010 to 2020 financial statements are selected as a dataset. It is now relevant to the Vietnamese economy to develop an approach to assess the financial performance of the Vietnamese securities market, which will offer investors useful information. Lenders, company governance, and investors all make choices. The methods of research are as follows: data collecting method, qualitative method, variable description, conceptual model development using quantitative method, experimental method. Testing the study's efficacy and applicability is appropriate for reducing observation and research time using the gathered data set. Furthermore, recognizing the limits of past studies' techniques, the report proposes to perform profit situation forecasting research approach by industry and try to investigate further the section where the model fits.

Keywords: Machine learning, Forecasting, Random forest, Profit performance, Supervised Learning.

1. INTRODUCTION

1.1 Background of the research project

Unprecedented events were happening in every area of society in 2020, including the economy, education, work, and, not to be forgotten, life. Businesses are affected negatively by the complex Covid outbreak, which is spreading quickly and spawning a lot of new, difficult strains. Financial performance analysis plays a significant role in assisting investors in determining whether or not to invest in a certain company. (Srinivasan et al, 2020). Therefore, maintaining high financial performance has become key to the survival and development of enterprises. As a result, many businesses strive to improve their financial accounts, making it difficult for investors, particularly inexperienced ones, to determine which companies have potential. During the period of Covid, there was a rapid development in digital transformation, which is contributing to the diversification of financial activities, and the use of Machine Learning methods in applied finance is becoming increasingly popular not only all over the world but also in Vietnam. After referring to previous studies with traditional approaches, it is found to to run into many limitations when running, displaying, and predicting models. Therefore, Machine Learning with multiple variables is decided to use to meet the needs of digital transformation and overcome limitations from traditional methods and increase predictive power for users. This study also makes an effort to provide an output model that may assist managers of companies in implementing specific procedures that are advantageous to both enterprises and shareholders during special situation as the Covid 19 timeframe. Additionally, it aids investors in decision-making in situations similar to those.

1.2 Purpose of the study

The purpose of this study, "Forecasting Profitability of Vietnamese Listed Companies with Machine Learning Techniques," is motivated by the importance of considering various factors when deciding whether to invest in a company, such as financial performance, industry trends, and the economic and political climate. Among these factors, a company's financial performance and overall health are crucial. Indicators like revenue growth, profitability, debt levels, and cash flow provide valuable insights into a company's financial strength. However, some companies manipulate their financial performance to appear better, making it challenging for investors, especially inexperienced ones, to make informed decisions. In light of this, the main focus of this study is on profitability as the target variable.

1.3 Objectives of the study

Identify and evaluate financial indicators or features influencing corporate earnings in order to better precisely estimate business profits. Profit is one of the most crucial criteria for outside investors and internal management to understand when examining financial performance of a firm. As a result, using Machine Learning applications to guide the information group to the aim of searching for elements impacting business advantages. Machine Learning was designed to handle a vast data collection with a high variety, releasing other factors outside of the underlying element and maximizing prediction accuracy. Furthermore, the topic is based on current data from Vietnamese firms to be relevant and appropriate to the Vietnamese economy and give important information to internal and foreign investors.

According to multiple earlier stories, during the COVID-19-caused crash in March 2020, the stocks of natural gas, food, healthcare, and software all experienced significant gains.(Mazur and associates, 2020). Additionally, this study looks at how well machine learning methods forecast the financial success of Vietnamese listed firms and makes suggestions on the strategy for each industry and also re-evaluate the above statement.

Factors influencing company profits that are subjective (financial statements, business sector, business size, growth rate, capital asset investment ratio, liquidity ratio, leverage ratio ...). Profit is influenced by both objective and macro variables (GDP, CPI, interest rate). This paper would select financial information of 516 enterprises listed on the Vietnam stock exchange and 11-year research period from 2010 to 2020 as a dataset.

Some research questions:

- How effective are machine learning techniques for predicting the financial performance of Vietnamese-listed companies?
- Which financial indicators or features are most important in predicting the financial performance of Vietnamese listed companies using machine learning?
- How do different types of industry (e.g., yearly) affect the accuracy of machine learning models in predicting the financial performance of Vietnamese listed companies?

- Which situation does this machine learning model suitable?
- How do different machine learning algorithms, such as regression, decision trees, and neural networks, compare in their ability to predict the financial performance of Vietnamese-listed companies?

By answering these questions, this study provides valuable insights into the application of machine learning in finance and offers practical implications for stakeholders in the Vietnamese securities market.

1.4 Research Methodology

Methods of data collection, qualitative methods: Qualitative methods were used to identify and analyze the various factors impacting enterprise profitability, and to explore new factors that may be relevant. These factors were then integrated to provide a comprehensive understanding of the enterprise's performance.

Building a preliminary model using quantitative methods: Profit forecasting for enterprises based on Machine Learning models using assumptions about previously known profitability criteria. From there, the direction of future profit movement may be determined.

Empirical research will be conducted to evaluate the efficacy and usefulness of the study's findings in reducing observation and research time with the acquired data collection. Additionally, the study plans to utilize Random Forest analysis, as it is suitable for the high level of data randomness observed in the Vietnamese dataset. The study also aims to compare the performance of Random Forest to other algorithms such as Neural Networks, which are commonly used in financial classification topics. The comparison will provide insights into the relative strengths and weaknesses of these algorithms in the context of financial forecasting in Vietnam.

The expected result of this scientific paper is to provide a comprehensive analysis of the factors impacting the performance of a particular financial or banking system, using advanced machine learning techniques. The study aims to give firms more accurate profit estimates. The results of the study will help financial and banking institutions managers and investors to make informed decisions based on data-driven insights and improve their overall performance hence reducing company risks. Additionally, the study may contribute to the academic literature on machine learning in finance and banking, by highlighting the potential of these techniques to provide valuable insights and improve decision-making in the industry.

2. THEORETICAL FRAMEWORK

2.1 Literature Reviews

One of the key components of performance evaluation is profitability, which displays the profit percentage in relation to sales, equity, and asset investments. However, concentrating just on the quantitative component of profit is insufficient; it is also vital to assess if the quality of that profit is truly sustainable and reliable.

Many businesses in Vietnam adjust profits and build a good financial performance to lower corporate income tax and attract outside investment, potentially sacrificing genuine profitability. This may have an impact on the enterprise's business performance evaluation. Furthermore, there are objective affecting elements outside of the corporate environment, known as macro variables. There is state management in the context of a market economy. By analyzing subjective and objective factors that affect earnings, businesses can forecast earnings, which shows the company's performance, and make informed decisions about resource management and investment.

Vietnamese studies

There are many studies not only in Vietnam but also around the world on forecasting the financial performance of corporations. Specifically, domestic studies on factors affecting business operations and profits such as

Factors affecting the profit of enterprises listed on HOSE (Nguyen Hoang Anh, Nguyen Thi Tu, 2017) used variables such as Enterprise size (Size), Financial leverage (LEV), Investment ratio capital asset investment (PPE). After the regression, the author has found that the higher the firm size, the better the earnings quality and the financial leverage is inversely related to the earnings quality.

Measuring business performance through market value index and book value index by machine learning method (Nguyen Anh Phong, 2021). The author used the variables return on assets, return on equity to measure the business' performance. In addition, the author also used some other control variables such as Age of the enterprise (Age) which I also studied and applied in this study.

The literature review highlights the impact of economic value added (EVA) and return on assets (ROA) on the creation of shareholder value in Vietnamese firms, as studied by Tran Thi Thuy Linh (2017). The findings suggest that using ROA as a measure of business performance can be distorted by managers under performance pressure or to serve their own interests. Managers may employ accounting methods that reduce costs or decrease asset value in the short term, artificially inflating ROA. However, such practices may not be beneficial in the long term and can lead to misguided investment decisions. Additionally, the study points out that ROA, calculated as net income divided by total assets, fails to consider the firm's cost of capital, neglecting the opportunity cost for investors. Furthermore, it overlooks the risks faced by the enterprise and relies solely on book value without considering market value. As a result, ROA may not accurately reflect the true value of the business in its current state. Therefore, this paper will combine the traditional measure of performance such as ROA and other measures with machine learning method to gain higher accuracy in predicting.

Foreign studies

Besides, there are also international studies on the elements affecting the performance of financial situation enterprises through profitability. Additionally, there are many papers also using supervised learning, which is a branch of machine learning, to forecast financial performance:

In addition to the commonly used traditional profitability measures, several other factors have been found to impact profitability, such as financial leverage, company size, liquidity, inventory efficiency, and age. For instance, a study on profitability factors in Malaysian listed companies by Alarussi and Alhaderi (2018) aimed to identify factors that affect profitability, including company size (measured by total revenue), liquidity, and financial leverage. The findings of the research demonstrated a strong positive relationship between profitability and total revenue, while liquidity did not exhibit any significant relationship with profitability. This highlights the importance of considering various factors beyond the traditional profitability measures when analyzing profitability in different industries and countries.

Factors Influencing the Companies' Profitability (Camelia Burja, 2011). Research shows that among the factors that have a good influence on profitability are inventory efficiency, debt level, and capital efficiency. In addition, the study also demonstrates a close relationship between the company's performance and the way available resources are managed.

Determinants of corporate profitability: An empirical study of Indian drugs and pharmaceutical industry (Chander, S., & Aggarwal, P, 2008). The main objective of this study was to examine the relationship between Indian drugs and pharmaceutical companies in terms of characteristics and profitability. The author used factors such as Size, Age, Liquidity and R&D intensity in this study. The results show that variables such as Age, Liquidity Ratio and R&D intensity have positive results on the profitability of enterprises.

Prediction Method of Enterprise Return on Net Assets Based on Improved Random Forest Algorithm (Yong-Hua Cai et al. , 2020) proposes a precise prediction technique for enterprise return on net assets using the improved random forest algorithm. The algorithm is optimized using the simulated annealing algorithm and compared to other algorithms, showing good predictive impact. This helps the capital market assess business performance.

Using Machine Learning to Forecast Future Earnings (Xinyue Cui, Zhaoyu Xu, Yue Zhou, 2020) the model in this study was able to serve as a convenient auxiliary tool for analysts to conduct predictions better than the traditional statistical models that are widely used in the industry including Logistic Regression. This model has made significant progress in both accuracy and speed of prediction.

Predicting profitability using machine learning (V Anand, R Brunner, K Ikegwu, T Sougiannis, 2019). The study explores whether a method from machine learning, a classification tree, can produce predictions of out-of-sample returns that outperform random walk predictions. The authors implement a machine learning approach using a large sample of US companies with required data valid for the phase 1963-2017 and generate out-of-sample predictions of changes in direction (increase or decrease) in five profitability measures: return on equity (ROE), return on assets (ROA), return on operating assets (RNOA), cash flow from operations (CFO) and free cash flow (FCF). The results show that their machine learning method achieves classification accuracy ranging from 57-64% compared to 50% for random walk, and the difference in classification accuracy rate between machine learning and Random walk is very meaningful. In summary, the study provides some evidence that machine learning methods have the potential to be useful in predicting profitability.

2.2. Machine learning in finance

Arthur Samuel defined machine learning as "the branch of research that makes computers capable of learning without being explicitly programmed." And, in the words of Muskaan et al., "Machine learning methods have been proposed as alternative approach to statistical methods by many researchers in their academic literature."

Machine learning is the process of building models and making predictions from data in an iterative manner using a variety of techniques. The data set is divided into two pieces by machine learning: the training set and the test set. Algorithms employ training data to build machine learning models. The test data set will be used to assess the correctness of the produced model. Machine learning algorithms such as ANN - Artificial Neural Networks, SVM - Support Vector Machines, KNN - K-nearest neighbours, logistic regression, AdaBoost, LDA - Linear Discriminant Analysis decision tree, and random forest are commonly used in financial analysis.

While econometrics is a powerful tool for understanding the relationships between variables, it has limitations in terms of predictive accuracy. Machine learning, on the other hand, is designed specifically for prediction and has shown tremendous success in a variety of domains, including finance. By using machine learning algorithms to analyze financial data, we can uncover patterns and relationships that may not be apparent through traditional econometric models. Additionally, machine learning models are capable of handling large volumes of data, including unstructured data, which is becoming increasingly important in the era of big data. Therefore, the application of machine learning in financial analysis offers a powerful tool for accurately predicting financial outcomes and identifying potential risks or opportunities.

2.3. Supervised learning

According to a reliable source, IBM, Supervised learning is a type of machine learning that involves the use of labeled data to train a model to predict outcomes for new, unseen data and the main goal of determining the relationship between input and output variables. In supervised learning, the model is presented with a dataset where both the input features and the corresponding output values are known. The model then learns the relationship between the input features and output values and can make predictions on new data based on this learned relationship. Some common examples of supervised learning algorithms include regression, A Random Forest Algorithm, decision trees, and neural networks. Supervised learning is widely used in various domains, including finance, where it can be applied to tasks such as credit risk assessment and fraud detection.

2.4. Random forest

Random forest is a machine learning algorithm that belongs to the ensemble learning family. It is a decision tree-based method that constructs multiple decision trees and combines their results to generate a final output. In a random forest model, each decision tree is trained on a random subset of the training data, and the final output is obtained by aggregating the predictions of all trees. Random forest is a popular algorithm for classification and regression tasks and has several advantages, including high accuracy, aids in overcoming the problem of overfitting, robustness to outliers and noise, and the ability to handle large datasets. It is also relatively easy to use and does not require extensive parameter tuning.

Based on the definition of Breiman (2001): random forest is "a classifier consisting of a collection of tree-structured classifiers {h(x, Θk), k = 1, . . .} where the {k} are independently identically distributed random vectors and each tree casts a unit vote for the most popular class at input x".

Random forest can determine feature importance, which helps to identify important and unimportant features. This may allow us to focus on the essential ones and potentially remove low-impact features.

2.5 Neural Network

Information about other neural network models can be found in any introductory book in this field, e.g., Kung (1993); Rumelhart and McClelland (1986); Hertz et al. (1991). A neural network is a computational model that is inspired by the structure and function of the human brain. It consists of interconnected nodes, or neurons, that receive input signals, perform mathematical operations on them, and produce output signals.

In a neural network, the input layer receives input data, which is then passed through one or more hidden layers before reaching the output layer. Each layer consists of a set of neurons that perform a linear or nonlinear function on the input data. The weights and biases of these neurons are adjusted during the training process to minimize the difference between the predicted output and the actual output.

Neural networks have been widely used in financial forecasting because they are able to capture complex patterns and nonlinear relationships in financial data. They can be trained to predict stock prices, exchange rates, and other financial indicators with high accuracy.

In a thesis, a neural network may be used as a tool for analyzing financial data and making predictions about future market trends. The structure and parameters of the neural network, as well as the training and testing methods used, should be carefully selected and justified in order to produce reliable results. Additionally, the limitations and potential sources of error in the neural network approach should be discussed.

2.6. ROC-Curve and AUC

The ROC (Receiver Operating Characteristics) curve is a graphical representation of a classification model's performance. It shows the sensitivity of the model to the ratio between true positives and false negatives. A perfect model would have a curve that runs from the bottom left corner to the top left corner to the top right corner, meaning it would have a true positive rate of 1 and a false positive rate of 0. The closer the curve comes to the top left corner, the better the model performs. The ROC curve is useful for comparing different models and determining the optimal threshold for classification. It can also be used to calculate the area under the curve (AUC), which is a metric for evaluating the overall performance of a model. The Area Under the Curve (AUC) is a crucial metric that evaluates a classifier's ability to distinguish between classes. A higher AUC indicates that the model has better positive and negative discrimination performance. A perfect classifier is achieved when the AUC is equal to 1. However, if the algorithm randomly guesses, the AUC will be equal to 0. Therefore, AUC is an essential metric to determine the performance of a classifier.

2.7. Enterprise profit

Profit is the difference between a company's income and costs in a certain time. Profit has an impact on the company's liquidity and financial situation. The profit gained is used to reinvest, grow, and develop the business in the future. This is also one of the factors used to evaluate enterprise business activity and assist investors in making investment decisions. Profits may be classified into two types: economic profit and accounting profit. Accounting profit is only concerned with monetary expenses, not sunk and opportunity costs, as economic profit is. This study simply takes into consideration the accounting earnings of businesses. One of the benchmarks that can be used to measure a company's performance is profitability (Pandian and Narendran, 2015, Rouf, 2010; Wallace & Naser, 1995; Meek, et al., 1995). Similarly, those used in medium-level financial institutions, unlike rural banks (BPR). Rouf (2010) and Wallace & Naser (1995) stated that Profitability Ratios are ratios used to assess a company's ability to find profits or profits in a given financial period, while Return On Assets (ROA) is a very important profitability ratio to determine how effective financial institutions e.g., Business Process Re-engineering is able to generate profits from total assets owned, due to the greater ROA shows the better level of profitability.

2.8. Real state of affairs

In today's highly competitive business environment, companies need to continuously optimize their profits to ensure long-term growth and success. Accurately forecasting profits and revenues is crucial in this regard. Machine learning algorithms can now be used to predict trends and forecast financial performance with greater accuracy, thanks to the increasing availability of historical data. The purpose of this study is to apply machine learning techniques to forecast the financial performance of listed companies in Vietnam. The study will use a dataset containing historical financial data for these companies to train a machine learning model, which will then be evaluated using out-of-sample tests to identify the most accurate and effective machine learning methods for forecasting financial performance in the Vietnamese market. By leveraging the power of machine learning, this study aims to provide valuable insights that can help investors and businesses make more informed decisions and achieve better financial outcomes.

While many studies have shown that using historical data sets to predict corporate financial results is feasible with a reasonable error rate of 2-8%, the biggest challenge for this approach is not making predictions but collecting data and building models with a large number of variables. With a huge amount of variables and data, it can take a lot of time and effort to build a simple model and get results, which may become outdated by the time the study is completed. Therefore, it is essential to strike a balance between the accuracy of the model and the time and resources needed to build it. The profit that a company earns in a particular period of time depends on several factors, such as the amount of time and money invested in various projects and policies. To predict a company's profit over a period of time, we need to train a machine learning model using a dataset containing historical data about the profits generated by the company. Overall, this study is a critical step towards helping businesses in Vietnam optimize their profits and achieve long-term growth and success. As well as it is hoped to be provided valuable insights that can help investors and businesses make more informed decisions and achieve better financial outcomes.

3. FRAMEWORK OF PREDICTION MODEL

3.1. Data preparation

This study employs a secondary data source consisting of two types of data, namely micro data and macro data. Micro data is collected from Thomson Reuters, covering the financial figures of 684 Vietnamese enterprises listed on two stock exchanges - HOSE (Hochiminh Stock Exchange) and HNX (HaNoi Stock Exchange) - from 2009 to 2020. These enterprises operate in ten different industries, including Utilities, Basic Materials, Industrials, Consumer Cyclicals, Consumer Non-Cyclicals, Energy, Healthcare, Real Estate, Technology, and Academic & Educational Services.

The macro data used in this study includes three indicators - GDP, CPI, and Real interest rate of Vietnam central bank. - which were collected from the World Bank from 2009 to 2020.

3.2. Preprocessing data

3.2.1. Data Cleaning

In the data cleaning process, businesses listed on the stock exchange for less than 5 years were dropped to ensure that the data is not too much missing and after removed 168 companies from original data of 684 enterprises, the dataset got 516 remained company. Then, the count_missing function was used to check the missing value of the variables, and variables with a missing value rate above 7% were identified, including Quick ratio, Cash conversion cycle, Inventory turnover ratio, FA turnover ratio, Operating margin, BEP, Expense of revenue ratio, and DSO. And the ratio of missing value variables from 12.46% or less.

To address the problem of missing values of these 9 variables, the interpolate function was used for filling missing values. The missing values in the first and last years will be filled by the backfill and first-fill method, while the missing values in the middle will be filled by the regression method. However, the variable avg inventory days still had 48 missing values, and 4 companies - VNF.HN, TVC.HN, VNT.HN, VNL.HM - could not be filled due to empty full and were removed completely.

After filling data and removing incomplete data, the missing value ratio was checked again, and variables with missing values were dropped by removing all rows with at least 1 missing value. After the cleaning process, the remaining data sets included businesses in 9 industries, except Academic & Educational Services.

3.2.2. Data transforming

At first, converting some unreasonable negative values into positive variables such as: Cost of revenue, Accounts receivable, Average Inventory days, Average Payable days, Average Receivable days. Next, I create 20 independent variables from the available data source. After the calculation, the Growth variable is missing data because there is no 2008 to calculate, so the 2009 drop should be performed on all variables. Age variable is calculated from the year of company establishment, companies established from 2011-2020 will be converted to NaN and dropped.

The Target variable is created based on the return on asset (ROA) variable, when compared to the average variable. The value of Target is 1 if return on asset is greater than average return on asset ratio; and zero if return on asset is less average return on asset ratio. The profit forecast associated with Operating cash flow shows the business's ability to pay its debts and reduce the risk of fraudulent financial statements for investors. 18 feature variables (independent variables) and 1 target variable (dependent variable) after being calculated are merged into 1 data frame with 20 columns and 5676 rows.

There were 13 variables with outliers found: Growth, PPE, Liquidity, Quick ratio, Inventory turnover ratio, FA turnover ratio, DSO, Capital intensity, Expense of revenue ratio, Operating margin, Net profit margin, ROE, CCC. Dealing with outlier by creating a function to replace outliers with first or thirrd quantile in order to help the model distinguish between class 1 and class 0 which would create better results:

- Upper bound = First quantile value + 3*Interquartile range
- Lower bound = Third quantile value - 3*Interquartile range

If outlier value < lower bound, it would be assigning a new value to the variable with lower bound. Or if it higher than upper bound it would be assigning a new value to the variable with upper bound,

3.3. Variables

Numerous studies have investigated the factors that impact the profitability of firms, including size, working capital management, age of the firm, leverage, and liquidity. Building on this existing literature, this study explores the relationship between profitability and 24 observed variables, including 23 feature variables (independent variables) and one target variable (dependent variable). These variables encompass various factors such as growth, firm size (as measured by total sales), age, liquidity, leverage (debt equity ratio and leverage ratio), PPE (as measured by Tangible Fixed Asset and sales), quick ratio, inventory turnover ratio, fixed assets turnover ratio, company efficiency (total assets turnover ratio), DSO, capital intensity, expense revenue ratio, operating margin, net profit margin, BEP, ROA, ROE, EPS, CCC, GDP, CPI, and interest rates.

To conduct the study, data from 516 companies listed on Vietnam's two exchanges, HOSE (Hochiminh Stock Exchange) and HNX (HaNoi Stock Exchange), were extracted for the period from 2010 to 2020. The target variable in this analysis is profitability.

By examining these variables and their relationship with profitability, the study aims to provide valuable insights into the key drivers of financial performance for firms in Vietnam. This information can help investors and businesses make more informed decisions and ultimately achieve better financial outcomes.

Target variable

The target variable is based on the return on asset (ROA) variable.This ratio serves as a crucial indicator of profitability for businesses, specifically measuring the net profits in relation to total assets. This metric provides valuable insights into how effectively a company manages its assets to generate revenue. It serves as a key evaluation tool for assessing the competence and operational performance of banks, as it quantifies the profits derived from the bank's invested assets (Jahan, 2012; Golin, 2001). ROA can be used as dependence variable with the below condition:

- ROA > 0: Target = 1.
- ROA < 0 or ROA = 0: Target = 0.

While the formula is

$$ROA = \frac{Net\ income}{Total\ asset}.$$

The purpose of this model is the high accuracy and high recall of class 0 and class 1 that forecasts which businesses having well performance or being unprofitable.

Extensive literature supports the significance of Return on Assets (ROA) as a key financial ratio for assessing a company's profitability. Scholars such as Golbert and Rai (1996), Akbas (2012), Eng (2013), Kasmir (2014), and Vasanth (2015) have emphasized the role of ROA in measuring a company's ability to generate profits relative to its income, assets, and capital stock. ROA serves as an indicator of management's efficiency in generating income from asset utilization and provides insights into the company's financial performance and operational efficiency. By considering these findings, the proposed model aims to leverage ROA as a vital metric for accurately evaluating the financial performance of listed companies.

Independent variables

This research chose to find as much as possible any related variables that found in previous papers and researches and entered it into the model in order to found new determinants that effect financial performance prediction. And the 18 dependent variables are:

3.3.1. Sale growth rate (Growth)

$$Growth = \frac{Sale_t - Sale_{t-1}}{Sale_{t-1}}$$

Growth is the growth rate of a business's revenue. Businesses with this high ratio are often in a strong growth phase. The relation of Growth variable is found that the quality of earnings is also impacted by the joint's growth pace. Firms with high this ratio, according to Nissimi and Penman (2001), Dechow and et al. (2011), Gopalan and Jayaraman (2012), and Nguyen Hoang Anh (2017), have lower-quality earnings.

3.3.2. Size

$$Size = ln(Market\ value)$$

The size of a business is a significant factor that has been studied in various research works. This finding has been supported by several studies, such as those conducted by Ball and Foster (1982) and Liu et al. (2017). However, it is important to note that this is not always the case, as other researchers have reported a negative relationship between firm size and profitability. For instance, Gopalan and Jayaraman (2012) and Watts and Zimmerman (1990) found that larger companies may actually be less profitable than smaller ones. These conflicting results highlight the need for further research to better understand the relationship between firm size and profitability.

3.3.3. Age

The company's age variable is calculated in 2020 minus the company's founding year. The longer a business has been in operation, the greater its profitability. This is because long-standing businesses have created a reputation, have high product recognition and experience in production. Chander and Aggarwal (2008) show a positive influence of age on firm profitability.

3.3.4. Liquidity ratio (Liq)

$$Liq = \frac{Total\ current\ asset}{Current\ liabilities}$$

Liquidity ratio indicates a company's ability to pay short-term debts in the short term. Cerf (1961) found that the higher the liquidity ratio, the higher the profit of the enterprise. According to Cerf's research, companies with higher liquidity ratios tend to exhibit higher profitability. This finding suggests that maintaining adequate levels of liquidity can contribute to a company's ability to generate profits.

3.3.5. Capital investment ratio (PPE)

$$PPE = \frac{Tangible\ Fixed\ Asset}{Sales}$$

Capital investment ratio shows the ratio of the company's investment in tangible assets. This high ratio makes it easy for investors to track and encourages managers to adjust profits accordingly..

3.3.6. Quick ratio

$$Quick\ ratio = \frac{Total\ current\ asset\ -\ Inventory}{Current\ liabilities}$$

Quick ratio shows the ability of a business to pay its short-term debts in the short term, regardless of its inventory.

3.3.7. Inventory turnover ratio

$$Inventory\ turnover\ =\ \frac{365}{Average\ inventory\ days}$$

Inventory turnover ratio measures the number of times a company sells new inventory and replacements. For a business with low inventory turnover ratio, inventory costs will increase which can potentially reduce profits.

3.3.8. Fixed asset turnover ratio

$$Fixed\ asset\ turnover\ ratio\ =\ \frac{Sales}{Fixed\ assets}$$

Fixed asset turnover ratio measures the efficiency of a company's fixed assets such as machinery and equipment. A higher fixed asset turnover ratio indicates that the company is effectively utilizing its fixed assets to generate sales, which can lead to increased profitability. Conversely, a lower ratio may suggest underutilization or inefficiency in utilizing fixed assets.

3.3.9. Total asset turnover ratio

$$Total\ asset\ turnover = \frac{Sales}{Total\ asset}$$

The relationship between total asset turnover and profitability has been consistently found to be positive in several studies. Alarussi and Alhaderi (2018), Alghusin (2015), and Agiomirgianakis et al. (2006) have all reported evidence supporting this relationship. This suggests that higher levels of total asset turnover, which measures the efficiency of utilizing assets to generate revenue, are associated with increased profitability.

3.3.10. Days sale outstanding (DSO)

$$DSO\ =\ \frac{Account\ receivable}{Sale/365}$$

DSO is too high, potentially bad debt, reducing the company's profit. Days Sales Outstanding (DSO) is a financial metric that reflects the average time it takes for a company to collect payment from its customers after making a sale. A high DSO indicates that it takes longer for the company to collect its receivables, which can have negative implications for the company's profitability. When DSO is too high, it may indicate potential issues such as bad debts or delays in customer payments, which can impact the company's cash flow, overall profitability and reducing the company's profit.

3.3.11. Capital intensity

$$Capital\ intensity\ =\ \frac{Total\ asset}{Sales}$$

Capital intensity is positively correlated with profitability, according to Gopalan and Jayaraman (2012). Capital intensity refers to the need for large amounts of investment for production of some industries which require significant investments in order to carry out their production activities effectively. Industries that have high capital intensity typically involve substantial expenditures on equipment, infrastructure, and other fixed assets. This higher level of investment can contribute to increased productivity, efficiency, and ultimately profitability for companies operating in these industries. The positive correlation suggests that as capital intensity increases, the potential for higher profitability also increases.

3.3.12. Expense revenue ratio

$$Expense\ revenue\ =\ \frac{Costs\ of\ revenue}{Sales}$$

Expense revenue ratio measures the performance between revenue and expenses. It provides insights into the efficiency and profitability of the company's operations by comparing the amount spent on expenses to the revenue generated. A lower expense revenue ratio indicates that a company is able to generate higher revenue relative to its expenses, indicating better financial performance.

3.3.13. Operating margin

$$Operating\ margin\ =\ \frac{EBIT}{Sales}$$

Operating margin is a profitability ratio that measures operating income per dollar of sales. This ratio is commonly used to assess a company's ability to control costs, manage its operations effectively, and generate sustainable profits.

3.3.14. Basic earning power (BEP)

$$BEP\ =\ \frac{EBIT}{Total\ asset}$$

Basic earning power indicates the ability of a business' assets to generate operating income against the effects of taxes and debt. This ratio is used to compare a business's ability to make money when taxes and debt are different.

3.3.15. Cash conversion cycle (CCC)

Cash conversion cycle = Average inventory days + Average receivable days - Average payable days

Cash conversion cycle is the time period from when the company pays its creditors to when it receives money from its customers. And it is important for assessing a company's efficiency in managing its working capital and cash flow. A shorter cash conversion cycle indicates that a company is able to convert its investments into cash quickly, which is generally considered as it allows for better liquidity and financial stability. Conversely, a longer cash conversion cycle may indicate inefficiencies in some processes, which can impact a company's profitability

3.3.16. Gross Domestic Index (GDP)

The Gross Domestic Index is a fundamental macroeconomic indicator that quantifies the overall economic activity within a country. Extensive research has shown that there is a positive relationship between GDP growth and business profitability (Pasiouras & Kosmidou, 2007; Demirgüc-Kunt & Huizinga, 1999; Bikker & Hu, 2002; Athanasoglou et al., 2008). Therefore, in this study, the annual growth rate of GDP is employed as a potential determinant of profitability.

3.3.17. Consumer Price Index (CPI)

Consumer Price Index indicating an increase in the general price level, can lead to production instability and affect business profitability. It creates challenges in managing expenses as the cost of production rises due to higher prices of inputs. Additionally, currency devaluation associated with a high CPI can increase costs for businesses reliant on imports and impact their profitability.

3.3.18. Real interest rate

Real interest rate is the lending rate as measured by the GDP deflator. Real interest rates are influenced by terms and conditions across countries, so comparability is limited.

3.4. Drawing correlation matrix of variables

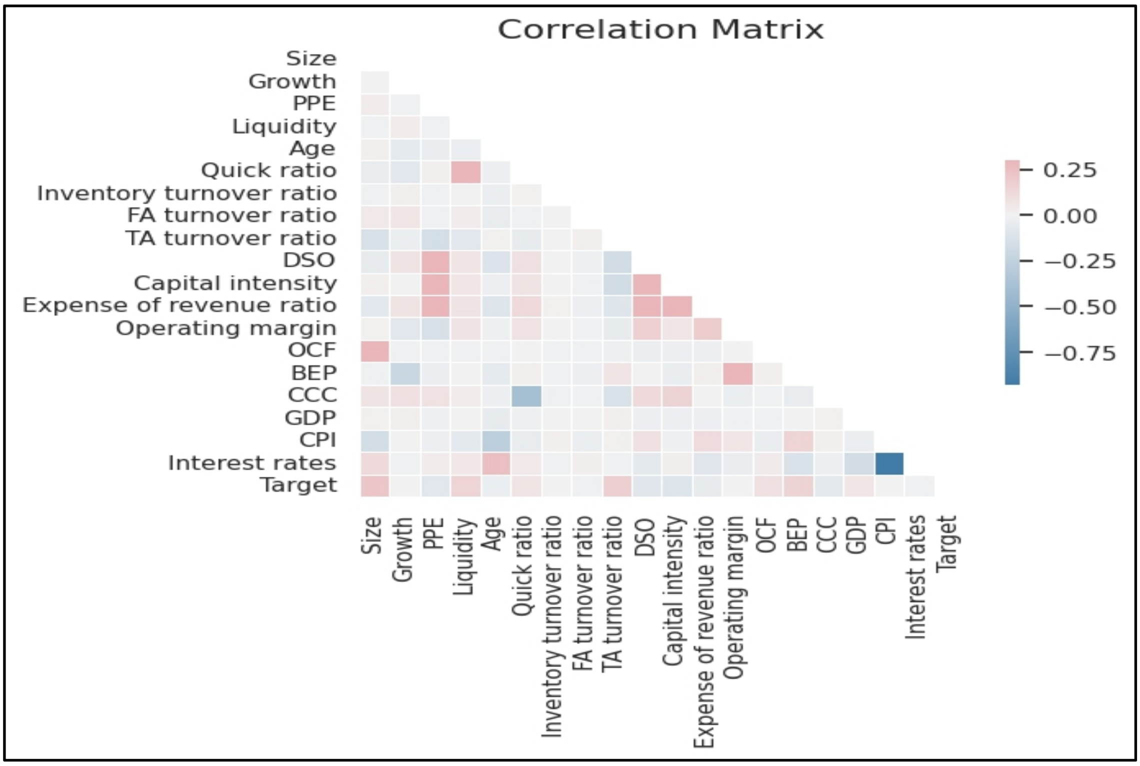


Figure 1. Correlation matrix of variables

Overall, the correlation coefficients between feature variables range from -0.25 and 0.25, it indicates low moderate correlations between the features themselves and that can avoid the problem of multicollinearity, which makes it difficult for the model to determine their individual effects on the target variable.

Table 1: Variables description

	count	mean	std	min	25%	50%	75%	max
Size	5.349.000.000	26.206.681	1.706.232	21.716.518	25.040.748	26.061.765	27.231.882	33.590.443
Growth	5.422.000.000	394.154	3.452.518	-997.369	-89.543	71.820	239.977	87.358.047
PPE	5.478.000.000	805.668	3.953.259	0	69.949	188.784	479.205	82.046.364
Liquidity	5.500.000.000	2.322.255	2.593.976	97.129	1.124.654	1.505.083	2.410.178	19.737.054
Age	5.671.000.000	14.539.058	8.030.378	0	9.000.000	13.000.000	19.000.000	60.000.000
Quick ratio	5.447.000.000	1.322.715	3.531.191	-31.298.624	496.845	898.045	1.564.948	33.586.066
Inventory turnover ratio	5.632.000.000	859.604.250	15.613.972.332	2.441	2.473.381	4.971.114	11.715.217	381.933.400.047
FA turnover ratio	5.493.000.000	60.869.434	550.442.808	916	1.914.813	6.022.258	17.069.339	15.679.062.105
TA turnover ratio	5.488.000.000	1.208.536	1.208.844	257	464.491	919.167	1.561.646	12.733.542
DSO	5.493.000.000	270.502.417	1.073.990.343	0	26.725.062	63.735.184	158.891.812	12.188.339.931
Capital intensity	5.488.000.000	3.561.490	16.194.301	78.533	640.350	1.087.942	2.152.895	246.062.040
Expense of revenue ratio	5.493.000.000	1.431.539	4.363.704	0	453.810	773.390	1.170.790	62.679.250
Operating margin	5.493.000.000	156.534	1.286.881	-22.413.873	23.277	60.170	141.389	22.582.316
OCF	5,50E+09	1,44E+17	9,74E+17	-9,98E+18	-5,79E+15	1,78E+16	9,45E+16	2,79E+19
BEP	5.500.000.000	92.343	252.319	-10.587.992	22.247	58.002	114.819	7.906.537
CCC	5.632.000.000	381.484.782	1.876.425.356	-25.121.662.701	50.453.632	114.445.307	225.237.739	26.335.734.538
GDP	5.676.000.000	6.001.623	1.131.059	2.905.836	5.421.883	6.240.303	6.812.246	7.075.789
CPI	5.676.000.000	5.821.202	4.809.275	631.201	2.795.824	3.539.628	9.094.703	18.677.732
Interest rates	5.676.000.000	3.797.918	2.944.192	-3.551.709	2.294.892	4.825.874	5.814.896	7.322.258
Target	5.676.000.000	385.483	486.752	0	0	0	1.000.000	1.000.000

4. BUILDING MODELS AND VISUALIZATION OF DATA

The models are using 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 80:20. The split ratio of 80-20 is commonly used, where 80% of the data is used for training and 20% is used for testing. And the 80-20 split is a reasonable compromise between having enough data for training and testing the model's performance on new data. The training set is utilized to train the machine learning model, while the testing set is used to evaluate the model's performance. Different machine learning algorithms such as Logistic Regression, K-nearest neighbours, Decision Tree, Support vector Machine (Linear and RBF Kernel), Neural Network, and Random forest are tested to find the most optimal algorithm.

Table 2: *Various models' scores to search for the best one*

Algorithm	**Model score**
Logistic Regression	75.71%
K-Nearest Neighbors	76.09%
Decision Tree	71.82%
Support Vector Machine (Linear Kernel)	75.52%
Support Vector Machine (RBF Kernel)	80.05%
Neural Network	79.70%
Random Forest	80.46%

The score shows that Random forest is the best model to forecast corporate profits with 80.46%. This can be explained random forest is suitable for Vietnamese dataset, which is high level of randomness.

To determine the optimal number of trees for the Random Forest model, a GridSearch is conducted across a range of values from 1 to 100. The results demonstrate that the most effective number of trees is reached at the 81st iteration. Following this analysis, the Random Forest model is then trained with this optimal number of trees to ensure the highest level of accuracy possible.

Research model:

Target = β_0 + β_1Growth$_{it}$ + β_2Size$_{it}$ + β_3Age$_{it}$ + β_4Liq$_{it}$ + β_5Ppe$_{it}$ + β_6Quick$_{it}$ + β_7Inv turnover ratio$_{it}$ + β_8FA turnover ratio$_{it}$ + β_9TA turnover ratio$_{it}$ + β_{10}DSO$_{it}$ + β_{11}Capital intensity$_{it}$ + β_{12}Expense revenue ratio$_{it}$ + β_{13}Operating margin$_{it}$ + β_{14}BEP$_{it}$ + β_{15}CCC$_{it}$ + β_{16}GDP$_{it}$ + β_{17}CPI$_{it}$ + β_{18}Interest Rate$_{it}$ + ε_{it}.

In there:

- Target variable: I choose to use return on total assets (ROA) as the condition to create binary values (0 and 1).
- Dependent variables: Capital investment rate (PPE), Inventory turnover ratio, Fixed asset turnover ratio, total asset turnover ratio, day sales outstanding (DSO), Capital intensity, Expense revenue ratio, operating margin, basic earning power (BEP), cash conversion cycle (CCC).
- Control variables: business's size variable (Size), business's Age variable (Age), sale growth variable (Growth), liquidity variable (Liq).
- In addition, I also use macro variables: Gross Domestic Product (GDP), Consumer Price Index (CPI), and Interest Real Rate (Interest Rate).

The above model includes data from businesses with industries in the dataset from 2010 -2020. In addition, we disaggregate data by industry to compare differences with variables like the model above. For model decomposition by industry, which is formed 9 models corresponding to 9 industries: Utilities, Basic Materials, Industrials, Consumer Cyclicals, Consumer Non-Cyclicals, Energy, Healthcare, Real Estate, and Technology.

5. RESULTS AND MODEL EVALUATION

First, the model will evaluate the predictive model for the entire business. When the number of estimators is 75, the model has a relatively decent accuracy index (0.76), as shown in Figure 2.

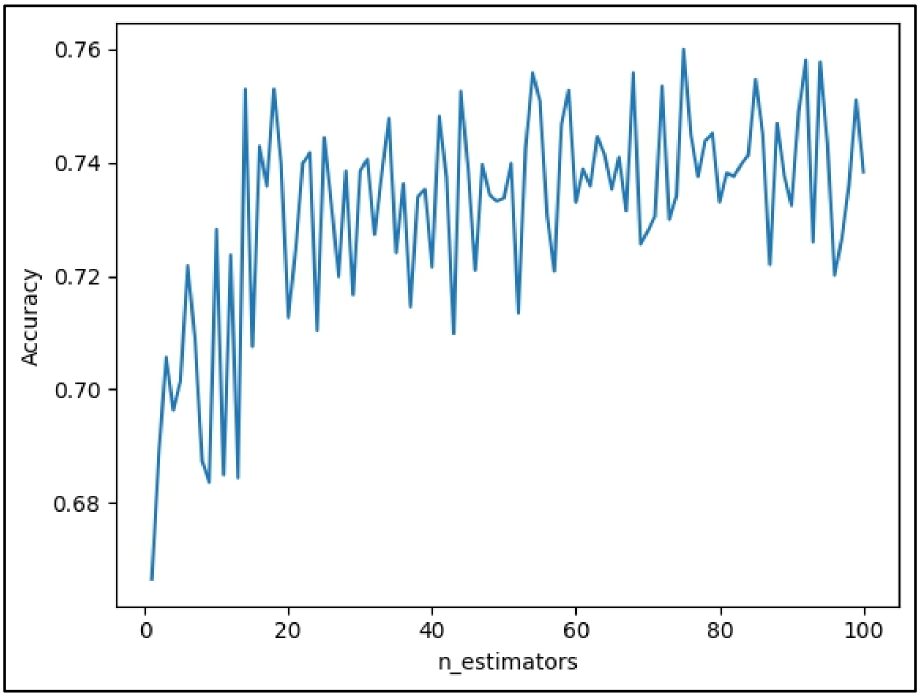


Figure 2. Chart of accuracy line with the various number of n_estimators in the total industries model

Also, the ROC-AUC findings of this model are pretty good, as shown in Figure 3. The 5-fold AUC values vary from 0.88-0.91 which are high, showing that the model is doing well.

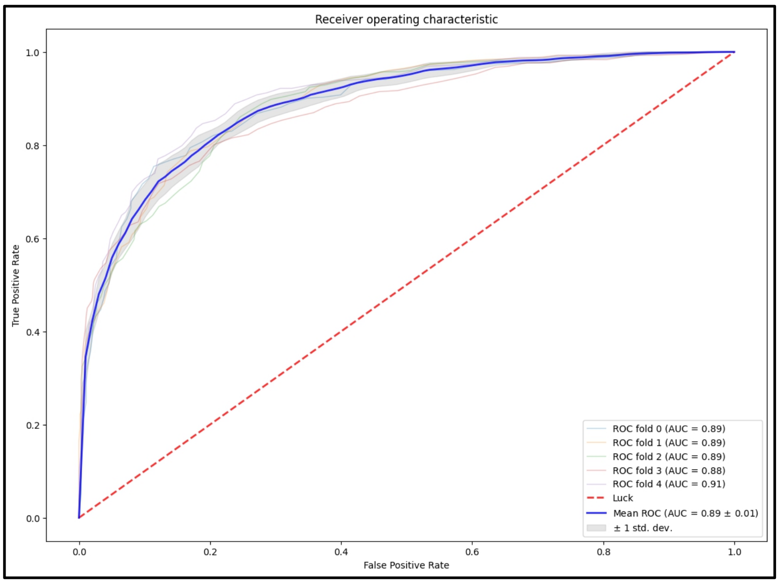


Figure 3. Chart of ROC curve in the total industries model

Formulae for Classification report:

$$\text{Accuracy}=\frac{TP+TN}{n}$$

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

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

Furthermore, the study found that the Precision and the Recall of both groups were good based on the indications in the classification report (Table 3). Demonstrates the model's ability to anticipate the accuracy of profit and loss estimates for the market as a whole. In addition, with the precision and recall of Loss class are 0.82 and 0.84, the capacity to foresee nearly entire business losses in the market. The Profit group's Recall index of 0.69 is relatively good, indicating that the model overlooks nearly 70 percent investment possibilities from organizations that are actually lucrative in the market but are expected to lose money.

Based on the table 3 classification report, the model has a decent overall accuracy of 0.78. However, when we look at the precision and recall values for the target class of 1 which we expected to, we can see that they are lower than for class 0. This suggests that the model is better at correctly identifying instances of firm that are actually unprofitable than another one.

However, when examining the Precision and Recall values more closely, it becomes clear that the model's performance varied depending on the target class being predicted. For instance, the model achieved a Precision score of 0.72 for the Profit class, indicating that when it predicted a company would be profitable, it was correct approximately 72% of the time. Similarly,the Recall score of 0.69 implies that the model identified around 69% of all instances of actually profitable companies in the dataset, offering investors a valuable reference for investment decisions.

On the other hand, the model's performance was more impressive when predicting the Loss class. The Precision score of 0.82 suggests that the model accurately predicted a loss 82% of the time, while the Recall score of 0.84 indicates that it correctly identified approximately 84% of all instances of companies expected to incur losses. This can potentially help investors avoid investing in failing companies.

Overall, while the model showed promise in accurately identifying profitable and unprofitable companies in the market, further refinement may be needed to enhance its ability to predict profitable companies. Future research could explore alternative modeling techniques or feature engineering methods to achieve more accurate predictions.

$$\text{F1-score} = \frac{2* precision * recall}{precision + recall}$$

Table 3: *Classification report of the total industries model*

	precision	recall	f1-score	support
0.0	0.82	0.84	0.83	657
1.0	0.72	0.69	0.71	397
accuracy			0.78	1054
macro avg	0.77	0.76	0.77	1054
weighted avg	0.78	0.78	0.78	1054

Decomposition of the industries

This study recognizes that the measurement rules for ROA can vary across different industries because the asset structure and ROA can vary significantly across industries. Therefore, there is a need to analyze ROA decomposition by industries to obtain meaningful insights and accurate comparisons.

Table 4: Features' importance of the model decomposed by industries

Decomposition of the industry	Obs	Result						
		Test-set obs	Accuracy	AUC score	Precision		Recall	
					Profit	Loss	Profit	Loss
All industries	**5002**	**1001**	**78%**	**89%**	**72%**	**82%**	**69%**	**84%**
Basic materials	841	169	78%	82%	74%	80%	72%	82%
Consumer cyclicals	769	154	79%	90%	80%	79%	74%	84%
Consumer non-cyclicals	478	96	73%	88%	71%	74%	62%	79%
Energy	375	75	82%	90%	79%	84%	79%	84%
Health care	161	33	82%	95%	86%	80%	75%	89%
Industrials	1461	293	82%	91%	76%	85%	70%	88%
Real estate	534	107	72%	85%	62%	79%	67%	76%
Technology	166	34	60%	87%	78%	54%	37%	88%
Utilities	217	44	71%	87%	67%	73%	47%	86%

Table 4 presents significant variations in the AUC scores among different industries. Notably, all the sectors exhibit comparatively high AUC scores (ranging from 82 percent to 95 percent) and Health Care sector got the highest score when compared to other industries, suggesting that the predictive model for this sector is exceptionally effective in identifying profitable firms. This finding is consistent with a study by Mieszko Mazur et al. (2021), which demonstrates that during the COVID-19 crash in March 2020, natural gas, food, healthcare, and software stocks experienced high positive returns.

Table 5. Classification report of the Healthcare industry model

	precision	recall	f1-score	support
0.0	0.80	0.89	0.84	18
1.0	0.86	0.75	0.80	16
accuracy			0.82	34
macro avg	0.83	0.82	0.82	34
weighted avg	0.83	0.82	0.82	34

The healthcare industry forecasting model has an accuracy score of 0.82, which indicates a high ability to predict companies in this industry reliably, and a recall value of 0.75, which indicates a good ability to correctly anticipate businesses in this sector. There is very little chance of missing out on highly profitable firms. The healthcare industry has a strong data gathering system, few missing data points, and consistent growth with little fluctuation. Additionally, the government and market systems strictly control the healthcare industry to reduce any negative effects on community services. As a consequence, the model has the highest accuracy of 0.82 and the highest AUC value (0.94-0.98, Figure 4) when compared to the other industries.

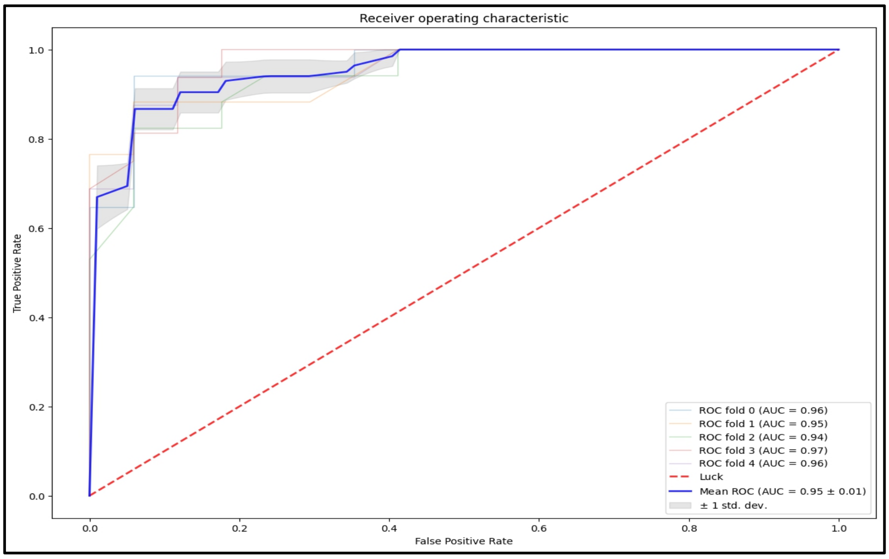


Figure 4. Chart of ROC curve of the Healthcare industry model

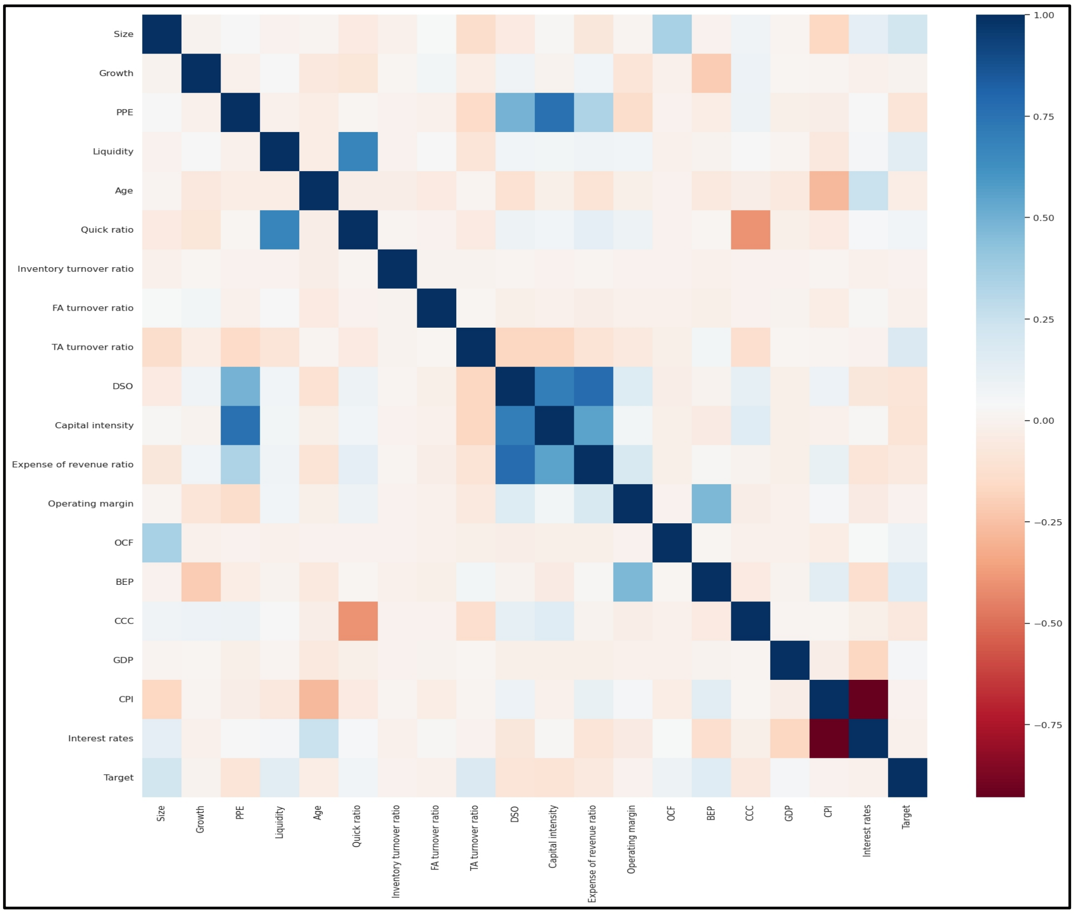


Figure 5. Heatmap of features’ importance in the model decomposed by industries

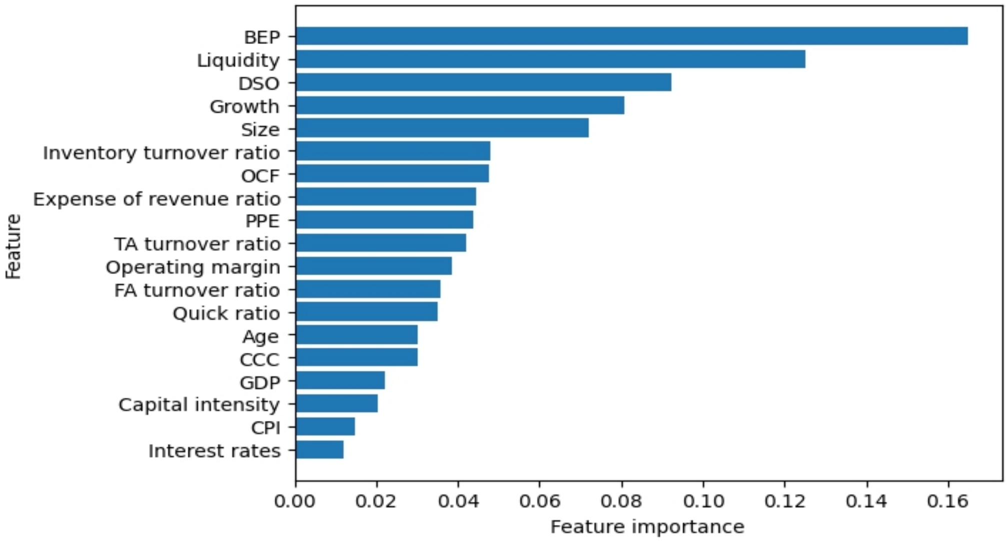


Figure 6: Feature importance values occupied in each variable for the total industries model

Figure 5 and figure 6 provides a clear illustration of the impact of various factors on profitability measures. The factors are divided into three categories based on their degree of impact (ranging from 2-16 percent). The primary set of factors, with the highest impact, consists of fundamental metrics commonly used to assess a company's performance,, such as Basic earning power (BEP), Liquidity, Growth, Size and Days Sales Outstanding (DSO). The second most important category of factors includes other financial indicators like Quick ratio, Cash Conversion Cycle (CCC), and Age. Finally, the remaining factors are macro variables, which have a lower impact on profitability measures but are still significant in the context of classification by industry.

It is worth noting that the impact of macro variables is more pronounced when analyzing profitability measures by industry. This is because different industries are subject to different macroeconomic conditions, and these conditions can have a significant impact on their profitability. On the other hand, when analyzing profitability measures by year, all companies are subject to the same systematic risk, so the impact of macro factors is less pronounced. These findings are consistent with previous research, such as the study by (Mieszko Mazur et al., 2021), which found that certain industries, like healthcare, outperformed others during the COVID-19 pandemic.

6. DISCUSSION AND CONCLUSION

6.1 Conclusion

This essay explores the potential of using the Machine Learning (ML) approach, specifically the Random Forest method, to generate out-of-sample profitability forecasts that surpass traditional regression-based methods. ML approaches are preferred due to their ability to handle nonlinearity in data and their superior prediction accuracy. Using a large sample of Vietnamese firms, the study generates out-of-sample predictions for directional changes in over twenty profitability measures. Results show that the ML method achieves high classification accuracy between 60% and 82%, which is significantly higher than the accuracy of random walk forecasts.

The analysis further examines the performance by industry and reveals that the healthcare sector achieves the highest accuracy of approximately 82%, while the real estate sector exhibits the lowest accuracy at around 72%. While all industries in the dataset show commendable accuracy scores, the study identifies the healthcare, energy, and industrial sectors as particularly suitable for the applied ML model.

The essay also highlights the significance of the healthcare industry, which has attracted considerable investor interest and generated substantial income. The study attributes this success to the explicit and limited missing values in the healthcare dataset, which contribute to more robust evaluation processes. The availability of advanced medical equipment is emphasized as a critical factor in enhancing the efficiency and quality of medical work, facilitating rapid, safe, and effective diagnosis and treatment by healthcare professionals.

6.2 Future study

Additionally, this Machine Learning model has little issues with anticipated profit companies. The results reveal that the model is unable to forecast which companies would actually have money, This needs to be improved in the future.

This study can be further improved by adding more features, employ different ML methods which are not used in this study, explode the feature space to properly handle categorical variables, and expand the set of hyperparameters over. Future work can also explore the use of accruals as features in the prediction of cash flows. Finally, the results can be compared to the results of a recent study by Vorst and Yohn (2018), who perform out-of-sample predictions using a regression methodology, but without using the random walk as a benchmark.

As a result of the significant number of decision trees employed in the testing process, random forests are widely considered to be a powerful and accurate approach to data analysis. One of the primary advantages of this method is that it does not suffer from overfitting issues, as it averages out the results of all forecasts, canceling out any potential bias. This approach is particularly applicable to this research, given the significant number of variables involved and the need to account for missing data. In dealing with missing data, two methods can be employed: replacing continuous variables with mean values and estimating the average of missing values. This can help identify the most significant features contributing to the classifier and help inform future research.

However, this method does have some limitations. Due to the large number of decision trees involved, it can be time-consuming to make predictions. This is because each tree in the forest must make a prediction for the same input, and then vote on the outcome. This process can be lengthy and complex, particularly compared to decision trees, which allow for quicker decision-making by following a clear path through the tree. Moreover, the dataset has not been updated yet because some objective reasons.

Additionally, while the dataset used in this study provided a valuable starting point for analysis, there are limitations to its accuracy and completeness. For example, there were some unreasonable data points such as inventory being higher than current assets. As such, future research should seek to improve the dataset by updating it or finding ways to better handle such outliers. Furthermore, The dataset used in this study is not the most recent one, which could limit the applicability of the findings to current market conditions.

Overall, while the results of this study provide valuable insights into the factors that contribute to company profitability, there is always room for improvement. In future research, it may be possible to improve results by sacrificing model interpretability, for example, by adding more features, expanding the set of hyperparameters over which to search, or exploring the use of accruals as features in the prediction of cash flows. Comparing the results to those of other studies, such as Vorst and Yohn's (2018) out-of-sample predictions using regression methodology without using the random walk as a benchmark, could also provide further insights.

REFERENCES

Vietnamese researches:

[1] Hieu, P. M., & Do Quyen, N. (2021). Earnings quality measurements and determinants: the case of listed firms in Vietnam. Journal of International Economics and Management, 21(3), 22-46.

[2] Linh, T. T. T. (2019). The impact of economic value added (EVA) and return on assets on created shareholder value of Vietnamese firms. VNUHCM Journal of Economics, Business and Law, 3(4), 402-417.

[3] Nguyen Anh Phong, Phan Huy Tam, Nguyen Ngoc Hieu, 2021, Đo lường hiệu quả hoạt động doanh nghiệp qua chỉ số giá trị thị trường và chỉ số giá trị sổ sách bằng phương pháp máy học, Tạp chí Kinh tế châu á - Thái Bình Dương.

[4] Nguyen Hoang Anh, Nguyen Thi Tu, 2017, Factors affecting the profit of enterprises listed on HOSE.

[5] Vu, H. T. (2018). *Basic Machine Learning.* Vietnam: Vietnamese Education Publisher.

Foreign research:

[6] Agiomirgianakis, G., Voulgaris, F., & Papadogonas, T. (2006). Financial factors affecting profitability and employment growth: The case of Greek manufacturing. *International Journal of Financial Services Management*, 1(2/3), 232-242.

[7] Alarussi, A. S., & Alhaderi, S. M. (2018). *Factors affecting profitability in Malaysia*. Journal of Economic Studies, 45(3), 442-458.

[8] Anand, V., Brunner, R., Ikegwu, K., & Sougiannis, T. (2019). *Predicting Profitability Using Machine Learning*. Retrieved from

[9] A. Munde, Nandita Mishra(2022), Corporate performance: SMEs performance prediction using the decision tree and random forest models

[10] Ball, R., & Foster, G. (1982). *Corporate Financial Reporting: A Methodological Review of Empirical Research*. Journal of Accounting Research, 20, 161–234.

[11] Barton, J., & Waymire, G. (2004). *Investor protection under*

unregulated financial reporting. Journal of Accounting and Economics, 38(1), 65–116.

[12] Breiman, L. (2001). *Machine Learning*, 45(1), 5-32.

[13] Brigham, E., & Houston, J. *Fundamentals of financial management*.

[14] Burja, C. (2011). Factors Influencing the Companies' Profitability.

[15] Cerf, A. R. (1961). *Corporate reporting and investment decisions*. Berkeley, University of California Press.

[16] Chander, S., & Aggarwal, P. (2008). *Determinants of corporate profitability: An empirical study of Indian drugs and pharmaceutical industry*. Paradigm, 12(2), 51-61.

[17] Eric Zhang (2021), Forecasting Financial Performance of Companies For Stock Valuation

[18] Gopalan, R., & Jayaraman, S. (2012). *Private Control Benefits and Earnings Management : Evidence from Insider Controlled Firms.* Journal of Accounting Research, 50(1), 117–157.

[19] Goyal, C. (2021). *Feature Engineering – How to Detect and Remove Outliers (with Python Code)*. Retrieved from https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/

[20] H. Chen (2009), Model for Predicting Financial Performance of Development and Construction Corporations

[21] Lewinson, E. (2020). *Python for Finance Cookbook*. Birmingham, UK: Packt Publishing Ltd.

[22] Liu, M., Shi, Y., Wilson, C., & Wu, Z. (2017). *Does family involvement explain why corporate social responsibility affects earnings management?.* Journal of Business Research, 75, 8–16

[23] Maoze Zhou, Hongjiu Liu, Yanrong Hu (2022), Research on corporate financial performance prediction based on self-organizing and convolutional neural networks

[24] Nouri, Y. (2018). *Random Forest & K-Fold Cross Validation.*

[25] Rahman, M. M., Hamid, M. K., & Khan, M. A. M. (2015). Determinants of bank profitability: Empirical evidence from Bangladesh. International journal of business and management, 10(8), 135.

[26] *sklearn.model_selection.RandomizedSearchCV.* Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

[27] Suardana, I. B. R., Astawa, I. N. D., & Martini, L. K. B. (2018). Influential factors towards return on assets and profit change (study on all BPR in Bali province). International journal of social sciences and humanities, 2(1), 105-116.

[28] Watts, R. L., & Zimmerman, J. L. (1990). *Positive Accounting Theory: A Ten Year Perspective*. The Accounting Review, 65(1), 131–156.

[29] Xinyue, C., Zhaoyu, X., & Yue, Z. (2020). *Using Machine Learning to Forecast Future Earnings.*

[30] Yong-Hua Cai, Qi Yin, Q. Su, Xinyu Huang, Yin Zhang, T. Liu (2020), Prediction Method of Enterprise Return on Net Assets Based on Improved Random Forest Algorithm

APPENDICES

Appendix 1. Table of raw data

Variable	Code
Date Became Public	"Public Date" sheet, "Date Became Public" column
Organization Founded Year	"Public Date" sheet, "Organization Founded Year" column
Company Code	"Name" sheet, "Identifier" column
Industry Code	"Name" sheet, "TRBC Economic Sector Name" column
Return On Asset	"ROA" sheet
Total Asset	"TA" sheet
Market Value	"MV" sheet
Total Liability	"Total Lia" sheet
Sales	"Sales" sheet
Tangible Fixed Asset	"Tangible FA" sheet
Total Current Asset	"Total CA" sheet
Current Liability	"Total Current Lia" sheet
Fixed Asset	"Fixed assets" sheet
Accounts Receivable	"Accounts Receivable" sheet
Cost Of Revenue	"Cost of revenue" sheet
Earning Before Interest And Tax	"EBIT" sheet
Equity	"Equity" sheet
Average Receivable Day	"Avg. Receivable days" sheet
Average Payable Day	"Avg. Payable days" sheet
Average Inventory Day	"Avg. Inventory days" sheet
Gross Domestic Product	"GDP" sheet
Consumer Price Index	"CPI" sheet
Interest Real Rate	"Interest rates" sheet

Appendix 2. Table of intermediate variables

Variable	Code	Formula
Net Income	NI	ROA * TA

Inventory	Inv	Sales/Inventory Turnover Ratio

Appendix 3. Table of independent variables

Variable	**Code**	**Formula**
Growth Of Sales	Growth	$\frac{Sale_t - Sale_{t-1}}{Sale_{t-1}}$
Market Size Of Company	Size	*In(Market value)*
Age Of Company	Age	*2020 - Founding year*
Liquidity Ratio	Liq	$\frac{Total\ current\ asset}{Current\ liabilities}$
Property, Plant, & Equipment Turnover Ratio	PPE	$\frac{Tangible\ Fixed\ Asset}{Sales}$
Quick Ratio	Quick ratio	$\frac{Total\ current\ asset - Inventory}{Current\ liabilities}$
Inventory Turnover Ratio	Inventory turnover ratio	$\frac{365}{Average\ inventory\ days}$
Fixed Asset Turnover Ratio	FA turnover ratio	$\frac{Sales}{Fixed\ assets}$
Total Asset Turnover Ratio	TA turnover ratio	$\frac{Sales}{Total\ asset}$
Days Sale Outstanding	DSO	$\frac{Account\ receivable}{Sale/365}$
Capital intensity	Capital intensity	$\frac{Total\ asset}{Sales}$
Expense Of Revenue Ratio	Expense of revenue ratio	$\frac{Costs\ of\ revenue}{Sales}$
Operating margin	Operating margin	$\frac{EBIT}{Sales}$
Basic Earning Power Ratio	BEP	$\frac{EBIT}{Total\ asset}$
Cash Conversion Cycle	CCC	*Average inventory days + Average receivable days - Average payable days*
Gross Domestic Product	GDP	*None*
Consumer Price Index	CPI	*None*
Interest Real Rate	Interest rates	*None*

Appendix 4. Table of target variables

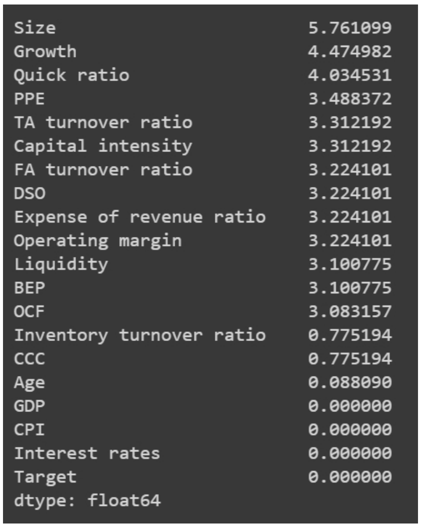
Variable	**Code**
Return on assets	ROA

Appendix 5. Variables' importance in the model decomposed by industries

	All industries	Basic materials	Consumer cyclicals	Consumer non-cyclicals	Energy	Health care	Industrials	Real estate	Technology	Utilities
ROA	0.091	0.111	0.101	0.091	0.074	0.154	0.075	0.042	0.053	0.075
Net profit margin	0.076	0.081	0.076	0.034	0.064	0.109	0.062	0.094	0.045	0.048
ROE	0.076	0.073	0.064	0.059	0.07	0.094	0.059	0.063	0.031	0.044
DSO	0.056	0.054	0.058	0.046	0.057	0.033	0.048	0.026	0.051	0.031
Growth	0.051	0.062	0.059	0.051	0.057	0.021	0.052	0.048	0.052	0.029
Leverage	0.048	0.046	0.047	0.062	0.031	0.059	0.059	0.047	0.029	0.043
CCC	0.047	0.042	0.041	0.044	0.05	0.025	0.048	0.043	0.047	0.055
BEP	0.045	0.046	0.037	0.035	0.037	0.018	0.043	0.041	0.059	0.014
PPE	0.045	0.038	0.043	0.033	0.075	0.043	0.047	0.043	0.028	0.052
Expense of revenue ratio	0.044	0.042	0.036	0.047	0.04	0.025	0.039	0.086	0.064	0.062
EPS	0.043	0.049	0.054	0.045	0.053	0.019	0.049	0.055	0.042	0.032
Inv turnover ratio	0.041	0.040	0.034	0.052	0.039	0.042	0.048	0.047	0.036	0.072
Liquidity	0.039	0.030	0.032	0.038	0.035	0.038	0.037	0.036	0.051	0.034
Size	0.039	0.039	0.033	0.04	0.038	0.027	0.044	0.047	0.041	0.075
Operating margin	0.039	0.036	0.039	0.032	0.026	0.023	0.037	0.052	0.035	0.049
FA turnover ratio	0.038	0.036	0.038	0.034	0.067	0.045	0.047	0.047	0.038	0.105
Quick ratio	0.036	0.032	0.035	0.034	0.039	0.033	0.037	0.038	0.035	0.05
TA turnover ratio	0.035	0.027	0.043	0.04	0.037	0.058	0.042	0.023	0.061	0.033
Capital intensity	0.033	0.029	0.043	0.041	0.023	0.047	0.042	0.031	0.063	0.008
Age	0.027	0.023	0.030	0.032	0.029	0.022	0.028	0.039	0.085	0.013
CPI	0.018	0.02	0.018	0.02	0.023	0.02	0.022	0.025	0.016	0.015
Interest	0.018	0.022	0.020	0.017	0.018	0.033	0.02	0.014	0.019	0.049

rates										
GDP	0.017	0.021	0.017	0.02	0.017	0.01	0.019	0.016	0.018	0.013

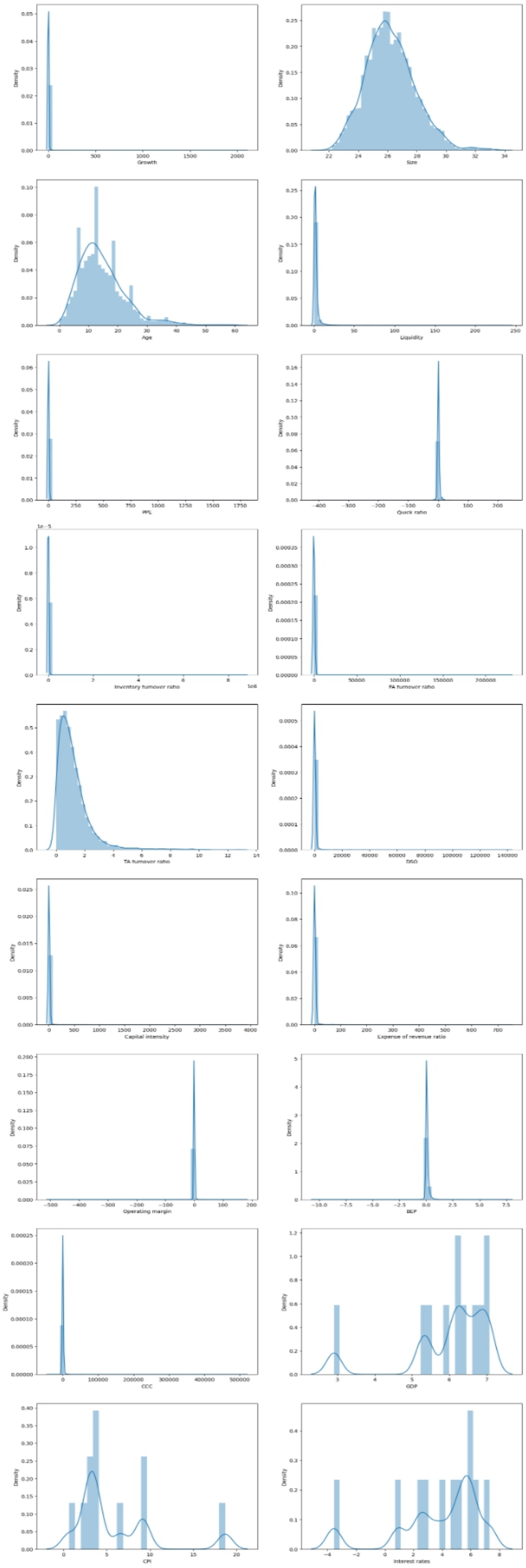
Appendix 6. Proportion of missing values in each column



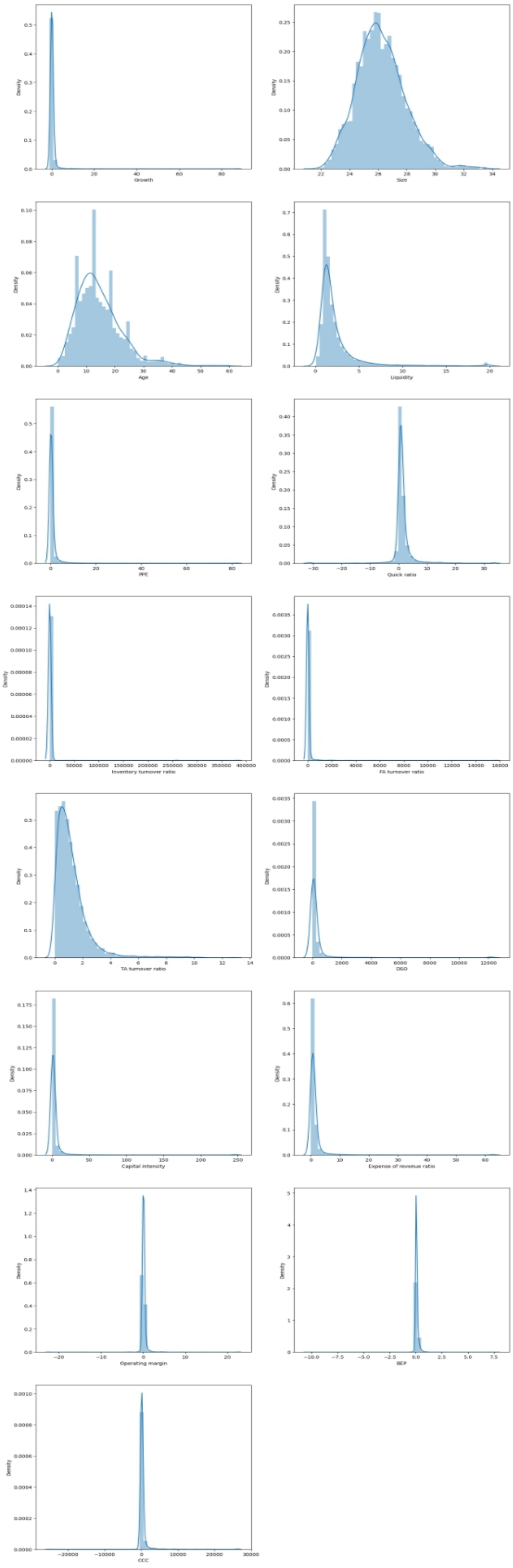
```
Size                         5.761099
Growth                       4.474982
Quick ratio                  4.034531
PPE                          3.488372
TA turnover ratio            3.312192
Capital intensity            3.312192
FA turnover ratio            3.224101
DSO                          3.224101
Expense of revenue ratio     3.224101
Operating margin             3.224101
Liquidity                    3.100775
BEP                          3.100775
OCF                          3.083157
Inventory turnover ratio     0.775194
CCC                          0.775194
Age                          0.088090
GDP                          0.000000
CPI                          0.000000
Interest rates               0.000000
Target                       0.000000
dtype: float64
```

Appendix 7. Chart of values' distribution before fixing outliers in the total industries

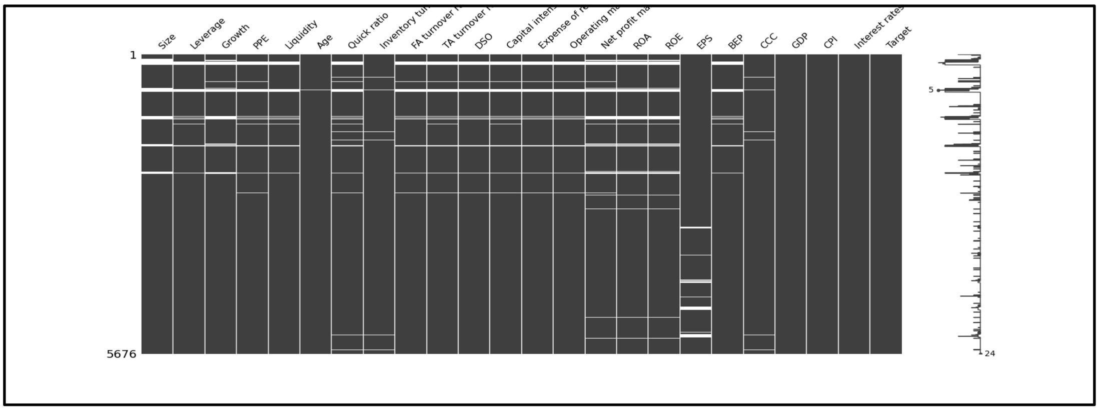
model



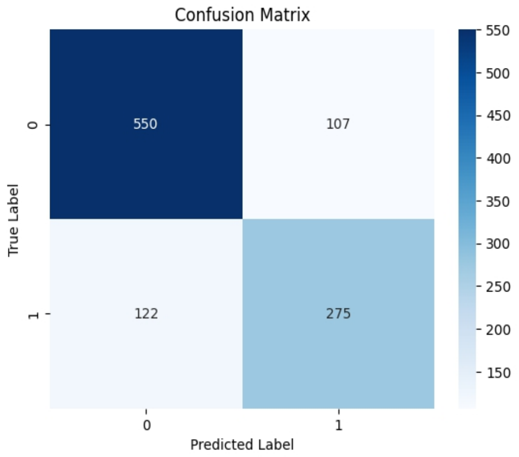
Appendix 8. Chart of values' distribution after fixing outliers in the total industries model



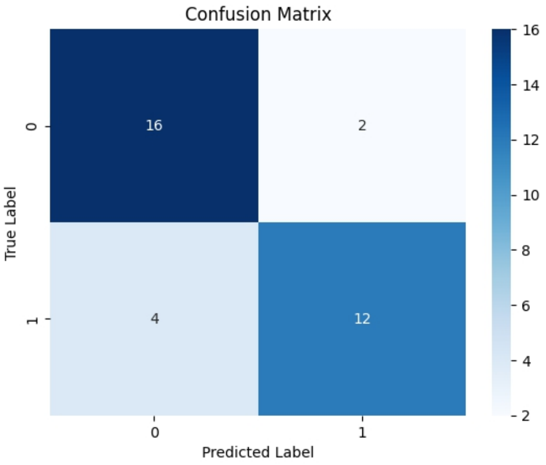
Appendix 9. Chart of missing values' distribution in the total industries model



Appendix 10. Chart of confusion matrix in the total industries model



Appendix 11. Chart of confusion matrix in the Healthcare model


Confusion Matrix
True Label
Predicted Label