



VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
UNIVERSITY OF ECONOMICS AND LAW

FINAL REPORT

**SCIENTIFIC RESEARCH TOPIC STUDENTS
PARTICIPATE IN THE AWARD "UEL SCIENCE"
2023**

Topic:

PROFIT PREDICTION OF LISTED ENTERPRISES IN VIETNAM BY MACHINE LEARNING

Scientific field: Economic field

Specialization: Finance - Banking

TP.HCM, March 2022



VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
UNIVERSITY OF ECONOMICS AND LAW

FINAL REPORT

**SCIENTIFIC RESEARCH TOPIC STUDENTS PARTICIPATE
IN THE AWARD "UEL SCIENCE"**

2022

Topic:

**PROFIT PREDICTION OF LISTED ENTERPRISES IN
VIETNAM BY MACHINE LEARNING**

No	Name	ID	Majors	Mission	Phone	Email
1	Trần Thanh Phúc	K19414 1740	Banking and Finance	Leader	08533 45213	phuctt19414c@ st.uel.edu.vn
2	Phạm Quỳnh Hương	K19414 1724	Banking and Finance	Member	03458 09036	huongpq19414c @st.uel.edu.vn
3	Nguyễn Thị Huệ Minh	K19414 1733	Banking and Finance	Member	03365 20966	minhnth19414c @st.uel.edu.vn
4	Hồ Thùy Dung	K19414 1718	Banking and Finance	Member	03752 88743	dunght19414c @st.uel.edu.vn
5	Võ Thụy Uyên Nhi	K19414 1736	Banking and Finance	Member	07725 22418	nhivtu19414c@ st.uel.edu.vn

TP.HCM, March, 2022

ABSTRACT

Data science and big data are growing prominence these days, and they may provide a variety of choices to advance the field of finance and commercial planning. Machine Learning is a popular area of computer science that has grown in popularity over the last two decades. When compared to traditional forecasting approaches, using Machine Learning in profit forecasting of listed businesses is a novel and successful strategy such as Regression, Random Forest, Random Walk, Classification, and so on. Capturing this trend, using the Random Forest approach based on Machine Learning to assess and anticipate the profitability of listed firms, while also using real historical data from Vietnamese enterprises' financial statements. It is now relevant to the Vietnamese economy in order to create a model to anticipate the future movement of profits of listed firms, consequently providing helpful information to investors. Investors, corporate governance, and lenders make decisions. 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 team proposes to perform profit forecasting research using two ways: by year and by industry.

TABLE OF CONTENTS

ABSTRACT	1
TABLE OF CONTENTS	3
LIST OF TABLES	5
LIST OF PICTURES	6
LIST OF ACRONYMS	8
INTRODUCTION	8
1.1 Background and justification for the research project	9
1.2 Study overview	9
1.2.1 Domestic studies	9
1.2.2 Foreign studies	10
1.3 Objectives of the study	10
1.4 Research subjects	11
1.5 Research scope	11
1.6 Research Methods	11
1.7 Expected results	11
1.8 New point of research	11
CONTENT	13
CHAPTER 1: THEORETICAL BASIS AND REALITY OF THESIS	13
1.1. Theoretical basis	13
1.1.1. Machine learning	13
1.1.2. Reasons for choosing Machine learning	13
1.1.3. Supervised learning	13
1.1.4. Random forest	13
1.1.6. ROC Curve	14
1.1.7. AUC	14
1.1.8. Enterprise profit	14
1.1.9. Operating cash flow	15
1.2. Real state of affairs	15
CHAPTER 2: PREPROCESSING DATA	16
2.1. Data resources	16
2.2. Preprocessing data	16
2.2.1. Data Cleaning	16
2.2.2. Data transforming	16
2.3. Variables	17

2.3.1. Target variable	17
2.3.2. Sale growth rate (Growth)	17
2.3.3. Size	17
2.3.4. Age	18
2.3.5. Liquidity ratio (Liq)	18
2.3.6. Leverage ratio (Lev)	18
2.3.7. Capital investment ratio (PPE)	18
2.3.8. Quick ratio	18
2.3.9. Inventory turnover ratio	18
2.3.10. Fixed asset turnover ratio	18
2.3.11. Total asset turnover ratio	19
2.3.12. Days sale outstanding (DSO)	19
2.3.13. Capital intensity	19
2.3.14. Expense revenue ratio	19
2.3.15. Operating margin	19
2.3.16. Net profit margin	19
2.3.17. Basic earning power (BEP)	19
2.3.18. Return on total asset (ROA)	20
2.3.19. Return on equity (ROE)	20
2.3.20. Earning per share (EPS)	20
2.3.21. Cash conversion cycle (CCC)	20
2.3.22. Gross Domestic Index (GDP)	20
2.3.23. Consumer Price Index (CPI)	20
2.3.24. Real interest rate	20
CHAPTER 3: BUILDING MODELS AND VISUALIZATION OF DATA	21
CHAPTER 4: RESULTS AND MODEL EVALUATION	23
4.1 Decomposition of the years	24
4.2 Decomposition of the industries	27
DISCUSSION AND CONCLUSION	30
5.1 Conclusion	30
5.2 Discussion and future study	31
REFERENCES	32
APPENDICES	34

LIST OF TABLES

Table name	Page
Table 3. Table of various models' scores in order to search for the best one	23
Table 4.1.1. Table of features' importance of the model decomposed by years	26
Table 4.2.1. Table of features' importance of the model decomposed by industries	29

LIST OF PICTURES

Picture name	Page
Figure 2.4. Correlation matrix of variables	22
Figure 4.1. Chart of accuracy line with the various number of n_estimators in the total industries model	25
Figure 4.2. Chart of ROC curve in the total industries model	25
Figure 4.3. Classification report of the total industries model	26
Figure 4.1.1. Classification report of the model decomposed by years	27
Figure 4.1.2. Chart of ROC curve of the year 2020 model	27
Figure 4.1.3. Heatmap of features' importance in the model decomposed by years.	28
Figure 4.2.2. Classification report of the Healthcare industry model	30
Figure 4.2.3. Chart of ROC curve of the Healthcare industry model	31
Figure 4.2.4. Heatmap of features' importance in the model decomposed by industries	31

LIST OF ACRONYMS

Avg. Inv days	Average Inventory days
Avg. Payable days	Average Payable days
Avg.Receivable	Average Receivable
FA turnover ratio	Fixed asset turnover
DSO	Days sale outstanding
CCC	Cash conversion cycle
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

INTRODUCTION

1.1 Background and justification for the research project

Along with the rapid development of digital transformation, which is contributing to the diversification of financial activities, the use of Machine Learning methods in applied finance is becoming increasingly popular in countries all over the world, particularly in Vietnam, where the financial market is emerging and vibrant. The research team comprehended the concept and carried out the study "Forecasting profitability of listed firms in Vietnam over time using Machine Learning." According to the group's knowledge, the studies on this issue employed a lot of traditional approaches but ran into a lot of constraints when it came to running and displaying model and predicting findings. As a result, in order to increase performance and create an efficient model, the team used Machine Learning as a research approach to fulfill the demands of digital transformation and overcome limits imposed by traditional methods and increasing predictive effectiveness.

1.2 Study overview

Profit has always been one of the most important aspect for users of financial information, such as investors, creditors, and administrators, because it is the consequence of production and business. 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.

Companies in Vietnam build a positive image to attract outside investment by altering profits to minimize corporate income tax. As a result, firms may change subjective aspects to generate high 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. These elements have access to the enterprise's business processes and are in close contact with one another. The aforesaid components are then studied and analyzed to find the subjective and objective aspects that influence business earnings. On that basis, forecast business earnings in order to make management or investment decisions in the context of market processes managed by the state.

1.2.1 Domestic studies

There are many studies on forecasting corporate profits.

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 (ROA), Return on Equity (ROE) to measure the performance of the business. In addition, the author also used some other control variables such as

Operating Cash Flow (OCF), Age of the enterprise (Age) which we also studied and applied in this study.

1.2.2 Foreign studies

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

Factors affecting profitability in Malaysia (Alarussi, A.S. and Alhaderi, S.M, 2018). This study aims to identify profitability factors in Malaysian listed companies specifically through factors such as company size (measured by total revenue), liquidity and financial leverage. The results highlight a strong positive relationship between total revenue and profitability. The results also show a negative relationship between leverage and profitability. Also, liquidity did not show any significant relationship with profitability.

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.

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, financial leverage, capital efficiency. In addition, the study also demonstrates a close relationship between the company's performance and the way available resources are managed.

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 period 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.

1.3 Objectives of the study

Identify and evaluate factors 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 a firm. As a result, by using Machine Learning applications to guide the information group to the aim of searching for elements impacting business advantages. Apply Machine Learning to research in order to be able 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 in order to be relevant and appropriate to the Vietnamese economy and give important information to internal and foreign investors.

1.4 Research subjects

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).

1.5 Research scope

Scope of space: 512 enterprises listed on Vietnam stock exchange.

Time scope: 11-year research period from 2010 to 2020.

1.6 Research Methods

Methods of data collection, qualitative methods: analyzing and integrating elements impacting the enterprise's profit, and seeking for new factors.

Description of variables: dependent variable and independent variable.

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: to evaluate the study's efficacy and usefulness in terms of reducing observation and research time with the acquired data collection.

1.7 Expected results

If effectively applied and investigated, the study will give firms with more accurate profit estimates. At the same time, the study examines related uncertainties and factors affecting the profits of publicly traded companies, such as economic figures, enterprises' subjective factors, which assists managers and investors in making sound investment decisions, hence reducing company risks.

1.8 New point of research

The analysis in two directions is a new element of the research model that will assist the analyst understand the nature of his model and generate solid forecasts in the direction of industry analysis (specifically, the profit forecast of each enterprise in the industry) or annual analysis (profit forecasts for all publicly traded firms on the stock exchange each year) to aid in making better investment selections. Specifically:

- Industry analysis aids in determining which industries are best projected for the model, as well as which elements, such as macro variables, have a significant impact on each industry and the market economy, and how much influence each type of industry has.
-

- Year-to-year 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.

CONTENT

CHAPTER 1: THEORETICAL BASIS AND REALITY OF THESIS

1.1. Theoretical basis

1.1.1. Machine learning

Machine Learning, according to Arthur Samuel, is "the branch of research that makes computers capable of learning without being explicitly programmed."

Machine Learning, as defined by Tom Mitchell, is "a computer program that learns from experience E to do task T , and its effectiveness is evaluated by P , if its efficacy in executing task T is measured by performance P , enhanced by experience E ."

In this study, $T E P$ is defined as:

- Task T is to determine whether the business is profitable or not.
- Experience E is the characteristic to classify businesses that profit from available data.
- The performance P metric is the accuracy of the determination process.

Machine learning is the activity of learning from data in an iterative fashion using various algorithms to develop models and predict outcomes. 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.

Popular machine learning algorithms such as: Artificial Neural Networks - ANN, Support Vector Machines - SVM, Genetic Programming - GPN, K-nearest neighbors (KNN), Logistic regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Tree, Random Forest...

1.1.2. Reasons for choosing Machine learning

Econometrics is mainly concerned with model interpretation, which is the evaluation of the effects of independent variables on a large number of dependent variables in an econometric model. Machine learning is exclusively concerned with the model's prediction results; ML models attempt and fail to find the best accurate prediction. Machine learning is better suited to prediction than econometrics.

1.1.3. Supervised learning

Supervised learning is a method of Machine Learning that uses labeled data with the main goal of determining the relationship between input and output variables. Supervised learning is divided into two problems: regression and classification. Regression is used to predict continuous data, while classification is used for discrete data.

1.1.4. Random forest

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

The random forest classification algorithm is an enhanced version of the decision tree classification technique. This algorithm, which is built from several decision trees, aids in overcoming the problem of overfitting. To pick trees in the forest, the algorithm utilizes a voting mechanism.

There are two ways to vote on a random forest. One is to choose the decision tree with the highest number of votes. The second is to select the results based on the proportion of votes, these votes are the weight of the results.

Random forest generates random trees by: bootstrapping and criteria selection.

- Bootstrapping technique: Each tree is created with a unique data set that is made up of a subset of the same size of the available data.

- Criteria: Decision trees will choose features for the tree to branch.

The random forest can indicate the importance of the model's features. It helps us to know the most important features and the unimportant features. This helps to focus more on the essentials and can be considered to remove low-impact features.

1.1.5. ROC Curve

ROC Curve stands for Receiver Operating Characteristics Curve, which is a metric to evaluate the performance of a classification model. Each point along the ROC Curve corresponds to a classification model. The ROC curve shows the sensitivity of the classification model to the ratio between true positive and false negative. The ROC that compares two operating characteristics as the criteria change is called a relative operating characteristic curve. These two performance characteristics are True Positive Rate (TPR) and False Positive Rate (FPR). An ROC with a ratio between True Positive and False Positive corresponding to 1:0 is considered ideal.

Important points in ROC Curve:

- TPR = 0, FPR = 1: the model predicts all cases to be negative class
- TPR = 1, FPR = 1: The model predicts all cases to be positive class
- TPR = 1, FPR = 0: Ideal model with 0 false classifications.

1.1.6. AUC

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.

1.1.7. 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.

1.1.8. Operating cash flow

Operating cash flow (OCF) is the amount of money generated from business activities of a business during a certain period. This is an important metric on a business' cash flow statement because it shows the financial health of the business. This cash flow represents a company's liquidity including its ability to pay debts and pay expenses. If $OCF > 0$, the business pays its expenses and repays its debt with the cash it receives. In contrast, with $OCF < 0$, the business must find other sources of money (such as loans, or liquidation of assets) to pay these costs. Operating cash flow helps to check the real income situation of the business, if the business has extremely high profits but the OCF is negative, it is likely that the business has used accounting tricks for financial statements.

1.2. Real state of affairs

Profit optimization is always the top concern of businesses. Therefore, methods to predict profits and revenue of enterprises need to be improved and gradually increased accuracy. Traditionally, doing fundamental analysis requires three stages. The first and most important stage is to accurately predict the future profitability of the company over a given period of time. Without accurate predictions, the calculation of the next period's intrinsic values will be inaccurate and lead to high-risk, wrong investment choices in subsequent steps, when the value market is compared to intrinsic value. Fundamental analysis requires out-of-sample forecasts of profitability in addition to accurate forecasts because these are the forecasts that investors demand in practice.

Many studies and many tests have really shown that using historical data sets, it is possible to easily predict and identify trends in corporate financial results with reasonable error from 2-8%. However, the biggest challenge for this approach is not making predictions on their own, but collecting data and building models with a large amount of variables. Just building a simple model and getting results takes hours, not months or even days. Therefore, with a huge amount of variables and data, it will take a lot of time and effort. On the other hand, when the study is completed, the results may be outdated and no longer useful and timely.

The task of predicting profits for a particular period is very important. Because if a business can generate a lot of profit with the amount of capital, policies, and projects implemented, then a business can earn more than the expected profit provided that the value can be achieved. guess. And investors will also be easier to make decisions. The profit that a company earns in a particular period of time depends on several factors such as the amount of time and money.... So to predict the profit of a company over a period of time. Specifically, we need to train a machine learning model with a dataset containing historical data about the profits generated by the company.

Monahan (2017) examined empirical research that revealed that the out-of-sample profitability estimates provided by various regression models are not more accurate than those generated by the regression model. Furthermore, Bradshaw et al. (2012) demonstrate that even financial experts' profitability estimations are no more accurate than those obtained by a random walk model. According to Monahan (2017), the

random walk model's greater performance over existing methodologies is problematic from both a theoretical and practical standpoint. Furthermore, the team views the failures of previous approaches as an impetus to employ new ways that are more optimum as well as optimal in terms of limiting the weaknesses of basic analysis.

CHAPTER 2: PREPROCESSING DATA

2.1. Data resources

The data used is a secondary data source with 2 types of micro data and macro data:

Micro data is collected from Thomson Reuter from 2009 to 2020. The data used are financial figures of 684 Vietnamese enterprises listed on 2 exchanges HOSE (Hochiminh Stock Exchange) and HNX (HaNoi Stock Exchange). These businesses operate in 10 industries such as: Utilities, Basic Materials, Industrials, Consumer Cyclical, Consumer Non-Cyclical, Energy, Healthcare, Real Estate, Technology, Academic & Educational Services.

Macro data includes 3 indicators GDP, CPI and Real interest rate of Vietnam central bank. These indicators are collected from the World Bank from 2009 to 2020.

2.2. Preprocessing data

2.2.1. Data Cleaning

To ensure that the data is not too much missing, we drop businesses listed on the stock exchange for less than 5 years. Then check the missing value of the variables with the `count_missing` function. The ratio of missing value variables from 12.46% or less, the variables with the missing value rate above 7% are Quick ratio, Cash conversion cycle, Inventory turnover ratio, ROE, FA turnover ratio, Operating margin, BEP, Expense of revenue ratio, and DSO. Use the fill missing value method to reduce the problem of missing values of these 9 variables with the `interpolate` function. The missing values in the first and last years will be filled by the `backfill` and `first fill` method. The missing values in the middle will be filled by the `regression` method. After filling data, only variable `avg inventory days` has 48 missing values and 4 companies VNF.HN, TVC.HN, VNT.HN, VNL.HM cannot be filled due to empty full and will be removed completely. Check the missing value value ratio again, there are variables with missing value. Drop all rows with at least 1 missing value. After cleaning, the remaining data sets are businesses in 9 industries except Academic & Educational Services.

2.2.2. Data transforming

We convert some unreasonable negative values into positive variables such as: Cost of revenue, Accounts receivable, Avg. Inv days, Avg. Payable days, Avg.Receivable days. Next, we 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 from Net income and Operating cash flow. The value of Target is 1 if Net income and Operating cash flow are greater than 0; and zero if Net income and Operating cash flow are less than 0. 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. 23 feature variables (independent variables) and 1 target variable (dependent variable) after being calculated are merged into 1 dataframe with 21 columns and 5676 rows.

We found 13 variables with outliers: 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 assigning a new value to the variable with:

- Highest values = Mean value + 3*Standard deviation.
- Lowest values = Mean value - 3*Standard deviation.

2.3. Variables

The study uses 24 observed variables including 23 feature variables (independent variables) and 1 target variable (dependent variable). The independent variables used are: Growth, Size, Age, Liq, Lev, PPE, Quick ratio, Inv turnover ratio, FA turnover ratio, TA turnover ratio, DSO, Capital intensity, Expense revenue ratio, Operating margin, Net profit margin, BEP, ROA, ROE, EPS, CCC, GDP, CPI, Interest rates. The target variable is Target.

2.3.1. Target variable

The target variable is created from 2 values net income and operating cash flow with the below condition:

- Net income > 0 and Operating cash flow > 0: Target = 1.
- Net income < 0 and Operating cash flow < 0: Target = 0.

2.3.2. 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 growth rate of the joint also affects the quality of profits. Specifically, Nissimi and Penman (2001), Dechow and et al (2011), Gopalan and Jayaraman (2012), Nguyen Hoang Anh (2017) said that firms with high this ratio have lower quality of profits.

2.3.3. Size

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

Size is the size of the business. Large companies are often highly profitable. This is also found in the studies of Ball and Foster (1982), Liu and et al (2017). However, Gopalan and Jayaraman (2012) and Watts and Zimmerman (1990) show a negative relationship between firm size and profitability.

2.3.4. 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

2.3.5. Liquidity ratio (Liq)

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

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.

2.3.6. Leverage ratio (Lev)

$$Lev = \frac{\text{Total liabilities}}{\text{Total asset}}$$

A highly leveraged company means a lot of debt, which makes it possible for managers to tamper with financial statements to inflate profits. Gopalan and Jayaraman (2012) suggest that financial leverage is negatively correlated with profitability. However, Barton and Waymire (2003) find profitability and financial leverage positively correlated.

2.3.7. Capital investment ratio (PPE)

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

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.

2.3.8. Quick ratio

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

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

2.3.9. Inventory turnover ratio

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

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.

2.3.10. Fixed asset turnover ratio

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

Fixed asset turnover ratio measures the efficiency of a company's fixed assets such as machinery and equipment.

2.3.11. Total asset turnover ratio

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

Total asset turnover ratio measures how effectively a company's total assets are used in generating revenue. Alarussi and Alhaderi (2018), Alghusin (2015) and Agiomirgianakis and et al (2006) show that total asset turnover has a positive effect on profitability.

2.3.12. Days sale outstanding (DSO)

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

Days sales outstanding measures the average time it takes for a company to collect money from making a sale. DSO is too high, potentially bad debt, reducing the company's profit.

2.3.13. Capital intensity

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

Capital intensity refers to the need for large amounts of investment for production of some industries. Gopalan and Jayaraman (2012) find a positive correlation between capital intensity and profitability.

2.3.14. Expense revenue ratio

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

Expense revenue ratio measures the performance between revenue and expenses.

2.3.15. Operating margin

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

Operating margin is a profitability ratio that measures operating income per dollar of sales.

2.3.16. Net profit margin

$$\text{Net profit margin} = \frac{\text{Net income}}{\text{Sales}}$$

Net profit margin is a profitability ratio that measures net income per dollar of sales.

2.3.17. Basic earning power (BEP)

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

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.

2.3.18. Return on total asset (ROA)

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

ROA measures profitability per dollar of assets. ROA can result from using a lot of debt, which will result in a fairly low net income.

2.3.19. Return on equity (ROE)

$$ROE = \frac{Net\ income}{Total\ equity}$$

ROE measures profitability on common equity. This ratio is of great interest to investors. The higher the ROE, the more efficient the company is in using shareholder capital and increasing profits, which is a great attraction for investors.

2.3.20. Earning per share (EPS)

Earning per share is the profit earned from a share. EPS is used to assess a company's profitability.

2.3.21. 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.

2.3.22. Gross Domestic Index (GDP)

The Gross Domestic Index reflects the growth rate of gross domestic product year over year. A growing economy helps to increase business efficiency, which in turn also affects profits.

2.3.23. Consumer Price Index (CPI)

Consumer Price Index is the rate of increase in the general price level of goods and services over time and the devaluation of a currency. High CPI increases the instability of production, affecting the profits of enterprises.

2.3.24. 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.

2.4. Drawing correlation matrix of variables

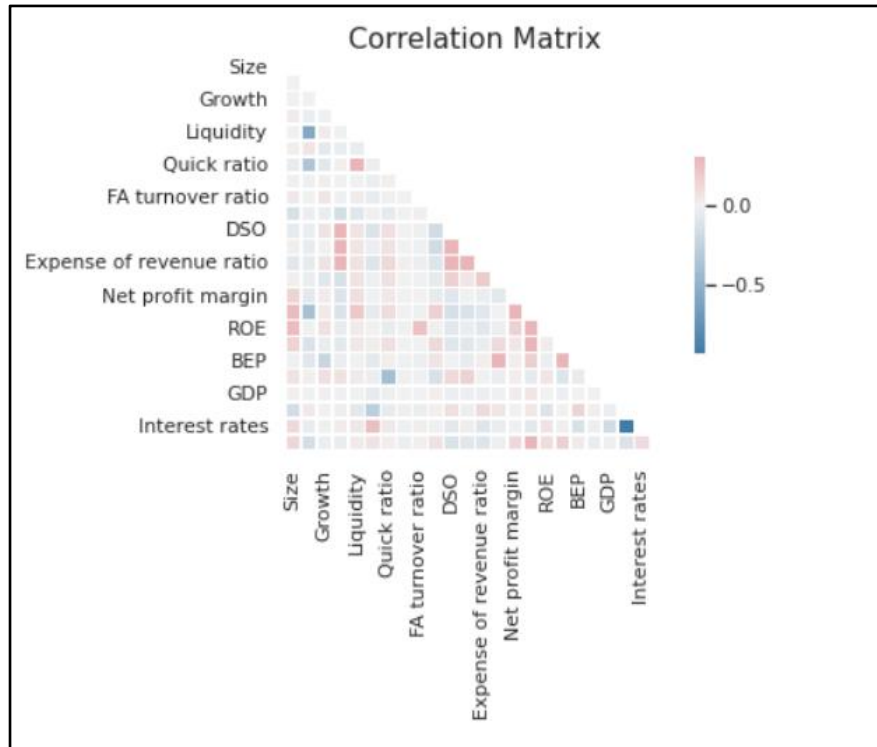


Figure 2.4. Correlation matrix of variables

CHAPTER 3: BUILDING MODELS AND VISUALIZATION OF DATA

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 80:20. 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, K-nearest neighbors, Decision above, Support vector Machine (Linear and RBF Kernel), Neural Network, and Random forest.

Algorithm	Model score
Logistic Regression	70.93%
K-Nearest Neighbors	67.93%
Decision Tree	66.43%
Support Vector Machine (Linear Kernel)	71.83%

Support Vector Machine (RBF Kernel)	71.03%
Neural Network	70.73%
Random Forest	74.53%

Table 3. Table of various models' scores in order to search for the best one

The score shows that Random forest is the best model to forecast corporate profits with 74.53%.

We perform GridSearch from 1 to 100 random forest to find the optimal number of trees. The results show that the model stopping at the 81nd tree is optimal. We train the random forest model with the 81nd optimal number of trees.

Research model:

$$\text{Target} = \beta_0 + \beta_1 \text{Growth}_{it} + \beta_2 \text{Size}_{it} + \beta_3 \text{Age}_{it} + \beta_4 \text{Liq}_{it} + \beta_5 \text{Lev}_{it} + \beta_6 \text{Ppe}_{it} + \beta_7 \text{Quick}_{it} + \beta_8 \text{Inv turnover ratio}_{it} + \beta_9 \text{FA turnover ratio}_{it} + \beta_{10} \text{TA turnover ratio}_{it} + \beta_{11} \text{DSO}_{it} + \beta_{12} \text{Capital intensity}_{it} + \beta_{13} \text{Expense revenue ratio}_{it} + \beta_{14} \text{Operating margin}_{it} + \beta_{15} \text{Net profit margin}_{it} + \beta_{16} \text{BEP}_{it} + \beta_{17} \text{ROA}_{it} + \beta_{18} \text{ROE}_{it} + \beta_{19} \text{EPS}_{it} + \beta_{20} \text{CCC}_{it} + \beta_{21} \text{GDP}_{it} + \beta_{22} \text{CPI}_{it} + \beta_{23} \text{Interest Rate}_{it} + \varepsilon_{it}$$

In there:

- Target variable: We use 2 figures net income (NI) and operating cash flow (OCF) 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, net profit margin, basic earning power (BEP), return on total asset (ROA), return on equity (ROE), earning per share (EPS), cash conversion cycle (CCC).
- Control variables: business's size variable (Size), business's Age variable (Age), sale growth variable (Growth), financial leverage (Lev), liquidity variable (Liq).
- In addition we also use macro variables: Gross Domestic Product (GDP), Consumer Price Index (CPI), Interest Real Rate (Interest Rate).

The above model includes data of businesses with industries in the dataset from 2010 - 2020. In addition, we disaggregate data by year and by industry to compare differences with variables like the model above. Specifically, for model decomposition by year, we form 11 models corresponding to 11 years from 2010 to 2020. For model decomposition by industry, we form 9 models corresponding to 9 industries: Utilities, Basic Materials, Industrials, Consumer Cyclicals, Consumer Non-Cyclical, Energy, Healthcare, Real Estate, Technology.

CHAPTER 4: RESULTS AND MODEL EVALUATION

First, we will evaluate the predictive model for the entire business. When the number of estimators is 81, the model has a relatively decent accuracy index (0.71), as shown in Figure **.

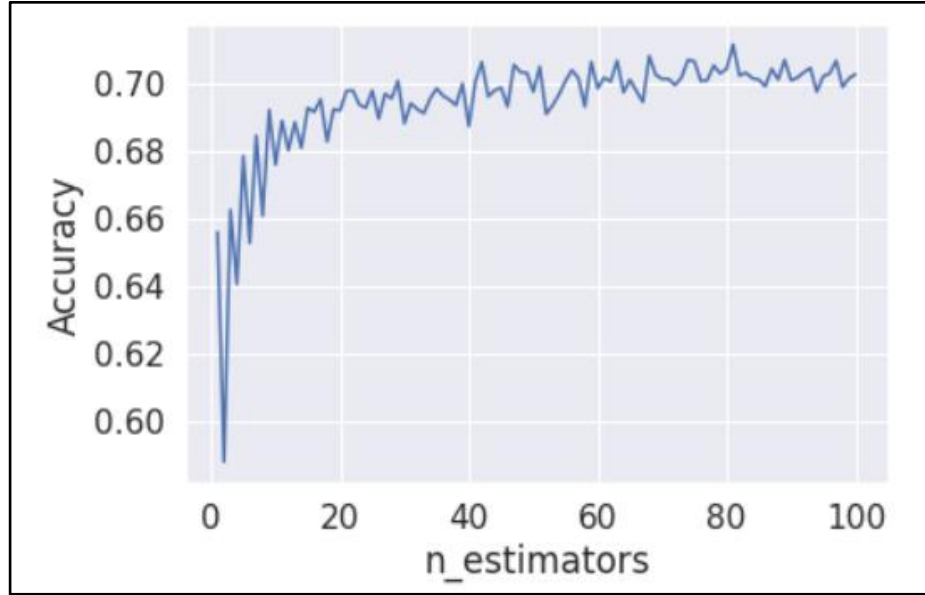


Figure 4.1. Chart of accuracy line with the various number of $n_estimators$ in the total industries model

Also, the ROC-AUC findings of this model are pretty good, as shown in Figure **. The 5-fold AUC values vary from 0.75-0.81, showing that the model is doing well.

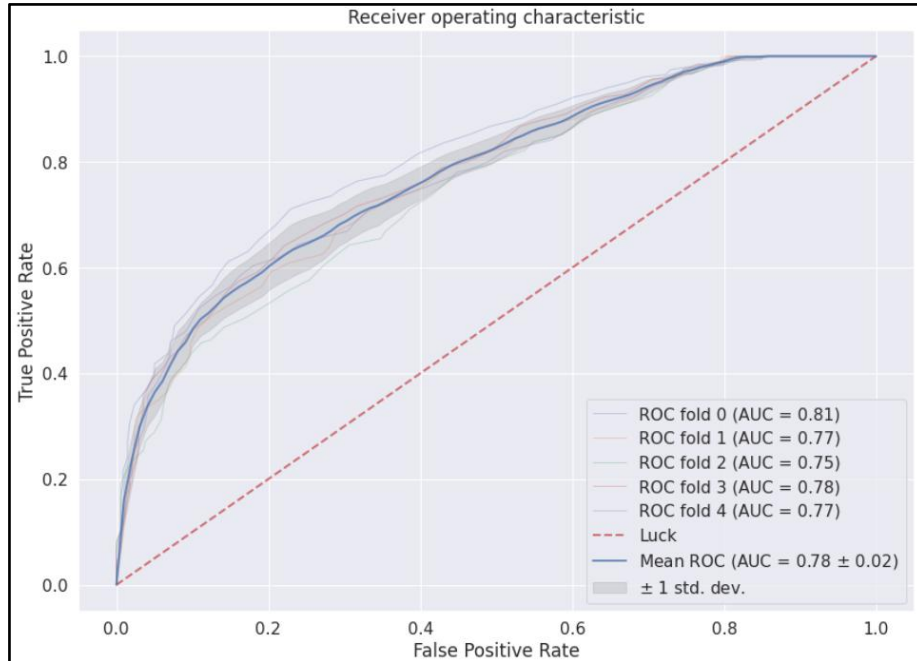


Figure 4.2. Chart of ROC curve in the total industries model

Furthermore, We discovered that the Precision of the Profit group and the Recall of the Loss group were fairly good based on the indications in the categorization report (Figure 4.3). Demonstrates the model's ability to anticipate the accuracy of profit estimates for the market as a whole. In addition, the capacity to foresee nearly entire business losses in the market. However, the Profit group's Recall index is relatively

low, indicating that the model overlooks many investment possibilities from organizations that are actually lucrative in the market but are expected to lose money.

Formulae for Classification report:

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

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

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

	precision	recall	f1-score	support
0.0	0.37	0.94	0.54	338
1.0	0.87	0.20	0.32	663
accuracy			0.45	1001
macro avg	0.62	0.57	0.43	1001
weighted avg	0.70	0.45	0.40	1001

Figure 4.3. Classification report of the total industries model

4.1 Decomposition of the years

Phân rã theo năm	Ob s	Training set							Test set						
		Train- set obs	Accura cy	AUC score	Precision		Recall		Test- set obs	Accura cy	AUC score	Precision		Recall	
					Profit	Loss	Profit	Loss				Profit	Loss	Profit	Loss
2010	393	314	61%	100%	61%	0%	99%	0%	79	48%	76%	48%	0%	100%	0%
2011	430	344	46%	100%	83%	41%	16%	95%	86	53%	71%	75%	51%	14%	95%
2012	445	356	55%	100%	89%	43%	37%	91%	89	52%	81%	96%	32%	38%	95%
2013	459	367	65%	100%	80%	51%	60%	73%	92	60%	72%	79%	44%	54%	71%
2014	470	376	34%	100%	100%	32%	3%	100%	94	35%	71%	100%	34%	2%	100%
2015	504	403	54%	100%	89%	40%	38%	90%	101	59%	72%	88%	45%	45%	88%
2016	490	392	58%	100%	79%	43%	49%	75%	98	58%	76%	84%	42%	48%	81%
2017	476	380	57%	100%	87%	45%	39%	90%	96	57%	80%	91%	47%	35%	94%
2018	454	363	49%	100%	81%	37%	32%	85%	91	46%	70%	86%	33%	29%	88%

2019	447	357	33%	100%	84%	28%	10%	95%	90	40%	73%	78%	36%	12%	94%
2020	434	347	42%	100%	89%	33%	20%	94%	87	52%	81%	100%	41%	28%	100%

Table 4.1.1. Table of features' importance of the model decomposed by years

When the number of estimators is 55, the algorithm achieves optimal accuracy (0.75), as illustrated in Figure **. In general, the model is of high quality. The accuracy value, on the other hand, merely measures the model prediction performance in general.

When we use the aforementioned accuracy, it just shows us what percentage of the data is correctly classified, not whether the classification results match the real value, if there is a mismatch or not? To be able to evaluate the model's performance in more detail, the team is considering adding a confusion matrix, as illustrated in Figure 4.1.1.

From the 4 indicators of the confusion matrix (TP, TN, FP, FN), we have a classification report to evaluate the reliability of the model.

	precision	recall	f1-score	support
0.0	0.39	1.00	0.56	29
1.0	1.00	0.21	0.34	58
accuracy			0.47	87
macro avg	0.69	0.60	0.45	87
weighted avg	0.80	0.47	0.41	87

Figure 4.1.1. Classification report of the model decomposed by years

The forecast model gives good results with a high precision value of 85% and $y = 1$ indicates that the accuracy of the points obtained is high, indicating that the rate of selecting the proper firm provides high profits of about 85%. However, with a recall rate of 43%, we can see that we have relatively zoned out for fear of risk, resulting in missing out on many really profitable companies.

Our testing sample ranges from 2010 to 2020. In addition to presenting the model's overall AUC, the team also presents the AUC score for each year, which is displayed in Table 4.1.1. In general, the AUC ratings for each year are in the 70-81% range

(highest in 2012 and

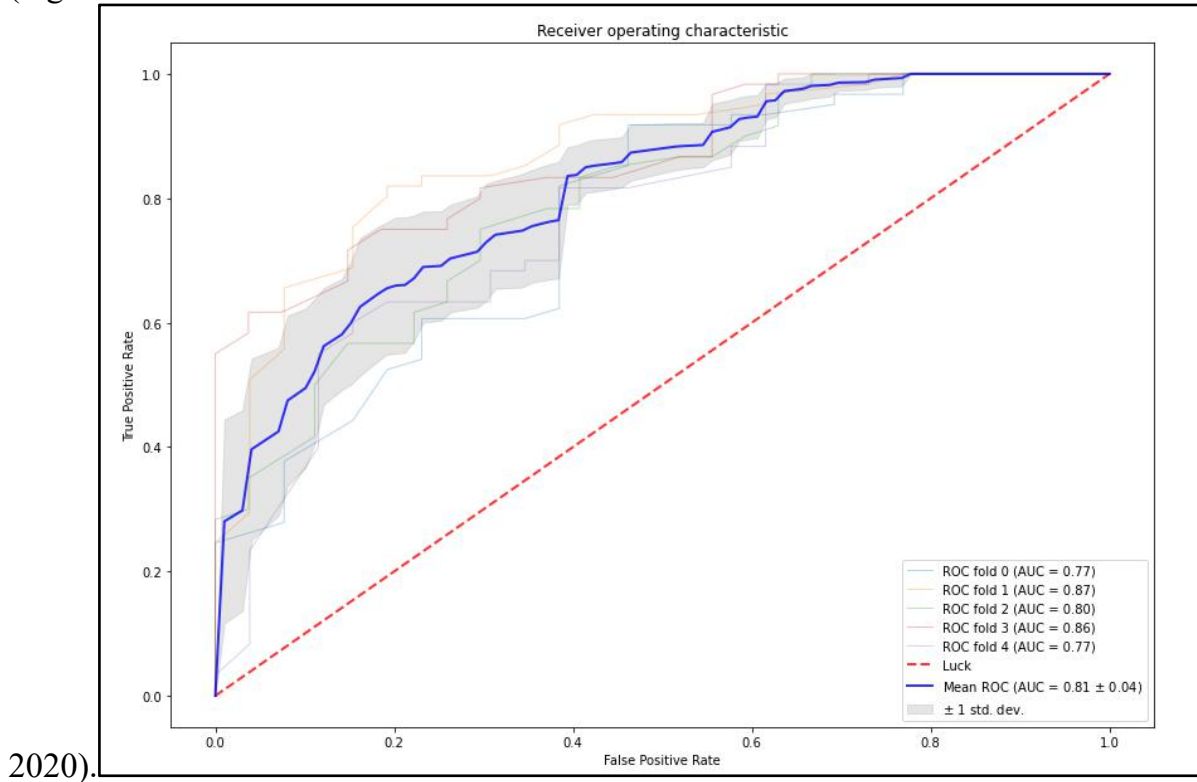


Figure 4.1.2. Chart of ROC curve of the year 2020 model

With the year 2020 model, we evaluated the model's output quality using ROC utilizing cross-validation and obtained the findings given in Figure 4.1.2.

ROC curves normally have a Y axis for true positive rate and an X axis for false positive rate. This suggests that the "ideal" point is in the plot's upper left corner, with a false positive rate of zero and a real positive rate of one. This is far from practical, but it does imply that a greater area under the curve (AUC) is typically preferable.

It displays the ROC response of several datasets created by 5-fold cross validation. Using all of these curves, the mean area under the curve can be calculated and the variance of the curve can be shown when the training set is divided into different subsets. This approximately depicts how changes in the training data impact the classifier's output, as well as the difference between the splits created by the 5-fold cross-validation versus each other.

The 5-fold AUC scores range from 0.77-0.87, indicating that the model's performance is stable.

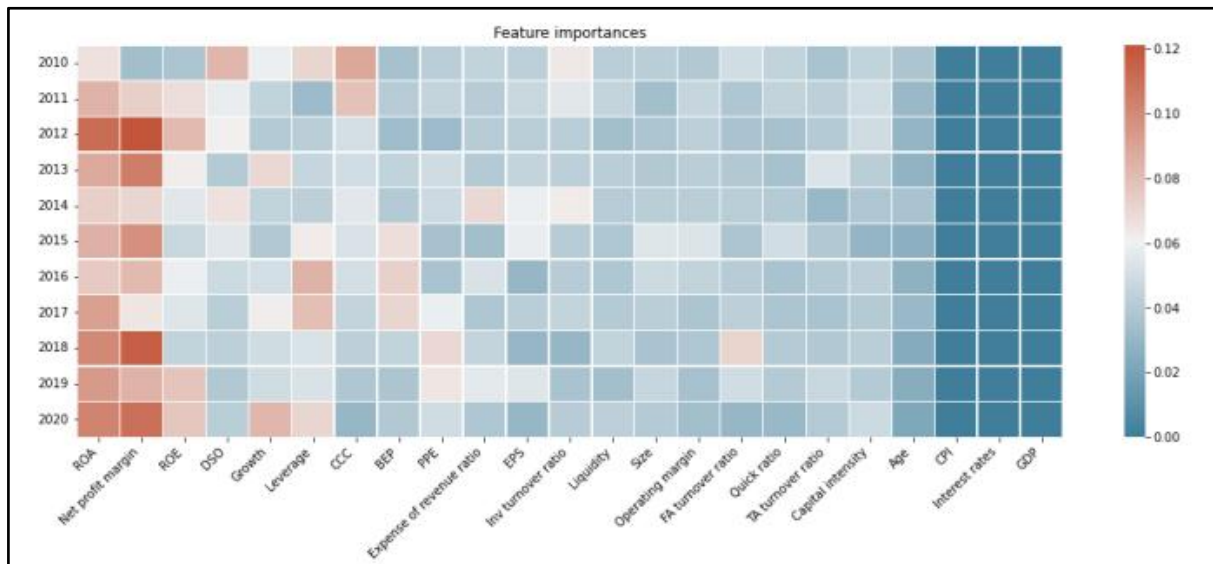


Figure 4.1.3. Heatmap of features' importance in the model decomposed by years.

We next examined the model's feature importances. Figure 4.1.3 shows that, in general, the impact ratio of the factors does not change much over time. ROA, Net profit margin, and ROE are more essential than other factors and grow year after year from 2010 to 2020 (particularly, in the Appendix 5, ROA: 0.067-0.103, Net profit margin: 0.034-0.11, ROE: 0.037-0.77). This demonstrates the need of paying more attention to these values, as changes in them in the future might have an impact on the profitability of businesses. Furthermore, it can be observed that macro variables have no effect on the analysis over time. The remaining factors have an influence but are not significant.

4.2 Decomposition of the industries

Decomposition by industries	Obs	Training set							Test set						
		Train-set obs	Accuracy	AUC score	Precision		Recall		Test-set obs	Accuracy	AUC score	Precision		Recall	
					Profit	Loss	Profit	Loss				Profit	Loss	Profit	Loss
All industries	5002	4001	45%	100%	85%	37%	20%	93%	1001	46%	78%	86%	38%	22%	93%
Basic materials	841	672	37%	100%	94%	32%	11%	99%	169	50%	80%	94%	40%	26%	96%
Consumer cyclicals	769	615	33%	100%	94%	30%	7%	99%	154	36%	76%	100%	33%	6%	100%
Consumer non-cyclicals	478	382	35%	100%	100%	34%	2%	100%	96	46%	77%	100%	43%	9%	100%
Energy	375	300	50%	100%	86%	27%	43%	75%	75	52%	64%	90%	25%	46%	79%
Health care	161	128	50%	100%	92%	32%	36%	91%	33	76%	89%	95%	46%	73%	86%
Industrials	1461	1168	48%	100%	78%	43%	18%	92%	293	48%	73%	85%	41%	21%	94%

Real estate	534	427	55%	100%	57%	54%	24%	83%	107	43%	63%	57%	39%	20%	77%
Technology	166	132	39%	100%	100%	36%	6%	100%	34	65%	66%	100%	60%	25%	100%
Utilities	217	173	85%	100%	92%	13%	92%	13%	44	25%	83%	83%	16%	14%	86%

Table 4.2.1. Table of features' importance of the model decomposed by industries

According to Table 4.2.1, the AUC of each industry varies greatly. In comparison to all industries, the AUC in real estate, energy, and technology is quite low (lower than 12-15 percent). Besides, the health-care business has the greatest AUC index (89% - 11 percent higher than the all industries model). From there, the Health care sector forecasting model may be rated as extremely good in terms of predicting lucrative firms in this discipline.

	precision	recall	f1-score	support
0.0	0.46	0.86	0.60	7
1.0	0.95	0.73	0.83	26
accuracy			0.76	33
macro avg	0.71	0.79	0.71	33
weighted avg	0.85	0.76	0.78	33

Figure 4.2.2. Classification report of the Healthcare industry model

The Healthcare industry forecasting model has an accuracy index of 0.95, indicating a high ability to accurately forecast firms in this industry, and a recall index of 0.73, indicating a strong ability to properly anticipate companies in this sector. The likelihood of losing out on truly lucrative businesses is minimal. The healthcare business has a robust data collection and few missing numbers, making it a stable growth and low volatility market with few outliers. Furthermore, the nature of the healthcare business is closely regulated by the government and market processes in order to mitigate detrimental effects on community services. As a result, the model is rated as the best and most stable in comparison to the other industries, with a very high AUC value (0.78-0.92, Figure 4.2.3) and a high accuracy of 0.76.

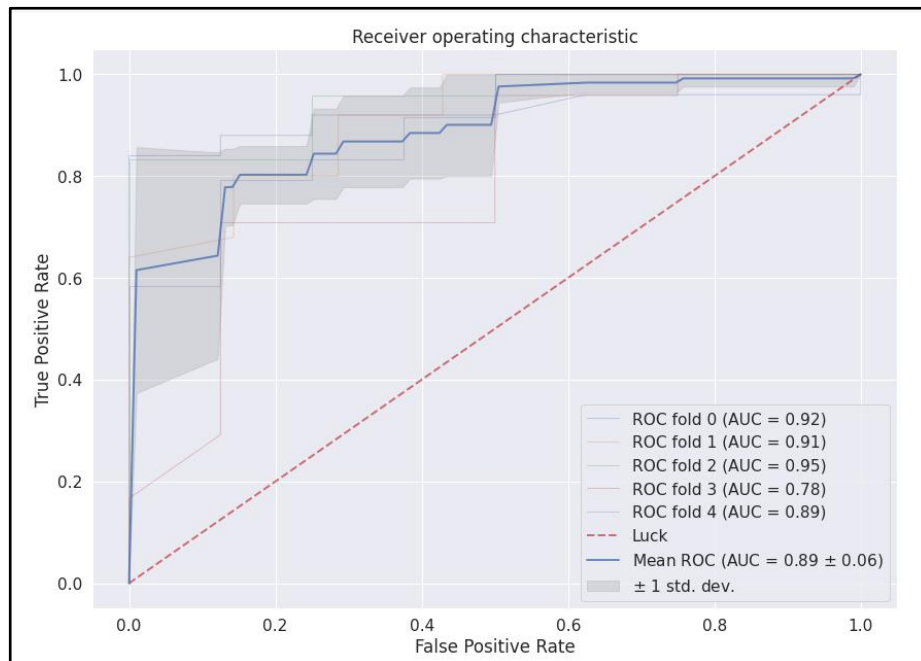


Figure 4.2.3. Chart of ROC curve of the Healthcare industry model

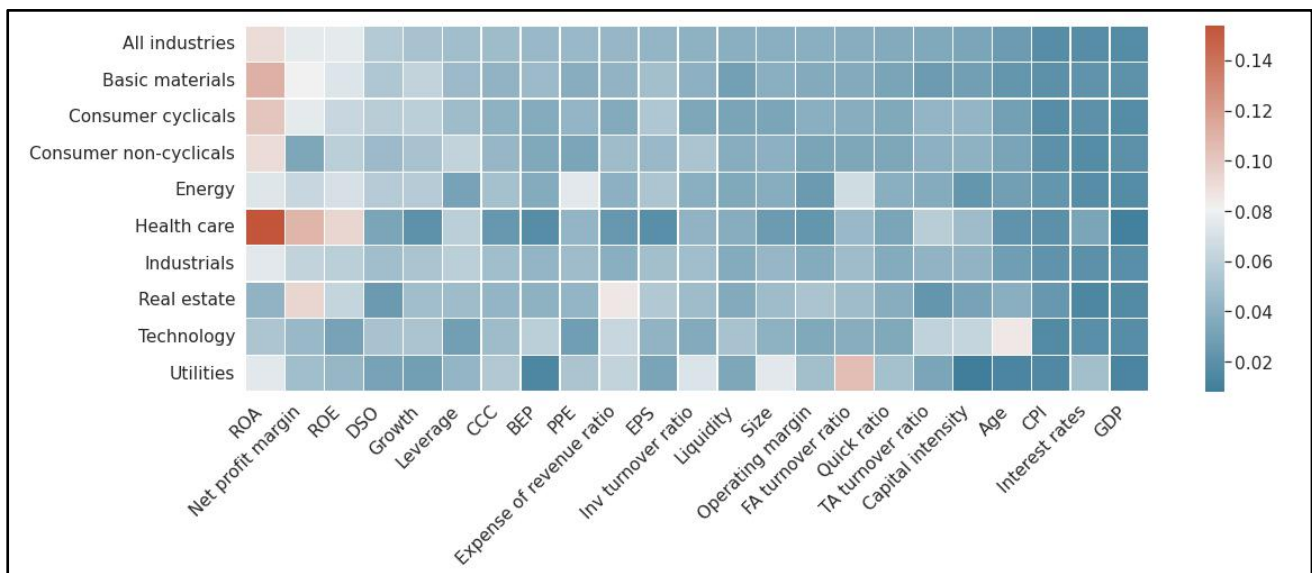


Figure 4.2.4. Heatmap of features' importance in the model decomposed by industries

The influence of the factors may be seen in Figure 4.243. In general, characteristics are classified into three classes based on their effect degree (2-14 percent). The most essential set of characteristics are the fundamental metrics used to assess the company's success, such as ROA, Net profit margin, ROE, and DSO. Other financial indicators such as Growth, Leverage, CCC, EPS... are the second most essential category of attributes. The remaining characteristics are macro variables. Although the level of impact of macro variables is low, it does show the influence of variables when classifying by industry; when looking at the year, all companies are subject to the same systematic risk, so the annual forecast model does not show the influence of macro factors, which is a significant difference when compared to the model decomposed by industries.

CHAPTER 5: DISCUSSION AND CONCLUSION

5.1 Conclusion

We investigate whether a machine learning (ML) approach, random forests can provide out-of-sample profitability forecasts that outperform classical forecasts. We are motivated to use the Machine Learning method because a) the literature and training show that traditional regression methods cannot produce out-of-sample forecasts that are superior to random walk forecasts, and Logistic Regression, K-Nearest Neighbors, Decision Tree, Support Vector Machine (Linear Kernel), Support Vector Machine (RBF Kernel), Neural Network, and, in particular, Random Forest. b) Because they are focused on prediction and optimizing prediction accuracy, ML approaches have certain benefits over regression methods in generating out-of-sample predictions.

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 2010-2021 and generate out-of-sample predictions of directional changes (increases or decreases) in more than twenty variables profitability measures: quick ratio, cash conversion cycle (CCC), inventory turnover ratio, return on equity (ROE), fix asset turnover ratio (FA turnover ratio), operating margin, basic earning power (BEP), expense of revenue ratio, days sales outstanding (DSO), size, net profit margin, return on assets (ROA), growth, earnings per share (EPS), Property, Plant, & Equipment Turnover Ratio (PPE), total asset turnover ratio (TA turnover ratio), capital intensity, leverage, liquidity, age , gross domestic product (GDP), consumer price index (CPI), interest real rates (interest rates), target. Results based on a minimum set of independent variables show that our ML method achieves classification accuracies ranging from 47% to 76% for our profitability measures, and that the differences in proportions of accurate classifications between ML and random walk are highly significant.

After obtaining the study's findings by sector and year, we can see that, by industry, AUC on the health care test set is the best at roughly 89 percent, while real estate is the worst at about 63 percent. Otherwise, when analyzing the model by year, the most accurate years are 2012 and 2020, with accuracies of roughly 81 percent. The efficacy of the data collection and the reality of each industry must be the cause of these results.

As we can see, the Vietnamese health care industry has attracted a big number of investors while also creating a high level of income and accuracy. The profit forecasting model of the health care industry is the best because of the good data set, few missing values, the industry's steady growth, little volatility, few outliers, the nature of the Health care industry is tightly controlled by the state and the market mechanism. schools to combat negative impacts on health care and public health services, that might help with the evaluation process. Because medical equipment is a significant component in deciding the efficiency and quality of medical work, it actively assists doctors in properly diagnosing and treating patients. It is quick, safe, and effective. Today, medical equipment not only helps prolong the senses, but also allows doctors to easily access and treat injuries inside the body, such as: laparoscopic

surgery, surgery, robotics... can even replace the human brain (using artificial intelligence) to help make the most informed, correct, and effective decisions in diagnosing, treating, and caring for people's health sick. However, the data set of real estate not only contains more missing values, but it also includes primarily short-term loans, which result in high finance leverage and, as a result, worse accuracy and low AUC score.

Furthermore, we judge that the results obtained from the study by the year are not as good as by the industries. AUC scores, which get from model evaluation by year, are changing around 71% and 81% and have no significant change over year and its accuracy, which ranges between 35% and 60%, is quite low when compared to the model by industries. In addition, the macro features' importance in the year model equal zero over years. This is because these variables affects in the same way on companies and industries and the systematic risk in that year is the same to all business, so the difference cannot be seen and the forecast is accurate. However, for different people and different purposes, we may choose different ways of modeling. As researchers require high accuracy prediction for investment and taking risk, evaluation model by fields is more suitable with higher AUC score and higher accuracy.

5.2 Discussion and future study

Because of the large number of decision trees involved in the test, random forests are regarded as an accurate and powerful approach. It does not have any overfitting issues. The fundamental reason for this is that it takes the average of all forecasts, canceling out the bias. The approach is applicable to our research. Missing values can also be handled using random forests. We use an effective approach to dealing with these numbers: utilizing linear regression method to replace continuous variables missing values. That may obtain relative feature importance, which aids in determining which characteristics contribute the most to the classifier.

However, this method has some limitations. Since there are so many decision trees in random forests, it takes a long time to make predictions. Every time it makes a forecast, all of the trees in the forest must make a prediction for the same supplied input and then vote on it. This entire procedure takes time. Models are more perplexing than decision trees, which allow you to readily make judgments by following a path through the tree. In addition, the selected data has some unreasonable points such as net income (NI) is higher than earnings before interest and taxes (EBIT) and higher than sales; inventory is higher than current assets. This should be improved in the next extended researches.

Additionally, our Machine Learning model has issues with anticipated loss companies, and there is a precision-to-recall mismatch. The results reveal that the model is unable to forecast which companies would lose money, but it can anticipate which companies will benefit in the future. This needs to be improved in the future.

We feel that these results can be further improved, especially if we are willing to sacrifice model interpretability. For example, we could add more features, employ different ML methods, explode the feature space to properly handle categorical variables, and expand the set of hyperparameters over which we searched. Future work can also explore the use of accruals as features in the prediction of cash flows. Finally, our 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.

REFERENCES

- 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.
- Alarussi, A. S., & Alhaderi, S. M. (2018). *Factors affecting profitability in Malaysia*. *Journal of Economic Studies*, 45(3), 442-458.
- Alarussi, A.S, & Alhaderi, S.M. (2018). *Factors affecting profitability in Malaysia*.
- Anand, V., Brunner, R., Ikegwu, K., & Sougiannis, T. (2019). *Predicting Profitability Using Machine Learning*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3466478
- Anand, V., Brunner, R., Ikegwu, K., & Sougiannis, T. (2019). *Predicting profitability using machine learning*.
- Ball, R., & Foster, G. (1982). *Corporate Financial Reporting: A Methodological Review of Empirical Research*. *Journal of Accounting Research*, 20, 161–234.
- Barton, J., & Waymire, G. (2004). *Investor protection under unregulated financial reporting*. *Journal of Accounting and Economics*, 38(1), 65–116.
- Breiman, L. (2001). *Machine Learning*, 45(1), 5-32.
- Brigham, E., & Houston, J. *Fundamentals of financial management*.
- Burja, C. (2011). Factors Influencing the Companies' Profitability.
- Cerf, A. R. (1961). *Corporate reporting and investment decisions*. Berkeley, University of California Press.
- Chander, S., & Aggarwal, P. (2008). *Determinants of corporate profitability: An empirical study of Indian drugs and pharmaceutical industry*. *Paradigm*, 12(2), 51-61.
- Gopalan, R., & Jayaraman, S. (2012). *Private Control Benefits and Earnings Management : Evidence from Insider Controlled Firms*. *Journal of Accounting Research*, 50(1), 117–157.
- 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/>
- Lewinson, E. (2020). *Python for Finance Cookbook*. Birmingham, UK: Packt Publishing Ltd.
- 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
-

- Nouri, Y. (2018). *Random Forest & K-Fold Cross Validation*. Retrieved from <https://www.kaggle.com/code/ynouri/random-forest-k-fold-cross-validation/notebook>
- sklearn.ensemble.RandomForestClassifier*. Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- sklearn.model_selection.RandomizedSearchCV*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
- Vu, H. T. (2018). *Basic Machine Learning*. Vietnam: Vietnamese Education Publisher.
- Watts, R. L., & Zimmerman, J. L. (1990). *Positive Accounting Theory: A Ten Year Perspective*. *The Accounting Review*, 65(1), 131–156.
- Xinyue, C., Zhaoyu, X., & Yue, Z. (2020). *Using Machine Learning to Forecast Future Earnings*.
-

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
Operating Cash Flow	"OCF" sheet
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
Earning Per Share	"EPS" 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	$\frac{\text{Sale}_t - \text{Sale}_{t-1}}{\text{Sale}_{t-1}}$
Market Size Of Company	Size	ln(Market value)
Age Of Company	Age	2020 - Founding year
Liquidity Ratio	Liq	$\frac{\text{Total current asset}}{\text{Current liabilities}}$
Leverage Ratio	Lev	$\frac{\text{Total liabilities}}{\text{Total asset}}$
Property, Plant, & Equipment Turnover Ratio	PPE	$\frac{\text{Tangible Fixed Asset}}{\text{Sales}}$
Quick Ratio	Quick ratio	$\frac{\text{Total current asset} - \text{Inventory}}{\text{Current liabilities}}$
Inventory Turnover Ratio	Inventory turnover ratio	$\frac{365}{\text{Average inventory days}}$
Fixed Asset Turnover Ratio	FA turnover ratio	$\frac{\text{Sales}}{\text{Fixed assets}}$
Total Asset Turnover Ratio	TA turnover ratio	$\frac{\text{Sales}}{\text{Total asset}}$
Days Sale Outstanding	DSO	$\frac{\text{Account receivable}}{\text{Sale}/365}$
Capital intensity	Capital intensity	$\frac{\text{Total asset}}{\text{Sales}}$
Expense Of Revenue Ratio	Expense of revenue ratio	$\frac{\text{Costs of revenue}}{\text{Sales}}$
Operating margin	Operating margin	$\frac{\text{EBIT}}{\text{Sales}}$

Net profit margin	Net profit margin	$\frac{\text{Net income}}{\text{Sales}}$
Return On Total Asset	ROA	$\frac{\text{Net income}}{\text{Total asset}}$
Return On Equity	ROE	$\frac{\text{Net income}}{\text{Total equity}}$
Earnings Per Share	EPS	None
Basic Earning Power Ratio	BEP	$\frac{\text{EBIT}}{\text{Total asset}}$
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
Net Income	NI
Operating Cash Flow	OCF

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

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
ROA	0.067	0.085	0.112	0.088	0.074	0.086	0.076	0.092	0.101	0.094	0.103
Net profit margin	0.034	0.074	0.121	0.105	0.071	0.098	0.082	0.065	0.117	0.085	0.11
ROE	0.037	0.068	0.082	0.062	0.055	0.047	0.059	0.054	0.044	0.078	0.077
DSO	0.083	0.058	0.061	0.04	0.067	0.055	0.048	0.042	0.043	0.039	0.042
Growth	0.059	0.044	0.04	0.07	0.044	0.039	0.051	0.062	0.049	0.049	0.083
Leverage	0.071	0.032	0.042	0.046	0.043	0.063	0.085	0.08	0.052	0.052	0.071
CCC	0.089	0.079	0.051	0.05	0.055	0.052	0.051	0.045	0.043	0.038	0.03

BEP	0.035	0.041	0.033	0.044	0.04	0.068	0.073	0.071	0.044	0.037	0.039
PPE	0.042	0.045	0.032	0.049	0.048	0.035	0.036	0.059	0.07	0.065	0.049
Expense of revenue ratio	0.044	0.041	0.041	0.04	0.07	0.034	0.052	0.038	0.045	0.056	0.038
EPS	0.043	0.047	0.042	0.045	0.059	0.058	0.030	0.042	0.03	0.054	0.03
Inv turnover ratio	0.064	0.055	0.042	0.043	0.063	0.041	0.041	0.045	0.03	0.036	0.041
Liquidity	0.042	0.045	0.034	0.042	0.041	0.038	0.038	0.04	0.044	0.034	0.043
Size	0.042	0.034	0.037	0.039	0.042	0.054	0.048	0.042	0.036	0.046	0.04
Operating margin	0.039	0.046	0.043	0.042	0.042	0.053	0.044	0.037	0.038	0.035	0.034
FA turnover ratio	0.05	0.038	0.036	0.039	0.042	0.037	0.041	0.043	0.072	0.05	0.03
Quick ratio	0.044	0.044	0.035	0.035	0.04	0.05	0.036	0.037	0.04	0.04	0.031
TA turnover ratio	0.036	0.043	0.04	0.053	0.031	0.039	0.04	0.036	0.039	0.047	0.04
Capital intensity	0.044	0.05	0.049	0.042	0.038	0.029	0.043	0.04	0.042	0.04	0.048
Age	0.037	0.031	0.029	0.028	0.036	0.026	0.027	0.031	0.024	0.025	0.022
CPI	0	0	0	0	0	0	0	0	0	0	0
Interest rates	0	0	0	0	0	0	0	0	0	0	0
GDP	0	0	0	0	0	0	0	0	0	0	0

Appendix 6. 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 rates	0.018	0.022	0.020	0.017	0.018	0.033	0.02	0.014	0.019	0.049
GDP	0.017	0.021	0.017	0.02	0.017	0.01	0.019	0.016	0.018	0.013

Appendix 7. Table of percent missing values occupied in each variable for the total industries model

	(%)
Quick ratio	12.455955
CCC	10.095137
Inventory turnover ratio	9.355180
ROE	9.249471
FA turnover ratio	8.192389
Operating margin	7.998591
BEP	7.892882
Expense of revenue ratio	7.716702
DSO	7.646230
Size	5.761099
Net profit margin	5.056378
ROA	4.844961
Growth	4.474982
EPS	4.439746
PPE	3.488372
TA turnover ratio	3.312192
Capital intensity	3.312192
Leverage	3.100775
Liquidity	3.100775
Age	0.088090
GDP	0.000000
CPI	0.000000
Interest rates	0.000000
Target	0.000000

Appendix 8. Tables of features' importance in the total industries model (1), the two models decomposed by industries (Healthcare industry) (2), decomposed by years (year 2020) (3)

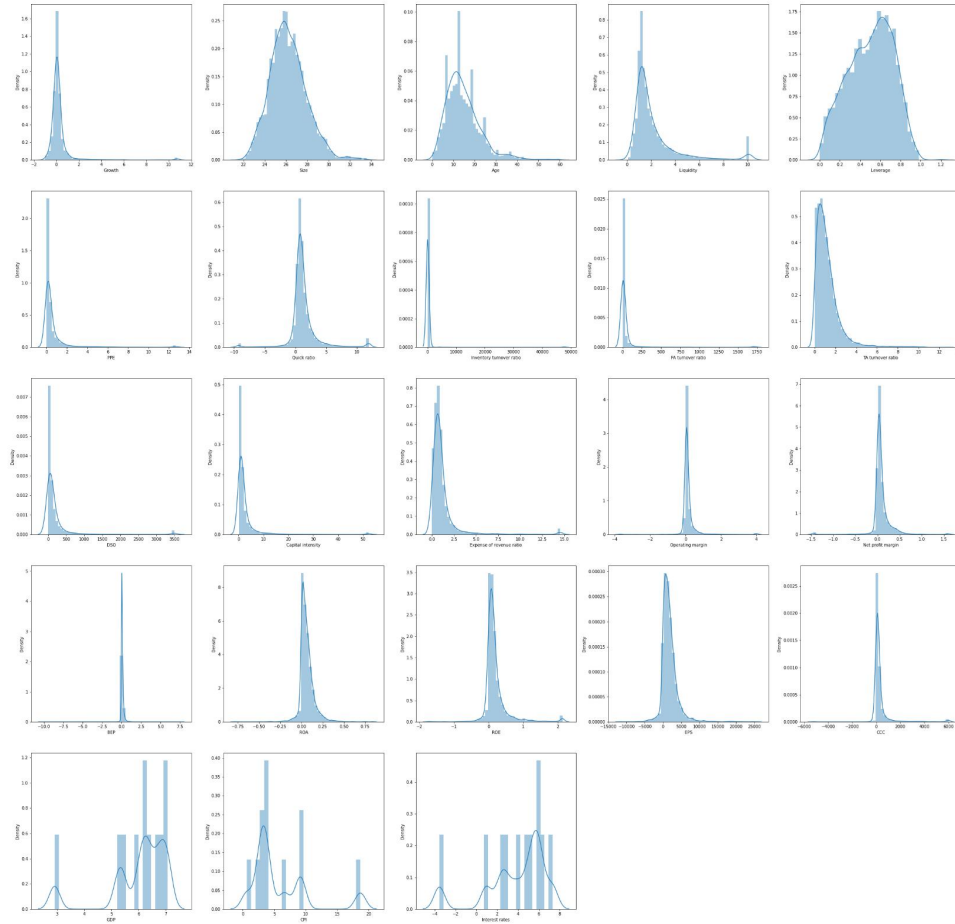
(1)

(2)

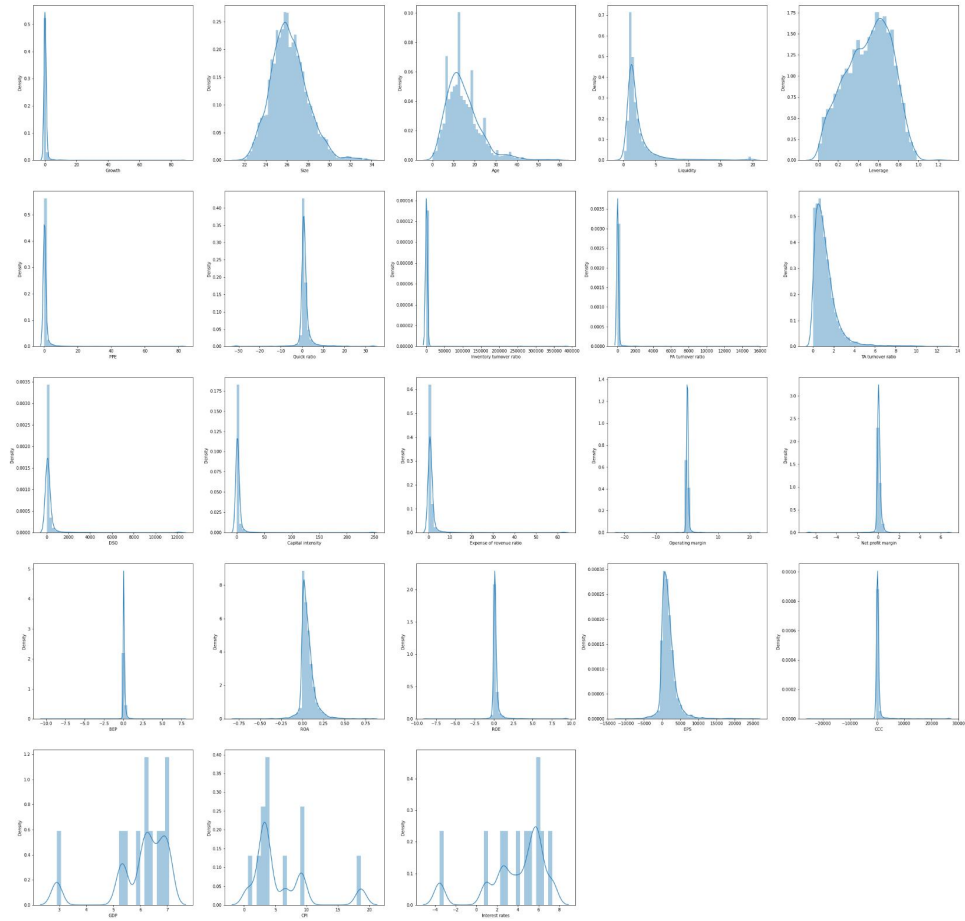
(3)

(%)		(%)		(%)	
ROA	0.089889	ROA	0.152503	Net profit margin	0.109058
Net profit margin	0.077182	Net profit margin	0.107874	ROA	0.098060
ROE	0.076106	ROE	0.092146	Growth	0.083994
DSO	0.056026	TA turnover ratio	0.060019	ROE	0.081257
Growth	0.050740	Leverage	0.058353	Leverage	0.071470
Leverage	0.048209	FA turnover ratio	0.051188	PPE	0.048940
CCC	0.047006	Capital intensity	0.047730	Capital intensity	0.048569
BEP	0.044831	PPE	0.043960	BEP	0.041972
PPE	0.044354	Inventory turnover ratio	0.041703	TA turnover ratio	0.041967
Expense of revenue ratio	0.043395	Liquidity	0.037320	Expense of revenue ratio	0.041538
EPS	0.043180	Interest rates	0.033550	Inventory turnover ratio	0.041219
Inventory turnover ratio	0.040494	DSO	0.032460	DSO	0.040435
Liquidity	0.039150	Quick ratio	0.032281	Liquidity	0.039447
Size	0.039076	Size	0.027716	Size	0.037474
Operating margin	0.038699	Expense of revenue ratio	0.025562	Operating margin	0.035791
FA turnover ratio	0.038274	CCC	0.023973	FA turnover ratio	0.032517
Quick ratio	0.036344	Operating margin	0.023077	CCC	0.030755
TA turnover ratio	0.035160	Growth	0.021283	EPS	0.028522
Capital intensity	0.032499	Age	0.021095	Quick ratio	0.027636
Age	0.026629	CPI	0.019816	Age	0.019378
CPI	0.018445	EPS	0.018863	GDP	0.000000
Interest rates	0.017795	BEP	0.017713	CPI	0.000000
GDP	0.016520	GDP	0.009817	Interest rates	0.000000

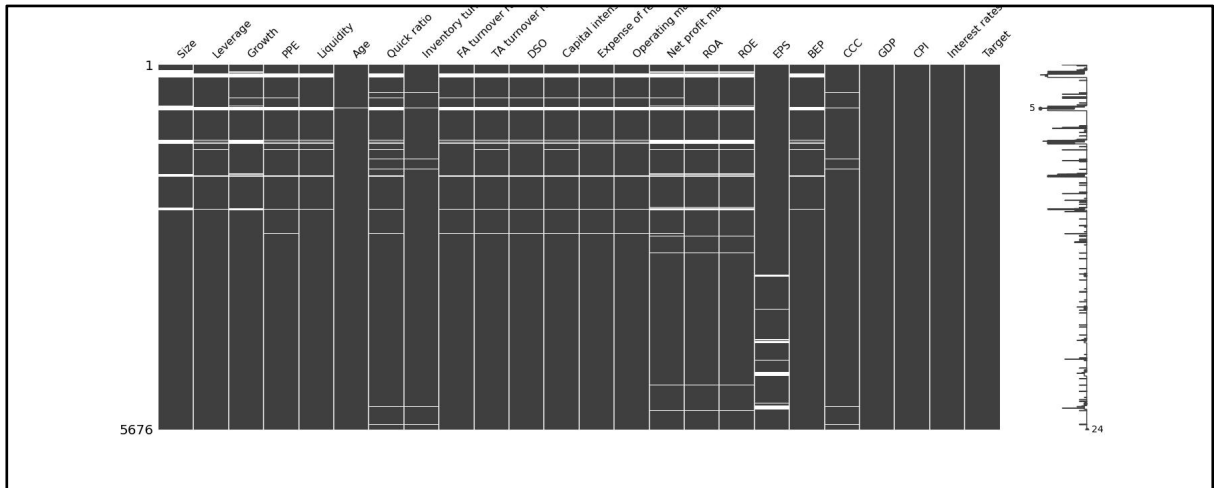
Appendix 9. Chart of values' distribution before fixing outliers in the total industries model



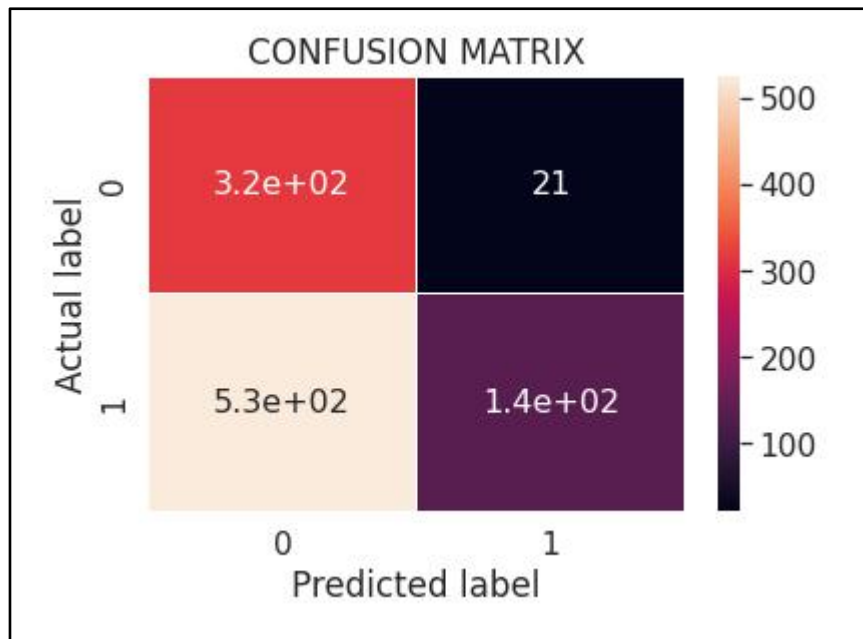
Appendix 10. Chart of values' distribution after fixing outliers in the total industries model

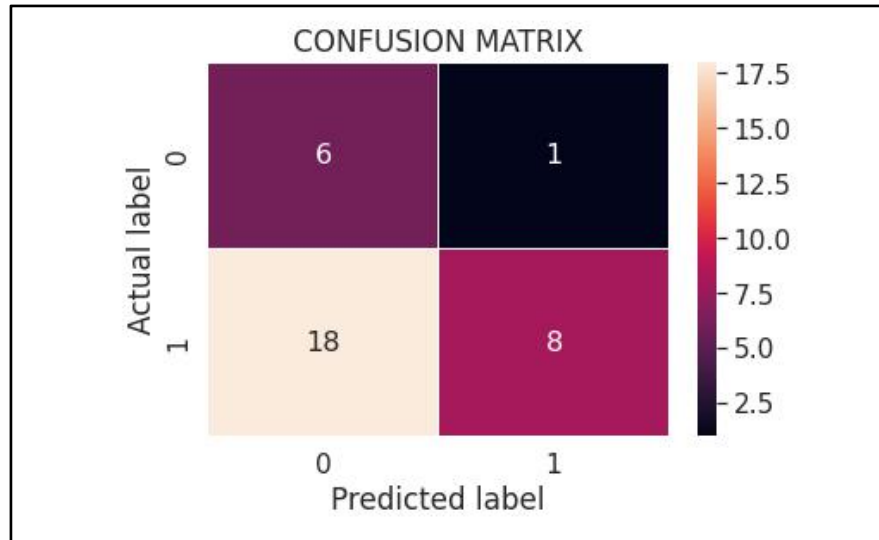


Appendix 11. Chart of missing values' distribution in the total industries model



Appendix 12. Chart of confusion matrix in the total industries model



Appendix 13. Chart of confusion matrix in the Healthcare model**Appendix 14. Chart of confusion matrix in the year 2020 model**