Science & Technology Development Journal – Economics - Law and Management 2023, 7(4):4814-4833

Open Access Full Text Article

Research article

Application of machine learning in classification of overinvestment: Evidence from listed firms in Vietnam stock exchange market

Tam Phan Huy[1,2,*], Tram Linh Dinh Ngo[1,2], Anh Ngoc Thi Nguyen[1,2], Nghia Trong Quoc Nguyen[1,2], Linh Hoang Thao[1,2], Thanh Van Trinh[1,2]

ABSTRACT

Studies have consistently demonstrated that both overinvestment and underinvestment exert adverse effects on the overall efficacy of business operations, showcasing the significance of understanding and addressing these phenomena in the realm of scholarly research. Therefore, in this study, we aim to develop an accurate machine-learning model to identify overinvestment in firms listed on the HSX and the HNX stock exchanges in Vietnam. We decided to conduct a comparison to identify the optimal model for classifying firms of overinvestment or not, including Logistic Regression, K-Nearest Neighbor (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree, and Random Forest. Using a sample of 658 non-financial listed companies in Vietnam between 2011 and 2021, our result shows that the most importance predictor variable is "FCF" (free cash flow), with an importance value of 0.14. Although both logistic regression and random forest (RD) algorithms demonstrate high accuracy in identifying firms with overinvestment, the Random Forest algorithm exhibits slightly higher precision and recall for class 1 (overinvestment firms) when compared to Logistic Regression. By contrast, the accuracy performance of the four models (NB, KNN, DT, and SVM) is low, ranging from 0.53 to 0.67. At the microeconomic level, this research can help businesses gain insights into their financial performance, identify areas for improvement, and take proactive measures to avoid financial distress and improve profitability by identifying potential cases of overinvestment. Overall, the study provides a valuable contribution to the field of financial analysis using machine learning techniques. We firmly believe that the findings of this research will serve as a significant scholarly reference for future investigations in the field and explore other importance predictors of overinvestment in Vietnam and other emerging markets.

Key words: Classification, Overinvestment, Machine learning

[1]*University of Economics and Law, Ho Chi Minh City, Vietnam*

[2]*Vietnam National University, Ho Chi Minh City, Vietnam*

Correspondence

Tam Phan Huy, University of Economics and Law, Ho Chi Minh City, Vietnam

Vietnam National University, Ho Chi Minh City, Vietnam

Email: tamph@uel.edu.vn

History

- Received: 13-6-2023
- Accepted: 14-12-2023
- Published Online: 31-12-2023

DOI :

https://doi.org/10.32508/stdjelm.v7i4.1269

INTRODUCTION

Although the study of overinvestment and the recognition of an enterprise's overinvestment situation is a critical topic in today's volatile world. It is well understood that effective investment can increase the development of a company and promote long-term growth of any enterprise. From an economic perspective, an investment is the purchase of goods that are not consumed today but will be used to generate wealth in the future [1]. Investment plays an essential role in economic development in that it is an asset or item acquired with the intention of generating income or recognition. Overinvestment can occur when a business had to spend more than it needs to stay afloat [2]. It is essential to know if a company is overinvested because overinvestment will harm business performance [3]. Previous studies and practice have demonstrated that overinvestment is an important issue and should be studied.

Recent studies on global overinvestment have primarily been conducted in the United States and China, and they are contentious in a variety of ways. Typically, studies in China focus on assessing overinvestment behavior and the factors that explain this overinvestment behavior [4–7]. Overinvestment is closely related to the use of corporate debt. According to empirical research, businesses with high financial leverage tend to overinvest. Because once the enterprise raises debt to finance investment, the risk shifts from the owner to the creditor, shareholders are more daring in investment decisions, which can easily lead to overinvestment [8–10]. However, in Vietnam, in particular, and in the world in general, until recent years, there has not been any scientific research focusing on the application of machine learning to forecast overinvestment.

Significant research has been conducted on the factors that influence overinvestment, but there is still

Cite this article : Huy T P, Ngo T L D, Nguyen A N T, Nguyen N T Q, Thao L H, Trinh T V. **Application of machine learning in classification of overinvestment: Evidence from listed firms in Vietnam stock exchange market**. *Sci. Tech. Dev. J. - Eco. Law Manag.* 2023, 7(4):4814-4833.

a lack of research on how to apply machine learning in classifying overinvestment. Future researchers may continue to analyze machine learning methods to classify overinvestment in companies. One research gap in the area of applying machine learning methods to classify overinvestment in firms in Vietnam is the lack of studies that compare the performance of different machine learning algorithms on this task. Although employed neural networks and fuzzy logic in constructing financial risk analysis models, there has been insufficient attention to alternative algorithms that could be similarly effective, such as support vector machines, neural networks, or random forests. A comparative analysis of the performance of different algorithms in this context could help identify the most effective approach for predicting overinvestment in Vietnamese firms using machine learning methods. By addressing these research gaps, we can better understand the relationship between factors that influence overinvestment and itself. Therefore, we decided to conduct this study: Application of Machine Learning in classification of overinvestment: Evidence from listed firms in Vietnam stock exchange market. Our study clarifies how machine learning works in classifying overinvested companies. In addition, in this study, we also build a completely new model for measuring overinvestment based on the existing foundational theories.

This research focuses on the main objective to compare the performance in classify overinvestment companies of six classification algorithms: logistic regression [11], support vector machine (SVM), decision tree [12], random forest [13], Naive Bayes (NB) and K-nearest neighbor (KNN). With the aforementioned comparison, we indicate which is the most suitable algorithm for classifying overinvestment companies.

LITERATURE REVIEW

Investment and Overinvestment

Investment refers to the allocation of funds or resources towards a specific objective, whether it be obtaining an asset, supporting a production process, or establishing a new business abroad; and this pursuit is typically motivated by the desire to realize future profits or gains [14,15]. To put it differently, the act of investing involves directing resources towards particular projects in the hope of obtaining a return on investment [16]. Additionally, an investor may seek to exert influence over corporate governance and establish a lasting interest in an enterprise operating in a distinct economic environment posit that a firm's investment policy is unaffected by its financing decisions within a perfect market [17]. However, in the natural world, factors such as asymmetric information and agency costs create problems such as underinvestment or overinvestment, where a company invests less or more than the optimal amount, respectively.

Overinvestment is a term used to describe the situation when a company or an economy spends too much on capital goods or projects that do not generate enough returns to justify the investment. Extraordinary investing was first introduced by [2]. based on the view of free cash flow. Jensen claimed that when a company has more free cash flow than it requires to maintain its operations and invest in projects with positive net present value (NPV > 0), this can lead to abnormal investment behaviors. Jensen & Meckling argue that due to the separation between ownership and management rights in most modern enterprises, there is always competition for power, even a conflict of interest between shareholders and investors manages [18]. Managers favor projects that benefit themselves over shareholders, which creates the problem of overinvestment. Degryse & De Jong and Richardson propose two concepts related to the abnormal investment status of firms: underinvestment and overinvestment [19,20]. Information asymmetry in the market leads to underinvestment, while agency problems lead to overinvestment.

Theoretical Basis

The research problem pertains to the different theoretical frameworks that are associated with overinvestment, namely: Capital market imperfections theory, Agency Theory, Free Cash Flow Theory, Behavioral finance theory, and Resource dependence theory.

Capital market imperfections theory

The theory of capital market imperfections posits that market frictions and information asymmetry can result in a misallocation of resources and reduced economic performance by causing overinvestment in specific industries or sectors [21]. Financial constraints arising from market frictions and information asymmetry, such as transaction costs, adverse selection, and asymmetric information, can limit access to external financing at a reasonable cost. This can lead to overinvestment when companies resort to internal funds to finance their investment projects. In some cases, companies may overinvest in new projects simply to utilize their excess cash, even if those projects are not profitable in the short term [21–23]. The pursuit of growth opportunities and profitability can also contribute to overinvestment behavior [21–23]. For instance, firms may overinvest in new projects to take advantage of potential future growth, even if those projects are not profitable in the short term[23]. Similarly, companies may overinvest in projects with expected high profitability, which may not be sustainable in the long term[23]. These behaviors can lead to a misallocation of resources and a reduction in overall economic performance. Moreover, the pecking order theory of capital structure can exacerbate overinvestment as companies may issue more debt or equity when internal funds are insufficient, increasing financing costs and signaling negative news to the market. Finally, market frictions such as transaction costs, regulatory barriers, or imperfect competition can also contribute to overinvestment[22]. In conclusion, the theory of capital market imperfections provides valuable insights into the causes and consequences of overinvestment, highlighting the dangers of excessive reliance on internal capital and the importance of external financing.

Agency Theory

Jensen & Meckling are credited with developing agency theory, modeling the theory within the principal-agent relationship framework.[18]. Overinvestment in companies is linked to agency theory through the concepts of agency costs and agency problems. Agency costs refer to the expenses incurred by a company when managers prioritize their own interests over those of shareholders, leading to overinvestment in projects that benefit the managers more than the shareholders[18].This can happen when managers are motivated to pursue projects that are not in the best interests of shareholders, such as projects that increase their salaries or bonuses[2]. If a company invests excessively in capital goods or projects that do not generate sufficient returns, it may have to reduce costs elsewhere, such as by cutting employee salaries or bonuses, to offset the losses. Furthermore, investing in projects that do not yield adequate returns can result in a decline in shareholder value, which can negatively impact the company's financial performance in the long run[2,18]. Effective corporate governance mechanisms that align the interests of managers and shareholders can mitigate overinvestment due to agency problems[2].

Free Cash Flow Theory

The concept of free cash flow refers to the amount of cash that a company has available after paying for its operating expenses, capital expenses, and dividends. This cash surplus can be used to maintain assets and new investments[20,24]. Posit that free cash flow has the potential to serve as an indication of overinvestment. The hypothesis regarding free cash flow is rooted in agency theory, which suggests that managers may be inclined to invest this surplus cash into projects that could ultimately lower profits and shareholder value, but provide them with greater control and status within the organization.[2,18,25,26] This is more evident for companies with high free cash flow but poor growth prospects, which encourages managers to overinvest. Although this investment enhances the manager's personal benefits, it destroys the company's value, reducing shareholder wealth. Richardson[20]. finds that overinvestment is mainly concentrated in firms with the highest free cash flow. Additionally, growth opportunities and capitalization can contribute to overinvestment behavior under this theory is a technology company that has excess cash flows from its profitable operations[2]. When companies face financial difficulties and cannot secure financing at a reasonable cost, the company can invest in a variety of growth opportunities, such as expanding into new markets or developing new products. new products without thoroughly evaluating the profitability or sustainability of these investments. The firm may also use its excess cash to repurchase shares, further increasing its capitalization. However, if these investments do not generate sufficient returns, the company may be considered to have overinvested, as it has allocated resources towards projects that do not create value for its shareholders[2,18].

Asymmetric Information Theory

In 1970, Akerlof introduced the concept of asymmetric information in his study The market for "lemons": Quality uncertainty and the market mechanism. This theory suggests that buyers have less information about the quality of a product they purchase, leading to mispricing by the seller. Asymmetric information can result in overinvestment when managers possess better information about potential project returns than outside investors[27,28]. This can lead to risky investments that are not in the best interest of the company or its shareholders due to differences in available information. Asymmetric information can lead regulators to engage in unethical or illegal practices for personal gain. However, companies can improve financial reporting and disclosure transparency to reduce information asymmetry. Providing more information to investors can help reduce managers' information advantage and increase market efficiency[27]. In conclusion, asymmetric information theory highlights challenges in managing information available to managers and external investors and the risks of excessive investment in risky ventures. To mitigate overinvestment risk, businesses can improve disclosure practices and understand the role of asymmetric information.

Behavioral finance theory

Behavioral finance theory suggests that psychological biases and emotional factors can influence overinvestment, and cognitive limitations can impact how investors perceive and process information[29]. Overly optimistic managers may fail to assess the risks and uncertainties involved in investment projects, and become emotionally attached to their projects, leading to excessive investments in projects that may not generate sufficient returns. Herding behavior, where investors follow the decision of others, can cause many companies to invest in the same industry or market, leading to oversupply and reduced profits[29,30]. The sunk cost fallacy, where companies continue to invest in a project with little chance of success, is an example of this phenomenon. The disposition effect, where investors hold on to losing investments for too long in the hope of recouping their losses, can contribute to overinvestment[31]. Anchoring bias, where investors rely too heavily on a single piece of information in their investment decision-making process, can also lead to suboptimal investment decisions[29]. Understanding the potential causes and consequences of overinvestment due to psychological biases and cognitive limitations can help policymakers and investors develop strategies to reduce the risk of overinvestment, make more rational investment decisions, and promote higher market efficiency[29–31].

Resource dependence theory

Resource dependency theory explains that organizations may invest heavily in external resources such as raw materials, technology, or skilled labor to create value for their customers and generate profit. However, the availability, quality, and cost of these resources may be uncertain and beyond the control of the company, leading to overinvestment in certain areas. This can result in a scenario where organizations continue to invest in a resource, even when it is no longer valuable or necessary, making them reluctant to reduce investment. To minimize the risk of overinvestment, companies can develop alternative sources of supplies or products and invest in acquiring or developing critical resources, even if the returns are uncertain. However, over-investing in resources can harm a company's profitability, stock prices, or even lead to bankruptcy if the returns on investments do not materialize or resources become outdated[32,33].

Previous studies

Richardson examines the extent of firm level overinvestment of free cash flow[20]. Using an accounting-based framework to measure over-investment and free cash flow, he found evidence that, consistent with agency cost explanations, overinvestment is concentrated in firms with the highest levels of free cash flow. Further tests examine whether firms' governance structures are associated with over-investment of free cash flow. The evidence suggests that certain governance structures, such as the presence of activist shareholders, appear to mitigate overinvestment. Hao et al., and Nghia et al., both employ a measure of overinvestment based on Richardson's model.[20,34,35]. Hao et al., practiced with 650 real estate companies listed in China between 2010 - 2015, successfully proved that overinvestment is a common practice (33.54% of real estate companies) and debt structure has a limited effect on overinvestment thereby providing policy implications for mitigating this problem[34]. Nghia et al., conduct a study that investigates the detrimental impact of overinvestment on firm performance and the moderating role of debt and dividend in mitigating agency costs resulting from overinvestment[35]. The research comprises all of Vietnam's non-financial companies that are listed on HSX and HNX from 2006 to 2016. The study employs two specific measurements of overinvestment, namely HP Filter and the positive error terms obtained from the subequation of Overinvestment Estimation. The findings reveal that overinvestment has a negative impact on profitability in Vietnamese enterprises. However, the harmful effect of overinvestment can be alleviated by the use of debt or the payout of dividends. Nevertheless, when combined, the separate influences of the two-variable interaction tend to be weakened. Overall, there are still limitations in the number of research studies related to the issue of overinvestment classification using machine learning models. While in recent years, machine learning algorithms have become increasingly popular as prediction tools in various industries such as finance, economics, healthcare, and marketing.

Machine learning (ML) is a type of computational intelligence that employs pre-programmed algorithms to examine input data and acquire knowledge from it through supervised or unsupervised methods, enabling it to produce output values that fall within an acceptable range. ML algorithms are adept at managing large and intricate datasets while also being capable of capturing non-linear relationships between variables. The effectiveness of ML has been demonstrated over the past decade, and its feasibility has been demonstrated as a substitute for classical statistical models in various research applications including mathematical problems forecasting, regression, and classification [36].

Several studies around the world have explored the application of machine learning in predicting financial and investment problems. For example, Lakhal et al. utilized machine learning techniques such as Logistic regression, Discriminative analysis, Neural networks, Boosting, AdaBoos, and RF to classify two basic investment models by Richardson and Biddle et al. and determine the impact of CSR performance on investment performance. [20,37,38] Their findings suggest that Richardson's method yields better investment efficiency results. Özlem & Tan examine the motives behind firms' decisions to hold cash and cash equivalents, and why they refrain from redistributing or reinvesting their cash [39]. The authors conduct an extensive literature review on the utilization of machine learning algorithms, including MLR, KNN, SVM, DT, extreme gradient boosting algorithm (XGB), and multilayer neural network (MLNN) methods, to predict the cash holding policy of 211 Turkish listed companies in Borsa Istanbul from 2006 to 2015. Their study revealed that DT and XGB models demonstrated superior performance compared to the other models, with an R2 value of 0.73.

Although the study provides valuable insights, it is subject to certain limitations that need to be considered. Firstly, the research primarily centers on Turkish firms and their attributes, thus, the outcomes may not be applicable to other countries or regions. Moreover, the time frame of the study is from 2006 to 2015, and as a result, the findings may not accurately reflect the current market situation or changes. Lastly, the study did not consider macroeconomic variables, including gross domestic product growth, interest rates, and oil prices, which could have an impact on the results. Wu et al. concentrated on Taiwan's high-tech industry to predict cash holdings using DT techniques in the domain of financial forecasting with machine learning [40]. Their research showed that among all the DTs, RF had the highest prediction accuracy. In a similar vein, Moubariki et al. conducted research on the cash management of the public sector and concluded that DT was the most effective predictive approach [41]. Likewise, Bae explored the predictive dividend policy decisions of Korean companies, utilizing SVMs, DTs, and neural networks, and determined that SVM was the most efficient technique [42]. Using Gaussian process and radial neural network models, Gholamzadeh et al. carried out a research investigation to predict financial constraints of companies on the Tehran Stock Exchange. Their study found that machine learning methods are appropriate for anticipating financial difficulties experienced by firms [43]. In addition, utilizing RF, quadratic discriminant analysis, and linear discriminant analysis, Mousa et al. forecasted the financial performance of 63 listed banks in emerging international markets [44]. According to their results, the RF approach produced the most precise predictive models. Furthermore, including disclosure tone factors in addition to financial variables enhanced the models' precision and quality.

In Vietnam, there are studies on using machine learning models to support and predict financial-related problems. In another study, Tran et al. utilized empirical evidence from listed companies in Vietnam between 2010 to 2021 to predict financial hardship using machine learning algorithms [12]. The research evaluated the predictive capability of different machine learning models and utilized SHAP values to interpret the obtained results. According to the study, XGB and random forest exhibited better recall and F1 scores compared to other models. Conversely, logistic regression, artificial neural network, and SVM showed elevated Type I errors. The random forest model had the highest AUC value (0.9788), signifying its superior classification performance in comparison to the remaining models. However, in Vietnam, there are still no specific studies on the application of machine learning models to classify overinvestment.

METHODOLOGY

Data

The present study utilized data from all companies listed on the two major Vietnamese stock exchanges, namely the Ho Chi Minh City Stock Exchange (HSX) and the Hanoi Stock Exchange (HNX). The data were obtained from the Refinitiv Eikon database and covered a period of a decade, from 2010 to 2020. Following the process of filtering and cleaning, the study obtained 6755 observations from a total of 717 listed companies that were listed after 2009. After the exclusion of financial enterprises, the remaining sample consisted of 658 non-financial enterprises. Subsequently, the infinite variables and missing value data were removed, resulting in a final set of 1707 valid data.

Regarding the data collection for about 10 years, there are many reasons for the choice of the authors. First of all, the longer the data collected over a period of time, the more observations it will have, meaning the more accurate the results will be, in case of fluctuations and data errors. The second reason, which is also the main reason for 10 years, is the economic crisis cycle according to Dr. Nguyen Duc Thanh, Director of the Institute for Economic and Policy Research (VEPR). In Vietnam, the last two economic crises were in 1997 and 2008. Also according to his sharing from the end of 2018 and the beginning of 2019, the Vietnamese market is showing many potential crisis factors. Choosing 10 years as a way for the team to review economic indicators, eliminate short-term fluctuations and provide the most intuitive, general results.

Empirical framework

Our empirical framework is built based on the combination of two research models, the traditional model developed from the original study of Hao et al. and the modern model applying machine learning in predicting a company's overinvestment [34].

The methodology of this paper is drawn from the model-construction approach developed by Hao et al. and Richardson [20,34]. We propose to use model (1) as a means to estimate firms' level of overinvestment.

$$\begin{aligned} \ln v_{i,t} &= \alpha_0 + \alpha_1 \ln v_{i,t-1} + \alpha_2 Dar_{i,t-1} \\ &+ \alpha_3 Cash_{i,t-1} + \alpha_4 Growth_{i,t-1} \\ &+ \alpha_5 Size_{i,t-1} + \alpha_6 Age_{i,t-1} \\ &+ \alpha_7 Ret_{i,t-1} + \varepsilon \end{aligned} \tag{1}$$

In this study, we undertook a rigorous data preprocessing protocol for datasets procured from the Ho Chi Minh City Stock Exchange (HSX) and the Ha Noi Stock Exchange (HNX). The initial phase entailed the amalgamation of multiple pertinent datasets into a consolidated repository, as elucidated in Figure 1. Subsequently, an exhaustive data cleansing process was executed, encompassing the expurgation of infinite values, the amelioration of null entries, the eradication of extraneous symbols and special characters, and the judicious application of imputation techniques to rectify missing values. This exacting data preprocessing regimen serves as the linchpin for ensuring the integrity, quality, and reliability of the dataset, thus establishing a robust foundation conducive to precise and profound analysis. By meticulously preparing the data, we were able to harness a gamut of machine learning algorithms for the express purpose of anomaly detection, thereby affording us profound insights into the behavioral intricacies of financial data within the Vietnamese stock exchanges. If the residual value (ε) is greater than zero, it suggests the presence of overinvestment. Where $\mathrm{Inv}_{i,t}$ is new investment from firm i in year t, scaled by total assets. This variable depends on the lagged new investment ($\mathrm{Inv}_{i,t-1}$); the asset liabilities rate is measured as the total liabilities to total assets at the beginning of the year ($\mathrm{Dar}_{i,t-1}$). The firm's growth opportunities ($\mathrm{Growth}_{i,t-1}$) are measured as the growth rate of the annual sales revenue. The firm's cash holding rate ($\mathrm{Cash}_{i,t-1}$); the number of years from IPO to the end of the last year ($\mathrm{Age}_{i,t-1}$); the log of a firm's total assets ($\mathrm{Size}_{i,t-1}$); and the dividend distribution rate of the previous year ($\mathrm{Ret}_{i,t-1}$). All of these variables are lagged one year.

To develop our new model, we combined previous research studies. The model includes Manager Confidence [45], Financial Constraints (FC), Agency Problems (AP), Size of the Firm (SOF), Growth Opportunities (GO), Profitability [33], and Capitalization (CL). These variables are used to determine the presence of overinvestment, and lagged variables are also taken into account.

The modern research model is built based on evaluating the factors influencing overinvestment, especially in Vietnam. Inheriting from the model in the article of Hao et al. and Richardson, we continue to apply the old variables and introduce new ones that are suitable for the practice in the Vietnamese market [20]. The details of eight variables are as follows:

Overinvestment (OInv)

Overinvestment is a dependent variable; the results are expressed in 2 forms as 1 - overinvested and 0 - not overinvested. As Table 1 referred, regression model for running overinvestment variable is:

$$\begin{aligned} O\ln v_{i,t} &= \alpha_0 + \alpha_1 MC_{i,t-1} + \alpha_2 FC_{i,t-1} \\ &+ \alpha_3 AP_{i,t-1} + \alpha_4 SOF_{i,t-1} + \alpha_5 GO_{i,t-1} \\ &+ \alpha_6 PF_{i,t-1} + \alpha_7 CL_{i,t-1} + \varepsilon \end{aligned} \tag{2}$$

Where, the inexplicable remainder is ε . If ε carries the sign (+) the enterprise is overinvested, the result is displayed as 1. If ε carries the sign (-) is the underinvested enterprise, the result is displayed as 0.

Manager overconfidence (MO)

The manager's overconfidence is understood as his willingness to make high-risk decisions that may not be met by his ability [53]. These managers tend to overtrust their ability to make accurate predictions and decisions. In such cases this overconfidence can lead to wrong investment decisions. Research by Grinblatt & Keloharju shows that overconfident individual investors tend to seek sensations leading to overinvestment in stock transactions, which means underperformance [46]. Directly assessing a manager's overconfidence can be challenging, but various metrics have been used in research studies. It can be

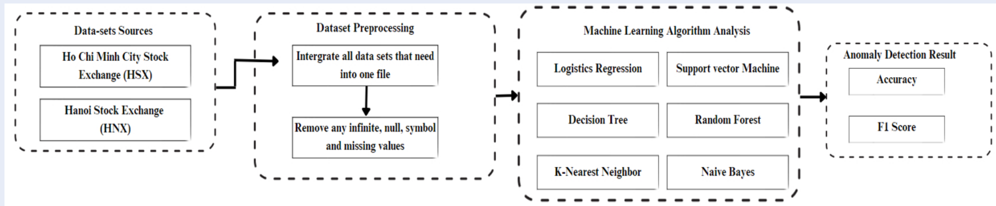

Figure 1: Data Processing (Source: Authors)

Table 1: Summary table of variables

Variables	Measurement	Research
Overinvestment	$O\ln v_{i,t} = \alpha_0 + \alpha_1 MC_{i,t-1} + \alpha_2 FC_{i,t-1} + \alpha_3 AP_{i,t-1} + \alpha_4 SOF_{i,t-1} + \alpha_5 GO_{i,t-1} + \alpha_6 PF_{i,t-1} + \alpha_7 CL_{i,t-1} + \varepsilon$	The authors
Volatility of Firm Earnings (Vol) Stock Price Volatility (Spv)	vol = $\sigma\sqrt{T}$	[11,46]
Debt-to-Asset Ratio	(Total debt)/(Total assets)	[21]
Z – score	Z" = 6.56 X1 + 3.26 X2 + 6.72 X3 + 1.05 X4 + 3.25	[47]
Free Cash Flow (FCF)	Net cash flow from operating activities - CAPEX - Interest expense	[48]
Size of the Firm (SOF)	Total assets= Short-term assets + Long-term assets ln(Total Assets)	[49]
Growth Opportunities (GO)	(Net Revenue$_t$ - Net Revenue$_{t-1}$)/(Net Revenue$_t$)	[50]
Profitability[33]	ROA=(Profit after tax)/(Total assets)	[51]
Capitalization (CL)	Number of shares outstanding x value of shares at time	[52]

Source: Authors.

measured through Corporate Earnings Volatility and Stock Price Volatility[11]. To measure volatility, we use the formula:

$$vl = \sigma\sqrt{T}$$

Where:

- vol = volatility over some interval of time
- σ = standard deviation of net income.
- T = number of periods in the time horizon
- This formula was proposed by the mathematician Benoît Mandelbrot to measure any volatility of a financial asset over a certain period of time.

Financial Constraints (FC)

When firms experience financial constraints, such as limited access to credit, they may engage in excessive investments to demonstrate their creditworthiness to lender[23]. Furthermore, financial constraints can make companies more risk-aware, making them invest in low-risk ventures even when returns are suboptimal[54]. Financial constraints are often difficult to measure directly, but there are several commonly used representations that have been used in empirical studies. For example, the debt-to-asset ratio and the Z-score[21,47].

To measure the debt-to-asset ratio, we use the formula:

Debt to asset ratio = (Total debt)/(Total assets)

For the Z - Score we apply to listed companies:

Z" = 6.56 X1 + 3.26 X2 + 6.72 X3 + 1.05 X4 + 3.26

Where:

- X1: Current assets/Total assets
- X2: Earning after tax/Total assets.
- X3: EBIT/ Total assets
- X4: Market capitalization of common shares/ Total book value of debt

The results will satisfy the following conclusions:

- If Z" > 2.6: The company is in a safe zone and has no risk of bankruptcy.

- If 1.1 < Z" < 2.6: The company is in an alert zone and may have a risk of bankruptcy.
- If Z" < 1.1: The company is in a danger zone and has a high risk of bankruptcy.

Agency Problems (AP)

Agency problems arise when there is a division between ownership and control in an organization. Managers can invest in projects that serve their personal interests instead of the interests of shareholders, lead to overinvestment [18]. Research by Richardson shows that the problem of agency costs, overinvestment is often concentrated in companies with the highest free cash flow [20]. Therefore, to measure overinvestment the representative selection group is free cash flow. Specifically:

FCF = Net cash flow from operating activities - CAPEX - Interest expense

The theoretical basis of this formula is the basic principle of cash flow in corporate finance, that is, the ability of the business to generate free cash flow after deducting fixed and overhead costs. capital.

Size of Firm (SOF)

Size of firm is a term for size that has an important influence on a firm's ability to generate revenue (Babalola & development, 2013). Previous studies have shown that a firm's size has an impact on overinvestment. Titman *et al.* and Harford & Li all conclude that larger firms tend to overinvest compared to smaller companies [13,49]. To measure or distinguish the size of companies, we use the criterion of total assets through which the author compares the value of this company with other companies to get an overall view of the position. position and size of the firm in the industry:

Total assets= Short-term assets + Long-term assets

Growth Opportunities (GO)

Growth opportunity is the ability and potential of a business to develop in the future. Miller & Modigliani asserted the influence of growth opportunities on firm value [55]. Firms that possess significant growth prospects might have a higher tendency to engage in overinvestment as they have a greater number of potential investments at their disposal [50].

To measure the growth opportunity of the firm, we use the revenue growth rate, similar to the growth opportunity representation in the study [56]. The index is calculated using the formula commonly used in financial statements:

(Net Revenue$_t$ - Net Revenue$_{t-1}$)/(Net Revenue$_t$)

Profitability [33]

Profitability is the degree to which a business makes a profit. High profits can lead to companies looking for new investment opportunities, which can easily lead to many wrong investment decisions due to subjective reasons as they possess more resources that could be utilized. To measure corporate profitability, we estimate ROA, similar to the proposal of Adyani & Sampurno considering the bank's profitability is measured by ROA at the end of year t. [57]. The specific formula is as follows:

ROA = (Profit after tax)/(Total assets)

Capitalization (CL)

Capitalization is a financial concept used to value a company's market value.Companies with high levels of capitalization tend to overinvest. Conversely, companies with lower levels of capitalization may be more conservative with their investments due to limited resources. To measure capitalization, we use the formula given by Fama & French [52]:

Market capitalization = Number of shares outstanding x Market price of each share

In recent times, machine learning algorithms have become increasingly popular as prediction tools, even within the finance industry. In order to forecast overinvestment, we utilized and compared several machine learning algorithms, including logistics regression, support vector machine, decision tree, random forest, K-Nearest neighbor, and Naive Bayes. In this study, the author applies the following machine learning algorithms: Logistics Regression, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, Decision Tree and Random Forest. These machine learning algorithms will be comparing performance based on accuracy, precision, recall, F1 Score, and time consuming.

RESULT AND DISCUSSION

Descriptive Statistics

The number of enterprises listed on the HSX exchange is 360/658, accounting for 54.71%, and it is accounting for 45.29% for HNX (298 enterprises). The study examined various financial and non-financial variables of listed firms, including overinvestment, debt ratio, FCF, ROA, ROE, quick ratio, capitalization, manager score, growth opportunities, OCF ratio, retain and z score. After removing the observations with the missing value, the data was utilized, including 1707 observations with descriptive statistics as follows:

Collected data is indicated as qualitative data 0 (non-overinvestment) and 1 (overinvestment), these results

Table 2: Descriptive statistics of observations

Variables	Obs	Mean	Std. Dev.	Min	Med	Max
Debt ratio	1707	0.381	0.431	0.000	0.275	2.814
FCF	1707	0.015	0.137	-0.625	0.018	0.861
ROA	1707	0.092	0.081	-0.006	0.073	0.994
ROE	1707	0.182	0.126	-0.012	0.163	1.269
Quick ratio	1707	0.487	0.634	0.001	0.263	4.667
Capitalization	1707	14.402	45.517	0.008	2.674	576.794
Manager confidence	1707	0.245	1.783	-21.353	0.059	27.134
Growth Oppotunities	1707	0.125	0.378	-1.039	0.077	2.813
OCF ratio	1707	0.248	0.330	-0.620	0.133	2.265
Retain	1707	3.569	13.854	-117.148	0.579	114.775
Z score	1707	2.984	2.046	-0.238	2.470	13.624

Source: Author's calculation

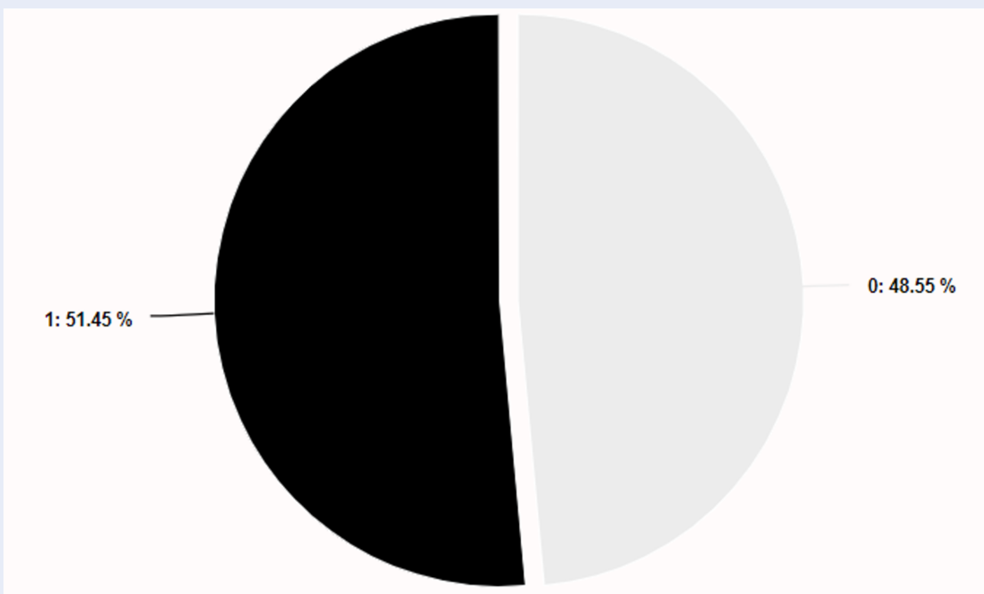

Figure 2: Frequency distribution of overinvestment. Source: Author's calculation

are calculated based on the traditional model to compare with the value running in the model. According to the results of running analysis from Stata, there are 827/1707 observations of overinvestment at 48.85%. These values are used to assess the accuracy of machine learning algorithms. The debt ratio of enterprises is at a low level with a range from 0 to 2.814, a mean value of 0.381, median value of 0.275, and volatility of 0.431. Additionally, the free cash flow represented as a ratio to total assets, ranges from -0.626 to 0.861, with a mean value of 0.015 and a median value of 0.018, indicating positive news about the business's cash flow. The return on assets has a mean value of 0.092 and a median value of 0.074, indicating the asset utilization efficiency of listed firms. The return on equity has a mean value of 0.182 and a median value of 0.163, indicating an efficient use of equity. However, there are doubts about the firms' liquidity as the quick ratio has a mean value of 0.614, a median value of 0.274, and a standard deviation of 1.392.

The market capitalization value fluctuates widely from 0.008 trillion VND to 576.794 trillion VND with a mean value 14.402 trillion VND, a median value 2.674 trillion VND, and a value bias of 45.517 trillion VND. The CEO confidence index has a range from -21.353 to 27.134, with a median value of 0.059 greater than 0, indicating that CEOs are confident with their investment decisions, and may lead to overinvestment decisions. The net sales growth rate has a mean of 0.125, a median of 0.077, and a degree of variation of 0.378. The OCF ratio has a mean value of 0.248, a median value of 0.133, and a degree of variation of 0.330, indicating the ability to cover short-term debts of enterprises with net operating cash is quite low. The income retained after paying dividends to shareholders has median value 0.579 billion VND and mean value 3.569, indicating that most companies keep possession of profits to continue reinvesting to expand their markets, which also leads to the possibility of overinvestment of the business. Finally, the Z-score of the firm's probability of bankruptcy ranges from -0.238 to 13.624. The median value of 2.470 clearly shows that more than 50% of the observations are in the warning and danger zones.

In conclusion, analyzing financial metrics such as debt ratio, free cash flow, return on assets, return on equity, quick ratio, capitalization, CEO confidence index, growth opportunities, OCF ratio, income retained, and Z-score can provide valuable insights into the overinvestment of listed companies. By looking at these metrics and making comparisons between companies, investors and analysts can make informed decisions about investment opportunities.

Machine Learning model

There are six classification reports for six models: Logistic Regression, K-Nearest Neighbor, Naive Bayes, Support Vector Machine, Decision Tree and Random Forest. All models were trained to classify data into two classes, labeled 0 and 1. Class 0 represents all companies that do not overinvestment and class 1 represents all companies that do overinvestment. The report evaluates the performance of the model based on precision, recall, F1-score, support, and accuracy metrics. Precision is a measure of how many of the instances classified as positive are actually positive. Recall is a measure of how many of the actual positive instances are correctly identified as positive. The F1-score is a weighted average of precision and recall that provides a single measure of overall performance. Support is the number of instances of each class in the dataset.

Logistic regression model classification report is evaluated on a dataset containing 513 instances. The report provides various performance metrics for the model. Overall accuracy is 0.68, which means that it correctly predicted the class label for 68% of the instances in the dataset. The precision for class 0 is 0.66, which means that when the model predicts an instance to be in class 0, it is correct 66% of the time. The recall for class 0 is 0.74, which means that out of all the instances that actually belong to class 0, the model correctly identified 74% of them. The F1-score for class 0 is 0.70, which is the harmonic mean of precision and recall for class 0. Similarly, the precision for class 1 is 0.70, which means that when the model predicts an instance to be in class 1, it is correct 70% of the time. The recall for class 1 is 0.62, which means that out of all the instances that belong to class 1, the model correctly identified 62% of them. The F1-score for class 1 is 0.66. The macro avg of F1-score for both classes are 0.68, which is the average of these metrics across both classes.

With the K-Nearest Neighbors algorithm, the precision for class 0 is 0.61, which means that 61% of the instances predicted to be in class 0 are actually in class 0. The recall for class 0 is 0.67, which means that 67% of the instances in class 0 are correctly identified as class 0. The F1-score for class 0 is 0.64, which is the harmonic mean of precision and recall for class 0. For class 1, the precision is 0.63, recall is 0.57, and f1-score is 0.60. The overall accuracy of the model is 0.62, which means that 62% of the instances in the dataset are correctly classified by the model. The macro-average of precision, recall, and f1-score is the unweighted mean of these metrics across both classes, which is 0.62 in this case.

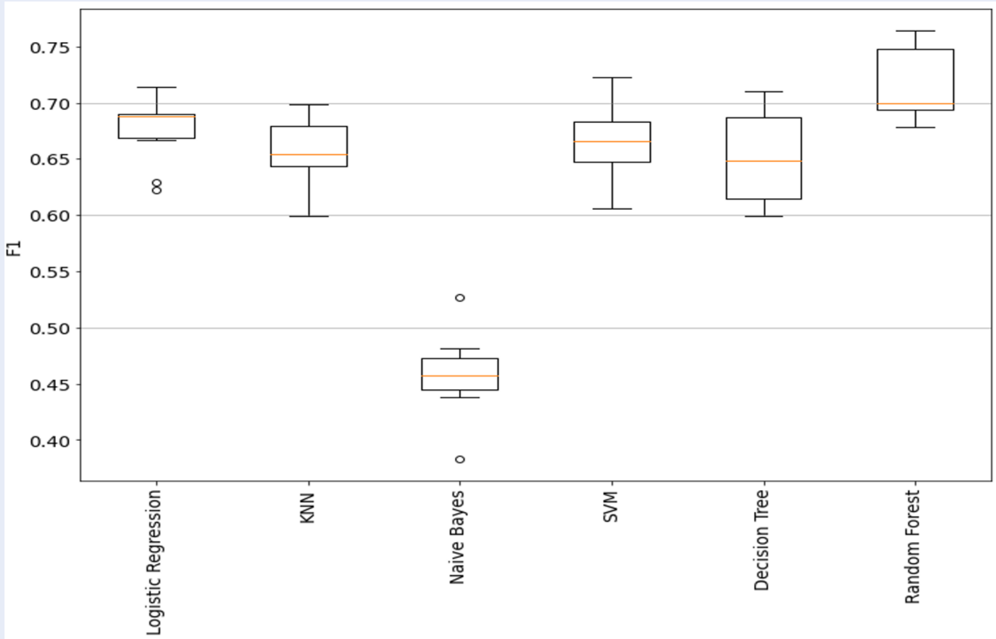

Figure 3: Algorithm Comparison - F1-Score. Source: Author's calculation

Naive Bayes model identified instances labeled as 1, with a precision of 0.52 and a recall of 0.93. Besides, the model performed poorly in identifying instances labeled as 0, with a precision of 0.66 and a recall of 0.13. The overall accuracy of the model was 0.53, indicating that the model correctly classified 53% of instances. The macro average F1-score was 0.44, which is the average F1-score across the two classes. In summary, the Naive Bayes model had a relatively good performance in identifying instances labeled as 1 but performed poorly in identifying instances labeled as 0. Therefore, the model may need to be further improved to achieve better overall performance on this dataset.

The SVM model performed relatively well in identifying instances labeled as 0, with a precision of 0.65 and a recall of 0.75. The model also performed well in identifying instances labeled as 1, with a precision of 0.70 and a recall of 0.60. The overall accuracy of the model was 0.67, indicating that the model correctly classified 67% of instances. The macro average F1-score was 0.67, which is the average F1-score across the two classes. In summary, the SVM model had a good overall performance, with high accuracy and reasonable precision and recall scores for both classes. Therefore, the SVM model can be considered a good choice for this classification task.

The Decision Tree model had similar precision, recall, and F1-score for both classes, with values around 0.63. The overall accuracy of the model was also 0.63, indicating that the model correctly classified 63% of instances. The macro average F1-score was 0.63, which is the average F1-score across the two classes. The Decision Tree model had a moderate overall performance, with similar precision, recall, and F1-score for both classes, and an accuracy of 63%. Despite the model have not performing as well as some other classification algorithms, it can still be useful for certain applications and datasets.

The Random Forest model performed reasonably well in identifying instances labeled as 0, with a precision of 0.68 and a recall of 0.74. The model also performed well in identifying instances labeled as 1, with a precision of 0.71 and a recall of 0.64. The overall accuracy of the model was 0.69, indicating that the model correctly classified 69% of instances. The macro average F1-score was 0.69, which is the average F1-score across the two classes.

The use of cross-validation technique is essential in evaluating the accuracy and F1-score of 6 classification models as it provides a robust and unbiased estimate of 6 models performance. In this study, we have used 10-fold cross-validation to evaluate the performance of 6 classification models, namely Logistic Regression, K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Decision Tree and Random Forest. Our results indicate that Naive Bayes has the lowest

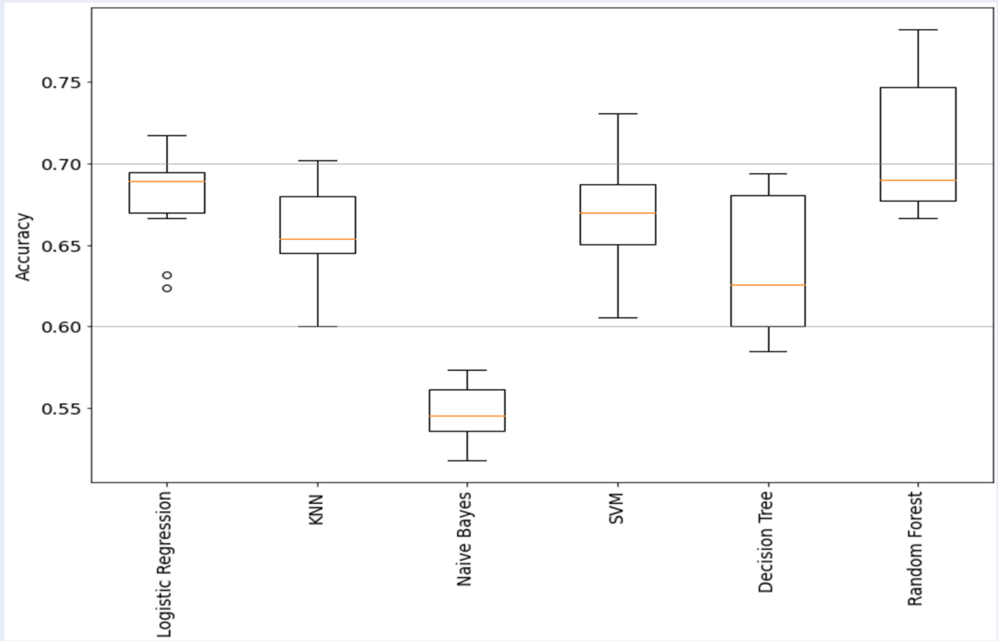

Figure 4: Algorithm Comparison - Accuracy. Source: Author's calculation

accuracy rate of about 55%. On the other hand, Random Forest and Logistic Regression algorithms show the highest accuracy rates of approximately 68% and 69% (Figure 2). Although both algorithms have similar accuracy rates, Random Forest exhibits greater variability as indicated by the larger spread of its results, with some instances yielding an accuracy rate of up to 75% (Figure 4). Considering these findings, we can conclude that Random Forest is the superior model among the six models evaluated in this study. This is due to its consistently high accuracy rates across different cross-validation folds, despite its higher variability compared to Logistic Regression. Therefore, we recommend the use of Random Forest for classification tasks that require high accuracy rates. Figure 3 depicts the F1-score of the 6 classification algorithms evaluated in our study. Our findings indicate that Naive Bayes has the lowest F1 Score, measuring less than 50% according to Table 2. In contrast, the Random Forest algorithm has the highest F1 Score, at around 70%. However, as shown by the large spread of its results, with a variation of up to 75%, Random Forest can be considered a model with moderate strength.

Accuracy is an important metric in classification because it directly measures the percentage of correctly classified instances, which is a fundamental goal of many classification problems. The accuracy of a classification model is determined by the number of true positives (correctly predicted positive instances), true negatives (correctly predicted negative instances), false positives (incorrectly predicted positive instances), and false negatives (incorrectly predicted negative instances). Accuracy is calculated as the ratio of the number of correctly classified instances to the total number of instances. A high accuracy means that the model is able to correctly classify most of the companies. While accuracy alone may not always provide a complete picture of the performance of a classification model, it is a crucial metric that is often used to evaluate the effectiveness of a model. Moreover, accuracy can be a useful metric when comparing different models or when assessing the impact of different features on the classification performance. F1-score is also an important metric for evaluating the performance of classification models because it considers both precision and recall, which are two important aspects of classification performance. Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positive predictions among all actual positive instances. F1-score shows the harmonic mean of precision and recall and provides a single metric that balances both precision and recall. F1-score is useful because it provides a single metric that considers both precision and recall and provides a balanced measure of classification performance. This makes it particularly useful in situations where both precision and recall are important, or where there is a trade-off between the two. The Random Forest model had a good overall performance, with high accuracy and reasonable precision and recall scores for both classes. Therefore, the Random Forest model can be considered a good choice among the 6 proposed models for overinvestment classification.

Figure 5 describes feature importances of Random Forest algorithm, which represents the relative importance of each predictor variable in the Random Forest model. The values provided indicate the contribution of each variable to the model's accuracy or predictive power. The most important predictor variable is "FCF" (free cash flow), with an importance value of 0.14. This suggests that free cash flow is a critical factor in predicting whether companies are overinvesting or not, likely indicating that companies with higher free cash flow tend to overinvestment. In contrast, "industry" appears to be the least important variable, with an important value of 0.055. This suggests that the industry in which a company operates may not be a critical factor in overinvestment of firms. Overall, these findings provide insights into the factors that contribute to overinvestment of firms and may have practical implications for financial decision-making.

DISCUSSION

Both Logistic Regression algorithm and Random Forest algorithm perform similarly in terms of average accuracy (0.68), but there are some differences in their performance regarding other metrics. For class 0, Random Forest has slightly higher precision (0.68 vs. 0.66) compared to Logistic Regression, while both algorithms have the same recall (0.74). For class 1, Random Forest outperforms Logistic Regression in terms of both precision (0.71 vs. 0.7) and recall (0.64 vs. 0.62). The F1 score, which combines precision and recall into a single metric, is also slightly higher for Random Forest (0.69 vs. 0.68).

In the context of this classification problem, class 0 represents firms with no overinvestment, while class 1 represents firms with overinvestment. Overinvestment occurs when a company invests too much capital in its operations or assets, which can lead to inefficient resource allocation and diminished returns. When interpreting the results, it's essential to consider the implications of each class. For instance, a high recall for class 0 indicates that the algorithm can correctly identify a large proportion of firms without overinvestment. On the other hand, high precision for class 1 suggests that the algorithm can accurately pinpoint firms with overinvestment. As mentioned earlier, both Logistic Regression and Random Forest perform similarly in terms of average accuracy, but there are some differences in their performance concerning precision and recall for each class. The Random Forest algorithm has slightly better precision and recall for class 1 (overinvestment firms) than Logistic Regression, which may be beneficial in identifying and addressing potential overinvestment cases.

Based on the results, the Random Forest algorithm appears to be the better choice for classifying overinvestment firms. However, it is essential to consider other factors, such as interpretability, computation time, and ease of implementation. It is also crucial to perform further evaluation using techniques like cross-validation and testing on different datasets to ensure the chosen algorithm's robustness. The importance of each predictor variable in the Random Forest model was also analyzed. The feature importance analysis indicated that free cash flow was the most important independent variable, followed by growth opportunity, ROE, and management confidence. These results suggest that firms with higher growth opportunities, better profitability, more significant free cash flow, and higher management confidence are more likely to be classified as overinvestment. Our study has important implications for researchers and practitioners interested in understanding the factors that contribute to firms being classified as overinvestment or not. The use of machine learning models can provide valuable insights for financial decision-making of firms.

Previous studies on overinvestment have used various methods, such as degree of Richardson used free cash flow as a measure of overinvestment and found that overinvestment is negatively related to future profitability, Effect of debt and dividends on the relationship between investment overcapacity and performance, regression statistics [20,35,58]. However, none of these studies applied machine learning techniques, despite the increasing popularity of machine learning in financial research. Inheriting the previous methods, our study addresses this gap by introducing machine learning algorithms as a new approach to overinvestment classification that complements the existing literature by showing the potential of machine learning in improving the accuracy and efficiency of overinvestment classification.

This study has not confirmed previous research on using machine learning machine learning algorithms to make predictions in the financial field, it could nevertheless be argued include differences in sample characteristics such as industry characteristics, sample size

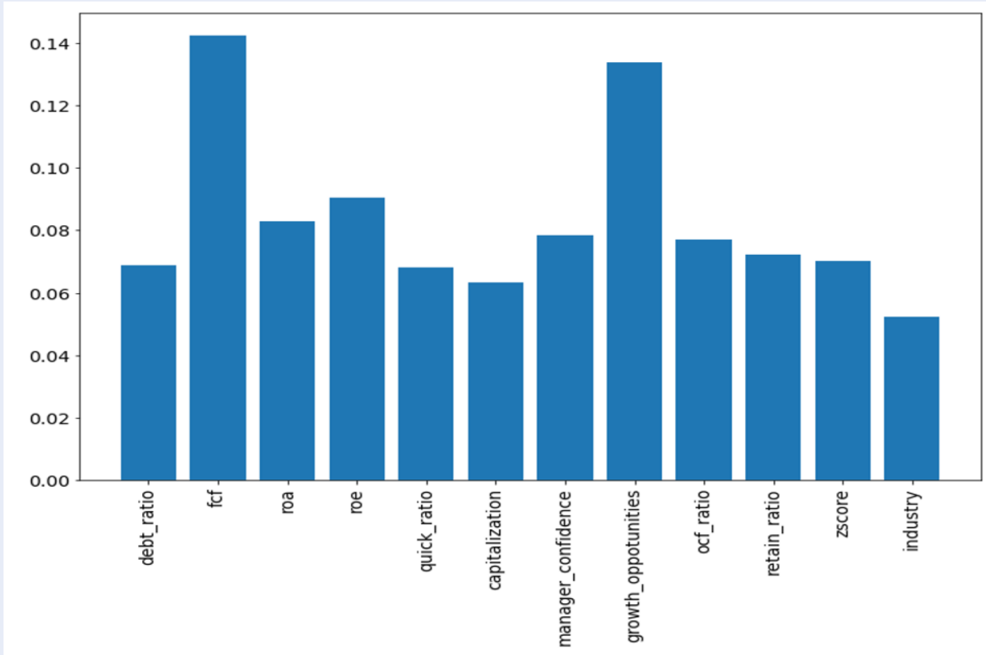

Figure 5: Feature Importances - Random Forest Model. Source: Author's calculation

or environment and geographical location. It serves to compare and identify algorithms suitable for each different environment. Özlem and Tan found that decision tree was the best performing model among multiple machine learning models, including multiple linear regression (MLR), K-nearest neighbors (KNN), support vector machine (SVM), and DT [39].

There could be several reasons for these differences in findings. One possible reason is the difference in the research environment or context. Each study may have used different datasets with variations in sample size, data quality, and industry characteristics, which can affect the performance of machine learning models. Additionally, the specific variables used in the machine learning models may differ across studies, leading to variations in the classification accuracy. Another possible reason for the differences is the choice of machine learning techniques and their parameter settings. Different studies may have used different algorithms, feature engineering techniques, and model hyperparameters, which can impact the performance of the models. The performance of machine learning models is also sensitive to the specific dataset and its characteristics, as well as the availability of data for model training and validation.

The finding that free cash flow was the most important independent variable in predicting overinvestment aligns with several prior studies that have identified FCF as a significant determinant of overinvestment. This is consistent with Richardson, Chen et al. and Jensen [2,20,24]. In addition, Growth Opportunity is equally important and also affects overinvestment, this substantiates previous findings in the literature Farooq et al [58]. In addition, Smith proposed the "free cash flow hypothesis" which suggests that firms with high levels of FCF are more likely to engage in overinvestment activities due to the availability of excess cash that may not be efficiently utilized for productive investments [59]. Similarly, Jensen argued that managers may have incentives to overinvest in order to pursue their own interests at the expense of shareholders, particularly when they have access to abundant internal funds such as FCF [2]. Furthermore, other studies have also found that FCF is positively correlated with overinvestment. For instance, studies by Lang et al., Opler et al., and Almeida & Campello have reported similar findings, suggesting that FCF has a significant impact on firms' overinvestment behavior [60–62]. Free cash flow is the most effective tool for predicting overinvestment because it measures a company's ability to generate what investors care about most, which is cash available to distribute to shareholders, creditors, and reinvest back into the business. Companies with high free cash flow have more resources to invest in new projects or acquisitions, which can lead to overinvestment. Additionally, free cash flow can be used to fund share buybacks or dividends, which can also contribute to overinvestment.

The study clearly achieved this objective by comparing the performance of six different classification algorithms, namely LR, SVM, DT, RF, NB, and KNN, in terms of accuracy, precision, recall, and F1 score. The findings in the study clearly highlight that Random Forest [13]. outperforms Logistic Regression [11]. in terms of precision and recall for classifying overinvestment companies (class 1), indicating that RF may be the most suitable algorithm for this particular classification problem. The study provides evidence that the performance of different algorithms may vary depending on the specific problem, and in the context of overinvestment classification, RF may be more effective than LR. Therefore, the results of the study are consistent with the stated objectives of comparing the performance of different classification algorithms and providing insights on the most suitable algorithm for classifying overinvestment companies. The final result of the Random Forest algorithm is aggregated from many decision trees, so the information from the trees will complement each other, leading to a model with low bias and low variance, or a model with high results prediction. The idea of aggregating decision trees of the Random Forest algorithm is similar to the idea of Crowd Intelligence proposed by Wu at al [40]. Crowd intelligence says that usually synthesizing information from a group is better than from a kernel. In the Random Forest algorithm, it also synthesizes information from a group of decision trees and the results are better than the Decision Tree algorithm with 1 decision tree.

Random forest processing involves aggregating diversity of opinion, partitioning, decentralization, and aggregation to produce classification results. The randomness in the process helps Random Forest come to the best conclusion because the random sample selected is representative of the population and many different points of view. Each tree is built off of a randomly selected subset of the data and predictors. Therefore, each tree is built based on completely different information from every other tree. By utilizing different training sets and randomly selecting the subset of predictors at each split, the algorithm ensures that each tree is independent from every other. This actually has the effect of decorrelating the trees. Decentralization is inherent in the fact that each tree is built with different training data and different predictors to choose from at each split. The last step of the algorithm is to take the mode (classification). Some mechanism exists to turn private judgments into a collective decision.

One of the strengths of our study is that we leverage the power of machine learning algorithms, which are known for their ability to process large volumes of data and uncover complex patterns that may not be easily discernible through traditional approaches. By harnessing the capabilities of machine learning, we have achieved improved accuracy in overinvestment classification, which is a significant advancement in the field of overinvestment research. Furthermore, our study aligns with the current trend of utilizing machine learning in finance research and processing financial data. Machine learning has gained significant traction in recent years due to its potential to extract valuable insights from large and complex datasets. By applying machine learning techniques in the context of overinvestment classification, our study contributes to the growing body of literature on the use of machine learning in finance, showcasing its applicability and effectiveness in solving financial decision-making problems.

Despite the significant contributions of our study, there are certain limitations that warrant further investigation. First, our study focuses on a specific context and the generalizability of our findings to other regions may be limited. Further research could explore the application of machine learning in overinvestment classification in different contexts to validate the robustness of our outcomes. Second, our study employs historical data, and the dynamic nature of financial markets may affect the performance of machine learning algorithms in real-time scenarios. Future research could explore the real-time applicability of machine learning in overinvestment classification using up-to-date data. However, it is essential to acknowledge that our study has limitations. The sample size used in this study was relatively small, and the predictor variables used may not be exhaustive or representative of all factors contributing to overinvestment. Future research could explore the use of additional variables or consider different classification methods to further investigate the phenomenon of overinvestment. Besides that, one limitation is the predictor variables used in our study may not be exhaustive, and there may be other factors that contribute to overinvestment that were not included in our analysis. Future research could explore the use of additional variables or consider different classification methods to further investigate the phenomenon of overinvestment and provide a more comprehensive understanding of the factors at play. Despite these limitations, our study suggests that Random Forest is an effective model for classifying firms as overinvestment or not.

Overall, our results suggest that Random Forest is an effective model for the classification of firms as overinvesting or not. The feature important analysis also provides insights into the factors contributing to overinvestment, which can inform financial decision-making for firms. In terms of addressing overinvestment, the following recommendations can be made: To ensure that resources are being used efficiently and to avoid overinvestment, companies should regularly monitor their capital allocation strategies and investment decisions. This can involve conducting detailed analyses of their cash flow statements, identifying areas where resources may be underutilized or misallocated, and implementing measures to address these issues. By doing so, companies can optimize their resource allocation and avoid the negative consequences of overinvestment, such as reduced profitability and decreased shareholder value.

Investors and analysts can also use the classification results from our study to identify firms with potential overinvestment and exercise caution when making investment decisions. By paying attention to the classification results, investors can avoid investing in firms that have a higher risk of overinvestment and instead focus on investing in firms that are more likely to provide sustainable returns over the long term. Furthermore, investors can use the results to guide their engagement with companies, encouraging them to prioritize efficient resource allocation and avoid overinvestment.

For firms that are identified as having overinvestment, it is recommended that they conduct thorough internal reviews and reassess their investment strategies. This can involve reviewing their capital allocation policies, identifying areas where resources are being misallocated, and implementing changes to optimize their resource utilization. By doing so, firms can improve their financial performance, increase shareholder value, and position themselves for long-term success. Overall, it is essential for companies, investors, and analysts to be aware of the risks associated with overinvestment and take proactive steps to avoid it.

CONCLUSION

This research paper contributes to the field of finance by assessing the implementation of machine learning algorithms for overinvestment classification in listed firms on the Vietnam stock exchange market. The study adds to the existing literature on overinvestment and presents a practical tool for companies to detect overinvestment and establish management strategies. The results of the study highlight the importance of using machine learning algorithms to identify overinvestment, a complex financial problem, and provide insights for financial decision-making. Our study aimed to compare the performance of six classification algorithms in classifying overinvestment companies and provide insights for financial decision-making. The results of our study indicate that while logistic regression[11]. and random forest[13]. perform similarly in terms of average accuracy, there are some differences in their performance concerning precision and recall for classifying overinvestment companies. Based on our findings, Random Forest[13]. appears to be the most suitable algorithm for classifying overinvestment companies, as it demonstrated slightly higher precision and recall compared to Logistic Regression[11]. for class 1 (overinvestment firms). Our study's findings provide further support to the existing literature, reinforcing the notion that FCF plays a crucial role in driving overinvestment behavior among firms. The consistency of our results with prior research adds to the robustness and validity of our study.

Our study has important implications for researchers and practitioners interested in understanding the factors that improve firms being classified as overinvestment or not. The use of machine learning models, specifically Random Forest in this case, can provide valuable insights into the financial decision-making of firms. Regular monitoring of capital allocation strategies and investment decisions proposed for companies to ensure efficient resource utilization and avoid overinvestment. Investors and analysts can also utilize the classification results from our study to identify firms with potential overinvestment and exercise caution in their investment decisions. Firms recognized as having overinvestment can conduct thorough internal reviews and reassess their investment strategies to maximize returns and reduce inefficiencies.

However, the study has certain limitations. The focus is primarily on firms listed on the Vietnam stock exchange and their attributes from 2010 to 2020. Future research can extend the timeframe and include non-financial variables such as Organizational Culture and Innovation and Technology. Additionally, industry classification can be included to examine companies on a sectoral basis in future research. In addition to expanding the time period, the number of countries studied can also be increased. Researchers can conduct a cross-country analysis by categorizing overinvestment in developed and emerging markets to identify if there are variations in the extent of overinvestment across markets. It is important to note that further evaluation using techniques like cross-validation and testing on different datasets is necessary to ensure the robustness of the chosen algorithm. Additionally, other factors such as interpretability, computation time, and ease of implementation should also be considered when selecting a suitable algorithm for a specific problem. Furthermore, researchers can incorporate this information to improve regression models and explore the overinvestment tendencies of companies.

In conclusion, our study contributes to the literature on overinvestment by comparing the performance of different classification algorithms and providing insights on the most suitable algorithm for identifying overinvestment companies. Companies should regularly monitor their capital allocation strategies and investment decisions to ensure efficient use of resources and avoid overinvestment. Investors and analysts can use these classification results to identify firms with potential overinvestment and exercise caution when making investment decisions. Firms identified as having overinvestment can conduct thorough internal reviews and reassess their investment strategies, focusing on maximizing returns and reducing inefficiencies. The findings have practical implications for financial decision-makers and highlight the value of machine learning approaches in addressing complex financial problems. Future research can build on our outcomes and explore other machine learning techniques or incorporate additional variables to enhance the accuracy of overinvestment classification models.

FUNDING

The research is funded by the University of Economics and Law, Vietnam National University, Ho Chi Minh City, Vietnam.

ABBREVIATIONS

AP Agency Problems
AUC Area Under Curve
CSR Corporate Social Responsibility
DT Decision Tree
FC Financial Constraints
FCF Free cash flow
GO Growth Opportunities
HNX Hanoi Stock Exchange
HSX Ho Chi Minh City Stock Exchange
IC Industry Characteristics
KNN K-Nearest Neighbor
LR Logistic Regression
MC Manager Confidence
ML Machine learning
MLNN Multilayer Neural Network
MLR Multiple linear regression
MO Managerial Overconfidence
NB Naive Bayes
PF Profitability
RF Random Forest
ROA Return on assets
ROE Return on equity
SHAP Shaplely Additive Explanations
SOF Size of the Firm
SVM Support Vector Machine
XGB Extreme Gradient Boosting Algorithm

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest

AUTHOR CONTRIBUTIONS

Phan Huy Tam: Analyzing and interpreting data, provide technical support, Reviewing and providing feedback on the manuscript. Ngo Dinh Linh Tram: Abstract, Introduction, Methodology, Result. Nguyen Thi Ngoc Anh: Literature Review, Methodology, Result. Nguyen Quoc Trong Nghia: Methodology, Result, Conclusion. Hoang Thao Linh:Methodology, Result, Conclusion. Trinh Van Thanh: Reference, Methodology, Result.

APPENDIX

Figure 6

REFERENCES

1. Ross CL, Barringer J, Yang J. Megaregions: Literature review of the implications for US infrastructure investment and transportation planning. 2008;.
2. Jensen MC. Agency costs of free cash flow, corporate finance, and takeovers. The American economic review. 1986;76(2):323-9;.
3. Nguyen Trong N, Nguyen CT. Firm performance: the moderation impact of debt and dividend policies on overinvestment. Journal of Asian Business and Economic Studies. 2021;28(1):47-63;Available from: https://doi.org/10.1108/JABES-12-2019-0128.
4. Ding S, Knight J, Zhang X. Does China overinvest? Evidence from a panel of Chinese firms. The European Journal of Finance. 2019;25(6):489-507;Available from: https://doi.org/10.1080/1351847X.2016.1211546.
5. Liu X. Can Overconfident Executives Restrain Overinvestment? Modern Economy. 2017;8(8):1056-68;Available from: https://doi.org/10.4236/me.2017.88073.
6. Shi M. Overinvestment and corporate governance in energy listed companies: Evidence from China. Finance Research Letters. 2019;30:436-45;Available from: https://doi.org/10.1016/j.frl.2019.05.017.
7. Zhang D, Cao H, Dickinson DG, Kutan AM. Free cash flows and overinvestment: Further evidence from Chinese energy firms. Energy Economics. 2016;58:116-24;Available from: https://doi.org/10.1016/j.eneco.2016.06.018.

Classification	Not Overinvestment				Overinvestment					
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support	Accuracy	Macro Avg F1
Logistic Regression	0.66	0.74	0.70	257	0.70	0.62	0.66	256	0.68	0.68
K-Nearest Neighbors	0.61	0.67	0.64	257	0.63	0.57	0.60	256	0.62	0.62
Naive Bayes	0.66	0.13	0.21	257	0.52	0.93	0.66	256	0.53	0.44
Support Vector Machine	0.65	0.75	0.70	257	0.70	0.60	0.65	256	0.67	0.67
Decision Tree	0.63	0.64	0.63	257	0.63	0.62	0.63	256	0.63	0.63
Random Forest	0.68	0.74	0.70	257	0.71	0.64	0.67	256	0.69	0.69

Source: Author's calculation

Figure 6: Classification report

8. Childs PD, Mauer DC, Ott SH. Interactions of corporate financing and investment decisions: The effects of agency conflicts. Journal of financial economics. 2005;76(3):667-90;Available from: https://doi.org/10.1016/j.jfineco.2004.06.012.
9. Lyandres E, Zhdanov A. Underinvestment or overinvestment: the effects of financial leverage on investment. European Finance Association. 2005;33;.
10. Mauer DC, Sarkar S. Real options, agency conflicts, and optimal capital structure. Journal of banking & Finance. 2005;29(6):1405-28;Available from: https://doi.org/10.1016/j.jbankfin.2004.05.036.
11. Malmendier U, Tate G. Who makes acquisitions? CEO overconfidence and the market's reaction. Journal of financial Economics. 2008;89(1):20-43;Available from: https://doi.org/10.1016/j.jfineco.2007.07.002.
12. Tran K, Le H, Nguyen T, Nguyen D. Explainable Machine Learning for Financial Distress Prediction: Evidence from Vietnam. Data 2022, 7, 160. s Note: MDPI stays neutral with regard to jurisdictional claims in published …; 2022;Available from: https://doi.org/10.3390/data7110160.
13. Harford J, Li K. Decoupling CEO wealth and firm performance: The case of acquiring CEOs. The Journal of Finance. 2007;62(2):917-49;Available from: https://doi.org/10.1111/j.1540-6261.2007.01227.x.
14. Laopodis NT. Understanding investments: Theories and strategies: Routledge; 2020;Available from: https://doi.org/10.4324/9781003027478.
15. Shan W, Wang L. The Concept of "Investment": Treaty Definitions and Arbitration Interpretations. Handbook of International Investment Law and Policy. 2021:23-44;Available from: https://doi.org/10.1007/978-981-13-3615-7_125.
16. Boffo R, Patalano R. ESG investing: practices, progress and challenges. Éditions OCDE, Paris. 2020;.
17. Modigliani F, Miller MH. The cost of capital, corporation finance and the theory of investment. The American economic review. 1958;48(3):261-97;.
18. Jensen MC, Meckling WH. Theory of the firm: Managerial behavior, agency costs and ownership structure. Corporate governance: Gower; 2019. p. 77-132;.
19. Degryse H, De Jong A. Investment and internal finance: Asymmetric information or managerial discretion? International journal of industrial organization. 2006;24(1):125-47;Available from: https://doi.org/10.1016/j.ijindorg.2005.03.006.
20. Richardson S. Over-investment of free cash flow. Review of accounting studies. 2006;11:159-89;Available from: https://doi.org/10.1007/s11142-006-9012-1.
21. Rajan RG, Zingales L. What do we know about capital structure? Some evidence from international data. The journal of Finance. 1995;50(5):1421-60;Available from: https://doi.org/10.1111/j.1540-6261.1995.tb05184.x.
22. Damodaran A. Financing innovations and capital structure choices. Journal of Applied Corporate Finance. 1999;12(1):28-39;Available from: https://doi.org/10.1111/j.1745-6622.1999.tb00657.x.
23. Myers SC, Majluf NS. Corporate financing and investment decisions when firms have information that investors do not have. Journal of financial economics. 1984;13(2):187-221;Available from: https://doi.org/10.1016/0304-405X(84)90023-0.
24. Chen F, Hope O-K, Li Q, Wang X. Financial reporting quality and investment efficiency of private firms in emerging markets. The accounting review. 2011;86(4):1255-88;Available from: https://doi.org/10.2308/accr-10040.
25. Jensen M. Eclipse of the modem corporation. Harvard Business Review. 1989;67(6);.
26. Jensen MC. The modern industrial revolution, exit, and the failure of internal control systems. the Journal of Finance. 1993;48(3):831-80;Available from: https://doi.org/10.1111/j.1540-6261.1993.tb04022.x.
27. Akerlof GA. The market for "lemons": Quality uncertainty and the market mechanism. Uncertainty in economics: Elsevier; 1978. p. 235-51;Available from: https://doi.org/10.1016/B978-0-12-214850-7.50022-X.
28. Cleary S. The relationship between firm investment and financial status. The journal of finance. 1999;54(2):673-92;Available from: https://doi.org/10.1111/0022-1082.00121.
29. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. science. 1974;185(4157):1124-31;PMID: 17835457. Available from: https://doi.org/10.1126/science.185.4157.1124.
30. Barberis N, Thaler R. A survey of behavioral finance. Handbook of the Economics of Finance. 2003;1:1053-128;Available from: https://doi.org/10.1016/S1574-0102(03)01027-6.
31. Shefrin H, Statman M. The disposition to sell winners too early and ride losers too long: Theory and evidence. The Journal of finance. 1985;40(3):777-90;Available from: https://doi.org/10.1111/j.1540-6261.1985.tb05002.x.
32. Hillman AJ, Withers MC, Collins BJ. Resource dependence theory: A review. Journal of management. 2009;35(6):1404-27;Available from: https://doi.org/10.1177/0149206309343469.
33. Pfeffer J, Salancik G. The external control of organizations: A resource dependence perspective: Stanford Business Books. Stanford; 2003;.
34. Hao X, Wang Y, Peng S. The effect of debt structure on overinvestment-Based on Chinese real estate listed compa-

nies. Journal of Asia Entrepreneurship and Sustainability. 2020;16(3):122-44;.

35. Nghia NT, Le Khang T, Thanh NC, editors. The Moderation Effect of Debt and Dividend on the Overinvestment-Performance Relationship. Beyond Traditional Probabilistic Methods in Economics 2; 2019: Springer;Available from: https://doi.org/10.1007/978-3-030-04200-4_81.
36. Twumasi C, Twumasi J. Machine learning algorithms for forecasting and backcasting blood demand data with missing values and outliers: A study of Tema General Hospital of Ghana. International Journal of Forecasting. 2022;38(3):1258-77;Available from: https://doi.org/10.1016/j.ijforecast.2021.10.008.
37. Biddle GC, Hilary G, Verdi RS. How does financial reporting quality relate to investment efficiency? Journal of accounting and economics. 2009;48(2-3):112-31;Available from: https://doi.org/10.1016/j.jacceco.2009.09.001.
38. Lakhal N, Guizani A, Sghaier A, El mine Abdelli M, Slimene IB, editors. The impact of CSR performance on Efficiency of Investments using Machine Learning. International conference on business and finance 2021; 2021;.
39. Özlem Ş, Tan OF. Predicting cash holdings using supervised machine learning algorithms. Financial Innovation. 2022;8(1):1-19;PMID: 35601748. Available from: https://doi.org/10.1186/s40854-022-00351-8.
40. Wu H-C, Chen J-H, Wang P-W. Cash holdings prediction using decision tree algorithms and comparison with logistic regression model. Cybernetics and Systems. 2021;52(8):689-704;Available from: https://doi.org/10.1080/01969722.2021.1976988.
41. Moubariki Z, Beljadid L, Tirari MEH, Kaicer M, Thami ROH, editors. Enhancing cash management using machine learning. 2019 1st international conference on smart systems and data science (ICSSD); 2019: IEEE;Available from: https://doi.org/10.1109/ICSSD47982.2019.9002731.
42. Bae JK. Forecasting Decisions on Dividend Policy of South Korea Companies Listed in the Korea Exchange Market Based on Support Vector Machines. J Convergence Inf Technol. 2010;5(8):186-94;Available from: https://doi.org/10.4156/jcit.vol5.issue8.20.
43. Gholamzadeh M, Faghani M, Pifeh A. Implementing machine learning methods in the prediction of the financial constraints of the companies listed on Tehran's stock exchange. International Journal of Finance & Managerial Accounting. 2021;6(20):131-44;.
44. Mousa GA, Elamir EA, Hussainey K. Using machine learning methods to predict financial performance: Does disclosure tone matter? International Journal of Disclosure and Governance. 2022:1-20;Available from: https://doi.org/10.1057/s41310-021-00129-x.
45. Thissen U, Van Brakel R, De Weijer A, Melssen W, Buydens L. Using support vector machines for time series prediction. Chemometrics and intelligent laboratory systems. 2003;69(1-2):35-49;Available from: https://doi.org/10.1016/S0169-7439(03)00111-4.
46. Grinblatt M, Keloharju M. Sensation seeking, overconfidence, and trading activity. The Journal of Finance. 2009;64(2):549-78;Available from: https://doi.org/10.1111/j.1540-6261.2009.01443.x.
47. Altman EI, Hartzell J, Peck M. A scoring system for emerging market corporate debt. Salomon Brothers. 1995;15(May);.
48. Chung R, Firth M, Kim JB. FCF agency costs, earnings management, and investor monitoring. Corporate Ownership and Control. 2005;2(4):51-61;Available from: https://doi.org/10.22495/cocv2i4p4.
49. Titman S, Wei KJ, Xie F. Capital investments and stock returns. Journal of financial and Quantitative Analysis. 2004;39(4):677-700;Available from: https://doi.org/10.1017/S0022109000003173.
50. Gompers PA, Ishii J, Metrick A. Extreme governance: An analysis of dual-class firms in the United States. The Review of Financial Studies. 2010;23(3):1051-88;Available from: https://doi.org/10.1093/rfs/hhp024.
51. Graham JR, Harvey CR. The theory and practice of corporate finance: Evidence from the field. Journal of financial economics. 2001;60(2-3):187-243;Available from: https://doi.org/10.1016/S0304-405X(01)00044-7.
52. Fama EF, French KR. The capital asset pricing model: Theory and evidence. Journal of economic perspectives. 2004;18(3):25-46;Available from: https://doi.org/10.1257/0895330042162430.
53. Gervais S, Heaton JB, Odean T. Overconfidence, compensation contracts, and capital budgeting. The Journal of Finance. 2011;66(5):1735-77;Available from: https://doi.org/10.1111/j.1540-6261.2011.01686.x.
54. Stein E. Without good reason: The rationality debate in philosophy and cognitive science: Clarendon Press; 1996;.
55. Miller MH, Modigliani F. Dividend policy, growth, and the valuation of shares. the Journal of Business. 1961;34(4):411-33;Available from: https://doi.org/10.1086/294442.
56. Lins KV, Servaes H, Tufano P. What drives corporate liquidity? An international survey of cash holdings and lines of credit. Journal of financial economics. 2010;98(1):160-76;Available from: https://doi.org/10.1016/j.jfineco.2010.04.006.
57. Adyani LR, Sampurno RD. Analisis faktor-faktor yang mempengaruhi profitabilitas (ROA). Jurnal Dinamika Ekonomi Pembangunan. 2011;7(1):46-54;.
58. Farooq S, Ahmed S, Saleem K. Overinvestment, growth opportunities and firm performance: Evidence from Singapore stock market. Corporate Ownership and Control. 2015;12(3):454-67;Available from: https://doi.org/10.22495/cocv12i3c4p6.
59. Smith RL, Kim J-H. The combined effects of free cash flow and financial slack on bidder and target stock returns. Journal of business. 1994:281-310;Available from: https://doi.org/10.1086/296633.
60. Almeida H, Campello M. Financial constraints, asset tangibility, and corporate investment. The Review of Financial Studies. 2007;20(5):1429-60;Available from: https://doi.org/10.1093/rfs/hhm019.
61. Lang LH, Stulz R, Walkling RA. A test of the free cash flow hypothesis: The case of bidder returns. Journal of financial economics. 1991;29(2):315-35;Available from: https://doi.org/10.1016/0304-405X(91)90005-5.
62. Opler T, Pinkowitz L, Stulz R, Williamson R. The determinants and implications of corporate cash holdings. Journal of financial economics. 1999;52(1):3-46;Available from: https://doi.org/10.1016/S0304-405X(99)00003-3.

Tạp chí Phát triển Khoa học và Công nghệ – Kinh tế-Luật và Quản lý 2023, 7(4):4814-4833

Open Access Full Text Article

Bài nghiên cứu

Ứng dụng học máy trong việc phân loại đầu tư quá mức: Bằng chứng từ các công ty niêm yết trên thị trường chứng khoán Việt Nam

Phan Huy Tâm[1,2,*], Ngô Đình Linh Trâm[1,2], Nguyễn Thị Ngọc Ánh[1,2], Nguyễn Quốc Trọng Nghĩa[1,2], Hoàng Thảo Linh[1,2], Trịnh Văn Thanh[1,2]

TÓM TẮT

Các nghiên cứu thực nghiệm đã liên tục chứng minh rằng cả đầu tư quá mức và đầu tư dưới mức đều có ảnh hưởng tiêu cực đến hiệu quả tổng thể của hoạt động kinh doanh, cho thấy tầm quan trọng của việc hiểu và giải quyết những hiện tượng này trong lĩnh vực nghiên cứu học thuật. Do đó, trong nghiên cứu này, chúng tôi mong muốn phát triển một mô hình học máy chính xác để xác định đầu tư quá mức ở các công ty niêm yết trên sàn chứng khoán HSX và HNX của Việt Nam. Chúng tôi quyết định tiến hành so sánh để xác định mô hình tối ưu nhất cho việc phân loại các công ty có đầu tư quá mức hay không, bao gồm Hồi quy Logistic (LR), K-Nearest Neighbor (KNN), Naive Bayes (NB), Máy Vector Hỗ trợ (SVM), Cây quyết định (DT) và Rừng ngẫu nhiên (RF). Sử dụng mẫu 658 công ty niêm yết phi tài chính tại Việt Nam từ năm 2011 đến 2021, kết quả của chúng tôi cho thấy biến dự báo quan trọng nhất là "FCF" (dòng tiền tự do), với giá trị quan trọng là 0.14. Mặc dù cả hai thuật toán hồi quy logistic (LR) và rừng ngẫu nhiên (RD) đều cho thấy độ chính xác cao trong việc xác định các công ty có đầu tư quá mức, thuật toán Rừng ngẫu nhiên lại thể hiện độ chính xác và độ nhạy cao hơn cho lớp 1 (các công ty đầu tư quá mức) so với Hồi quy Logistic. Ngược lại, hiệu suất độ chính xác của bốn mô hình (NB, KNN, DT và SVM) thấp, dao động từ 0.53 đến 0.67. Ở cấp độ vi mô, nghiên cứu này có thể giúp doanh nghiệp hiểu biết về hiệu suất tài chính của mình, xác định các lĩnh vực cần cải thiện, và áp dụng các biện pháp chủ động để tránh khó khăn tài chính và cải thiện lợi nhuận bằng cách xác định các trường hợp có thể của đầu tư quá mức. Tổng thể, nghiên cứu cung cấp một đóng góp có giá trị cho lĩnh vực phân tích tài chính sử dụng các kỹ thuật học máy. Chúng tôi tin tưởng rằng kết quả của nghiên cứu này sẽ là một tài liệu tham khảo học thuật quan trọng cho các cuộc điều tra tương lai trong lĩnh vực và khám phá các biến dự báo quan trọng khác của đầu tư quá mức ở Việt Nam và các thị trường mới nổi.

Từ khoá: Phân loại, Đầu tư quá mức, Học máy

[1]*Đại học Kinh tế-Luật, Thành phố Hồ Chí Minh, Việt Nam*

[2]*Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam*

Liên hệ

Phan Huy Tâm, Đại học Kinh tế-Luật, Thành phố Hồ Chí Minh, Việt Nam

Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

Email: tamph@uel.edu.vn

Lịch sử

- Ngày nhận: 13-6-2023
- Ngày chấp nhận: 14-12-2023
- Ngày đăng: 31-12-2023

DOI : https://doi.org/10.32508/stdjelm.v7i4.1269

Trích dẫn bài báo này: Tâm P H, Trâm N D L, Ánh N T N, Nghĩa N Q T, Linh H T, Thanh T V. **Ứng dụng học máy trong việc phân loại đầu tư quá mức: Bằng chứng từ các công ty niêm yết trên thị trường chứng khoán Việt Nam**. *Sci. Tech. Dev. J. - Eco. Law Manag.* 2023, 7(4):4814-4833.