VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

FINAL ESSAY

SUBJECT

Machine Learning and Artificial Intelligence in Finance

TOPIC

Predicting cash holdings using machine learning algorithms

STUDENT

Nguyen Thi Tuyet Ngan – K214051252

LECTURER

Master Phan Huy Tam

Ho Chi Minh City, June 16, 2024

Contents

List of Tables

List of Figures

Abstract

This research predicts the cash holding ratios of companies in Vietnam from 2018 to 2022. The research model is constructed with six influencing factors: company size (Size), net working capital (NWC), cash flow (CF), financial leverage (LEV), return on assets (ROA), and liquidity (LIQ). The main machine learning models used in the study include Random Forest, Gradient Boosting, Neural Network, and XGBoost. The research results indicate that among the applied models, Random Forest and Gradient Boosting provide the most accurate predictions of companies' cash holding ratios. In contrast, the Neural Network and XGBoost models do not achieve the same level of accuracy as the former two algorithms. Therefore, the study, with a large sample of companies in Vietnam, will provide a robust database for financial decision-making and corporate governance.

1. Introduction

Predicting the cash holdings of businesses is a critical issue in corporate finance, directly impacting financial management strategies and investment decisions. Cash holding ratios reflect not only the financial health of a company but also its ability to withstand business and market risks. However, accurately predicting this ratio poses significant challenges due to the volatility of economic factors and internal management decisions.

Currently, traditional methods such as linear regression and quantitative financial analysis have been applied to predict cash holding ratios, but these methods often face limitations in handling nonlinear models and the complexity of modern financial data. With the advancement of machine learning technology, particularly sophisticated algorithms like Random Forest (RF), Support Vector Machines (SVM), and neural networks, there is great potential for improving the accuracy of these predictions.

However, the selection and application of the appropriate machine learning algorithm for this specific problem have not been thoroughly researched and exploited. Previous studies often focus on one or a few specific models without a comprehensive and systematic comparison between different models. Furthermore, the variables influencing cash holding ratios have not been fully explored and scientifically identified.

Therefore, this research aims to evaluate and compare the effectiveness of various machine learning algorithms in predicting the cash holding ratios of businesses, while also identifying the key variables that affect this ratio. This study will provide a more comprehensive and systematic view of the application of machine learning algorithms in financial prediction, contributing to the improvement of current corporate financial management models.

The paper consists of seven sections, with the remainder of the research organized as follows. Section 1 introduces the topic. Section 2 discusses relevant theories and previous studies. Section 3 outlines the research methodology, including research models, variable measurements, and data sources. Section 4 presents the data processing procedures. Section 5 provides descriptive statistics and data visualization. Section 6 covers machine learning techniques and their results. Finally, Section 7 discusses the conclusions and recommendations based on the research findings.

2. Theoretical Background and Literature Review

2.1. Theories on Cash Holdings

2.1.1. Trade-Off Model

The trade-off model, as explored in the research by Ferreira and Vilela (2004), suggests that companies determine their optimal level of cash holdings by balancing the marginal costs and benefits associated with holding cash. Holding cash offers several advantages. Firstly, it lowers the risk of financial distress, serving as a safety net for unexpected losses or increased costs of external financing. Secondly, cash reserves enable companies to pursue optimal investment opportunities even when facing financial constraints. If external funding is expensive, firms might need to forgo profitable investment projects. Lastly, holding cash helps reduce the costs related to raising external funds and improves the liquidity of the company's assets. Thus, cash serves as a cushion between the company's resources and its capital utilization.

2.1.2. Pecking Order Theory

The pecking order theory, introduced by Myers and Majluf (1984), asserts that companies follow a specific hierarchy when financing their investments: they use retained earnings first, then debt, and finally equity. This order aims to minimize the costs arising from asymmetric information and other financial expenses. According to this theory, companies do not target a specific level of cash holdings. Instead, cash is used as a buffer between retained earnings and investment needs. Therefore, when the current operating cash flow is sufficient to fund new investments, companies will pay down debt and accumulate cash. If retained earnings are insufficient to cover investments, companies will utilize their cash reserves and turn to debt if necessary.

2.2.3. Free Cash Flow Theory

According to Jensen (1986), managers are motivated to accumulate cash to enhance their control over corporate assets and influence investment decisions. Holding substantial cash reserves allows managers to avoid seeking external financing, thereby circumventing the need to provide comprehensive details about company investment projects to the capital market. This discretion potentially enables managers to pursue inefficient investments that could adversely affect shareholder interests.

2.2. Theories on machine learning

2.2.1. Random Forest

Random Forest is a machine learning algorithm that builds multiple decision trees and combines their outputs to make final predictions. Each decision tree in the Random Forest is constructed independently using a random subset of the training data and a random subset of the features to minimize overfitting. This algorithm is suitable for prediction, classification, and regression tasks on both structured and unstructured data.

2.2.2. Gradient Boosting

Gradient Boosting is a machine learning technique where predictive models are built sequentially, with each model attempting to improve upon the errors of the previous one. This algorithm is a supervised learning method where the weights of each data sample are adjusted to minimize prediction errors along the gradient of the loss function. Gradient Boosting is commonly used for prediction tasks with structured data.

2.2.3. Neural Network

Neural Network is a machine learning model inspired by the human brain's functioning. It consists of a network of nodes organized into layers, where nodes hold weights and propagate information from the input layer through hidden layers to the output layer. The training process of a neural network involves adjusting these weights so the model can learn and generalize complex patterns from training data. Neural networks are widely used in image recognition, natural language processing, and time series prediction.

2.2.4. XGBoost

XGBoost (Extreme Gradient Boosting) is an optimized machine learning algorithm known for its speed and efficiency, especially in regression and classification prediction problems. XGBoost combines gradient boosting with specialized optimization techniques such as tree ensembles to improve prediction accuracy and reduce model errors. This algorithm is particularly suitable for large-scale, structured, and unstructured data and can be used for various machine learning purposes.

2.3. Literature Review

In recent years, machine learning algorithms have significantly advanced in the field of corporate finance. For example, Wu et al. (2021) utilized Decision Tree (DT) methods to predict cash holdings in the high-tech industry in Taiwan, employing J48, logistic model tree (LMT), random forest (RF), REP tree, simple CHART, extra tree, and BF tree. Their findings highlighted RF as the most accurate predictor among these DT methods. Moubariki et al. (2019) analyzed cash management in the public sector using DT, RF, and neural networks, demonstrating DT as the optimal prediction method. Gholamzadeh et al. (2021) predicted financial constraints for listed firms on the Tehran Stock Exchange with Gaussian process and radial neural network approaches, confirming the suitability of machine learning methods for financial constraint predictions. Key variables such as institutional ownership percentage, return on assets, financial leverage, operating cash flow to assets, and company value were identified as crucial in forecasting financial constraints. From a financial perspective, Saddour (2006) illustrates that cash reserves tend to increase with operational risks and growth prospects but decrease with higher financial leverage. Growing companies typically show a negative relationship between cash reserves and factors like company size, current assets, and short-term debts. Cash reserves increase with company size, investments, dividend distributions, and decrease with reliance on trade credit and expenditures in research and development. Drobetz and Grüninger (2007) find that tangible assets and company size have a negative association with cash holdings. However, dividend payouts, operational cash flows, and CEO dual roles exhibit a positive connection with cash reserves. Positive correlations are observed between cash holdings and factors such as cash flow, financial leverage, return on assets (ROA), and investments in fixed assets. Conversely, there is a negative correlation between net working capital and cash reserves. The opportunity for investment and company size positively impacts cash holdings (Ogundipe et al., 2012). Chen & Liu (2013) suggest that financially constrained firms tend to hold more cash. Companies with lower financial leverage, net working capital, and capital expenditures are more inclined to increase their cash reserves. Additionally, China's economic landscape differs from developed countries, with state-owned enterprises playing a dominant role in the manufacturing sector, heavily influenced by government policies. Finally, Kafayat et al. (2014) indicate that firm size, net working capital, and dividend payments are positively associated with cash holdings but negatively correlated with financial leverage and capital expenditures. Net working capital shows a positive correlation with dividend payments but a negative correlation with firm size, financial leverage, and capital expenditures. Different industries adopt varying approaches to cash management.

3. Methodology

3.1. Data

The dataset contains financial information from 2018 to 2022 sourced from FiinPro. This time span provides users with an in-depth view of the changes and developments of businesses over a 5-year period, which is crucial for analyzing financial trends and business performance.

The dataset includes all publicly listed companies across Vietnam's three major stock exchanges: HOSE (Ho Chi Minh City Stock Exchange), HNX (Hanoi Stock Exchange), and UPCOM (Unlisted Public Company Market). This ensures comprehensive coverage of the entire Vietnamese stock market, offering a holistic perspective on the financial activities of listed companies.

3.2. Variable measurements

Size (SIZE): The size of a company influences various aspects of its business operations. Larger companies often benefit from better financial risk management and bargaining power, which help reduce costs and enhance competitiveness. Scale also reflects strong brand presence and the ability to attract investment capital, while providing resources for research and development investments to maintain industry leadership.

$$\text{Size} = \ln(\text{Total Assets})$$

Net Working Capital (NWC): Net working capital is a crucial financial metric that indicates a company's liquidity and operational efficiency. It represents the difference between a company's current assets (such as cash, accounts receivable, and inventory) and its current liabilities (including accounts payable and short-term debt). A positive NWC indicates that a company has enough current assets to cover its short-term liabilities, reflecting healthy liquidity. Conversely, a negative NWC suggests potential liquidity issues and may indicate inefficient management of working capital. Efficient management of NWC ensures that a company can meet its short-term financial obligations and sustain smooth operational cycles.

$$\text{NWC} = (\text{Current Assets} - \text{Current Liabilities}) / \text{Net Assets}$$

Cash Flow (CF): Cash flow impacts a company by providing insights into its profitability and financial health. It influences the company's ability to make spending decisions, investments, debt payments, and business development. Positive cash flow enhances profitability and operational expansion capabilities, whereas negative cash flow can pose financial risks and limit growth opportunities.

$$\text{CF} = \text{Operating Cash Flow} / \text{Total Assets}$$

Financial leverage (LEV): Financial leverage measures the extent to which a company uses debt relative to its total assets in its business operations. This metric can significantly impact a company in various ways. Employing high financial leverage can enhance profitability and support expansionary investments, but it also introduces higher financial risk, especially during fluctuations in the financial markets. Conversely, companies with low leverage typically exhibit higher financial stability, rely less on debt financing, and have better risk management capabilities in challenging business conditions.

LEV = Debt / Total Assets

Return on Assets (ROA): ROA measures how effectively a company generates profit from its assets. ROA is calculated by dividing the company's after-tax profit by its total assets. A high ROA indicates that the company is achieving higher profitability relative to its asset base, demonstrating efficient asset utilization and effective management strategies. Conversely, a low ROA suggests inefficient asset management and lower profitability relative to the amount of assets invested. ROA is a key metric for investors, analysts, and managers to evaluate a company's financial performance and compare it against industry standards, providing insights into operational efficiency and profitability ratios.

ROA = Profit after tax / Total assets

Liquidity (LIQ): Liquidity measures a company's ability to meet its short-term financial obligations. High liquidity indicates that a company has sufficient cash or easily convertible assets to cover its short-term liabilities promptly. This enhances financial flexibility, reduces the risk of default or financial distress, and improves the company's ability to seize opportunities such as strategic investments or acquisitions. Conversely, low liquidity may indicate difficulty in meeting immediate obligations, potentially leading to increased borrowing costs or missed business opportunities. Therefore, maintaining adequate liquidity is crucial for a company's financial health and operational efficiency.

LIQ = Current assets / Current liabilities

Cash holding ratio (CASH): It directly impacts liquidity management by ensuring the company can meet short-term obligations promptly and effectively. Moreover, maintaining an optimal cash reserve enhances financial flexibility, mitigates risks, and supports strategic decision-making, such as seizing investment opportunities or navigating economic uncertainties. Overall, a balanced cash holding ratio is essential for fostering stability, minimizing borrowing costs, and enhancing investor confidence in the company's operational resilience and growth prospects.

CASH = Cash and Cash equivalents / Total assets

3.3 Model

Based on foundational theories and previous studies, the machine learning model in this paper is built with 6 independent variables: Size, NWC, CF, LEV, ROA, and LIQ. These 6 variables are used to predict CASH in the dataset. These independent variables were chosen based on their theoretical importance and empirical evidence demonstrating their impact on the amount of cash that businesses hold. Company size (Size) was chosen because larger companies often have easier access to capital markets, leading to lower cash reserves compared to smaller companies. Net working capital (NWC) is also an important variable, as high net working capital can reduce the need for holding cash since the company can easily convert current assets into cash. Cash flow (CF) is another factor, as strong operating cash flow helps the company feel more confident in reducing cash reserves, thanks to the ability to generate cash from business operations. Financial leverage (LEV) is also considered, as companies with high leverage may hold more cash to ensure debt repayment capabilities in emergency situations. Return on assets (ROA) is another important variable, as high ROA indicates good asset utilization efficiency, helping the company need less cash reserves due to the ability to reinvest profits. Finally, the liquidity (LIQ) of assets is also considered, as highly liquid assets can reduce the need for holding cash due to the ability to quickly convert into cash when needed. Thus, selecting these 6 variables based on both financial theory and empirical evidence will provide a more comprehensive and accurate view of the factors affecting a company's cash management.

4. Data processing

4.1. Missing value

Missing values are data points that are absent for a specific variable in a dataset. They can be represented in various ways, such as blank cells, null values, or special symbols like “NA” or “unknown.” These missing data points pose a significant challenge in data analysis and can lead to inaccurate or biased results. There are several common approaches to handling missing values in a dataset, such as imputing with mean, median, etc. In this study, all missing values will be removed entirely. A total of 869 rows containing "NA" were removed, accounting for approximately 10% of the total dataset. Prior to removing missing values, the dataset contained 8,407 rows; after removal, it contained 7,538 rows.

4.2. Duplicate

The presence of duplicate data in a dataset can have several effects on data analysis. Firstly, it can introduce bias into the analysis results by disproportionately inflating the importance of certain instances. This can lead to skewed estimates, misleading patterns, and inaccurate conclusions. Secondly, duplicate data can overemphasize certain patterns or relationships, making them appear more significant than they actually are. This can result in an inflated sense of correlation or predictive power. Lastly, duplicate data reduces the representativeness and diversity of the dataset, potentially compromising the generalizability of the findings. In this study, there is no duplicated data.

4.3. Outliers

Outliers can distort statistical measures such as the mean and standard deviation, leading to biased estimates of central tendency and variability. They can also impact the results of statistical tests, potentially leading to incorrect conclusions or misleading interpretations. Outliers can similarly affect the performance of machine learning models by exerting a pull on the estimated regression line or decision boundary towards them, thereby reducing the accuracy of predictions. In this dataset, outliers were identified using the z-score. After removing outliers, the dataset was reduced from 7538 rows to 7372 rows, indicating that 166 outliers were removed.

5. Data visualization

5.1. Descriptive analysis

Table 1. Descriptive analysis

	CASH	Size	NWC	CF	LEV	ROA	LIQ
count	7372	7372	7372	7372	7372	7372	7372
mean	0.079404	27.02736	0.286986	0.043104	0.621683	0.027448	3.136261
std	0.095179	1.549533	7.420806	0.216021	1.827435	0.379944	10.09121
min	0.000006	16.72152	-402.307	-9.72564	0.000622	-21.1431	0.001221
25%	0.017054	25.94339	0.127315	-0.01977	0.288987	0.006973	1.069027
50%	0.047095	26.96921	0.415752	0.040902	0.494351	0.033605	1.430754
75%	0.105818	28.04799	0.69614	0.120792	0.682789	0.074629	2.470095
max	0.855117	31.44205	249.9714	2.453826	79.28045	2.873531	408.7314

Source: Author

Firstly, when looking at the cash holding ratio of businesses, the average value is around 7.9%. This ratio indicates a relatively safe and liquid position of companies, reflecting their ability to meet short-term payment needs and handle emergencies without needing to sell assets or incur debt. This ratio may also imply that businesses are maintaining a reasonable cash reserve to balance between using cash for investment and retaining enough cash to ensure financial stability. The dataset also shows diversity in cash holdings among businesses, with some holding almost no cash while others have a significant portion of their assets in cash.

Regarding the size of the companies, the dataset shows diversity in their scales. The average net working capital is 28.69%, indicating that most businesses have positive net working capital, meaning their current assets exceed their current liabilities. This generally reflects good liquidity and the ability to meet short-term obligations without relying on debt or selling long-term assets.

Considering the cash flow of businesses, the average value is around 0.043, indicating that, on average, businesses have positive cash flow, although not very large. A positive average cash flow shows that businesses are generally generating enough cash from their operations to cover daily operating expenses and potentially have some surplus for investment or reserves.

Regarding the financial leverage of businesses, the average value is around 0.62, indicating that businesses are using a significant amount of debt in their capital structure. This level of leverage implies that, on average, companies finance about 62% of their assets with debt. This can reflect a strategy of using debt to take advantage of investment and growth opportunities. However, high leverage also comes with certain risks, including liquidity risk and the potential for difficulty in repaying debt if cash flow is unstable. An average leverage of 0.62 suggests that businesses are balancing the use of debt for growth with managing financial risk, but close monitoring is needed to avoid excessive debt that could lead to financial stress.

Regarding the return on assets (ROA) of businesses, the average value is about 0.027. This indicates that, on average, businesses generate 2.7% profit on their total assets. An average ROA of 0.027 suggests a modest level of profitability, reflecting that each dollar of assets generates only a small portion of profit. Although this ROA is not very high, it still indicates that businesses are generating profit from their assets, albeit with potential to improve asset utilization efficiency. A low ROA may also imply the need to reassess asset management strategies and seek opportunities to enhance operational efficiency to increase profitability in the future.

Regarding the liquidity of businesses, the average value is about 3.13. This indicates that, on average, businesses have the ability to cover their short-term obligations more than three times over. An average liquidity ratio of 3.13 reflects that businesses are maintaining a very high level of liquidity, sufficient to cover short-term debts and handle emergencies without financial difficulties. Such high liquidity indicates financial safety and good management of short-term assets, enabling businesses to maintain stable operations and flexibility in seizing new business opportunities. However, excessive liquidity may also suggest that businesses might be holding too much cash or other short-term assets, which could be invested in higher-return opportunities to optimize profits.

5.2. Univariate Analysis

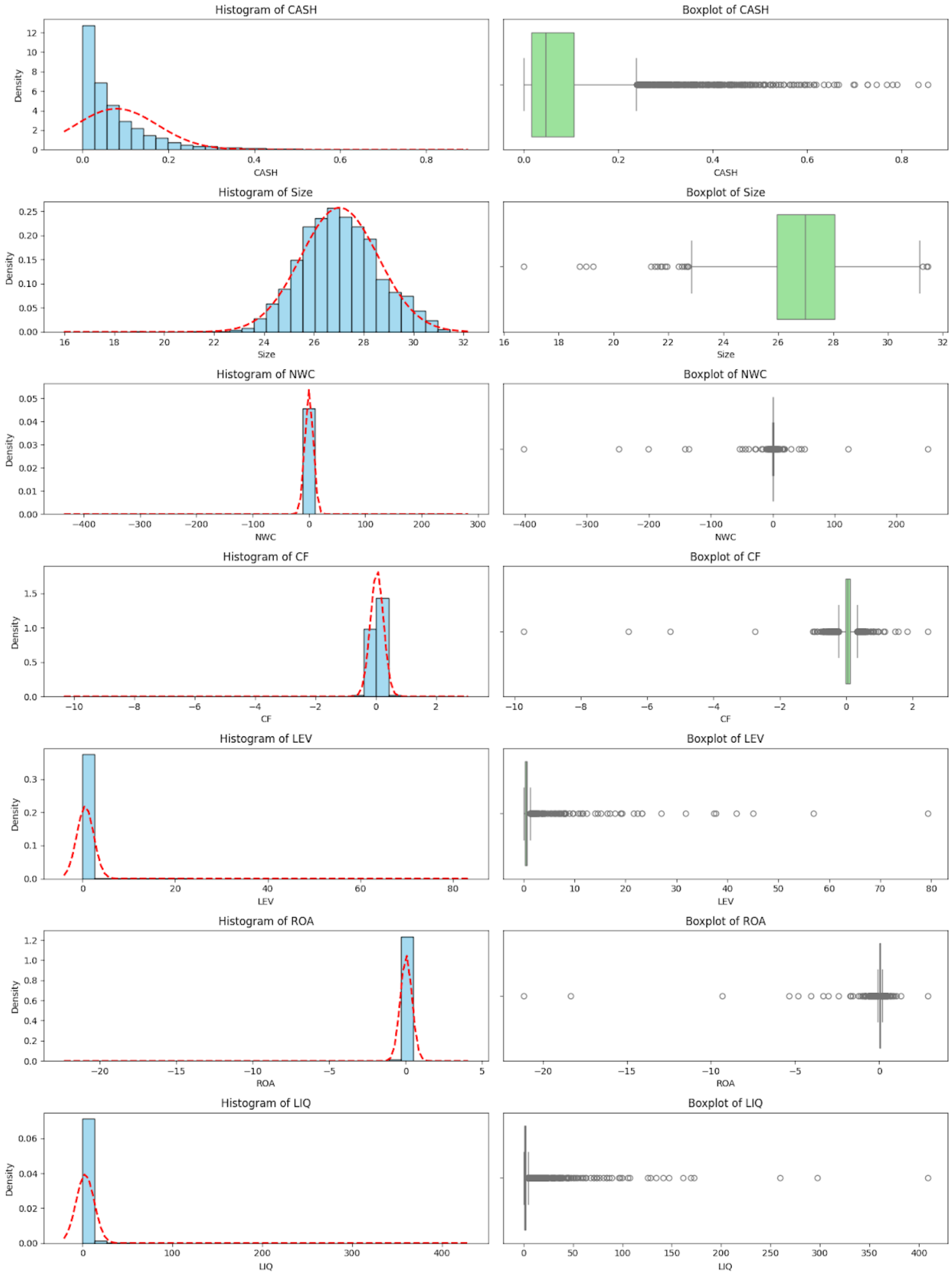

Figure 1. Univariate Analysis

When looking at the cash holdings ratio (CASH) of businesses, both the histogram and boxplot reveal a right-skewed distribution with numerous outliers extending beyond 0.3. This indicates that while most companies maintain a relatively low cash ratio, some hold significantly higher cash reserves, causing a skewness in the distribution. The presence of many outliers suggests that some businesses have cash holdings well above the majority of the sample, possibly due to specific business strategies, industry characteristics, or unusual financial situations. These companies may be maintaining high cash reserves as a precaution against risks, for future investments, or because they do not currently have attractive investment opportunities to utilize this cash more effectively. The skewed distribution and presence of outliers underscore the diversity in cash management strategies among businesses.

Examining the size (Size) of businesses, we observe a relatively normal distribution with values evenly distributed. However, there are still a few outliers below the average level. This indicates that while most companies are of comparable size, there are some significantly smaller businesses within the dataset.

In terms of net working capital (NWC), a sharp peak is evident with many outliers surrounding this peak area. This suggests that the majority of businesses have stable NWC levels within a narrow range, reflecting relatively consistent liquidity and working capital management. However, the presence of numerous outliers indicates that some companies have very high or very low NWC levels. This could be due to specific factors such as industry characteristics, business strategies, or fluctuations in financial management. Companies with extremely high NWC may be holding excessive short-term assets compared to short-term liabilities, while those with very low NWC may struggle to meet short-term financial obligations, posing liquidity risks. This highlights the importance of effective NWC management to maintain liquidity and financial stability.

Regarding cash flow (CF), the graph shows a slight left-skewed distribution with scattered outliers at very high negative values. This indicates that most businesses have positive cash flows, reflecting their ability to generate cash from operational activities consistently. However, the presence of significant negative outliers suggests that some companies are facing serious difficulties in generating cash flow, possibly due to poor operational efficiency, high costs, or other financial issues. These businesses require special attention and may need to adjust their business strategies or financial restructuring to improve cash flow and avoid liquidity risks.

When examining Return on Assets (ROA), the graph also shows a sharp peak, indicating that the majority of companies have ROA fluctuating around 0. However, there are some companies with ROA below -20, indicating significant challenges in generating profits from assets. This could be attributed to poor operational efficiency, suboptimal asset management, or high costs. These companies need to reconsider their asset management strategies and operational efficiency to improve profitability. The majority of companies with ROA near 0 indicate low profitability, requiring improvement measures to enhance asset utilization efficiency.

Looking at liquidity (LIQ), the graph reveals a right-skewed distribution with various values. Some companies have the highest LIQ index around 400, indicating their ability to quickly convert assets into cash when needed. This demonstrates financial flexibility and the ability to promptly meet short-term payment obligations.

5.3. Multivariate analysis

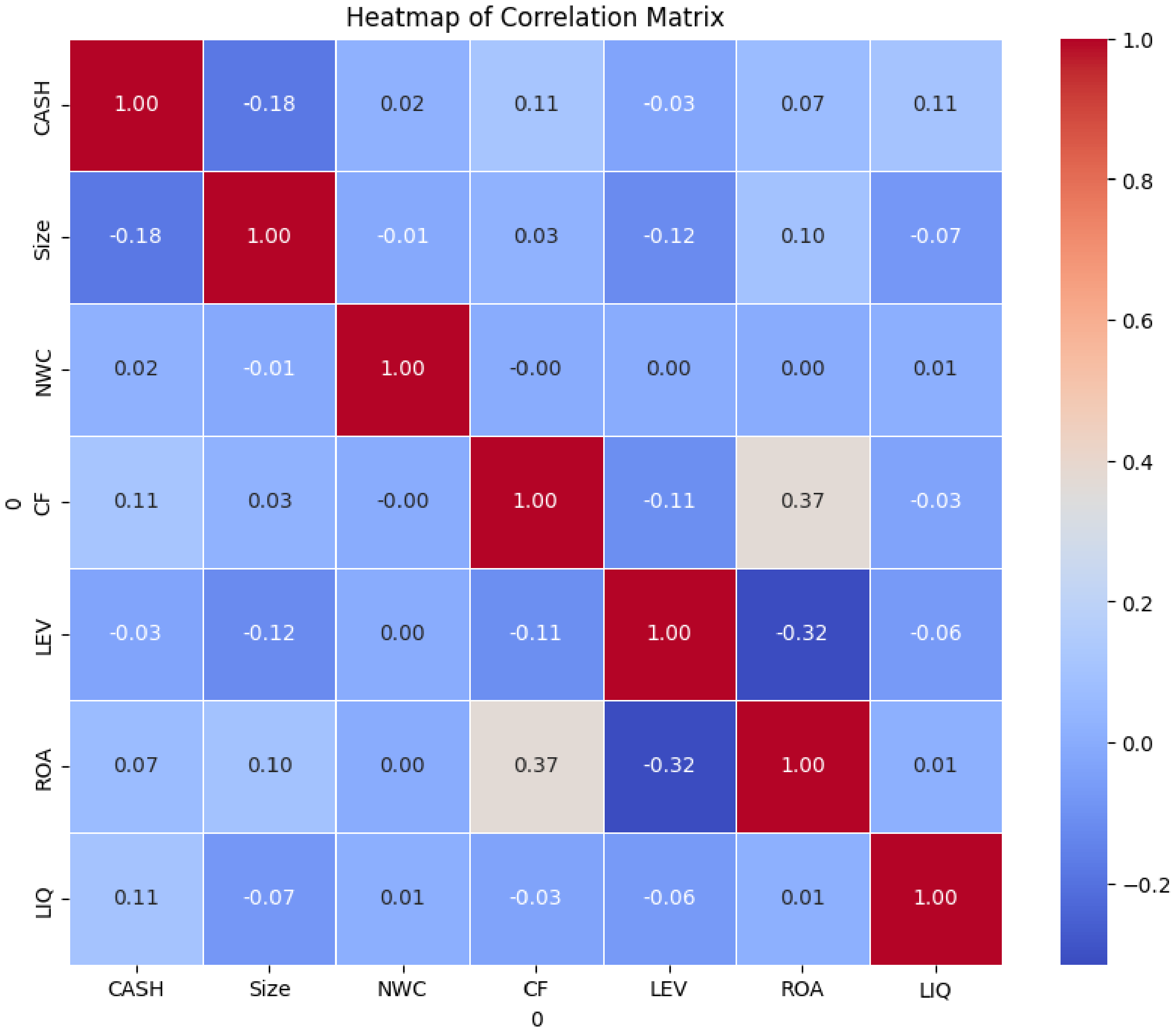

Figure 2. Multivariate analysis

From the heatmap, we can see various interactions between variables. The two variables with the highest correlation in the dataset are CF and ROA, followed by LEV and ROA. However, CF positively impacts ROA, indicating that an increase in cash flow from operating activities tends to accompany growth in profitability from assets. On the other hand, LEV negatively affects ROA, suggesting that high financial leverage may reduce the efficiency of generating profits from assets. Most variables in the dataset do not exhibit excessively strong or near-perfect correlations with each other, which is suitable for inclusion in machine learning experiments in future articles.

Regarding variables correlated with CASH, it is obvious that NWC has the least impact on CASH. This implies that the net working capital of short-term assets and liabilities does not strongly influence cash holdings in companies. Companies may maintain different balances between short-term assets and cash. The heatmap also show that SIZE has the strongest negative impact on CASH. This suggests that larger companies tend to hold proportionately less cash compared to smaller companies. This is because larger companies may have more investment opportunities or need funds for activities such as fixed investments or expansions, thereby reducing their cash holdings.

6. Result and discussion

Table 2. Machine learning result

Model	**R-squared (R2)**	**Mean Squared Error (MSE)**	**Mean Absolute Error (MAE)**	**Rootn Mean Squared Error (RMSE)**
Random Forest	0.190086	0.007426	0.057161	0.086177
Gradient Boosting	0.195125	0.007380	0.057191	0.085908
Neural Network	0.151997	0.007776	0.059367	0.088180
XGBoost	0.103185	0.008223	0.059160	0.090682

Source: Author

6.1. R-squared (R2)

R-squared is an indicator that measures how well the model explains the variability of the dependent variable. The higher the R2 value, the better the model is at explaining the data's variability.

- Gradient Boosting has the highest R2 value (0.195125), indicating it explains the data variability the best among the models.
- Random Forest has a very close R2 value (0.190086), making it a good choice as well.
- Neural Network and XGBoost have lower R2 values, with XGBoost having the lowest (0.103185), indicating poorer data explanation capabilities.

6.2. Mean Squared Error (MSE)

MSE measures the average squared difference between the predicted values and the actual values. The lower the MSE, the more accurate the model.

- Gradient Boosting has the lowest MSE (0.007380), showing the highest accuracy.
- Random Forest also has a low MSE (0.007426), very close to Gradient Boosting.
- Neural Network has a slightly higher MSE (0.007776).

- XGBoost has the highest MSE (0.008223), indicating the lowest accuracy.
-

6.3. Mean Absolute Error (MAE)

MAE measures the average absolute difference between the predicted values and the actual values. The lower the MAE, the more accurate the model.

- Random Forest has the lowest MAE (0.057161), indicating the highest accuracy.
- Gradient Boosting has a very close MAE (0.057191).
- Neural Network has a higher MAE (0.059367).
- XGBoost has a slightly higher MAE (0.059160) but still lower than Neural Network.

6.4. Root Mean Squared Error (RMSE)

RMSE is the square root of MSE and is commonly used to evaluate model accuracy. The lower the RMSE, the more accurate the model.

- Gradient Boosting has the lowest RMSE (0.085908), indicating the highest accuracy.
- Random Forest also has a low RMSE (0.086177), very close to Gradient Boosting.
- Neural Network has a higher RMSE (0.088180).
- XGBoost has the highest RMSE (0.090682), indicating the lowest accuracy.

6.5. Insights and Conclusions

- Gradient Boosting emerges as the best model among the ones evaluated, based on all four-performance metrics (R2, MSE, MAE, RMSE). This shows that it has the best ability to explain data variability and provides the most accurate predictions.
- Random Forest is also a strong candidate, nearly matching Gradient Boosting in most metrics. It can serve as a good alternative when needed.
- Neural Network performs lower than Gradient Boosting and Random Forest but remains acceptable in certain contexts.
- XGBoost shows the poorest performance among the models, especially in terms of data variability explanation (R2) and accuracy (MSE, RMSE).

6.6. Recommendations

6.6.1. Application

If the goal is to achieve the highest accuracy in predictions and explain the variability of the data effectively, Gradient Boosting is the top choice. This model has the highest R-squared (R2) value and the lowest error metrics such as MSE, MAE, and RMSE, indicating the most accurate and stable predictions. However, it is important to note that Gradient Boosting may require longer training times and more computational resources compared to some other models.

Random Forest is also an excellent choice, nearly matching Gradient Boosting in performance. With R2 and error metrics very close to those of Gradient Boosting, Random

Forest offers a robust alternative that can be easily parallelized, helping to reduce training time. This makes Random Forest an ideal choice when needing a powerful model but also need reasonable training times and computational resource requirements.

If the project involves handling complex and non-linear data, Neural Network might be the right choice. Although its performance is slightly lower than Gradient Boosting and Random Forest, it still provides acceptable accuracy. Neural Networks scale well with large datasets and can be optimized using specialized hardware like GPUs. However, deploying Neural Networks can be more complex and may require higher computational resources.

Lastly, if the project need a model with excellent optimization capabilities and efficient memory management, XGBoost is a worthy consideration. Despite having the lowest performance among the analyzed models, XGBoost stands out for its parallel processing capabilities and efficient memory management. XGBoost is a powerful tool in data science competitions and is easy to deploy in production environments.

6.6.2. Improving model

6.6.2.1. Gradient Boosting

- Hyperparameter Tuning: Use Grid Search or Random Search to find the best values for parameters such as learning rate, number of trees (n_estimators), and tree depth (max_depth).
- Reduce Overfitting: Use regularization techniques like L2 regularization and adjust parameters such as min_samples_split and min_samples_leaf.
- Speed Up Training: Employ parallelization techniques or leverage more powerful hardware like GPUs.

6.6.2.2. Random Forest

- Hyperparameter Tuning: Optimize parameters such as the number of trees (n_estimators), tree depth (max_depth), and the number of features considered when splitting (max_features).
- Dimensionality Reduction: Use feature selection techniques to reduce the number of input features, improving training efficiency.
- Reduce Complexity: Decrease the number of trees or adjust parameters to avoid overfitting.

6.6.2.3. Neural Network

- Optimize Model Architecture: Experiment with the number of hidden layers, the number of neurons per layer, and different activation functions.
- Hyperparameter Tuning: Optimize parameters like learning rate, batch size, and the number of epochs.

- Use Dropout and Batch Normalization: Apply Dropout to prevent overfitting and Batch Normalization to improve training speed and model stability.
- Data Augmentation: Use data augmentation techniques if working with image or text data.

6.6.2.4. XGBoost

- Hyperparameter Tuning: Use Grid Search, Random Search, or Bayesian Optimization to find optimal values for learning rate, max_depth, n_estimators, and subsample.
- Reduce Overfitting: Apply regularization techniques and adjust parameters like gamma and min_child_weight.
- Speed Up Training: Utilize GPUs for faster training, especially with large datasets.

7. Conclusion

In this study, the author applied several machine learning models to predict the cash holding ratio of enterprises, including Random Forest, Gradient Boosting, Neural Network, and XGBoost. By analyzing various performance metrics such as R-squared (R2), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), author was able to evaluate the effectiveness of each model in capturing the underlying patterns in the data and making accurate predictions. Our analysis revealed that Gradient Boosting emerged as the best-performing model among the ones evaluated, achieving the highest R-squared value of 0.195125, indicating a relatively better ability to explain the variance in the cash holding ratio compared to the other models. Additionally, it had the lowest MSE, MAE, and RMSE values, highlighting its superior accuracy and reliability in predictions. Random Forest also performed reasonably well, with an R-squared value of 0.190086 and slightly higher MSE, MAE, and RMSE values than Gradient Boosting, indicating that it is a viable model for predicting cash holding ratios, though slightly less accurate. Neural Network showed a lower performance, with an R-squared value of 0.151997, and its higher MSE, MAE, and RMSE values compared to Gradient Boosting and Random Forest suggest that it may not be as effective in capturing the complex relationships within the data for this particular task. XGBoost was the least effective model, with the lowest R-squared value of 0.103185 and the highest MSE, MAE, and RMSE values, indicating that it struggled the most with predicting the cash holding ratio accurately in this context. In conclusion, machine learning models, particularly Gradient Boosting and Random Forest, show promise in predicting the cash holding ratio of enterprises. These models can capture significant relationships within financial data and provide relatively accurate predictions, which can be valuable for financial analysts, investors, and decision-makers in understanding and anticipating corporate cash management behaviors. The insights gained from this study highlight the potential of machine learning in financial analysis and forecasting. Future research could explore further optimization of these models, inclusion of additional relevant features, and application of alternative machine learning techniques to enhance prediction accuracy and robustness. This study demonstrates the practical utility of machine learning in financial decision-making and sets the stage for more advanced applications in corporate finance.

REFERENCES

Chen, S., & Liu, S. (2013). Corporate Cash Holdings: Study of Chinese Firms.

Drobetz, W., & Grüninger, M. C. (2007). Corporate cash holdings: Evidence from Switzerland. *Financial Markets and Portfolio Management, 21*, 293-324.

Ferreira, M. A., & Vilela, A. S. (2004). Why do firms hold cash? Evidence from EMU countries. *European financial management, 10*(2), 295-319.

Gholamzadeh, M., Faghani, M., & Pifeh, A. (2021). 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, 6*(20), 131-144.

Jensens, M. C. (1986). Agency costs of free cash flow, corporate finance and takeovers. *American Economic Review, 76*(3), 323-329.

Kafayat, A., Rehman, K. U., & Farooq, M. (2014). Factors effecting corporate cash holding of non-financial firms in Pakistan. *Acta Universitatis Danubius. Œconomica, 10*(3).

Moubariki, Z., Beljadid, L., Tirari, M. E. H., Kaicer, M., & Thami, R. O. H. (2019). *Enhancing cash management using machine learning.* Paper presented at the 2019 1st international conference on smart systems and data science (ICSSD).

Myers, S. C., & Majluf, N. S. (1984). Corporate financing and investment decisions when firms have information that investors do not have. *Journal of financial Economics, 13*(2), 187-221.

Ogundipe, L. O., Ogundipe, S. E., & Ajao, S. K. (2012). Cash holding and firm characteristics: Evidence from Nigerian emerging market. *Journal of Business Economics and Finance, 1*(2), 45-58.

Özlem, Ş., & Tan, O. F. (2022). Predicting cash holdings using supervised machine learning algorithms. *Financial Innovation, 8*(1), 44.

Saddour, K. (2006). *The determinants and the value of cash holdings: Evidence from French firms*. Retrieved from

Wu, H.-C., Chen, J.-H., & Wang, P.-W. (2021). Cash holdings prediction using decision tree algorithms and comparison with logistic regression model. *Cybernetics and Systems, 52*(8), 689-704.