Science & Technology Development Journal – Economics - Law and Management 2024, 8(3):5494-5512

Research article

Open Access Full Text Article

Credit rating by clustering algorithm in the Vietnam Stock Exchange market

Tam Phan Huy[1,2,*], Thuy Chu Quang[1,2]

ABSTRACT

This study employs the K-means clustering algorithm to develop a corporate credit rating framework tailored to the Vietnamese market. By analyzing financial data from 568 non-financial firms listed on the Ho Chi Minh City Stock Exchange and the Hanoi Stock Exchange between 2019 and 2023, the research identifies vital financial indicators, including financial health ratios, management efficiency ratios, growth ratios, and dividend payout ratios. The K-means clustering model effectively categorizes these companies into six distinct clusters, each representing different levels of financial performance and credit risk. The clusters range from A+ (very low credit risk) to C (very high credit risk), providing a clear differentiation based on financial stability and operational efficiency. This systematic approach offers valuable insights for investors, managers, and government agencies, enhancing their ability to make informed decisions. Despite some limitations, such as reliance on historical data and sensitivity to initial cluster centroids, the K-means clustering model proves to be a robust starting point for assessing the creditworthiness of companies. This research contributes to the growing body of literature on machine learning applications in credit rating by demonstrating the superiority of clustering algorithms over traditional methods. It highlights how financial health and management efficiency indicators can be integrated into a data-driven framework to enhance credit risk assessment. The results suggest that the K-means clustering approach improves the accuracy of credit ratings and promotes transparency and efficiency in the financial market. Furthermore, the proposed framework can be a foundation for developing more sophisticated models, incorporating additional financial and non-financial variables. Future research could expand on this by integrating real-time data and exploring the impact of external economic factors on credit risk. By leveraging advanced machine learning techniques, this study paves the way for more reliable and comprehensive credit rating systems, ultimately supporting the stability and growth of financial markets in emerging economies like Vietnam.

Key words: K-Means, Credit Rating, Clustering, Vietnam

[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: tamphan.ntc@gmail.com

History

- Received: 17-5-2024
- Revised: 23-7-2024
- Accepted: 27-9-2024
- Published Online: 30-9-2024

DOI :

https://doi.org/10.32508/stdjelm.v8i3.1417

INTRODUCTION

In today's fiercely competitive market, all enterprises must utilize their resources efficiently. Companies with high financial leverage ratios often mobilize short-term capital through credit[1]. Some surveys also indicate that most businesses utilize credit[2]. In the banking sector, efficiency and productivity can be measured by the profits from loans extended to customers. As a result, the credit rating process, used to measure credit risk, has become an important issue in recent years[3]. With accurate business credit ratings, investors and financial institutions can make better investment and lending decisions. Additionally, credit ratings serve as a reference channel, increasing transparency in the market. Current credit rating methods and indicators often rely on financial statements and credit information of businesses[4]. The evaluation mainly focuses on borrowing situations, operational efficiency, debt collection ability, and asset utilization efficiency. Globally, credit ratings are usually performed by large and well-established credit rating agencies such as Standard & Poor's (S&P), Moody's, and Fitch Group. In Vietnam, many banks have developed and implemented their own internal credit scoring systems tailored to their specific needs and criteria. The Credit Information Centre (CIC) under the State Bank of Vietnam is a notable entity that provides credit information for customers who have borrowed from the commercial banking system. However, it does not perform business credit ratings. These internal systems and thallowsormation from CIC allow banks to better manage and assess the credit risk of their clients. Although domestic credit ratings have been implemented, they still face limitations in terms of data and tools, so only a few units perform this activity professionally and publicly. In academics, few published research works related to domestic business credit ratings have been published.

Cite this article : Huy T P, Quang T C. **Credit rating by clustering algorithm in the Vietnam Stock Exchange market**. *Sci. Tech. Dev. J. - Eco. Law Manag.* 2024; 8(3):5494-5512.

Moreover, the increasing risks in lending highlight the necessity for robust corporate credit ratings. Currently, most credit ratings are conducted internally by commercial banks, which means that external investors do not have access to comprehensive credit information. This lack of transparency can lead to uninformed investment decisions and increased financial instability. Therefore, establishing a standardized and publicly accessible credit rating system is crucial for providing investors with the information they need to make well-informed decisions, ultimately promoting a more stable and transparent financial market.

Thus, business credit ratings in Vietnam are a fascinating and practical topic in the financial field. Research on this subject will help us better understand the credit rating process, the factors affecting this process, and the methods for evaluating business credit rankings. Futhermore, with a reasonable credit rating basis, financial institutions can make decisions on granting loans or raising credit limits for businesses, and investors can gain a broader perspective on businesses' financial stability, enabling them to make informed investment decisions.

Currently, most business credit risk ratings are conducted by experts, but this method is not immune to human risks and disagreements among experts. Therefore, applying machine learning to the business credit rating process can help reduce workload, minimize disagreements and human risks, and increase evaluation accuracy. Through machine learning algorithms, we can perform calculations of financial indicators for thousands of businesses and visualize analyses automatically and quickly. In the long run, by combining theoretical foundations with computational power, financial institutions with clear data structures and fast information updates will be able to proactively assess business credit ratings in real time.

The objective of this research is to develop a corporate credit rating framework specifically tailored for the Vietnamese market, utilizing the K-means clustering algorithm. This framework leverages data from the financial statements of non-financial firms listed on the Ho Chi Minh City Stock Exchange and the Hanoi Stock Exchange from 2019 to 2023. By analyzing key financial indicators such as financial health ratios, management efficiency ratios, growth ratios, and dividend payout ratios, the framework aims to categorize companies into distinct clusters that reflect their credit risk levels. This systematic and data-driven approach will provide investors, lenders, and other stakeholders with a clearer understanding of these companies' creditworthiness and financial stability, thereby promoting more informed decision-making and contributing to a more transparent and efficient financial market.

LITERATURE REVIEW

Background theories

Credit rating through clustering is an innovative approach that combines both financial theories and machine learning techniques to assess the creditworthiness of businesses. The foundational financial theories related to this topic include the Modigliani-Miller theorem, the Trade-off theory, and the Pecking Order theory. These theories focus on firms' capital structure, the implications of their financing choices on overall credit risk, and the foundation of machine learning and clustering algorithms[5,6].

Modigliani-Miller Theory

The Modigliani-Miller (M-M) theorem, proposed by Franco Modigliani and Merton Miller in 1958, is an influential financial theory that lays the groundwork for understanding the relationship between a firm's capital structure and its credit risk. The M-M theorem posits that a firm's value is independent of its capital structure under certain assumptions such as no taxes, no bankruptcy costs, and perfect capital markets[5]. In other words, the choice between debt and equity financing does not impact a firm's overall value.

In the context of the research topic on credit rating by clustering, the Modigliani-Miller theorem is crucial in establishing the fundamental principles of capital structure and financing choices. Despite the theorem's assumptions not holding in the real world, it still provides a theoretical foundation that helps researchers and practitioners understand how different financing choices may affect a firm's credit risk. Researchers can identify relevant financial ratios and indicators that reflect a company's credit risk by examining the deviations from the M-M theorem's assumptions, such as the presence of taxes and bankruptcy costs. For example, higher leverage ratios, which represent the proportion of debt in a firm's capital structure, may indicate a higher credit risk due to the increased likelihood of financial distress and bankruptcy. These financial ratios can then be used as input features for clustering algorithms, which group companies with similar financial profiles and credit risk characteristics[7].

Trade-off theory

The Trade-off theory is a significant financial concept relevant to the research topic of clustering-based credit rating. This theory posits that companies strive to find an optimal balance between the advantages and disadvantages of debt financing to minimize their overall capital costs[8]. Debt financing's primary benefit stems from tax shields gained through interest payments, while the costs are associated with a heightened risk of financial distress and bankruptcy resulting from increased leverage.

In relation to credit rating through clustering, the Trade-off theory aids in pinpointing essential financial ratios and indicators that signify a company's credit risk. For example, a business with elevated leverage ratios may be more vulnerable to financial distress, while one with lower leverage ratios might possess a more stable capital structure and, consequently, reduced credit risk. Furthermore, the theory implies that companies with greater profitability and diminished bankruptcy risk will likely have superior credit ratings, as they can accommodate higher debt levels. By leveraging the insights offered by the Trade-off theory, researchers can select pertinent financial ratios, such as those about leverage, liquidity, and profitability, as input variables for clustering algorithms. Subsequently, these algorithms, including hierarchical clustering, k-means clustering, and density-based clustering, can be employed to categorize companies based on similar financial characteristics and credit risk profiles[7].

Pecking Order theory

The Pecking Order theory is another crucial financial concept relevant to the research topic of credit rating using clustering methods. This theory posits that firms prioritize their financing sources based on the information asymmetry and costs associated with each option, preferring internal financing first, followed by debt, and finally equity financing[9]. The rationale behind this order is that internal financing minimizes asymmetric information problems, while equity financing is considered the most expensive due to the adverse selection issue arising from information asymmetry.

In clustering-based credit rating, the Pecking Order theory helps identify vital financial ratios and indicators that reflect a company's credit risk. For instance, a firm that relies heavily on debt financing, as opposed to equity financing, may have a higher credit risk due to the potential for financial distress. On the other hand, companies with a greater reliance on internal financing and lower debt levels might exhibit lower credit risk. By incorporating the insights derived from the Pecking Order theory, researchers can choose relevant financial ratios, such as leverage, liquidity, and profitability ratios, as input features for clustering algorithms. These algorithms, including hierarchical clustering, k-means clustering, and density-based clustering, can then be utilized to group companies with similar financial characteristics and credit risk profiles[10].

In the context of credit rating by clustering, financial theories, such as the Modigliani-Miller theorem, Trade-off theory, and Pecking Order theory, can be employed to identify relevant financial ratios and indicators that reflect a company's credit risk. Key financial ratios include leverage ratios (e.g., debt-to-equity and debt-to-assets), liquidity ratios (e.g., current and quick ratios), profitability ratios (e.g., return on assets and return on equity), and efficiency ratios (e.g., asset turnover and inventory turnover)[11]. These financial ratios and indicators serve as the basis for clustering algorithms, which analyze patterns in large datasets to group companies with similar financial profiles and credit risk characteristics. Machine learning techniques, such as hierarchical clustering, k-means clustering, and density-based clustering, are particularly well-suited for this task.

Hierarchical clustering creates a tree-like structure, called a dendrogram, representing the hierarchical relationships between different clusters[12]. This approach allows for a more intuitive understanding of the relationships between clusters, which can be particularly helpful for credit rating purposes. K-means clustering is a popular centroid-based clustering algorithm that partitions the dataset into a predefined number of clusters by minimizing the within-cluster sum of squared distances[13]. This technique provides a simple and efficient way to group companies based on their financial ratios, thus facilitating comparisons of credit risk across different firms.

Combining these machine learning techniques and financial theories allows for a more comprehensive and data-driven approach to credit rating. This could potentially improve the accuracy and reliability of credit assessments and aid investors, lenders, and other stakeholders in their decision-making process.

The f oundation of machine learning

The foundation of machine learning lies in its ability to learn patterns and make predictions from data without explicit programming for each specific task.

Machine learning algorithms, such as clustering, classification, and regression, are designed to identify underlying structures in data, enabling more accurate and automated decision-making processes. In the context of credit rating, clustering algorithms like K-means play a crucial role in categorizing companies based on their financial profiles.

Clustering algorithms are unsupervised learning techniques that group data points based on similarity measures. K-means clustering, one of the most widely used clustering algorithms, partitions data into k distinct clusters by minimizing the within-cluster variance [13]. This algorithm operates iteratively, assigning each data point to the nearest cluster centroid and recalculating centroids until convergence. Various studies have demonstrated the effectiveness of K-means clustering in financial applications, including credit rating [14].

The advantage of using machine learning, mainly clustering algorithms, in credit rating, lies in its ability to handle large datasets and uncover complex patterns that may not be evident through traditional methods. Clustering algorithms can provide a more nuanced and data-driven credit risk assessment by analyzing a comprehensive set of financial indicators. This approach enhances the objectivity, consistency, and transparency of credit ratings, addressing many of the limitations associated with traditional expert-driven methods.

Incorporating machine learning into credit rating processes aligns with the broader trend of leveraging big data and advanced analytics in financial decision-making. As financial markets become increasingly complex, the ability to process and analyze large volumes of data efficiently is crucial for maintaining accurate and reliable credit assessments. Studies have shown that machine learning models, including clustering algorithms, outperform traditional statistical methods in various aspects of credit risk prediction [15,16].

By combining the theoretical foundations of capital structure with the analytical power of machine learning, credit rating through clustering represents a significant advancement in credit risk assessment. This innovative approach not only improves the accuracy and reliability of credit ratings but also provides valuable insights into the financial health and stability of businesses, ultimately supporting more informed investment and lending decisions.

Credit Rating Methods

One of the earliest and most prominent methods in this group of credit rating systems was developed by Moody's Investors Service in 1909 [17]. Moody's employed an alphabetical rating system to assess the debt repayment ability of businesses. In descending order, the ratings are Aaa, Aa, A, Baa, Ba, B, Caa, Ca, C, with Aaa being the safest and C being the most dangerous. This method uses the following primary criteria to evaluate a company's debt repayment ability:

- Debt and interest repayment capacity: This is the most crucial factor in assessing a company's ability to repay its debt. Moody's evaluates a company's capacity to repay its principal and interest based on its profitability, assets, and debt repayment history.
- Financial health: This criterion is assessed based on measurements of outstanding debt, net assets, profitability, and cash flow.
- Market and competition: Moody's assess the market in which a company operates, including its competitors, pricing power, and value creation for shareholders.
- Management and business strategy: This includes evaluations of innovation, adaptability to the business environment, and motivation to create value for shareholders.

In addition to Moody's credit rating method, Standard & Poor's (S&P) introduced its credit rating system in 1917 [17]. They also use an alphabetical rating system to assess the creditworthiness of businesses but employ different symbols to distinguish rating levels. The S&P credit rating method uses various criteria to evaluate a company's debt repayment ability, including:

- The company's financial situation: This is the most important factor used to assess a company's debt repayment ability. It includes indicators such as debt-to-total assets ratio, return on equity, free cash flow, and financial leverage.
- Product and service diversification: A company with diversified products and services is better able to mitigate risks than one focused on a single business area.
- Market position: A company's market position is assessed by examining market share and industry competition. A company with a strong market position is better able to maintain sales and profits.
- Management and business strategy: S&P also assesses the ability of the company's leadership to manage the business and its overall business strategy.

- External factors: S&P considers external factors such as the impact of the economic, political, and legal environment on the company.

Furthermore, Fitch Ratings introduced another credit rating method in 1913[17]. Like the other agencies, Fitch Ratings uses an alphabetical rating system with different symbols to distinguish rating levels. The Fitch Ratings method uses various evaluation criteria to assess a company's debt repayment ability, including:

- Financial Strength: This criterion assesses a company's financial ability, including its profitability, cash flow management, debt repayment capacity, and market opportunity seizing.
- Operating Performance: This criterion evaluates a company's ability to achieve its long-term operational objectives, including growth, profitability, and cost reduction.
- Business Profile: This criterion assesses a company's ability to maintain and grow its sales, profits, and market share in the industry, including strategic direction, human resource management, and customer relations.
- Risk Management: This criterion evaluates a company's ability to manage and control risks in its business operations, including credit risk, market risk, capital risk, and environmental risk.

Globally, major credit rating agencies such as Standard & Poor's (S&P), Moody's, and Fitch Ratings have established well-defined criteria for assessing the creditworthiness of companies. These criteria typically include debt and interest repayment capacity, financial health, market and competition, management and business strategy, and external factors. Debt and interest repayment capacity evaluate a company's ability to repay its principal and interest based on its profitability, assets, and debt repayment history. Financial health is assessed by measuring outstanding debt, net assets, profitability, and cash flow. Market and competition consider the market in which a company operates, including its competitors, pricing power, and value creation for shareholders. Management and business strategy evaluate the company's innovation, adaptability to the business environment, and motivation to create shareholder value. External factors consider the economic, political, and legal environments affecting the company.

In Vietnam, commercial banks have developed internal credit scoring systems to evaluate their clients, tailored to their specific needs and criteria. These internal systems typically include liquidity, leverage, profitability, and efficiency ratios. Liquidity ratios, such as the current ratio and quick ratio, assess a company's ability to meet short-term obligations. Leverage ratios, including debt-to-equity and debt-to-asset ratios, evaluate financial leverage. Profitability ratios, such as return on assets (ROA) and return on equity (ROE), measure financial performance. Efficiency ratios, like asset turnover and inventory turnover, gauge management efficiency.

Given these established criteria, the input variables for the K-means model in this study are selected to provide a comprehensive assessment of a company's financial performance. The variables include financial health ratios (quick ratio, current ratio, short-term liabilities to equity, short-term liabilities to asset, debt to equity, debt to asset, long-term debt to equity, and long-term debt to asset), management efficiency ratios (ROA, asset turnover, accounts receivable turnover, and payment period turnover), growth ratios (sales growth rate and EBIT growth rate), and the dividend payout ratio. These variables are essential for labeling the clusters obtained from the K-means algorithm and developing a robust credit rating system.

By incorporating these financial variables as inputs for the K-means model, this study aims to create a comprehensive credit rating system that accurately reflects various aspects of a company's financial performance and credit risk profile. The identified clusters will provide meaningful and reliable credit ratings for various stakeholders in the financial sector, ultimately promoting a more transparent and efficient financial market.

Despite the widespread use of traditional credit rating methods, these approaches have notable areas for improvement. Traditional methods often rely heavily on expert judgment, which can introduce subjectivity and potential biases into the credit rating process. This subjectivity can lead to consistency in ratings, especially when different experts assess the same company. Additionally, traditional methods may need to efficiently handle large datasets or rapidly changing financial environments, making it difficult to provide timely and accurate credit ratings. They also need to improve in their ability to uncover complex patterns and relationships within financial data, as they often focus on a narrow set of financial indicators and historical performance.

Machine learning techniques, particularly clustering algorithms like K-means, offer solutions to these limitations. Machine learning models can quickly process vast amounts of data and identify intricate patterns and relationships that human analysts may miss. By leveraging data-driven insights, machine learning can enhance the objectivity and consistency of credit ratings. Clustering algorithms, specifically, can group companies based on a comprehensive set of financial indicators, providing a more nuanced understanding of their credit risk profiles. This approach reduces the reliance on subjective expert judgment and improves the transparency and accuracy of the credit rating process.

Clustering Algorithm

This study employs the k-means algorithm as the primary machine learning technique to achieve the research objective. As discussed earlier, the k-means algorithm offers several advantages, including simplicity, computational efficiency, scalability, and proven effectiveness in various applications, particularly in finance and credit risk assessment. By utilizing k-means as the chosen machine learning algorithm, this research aims to effectively uncover patterns and groupings within the dataset, facilitating a deeper understanding of the relationships between financial and non-financial variables and credit ratings. Ultimately, the application of the k-means algorithm in this study is expected to contribute to improved credit rating prediction accuracy, providing valuable insights to support informed decision-making in the credit assessment process.

The k-means algorithm was chosen for this research topic on credit rating prediction for several reasons. First, the simplicity and computational efficiency of the k-means algorithm make it an attractive choice for researchers [10]. The algorithm's straightforward nature allows for rapid prototyping and experimentation, enabling researchers to quickly assess its potential utility in predicting credit ratings. Second, k-means has been proven effective in various applications, including finance and credit risk assessment. Its ability to identify patterns and groupings in data makes it suitable for uncovering distinct credit risk categories based on financial and non-financial variables. This feature can enhance the understanding of the underlying relationships between variables and credit risk, ultimately leading to better prediction accuracy.

Third, k-means is capable of handling large datasets efficiently [13]. As credit rating prediction often involves the analysis of large amounts of data from numerous companies, the algorithm's scalability is a critical factor. K-means can process large datasets quickly, making it suitable for this research context. Lastly, k-means has been successfully applied in previous credit rating research, showing promising results in comparison to other techniques [14,18]. Its previous success in the field adds credibility to its use in the current research topic and suggests that it may provide valuable insights into credit rating prediction. To summarize, the k-means algorithm's simplicity, effectiveness in various applications, scalability, and successful application in previous credit rating research make it a suitable choice for the current research topic. Its ability to efficiently handle large datasets and identify underlying patterns can contribute to improved credit rating prediction accuracy.

The k-means algorithm is an unsupervised machine learning technique widely employed for clustering and partitioning datasets into meaningful groups [10,13]. It aims to identify underlying structures and patterns in the data based on similarity among data points. The algorithm's simplicity, computational efficiency, and effectiveness in various applications make it a popular choice for researchers and practitioners [10].

The k-means algorithm operates by initializing a predetermined number of centroids (k), representing the centers of each cluster. These centroids are generally initialized randomly within the dataset's feature space [13]. The algorithm then iteratively assigns each data point to the nearest centroid, based on a distance metric, such as Euclidean distance [10]. Once all data points are assigned to their respective centroids, the centroids are recalculated to represent the meaning of all data points within each cluster. This process is repeated until convergence is reached, i.e., the centroids' positions stabilize, or a predefined number of iterations have been completed [13].

By partitioning the dataset into distinct groups, the k-means algorithm facilitates the identification of relationships between variables and allows researchers to uncover hidden patterns within the data [10]. In the context of credit rating prediction, the k-means algorithm can be applied to cluster companies based on their financial and non-financial characteristics, providing insights into the factors that drive credit risk and potentially contributing to improved prediction accuracy.

To evaluate the performance of the k-means algorithm in credit rating prediction, various performance metrics can be utilized. One standard method is the silhouette score„ which measures the clustering quality by computing the average distance between observations within the same cluster and comparing it

to the average distance to the nearest neighboring cluster [19]. A higher silhouette score indicates better-defined clusters and implies that the algorithm has effectively identified distinct risk categories in the context of credit rating prediction.

The elbow method is a popular technique to determine the optimal number of clusters (k) in k-means clustering. It involves plotting the variance explained or within-cluster sum of squared distances (WSS) as a function of the number of clusters and identifying the "elbow point," where adding more clusters does not significantly reduce the WSS [20]. The rationale behind the elbow method is that as the number of clusters increases, the WSS decreases since each additional cluster can capture a portion of the remaining variance. However, at some point, adding more clusters will not lead to a substantial decrease in the WSS, and the curve will begin to flatten. The elbow point represents the number of clusters at which the diminishing returns in variance reduction are no longer worth the added complexity of having more clusters [21]. To implement the elbow method, researchers can perform k-means clustering for a range of cluster values (e.g., k = 1 to k = 10) and compute the WSS for each value of k. By visualizing the WSS values on a line chart, the elbow point can be identified, representing the optimal number of clusters for the dataset.

In conclusion, employing the elbow method and silhouette score in this research provides a robust approach to determining the optimal number of clusters for the k-means algorithm in credit rating prediction. The elbow method allows us to identify the point where adding more clusters does not significantly reduce the within-cluster sum of squared distances, ensuring the model's simplicity without compromising its explanatory power. On the other hand, the silhouette score evaluates the quality of clustering by assessing the cohesion within clusters and the separation between them, ensuring that the chosen clusters are meaningful and well-defined.

By combining the elbow method and silhouette score, this research benefits from a comprehensive approach to cluster selection, balancing the trade-off between model complexity and prediction accuracy. These techniques enhance the reliability and validity of the credit rating predictions derived from the k-means algorithm. It contributes to a better understanding of the underlying relationships between variables and credit risk. Ultimately, this approach can lead to more accurate credit rating predictions, benefiting both financial institutions and companies in their decision-making processes.

Previous studies

In recent years, the application of machine learning techniques for predicting corporate credit ratings has become an increasingly popular research topic. A wide range of studies have explored various algorithms, input variables, and methodologies to improve the accuracy and reliability of credit rating predictions.

Early research laid the groundwork for using machine learning in credit rating prediction. Huang et al. [14] compared support vector machines (SVMs) to traditional statistical methods like linear discriminant analysis and logistic regression, while Altman and Sabato [22] explored hybrid models that combined logistic regression with SVM. Both studies found that machine-learning approaches outperformed conventional methods in accuracy and robustness.

Subsequent research has built upon these initial findings. Kim and Kang [15], for example, investigated the performance of decision trees, artificial neural networks (ANNs), and logistic regression in predicting Korean firms' credit ratings. Their study demonstrated that ANNs provided superior accuracy compared to the other methods. Similarly, other studies have compared various machine learning algorithms, such as logistic regression, decision trees, random forests, SVMs, ANNs, and k-nearest neighbors (KNN), to identify the best-performing models for credit rating prediction [23–25].

In terms of input variables, most studies have utilized financial ratios related to liquidity, leverage, profitability, and efficiency [16,26]. However, some research has also explored the incorporation of industry-specific variables, such as asset turnover and net profit margin as well as non-financial data like macroeconomic indicators and textual information from news articles [27]. These studies have found that the inclusion of industry-specific and non-financial variables can improve the accuracy of credit rating prediction models.

The performance of machine learning models in credit rating prediction has been assessed using various evaluation metrics, such as accuracy, precision, recall, and F1 score. Overall, the literature suggests that machine learning algorithms can effectively predict corporate credit ratings using financial ratios as input variables, and that incorporating industry-specific and non-financial variables may further enhance the accuracy of these models [14,16,22,25,27,28].

In summary, the growing body of literature on predicting corporate credit ratings using machine learning models has demonstrated the potential of these

approaches in providing more accurate and reliable predictions compared to traditional statistical methods. Researchers have explored various algorithms, input variables, and methodologies, and have found that a combination of financial ratios, industry-specific variables, and non-financial data can lead to improved performance in credit rating prediction. Future research may further refine these models and explore the potential of emerging machine learning techniques in this area.

Research Gaps

Despite the extensive research conducted on credit rating and risk assessment using machine learning techniques, several gaps remain that this study aims to address. Previous studies have predominantly focused on well-established markets and large corporations, leaving a significant gap in understanding the credit risk dynamics within emerging markets such as Vietnam. For instance, research by Huang et al.[14] and Altman and Sabato[22] primarily explored the use of support vector machines (SVMs) and logistic regression in more developed markets, thereby limiting the applicability of their findings to the Vietnamese context.

Furthermore, while studies by Kim and Kang[15] and Barboza et al.[16] have shown the efficacy of machine learning models such as artificial neural networks (ANNs) and decision trees in credit rating prediction, they often neglect the specific financial indicators relevant to smaller firms and emerging economies. This study bridges this gap by incorporating a comprehensive set of financial ratios specifically tailored to non-financial firms listed on the Ho Chi Minh City Stock Exchange and the Hanoi Stock Exchange.

Additionally, the existing literature, including works by Abdou and Pointon[23] and Galindo and Tamayo[24], has largely overlooked the practical implementation challenges and the need for a standardized and publicly accessible credit rating framework in emerging markets. This study addresses this issue by proposing a robust credit rating system based on the K-means clustering algorithm, which enhances prediction accuracy but also provides a transparent and systematic approach to credit risk assessment.

Moreover, while the integration of non-financial data and industry-specific variables has been explored to some extent[26,27], there is still a lack of research focusing on the unique financial environments of emerging markets. This study fills this void by analyzing key financial indicators such as liquidity ratios, leverage ratios, profitability ratios, and efficiency ratios, which are crucial for assessing the creditworthiness of companies in Vietnam.

In conclusion, this research contributes to the existing body of knowledge by addressing these critical gaps and providing a nuanced understanding of credit risk assessment in the Vietnamese market. By leveraging machine learning techniques and a detailed set of financial indicators, this study offers a practical tool for financial institutions, investors, and policymakers to make informed decisions, ultimately promoting a more transparent and efficient financial market.

METHODOLOGY

Data

In this study, we focus on non-financial firms listed on both the Ho Chi Minh City Stock Exchange and the Hanoi Stock Exchange from 2019 to 2023. The initial dataset comprised data collected from 692 firms. Upon inspection, observations with missing values or duplicates were identified and subsequently eliminated from the dataset. Consequently, the refined dataset encompassed 568 firms, resulting in 2,567 unique observations. The yearly distribution of companies within the dataset is as follows: 510 companies in 2018, 525 companies in 2019, 534 companies in 2020, 532 companies in 2021, and 466 companies in 2022. This comprehensive dataset offers a solid foundation for investigating the credit rating prediction of these non-financial firms using machine learning techniques.

Input Variables

The input data for the K-means model in this study comprises a comprehensive set of financial variables, which can be broadly categorized into four groups: financial health ratios, management efficiency ratios, growth ratios, and dividend payout ratio . These variables provide a detailed assessment of a company's financial performance and are essential criteria for labeling the clusters obtained from the K-means algorithm as described in Table 1.

Financial health ratios include the quick ratio, current ratio, short-term liability on equity, short-term liability on the asset, long-term debt on equity, long-term debt on the asset, debt on equity, and debt on asset. These ratios offer insights into a company's liquidity, solvency, and overall financial stability, capturing the its ability to meet its short-term and long-term obligations.

Management resource management comprise ROA, asset turnover, account receivable turnover, and payment period turnover. These ratios evaluate a company's ability to generate returns from its assets and

the efficiency with which it manages its operations. Efficient management of resources is a critical factor in assessing a company's creditworthiness, as it reflects the firm's capacity to generate profits and meet its financial commitments.

Growth ratios, including sales and EBIT growth rates, capture a company's ability to expand its operations and increase its earnings. Companies with strong growth potential are generally considered less risky, as their expanding revenue base allows them to service their debts better..

Lastly, the dividend payout ratio is important in determining a company's financial health and credit risk. This ratio measures the proportion of earnings paid out to shareholders as dividends, providing insights into a firm's ability to retain earnings for future growth and its commitment to returning value to shareholders.

By incorporating these financial variables as inputs for the K-means model, this study aims to develop a comprehensive credit rating system that accurately reflects the various aspects of a company's financial performance and credit risk profile. The identified clusters will be labeled based on their unique combination of these financial variables, providing a meaningful and reliable credit rating system for various stakeholders in the financial sector.

RESULTS & DISCUSSION

The elbow method graph displays a sharp decline in the SSE (sum of squared errors) from 900 to 400 as the number of clusters (k) increases from 1 to 5. After this point, the SSE continues to decrease, albeit at a slower rate, reaching around 300 at k=7.5. Beyond this point, the SSE exhibits a more gradual decline, decreasing to approximately 200 by the time k reaches 18.

Figure 1 suggests that the optimal value for k is around 6 clusters, as the most significant reduction in SSE occurs up to that point. Beyond k=6, the SSE decreases at a diminished rate, indicating that adding more clusters does not contribute substantially to the reduction of the within-cluster sum of squared distances. Therefore, selecting k=6 strikes a reasonable balance between model simplicity and its ability to capture the underlying patterns in the data, making it a suitable choice for credit rating prediction using the k-means algorithm.

According to Figure 2, upon analyzing the silhouette scores, we observe a gradual decline from 0.28 to approximately 0.25 as the number of clusters (k) increases from 1 to 5. The silhouette score remains relatively stable, fluctuating around 0.25, as k increases from 5 to 8. However, beyond k=8, the silhouette score experiences a sharp drop, decreasing to 0.2 as k continues to increase up to 20.

Considering the results from both the elbow method and silhouette score analyses, we can conclude that selecting k=6 is an appropriate choice for our credit rating prediction model. With the elbow method revealing a significant drop in SSE at k=6 and the silhouette score maintaining a relatively stable level around k=5 to k=8, it is reasonable to proceed with fitting the k-means model using k=6. This choice balances the trade-off between model complexity and performance, thus allowing us to effectively uncover the underlying relationships between variables and credit risk in our dataset.

In the three-dimensional space depicted in Figure 3, it is evident that the k-means clustering algorithm effectively partitions the data into distinct clusters with clear convergence. To further assess the differences between these six clusters, it is necessary to examine additional graphical representations or employ descriptive statistical methods, as discussed below. By doing so, we can better understand the criteria that set each cluster apart and solidify our confidence in the effectiveness of using k=6 in the k-means clustering algorithm for credit rating prediction.

Table 2: Number of observations for each cluster with K=6

Cluster	Observations
0	213
1	208
2	623
3	369
4	463
5	691

Table 2 displayed above provides a comprehensive overview of the distribution of observations within the six clusters generated by the k-means clustering algorithm. The different number of observations in each cluster suggests that the dataset comprises diverse patterns and relationships, which have been successfully captured by the algorithm. Cluster 0 contains 213 observations, indicating a group of companies with certain shared characteristics. Similarly, Cluster 1 comprises 208 observations, revealing another set of companies with distinct features. Cluster 2, the largest group with 623 observations, represents a significant portion of the dataset and highlights a more prevalent pattern among the companies. Cluster 3, consisting of 369 observations, and Cluster 4,

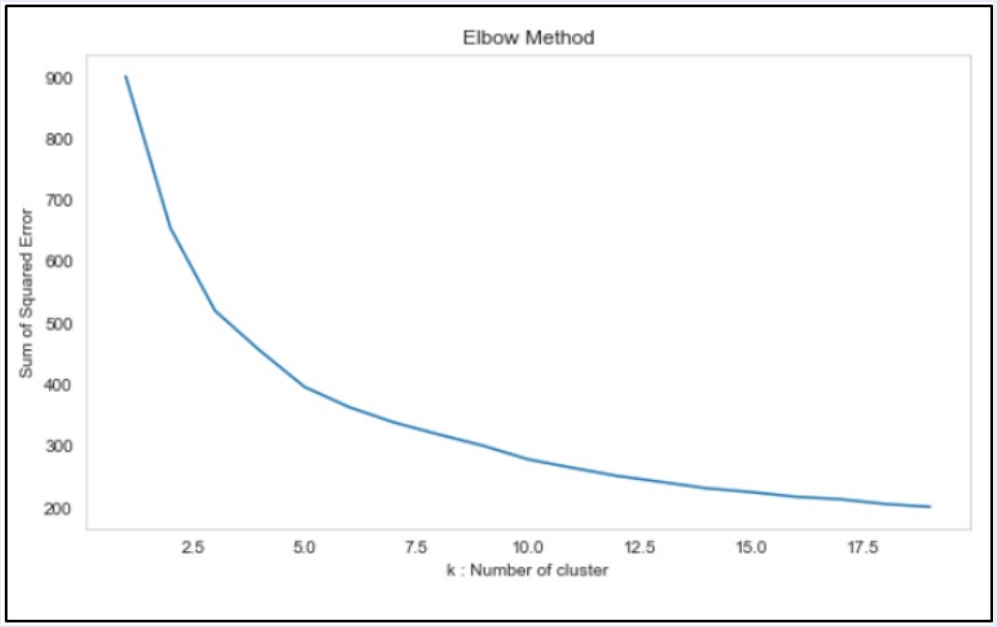

Figure 1: Sum of Squared Error by number of clusters (Source: Author's Calculation)

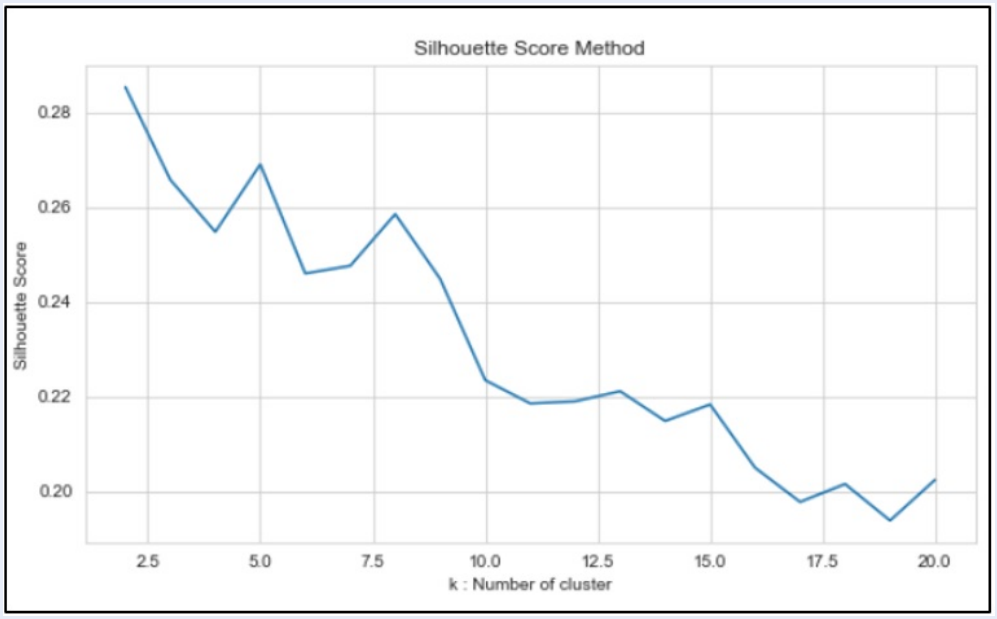

Figure 2: Silhouette Score by number of clusters (Source: Author's Calculation)

Table 1: Credit Rating Criteria and Measurement Methods

Criteria Group	Criteria	Measurement Method	Referenced Standards
Financial Health Ratios	Quick Ratio	Quick Assets / Current Liabilities	Standard & Poor's, Moody's, Fitch Ratings
	Current Ratio	Current Assets / Current Liabilities	Standard & Poor's, Moody's, Fitch Ratings
	Short-term Liabilities to Equity	Short-term Liabilities / Equity	Internal Standards of Vietnamese Commercial Banks
	Short-term Liabilities to Asset	Short-term Liabilities / Total Assets	Internal Standards of Vietnamese Commercial Banks
	Debt to Equity	Total Debt / Equity	Standard & Poor's, Moody's, Fitch Ratings
	Debt to Asset	Total Debt / Total Assets	Standard & Poor's, Moody's, Fitch Ratings
	Long-term Debt to Equity	Long-term Debt / Equity	Internal Standards of Vietnamese Commercial Banks
	Long-term Debt to Asset	Long-term Debt / Total Assets	Internal Standards of Vietnamese Commercial Banks
Management Efficiency Ratios	Return on Assets (ROA)	Net Income / Total Assets	Standard & Poor's, Moody's, Fitch Ratings
	Asset Turnover	Net Sales / Average Total Assets	Standard & Poor's, Moody's, Fitch Ratings
	Accounts Receivable Turnover	Net Credit Sales / Average Accounts Receivable	Internal Standards of Vietnamese Commercial Banks
	Payment Period Turnover	Number of Days in Period / Payables Turnover	Internal Standards of Vietnamese Commercial Banks
Growth Ratios	Sales Growth Rate	(Current Year Sales - Previous Year Sales) / Previous Year Sales	Standard & Poor's, Moody's, Fitch Ratings
	EBIT Growth Rate	(Current Year EBIT - Previous Year EBIT) / Previous Year EBIT	Standard & Poor's, Moody's, Fitch Ratings
Dividend Payout Ratio	Dividend Payout Ratio	Dividends / Net Income	Standard & Poor's, Moody's, Fitch Ratings

Source: by authors

with 463 observations, illustrate additional variations within the dataset. Lastly, Cluster 5 encompasses 691 observations, making it the second-largest group and pointing to another common pattern among the companies.

These varying cluster sizes demonstrate the k-means algorithm's effectiveness in identifying and segregating diverse patterns within the dataset. The k-means clustering algorithm with k=6 has resulted in the formation of six distinct clusters, which the author proposes to use as the basis for a new credit rating system. This system is outlined in the Table 3 and consists of the following credit ratings.

The K-means clustering algorithm applied in this study identified six distinct clusters (0, 1, 2, 3, 4, 5), each representing different levels of financial performance and credit risk. These clusters provide valuable insights into the financial health and creditworthiness of the companies analyzed, which can be understood through theoretical, empirical, and practical lenses.

- Cluster 0 (C): Companies in Cluster 0 exhibit significant liquidity challenges and lower management efficiency. The high levels of both

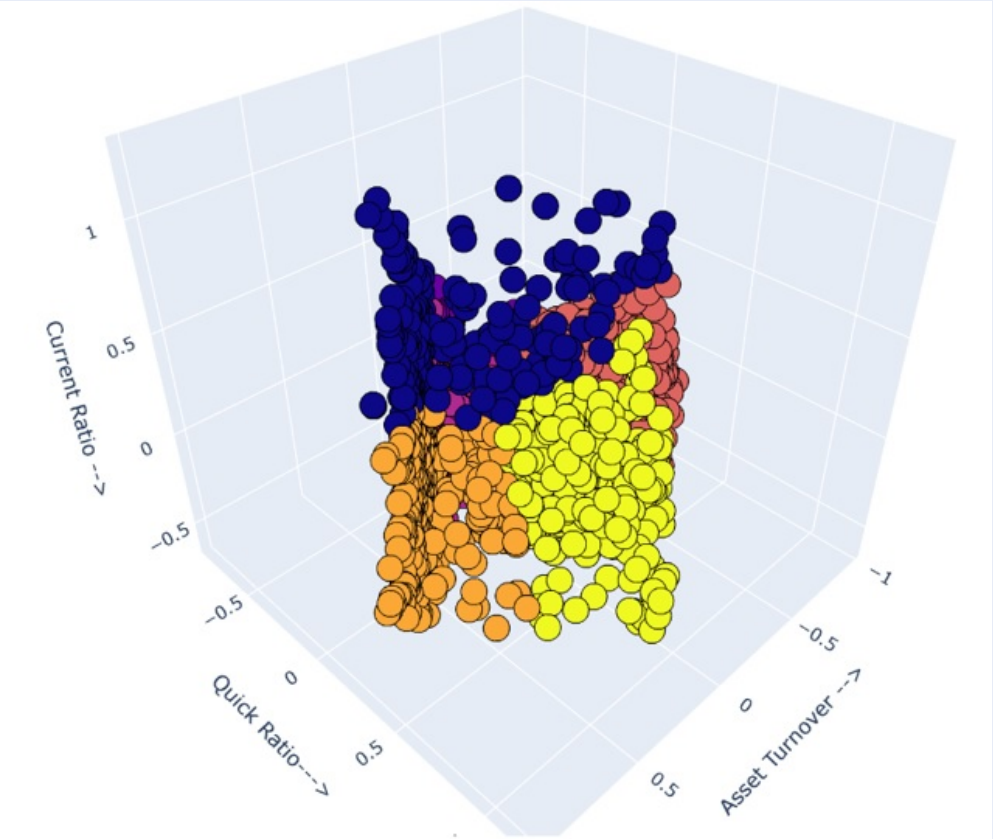

Figure 3: K-Mean Clustering Result with K=6 (Source: Author's Calculation)

Table 3: Suggested label for credit scoring.

Label	Description
A+	Very good (very low credit risk)
A	Good (low credit risk)
B+	Fairly good (credit risk in the middle range from fair to good)
	Average (medium credit risk)
C+	Poor (high credit risk)
C	Very poor (very high credit risk)

Source: Author's Suggested

short-term and long-term debt indicate a substantial credit risk. Theoretically, this aligns with the Pecking Order Theory[9], suggesting that companies facing financial distress are more reliant on debt. Empirically, the observed low return on assets (ROA) and subpar growth rates support categorizing these companies as high-risk. Practically, investors and financial institutions should approach these firms with caution, considering their high likelihood of financial instability.

- Cluster 1 (A+): This cluster is characterized by outstanding liquidity, low indebtedness, and strong financial health, positioning these companies as very low credit risk. The Trade-off Theory supports the high creditworthiness of firms with optimal leverage, which is evident in this cluster. Empirically, the high ROA and efficient management practices confirm the theoretical expectations. Practically, companies in

this cluster are attractive investment opportunities due to their financial stability and low risk of default.

- Cluster 2 (A): Companies in Cluster 2 also display robust financial health with above-average management efficiency and growth potential. However, their liquidity is not as strong as that in Cluster 1. This finding is consistent with the Modigliani-Miller Theorem, which suggests that firm value is independent of capital structure under certain conditions [5]. Empirically, the strong ROA and EBIT growth rate validate the theoretical foundation. Practically, these firms are still considered low-risk and are suitable candidates for investment, albeit with slightly higher caution than Cluster 1.
- Cluster 3 (B+): This cluster includes companies with mixed financial health and management efficiency. While they have reasonable liquidity, their high debt levels increase credit risk. The theoretical backing from the Trade-off Theory indicates that these firms balance the benefits of debt with the risk of financial distress. Empirically, the average ROA and above-average growth rates provide a nuanced understanding of their creditworthiness. Practically, these companies offer moderate investment potential but require a thorough risk assessment.
- Cluster 4 (B): Firms in Cluster 4 show weaker financial health and lower management efficiency, coupled with higher debt ratios. The Pecking Order Theory again explains the reliance on debt due to financial constraints. Empirically, their low ROA and mixed growth rates indicate medium credit risk. Practically, while investment in these firms is riskier, potential returns could be balanced against the higher risk, making them suitable for risk-tolerant investors.
- Cluster 5 (C+): Companies in this cluster have better financial health than those in Cluster 0 but still face significant credit risk due to lower management efficiency and growth rates. The theoretical implications align with the Trade-off Theory, indicating an ongoing struggle to maintain financial stability. Empirically, the findings of moderate ROA and low dividend payout ratios reinforce their classification. Practically, these firms are higher-risk investments, and investors should be cautious.

This proposed credit rating system aims categorizes companies based on their credit risk levels, as determined by the k-means clustering analysis. By assigning specific credit ratings to each cluster, the author has established a comprehensive framework to assess the creditworthiness of companies. The ratings range from A+ for those exhibiting shallow credit risk to C for companies with very high credit risk.

The suggested credit rating system provides a valuable tool for investors, financial institutions, and regulators to make informed decisions and assess the credit risk of different companies effectively. By leveraging the insights from the k-means clustering analysis, the proposed system captures the underlying relationships between financial and non-financial variables, contributing to determining credit risk levels.

The k-means clustering algorithm with k=6 has successfully grouped the data into six distinct clusters, each with different characteristics regarding financial health, management efficiency, growth potential, and dividend payout capacity. These clusters offer valuable insights into the various credit risk profiles and can aid in developing a credit rating system (see Appendix 1 & 2).

Upon examination of the clusters, it is evident that companies in Cluster 1 exhibit outstanding liquidity and low indebtedness, indicating strong financial health. However, they have lower growth rates and dividend payout ratios than the average. Cluster 2 companies, on the other hand, demonstrate above-average management efficiency and growth potential but have average liquidity and lower dividend payout ratios.

Clusters 3 and 4 present a more mixed picture, with companies in these groups showing weaker financial health and management efficiency, alongside varied growth potential. Both clusters have lower dividend payout ratios compared to the average. Companies in Cluster 5 display better financial health, average management efficiency, and higher growth rates, but their dividend payout ratios remain low. Finally, Cluster 0 companies face liquidity challenges and lower management efficiency, along with average growth rates and below-average dividend payout ratios.

These findings suggest that companies within each cluster share common financial and operational characteristics, which can help inform credit risk assessment and decision-making. It is crucial to note that further research, including the evaluation of additional graphs and the application of descriptive statistical methods, is necessary to validate the differences between clusters and refine the proposed credit rating system. Moreover, it is essential to consider external factors, such as market conditions and industry-specific risks, to ensure a comprehensive and accurate credit risk assessment.

Upon revisiting the clusters with the new naming convention, the author proposed the following credit rating suggestions: Cluster 1 as A+, Cluster 2 as A, Cluster 3 as B+, Cluster 4 as B, Cluster 5 as C+, and Cluster 0 as C. This rating system aligns with the companies' observed financial and operational characteristics within each cluster.

Companies in Cluster A+ (Cluster 1) demonstrate exceptional financial health, while those in Cluster A (Cluster 2) exhibit above-average management efficiency and growth potential. Cluster B+ (Cluster 3) and Cluster B (Cluster 4) include companies with varying financial health and management efficiency. Companies in Cluster C+ (Cluster 5) display better financial health and higher growth rates, but lower dividend payout ratios. Finally, Cluster C (Cluster 0) comprises companies facing liquidity challenges and lower management efficiency. The suggested credit rating system appears to be a logical classification based on the distinct characteristics observed in each cluster.

CONCLUSIONS & RECOMMENDATIONS

Conclusions

In conclusion, this study has made a significant contribution to the development of a credit rating system based on companies' financial and operational characteristics using the K-means clustering algorithm. The research objectives were successfully met, with the K-means model effectively clustering the companies into six distinct groups, each exhibiting unique financial and operational attributes. The author has suggested a credit rating system consisting of A+, A, B+, B, C+, and C labels, representing varying levels of credit risk.

The findings of this study provide valuable insights into the financial and operational features that distinguish companies with different credit risk profiles. By identifying these characteristics, the proposed credit rating system offers a practical tool for assessing credit risk, which various stakeholders, including financial institutions, credit rating agencies, and investors can use.

Furthermore, this research has demonstrated the potential of clustering techniques, notably the K-means algorithm, for addressing complex financial problems such as credit risk assessment. The methodology employed in this study can serve as a foundation for future research endeavors that aim to improve and refine credit rating systems.

The practical application of the K-means clustering model developed in this study can significantly enhance credit rating processes within various financial institutions. Commercial banks can implement this model to improve their internal credit scoring systems, allowing for more accurate risk management and loan pricing strategies by better segmenting corporate clients based on credit risk. Credit rating agencies in Vietnam can utilize this model to supplement traditional credit rating methods, providing a data-driven approach that complements expert assessments. Additionally, government and regulatory bodies, such as the State Bank of Vietnam, can use the model to monitor and evaluate the financial health of businesses within the economy, facilitating more informed policymaking.

To ensure the credibility and usability of the model, the results should be published and disseminated in a transparent manner. This can be achieved through periodic reports that detail the credit ratings of companies segmented by the identified clusters, making these reports accessible to investors, financial institutions, and other stakeholders. Furthermore, developing an online platform where stakeholders can access real-time credit ratings and updates will provide detailed insights into rated companies' financial health and risk profiles.

Several factors underscore the reliability of the K-means clustering model in assessing credit risk. The model is grounded in quantitative data, utilizing comprehensive financial indicators to ensure robust credit ratings. Using the elbow method and silhouette scores to determine the optimal number of clusters enhances the model's robustness and validity. Additionally, the clustering results align with established financial theories, providing empirical support for the model's conclusions. To maintain continuous reliability, it is essential to periodically update the model with new data and refine the input variables based on evolving market conditions and financial environments. Regular validation against actual financial outcomes will enhance the model's accuracy and credibility.

Recommendations

Overall, this study's findings contribute to the existing body of knowledge on credit risk assessment and offer a foundation for the development of more accurate and reliable credit rating systems. By addressing the identified limitations and recommendations, future research can continue to advance our understanding of credit risk and support improved decision-making processes in the financial sector.

For investors, focusing on companies categorized in clusters A+ and A, as they demonstrate robust financial health, efficient management, and promising growth potential. These companies will likely offer higher returns on investment and lower credit risk. Additionally, investors should consider diversifying their portfolio by including companies from clusters B+ and B, as they may present moderate risk and potential for growth. However, investors should cautiously approach investments in clusters C+ and C due to their relatively weaker financial health and management efficiency.

Managers of companies within clusters B+, B, C+, and C should improve their financial health and management efficiency. This may include enhancing liquidity management, reducing debt levels, optimizing working capital, and implementing cost control measures. Furthermore, managers should focus on sustainable growth strategies and aim for higher operational efficiency to increase profitability and competitiveness.

Government agencies can utilize the clustering results to understand the financial landscape better and identify potential areas of concern. This information can be used to develop targeted policies and regulations to promote a healthier financial environment for companies. Additionally, government agencies can support and incentivize companies in lower-ranked clusters to improve their financial stability and promote growth. This might include offering tax incentives, providing access to low-interest loans, or facilitating collaboration between companies and relevant stakeholders to foster innovation and technological advancements.

For Credit Rating Agencies, adopting the K-means clustering algorithm can lead to more accurate and reliable credit ratings. The algorithm's ability to handle large datasets efficiently and its robustness in identifying distinct credit risk profiles can improve the overall quality of credit assessments. Credit Rating Agencies can integrate this algorithm into their existing frameworks to complement expert evaluations, thereby enhancing the transparency and credibility of their ratings. Several policies and solutions should be considered to help Credit Rating Agencies achieve more accurate and reliable credit ratings using the K-means clustering algorithm. Firstly, Credit Rating Agencies should invest in advanced data analytics infrastructure to support the implementation of machine learning models. This includes acquiring the necessary hardware, software, and skilled personnel to manage and analyze large datasets. Additionally, staff training and development programs should be established to ensure they are proficient in the latest data analysis and machine learning techniques. Financial institutions should collaborate with credit rating agencies to share relevant financial data, enhancing the robustness of the clustering models. This collaboration can be facilitated through standardized data-sharing agreements that protect the confidentiality and integrity of sensitive information. Moreover, financial institutions should consider integrating these advanced credit rating models into their risk management and loan pricing strategies to optimize their credit assessment processes.

Government and regulatory bodies play a crucial role in fostering an environment conducive to adopting such advanced technologies. They should establish guidelines and regulations that encourage using data-driven credit rating methods while ensuring data privacy and security. Incentives, such as tax breaks or grants, could be provided to CRAs and financial institutions that invest in these technologies. Furthermore, regulatory bodies should promote transparency and standardization in credit rating practices to enhance the comparability and reliability of credit ratings across the market.

However, it is important to acknowledge that the proposed credit rating system may have limitations, and further research is needed to ensure its robustness and accuracy. Additional validation, incorporation of external factors, longitudinal analysis, and comparison with other methods are recommended to enhance the credit rating system's comprehensiveness and predictive power. While the K-means clustering model provides valuable insights, there are certain limitations to consider. First, the analysis is based on a set of financial ratios, which may not capture all aspects of a company's performance. Second, the model is sensitive to the initial cluster centroids, which can affect the results. Finally, the model relies on historical data, and thus may not accurately predict future performance or account for external factors such as economic or industry changes.

FUNDING

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

ABBREVIATIONS

SVM: Support Vector Machine
LDA: Linear Discriminant Analysis
LR: Logistic Regression
HOSE: Ho Chi Minh City Stock Exchange
HNX: Hanoi Stock Exchange

IPO: Initial Public Offering
ML: Machine Learning
ROA: Return on Assets
ROE: Return on Equity
EPS: Earnings Per Share
DPR: Dividend Payout Ratio
CR: Current Ratio
QR: Quick Ratio
DER: Debt to Equity Ratio
GPR: Gross Profit Ratio
NPM: Net Profit Margin
ATO: Asset Turnover Ratio

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest

AUTHORS' CONTRIBUTION

Tam Phan Huy: research ideas, data processing, data collecting, methodology, results interpreting, conclusion and implication writing.

Thuy Chu Quang: coordinator, data collecting, methodology, data visualizing, results interpreting, conclusion and implication writing, table and figure editing.

APPENDIXES

Figures 4 and 5

REFERENCES

1. Chung KJ, Chang SL, Yang WD. The optimal cycle time for exponentially deteriorating products under trade credit financing. The Engineering Economist. 2001;46(3):232-42;Available from: https://doi.org/10.1080/00137910108967575.
2. Scherr FC. Credit-granting decisions under risk. The Engineering Economist. 1992;37(3):245-62;Available from: https://doi.org/10.1080/00137919208903072.
3. Yilmaz MK, Kucukcolak A. Effects of Basel II standards on small-medium size enterprises: evidence from the Istanbul Stock Exchange. Am J Finance Account. 2009;1(4):408-31;Available from: https://doi.org/10.1504/AJFA.2009.031776.
4. Yamanaka S. Credit scoring method using estimated forward financial statements based on purchase order information. JSIAM Lett. 2019;11:33-6;Available from: https://doi.org/10.14495/jsiaml.11.33.
5. Modigliani F, Miller MH. The cost of capital, corporation finance and the theory of investment. Am Econ Rev. 1958;48(3):261-97;.
6. Myers SC, Majluf NS. Corporate financing and investment decisions when firms have information that investors do not have. J Financ Econ. 1984;13(2):187-221;Available from: https://doi.org/10.1016/0304-405X(84)90023-0.
7. Xu R, Wunsch DC. Clustering algorithms in biomedical research: a review. IEEE Rev Biomed Eng. 2010;3:120-54;Available from: https://doi.org/10.1109/RBME.2010.2083647.
8. Kraus A, Litzenberger RH. A state-preference model of optimal financial leverage. J Finance. 1973;28(4):911-22;Available from: https://doi.org/10.1111/j.1540-6261.1973.tb01415.x.
9. Myers SC. Capital structure puzzle. 1984;Available from: https://doi.org/10.3386/w1393.
10. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv. 1999;31(3):264-323;Available from: https://doi.org/10.1145/331499.331504.
11. Gitman LJ, Juchau R, Flanagan J. Principles of managerial finance. Pearson Higher Education AU; 2015;.
12. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J Classif. 2014;31:274-95;Available from: https://doi.org/10.1007/s00357-014-9161-z.
13. MacQueen J. Some methods for classification and analysis of multivariate observations. Proc Fifth Berkeley Symp Math Stat Probab. 1967;1(14):281-97;.
14. Huang Z, Chen H, Hsu CJ, Chen WH, Wu S. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decis Support Syst. 2004;37(4):543-58;Available from: https://doi.org/10.1016/S0167-9236(03)00086-1.
15. Kim MJ, Kang DK. Ensemble with neural networks for bankruptcy prediction. Expert Syst Appl. 2010;37(4):3373-9;Available from: https://doi.org/10.1016/j.eswa.2009.10.012.
16. Barboza F, Kimura H, Altman E. Machine learning models and bankruptcy prediction. Expert Syst Appl. 2017;83:405-17;Available from: https://doi.org/10.1016/j.eswa.2017.04.006.
17. Cantor R, Packer F. Determinants and impact of sovereign credit ratings. Econ Policy Rev. 1996;2(2);Available from: https://doi.org/10.1111/j.1468-036X.1996.tb00040.x.
18. Vellido A, Lisboa PJ, Vaughan J. Neural networks in business: a survey of applications (1992-1998). Expert Syst Appl. 1999;17(1):51-70;Available from: https://doi.org/10.1016/S0957-4174(99)00016-0.
19. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53-65;Available from: https://doi.org/10.1016/0377-0427(87)90125-7.
20. Kodinariya TM, Makwana PR. Review on determining number of clusters in K-means clustering. Int J. 2013;1(6):90-5;.
21. Ketchen DJ, Shook CL. The application of cluster analysis in strategic management research: analysis and critique. Strateg Manag J. 1996;17(6):441-58;Available from: https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G.
22. Altman EI, Sabato G. Modelling credit risk for SMEs: Evidence from the US market. Abacus. 2007;43(3):332-57;Available from: https://doi.org/10.1111/j.1467-6281.2007.00234.x.
23. Abdou HA, Pointon J. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intell Syst Account Finance Manag. 2011;18(2-3):59-88;Available from: https://doi.org/10.1002/isaf.325.
24. Galindo J, Tamayo P. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Comput Econ. 2000;15:107-43;Available from: https://doi.org/10.1023/A:1008699112516.
25. Min JH, Lee YC. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Syst Appl. 2005;28(4):603-14;Available from: https://doi.org/10.1016/j.eswa.2004.12.008.
26. Kovalerchuk B, Vityaev E. Data mining for financial applications. Data Min Knowl Discov Handb. 2005;1203-24;Available from: https://doi.org/10.1007/0-387-25465-X_57.
27. Yu L, Wang S, Lai KK. A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates. Comput Oper Res. 2005;32(10):2523-41;Available from: https://doi.org/10.1016/j.cor.2004.06.024.
28. Oreski S, Oreski G. Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Syst Appl. 2014;41(4):2052-64;Available from: https://doi.org/10.1016/j.eswa.2013.09.004.

Current Ratio

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	1.249014	0.857532	0.13	0.75	1.05	1.420	5.73
1	208.0	6.950048	5.258556	1.75	3.58	5.36	8.055	29.41
2	623.0	2.768700	2.117755	0.37	1.57	2.21	3.320	17.78
3	369.0	1.235474	0.307692	0.19	1.09	1.22	1.420	2.41
4	463.0	1.149244	0.245689	0.25	1.05	1.15	1.275	2.05
5	691.0	1.799190	0.727213	0.27	1.30	1.67	2.195	5.34

Quick Ratio

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	0.785399	0.533702	0.01	0.4400	0.690	0.98	3.35
1	208.0	3.910048	2.302878	0.50	2.2275	3.215	5.15	10.00
2	623.0	1.370963	0.991135	0.06	0.7300	1.130	1.72	9.35
3	369.0	0.695962	0.322262	0.03	0.4500	0.680	0.88	1.62
4	463.0	0.600130	0.271519	0.08	0.3950	0.590	0.79	1.38
5	691.0	1.056715	0.558862	0.06	0.6700	1.020	1.36	4.81

Short-term Liabilities to Equity

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	1.008480	0.751633	0.055429	0.423098	0.813385	1.413927	3.962566
1	208.0	0.147958	0.086888	0.016252	0.080086	0.124993	0.204092	0.427761
2	623.0	0.429200	0.263983	0.030697	0.224266	0.375488	0.584210	1.598543
3	369.0	1.962069	1.119835	0.660347	1.153185	1.618127	2.474522	5.987908
4	463.0	2.415175	1.108245	0.881542	1.566512	2.177750	2.940786	5.888106
5	691.0	0.753455	0.418643	0.088264	0.444008	0.661971	0.960715	2.613387

Short-term Liabilities to Asset.

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	0.254304	0.114913	0.022833	0.152942	0.263483	0.348461	0.473744
1	208.0	0.113156	0.057034	0.015992	0.069363	0.102883	0.156146	0.292874
2	623.0	0.256920	0.120000	0.029783	0.166092	0.258035	0.346322	0.615169
3	369.0	0.571548	0.118188	0.283094	0.487540	0.568901	0.654615	0.851655
4	463.0	0.622120	0.106021	0.383086	0.542269	0.619415	0.700412	0.849648
5	691.0	0.358073	0.116947	0.068836	0.275002	0.354378	0.441235	0.693855

Debt on Equity

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	1.833243	1.264268	0.499222	1.037094	1.500820	2.086185	8.842545
1	208.0	0.052819	0.093467	0.000000	0.000000	0.003652	0.065736	0.611728
2	623.0	0.160101	0.180212	0.000000	0.000000	0.112836	0.253899	1.060719
3	369.0	1.039450	0.724339	0.000000	0.550061	0.869829	1.472228	4.585333
4	463.0	1.431520	0.706615	0.044589	0.929022	1.321126	1.777834	4.216249
5	691.0	0.327490	0.253585	0.000000	0.127207	0.284606	0.489233	1.468681

Debt on Asset

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	0.470990	0.097547	0.280033	0.396494	0.462748	0.532574	0.756367
1	208.0	0.037748	0.062159	0.000000	0.000000	0.003311	0.053012	0.363810
2	623.0	0.092953	0.093415	0.000000	0.000000	0.072108	0.152994	0.445239
3	369.0	0.307641	0.148741	0.000000	0.210302	0.319588	0.401976	0.752839
4	463.0	0.384554	0.136564	0.012194	0.287844	0.386639	0.476051	0.734588
5	691.0	0.158812	0.104418	0.000000	0.073650	0.158558	0.236323	0.411688

Long-term Debt on Equity

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	1.378774	1.089121	0.312612	0.657698	1.115807	1.719013	8.308940
1	208.0	0.024746	0.070434	0.000000	0.000000	0.000000	0.005613	0.502721
2	623.0	0.064562	0.127442	0.000000	0.000000	0.000000	0.077193	0.907584
3	369.0	0.186735	0.284595	0.000000	0.000000	0.048108	0.267556	1.728118
4	463.0	0.220851	0.308953	0.000000	0.006356	0.088834	0.334022	1.967004
5	691.0	0.099784	0.152461	0.000000	0.000000	0.023061	0.147575	1.250674

Sale Growth Rate

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	9.832394	37.351098	-89.50	-9.900	8.86	21.57	199.04
1	208.0	-0.322188	50.075472	-100.00	-25.690	1.76	19.20	263.28
2	623.0	7.847846	30.093048	-103.92	-4.885	3.91	14.75	233.34
3	369.0	9.326775	31.488429	-70.50	-5.940	6.38	22.21	295.72
4	463.0	13.750086	41.013789	-100.00	-7.710	10.51	30.07	355.93
5	691.0	11.198321	44.949804	-94.29	-13.460	6.89	24.31	273.91

EBIT Growth Rate

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	3.633366	88.100567	-467.45	-19.9100	7.390	34.830	461.12
1	208.0	-18.689837	120.590595	-469.95	-73.1975	-14.955	25.895	452.34
2	623.0	10.914039	49.254717	-75.96	-11.3600	3.410	24.810	499.31
3	369.0	11.974946	49.813616	-90.30	-12.6200	5.130	23.900	409.64
4	463.0	7.306981	93.228668	-509.60	-23.4750	2.260	37.320	493.23
5	691.0	8.294373	102.852295	-503.73	-35.8700	1.080	34.950	499.78

Dividend Payout Ratio

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	21.005587	28.631616	0.00	0.000	0.00	41.18	98.22
1	208.0	3.743125	10.739039	0.00	0.000	0.00	0.00	53.06
2	623.0	65.336950	17.767126	29.83	50.985	64.76	78.92	99.95
3	369.0	67.929431	16.823543	34.22	54.500	68.54	81.70	98.65
4	463.0	4.308294	10.143728	0.00	0.000	0.00	0.00	47.96
5	691.0	4.762171	10.453051	0.00	0.000	0.00	0.00	39.21

ROA

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	3.169953	3.864279	-10.76	1.1600	2.910	4.9600	23.83
1	208.0	5.362644	8.985155	-36.97	0.8675	4.095	8.6175	42.58
2	623.0	10.521621	7.270606	1.03	5.7850	8.490	13.1300	54.15
3	369.0	4.810650	3.305776	0.20	2.3200	4.100	6.5200	19.31
4	463.0	2.744881	4.067027	-11.91	0.4650	2.060	4.4850	21.41
5	691.0	5.369190	7.389523	-51.72	1.1980	4.120	8.2800	38.88

Account Receivable Turnover

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	14.907887	26.277373	0.30	3.640	6.400	13.1300	172.36
1	208.0	13.059418	31.652645	0.09	2.265	5.325	9.1175	271.35
2	623.0	30.809530	62.258080	0.00	5.700	10.600	21.0450	317.17
3	369.0	9.213401	10.687214	0.35	2.810	5.390	10.5600	65.81
4	463.0	11.661102	27.644446	0.26	2.905	5.730	9.7900	317.17
5	691.0	11.625980	28.895539	0.01	2.720	5.100	9.5450	317.17

Payment Period Turnover

Class	count	mean	std	min	25%	50%	75%	max
0	213.0	8.827624	21.690545	0.30	3.380	5.610	8.9400	305.39
1	208.0	30.644279	52.679121	0.18	6.035	11.785	27.7975	305.39
2	623.0	23.950042	36.679793	0.00	7.680	12.910	23.9900	305.39
3	369.0	11.539946	12.549028	0.00	4.590	8.270	13.9000	92.27
4	463.0	13.267970	19.632760	0.00	4.040	7.820	13.4000	180.51
5	691.0	12.422200	16.710830	0.02	4.625	7.320	12.9500	151.61

Figure 4: Descriptive Statistics Of Clusters By Variables

	Financial health factors	Management efficiency	Growth potential	Dividend payout capacity
Cluster 1	Outstanding liquidity, more than twice the average level. Short-term payables over equity are about 15% of the average. Total debt over total assets or total debt over total equity are both very low, about 15% of the average. Long-term and short-term debt ratios are approximately equal. More assets than equity.	ROA is approximately around 75% of the average, the number of collections per year is approximately the average asset turnover, but the number of disbursements is double.	Relatively low revenue growth rate, and low pre-tax profit and interest rate growth rate, about 3 times the average.	Dividend payout ratio is about 6-7 times lower than the average
Cluster 2	Short-term liquidity is approximately average, not dependent on inventory, short-term payables over equity is about 42%, about 50% of the average. Total debt over equity is about 20% and over assets about 37% compared to the average. Short-term debt is higher than long-term debt, but the difference is not significant. Assets are twice the equity.	Outstanding ROA, twice the average, high accounts receivable turnover, double the average, and payment ability is 2/3 of the average.	Positive revenue growth rate, below average but not too much, EBIT growth rate is 20% higher than the average	Dividend payout ratio is about 2 times lower than the average
Cluster 3	Liquidity is available but less than half of the average, loss of short-term liquidity if inventory value is excluded. Short-term payables over equity are approximately twice the average. Total debt over equity or assets are both about 1.25 times the average. Assets are three times the equity.	ROA is around 70% of the average. The cash collection cycle is lower than the payment cycle, and both are below 75% of the average.	Revenue growth rate is at an average level, but EBIT growth rate is about 1.25 to 1.5 times the average.	Dividend payout ratio is about 2 times lower than the average.
Cluster 4	Short-term liquidity is available but very low, more than half lower than the average, dependent on inventory value. Short-term payables over equity are approximately 2 to 2.4 times the average. Total debt over equity or assets is about 1.5 times the average. Assets are three times the equity.	ROA is low, ranging from 50% to 75% of the average. The number of collections and disbursements is about once a month, at an average level.	Revenue growth rate is 1.6 times the average, but EBIT growth rate is about 50% lower than the average.	Dividend payout ratio is about 6-7 times lower than the average.
Cluster 5	Short-term liquidity is available and 50% higher than the average, not dependent on inventory value. Short-term payables over equity are approximately 75%. Total debt over equity is below 50% of the average, with short-term debt being three times the long-term debt. Assets are twice the equity.	ROA is approximately average. The number of collections and disbursements is about once a month, at an average level.	Revenue growth rate is 1.5 times the average, but EBIT growth rate is about 50% lower than the average.	Dividend payout ratio is about 6-7 times lower than the average.
Cluster 0	short-term liquidity is available but less than 50% of the average, losing liquidity when inventory value is excluded. Short-term payables are approximately equal to equity, equivalent to the average. Total debt over equity or total assets is more than twice the average, with most of the debt being long-term.	ROA is 60% lower than the average. The number of collections for accounts receivable is approximately twice that of short-term payables, about 50% of the average.	Revenue growth rate is equivalent to the average, but EBIT growth rate is about 50% lower than the average.	Dividend payout ratio is about 50% lower than the average.

Figure 5: Descriptive By Clusters

Tạp chí Phát triển Khoa học và Công nghệ – Economics-Law and Management 2024, 8(3):5494-5512

Open Access Full Text Article

Bài nghiên cứu

Xếp hạng tín dụng bằng thuật toán phân cụm tại thị trường Chứng khoán Việt Nam

Phan Huy Tâm[1,2,*], Chu Quang Thuy[1,2]

TÓM TẮT

Nghiên cứu này áp dụng thuật toán phân cụm K-means để phát triển khung xếp hạng tín dụng doanh nghiệp cho thị trường Việt Nam. Bằng cách phân tích dữ liệu tài chính từ 568 công ty phi tài chính niêm yết tại thị trường Chứng khoán Thành phố Hồ Chí Minh (HOSE) và thị trường Giao dịch Chứng khoán Hà Nội (HNX) trong giai đoạn từ 2019 đến 2023, nghiên cứu xác định các chỉ số tài chính quan trọng bao gồm tỷ lệ sức khỏe tài chính, tỷ lệ hiệu quả quản lý, tỷ lệ tăng trưởng và tỷ lệ chi trả cổ tức. Mô hình phân cụm K-means cho thấy tính hiệu quả trong phân loại các doanh nghiệp này thành sáu cụm khác nhau, mỗi cụm đại diện cho các mức độ hiệu suất tài chính và rủi ro tín dụng khác nhau. Các cụm này được xếp từ A+ (rủi ro tín dụng rất thấp) đến C (rủi ro tín dụng rất cao), cung cấp sự phân biệt rõ ràng dựa trên sự ổn định tài chính và hiệu quả hoạt động. Cách tiếp cận hệ thống này mang lại những hiểu biết có giá trị cho các nhà đầu tư, nhà quản lý và các cơ quan chính phủ, nâng cao khả năng đưa ra quyết định thông minh. Mặc dù có một số hạn chế như phụ thuộc vào dữ liệu lịch sử và độ nhạy cảm đối với các tâm cụm ban đầu, mô hình phân cụm K-means chứng minh là một điểm khởi đầu mạnh mẽ để đánh giá độ tín nhiệm của các công ty. Nghiên cứu này đóng góp vào tài liệu ngày càng tăng về các ứng dụng học máy trong xếp hạng tín dụng bằng cách chứng minh sự vượt trội của các thuật toán phân cụm so với các phương pháp truyền thống. Nghiên cứu nêu bật cách các chỉ số sức khỏe tài chính và hiệu quả quản lý có thể được tích hợp vào một khung dữ liệu để nâng cao đánh giá rủi ro tín dụng. Kết quả gợi ý rằng cách tiếp cận phân cụm K-means không chỉ cải thiện độ chính xác của xếp hạng tín dụng mà còn thúc đẩy tính minh bạch và hiệu quả trong thị trường tài chính. Hơn nữa, khung đề xuất có thể đóng vai trò là nền tảng để phát triển các mô hình phức tạp hơn, tích hợp thêm các biến tài chính và phi tài chính. Nghiên cứu trong tương lai có thể mở rộng điều này bằng cách tích hợp dữ liệu theo thời gian thực và khám phá tác động của các yếu tố kinh tế bên ngoài đối với rủi ro tín dụng. Bằng cách tận dụng các kỹ thuật học máy tiên tiến, nghiên cứu này mở đường cho các hệ thống xếp hạng tín dụng đáng tin cậy và toàn diện hơn, hỗ trợ sự ổn định và phát triển của các thị trường tài chính tại các nền kinh tế đang nổi như Việt Nam.

Từ khoá: K-Means, Xếp hạng tín dụng, Phân cụm, Việt Nam

[1]*Trường Đại học Kinh tế - Luật, Tp. Hồ Chí Minh, Việt Nam*

[2]*Đại học Quốc gia Tp. Hồ Chí Minh, Tp. Hồ Chí Minh, Việt Nam.*

Liên hệ

Phan Huy Tâm, Trường Đại học Kinh tế - Luật, Tp. Hồ Chí Minh, Việt Nam

Đại học Quốc gia Tp. Hồ Chí Minh, Tp. Hồ Chí Minh, Việt Nam.

Email: tamphan.ntc@gmail.com

Lịch sử

- Ngày nhận: 17-5-2024
- Ngày sửa đổi: 23-7-2024
- Ngày chấp nhận: 27-9-2024
- Ngày đăng: 30-9-2024

DOI : https://doi.org/10.32508/stdjelm.v8i3.1417

Trích dẫn bài báo này: Tâm P H, Thuy C Q. **Xếp hạng tín dụng bằng thuật toán phân cụm tại thị trường Chứng khoán Việt Nam** . *Sci. Tech. Dev. J. - Eco. Law Manag.* 2024, 8(3):5494-5512.