Proceedings of ICMRED 2025 symposium: Effective Communication as a Powerful Management Tool
DOI: 10.54254/2754-1169/184/2025.BL23228

Enhancing Financial Market Efficiency Through Data Science: Mitigating Information Asymmetry

Xiang Meng

University of Toronto, Toronto, Canada
xiangm900@gmail.com

Abstract: Financial market efficiency is significantly influenced by the availability and quality of information, with information asymmetry posing a major barrier to optimal market functioning. This article reviews the role of data science in mitigating information asymmetry and enhancing market efficiency, comparing traditional approaches with modern data-driven methods (e.g., machine learning, NLP, and blockchain). It systematically evaluates traditional approaches used to measure and mitigate information asymmetry and highlights their limitations in accurately capturing complex market dynamics. Traditional approaches such as statistical testing, price behavior analysis, and asset pricing models provide fundamental insights but often fail to capture complex, non-linear market dynamics, such as adverse selection, moral hazard, and asset mispricing, due to their reliance on historical data and linear assumptions. In contrast, data science has revolutionized financial market analysis by combining machine learning, natural language processing (NLP), big data analytics, and blockchain technology to solve information imbalances. It enables real-time analysis of unstructured data, improves predictive modeling, and enhances transparency through sentiment analysis, algorithmic trading, and decentralized ledgers. It concludes that integrating data science with traditional finance theory significantly reduces information gaps, offering policymakers and investors tools to foster fairer, more efficient markets. This bridges theoretical finance with computational innovations, demonstrating how data science addresses longstanding limitations in measuring and improving market efficiency.

Keywords: Information Asymmetry, Market Efficiency, Efficient Market Hypothesis, Machine Learning, Natural Language Processing

1. Introduction

Market efficiency is a fundamental concept in financial economics, describing the degree to which asset prices fully incorporate available information. According to the Efficient Market Hypothesis (EMH), asset prices should accurately reflect all relevant data, thereby making it impossible to consistently achieve excess return. However, real-world financial markets often exhibit inefficiencies due to information asymmetry, a situation where certain investors have access to superior information relative to others. This imbalance can manifest in various forms, including private information, hidden information, and information impactedness, and distort asset pricing and confer unfair trading advantages to better-informed market participants. Classic examples like Akerlof's Lemon Market demonstrate how asymmetric information can drive quality assets out of markets, while the

DOI: 10.54254/2754-1169/184/2025.BL23228

Grossman-Stiglitz paradox highlights the inherent contradiction between efficient markets and the profit motives of information acquisition.

The Efficient Market Hypothesis (EMH) categorizes market efficiency into three forms—weak, semi-strong, and strong—yet empirical evidence consistently challenges these ideals. Behavioral phenomena like herding behavior and structural issues such as persistent bid-ask spreads reveal how psychological and institutional factors disrupt price discovery. Traditional approaches to measuring market efficiency primarily rely on statistical methods, price behavior analysis, and conventional asset pricing models. Despite their usefulness, these methods have significant limitations, particularly because they heavily depend on historical market data and assumptions and may fail to accurately capture rapidly evolving market conditions. Advances in data science now address these gaps. Machine learning models detect non-linear patterns in asset pricing that traditional models miss, while natural language processing analyzes sentiment in earnings calls and news to quantify previously unmeasurable market psychology. These tools not only enhance predictive accuracy but also provide regulators with real-time monitoring capabilities for systemic risks.

This paper aims to systematically evaluate how data science methodologies overcome the limitations of traditional approaches in measuring and mitigating information asymmetry. By contrasting conventional statistical tests with modern machine learning and NLP techniques, this study demonstrates how these advanced tools more effectively capture market inefficiencies such as adverse selection and moral hazard—problems theoretically established by the Lemon Market framework yet inadequately addressed through traditional analytical approaches. The review highlights data science's transformative role in enhancing market transparency while remaining grounded in the empirical evidence presented throughout the discussion.

2. Theoretical basis of information asymmetry and market efficiency

2.1. Information asymmetry: definitions and classifications

Information asymmetry occurs when one party in a financial transaction possesses superior or more complete information compared to another, creating imbalances that can lead to market inefficiencies. It generally manifests in four forms: private information, different information, hidden information, and information impactedness [1].

Private information refers to knowledge exclusively held by specific individuals or entities and unavailable to the public, such as insider knowledge used in trading [2]. Different information arises from varied perspectives and unique expertise held by different investors, influencing their judgment and investment decisions [3-4]. Hidden information refers to data that exists but remains undisclosed or difficult to detect, affecting transparency. Information impactedness occurs when relevant information is known to certain parties but difficult for others to verify, complicating transactions and contractual fairness [5].

2.2. Market efficiency: concept and definitions

In an efficient market, asset prices accurately reflect all available information, eliminating opportunities for consistently earning abnormal returns [6]. Definitions of market efficiency must be precise regarding the investor group being covered as well as the market under consideration [7]. For example, an established financial market may be more effective than an emerging one. Institutional investors who have data access and analytical skills may experience more market efficiency than individual investors who depend on public news and constrained resources. Definitions of market efficiency are also connected with what information is accessible to investors and reflected in the price [7]. This is linked to the Efficient Market Hypothesis (EMH), which defines market efficiency based on the extent to which prices reflect different information.

DOI: 10.54254/2754-1169/184/2025.BL23228

2.3. Efficient market hypothesis

The Efficient Market Hypothesis (EMH) suggests that when market efficiency is high, asset prices fully reflect all available information. There are three versions of the EMH: The weak form states that market prices reflect all past trading information, such as historical prices and trading volume, where future price movements are independent of past trends. The semi-strong form states that prices incorporate all public information, including financial statements and economic news; the strong form states that prices should reflect all information, both public and private [4].

Information asymmetry creates persistent exceptions to each EMH form. The weak-form efficiency fails when historical patterns like bid-ask spreads exhibit predictable anomalies [8]. The famous Lemon Market dilemma exemplifies semi-strong form inefficiency: even with complete public information on used automobile markets, dealers' concealed knowledge of vehicle quality causes pricing distortions that public data cannot address [9]. For strong-form efficiency, the widespread prohibition on insider trading recognizes that private information frequently enables anomalous gains while markets appear efficient to the public [2].

2.4. Consequences of information asymmetry: adverse selection and moral hazard

Information asymmetry significantly reduces market efficiency, influencing decisions within households, firms, and governments [10]. Information influences how people make decisions in their homes, companies, and governments [11]. Private information often arises in explanations of competitive advantage and resource-based theory, and hidden information is depicted as the major cause of adverse selection and moral hazard.

Adverse selection in financial markets refers to a situation where one party in a transaction has superior information compared to the other, leading to a mispricing of assets and inefficiencies in the market. Originally coined in insurance contexts, adverse selection describes situations when insured individuals exploit their private knowledge of their riskiness, information insurers cannot access when setting premiums [12]. In the financial market, adverse selection results from informational disparities between buyers and sellers, allowing lower-quality participants to dominate the market while better-informed participants withdraw. The classic example is the ‘lemon market’ theory introduced by Akerlof [9]. In this scenario, buyers, unable to distinguish good-quality products (“peaches”) from poor-quality ones (“lemons”), offer only average prices, discouraging sellers of high-quality products and eventually reducing the overall quality and efficiency of the market.

Moral hazard, another consequence of information asymmetry, occurs when one party entrusted with another’s interests prioritizes their benefit due to hidden information or actions [13]. For example, a financial advisor may sell unsuitable financial products, such as a risky mortgage, prioritizing personal commissions over client interest [13]. The information asymmetry causes one party to have more information than the other, then preventing the less-informed party from effectively monitoring or controlling the actions of the other party. So, the more informed party can take actions that the less informed party cannot monitor since their behavior is hidden. Therefore, this leads to moral hazard, making the market more inefficient (e.g., lemon market theory, Grossman-Stiglitz paradox, herding behavior).

3. Traditional methods for measuring market efficiency

Market efficiency refers to the degree to which asset prices reflect all available information. The methods measuring market efficiency primarily rely on statistical tests, price behavior analysis, and asset pricing models to determine how well financial markets incorporate information.

DOI: 10.54254/2754-1169/184/2025.BL23228

3.1. Statistical tests

Stock markets show a certain degree of predictability, and using unit root tests, variance ratio tests, and run tests can examine whether stock markets are efficient individually and collectively [6]. In the financial market, future returns could be estimated based on historical data on market prices or stocks [14]. The unit root tests and variance ratio tests are commonly used to determine whether a financial time series is stationary or random walk, which has important implications for modeling, forecasting, and understanding market behavior [15]. The phrase 'random walk' describes the inability to forecast because the stock prices can fluctuate without bound over time [14].

The Augmented Dickey-Fuller (ADF) Test is a baseline test, which is one of the unit root tests to examine whether stock price series are stationary or follow the random walk. The null hypothesis for the ADF test is the presence of a unit root, so not rejecting that hypothesis means the series follows a random walk [6]. The presence of a unit root for a stock price means the stock is consistent with the weak-form efficiency, which means the current financial asset values contain all accessible historical financial data at any moment [14]. Also, the random walk for stock price means the returns of stocks must be uncorrelated [6].

The random walk hypothesis (RWH) suggests asset price as a random process, with future price movements being unpredictable and independent of past movements. The most popular econometric methods for evaluating the random walk hypothesis are variance ratio (VR) tests [16]. It is more reliable and accurate than ADF tests [15]. The main idea behind VR testing is that if the return on a stock is completely random and consistent with RWH, the variance of the k-period returns equals k times that of the one-period return [16].

The Runs test is a non-parametric statistical test used to examine whether the time series exhibits independence and randomness [6]. Failure to reject the null hypothesis for the run test means the sequence of prices or returns is consistent with a random pattern [6]. By looking at whether the difference between the actual observable number of runs and the expected number of runs is statistically significant or not, we can determine whether the series follows the random walk process or not [17].

3.2. Price behavior analysis

Traditional methods for measuring market efficiency and information asymmetry often rely on transaction data such as bid-ask spreads, trading volume fluctuations, and market depth. These metrics provide valuable insights into how well markets function under conditions of asymmetric information, particularly about price discovery and liquidity.

The bid-ask spread, one of the most widely used indicators of market illiquidity, represents the average difference between the price the buyer pays at an early point in time (the bid) and the price the dealer sells at one point in time (the ask) [8]. When new information is released, the ask and bid prices fluctuate until their average becomes the new equilibrium value [18]. Therefore, in an efficient market, the bid-ask average swings at random. The considerable difference between spreads estimated from daily and weekly data implies informational inefficiency [18]. Also, a narrow bid-ask spread indicates high market efficiency and low transaction costs, implying that information is quickly reflected in asset prices [18]. Conversely, a wider bid-ask spread shows there may be significant information asymmetry between market participants.

Another important measure of market efficiency is trading volume fluctuations. The trading volume is the key indicator that assists investors in representing the total number of shares moved over a certain period, assisting investors in differentiating between genuine price movements and possible head fakes. High trading volumes often suggest that information is being rapidly

DOI: 10.54254/2754-1169/184/2025.BL23228

disseminated and absorbed by market participants, promoting market efficiency [19]. Therefore, a low trading volume period can indicate that information is not being fully integrated into asset prices.

4. Data science applications in ensuring market efficiency

The financial market's data volume has grown rapidly in recent years, and traditional data analysis methods have been unable to satisfy modern financial institutions for efficient and accurate information processing [20]. The integration of data science methodologies provides a new solution for studying market efficiency and information asymmetry. Key applications include natural language processing (NLP), machine learning, big data analytics, and blockchain technology.

4.1. Natural language processing in financial markets

Natural language processing (NLP) technology can effectively process and analyze vast amounts of unstructured data, assisting in the identification and assessment of subtle shifts in public opinion as well as offering data-driven support for investment decisions [20]. By analyzing the sentiment embedded in news articles, analyst reports, and social media posts, NLP tools assist in detecting subtle shifts in public perception and market mood, offering real-time, data-driven support for investment decisions [20]. Additional uses include the use of complicated sentiment analysis methods to assess psychological states such as panic and greed index, which are major psychological elements influencing market movements [21]. NLP models can gauge market sentiment and predict asset price movements by evaluating the sentiment expressed in news articles, social media posts, and analyst reports [21]. NLP facilitates the assessment of corporate disclosures and financial reports, enabling the measurement of transparency and detection of potential risks. This aids investors in making informed decisions and enhances overall market efficiency.

4.2. Machine learning in asset pricing and market prediction

Asset pricing fundamentally seeks to understand how risk is compensated through returns [22]. Precise stock price estimation offers crucial data for financial planning and investing choices [23]. Businesses can more efficiently plan their operations given the pattern of stock price changes, and investors can increase their productivity through stock trading [23]. Many machine learning methods, such as random foresting, CNN, and SVM, are ways to do financial forecasting.

The random forest method constructs multiple decision trees utilizing subsets of the dataset and aggregates their predictions to anticipate the upward and downward movement of the index. At each node, a random subset of variables is selected, and the best split is determined using the Gini index, continuing until each node contains a single class. The final prediction is based on majority voting among all decision trees [23].

Convolutional Neural Networks (CNNs) are effective in forecasting by extracting key features from sequential data, such as financial time series or textual data in sentiment analysis. Through convolutional layers, CNNs identify important patterns by applying kernels that capture relationships between data points, while pooling layers help reduce dimensionality and improve generalization [24]. This enables CNNs to recognize patterns across different positions, enhancing their ability to detect trends and predict future market movements efficiently [24].

Support Vector Machines (SVMs) are one of the most popular algorithms for predicting the direction of stock market indices and forecasting the movement of indices. The algorithm minimizes generalization error using structural risk minimization, ensuring better predictive performance [23]. By identifying key support vectors, SVMs effectively distinguish market trends and forecast future price directions [23].

DOI: 10.54254/2754-1169/184/2025.BL23228

5. Discussion

In the financial market, there are two different analysis methods, fundamental and technical analyses, used to predict market prices. Fundamental analysis predicts stock values using financial studies of firms or industries [23].

The goal of fundamental analysis is to buy an asset at a low price and sell it for a high price by calculating the difference between the item's actual value and its market price [23]. The traditional methodologies are essential in understanding market efficiency, yet they face inherent limitations in capturing complex, non-linear relationships within financial data. However, statistical tests such as ADF tests, run tests, and variance root tests are only for weak-form efficiency markets [14].

Compared to the traditional empirical methods in asset pricing, technical analyses such as machine learning allow for a far larger list of potential predictor variables and more detailed functional requirements [22]. This adaptability enables us to advance the boundaries of assessing risk premiums [22]. The most important accomplishment of data science is its ability to process and analyze large volumes of unstructured data in real-time. NLP techniques enable the extraction of sentiment and transparency measures from financial reports, news articles, and social media, providing a more comprehensive understanding of market sentiment and investor behavior [21]. Therefore, machine learning models improve asset pricing by identifying patterns and anomalies that traditional models may ignore, which can get a more accurate prediction of market movements.

Even with these progressions, obstacles still exist in incorporating data science techniques into the analysis of financial markets. Concerns such as data privacy, biases in algorithms, and the understanding of intricate machine learning models present major issues.

6. Conclusion

This paper explores how information asymmetry affects the efficiency of financial markets, contrasting conventional measurement techniques with contemporary data science methods. Although statistical analyses, price movement evaluations, and asset pricing frameworks offer important perspectives on market efficiency, they are progressively being enhanced by machine learning, natural language processing, and big data analytics. Incorporating data science into financial market research provides substantial benefits, such as better predictive modeling, advanced sentiment analysis, and the ability to process data in real-time. These developments help decrease information asymmetry, which in turn boosts market efficiency and transparency.

As financial markets keep evolving, the integration of established financial theories with data science techniques will be essential in influencing future studies on market efficiency. Policymakers, analysts in finance, and investors need to adjust to these technological developments to successfully manage the intricacies of contemporary financial markets. Through the adoption of data-driven strategies, financial organizations can improve their decision-making processes and help foster a more efficient and transparent global financial system.

References

[1] *Bergh, D. D., Ketchen, D. J., Orlandi, I., Heugens, P. P. M. A. R., & Boyd, B. K. (2019). Information Asymmetry in Management Research: Past Accomplishments and Future Opportunities. Journal of Management, 45(1), 122-158. https://doi.org/10.1177/0149206318798026*

[2] *Jaffe, J. F. (1974). Special Information and Insider Trading. The Journal of Business, 47(3), 410–428. http://www.jstor.org/stable/2352458*

[3] *Hambrick, D. C., & Mason, P. A. (1984). Upper Echelons: The Organization as a Reflection of Its Top Managers. The Academy of Management Review, 9(2), 193–206. https://doi.org/10.2307/258434*

[4] *Greene, C. Differential Information, Arbitrage, and Subjective Value. Topoi 40, 745–753 (2021). https://doi.org/10.1007/s11245-019-09661-6*

DOI: 10.54254/2754-1169/184/2025.BL23228

[5] Williamson, O. E. (1975). Markets and hierarchies: analysis and antitrust implications: a study in the economics of internal organization. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.
[6] Guidi, F., & Gupta, R. (2011). Are ASEAN stock markets efficient? Evidence from univariate and multivariate variance ratio tests.
[7] Efficiency, M. MARKET EFFICIENCY—DEFINITION, TESTS, AND EVIDENCE.
[8] Stoll, H. R. (1989). Inferring the components of the bid-ask spread: Theory and empirical tests. The Journal of Finance, 44(1), 115-134.
[9] Devos, J., Van Landeghem, H., & Deschoolmeester, D. (2012). The theory of the lemon markets in IS research. Information systems theory: explaining and predicting our digital society, vol. 1, 213-229.
[10] Kim, J. C., & Lee, B. (2005). When does a lemon market emerge? In Proceedings of KATP (pp. 9-28). Korea Association for Telecommunications Policies.
[11] Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. Journal of Management, 37(1), 39-67.
[12] Siegelman, P. (2003). Adverse selection in insurance markets: an exaggerated threat. Yale LJ, 113, 1223.
[13] Dowd, K. (2009). Moral hazard and the financial crisis. Cato J., 29, 141.
[14] Shaik, M., Kamdar, P., Nawaz, N., Rabbani, M. R., E Vahdati, S., Afzal Saifi, M., & Grewal, H. (2024). The global financial crisis's impact on stock market efficiency: a Fourier unit root test analysis. Cogent Economics & Finance, 12(1), 2392627.
[15] Nguyen, J., & Parsons, R. (2021). A Study of Market Efficiency in Emerging Markets Using Improved Statistical Techniques. Emerging Markets Finance and Trade, 1–13. https://doi.org/10.1080/1540496x.2021.1949981
[16] Hoque, H. A. A. B., Kim, J. H., & Pyun, C. S. (2007). A comparison of variance ratio tests of random walk: A case of Asian emerging stock markets. International Review of Economics & Finance, 16(4), 488–502. https://doi.org/10.1016/j.iref.2006.01.001
[17] AI-Powered Stock Forecasting Algorithm | I Know First |Market Efficiency: Testing Random Walk by the Runs Test and the Variance Ratio Test. (2022). Iknowfirst.com. https://iknowfirst.com/market-efficiency-testing-random-walk-by-the-runs-test-and-the-variance-ratio-test
[18] ROLL, R. (1984). A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market. The Journal of Finance, 39(4), 1127–1139. https://doi.org/10.1111/j.1540-6261.1984.tb03897.x
[19] NIckolas, S. (2020). Using Trading Volume to Understand Investment Activity. Investopedia. https://www.investopedia.com/ask/answers/041015/why-trading-volume-important-investors.asp
[20] Xiao, J., Wang, J., Bao, W., Deng, T., & Bi, S. (2024). Application progress of natural language processing technology in financial research. Financial Engineering and Risk Management, 7(3), 155-161.
[21] Ling Aifan, Peng Wei, Wang Qianqian, et al. Progress in the application of natural language processing technology in financial research [J]. Theory and Practice of System Engineering, 2024, 44 (01): 387-421.
[22] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273.
[23] Ayyildiz, N., & Iskenderoglu, O. (2024). How effective is machine learning in stock market predictions? Heliyon, 10(2), e24123. https://doi.org/10.1016/j.heliyon.2024.e24123
[24] Puh, K., & Babac, M. B. (2023). Predicting stock market using natural language processing. American Journal of Business, 38(2), 41–61. https://doi.org/10.1108/ajb-08-2022-0124