Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/115/2025.18521

Machine Learning Models in Quantitative Investment

Tianyi Zhu[1,a,*]

[1]*School of Business, Soochow University, No.50 Dong Huan Road, Suzhou, Jiangsu, 215021, P.R.China*
a. ztyzty77@outlook.com
**corresponding author*

Abstract: This paper explores the application of machine learning models in the realm of quantitative investment, emphasizing their potential to enhance decision-making processes and predictive accuracy in financial markets. Beginning with an overview of machine learning types—supervised, unsupervised, and semi-supervised—the paper delves into specific models commonly employed in investment strategies. These include supervised models such as Random Forests and Support Vector Machines, as well as unsupervised models like K-Means Clustering and Bayesian Networks. The practical applications and advantages of each model are discussed, with a focus on how they contribute to portfolio management, risk assessment, and the identification of profitable trading opportunities. While highlighting the transformative power of machine learning, the paper also acknowledges the inherent challenges, such as data quality and model interpretability, which practitioners must navigate. By examining the benefits and limitations of these technologies, this paper provides a comprehensive understanding of how machine learning can be effectively integrated into quantitative investment practices and offers insights into future research directions.

Keywords: Machine Learning, Quantitative Investment, Supervised Learning, Unsupervised Learning.

1. Introduction

Over the past few years, Artificial Intelligence (AI) has become a highly popular topic over the world. Recognizing its well-rounded capabilities, countries and corporations across the world is striving to develop advanced AI program. Machine learning models, particularly probabilistic models, have made tremendous contributions in this advancement [1]. In reality, AI is a comprehensively combined application of machine learning, statistical models, and other techniques [2]. This illustrates that machine learning is the key in AI development. Moreover, machine learning is of great importance not only for developing AI but also for maximizing returns, achieving scientific innovation, and numerous other applications [3].

The history of machine learning dates back centuries, to the 17th century, when Bayes, Laplace's derivation of least squares method and the creation of Markov chains took place. These early advancements provided the groundwork for modern machine learning practices. The proposal of modern concept of machine learning is associated with psychologist Frank Rosenblatt from Cornell University, who conducted pioneering researches on machine learning models in 1950s. Following his work, the field of machine learning began to boom [4].

As machine learning develops, it has also impacted quantitative investment. With the help of machine learning, quantitative investment strategies transfer from traditional methods to modern, data-driven approaches. In spite of the wide use of machine learning, this paper specifically focuses on its application in quantitative investment. First, this paper provides a brief introduction to machine learning and quantitative investment. Then it offers a deep insight into six common machine learning models and explains their potential applications in quantitative investment. Finally, it presents the key conclusions drawn from this analysis.

2. Machine Learning

Machine learning, as defined by Arthur Samuel, is the field of study that provides computers with the ability to learn without explicit programming [5]. It enables machines to deal with data more efficiently. Machine learning is mainly separated to three categories: supervised learning, unsupervised learning and semi-supervised learning.

Supervised learning uses labeled data for specific predictions. In contrast, unsupervised learning operates without labels, focusing on discovering hidden patterns or structures within data. Semi-supervised learning bridges the gap by combining labeled and unlabeled data, which is particularly useful when labeling data is costly or time-consuming.

Examples of supervised machine learning models include Random Forests (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and various neural networks (e.g. DNN, RNN, CNN, LSTM). Common unsupervised machine learning models include Bayesian Network, K-Means Clustering, Hierarchical Clustering, and Principal Component Analysis (PCA). Semi-supervised machine learning models are combinations of supervised machines learning models and unsupervised machine learning models, integrating characteristics of both.

This paper introduces 4 supervised machine learning models--Random Forests, Support Vector Machine, K-Nearest Neighbor, and Long Short-Term Memory--and 2 unsupervised machine learning models--Bayesian Network, K-Means Clustering--and explains their possible applications in quantitative investment. The emphasis on supervised machine learning models is due to their prevalent used in quantitative investment, driven by the nature of financial tasks in quantitative finance and the availability of labeled financial data. On the other hand, semi-supervised machine learning models are not covered in this paper because they encompass a large variety of hybrid models, making detailed explanations complex and expansive.

3. Quantitative Investment

Quantitative investment is an investment approach to investing that uses mathematical models, statistical analysis, and big data to inform trading and portfolio management decisions. Unlike traditional stock-picking methods, which rely on traditional stock-picking methods, like analyzing financial statements, quantitative investors use algorithms to process vast amounts of data, identify patterns, and uncover or opportunities in financial markets. A prominent figure in quantitative investment is James Simons, one of the world's highest-paid hedge fund managers, known for his success in applying mathematical and algorithmic strategies to achieve significant financial returns.

4. Supervised Machine Learning Models and Their Applications in Quantitative Investment

4.1. Random Forests

Random Forests is a supervised machine learning algorithm which was firstly proposed by Tin Kam Ho [6] at Bell Laboratory and further developed by Leo Breiman and Adele Cutler [7]. It is primarily used for classification and regression tasks and functions as an ensemble method which combines the predictions from multiple models to improve accuracy and robustness. Random Forests are built upon decision trees, which are simple models that split data based on certain conditions to make predictions. However, a single decision tree can be prone to noise and over-fitting, especially on complex datasets. To mitigate these issues, Random Forests generates a large number of decision trees and combines their predictions. Each tree is trained independently on a random subset of the data, and their collective outputs are aggregated for the final prediction. For classification tasks, the ensemble uses a majority vote, while for regression, it averages the predictions.

Studies have shown that Random Forests can effectively predict stock prices using Python software, supported by various essential library packages such as NumPy, Pandas and Scikit-Learn [8]. The models are evaluated using metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2), demonstrating substantial predictive accuracy.

In another study, Random Forests models directly linked market indicators—including daily open, close, high, low prices and volume—to trading decisions. The model's output decision among "buy", "sell" or "hold" options were shown to generate excess return after performance analysis [9].

Random Forests can also be enhanced by integrating them with various quantitative investment strategies. For example, the Bollinger Band Strategy which was developed by John Bollinger in the 1980s can be combined with Random Forests model to generate higher profits [10].

4.2. Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning model used primarily for classification, though it can also be adapted for regression tasks. SVM was first introduced by Vladimir N. Vapnik and Alexey Y. Chervonenkis in 1964 [11].

SVM is particularly effective for complex datasets where the separation between classes isn't straightforward. The primary objective of the SVM algorithm is to identify the optimal hyperplane in an N-dimensional space that can effectively separate data points into different classes in the feature space. The algorithm ensures that the margin between the closest points of different classes, known as support vectors, is maximized.

SVM has been successfully applied to forecast stock market movement direction, outperforming many other models due to its structural risk minimization approach. While other models are usually based on minimization of empirical risk, SVM seeks to minimize an upper bound of the generalization error instead of minimizing training error , making it less prone to overfitting [12].

Moreover, SVM is valuable in portfolio optimization. An important indicator to evaluate the performance of an investment portfolio is Sharpe Ratio which was firstly proposed by Economist William F. Sharpe in 1966. With the help of SVM model, an investment portfolio established on the basis of the fundamental indicators has a high Sharpe Ratio [13].

4.3. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple and intuitive machine learning model used for classification and regression tasks. The model was first introduced by Thomas Cover and Peter E.Hart [14]. KNN classifies data points based on the labels of their closest "neighbors". When a new data point needs to be classified, KNN examines the 'k' closest points in the dataset (its "neighbors") and assigns a label based on the majority label among those neighbors. The parameter 'k' (number of neighbors) is crucial; a small 'k' may make the model sensitive to noise, while a large 'k' can overly smooth decision boundaries. A common approach is to test several 'k' values and use cross-validation to select the optimal one. KNN is a simple machine learning model because it's highly interpretable and requires minimal model training. Moreover, KNN is a non-parametric method, meaning it does not assume a specific underlying data distribution.

Despite its straightforward nature, KNN may face limitations when used independently in quantitative investment. Because financial data often has many features, such as price movements, trading volume, economic indicators, and technical indicators, all interacting in complex ways. KNN struggles in high-dimensional spaces due to the "curse of dimensionality," when the number of features is large, the data points tend to become more dispersed, making it harder for KNN to find meaningful “neighbors”. Additionally, financial markets contain significant random noise, which can lead to overfitting, especially when ‘k’ is small. This can make the model unreliable for actual investment decisions, as it might capture random fluctuations rather than true patterns. Consequently, KNN is often combined with more sophisticated models in quantitative investment.

A common combination is the SVM-KNN model. The SVM model is used to classify data in large datasets by identifying separating surface while the KNN model is used to make predictions by identifying the k nearest neighbors closest to the data [15]. This combined model has shown improved performance compared to using individual models. An enhanced version, called FWSVM-FWKNN, further improves performance. FWSVM stands for feature weighted SVM, which assigns different weight values to different features according to some certain principles, , addressing feature importance. FWKNN applies the same concept to KNN by utilizing a feature-weighted matrix [16]. This combined, feature-weighted model has demonstrated superior performance over the standard SVM-KNN model.

4.4. Long Short-Term Memory

Long Short-Term Memory (LSTM), first introduced by Schmidhuber and Hochreiter in 1997 [17], is a type of recurrent neural network (RNN) architecture designed to model sequential data. Unlike standard RNNs, LSTMs are well-suited for learning from long-term dependencies in data because they overcome the problem of “vanishing gradients,” which can occur when training RNNs on long sequences. Standard RNNs pass hidden states from one cell to the next, which works for short-term dependencies but struggles when trying to remember information from far back in the sequence. LSTMs solve this by maintaining a cell state over time, and using gates that regulate the flow of information—deciding what to keep or discard—thereby effectively managing long-term dependencies. Because LSTMs are capable of learning the context of sequences, LSTMs are particularly valuable in tasks that require an understanding of order and context like predicting stock prices.

Although LSTMs are relatively recent in the field of quantitative investment, their application began to gain attention with research in 2018 [18]. Fischer and Krauss conducted pioneering work using Python, demonstrating that the LSTM model outperforms the Random Forests model. Further advancements were made in 2022, when another research conducted by Pushpendu Ghosh,Ariel Neufeld and Jajati Keshari Sahoo improves the LSTM model by giving it a multi-feature setting [19]. The enhanced LSTM model is proved to have better performance compared to older one with a single feature setting. The research also shows that the LSTM model outperforms Random Forests model and has an advantage compared to the memory-free methods.

LSTM model have also been used to forecast stock price with innovations such as the multi-layer sequential LSTM (MLSLSTM), introduced in 2023 [20]. MLSLSTM contains two major parts named data preparation and sequential LSTM. The sequential LSTM consists of 4 layers. Three of them are vanilla LSTM and the other layer is Dense layer. This deep architecture improves prediction accuracy, making MLSLSTM effective for stock price forecasting, as additional layers help capture more complex patterns in sequential data.

5. Unsupervised Machine Learning Models and Their Applications in Quantitative Investment

5.1. Bayesian Network

A Bayesian Network, first proposed by Judea Pearl in 1988 [21], is a type of probabilistic graphical model that represents a set of variables and their conditional dependencies using a directed acyclic graph (DAG). It combines principles from probability theory and graph theory, making it a powerful tool for reasoning under uncertainty. Bayesian Networks enable inference and allow the calculation of the probability of various outcomes based on available evidence.

Research has demonstrated that Bayesian Network can effectively forecast P/E ratio. It is also an early example for researchers to digitize stock price distribution by using the clustering algorithms [22]. The research demonstrates that a Bayesian Network can forecast P/E ratio with high accuracy and it outperforms traditional machine learning models.

Bayesian Network can also be applied to analyze relationships between the macroeconomy and the stock market. However, there is a limited body of research on this topic. A research took place in 2021 comes up with the conclusion that macroeconomy and the stock market operate independently and the strategy to use macroeconomy information to invest in the stock market will not bring the investors excess returns [23]. As a consequence, more researches are required in this area.

5.2. K-Means Clustering

K-Means Clustering is a popular unsupervised machine learning model used to group data points into a specified number of clusters. Originally proposed by James MacQueen in 1967 [24], the goal of K-means is to divide a dataset into distinct clusters based on the similarity of data points, so points in the same cluster are more similar to each other than to those in other clusters.

K-Means Clustering model has practical applications in constructing stock portfolios based on continuous trend features. The portfolios constructed with the help of K-Means Clustering model show a higher return and a lower risk than those of traditional portfolios [25]. Nevertheless, the quantity of related researches in this area is also small and further researches are needed.

6. Conclusion

This paper provides an overview aimed at giving machine learning freshmen a deep insight into six common machine learning models and their possible usage in quantitative investment. I t helps fill the gap for review articles that cater to beginners in machine learning who seek higher returns through investment strategies and need foundational knowledge of machine learning models and their applications. Besides, this paper highlights the current lack of study in unsupervised machine learning models' applications in quantitative investment. In order to be more proficient in applying machine learning to investment or even try to combine different machine learning models to generate more advanced models, further efforts must be made by learners.

Given the increasing popularity of the concept of AI, more and more people are thirsty for the knowledge of machine learning and its applications in various areas. It is essential for researchers to take on the responsibility of providing comprehensive reviews about machine learning and its usage in certain areas, including investments. This paper also underscores the importance of conducting further research into the application of unsupervised machine learning models to quantitative investment need to be conducted.

References

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

[2] Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. Expert Systems with Applications, 80, 340–355. https://doi.org/10.1016/j.eswa.2017.02.044

[3] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/tit.1967.1053964

[4] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

[5] Fradkov, A. L. (2020). Early History of Machine Learning. IFAC-PapersOnLine, 53(2), 1385–1390. https://doi.org/10.1016/j.ifacol.2020.12.1888

[6] Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. Nature, 521(7553), 452–459. https://doi.org/10.1038/nature14541

[7] Ghosh, P., Neufeld, A., & Sahoo, J. K. (2021). Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. Finance Research Letters, 46, 102280. https://doi.org/10.1016/j.frl.2021.102280

[8] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[9] Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 32(10), 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016

[10] Kühl, N., Goutier, M., Hirt, R., & Satzger, G. (2019). Machine Learning in Artificial Intelligence: Towards a Common Understanding. https://doi.org/10.48550/arXiv.2004.04686

[11] Liu, Y., Feng, H., & Guo, K. (2021). The Dynamic Relationship between Macroeconomy and Stock Market in China: Evidence from Bayesian Network. Complexity, 2021, 1–12. https://doi.org/10.1155/2021/2574267

[12] Macqueen, J. (1967). SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS. https://www.cs.cmu.edu/~bhiksha/courses/mlsp.fall2010/class14/macqueen.pdf

[13] Mahesh, B. (2018). Machine Learning Algorithms -A Review. International Journal of Science and Research (IJSR) ResearchGate Impact Factor, 9(1). https://doi.org/10.21275/ART20203995

[14] Md, A. Q., Kapoor, S., A.v., C. J., Sivaraman, A. K., Tee, K. F., H., S., & N., J. (2023). Novel optimization approach for stock price forecasting using multi-layered sequential LSTM. Applied Soft Computing, 134, 109830. https://doi.org/10.1016/j.asoc.2022.109830

[15] Meher, B. K., Singh, M., Birau, R., & Anand, A. (2024). Forecasting stock prices of fintech companies of India using random forest with high-frequency data. Journal of Open Innovation: Technology, Market, and Complexity, 10(1), 100180. https://doi.org/10.1016/j.joitmc.2023.100180

[16] Nayak, R. K., Mishra, D., & Rath, A. K. (2015). A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices. Applied Soft Computing, 35, 670–680. https://doi.org/10.1016/j.asoc.2015.06.040

[17] Pearl, J. (2014). Probabilistic reasoning in intelligent systems : networks of plausible inference. Morgan Kaufmann.

[18] Qin, Q., Wang, Q.-G., Li, J., & Ge, S. S. (2013). Linear and Nonlinear Trading Models with Gradient Boosted Random Forests and Application to Singapore Stock Market. Journal of Intelligent Learning Systems and Applications, 05(01), 1–10. https://doi.org/10.4236/jilsa.2013.51001

[19] Silva, N. F., de Andrade, L. P., da Silva, W. S., de Melo, M. K., & Tonelli, A. O. (2024). Portfolio optimization based on the pre-selection of stocks by the Support Vector Machine model. Finance Research Letters, 61, 105014. https://doi.org/10.1016/j.frl.2024.105014

[20] Tin Kam Ho. (1995, August 1). Random decision forests. IEEE Xplore. https://doi.org/10.1109/ICDAR.1995.598994

[21] Vapnik, V. N. , & Chervonenkis, A. . (1964). A note on one class of perceptrons. Automation and Remote Control, 25(1).

[22] Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., & Marks, D. (2023). Scientific discovery in the age of artificial intelligence. Nature, 620(7972), 47–60. https://doi.org/10.1038/s41586-023-06221-2

[23] Wu, D., Wang, X., & Wu, S. (2022). Construction of stock portfolios based on k-means clustering of continuous trend features. Knowledge-Based Systems, 252, 109358. https://doi.org/10.1016/j.knosys.2022.109358

[24] Yan, K., Wang, Y., & Li, Y. (2023). Enhanced Bollinger Band Stock Quantitative Trading Strategy Based on Random Forest. Artificial Intelligence Evolution, 22–33. https://doi.org/10.37256/aie.4120231991

[25] Zuo, Y., & Kita, E. (2012). Stock price forecast using Bayesian network. Expert Systems with Applications, 39(8), 6729–6737. https://doi.org/10.1016/j.eswa.2011.12.035