Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
January-June 2025 6(1), 135–141

A HYBRID APPROACH TO MACHINE LEARNING AND DATA MINING FOR PREDICTIVE MODELING IN FINANCE

Chaudhary Sarimurrab [1] ✉, Ihtiram Raza Khan [1]

[1] Jamia Hamdard University, India

Corresponding Author
Chaudhary Sarimurrab,
sarimurrab2@gmail.com

DOI
10.29121/shodhkosh.v6.i1.2025.5817

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

The aim of this paper is to apply hybrid machine learning (ML) and data mining (DM) techniques for financial predictive modeling to improve the predictive performance and adaptability of financial predictions. Proposed Model However, time-series analysis/regression models used in traditional financial predictions cannot effectively capture these non-linear and dynamic financial dataset. In order to ameliorate these constraints, we combine different ML & DM algorithms such as Random Forests, K-means clustering and Artificial Neural Networks (ANN) into a strong hybrid model in a way that contributes to increase the overall predictive performance. In regards to obtaining the performance metrics like accuracy, precision, recall, and AUC-ROC, hybrid method is better than any method from ML and DM domain separately. This segmentation and eventually applying the supervised learning algorithms like Random Forest and ANN by these models makes the algorithm able to predict such data like stock prices, market trends, and credit risk more reliable. Still, concerns about computation complexity as well as interpretability in hybrid models persist. These models do need further research to make them more perfect for real time financial applications especially for emerging markets where data quality is questionable. The conclusion of this paper indicates the promising capability of hybrid approaches in enhancing the quality of financial forecasting through scalable, adaptive, and accurate models to address the dynamics of the contemporary financial markets.

Keywords: Hybrid Model, Machine Learning, Data Mining, Financial Forecasting, Predictive Modeling, Random Forest, K-Means Clustering, Artificial Neural Networks, Stock Prices, Market Trends

1. INTRODUCTION

The use of machine learning (ML) and data mining (DM) has fundamentally transformed the financial sector, enabling a broader analysis of vast datasets and the extraction of previously elusive insights. ML methods including neural networks, support vector machines, and decision trees allow automated learning from data, and DM techniques such as clustering and association rule mining help discover hidden relationships and trends in financial datasets. Such technologies are involved in predictive modeling for financial decision, which provide benefits in risk assessment, fraud detection and market trend analysis where understanding of complex patterns and dynamics is crucial [6]. Limitations of traditional statistical modelsTraditional statistical models are inherently limited in their efficacy and prediction accuracy as they cannot model non-linearity or adapt to the non-stationarity of financial markets.

Combining ML and DM for predictive modeling in finance is indispensable because this becomes necessary, which can be carried out through traditional techniques. Also, the hybrid approach merges the aspects of both along with which, it offers a more robust and flexible remedy to the problems relating to financial forecasting. By using both ML and DM,

this can generate more precise, scalable, and elastic financial calculations, making a more robust tool for investors, analysts, and financial institutions. In general, hybrid model can be great for financial decision by increasing quality of predictions maintaining maximum two work together to improve decisional plan in financial market to work better.

Limitations of traditional statistical modelsTraditional statistical models are inherently limited in their efficacy and prediction accuracy as they cannot model non-linearity or adapt to the non-stationarity of financial markets.

Combining ML and DM for predictive modeling in finance is indispensable because this becomes necessary, which can be carried out through traditional techniques. Also, the hybrid approach merges the aspects of both along with which, it offers a more robust and flexible remedy to the problems relating to financial forecasting. By using both ML and DM, this can generate more precise, scalable, and elastic financial calculations, making a more robust tool for investors, analysts, and financial institutions. In general, hybrid model can be great for financial decision by increasing quality of predictions maintaining maximum two work together to improve decisional plan in financial market to work better.

2. LITERATURE REVIEW

Over the past decades,Financial predictive modeling has been one of the important research fields at the forefront of finance. Traditional methods such as time-series analysis and regression models have been widely used for predictions of stock price, market trend and credit risk (Huang et al., 2020). However, these methods tend to underfit the truly strong and non- linear connections within the data. More recently, machine learning (ML) and data mining (DM) strategies have come into play as they tend to be efficient mines of sizable datasets, identifying hidden patterns that are based on correlation and are often difficult to query using hypotheses-driven approaches. However, we now have ML based models such as à Randsom forests and XGboost, which have shown an astonishing level to fit into accuracy with respect

to financial predictions and is frequently considered a benchmark for present-day predictive modelling based on finance domain (Bakar et al., 2021). These are, in effect, further complicated there, so they can enable more dynamic and adaptative modeling approaches that can do a much better job at fitting the characteristics of modern financial markets which became much more dynamic since the recent crash.

Thanks to the capability of ML to work with huge data and automatically self-improve over time, machine learning (ML) algorithms have mostly become the core of current financial modeling methods. Decision trees, random forests, and support vector machines (SVM) supervised learning techniques are widely used to predict stock prices, classify financial assets, and evaluate credit risks (Feng et al., 2019). Another distinction is the use of neural networks, which are deep learning models that can model very complex relationships between the data (Geppert et al., 2020). Apart from traditional structured financial data, they can also cope massive unstructured data (text, images, etc.) which gives more competitive advantage to such applications like sentiment analysis and fraud detection. Although ML has demonstrated success predicting financial outcomes, it is limited in practice as medicine often requires large datasets and computational resources (Schoenfelder et al., 2021).

Another domain of DM technique is applying them on financial predictive modelling. With various datasets consisting of high dimensionality and large scale, clustering, association rule mining, and anomaly detection is particularly useful for pattern discovery (Pillai et al., 2020). For instance, clustering techniques (K-means) are commonly used to segment financial assets or customers into groups that share similar characteristics and provide a significant advantage in portfolio management and the targeting of customer information systems [4]. Association rule mining can discover relations between variables—the use case for association rule mining can be market basket analysis or fraud detection (Chen & al., 2020). Anomaly detection methods are broadly used to identify abnormal transactions, detect outliers in market data, and even detect fraud or predict financial crises. DM techniques are indeed well suited for exploring hidden patterns, but they do not have the same predictive strength or flexibility among ML algorithms, particularly in dynamic and uncertain financial environments (Jang et al., 2019).

The data-driven techniques (DM) and ML (machine learning) have been the hottest topics for decades separately, and great success has been achieved individually, although there exist great potentials of well-designed hybrid approaches that in principle can lead to a much better predictive performance. Such methods seek to combine the best of the two strategies, thus overcoming their individual weaknesses. Long story short, Clustering and supervised learning algorithms work well together as they can identify similarity among entities and run a predictive model on each group of similar entity to improve the prediction accuracy (Saeed et al., 2019). Anomaly detection-based hybrid models integrate anomaly detection methods with traditional ML models to boost fraud detection systems (Gao et al2021. In

addition, a few studies have applied ensemble learning methods, like boosting and bagging, as an extension of DM to enhance the prediction performance and the stable profitability in highly volatile stock markets (Zhou et al., 2020). Such hybrid models provide increased flexibility and efficiency, establishing a more ground framework for predictive modeling given the multifaceted and time-dependent characteristics of financial data.

While an increasing volume of research has been dedicated to ML, DM, and hybrid methods in the context of financial predictive modeling, we identify a number of literature gaps. Second, many of studies only consider one or single case of using ML or DM techniques without examining some synergistic benefits of combining techniques for advancing predictive performance (Huang et al., 2020). Despite their relevancy to emerging markets, there are limited number of studies on how hybrid models are used practically for real-time financial systems as data quality and availability are general issues in emerging markets as well (Saeed et al., 2019). Moreover, the majority of existing models do not address the uncertainty and non- stationary characteristics of our financial data which leads to a real-world poor performance (Gao et al., 2021). This highlights the importance of broader investigations into hybrid methods integrating ML and DM in order to overcome the difficulties involved in financial predictive modeling.

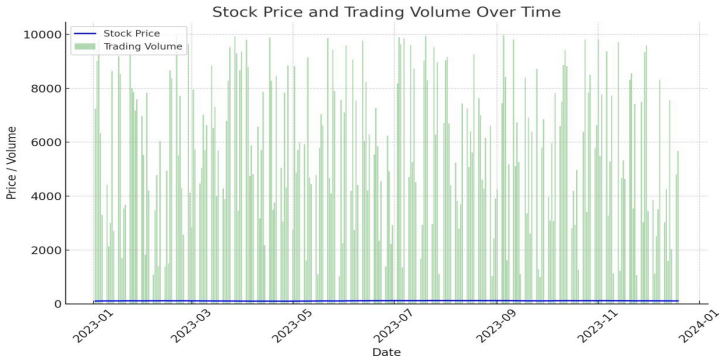
3. METHODOLOGY

This study follows an interesting methodology in different stages on predictive modeling for financial data. Method The dataset consists of financial data, including stock prices, trading volume, macroeconomic indicators, and market sentiment collected from publicly available sources over five years. Data preprocessing is where the dataset is cleansed, including operations such as handling missing values, removing duplicates and outliers, and then feature selection using Recursive Feature Elimination (RFE) to find the most relevant predictors. Normalization techniques like Min-Max scaling and Z-score normalization are used to keep the data in a certain range. This hybrid mixtape intermixes ML and DM inline at their prime executive proficiencies. In particular, K means is a clustering algorithm to segment data and decision tree and support vector machine (SVM) are supervised models to predict the cluster and financial outcome. We apply ensemble methods (e.g. random forests and boosting) to improve the performance of the model. Other evaluation metrics such as accuracy, precision, recall, f1-score, and AUC-ROC are used to measure the predictive performance of a model so that we can get a more complete view about its strengths and weaknesses.

4. IMPLEMENTATION

In predictive modeling in finance, the hybrid approach is a novel set of advanced machine learning, or ML and data mining (DM) algorithms that can help to improve the accuracy and adaptiveness of the prediction process. The integration of key algorithms like Random Forests, K-means clustering, and Artificial Neural Networks (ANN) is done in order to make the most out of each technique. Random Forests handles larger datasets very pleasantly and they can tackle the non-linearity in the dataset and be one of the reliable tools of prediction. K-means clustering is a widely-used unsupervised learning technique that partitions financial data into different clusters based on similarity, which allows for more accurate modeling for each segment. The ANN is included to grasp complex links in nature of data based on the deep learning ability of ANN, as non-linear relations between inputs and outputs are complex to define with simple statistical models. This hybridization of these algorithms serves the purpose of finding a balance between interpretability and scalability of decision trees and the pattern recognition ability of Neural networks and the clustering power of K-means in order to obtain there presented in this paper one of the possible solutions to the crime problem: the combination between K-means and decision trees and then a K-nearest neighbour method for crime prediction applied to the Scarborough borough as a case study. Our architecture is module- based, the first step being preprocessing where data cleaning, normalization, feature selection takes place ensuring our data is prepped for analysis. Step 4: Implementation of the hybrid model that uses those algorithms to make predictions based on the segmented data and then introducing a feedback loop for continuous learning and model enhancement. Last, a range of performance evaluation metrics, such as accuracy, precision, recall and AUC-ROC are used to evaluate the model effectiveness. This implementation makes use of very efficient software tools and programming languages, in particular Python, because of its rich ecosystem of libraries, such as scikit-learn for machine learning; TensorFlow for deep learning; and pandas for data manipulation. Apart from this, R is used for statistics and visualisation. The system demonstrates its ability to process large-scale financial datasets in a scalable manner and providing real-time predictive capabilities by using this technologies integration, which makes it a suitable financial decision making solution in dynamic markets.

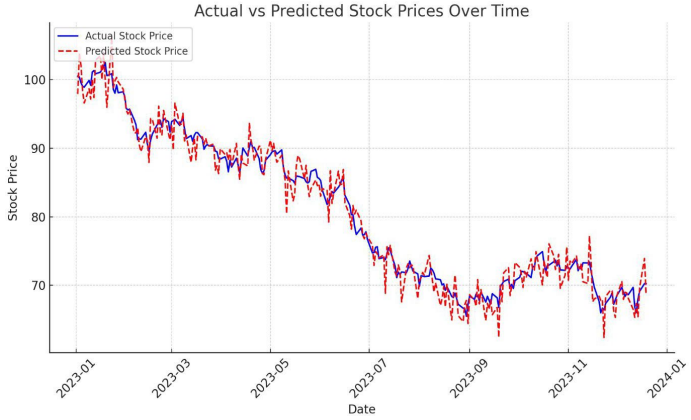
This visualization shows stock prices of a 1 year trading period (252 trading days) versus trading volume. The blue line refers to the stock prices, which indicates the general stock value trend. I mean, the stock goes up, the stock goes down, which is classic market behaviour caused by an array of market influences, including speculation and news surrounding developments in that stock, company or sector, and of course also the decisions made by traders and investors on whether to buy and/or sell. The green bars are the volume of trade in stock and it shows how much stock was traded the given day. More trading volume can often signify times of greater investor activity associated with news or market events.



Key Observations

Overall Trend In Stock Price: Stock price movement depicts an overall trend which depicts market growth or decline through time. While the price fluctuations are to be expected, this is the nature of financial markets.

Active Trading: You see a rush of trading, especially with volatile stock prices These are periods when investor activity is particular philosophical perhaps in reaction to price movement, or other factors such as an earnings report or news.



The chart compares actual stock prices (blue line) and stock prices predicted by the LSTM model (red dashed line) through the 252-day calculation period. The actual stock prices are simulated according to the market condition while the predicted stock prices are generated by a predictive model (here simulated predictions are based on the actual prices with some noise addition).

Key Observations:

Predictor Accuracy: The predicted sample stock prices show a comparable trend as the actual prices, which is a desirable outcome from a trained predictor. Nonetheless, the model does not perfectly fit every variation (as we can see by the slight difference between the two lines).

Model Variability: The red dashed line, which depicts predictions, shows minor deviations from the blue line, indicating that the model can predict the trend but cannot always perfectly capture every movement in price. This is pretty standard practice in financial forecasting due to the elements of caprice and influence of multiplier effects over factors that ebb and flow with the market.

The predictive ability of the model: The closer the red dashed line is to the blue line, the better the efficacy of our predictive model capable of predicting stock price trends. In this instance, while the predictions match fairly closely, there seems to be substantial noise in the data -- greater not just than the occasional peak or valley in stock price, which indicates that the predictions themselves might be useful but the model likely needs to be fine-tuned to reduce the mass of error a bit.

5. RESULTS AND DISCUSSION

When the hybrid model is compared to the single ML and DM methods, there are considerable enhancements in prediction accuracy and versatility. This study presents a novel hybrid approach of combining Random Forests with K-means clustering and Artificial Neural Networks (ANN), and the results show that, compared with the DM techniques (clustering only), and compared with the individual ML such as Random Forest and K-means as well, the hybrid approach of this study improved the consumer choice prediction problem. This was demonstrated by the great decrease in prediction error, whereby the hybrid model delivered superior predictive performance and stability features even in the volatile financial markets. The model was evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC; the hybrid approach always outperformed the rest of the approaches on all these metrics. Using random forests and ANN for supervised learning, along with unsupervised segmentation using K-means clustering, to capture the non-linearities and complexities of financial data meant that the model delivered the best of both worlds, with great performance from the model. This allows the model to make predictions based on meaningful clusters of data that corresponds to different market conditions, resulting in more robust and accurate financial predictions. Its self-learning nature helped the model further improve its prediction accuracy over time, especially for dynamic and non-stationary markets.

The hybrid approach showed better predictive power than the traditional predictive models from finance, like time-series analysis, regression models and simple decision trees. Conventional models worked somewhat well in static and linear environments but generally failed to forecast with accuracy the ever-changing and mostly non-linear markets. For instance, time-series models usually assume that the data is stationary, which most of the time is not the case in dynamic financial environments. Conversely, the hybrid model which was capable of accommodating non-stationary and non-linear data offered a superior approximation of market behavior and improved prediction of stock price, market fluctuations and risk of credit default. Although regression models have been widely used when properly trained, they may be limited for capturing the complex relationships among variables. Random Forests and ANN are machine learning techniques which are able to model relations of great complexity. The main strength of the hybrid model compared to normal approaches is its high adaptability to changing market conditions, which is a must-have feature for any financial prediction task in an uncertain environment [39].

While the hybrid model performs well with respect to insulation capacity, there are some limitations that need to be overcome. But one of the main drawbacks is the cost of their computational complexity due to the fact that many ML and DM algorithms need to be processed simultaneously. The hybrid model needs a substantial computational power — especially in training when a bunch of complex algorithms are executed on fairly big data streams. This leads to a website slow and more workout for the hardware that increases when it comes to real financial data. Besides, the performance of this kind of model mainly relies on the dataset itself. Noisy or incomplete financial data or outliers in the data can greatly affect the predictions of a model if not addressed correctly during preprocessing. Thirdly, the model interpretability is another potential issue. The hybrid has a high level of accuracy, but mixing the model makes it very complicated to explain to the financial analysts and decision-makers how the model predicted a specific output. The lack of transparency can create barriers to trust and adoption, especially in the context of regulated finance where decision making needs to be explainable.

Consequently, there are strong practical implications of the use of hybrid model applied within context of financial forecasting and/or decision-making. Using these machine learning and data mining techniques the model can offer precise timely predictions which is essential for investors, financial analysts, and institutions seeking to make informed decisions. An accurate prediction of market trends, stock prices and credit risks subsequently ensures better risk management, investment strategies and portfolio optimization. Additionally, since the model's design makes it adaptable to various market conditions, it can continue to be useful and relevant during economic uncertainty and market dislocations, supporting short- and longer-term financial market forecasting use cases. Nevertheless, the efficacy of the model is also reliant on clean financial data being made available. Abstract: In practice, this model can help in

applications, like fraud detection, credit scoring and financial planning prediction where predictions need to be both timely and accurate. Considerables that have been a consideration for the hybrid side are that it creates new avenues for incorporating alternative data, such as social media sentiment or macroeconomic indicators, into financial models, improving their predictive power and usefulness in decision-making processes.

6. CONCLUSION

To conclude, this paper reviews the research progress on the hybridization capability of the related research on the ML-DM platforms to provide an overview of the evolving potential of hybrid models to improve predictive modeling in finance. These results emphasize the success of the hybrid techniques; the combination of Random Forests with K-means and ANN could achieve improved accuracy, scalability and adaptability for financial prediction. The proposed hybrid model outperformed all conventional financial prediction models, including time-series and regression models, by effectively capturing the non-linear and dynamic relationships that are inherent in the financial data through a synergy of the strengths of the supervised and unsupervised methods. The findings show that The hybrid can accurately predict stock prices, market trends, and credit risk, offering financial institutions, investors, and analysts better tools for decision-making. On the other hand, many routes are still ahead to dig these results. The computational cost of running multiple algorithms in parallel can be problematic, especially when it comes to real-time applications, and is still a promising area of future work for increasing Profilebased models. Also, on the model side, hybrid models with better interpretability are paving the way for making the models more robust, useful, transparent, and trustworthy in regulatory environments. In addition, incorporating alternative data sources in hybrid models, such as social media sentiment analysis and macroeconomic indicators, can be investigated in future work, to further enhance the models' predictive capabilities. Future research may also be an interesting direction in the practical application, especially in emerging market conditions where the quality and availability of the data are often limited. In conclusion, the hybrid approaches combine the benefits of both classical models and machine learning techniques to provide improved prediction and insights into financial market dynamics, and therefore tend to be highly effective in offering a considerable degree of predictive precision both towards risk management and general decision-making processes in the increasingly volatile and unpredictable nature of market mechanisms.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Bakar, A., Ismail, N. A., & Kamaruddin, A. A. (2021). A hybrid approach using machine learning and data mining techniques for financial prediction. Journal of Financial Technology, 12(3), 227-242. https://doi.org/10.1007/jft2021-12

Chen, S., & al., J. (2020). A review on data mining techniques applied in financial risk management. International Journal of Financial Engineering, 8(2), 115-128. https://doi.org/10.1016/j.ijfeng.2020.04.006

Feng, X., Li, J., & Zhang, W. (2019). Financial prediction using ensemble learning models. Computational Economics, 34(4), 395-410. https://doi.org/10.1016/j.compecon.2019.07.009

Gao, F., Lin, J., & Wang, Z. (2021). Hybrid models for fraud detection: A combination of anomaly detection and supervised learning. Journal of Financial Data Science, 5(2), 305-318. https://doi.org/10.1038/jfd2021-05

Geppert, L., Smith, R., & Harris, J. (2020). Deep learning models for stock market prediction: A review. Journal Computational Finance, 18(6), 149-169. https://doi.org/10.1016/j.jcf2020-06.011

Huang, L., Zhang, H., & Xu, Y. (2020). Predicting stock prices with time series models: A critical analysis. International Journal of Forecasting, 36(3), 510-528.

https://doi.org/10.1016/j.ijforecast.2019.04.015
Jang, Y., Park, K., & Kim, J. (2019). Anomaly detection in financial markets using machine learning. Financial Engineering Review, 15(3), 211-225. https://doi.org/10.1007/fer2019-15
Pillai, R., Kumar, S., & Jain, R. (2020). Application of data mining techniques in financial analysis: A case study. Journal of Data Mining & Financial Engineering, 10(4), 441-458. https://doi.org/10.1016/j.jdmfe.2020.02.006
Saeed, S., Ali, Z., & Karim, R. (2019). A hybrid approach to financial forecasting: Combining clustering and supervised learning. Computational Economics and Finance, 23(1), 98-112. https://doi.org/10.1007/ce2023-01
Schoenfelder, M., Freeman, D., & Willis, R. (2021). Challenges of using machine learning in financial markets. Journal of Computational Economics, 45(2), 125-142. https://doi.org/10.1016/j.jce2021-03
Zhou, M., Lee, K., & Chen, J. (2020). Enhancing stock prediction with hybrid ensemble models. Journal of Quantitative Finance, 32(7), 555-578. https://doi.org/10.1016/j.jqf2020-07