12th International Conference on Business Economics, Marketing & Management Research (BEMM-2025)
Proceedings Book Series -PBS-Vol 22, pp.162-167

Data Mining and Machine learning models comparison for the prediction of financial stock prices

Cheima Ali Bensaad[#*1], Rashemi Bena[*2], Yasir Iqbal[#3]

[#]*School of Physics Engineering and Computer Science, University of Hertdforshire ,United Kingdom*
[1]Cheima.ali-bensaad@herts.ac.uk/cheima.alibensaad@gmail.com
[3]y.iqbal2@herts.ac.uk/yasiriqbaluk21@gmail.com

[*]*School of Computer science , Mathematics and Digital Media , London Metropolitan University United Kingdom*
[2] *rex0020@my.londonmet.ac.uk*

Abstract— **Accurately predicting stock prices is challenging due to the financial market's complexity and volatility. By integrating data mining and machine learning models, this research investigates and compare the predictive performance of four distinct models supervised, unsupervised, deep machine learning, and neural network models, all implemented in Python for predicting the stock prices of five different company portfolios, which holds significant financial and economic implications, especially in dynamic markets.**
The evaluation results show a significant variance in predictive performance.Multi-Layer Perceptron (MLP) Regression and K-Nearest Neighbours (KNN) achieved the highest metrics, with R-squared (R^2) values of 0.99968 and 0.96415, and low Root Mean Squared Errors (RMSE) of 0.00302 and 0.01402, respectively. The findings indicate that while each model exhibited unique strengths across various aspects of the prediction task, MLP consistently delivered the most precise results, followed by KNN, Support Vector Regression (SVR)with 61%, and at last the Long Short-Term Memory (LSTM) with only 20% which depicts different results from related works.
Additionally, visualisation comparisons revealed key strengths and limitations across models, underscoring their potential in enhancing stock price prediction and trend analysis. These insights can help investors make more informed decisions for short term within dynamic market environments.

Keywords— Long Short-Term Memory (LSTM), Multi-layer Perceptron Regression (MLP R), Support Vector Regression (SVR),K-Nearest Neighbors (KNN), stock priceprediction , Mean Square

I. INTRODUCTION

The prediction of stock prices has been a classical problem in finance and economics for decades. However, the introduction of new deep learning and machine learning techniquesdue to their promise of higher accuracy and improved trend analysiscan significantly enhance informed decision-making in stock buying and selling, portfolio optimization, and financial risk management. In this context, the primary aim of this research is to investigate the effectiveness and evaluate the performance of various predictive models in forecasting stock prices for five companiesincluding :Tesco, Marks & Spencer, Lloyds, Barclays, and EasyJet over an 18-months period, using data manually compiled from Yahoo Finance and using Python language. To address the research problem of "**What are the trends and effectiveness of different machine learning models in predicting the closing prices of a five-company portfolio? »**
The advent of modern computational technologies has introduced more sophisticated and accurate predictive methods, with machine learning and deep learning models playing an increasingly significant role in stock price forecasting[1]. For this reason, a model comparison is conducted to evaluate each model's strengths, weaknesses, and overall suitability for informed decision-making.
The background of this study, involving the introduction of predictive models, is compared with the following previous works. Medarhi et al. (2022) [2] attempted to predict stock prices using six machine learning methods. The results showed that no single machine learning technique had a clearly dominant performance; however, MLP and LSTM techniques demonstrated the best outcomes. A limitation of their study was the lack of a comprehensive comparison. In a detailed analysis by Gupta et al. (2023) [3], LSTM provided more accurate predictions across most sequence lengths compared to other models. The study by Mehtab et al. (2020) [4] applied Deep Learning and Natural Language Processing, and the final results indicated that among KNN, NIFTY, Random Forest, SVM, and

ISSN: 2961-6611

CNN, the LSTM model showed strong performance.Wide Traditional methods have been conducted for the analysis of the financial market, which highlighted time series forecasting[1], This initiated the thought of considering new approaches and models.

II. IMPLEMENTATION

This study follows the data cycle of the CRISP-DM data mining framework[5]. The dataset comprises historical stock price data, including features such as opening price, closing price, highest price, lowest price, and volume, collected from Yahoo Finance. Data preprocessing involved merging data from all companies, scaling using the Min-Max normalisation technique, and performing feature selection through correlation analysis. The analysis revealed that the closing price and volume are the most significant target variables [6]. Data visualisation was employed to explore distributions, trends, and patterns. In the modelling phase, deep learning, neural networks, and data mining approaches were applied, with model selection based on feature extraction. The dataset was split into 80% for training and 20% for testing and validation.The hyper-parameter changes used for the models [6]are shown in the figures below. These include the activation function, number of hidden layers, and number of neurons for MLP and LSTM; the value of *K* for KNN; and the kernel and complexity function for SVR. Dropout techniques were also applied to mitigate overfitting and underfitting. Model comparison and validation are based on regression metrics such as RMSE, MSE, and R^2.

```
# Define the MLP model
mlp = MLPRegressor(hidden_layer_sizes=(100, 50),
                   activation='relu',
                   solver='adam',
                   max_iter=1000,
                   random_state=42)

# Define the SVR model
svr = SVR(kernel='rbf', C=100, epsilon=0.1)
```

```
# Define the LSTM model
model = Sequential()
model.add(LSTM(units=50,
model.add(Dropout(0.2))
model.add(LSTM(units=50,
model.add(Dropout(0.2))
model.add(Dense(units=1)

# Compile the model
model.compile(optimizer='adam', loss='mean_squared_error')

# Define the KNN model
knn = KNeighborsRegressor(n_neighbors=5)
```

Fig. 01: Models implementing and hyperparameters choice in python (source: Auhors)

III. RESULTS AND DISCUSSIONS

Aspart of data visualisation and pre-analysis of the trend and patterns over a period of 10 years, the following graphsconfirm the stock market's volatility, with significant price increases in 2018 and fluctuations through 2020. When considering the portefeliomoving average (MA) it appears the stock of EZI is the most fluctuating with dramatic downtrend and cycle every towyears, still the prices is well higher than the other 04 stocks.

ISSN: 2961-6611

Fig. 02: Distribution and the trend of the historical closing prices of five companies (source:Authors via Python)

The volatility and the large difference in the prices of the portfolio assets have impacted the distribution of the data. The adj_close price distribution is right-skewed, indicating the existence of low-value stocks, while MA_10 and MA_50 exhibit less skewness, offering a smoother trend representation.The trend and pattern may exhibit the non-stationarity of the data [7]; however, using machine learning models does not need to deal with stationarity in advance, as it can capture the hiddencharacteristics of the data.

Understanding how effectively a model performs in predicting financial time series is crucial (Sezer et al.[8]. To assess this, the predicted values are plotted against the actual values to evaluate how well the models capture trends and align with historical data. The modelling results are presented in the following prediction figure, which highlights the performance of different models in predicting stock prices over an 18-month period of historical data.

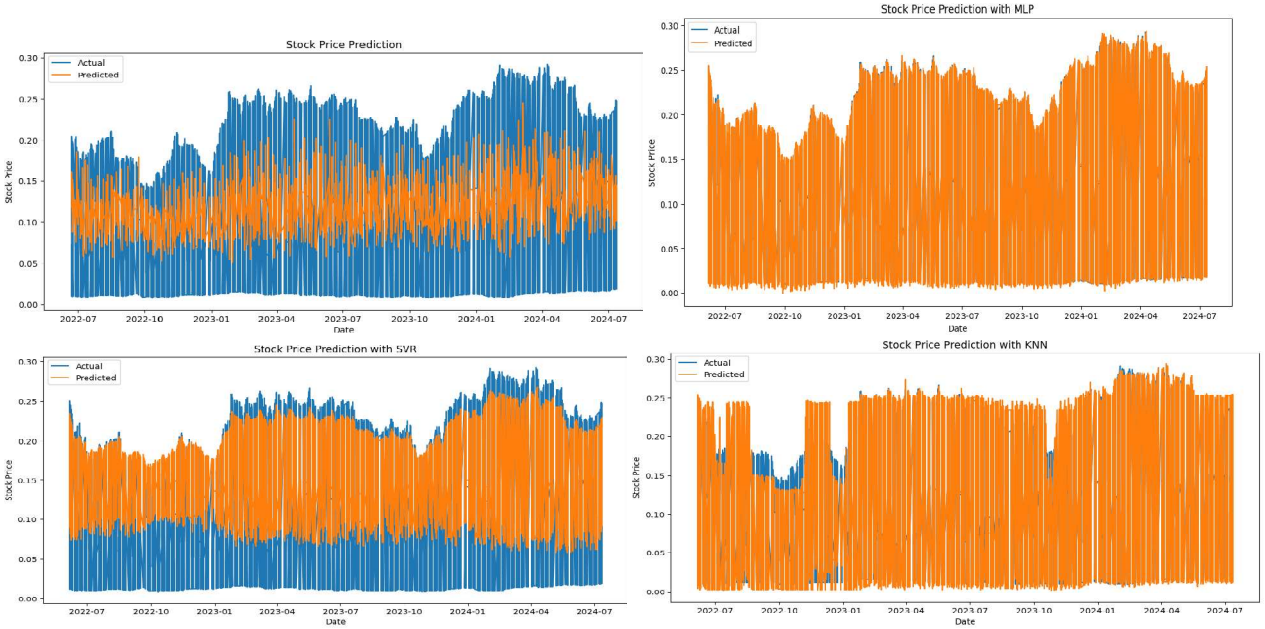

Fig.03 : Prediction of 5 company s stock prices with 4 machine learning models (source:Authors python output)
Caption : up-left :LSTM, Upright : MLP, donwleft : SVR, down right : KNN

Both MLP and KNN demonstrate strong predictive performance, with the predicted values (orange line) closely overlapping the actual values (blue line), particularly in dense regions. This high degree of overlap indicates that these models have learned effectively and offer the best fit for accurate, instantaneous stock price prediction. In contrast, the SVR model predictions generally follow the overall trend but display some notable deviations. LSTM, meanwhile, performs less effectively, struggling to capture smaller fluctuations and extreme price values [5].

This result deviates from findings in related studiesparticularly with regard to LSTMsuggesting that its predictions remain noisier than actual stock prices, thereby reducing accuracy, especially in the presence of extreme market anomalies. Additionally, the KNN model fails to fully capture trendsduring the mid-2023 period.

The following table presents validation results using regression metrics for both training and testing data to further highlight the performance of different models in stock price prediction.

Table (1):Evaluation matric of Machine learning models

Model	Train R^2	Train RMSE	Train MSE	Test R^2	Test RMSE	Test MSE
LSTM	0.2750	0.203768	0.0415	0.20263	0.00438	0.00438
MLP Regression	0.99967	0.00432	1.868×10^{-5}	0.99833	0.00302	9.129×10^{-6}
SVM	.96788	.04275	0.00183	0.61121	0.04617	0.00213
KNN	0.99968	0.00427	1.821×10^{-5}	0.96415	0.01402	0.00020

(source: Authors using Python)

ISSN: 2961-6611

The evaluation metrics further support the strong visual alignment observed for MLP and KNN, with both models achieving near-perfect R^2 scores of 99% and 96%, respectively. In contrast, while SVR and LSTM exhibit lower RMSE and MSE on training data compared to test datasuggesting good learning from the training sethey also indicate that these models may not generalize as well to unseen data. This discrepancy raises concerns about overfitting, where models perform well on training data but struggle with real-world predictions.

LSTM, in particular, raises concerns of underfitting, registering the lowest R^2 score of 0.20, which falls well below the acceptable range of [0.5–1]. This result highlights the need to re-evaluate and fine-tune the model's parameters to improve generalization performance [6]. The LSTM model was only able to explain 21% of the variability in unseen data, indicating poor predictive power.

Despite its popularity in time-series forecasting, LSTM struggles to handle the inherent volatility of stock market prices. During periods of sharp increases or decreases in actual prices, the model fails to capture the full magnitude of these changes. This exposes a clear gap in its ability to accurately predict extreme price movements, suggesting that further optimization is required to improve its practical, real-world performance.

LSTM is typically designed to handle sequential data and capture long-term dependencies through its unique architecture, which includes memory cells capable of storing information over extended periods [11] It is particularly effective in managing complex time-series data. However, in this study, LSTM did not perform as expected. One possible explanation is that the model failed to capture the stock price dynamics effectively, potentially due to the regularization technique appliedspecifically, a 20% dropout rate, which means that 20% of the neurons were randomly deactivated during training. Although dropout is intended to prevent overfitting, it may have impacted the model's ability to learn finer patterns. Nonetheless, the use of the Adam optimization algorithm was beneficial in minimizing the loss function, particularly the Mean Squared Error (Kumar et al [7].

In contrast, the MLP (Multi-Layer Perceptron) operates based on the backpropagation algorithm during training. Its architecture consists of an input layer, one or more hidden layers, activation functions, and an output layer. During the forward pass, input data is propagated through the network, and outputs are computed based on current weights. The difference between the predicted and actual output is calculated using a loss function, and this error is used to adjust the weights through backward propagation. This process is repeated iteratively until the model achieves optimal performance (Mehtab, 2020) [2].

Regarding SVR (Support Vector Regression), the model demonstrated an ability to learn and reproduce the general patterns of stock price movement. SVR attempts to approximate the functional relationship between independent and dependent variables by fitting a curve within a defined margin of tolerance around the actual data points [12] .Data points that fall outside this margin are penalized, encouraging the model to generalize well. In this study, SVR utilized the Radial Basis Function (RBF) kernel, which is commonly applied to capture non-linear relationships. However, despite using a regularization parameter (C) set to 100, the model did not perform as well as anticipated. This may suggest the need for further tuning of hyperparameters or kernel selection in future studies (Bhattacharjee & Bhattacharja) [10]

Unlike parametric models, KNN (K-Nearest Neighbors) does not assume any prior distribution of the underlying data. It is particularly effective at capturing local patterns, making it suitable for identifying short-term trends in stock prices. KNN works by memorizing the training data and, upon receiving a new input, calculates distances between this input and existing data points to find the nearest neighbors. The prediction is then based on the average (or majority) of these neighbors. This approach not only simplifies interpretation but also enhances understanding of the model’s outputs (Cheima Ali Bensaad &Beena)[6].

IV. CONCLUSION

This project provides insights into stock market forecasting through the application of various machine learning models and techniques. It offers a comprehensive analysis of how these models perform when applied to real-world financial datasets, demonstrating their potential in stock price prediction.

It is important to emphasise that the models used in this study are designed for prediction rather than forecasting future prices. Machine learning models, particularly neural networks, do not require stationary data in advance, as was evidenced in this implementation. The models were able to learn patterns and relationships within the data that

ISSN: 2961-6611

were useful for predicting the target variable, thereby reducing the need for extensive statistical analysis and lengthy steps typical of traditional statistical learning methods [7].

From the overall evaluation of the models, it is evident that machine learning techniques, especially deep learning models like MLP (Multi-Layer Perceptron), and artificial neural networks[9], provide significant predictive power. MLP regression was able to predict stock prices with the highest accuracy and the lowest error. Both MLP and KNN (K-Nearest Neighbors) effectively capture significant trends and patterns in stock prices.

The performance of MLP and KNN stands out compared to the other models, though there is no common factor or parameter that has been universally applied to both. This could be due to the inherent design of these models, which rely on distance and dispersion of the data, measured mathematically, to minimize errors. This is a likely reason for the small errors observed in these models'predictions.

SVR (Support Vector Regression), as shown in the evaluation, is able to capture the general direction of stock price movement but is sensitive to short-term fluctuations. The model fails to capture smaller variations, leading to more volatility in the predicted values.

While KNN provided a good fit, it was interesting to note that the number of neighbors (K) used in this case was relatively small, contrary to what is often recommended in the literature, which suggests that a higher K value leads to better results. This could be an area for future optimization.

In contrast to previous studies and theoretical concepts, which emphasize that LSTM models are particularly adept at capturing temporal dependencies and mitigating the vanishing gradient problem in recurrent neural networks [9],the LSTM model in this study did not perform as well, especially when compared to the works of Mehtab (2020) and Medarhi (2022) [2]. This highlights that while LSTM has the potential to outperform other models in theory, it may not always be the best fit for every financial data set or market condition.

A key challenge encountered in this study was overfitting and underfitting. Powerful models like LSTM and SVR are highly prone to overfitting on training data, where they perform well during training but fail to generalise to unseen data. To address this, it is crucial to investigate further into regularisation methods, techniques [6][13] and the fine-tuning of hyperparameters to improve the model's ability to generalise across diverse datasets.

REFERENCES

[1] B. D. McCullough, "A spectral analysis of transactions stock market data," *The Financial Review*, vol. 30, no. 4, pp. 823–842, 1995.
[2] I. Medarhri, M. Hosni, N. Nouisser, F. Chakroun, and K. Najib, "Stock price prediction using machine learning: A comparative study," *IEEE Xplore*, Oct. 7, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9934252
[3] A. Gupta, M. Jha, V. Patil, and P. M. Varalakshmi, "Stock price prediction using machine learning," *IEEE*, Mar. 3, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10101226
[4] J. S. Sidra Mehtab, "A robust predictive model for stock price prediction using deep learning and natural language processing," *SSRN*, 2020.
[5] P. Morgridge, *Data Mining Models*, Module Resources, University of Hertfordshire, 2024.
[6] CheimaAli Bensaad and Rashemi. Beena, *Predicting Stock Market Trends Using Machine Learning: A Comparative Analysis*, M.Sc. thesis, London Metropolitan University, UK, 2024.
[7] A. S. Kumar, B. A. Goyal, A. V. Bhardwaj, and P. S. Sharma, "Stock price prediction using time series," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 11, no. V, 2023.
[8] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Applied Soft Computing*, vol. 90, 2020, Art. no. 106181.
[9] G. G. Rajput and B. H. Kaulwar, "A comparative study of artificial neural networks and support vector machines for predicting stock prices in National Stock Exchange of India," *IEEE Xplore*, 2019.
[10] I. Bhattacharjee and P. Bhattacharjya, "Stock price prediction: A comparative study between traditional statistical approach and machine learning approach," *IEEE Xplore*, 2019.
[11] B. D. McCullough, "A spectral analysis of transactions stock market data," *SCISPACE*, vol. 30, no. 4, pp. 823–842, 1995.
[12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[13] K. I. R. Ghulam, F. H. Jamil, Z. Shah, and M. Hameed, "Karachi Stock Exchange price prediction using machine learning," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 8, no. 28, 2021.

ISSN: 2961-6611