ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 6| June 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2024: 8.402 || ISI Value: 1.188

AN EVALUATION OF ML TECHNIQUES ON NASDAQ DATASET FOR STOCK MARKET FORECASTING

Rakesh Kumar Mahapatro, Anooja Ali
School of CSE, REVA University, Bangalore, India - 560064

Article DOI: https://doi.org/10.36713/epra17331
DOI No: 10.36713/epra17331

ABSTRACT

The uncertainty of stock pricing has popularized stock market prediction as a common practice. Forecasting prices in the market are viewed as problematic, as the hypothesis of efficient markets (EMH) explains. According to the EMH, all accessible information is represented in market prices, and price variations are only the consequence of newly available information. The approach to prediction forecasts the market as either positive or negative based on a variety of input parameters. A combination of derived, fundamental, and pure technical data is utilized in stock forecasts to project future stock prices. Algorithms for machine learning (ML) are made to find patterns in data and utilize those patterns to forecast future events. K Nearest Neighbour (KNN) can process relationships between the numerical data, it is particularly effective in numerical prediction problems for predicting changes in stock value the following day. KNN categorizes freshly input data according to how similar it is to previously taught data, and it does this by clustering the data into coherence subsets or clusters. Using the KNN approach in conjunction with technical analysis, the closest neighbour search strategy yielded the desired outcome.

KEYWORDS: *Hypothesis, K Nearest Neighbour, Prediction, Stock prices,*

I. INTRODUCTION

Financial data is thought to be difficult to anticipate or predict, which is why many firms are interested in researching stock market predictions. The EMH is seen as filling the disconnect between financial knowledge and the world of finance. According to the EMH, stocks are never out of balance and make it hard for innovators to make predictions. Moreover, it has been confirmed that share prices do not follow a random walk, and further data is needed to make accurate stock predictions. Apart from buying and selling shares and stocks in the market, every stock has additional characteristics besides price, such as closing price, which is the most crucial factor in predicting a particular stock's price the next day. Every factor that influences stock movements over time has a link and exhibits certain behavior. Predicting stock prices has taken into account a variety of economic elements, including political stability and other unforeseen events.

The fundamental data depicts the company's operations and the state of the market, the pure technical data is based on historical stock data [1]. We believe that by combining this data about a business with its shares, we ought to be able to forecast the stock's future price [2]. A data set can be divided into a training and a testing set for classification techniques that use ML algorithms. KNN compares a given test object with its training set of data using similarity measures. A record with n characteristics is represented by each data object. KNN chooses k training data set records that are most similar to the unknown records to forecast a class label for the unknown records. ML enhances prediction by learning from historical data, identifying patterns, and applying this knowledge to new data.

The rest of the article is planned as follows; section 2 will represent the review of some literature that has already been published on using KNN for stock prediction, section 3 will describe the research procedure used and analysis that was conducted, section 4 will show the description of the data used and the results we obtained, and finally the conclusion is seen in section 6

II. LITERATURE REVIEW

Based on previous stock data, EMH theories state that the future stock price is unpredictable. When fresh data enters the system, the imbalanced stock is detected right away and swiftly removed with the appropriate price adjustment [3]. To predict the stock price in a semi-strong EMH, all available information is employed along with past data. The stock price is predicted using all available data in the strong EMH, including historical data as well as both public and private data like insider knowledge. However, the random walk theory contends that stock prices are independent of historical stock performance [4].

The author of the study [5] tried different lookback periods in their model to forecast future stock value and get more accurate predictions. Additionally, they have employed a variety of algorithms, including GRU and LSTM [6]. Using stock data from two well-known banks listed on the Nepal Stock Exchange (NEPSE), they discovered that GRU outperforms the other two models in terms of accuracy [7]. They have determined that a look-back time of five to fifteen provides findings that are significantly closer to the real value based on their examination of look-back periods.

Searches for the optimal method among a set of algorithms in its library are carried out using the Fast Library for Approximate Nearest Neighbours (FLANN). Time series analysis is a widely used method because the stock market is a very time-dependent industry with minute-by-minute price fluctuations. The ARIMA (Auto Regressive Integrated Moving Average) model, which was projected by [8], is a widely used technique for time-series modeling. However, because it is a linear model, it is not suitable for stock market prediction because it cannot account for the fluctuations in the data set caused by high market volatility.

The method of linear regression is classified as supervised learning in ML [9]. It predicts values well inside the range rather than categories; it creates a linear relationship among the dependent and independent variables; it is not very effective with non-linear data sets because of outliers; researchers have used this algorithm to predict stock market values and found that, when applied to daily stock values, there are significant issues that need to be addressed [10]. Information from the internet and economic news sources may have an impact on investor behavior, and stock movements may be predicted using ML algorithms [11-12]. The data sets are subjected to feature selection and spam tweet reduction to enhance prediction performance and quality.

An extensive investigation of the fundamental link between macroeconomic conditions and the KSE market has been conducted by the study reported in. Likewise, studies, such as [13], have shown that a person's mood is a major factor in their decision-making. If social media is used to gauge public sentiment, this may also be used to forecast how people would vote over whether to participate in the market and, consequently, how the market will perform.

Blockchain technology is being used by stock markets all around the world to transact business quickly. A portion of the nation is still getting ready to employ blockchain technology. The tracking of securities lending, margin financing, and system risk monitoring are all greatly enhanced by this technology [14]. Many banks and other financial institutions are devoting time and resources to this technology to enhance their offerings and give their customers safe and secure spaces. NASDAQ, Deutsche Bank, and DBS are a few well-known banks and financial organizations that have used blockchain technology [15].

III. METHODOLOGY

KNN is a straightforward and widely used ML technique primarily used for classification and regression tasks [16]. It works on the similarity principle, which holds that comparable occurrences will probably have similar results. KNN is a type of instance-based learning, meaning it does not construct an explicit model but instead memorizes the training dataset. When a prediction is needed for a new instance, the algorithm finds the K training examples closest in distance to the new instance. The closeness of instances is typically measured using distance metrics [17]. KNN averages the values of the K nearest neighbors to forecast the value for the new instance in regression problems.

KNN is a regression technique that forecasts a new data point's result by averaging the values of its K nearest neighbors [18]. This makes KNN regression effective and easy to understand. KNN regression doesn't assume any underlying data distribution. It directly uses the training data to make predictions [19]. The parameter K plays a crucial role in the algorithm's performance. A small K can lead to high variance (overfitting), while a large K can introduce high bias. However, its performance is highly dependent on the choice of K and the distance metric, and it can be computationally expensive.

Advantages of KNN Regression

- **Simplicity**: Easy to implement and understand.
- **Flexibility**: Can model complex relationships without requiring a parametric model.
- **Adaptability**: Suitable for various types of data, given an appropriate distance metric and K value.

Predicting stock prices is a complex task that can be approached using KNN regression. KNN can be applied to this task using Python and the scikit-learn library. The steps are mentioned in below table 1.

Table 1: Execution steps

Step 1:	**Import Libraries**: Essential libraries for data manipulation, modeling, and evaluation.
Step 2:	**Load and Preprocess the Data**: Read the dataset, select features, and split the data into training and testing sets.
Step 3:	**Apply KNN Regression**: Initialize the KNN regressor, train it on the training data, and predict the test data.
Step 4:	**Evaluate the Model**: Calculate the mean squared error to assess model performance and optionally visualize the results
Step 5:	**Optimize the Hyperparameters**: Use grid search with cross-validation to find the optimal number of neighbors (K).

IV. DATASET AND IMPLEMENTATION

The NASDAQ stock exchange provided the sample data, which was taken from Yahoo Finance. Seven chosen firms that were listed on the NASDAQ stock exchange had their stock data included in the study sample. The NASDAQ-listed stocks are represented in Table 2. The period of the data sample is from Jan 1, 2011, to April 16, 2017, each of these companies has six attributes including Date, Opening and adjusted closing price, High, Low, and state change of the stock. Based on the KNN algorithm, the primary component influencing the forecast process for a particular stock is the closing price. The details about the data used in training and testing are mentioned in Table 3.

Table 2: The Stocks used that are listed in NASDAQ

Stock name	Company
AMTD	TD Ameritrade
AMZN	Amazon.com Inc
DIS	The Walt Disney Company
SBUX	Starbux
TWLO	Twilio
Yahoo In	Yahoo In

Table 3: Final result and accuracy

Stock	Data used	Accuracy (in %)
AMTD	Train: 2538 Test: 1308	72.31
AMZN	Train: 2575 Test: 1271	74.18
DIS	Train: 2536 Test: 1310	67.29
SBUX	Train: 2576 Test: 1270	68.53
TWLO	Train: 138 Test: 64	65.27

Fig 1 is the stock movement representation of AMTD and fig 2 represents the predicted and actual trend in stock prices. We have considered the stock representation for DIS in fig 3 and fig4 is the actual and predicted representation.

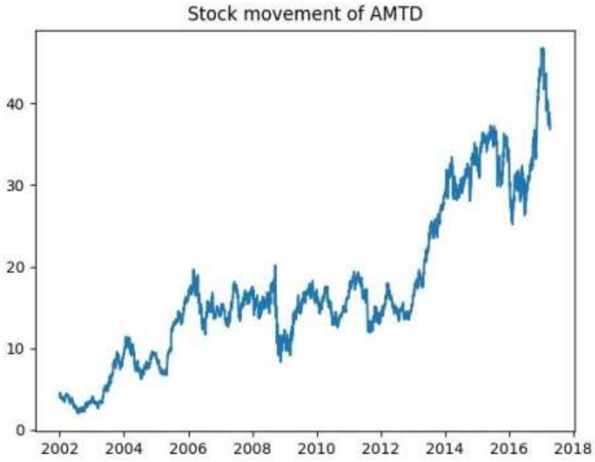

Figure 1: Stock movement representation of AMTD

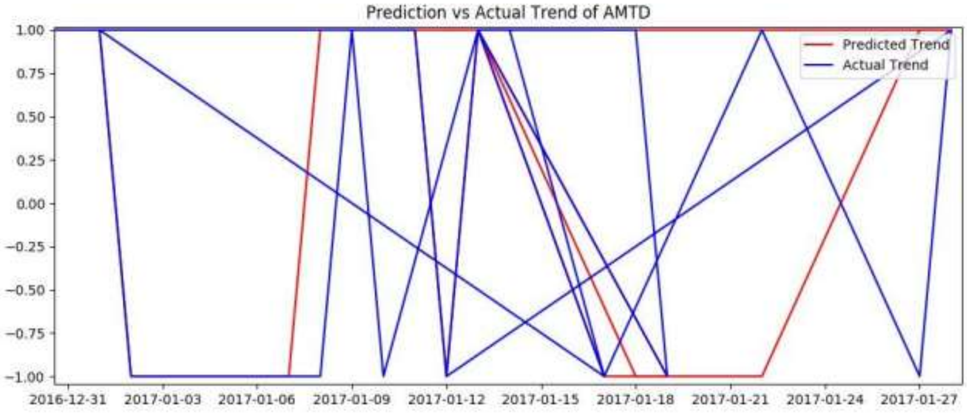

Figure 2: The predicted and actual trend in stock prices for AMTD.

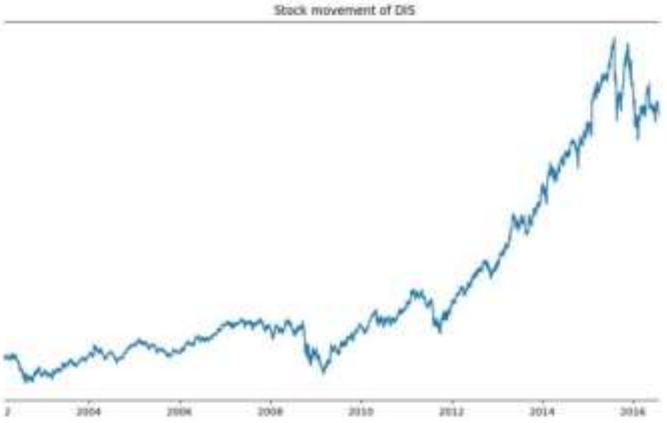

Figure 3: Stock movement representation of DIS

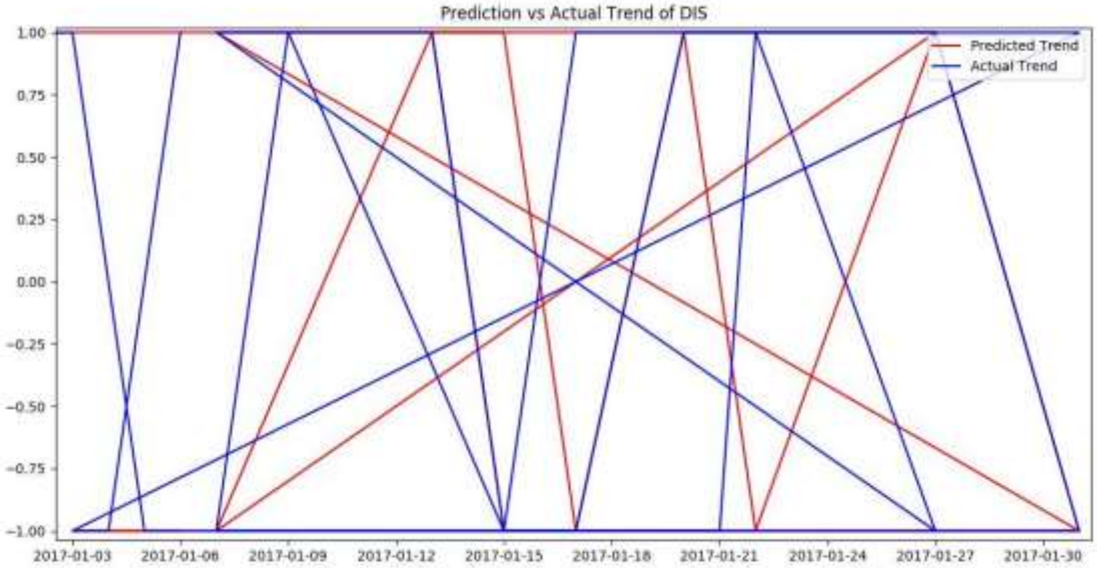

Figure 4: The predicted and actual trend in stock prices for DIS.

As depicted in the figures above, the prediction and the actual trend overlap in a lot of areas. The average accuracy of 70.236% is visible in the graphs. The fact that the actual and expected trends coincide indicates that the errors are quite tiny, meaning that the actual and predicted values are near to each other. This results in an excellent level of accuracy when utilizing the KNN to forecast the value of stocks.

V. CONCLUSION

The paper presents the results of a prediction technique applied to seven NASDAQ Stock Market-listed businesses. As a result, a solid model was created for the intended use with Yahoo Finance as the source of the data extraction. To conduct these tests on our training data sets, we used the prediction with, KNN using k=5. The findings were logical and understandable as KNN demonstrated high accuracy and stability. Furthermore, the outcomes of the predictions were quite close to the real values, based on data on actual stock prices. These logical outcomes for predictions using data mining techniques in practical situations suggest that decision-makers at different levels benefit from the usage of data mining techniques when employing KNN for statistical analysis of data. Therefore, we believe that KNN is a practical prediction model to use for stock forecasts. Moreover, this makes investments in the NASDAQ market less appealing, which would ultimately reduce the return on investment. The study also demonstrates how modern data mining methods provide the financial industry with insightful analysis of the stock market's predictions.

REFERENCES

1. *Mujeeb, S., Javaid, N., Akbar, M., Khalid, R., Nazeer, O., Khan, M. (2019). Big Data Analytics for Price and Load Forecasting in Smart Grids. In: Barolli, L., Leu, FY., Enokido, T., Chen, HC. (eds) Advances on Broadband and Wireless Computing, Communication and Applications. BWCCA 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 25. Springer, Cham. https://doi.org/10.1007/978-3-030-02613-4_7*
2. *S. Sharon Priya and Anooja Ali, "Localization of WSN using IDV and Trilateration Algorithm",* Asian Journal of Engineering and Technology Innovation, *vol. 4, no. 7, 2016.*
3. *S. Sneha, Applications of ANNs in Stock Market Prediction: A Survey, International Journal of Computer Science Engineering Technology, vol. 2(3), 2011.*

4. *E. F. Fama, Random Walks In Stock Market Prices, Financial Analysts Journal, vol. 21(5), pp. 55-59, 1965.*
5. *Arjun Singh Saud, Subarna Shakya, Analysis of look back period for stock price prediction with RNN variants: A case study on banking sector of NEPSE, Procedia Computer Science, Volume 167, 2020, Pages 788-798, ISSN 1877-0509.*
6. *Ali, Anooja, et al. "Conversational AI agent for educational institute using NLU and LSTM algorithm." AIP Conference Proceedings. Vol. 2742. No. 1. AIP Publishing, 2024. https://doi.org/10.1063/5.0183803*
7. *Htun, Htet Htet, Michael Biehl, and Nicolai Petkov. "Survey of feature selection and extraction techniques for stock market prediction." Financial Innovation 9.1 (2023): 26.*
8. *Box GEP, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. New York: Wiley; 2015*
9. *Ali, Anooja, et al. "Prognosis of chronic kidney disease using ML optimization techniques." 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS). IEEE, 2023.*
   DOI: 10.1109/ICAECIS58353.2023.10170077
10. *Bhuriya, D., Kaushal, G., Sharma, A., & Singh, U. (2017). Stock market predication using a linear regression. Paper presented at the 2017 international conference of electronics, communication and aerospace technology (ICECA)*
11. *Jere, Shreekant, et al. "Recruitment graph model for hiring unique competencies using social media mining." Proceedings of the 9th International Conference on Machine Learning and Computing. 2017.*
12. *Ali, Anooja, et al. "DPEBic: detecting essential proteins in gene expressions using encoding and biclustering algorithm." Journal of Ambient Intelligence and Humanized Computing (2021): 1-8.*
13. *Valle-Cruz, David, et al. "Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the h1n1 and the covid-19 periods." Cognitive computation 14.1 (2022): 372-387.*
14. *Mahapatro, Rakesh Kumar, Anooja Ali, and Nithin Ramakrishnan. "Blockchain segmentation: a storage optimization technique for large data." 2023 8th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2023.*
   DOI: 10.1109/ICCES57224.2023.10192631
15. *[b]. Ali, Anooja, et al. "A review of aligners for protein protein interaction networks." 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, 2017.*
   DOI: 10.1109/RTEICT.2017.8256879
16. *Wong, S. (2020, December). Stock Price Prediction Model Based on the Short-term Trending of KNN Method. In 2020 7th International Conference on Information Science and Control Engineering (ICISCE) (pp. 1355-1360). IEEE.*
17. *Ramachandra, H. V., et al. "Ensemble machine learning techniques for pancreatic cancer detection. "2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC). IEEE, 2023.*
   DOI: https://ieeexplore.ieee.org/abstract/document/10200380
18. *A Ali, VR Hulipalled, SS Patil and RA Kappaparambil, "DPCCG-EJA: detection of key pathways and cervical cancer related genes using enhanced Johnson's algorithm", Int J Adv Sci Technol, vol. 28, no. 1, pp. 124-138, 2019.*
19. *Huang, Fangjun. "Stock Price Prediction Based on Trend Characterization." Proceedings of the 2nd International Conference on Mathematical Statistics and Economic Analysis, MSEA 2023, May 26–28, 2023, Nanjing, China. 2023.*
20. *Qin, Wenxi. "Predictive Analysis of AAPL Stock Trend by Random Forest and K-NN Classifier." Highlights in Business, Economics and Management 24 (2024): 1418-1422.*