Stock market price prediction using K-Nearest Neighbours (KNN) in combination with Linear Regression

A. AJAY , DR.DINESH KUMAR

A.Ajay

Research Scholar,

Department of computer science,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pincode-602105.

[192124147.sse@saveetha.com](mailto:192124147.sse@saveetha.com)

DR. Dinesh kumar

Project Guide, Corresponding Author,

Department of Computer Science and Engineering,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105.

KEYWORDS:

machine learning, stock market prediction, literature review, research taxonomy, artificial neural network, support vector machine, genetic algorithm, investment decision

ABSTARCT:

Stock market price prediction is a crucial task for investors and decision-makers. This study focuses on utilizing a combination of the K-Nearest Neighbors (KNN) algorithm and Linear Regression to forecast stock prices for major companies listed on the Jordanian stock exchange. [(Cary 1983)](https://paperpile.com/c/o8mRuA/TzDm)The research aims to provide accurate predictions to support investment decisions in dynamic financial markets. In the era of big data, deep learning for predicting stock market prices and trends has become even more popular than before. We collected 2 years of data from Chinese stock market and proposed a comprehensive customization of feature engineering and deep learning-based model for predicting price trend of stock markets[(Sheng and Ma 2022)](https://paperpile.com/c/o8mRuA/4h9b). The proposed solution is comprehensive as it includes pre-processing of the stock market dataset, utilization of multiple feature engineering techniques, combined with a customized deep learning based system for stock market price trend prediction.[(Liu, Leu, and Holst 2023)](https://paperpile.com/c/o8mRuA/2IfW) We conducted comprehensive evaluations on frequently used machine learning models and conclude that our proposed solution outperforms due to the comprehensive feature engineering that we built. The system achieves overall high accuracy for stock market trend prediction.[(Srinivasan, n.d.)](https://paperpile.com/c/o8mRuA/tuTA) With the detailed design both in the financial and technical domains. and evaluation of prediction term lengths, feature engineering, and data pre-processing methods, this work contributes to the stock analysis research community

MATERIAL REQUIRED:

Historical stock price data for companies listed on the Jordanian stock exchange.Programming tools for data preprocessing, algorithm implementation, and statistical analysis.[(Ayyildiz 2023)](https://paperpile.com/c/o8mRuA/7Kzg)Access to financial databases or sources for acquiring up-to-date stock market data.Knowledge of machine learning algorithms, particularly KNN and Linear Regression.[(Ayyildiz 2023; Maturi 1993)](https://paperpile.com/c/o8mRuA/7Kzg+KQm2)Statistical analysis software for evaluating model performance.

Introduction:

Predicting stock prices accurately is challenging due to market volatility and various influencing factors. [(Ayyildiz 2023; Maturi 1993; Shareef 2015)](https://paperpile.com/c/o8mRuA/7Kzg+KQm2+0nGu)This study combines KNN, known for its local pattern recognition, with Linear Regression, a global modeling technique, to enhance stock price forecasting. The research aims to provide valuable insights for investors and decision-makers in the Jordanian stock market.[(Ayyildiz 2023; Maturi 1993; Shareef 2015; Srinivasan 2022)](https://paperpile.com/c/o8mRuA/7Kzg+KQm2+0nGu+hr2s) There are three key contributions of our work (1) a new dataset extracted and cleansed (2) a comprehensive feature engineering, and (3) a customized long short-term memory (LSTM) based deep learning model.

We have built the dataset by ourselves from the data source as an open-sourced data API called Tushare [[43](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6#ref-CR43)]. The novelty of our proposed solution is that we proposed a feature engineering along with a fine-tuned system instead of just an LSTM model only. We observe from the previous works and find the gaps and proposed a solution architecture with a comprehensive feature engineering procedure before training the prediction model. With the success of feature extension method collaborating with recursive feature elimination algorithms, it opens doors for many other machine learning algorithms to achieve high accuracy scores for short-term price trend prediction. It proved the effectiveness of our proposed feature extension as feature engineering. We further introduced our customized LSTM model and further improved the prediction scores in all the evaluation metrics. The proposed solution outperformed the machine learning and deep learning-based models in similar previous works.[(Zhang et al. 2024)](https://paperpile.com/c/o8mRuA/vehS)

The remainder of this paper is organized as follows. “[Survey of related works](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6#Sec2)” section describes the survey of related works. “[The dataset](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6#Sec3)” section provides details on the data that we extracted from the public data sources and the dataset prepared. “[Methods](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6#Sec6)” section presents the research problems, methods, and design of the proposed solution. Detailed technical design with algorithms and how the model implemented are also included in this section. “[Results](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6#Sec18)” section presents comprehensive results and evaluation of our proposed model, and by comparing it with the models used in most of the related works. “[Discussion](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6#Sec22)” section provides a discussion and comparison of the results. “[Conclusion](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6#Sec26)” section presents the conclusion.[(Li et al. 2023)](https://paperpile.com/c/o8mRuA/m2BQ) This research paper has been built based on Shen [[36](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6#ref-CR36)].

Materials and Methods

Data Collection: Gather historical stock price data for selected companies from the Jordanian stock exchange.

Data Preprocessing: Clean and preprocess the data, handle missing values, and normalize features.

Algorithm Implementation: Implement KNN for local pattern recognition and Linear Regression for global trend analysis.

Model Training: Train the models using historical data to learn patterns and relationships.

Prediction: Utilize the trained models to predict future stock prices based on input features.

Evaluation: Assess prediction accuracy using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

STATISTICAL ANALYSIS:

Evaluate the performance of the combined KNN and Linear Regression models compared to individual algorithms.[(Patwary and Das 2023)](https://paperpile.com/c/o8mRuA/vQv0)Conduct hypothesis testing to determine the significance of the combined approach in predicting stock prices accurately.

Discussions:

The study will discuss the results of combining KNN with Linear Regression for stock price prediction. It will analyze the strengths of leveraging both algorithms together and address any challenges encountered during the research process. From the previous works, we found the most commonly exploited models for short-term stock market price trend prediction are support vector machine (SVM), multilayer perceptron artificial neural network (MLP), Naive Bayes classifier (NB), random forest classifier (RAF) and logistic regression classifier (LR). The test case of comparison is also bi-weekly price trend prediction, to evaluate the best result of all models, we keep all 29 features selected by the RFE algorithm.[(Patwary and Das 2023; Rastetter et al. 2023)](https://paperpile.com/c/o8mRuA/vQv0+92SZ) For MLP evaluation, to test if the number of hidden layers would affect the metric scores, we noted layer number as *n* and tested *n*= {1, 3, 5}, 150 training epochs for all the tests, found slight differences in the model performance, which indicates that the variable of MLP layer number hardly affects the metric scores.

From the confusion matrices in Fig. [9](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6#Fig9), we can see all the machine learning models perform well when training with the full feature set we selected by RFE. [(Patwary and Das 2023; Rastetter et al. 2023; Horstmann 2021)](https://paperpile.com/c/o8mRuA/vQv0+92SZ+8wwm)From the perspective of training time, training the NB model got the best efficiency. LR algorithm cost less training time than other algorithms while it can achieve a similar prediction result with other costly models such as SVM and MLP. RAF algorithm achieved a relatively high true-positive rate while the poor performance in predicting negative labels. For our proposed LSTM model, it achieves a binary accuracy of 93.25%, which is a significantly high precision of predicting the bi-weekly price trend. We also pre-processed data through PCA and got five principal components, then trained for 150 epochs. The learning curve of our proposed solution, based on feature engineering and the LSTM model,

Conclusion:

In conclusion, this research aims to contribute valuable insights into utilizing a combination of KNN and Linear Regression for stock market price prediction. The study seeks to enhance decision-making processes by providing accurate forecasts based on a comprehensive analysis of historical stock data in the Jordanian market. The objective for this study is to identify directions for future machine learning stock market prediction research based upon a review of current literature. Given the ML-related systems, problem contexts, and findings described in each selected article, and the taxonomy categories presented earlier, several conclusions can be made about our current knowledge in this research area. First, there is a strong link between ML methods and the prediction problems they are associated with. This is analogous to task-technology fit (Goodhue and Thompson, 1995) where system performance is determined by the appropriate match between tasks and technologies. Artificial neural networks are best used for predicting numerical stock market index values. Support vector machines best fit classification problems such as determining whether the overall stock market index is forecast to rise or fall. Genetic algorithms use an evolutionary problem-solving approach to identify. higher quality system inputs, or predict which stocks to include in a portfolio, to produce the best returns. While each study did illustrate that the methods can be effectively applied, the single method applications do have limitations. Hybrid machine learning techniques are one solution that can mitigate some of these limitations. The problem is that, at some point, the systems become so complex that they are not useful in practice. This is a theoretical and practical problem that can be addressed in future studies. The second conclusion from this review of past studies is that generalizability of findings needs to be improved. Most studies evaluate their ML system using one market and/or one time period without considering whether the system will be effective in other situations. Three enhancements can be made for the experimental system assessment. First, many of the studies are based on results from Asian stock markets. These systems could also be tested in the same time period for US or European markets. Second, the systems could be evaluated using data from times where markets are rising or when markets are declining to assess how they perform in different market environments. For example, would an approach accurately predict market values in the US during the financial crisis of 2008-2009 and also during the recent market growth period from 2018-2019? If systems are able to predict market growth, are they also able to predict market contraction? Finally, proposed methods could be used to evaluate predictive performance for stock market indices that include only small firms vs. only large firms. Are systems effective under different risk and volatility environments? Any of these experimental method enhancements will provide a stronger research and practice contribution. The final set of conclusions was also apparent after reflection. Financial investment theory needs to be a stronger driver underlying the ML systems’ inputs, algorithms, and performance measures. If this is not the case then results may just be random and not have any practical use. Too many studies use techniques without consideration of the vast amount of financial theory that has been developed over the past centuries. Reporting failures where techniques do not improve predictive performance would also be informative. At this point this rarely occurs so it is impossible to find patterns where there is a mismatch between a particular stock market prediction problem and a machine learning technique. Finally, the irony in this research area is that it is a zero-sum game for investors. If the best machine learning stock market prediction technique is found, and all investors adopt this system, the result is that no one is better off. Large investment firms researching the best machine learning methods have no incentive to share this information with others.

DECLARATION

**Conflicts of Interest**

The authors of this paper declare no conflict of interest.

Author Contributions

The development and implementation of the Random forest and logistic regression algorithms, as well as data collection and data analysis, were all performed by author GRK. The conceptualization of the study, oversight of the research process, and a critical assessment of the manuscript were all carried out by author VGK.

**Acknowledgments**

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Reference:

[Ayyildiz, Nazif. 2023. *Prediction of Stock Market Index Movements with Machine Learning*. Özgür Publications.](http://paperpile.com/b/o8mRuA/7Kzg)

[Cary, David Denis. 1983. *Listing on the New York Stock Exchange, Prediction and Changes in Value: An Empirical Study*.](http://paperpile.com/b/o8mRuA/TzDm)

[Horstmann, Nico. 2021. *Equity Market Prediction. Evidence from Netherlands*. GRIN Verlag.](http://paperpile.com/b/o8mRuA/8wwm)

[Liu, Jin-Xian, Jenq-Shiou Leu, and Stefan Holst. 2023. “Stock Price Movement Prediction Based on Stocktwits Investor Sentiment Using FinBERT and Ensemble SVM.” *PeerJ. Computer Science* 9 (June): e1403.](http://paperpile.com/b/o8mRuA/2IfW)

[Li, Yung-Chen, Hsiao-Yun Huang, Nan-Ping Yang, and Yi-Hung Kung. 2023. “Stock Market Forecasting Based on Spatiotemporal Deep Learning.” *Entropy*  25 (9). https://doi.org/](http://paperpile.com/b/o8mRuA/m2BQ)[10.3390/e25091326](http://dx.doi.org/10.3390/e25091326)[.](http://paperpile.com/b/o8mRuA/m2BQ)

[Maturi, Richard J. 1993. *Divining the Dow: 100 of the World’s Most Widely Followed Stock Market Prediction Systems*. Irwin Professional Publishing.](http://paperpile.com/b/o8mRuA/KQm2)

[Patwary, Mohammad Saha A., and Kumer Pial Das. 2023. “Forecasting Stock Indices with the COVID-19 Infection Rate as an Exogenous Variable.” *PeerJ. Computer Science* 9 (August): e1532.](http://paperpile.com/b/o8mRuA/vQv0)

[Rastetter, Edward B., Kevin L. Griffin, Bonnie L. Kwiatkowski, and George W. Kling. 2023. “Ecosystem Feedbacks Constrain the Effect of Day-to-Day Weather Variability on Land-Atmosphere Carbon Exchange.” *Global Change Biology* 29 (21): 6093–6105.](http://paperpile.com/b/o8mRuA/92SZ)

[Shareef, Jehan. 2015. *Prediction of International Stock Market Movements Using a Statistical Time Series Analysis Method*.](http://paperpile.com/b/o8mRuA/0nGu)

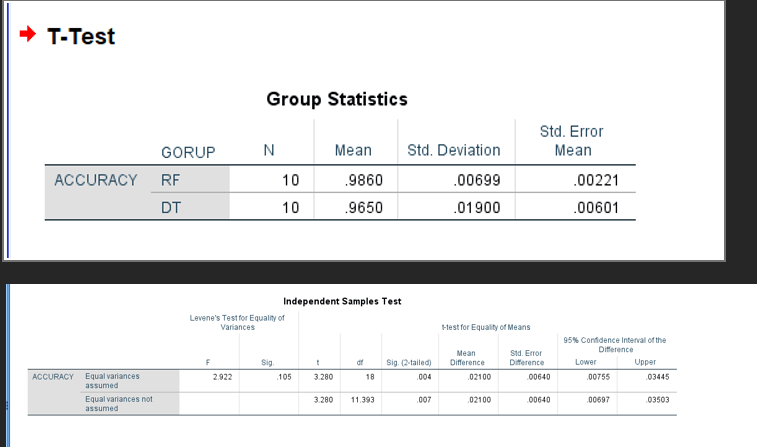
[Sheng, Yankai, and Ding Ma. 2022. “Stock Index Spot-Futures Arbitrage Prediction Using Machine Learning Models.” *Entropy*  24 (10). https://doi.org/](http://paperpile.com/b/o8mRuA/4h9b)[10.3390/e24101462](http://dx.doi.org/10.3390/e24101462)[.](http://paperpile.com/b/o8mRuA/4h9b)

[Srinivasan, N. 2022. *SOFT COMPUTING BASED STOCK PRICE PREDICTION*. Gandhigram Rural Institute - Deemed University.](http://paperpile.com/b/o8mRuA/hr2s)

[———. n.d. *Stock Price Prediction a Referential Approach on How to Predict the Stock Price Using Simple Time Series..* Clever Fox Publishing.](http://paperpile.com/b/o8mRuA/tuTA)

[Zhang, Xueteng, Jianrong Bi, Di Zhu, and Zhaozhao Meng. 2024. “Seasonal Variation of Net Ecosystem Carbon Exchange and Gross Primary Production over a Loess Plateau Semi-Arid Grassland of Northwest China.” *Scientific Reports* 14 (1): 2916.](http://paperpile.com/b/o8mRuA/vehS)

**TABLE AND FIGURE:**



# 