Comparison with Deep Learning Methods For Predicting Stock Prices

Colton Nutter, Nayeong Kong, Seonguk Kim*
Division of Natural Science, Applied Science, and Mathematics
Defiance College
Defiance, OH 43512, USA
**Corresponding author's email: skim [AT] defiance.edu*

Abstract— **Recently, machine learning has been an essential tool for analysis in diverse fields, including science, sports management, and economics. In particular, the stock market comprises a complex network of buyers and sellers engaged in stock trading. So, predicting stock prices has been developed using machine-learning techniques to significantly enhance such forecasts' accuracy. Recent advancements have improved the performance of several algorithms, such as Linear Regression, Support Vector Machines (SVM), and K-nearest neighbors (KNN) to predict stock prices. Stock price datasets typically contain information such as opening and closing prices, high and low values, dates, trading volume, and adjusted closing prices provided by Yahoo Finance. Based on the data, this research evaluates the prediction accuracy of each machine-learning method and presents the results through data visualizations, including box plots and tables. The compiled results will assist in identifying the most effective model for stock price prediction.**

Keywords: Stock price, Linear Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN)

I. INTRODUCTION

Machine learning has become an essential tool for research in various areas, such as biology, chemistry, and economics. Using the methods in machine learning enables us to understand more deeply from data and improve our ability to perform tasks based on the analysis [1], [2], [3], [4].

Machine learning has contributed to predicting stock prices in the stock market, a complicated correlation between buyers and sellers trading stocks. Adopting various machine learning techniques has significantly enhanced the accuracy of these predictions. For the work, this paper investigates specific machine learning algorithms in forecasting tasks: Linear Regression, Support Vector Machine (SVM), and K-nearest neighbors (KNN), [5].

Stock price datasets typically contain information on opening and closing prices, highest and lowest values, dates, transaction volumes, and adjusted closing prices. Our study aims to use linear regression, SVM, and KNN to predict closing stock prices with the data extracted from Yahoo Finance, which is known for its user-friendly platform and comprehensive datasets [6], [7], [8].

Our research focuses on achieving accurate company stock price predictions using these machine-learning techniques. We also aim to create data visualizations like box plots and tables to assist in evaluating the effectiveness of each model for stock price prediction. The compiled data and visualizations will help identify the most suitable model for stock price forecasting.

The paper consists as follows: Section 2 introduces and details the data and methodology used in our analysis, accompanied by a diagram that illustrates the process. In Section 3, we present the results of our work, utilizing tables and figures for clarity. Specifically, we employ box plots to compare the distribution of outcomes across different methods.

II. DESCRIPTION OF METHODOLOGY

A. Dataset and Software

In our study, we concentrated on evaluating the performance of stock price predictions for companies within the NASDAQ 100 index, covering a period from January 2020 to January 2021. We utilized an array of machine learning techniques to gauge how precisely each model could forecast the closing stock prices. The techniques we explored include Linear Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and an Ensemble method that integrates several predictive approaches. To obtain the necessary stock data, we employed the yfinance Python package, a powerful tool that allows for easy access to historical market data directly from Yahoo Finance and visualize the historical graph of the stock prices (See Figure 1). Our analysis was conducted using Python, specifically within the Anaconda Navigator environment, which provided a comprehensive platform for running our machine learning models and executing the analysis program. The key metrics we focused on to assess the predictive accuracy of our models were the accuracy percentages, Mean Squared Error (MSE), and R-squared (R2) scores. These metrics are crucial for understanding each model's effectiveness in predicting stock prices. The accuracy percentage gives a straightforward measure of how often the model's predictions were correct. In contrast, the MSE offers insight into the average squared difference between the predicted and actual stock prices, helping us identify models that minimize prediction errors. Lastly, the R2 score indicates how well the variations in stock prices are accounted for by the model, with higher values suggesting a better fit to the historical data. By leveraging these metrics and the robust capabilities of Python and yfinance, we

aimed to discover which machine learning techniques are most adept at forecasting stock price movements within the environment of the NASDAQ 100 index.

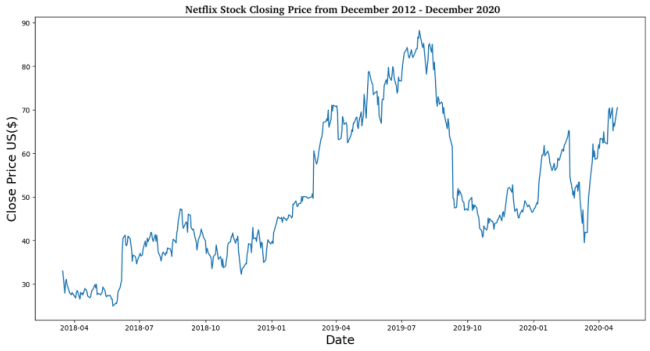

Figure 1. Netflix Stock Closing Price from Dec 2012 – Dec 2020

B. *Linear Regression*

Linear Regression is a statistical technique utilized to predict future stock prices by modeling the relationship between a dependent variable—stock's closing prices—and one or more independent variables, such as historical prices, trading volumes, or earnings. This method draws a linear connection between the variables and aims to find the optimal linear equation that minimizes the sum of the squared differences between the observed and predicted values, capturing the best-fit line through the data. We proceed with data preprocessing, which includes cleaning and normalizing the data to optimize it for the modeling process. This preparation is vital for the accuracy and reliability of the predictions made by the Linear Regression model.

C. *Support Vector Method (SVM)*

Support Vector Machine (SVM) is a machine learning technique employed to predict future stock prices by classifying data points into distinct categories based on historical prices, trading volumes, and other relevant variables. At its core, SVM searches for the optimal boundary that differentiates between categories of data, handling the complexities and non-linear relationships characteristic of financial markets. The initial step in leveraging SVM for our analysis involved selecting features that impact stock prices, followed by data preprocessing. This preprocessing is crucial for the model's ability to effectively classify the patterns present in stock market data. Our evaluation of SVM's performance centered on classification accuracy, utilizing metrics appropriate for assessing the precision of the model's predictions in categorizing stock price trends. By implementing SVM, we understand the factors driving stock prices and improve the precision of our forecasts.

D. *K-Nearest Neighbor (KNN)*

K-Nearest Neighbors (KNN) is a machine learning algorithm utilized to predict future stock prices by analyzing historical data patterns and features such as prices and trading volumes. KNN operates on the principle of proximity, identifying the 'k' nearest data points to a query point in the feature space and predicting outcomes based on the majority class or the average of their values. This method is particularly effective in the context of financial.

III. RESULTS

In analyzing stock data, box plots, also known as box-and-whisker plots, play a pivotal role across various predictive modeling techniques, including Linear Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) to evaluate model performance metrics like Accuracy, Mean Squared Error (MSE), and R2 score. These plots offer a graphical representation to scrutinize the distribution of residuals in Linear Regression—differences between actual stock returns and the model's predictions— highlighting the model's adherence to assumptions. Similarly, within the SVM framework, box plots are leveraged to explain the distribution of key features and the proximity of samples to the decision boundary for both classification and regression tasks, aiding in the identification of outliers and model performance through a visualization of residuals and margin distributions. For KNN, box plots aid the selection of the "k" parameter but also refine the quality of input features by pinpointing outliers and extreme values. By incorporating metrics such as Accuracy, MSE, and the R2 score into these box plots, we gain insight into the model's consistency, reliability, and overall prediction accuracy.

TABLE I. SPECIFIC DATA OF COLUMN/ROW

	KNN	Linear Regression	SVM
Accuracy	98.916	99.801	99.663
MSE	0.764	1.016	0.044
R2 Score	0.998	1.0	0.894

A. *Accuracy findings for Linear Regression, KNN, and Support Vector Machine*

Accuracy is a metric used to evaluate observational error. Accuracy refers to the proximity of a set of measurements (observations or readings) to their actual value.

Box plots visually summarize the accuracy of Linear Regression, KNN, and SVM models in forecasting stock prices. The accuracy is determined by how closely the model's predictions match actual values. In these plots, the median accuracy is marked by a line within each box, offering a comparison of each model's performance. The spread of the data, or the interquartile range (IQR), encapsulated by the box, illustrates the consistency of each model's accuracy.

The box plots of the accuracy (Figures 2, 3, 4) provide a statistical evaluation of how closely the predictions made by Linear Regression, KNN, and SVM models align with the actual stock prices. In analyzing these plots, we can draw the following observations:

1. Generally, each model achieves a high score, surpassing 99%.

2. Evaluating based on the Interquartile Range (IQR), Linear Regression exhibits the highest performance, followed by SVM, and KNN performs the least effectively. unavoidable.

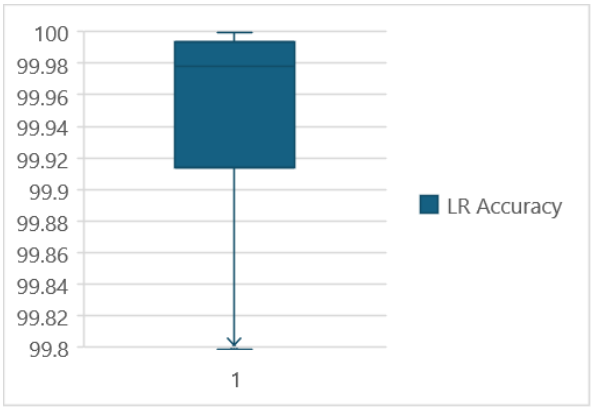

Figure 2. Box plots representing our accuracy findings for Linear Regression

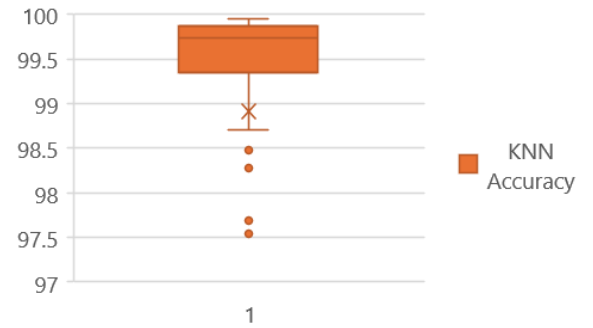

Figure 3. Box plots representing our accuracy findings for KNN

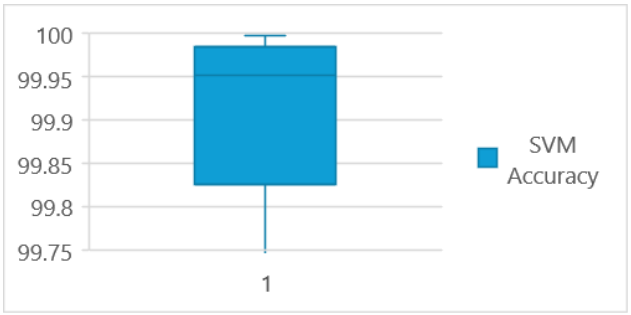

Figure 4. Box plots representing our accuracy findings for Support Vector Machine

B. Mean Squared Errors findings for Linear Regression, KNN, and Support Vector Machine.

The Mean Square Error (MSE) represents the average of the squared errors. A higher value indicates a more significant error. In this context, "error" refers to the disparity between the observed and predicted values. In Mean Squared Error (MSE) findings, box plots convey the spread and central tendency of errors across our predictive models. MSE quantifies the average squared difference between the estimated values by the model and the actual values, serving as a critical metric for prediction accuracy. A model's effectiveness is inversely related to its median MSE—visible as the line within each box; lower medians indicate higher accuracy. Outliers point to models' performances that are anomalously high or low in error, essential for evaluating model robustness.

In examining these box plots, the following observations can be made:

1. In the Linear Regression Model, the difference between observed and actual values is nearly negligible, indicating minimal error in the model.
2. The KNN model exhibits a comparatively significant error and a more expansive Interquartile Range (IQR), suggesting less consistency.
3. Although the error remains small in the SVM model, it is relatively more significant than the Linear Regression model.

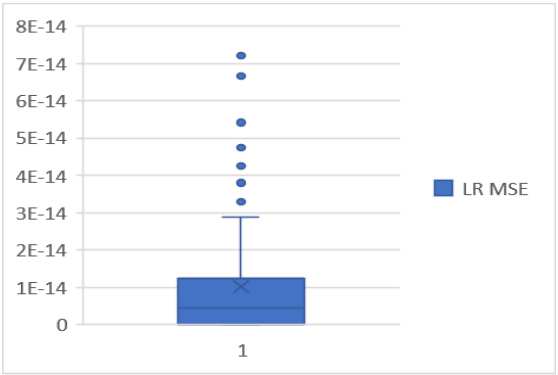

Figure 5. Box plots representing our Mean Squared Errors findings for Linear Regression

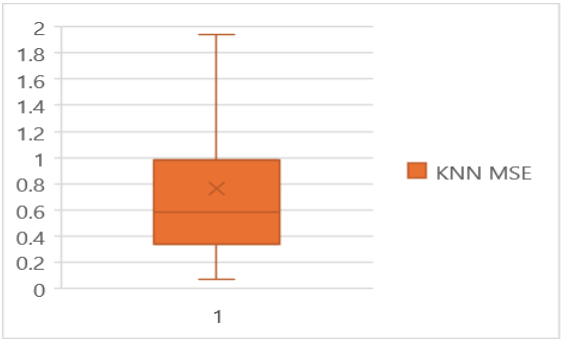

Figure 6. Box plots representing our Mean Squared Errors findings for KNN

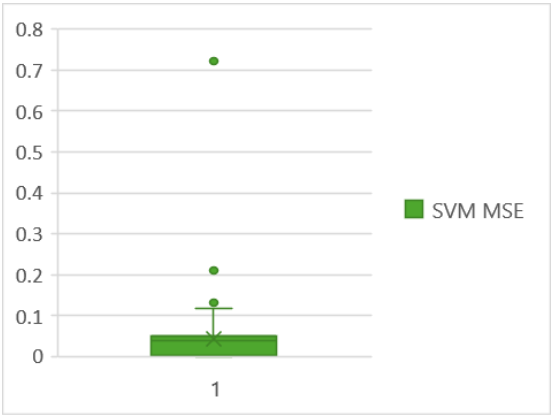

Figure 7. Box plots representing our Mean Squared Errors findings for Support Vector Machine

C. *R2 Score findings for Linear Regression, KNN, and Support Vector Machine*

The R2 score ranges from 0 to 100% (=1), and it is closely linked to the Mean Squared Error (MSE) but is distinct. R2 score of 100% indicates a perfect correlation between variables, implying no variance. Conversely, a low R2 value suggests a weak correlation, indicative of a regression model that may lack validity, though exceptions exist.

The box plots of the R2 scores (Figures 8, 9, 10) offer a statistical assessment of how accurately the predictions made by Linear Regression, KNN, and SVM models reflect the actual stock prices. Upon comparison of the figures, we have the following observations:

1. Linear Regression demonstrates the highest score and the most concentrated distribution.

2. KNN exhibits a commendable score, albeit with a slightly broader distribution (the interquartile range (IQR) is wide).

3. SVM also presents a respectable score, yet the median value (marked as '×') noticeably diverges from the mean. the average.

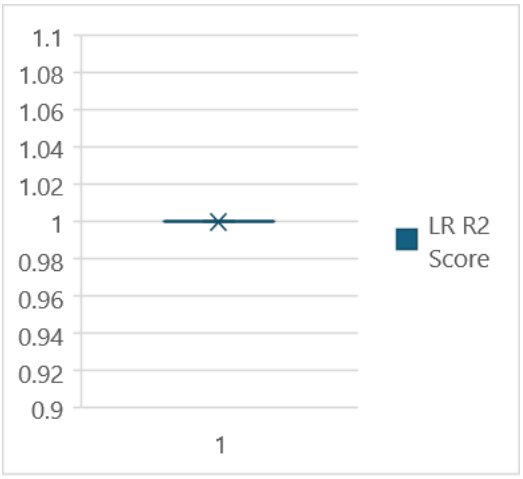

Figure 8. Box plots representing our R2 Score findings for Linear Regression

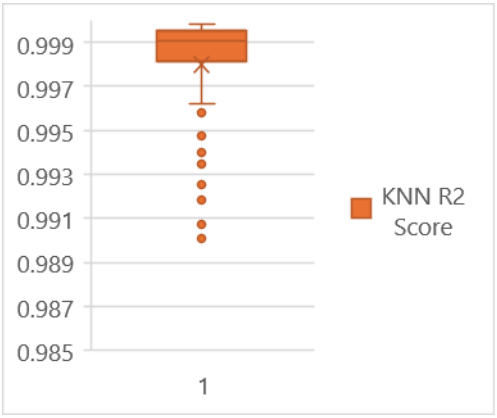

Figure 9. Box plots representing our R2 Score findings for KNN

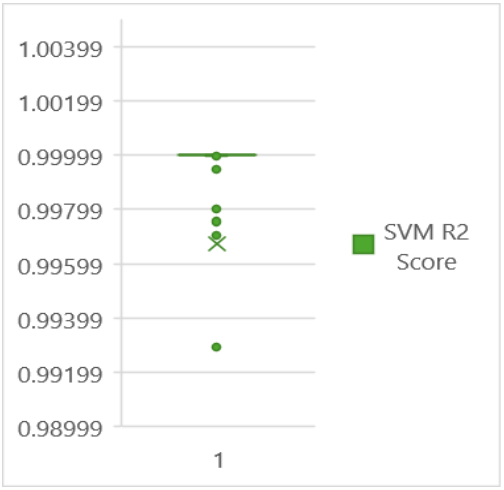

Figure 10. Box plots representing our R2 Score findings for Support Vector Machine

IV. Conclusion

Our analysis of stock price predictions within the NASDAQ 100 index, employing mathematical formulas to our machine learning model such as Linear Regression, KNN, and SVM, has illuminated the varying degrees of efficacy these models possess in navigating the dynamics of stock markets. The predictive accuracy of these models, assessed through metrics such as Accuracy, Mean Squared Error (MSE), and R-squared (R2) scores, suggests that a multi-faceted approach to stock market analysis can provide a more nuanced understanding of price movements. This study not only contributes to the evolving discourse on the application of machine learning in financial markets but also highlights the importance of adopting versatile analytical strategies to accommodate the unpredictability of stock prices. The implications of our research extend beyond academic inquiry, offering valuable insights for investors and financial analysts seeking to harness the predictive power of machine learning in making informed economic and investment decisions. As we look towards the future, the horizon for machine learning in

finance broadens with advancement in technology and artificial intelligence, promising ever more sophisticated tools for market analysis.

ACKNOWLEDGMENT (HEADING 5)

It was supported by the Summer Undergraduate Research Project from Defiance College.

REFERENCES

[1] Hai Phan, Seonguk Kim, Numerical Approaches of Pricing European Options in the Cox-Ross-Rubinstein Models, Universal Journal of Applied Mathematics, Vol.10, No.3, pp. 43-48, 2022.

[2] Destiny Rankins, Dewayne A. Dixon, Yeona Kang, Seonguk Kim " Analysis of the Convolutional Neural Network Model in Detecting Brain Tumor," International Journal of Computer and Information Technology, Volume 11 – Issue 4, August 2022.

[3] Borrellas P, Unceta I. The Challenges of Machine Learning and Their Economic Implications, Entropy (Basel), Vol 25, No.3, pp.275, 2021.

[4] Hock Chuan Yeo, Kumar Selvarajoo, Machine learning alternative to systems biology should not solely depend on data, Briefings in Bioinformatics, Vol 23, No.6, pp.1-6, 2022.

[5] Weiwei Jiang, Applications of deep learning in stock market prediction: Recent progress, Expert Systems with Applications, Vol.184, 2021.

[6] B. Panwar, G. Dhuriya, P. Johri, S. Singh Yadav and N. Gaur, Stock Market Prediction Using Linear Regression and SVM, 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, pp. 629-631, 2021.

[7] Li, Y., Pan, Y. A novel ensemble deep learning model for stock prediction based on stock prices and news. Int J Data Sci Anal Vol.13, pp.139–149, 2022.

[8] Hakob GRIGORYAN, A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA). Database Systems Journal vol. VII, no. 1/2016.