Journal of Research High School

Vol. 2024 (12) 537 – 544
ISSN 2688-3651 [online]

Predicting Stock Prices with Different AI Models (Linear & MLP regression)

Alexander Baruah[1] *

[1] Washington High School, Fremont, CA, USA
*Corresponding Author: Alexanderbaruah@gmail.com

Advisor: Odysseas Drosis, od84@cornell.edu

Received November 12, 2023; Revised July 14, 2024; Accepted August 13, 2024

Abstract

With the new trend of AIs popping up everywhere, the stock trading game has changed. Lots of stock traders have pivoted and started using AI to make educated decisions to maximize profit on stocks based on the data given to the AI model. This research explored the effectiveness of AI models called Linear regression and Multi-Layer Perceptron (MLP) regression for predicting stock prices. Utilizing historical stock price data from Cisco (CSCO), Tesla (TSLA), Apple (AAPL), Starbucks (SBUX), and Johnson & Johnson (JNJ), the study evaluated the models' performances based on Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). The results offer valuable insights for investors, suggesting that the choice between these models should be guided by specific trading strategies and practical considerations rather than solely performance metrics shown here.

Keywords: Stock price forecasting, Linear regression, Multi-layer perceptron regression

1. Introduction

Predicting stock prices accurately is really important because it affects how people invest their money. Even though there have been many advancements, it's still hard to predict stock prices due to the unpredictable nature of financial markets. This study tackled this problem by using advanced regression models, specifically Linear regression and Multi-Layer Perceptron (MLP) regression, to improve accuracy.

Accurate stock price predictions are crucial for investors and financial analysts. They rely on these predictions to make decisions that can lead to big financial gains or losses. That's why it's so important to develop reliable models for predicting stock prices.

Previous research (SH, 2024) suggests that MLP Regression is a valuable option for stock prediction AI with a score of (test set performance) 3.96 for the Mean Squared Error (MSE) metric and R-squared (R^2) metric 0.9994. For reference a larger score is worse for MSE, but better for R^2. Another paper (Zhou, 2024) discusses Linear regression on stock prediction and has proven that Linear regression is an okay AI model to be used for stock prediction with a score of 51.518 for MSE and 0.016 for R^2. This gives the misleading perception that MLP regression is better than linear regression. The reason why it is not a good indication of which model regression model is actually better is because of the fact that each model had predicted with different stocks, different feature engineering, different lengths of time, and so on. So this research attempted to fairly compare the 2 models with controlled variables to give each model an equal opportunity to assess which model is better and why.

By using historical stock price data from various companies, this study evaluated how well Linear and MLP regression models perform using Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) metrics. The results should provide useful insights into how these models can be used for different prediction timeframes, helping investors make better decisions. This research contributed to the development of AI-driven financial forecasting and set the stage for future studies that might use more data sources and different models.

Vol. 2024 (12) 537 – 544
ISSN 2688-3651 [online]

2. Data Collection and Preprocessing

This project utilized historical stock price data for any given stock over five years. The data preparation involves removing unnecessary columns like 'High,' 'Close,' 'Volume,' 'Dividends,' and 'Stock Splits' provided by Yahoo Finance because the only required input is the ‘Open’ price.

After obtaining the open price, it is necessary to preprocess the information so that the AI can use the given 'Open' prices to predict the outcome for the next day. Utilizing all the stock prices for a specific number of days leading up to the next prediction simplifies the data for the AI, allowing it to identify patterns in input values to predict the output value. For instance, one of the periods considered is three days. In this case, all values with a three days period before the next stock price are collected over five years.

Table 1. Imaginary Example

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Data 1	$120	$130	$140	$150	$160	$170	$180	$190	$200	$210
Data 2	$120	$130	$140	$150	$160	$170	$180	$190	$200	$210
Data 3	$120	$130	$140	$150	$160	$170	$180	$190	$200	$210
Data 4	$120	$130	$140	$150	$160	$170	$180	$190	$200	$210
Data 5	$120	$130	$140	$150	$160	$170	$180	$190	$200	$210
Data 6	$120	$130	$140	$150	$160	$170	$180	$190	$200	$210
Data 7	$120	$130	$140	$150	$160	$170	$180	$190	$200	$210

Blue = days before predicted stock (input value), Pink = Actual stock price (Output value)

After getting that necessary information, it’s important to insert these values into the AI model by splitting it into training and testing sets (the ratio of training: testing is 2:1). The two regression models - Linear regression and MLP regression - are then trained and evaluated to determine their performance in predicting stock prices. The stocks used for this experiment are Cisco (CSCO), Tesla (TSLA), Apple (AAPL), Starbucks (SBUX), and Johnson & Johnson (JNJ).

3. Feature Selection and Engineering

The ‘Open’ price of a stock served as a feature to predict its share value. The ‘Open’ price refers to the stock quote when trading starts. It also took into account the different number of days for tomorrow’s stock price. The program was considered 3, 6, 12, and 24 days before the predicted stock price.

This ruled choice of deploying a stock price prediction model that uses only the ‘Open’ price for the underlying stock and varies between 3, 6, 12, and 24 past days can be predominantly attributed to several practical considerations. Initially, the entire model is simplified by employing a single feature stream, which ensures ease of implementation and understanding. The “Open” price is an excellent data point that provides insight into fundamental characteristics and serves as a solid ground in historical stock prices.

On the other hand, short-term forecasts are best accomplished by taking into account a few past days and mainly relying on the “Open” price. First, because short-term stock price movements are influenced by multiple factors, using a set of features is convenient in that it simplifies the model without necessarily losing relevant information.

In addition, sticking to the ‘Open’ price alone and avoiding other features helps in eliminating possible noise or irrelevance. This methodology retains data integrity, thus preventing the entry of redundant variables that could compromise model performance.

While such a simple approach may be an adequate basis for predicting stock prices, it is also crucial to note that the model’s performance greatly depends on issues like the quality of data and the selection of regression methods.

Vol. 2024 (12) 537 – 544
ISSN 2688-3651 [online]

4. Model Selection and Evaluation

This experiment used two regression models to predict stock prices: Linear regression and Multi-Layer Perceptron (MLP) regression.

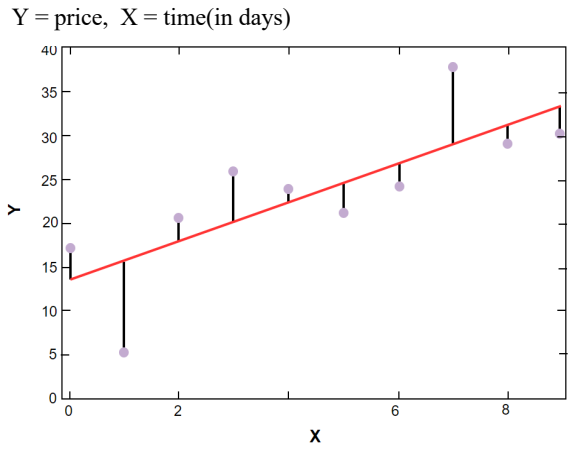

Figure 1: Linear regression example

Linear regression is a straightforward and extensively used regression technique that seeks to create a linear relationship between input features & the target variable (stock prices in this context). It seeks to fit a straight line that best represents the relationship between the historical stock prices and their corresponding input features. Think of it as drawing a line on a graph that comes closest to touching all the points representing past stock prices and their corresponding factors. Once we have this line, we can use it to make predictions about future stock prices based on the new values of those factors. To visualize the figure below could be our model where the purple dots are the actual stock price, the red line is the forecast price, and the goal is to minimize the distance between the purple dots and the red line.

A **Multi-layer Perceptron** (MLP) regression is like a virtual brain composed of many linked building parts. These are neurons, and they are arranged in layers. Each neuron is like a separate equation that changes based on incorrect results in the more suitable direction. Each neuron receives information, processes it, and creates an output. The output of one layer of neurons is sent into the next layer until it gets sent out as the final output value.

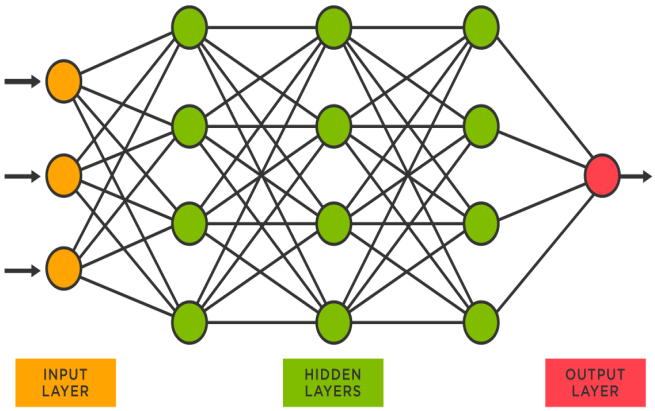

Figure 2. MLP regression example

5. Evaluation of Model Performance:

The performance of the regression models was evaluated using two key metrics: Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE).

MSE is the conventional measure used to calculate the average squared difference in stock prices between predicted and actual values. It measures error scatter and hence quantifies the quality of predictions, with lower MSE values representing better model performance.

$$MSE = \frac{\Sigma(Actual - Forecast)^2}{n}$$

Where n is the number of data points;
Σ represents summation notation;
Actual denotes the original or observed values;
Forecast represents the predicted values.

The purpose of MAPE is to calculate the accuracy percentage of regression models. It computes the mean percentage deviation from actual stock prices predicted by it. A lower MAPE value indicates a more precise model.

Vol. 2024 (12) 537 – 544
ISSN 2688-3651 [online]

$$MAPE = \frac{\Sigma(\frac{|Actual - Forecast|}{|Actual|}) \times 100}{n}$$

n: The number of data points;
Σ represents summation notation;
Actual denotes the original or observed values;
Forecast represents the predicted values;
100: Express as a percentage.

Evaluation metrics used for the regression models for the different periods (3 days, 6 days, 12 days, and 24 days):
mse_train: Mean Squared Error for the training set;
mse_test: Mean Squared Error for the testing set;
mape_train: Mean Absolute Percentage Error for the training set;
mape_test: Mean Absolute Percentage Error for the testing set

Evaluating both models in different period ranges as an input is to find which model paired up with a specific range of periods beforehand produces the most accurate results. The results with the lowest value are the best as it calculates how much error in predicted price vs. actual price.

6. Model Training and Optimization:

6.1 Model Training:

This project consists of two trained regression models, Linear regression and Multi-layer Perceptron (MLP) regression that predict stock prices for arbitrary ticker symbols. The models are trained with historical stock price data obtained from the Yahoo Finance API over 5 years. The data is processed before, and X for the input attributes as well as Y for target values are prepared to train models. Using the train_test_split function from sci-kit-learn, the data is divided into training and testing sets. In the case of Linear regression, sci-kit-learn's class named LinearRegression is used for training the model. In this class, we use the Ordinary Least Squares technique for weight estimation of a linear equation that will provide the best fit to data.

Y= W1X1 +W2X2+ W3X3 + b
Y is the value of the predicted target.
b is an intercept (bias) term.
W1, W2, and W3 represent the weights assigned to X1,X2, and X3 respectively.

MLP regression utilizes the class of sci-kit-learn, known as MLPRegressor. It is a Multi-Layer Perceptron type of neural network class. The model is trained through backpropagation to minimize the weights and biases in it.

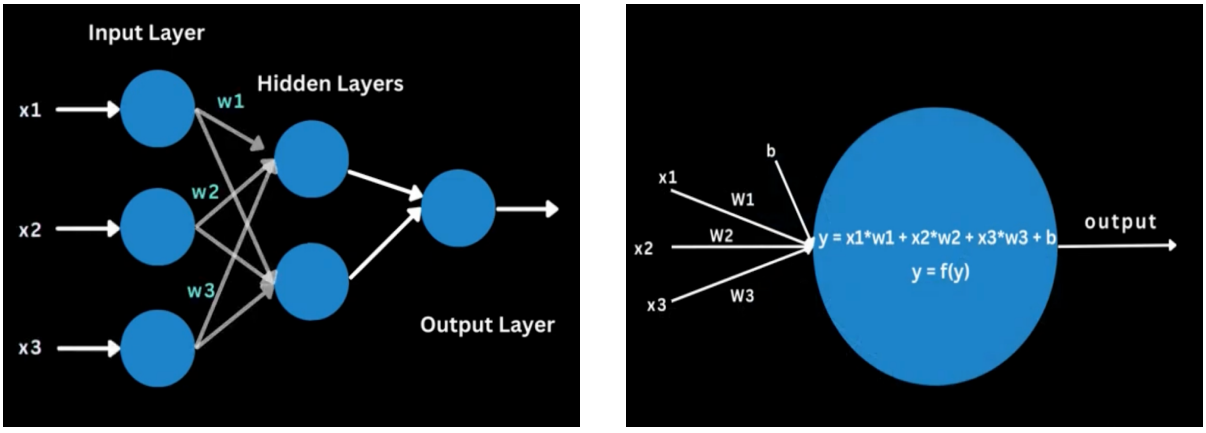

Figure 3. MLP regression equation visualization

Vol. 2024 (12) 537 – 544
ISSN 2688-3651 [online]

In Regression, the weights (W) are trained to fit the data by starting with initial guesses for W. Then calculating predictions using these guesses. Check by finding the errors (differences between predictions with weight and actual data). After that it's necessary to fix by adjusting W to minimize these errors in the suitable direction then repeat these steps until the weight is accurate. The final W values represent the best-fit linear equation to predict outcomes based on input features.

Table 2: Results for Linear Regression

Stock	Days	MSE for Training Set	MSE for Testing Set	MAPE for Training Set	MAPE for Testing Set
CSCO	3	0.4768	0.6979	1.15%	1.27%
CSCO	6	0.4941	0.6562	1.16%	1.23%
CSCO	12	0.567	0.5034	1.21%	1.15%
CSCO	24	0.5574	0.5164	1.20%	1.16%
AAPL	3	5.3072	5.815	1.48%	1.49%
AAPL	6	5.5077	5.5046	1.47%	1.53%
AAPL	12	5.9984	4.5564	1.55%	1.37%
AAPL	24	5.324	6.0755	1.50%	1.47%
TSLA	3	64.4893	51.4406	3.45%	3.36%
TSLA	6	50.5454	80.2694	3.18%	3.56%
TSLA	12	56.6197	68.3385	3.45%	3.67%
TSLA	24	56.8554	67.6055	3.46%	3.60%
SBUX	3	2.1019	2.2706	1.23%	1.28%
SBUX	6	2.2024	2.0874	1.27%	1.22%
SBUX	12	2.1392	2.1979	1.24%	1.27%
SBUX	24	2.0593	2.4666	1.26%	1.27%
JNJ	3	2.7226	3.1027	0.84%	0.86%
JNJ	6	2.8627	2.8059	0.85%	0.85%
JNJ	12	2.7517	3.0613	0.83%	0.89%
JNJ	24	2.9371	2.698	0.84%	0.86%

6.2 Results and Analysis

The stocks used for this analysis are Cisco (CSCO), Tesla (TSLA), Apple (AAPL), Starbucks (SBUX), and Johnson & Johnson (JNJ).

7. Discussion and Interpretation

7.1 General observations:

Overall, the differences in performance between MLP and Linear regression models are relatively minor. MLP often shows marginally better performance with shorter historical data intervals (3-day), while Linear regression tends to perform better or equally well over longer intervals (12-day and 24-day). This suggests that MLP can capture more complex patterns in the data, whereas Linear regression offers consistent performance over longer periods.

Table 3: Results for MLP Regression

Stock	Days	MSE for Training Set	MSE for Testing Set	MAPE for Training Set	MAPE for Testing Set
CSCO	3	0.5829	0.4734	1.19%	1.17%
CSCO	6	0.5678	0.5041	1.18%	1.19%
CSCO	12	0.4932	0.6573	1.17%	1.21%
CSCO	24	0.5536	0.5314	1.19%	1.20%
AAPL	3	5.286	5.8837	1.49%	1.47%
AAPL	6	5.4311	5.6663	1.45%	1.55%
AAPL	12	5.2942	5.9624	1.46%	1.54%
AAPL	24	5.6005	5.3712	1.47%	1.53%
TSLA	3	60.0362	60.6566	3.52%	3.38%
TSLA	6	58.699	62.762	3.75%	3.56%
TSLA	12	61.5865	58.5972	3.84%	3.95%
TSLA	24	61.2439	58.7821	3.68%	3.81%

7.2. Practical implications

Despite the differences, the practical implications of choosing one model over the other are minimal. For example, the maximum observed difference in percent error is 0.28% for TSLA over a 12-day interval. Such small discrepancies are unlikely to significantly impact decision-making in real-world trading, where other factors like market volatility, transaction costs, and external economic indicators play a more substantial role.

7.3 Application to specific stocks:

CSCO: The slight advantage of MLP with shorter intervals (3-day) might not

Vol. 2024 (12) 537 – 544
ISSN 2688-3651 [online]

Table 3: Results for MLP Regression_continued

Stock	Days	MSE for Training Set	MSE for Testing Set	MAPE for Training Set	MAPE for Testing Set
SBUX	3	2.2914	1.8772	1.29%	1.17%
SBUX	6	2.0327	2.4159	1.22%	1.31%
SBUX	12	2.1639	2.1633	1.27%	1.22%
SBUX	24	2.2684	2.0908	1.29%	1.22%
JNJ	3	2.7986	2.9232	0.85%	0.84%
JNJ	6	2.8951	2.714	0.85%	0.84%
JNJ	12	2.9858	2.5599	0.85%	0.84%
JNJ	24	2.6879	3.2836	0.85%	0.86%

justify its complexity. For longer intervals, Linear regression's slight edge could be preferred for its simplicity. AAPL: The better performance of Linear regression with 12-day and 24-day intervals makes it a suitable choice for medium to long-term predictions.

TSLA: The significant advantage of MLP with 12-day and 24-day intervals suggests it might be more effective in capturing TSLA's price patterns.

SBUX: MLP's consistent advantage across most intervals suggests it could be better suited for SBUX, particularly in short-term predictions.

JNJ: The minimal differences across all intervals indicate that either model could be used, with the choice potentially based on other factors like model interpretability and ease of implementation.

Table 4: Comparison between Linear Regression (LR) and MLP

Stock	3 days	6 days	12 days	24 days
CSCO	MLP - Better	MLP - Better	LR - Better	LR - Better
AAPL	MLP - Better	LR - Better	LR - Better	LR - Better
TSLA	LR - Better	Equal	LR - Better	LR - Better
SBUX	MLP - Better	LR - Better	MLP - Better	MLP - Better
JNJ	MLP - Better	MLP - Better	MLP - Better	MLP - Better

However, when considering the longer input or longer interval the Linear regression model has a consistently slight edge over the MLP regression when it comes to the tech companies (CSCO, TSLA, & AAPL) while the MLP regression model is good with the non-tech companies (SBUX & JNJ). The reasoning for this is that the tech companies have a trend of volatile characteristics while the non-tech companies are more static. This means that linear regression could capture the movement of volatile stock, while the MLP regression is more suited for static stocks.

While MLP and Linear regression models each have their strengths, the differences in their predictive capabilities are minor. The choice of model should be guided by specific trading strategies and practical considerations rather than solely by performance metrics.

7.4 Insights from stock price prediction models (prediction performance):

The MSE and MAPE values indicate the accuracy of the models in predicting stock prices. Measures with lower values imply better prediction performance. The performance of the models differs across various stocks and projection days, indicating that certain stocks could be easier to anticipate than others. Impact of Prediction Days: As the number of prediction days increases, the prediction performance worsens for both models.

7.5 Other ideas to consider when using stock prediction with AI in the real-world application

Model Selection: Based on the insights gained from the prediction models, traders and investors can choose the appropriate model (Linear regression or MLP regression) depending on the specific stock and prediction horizon. It is crucial to consider the prediction performance indicators to make wise selections.

Risk Management: Understanding the mistakes and uncertainties of stock price predictions is essential. Traders and investors should be cautious when making decisions based solely on predictions and consider other factors.

Diversification: As different stocks exhibit different levels of predictability, investors may consider diversifying their portfolios across various stocks to reduce overall risk and improve the chances of better prediction outcomes. This should be done to minimize risk by picking stocks with low error.

Vol. 2024 (12) 537 – 544
ISSN 2688-3651 [online]

Continuous Model Improvement: Predictive models can be continuously updated using new data and advanced modeling techniques. Investors and traders should monitor model performance and adapt accordingly.

7.6 Limitations and potential biases of the stock price prediction models

Table 5: Limitation & biases

Limitations	Biases
Lack of data: This data only captures daily stock prices, making it inferior to people who know about the market trends, specific financial data, Sentimental value, ETC.	**Sample bias:** If this model picked up on many unique market swings, it would have a high chance of not being able to properly assess normal market activity
Market changes: This model will only repeat previous market trends, making the model inaccurate when faced with a new change.	**Outlier insensitivity:** If previous data only assesses normal market activity, then it would be able to notice outliers
Data lag: Once again, this data only captures daily stock prices, which would be slow to get on a critical event.	**Outdated info:** This model collects data from a 5-year time period, which might not be the most reliable as new trends pop up every year

7.7 How the Models Could Be Improved or Extended in Future Research:

To improve the model, it is crucial to address the present shortcomings it presents. It's great to start addressing the limitations and biases already listed above. First, the biggest problem is data, as AI can only get as good as the data it gets. There are many kinds of data to choose from for stock prices. A good first choice is intraday stock prices (stock prices throughout the day) because the trading day is a whole realm of missed opportunities for the AI to understand what happens when people trade stocks. It also is a straightforward solution to data lag. Another of the biggest problems is the more human side of predicting based on sentimental analysis and market trends. A fitting solution for that could reside in the NLP (natural language processing) part of AI. For example, for stocks, a good source of data for the more human side of predicting stock would be extracting stock-related news titles that are relevant to the chosen stock. However, one drawback of extracting stock-related news titles is that it wouldn't work well with Linear regression, and the article title may be highly subjective, leading to the AI misinterpreting the information. Still, it could eventually become a positive force to the AI's prediction when the model is built properly. An important factor that would contribute well is the financial report as there would be many reliable indicators, such as financial metrics and sometimes the company's business strategy. Some good financial metrics include revenue, profit margins, debt levels, and growth rates. One thing to be aware of is that financial reports are issued every quarter of the year, and sometimes they could be manipulated to be more favorable than it is, or the financial reports might have a lack of context. Most of the time, financial reports would be a valuable resource for AI, especially for understanding the concept of quarterly sales, which properly assesses when the company is more or less valuable.

7.8 Additions to Experiment with

Hyperparameters: Regularization techniques, batch sizes, and epochs, and adding extra neuron/decision trees to AI models to improve predictability.

Models: There are more models than just Linear and MLP regression. Some other examples of other models might be decision trees, random forests, gradient boosting, recurrent neural networks (RNNs), or long short-term memory networks (LSTMs). It's also possible to combine the models together, which is called the ensemble technique.

Data: Change the 5-year time period to more relevant, like 2-3 years or older, 6-8 years, and see what works out in other areas like training to test ratios. Also, adding more kinds of data that weren't shown here could be beneficial, but it would need experimentation to check if it's a proper fit or would be a burden that adds extra noise.

8. Conclusion:

This study demonstrated that both Linear regression and Multi-Layer Perceptron (MLP) regression models are viable options for stock price prediction, each with its strengths and limitations. The marginal differences in

Vol. 2024 (12) 537 – 544
ISSN 2688-3651 [online]

performance metrics suggest that practical considerations, such as model simplicity and ease of implementation, should influence the choice of the model instead of considering the performance metrics. Although it's one small consistent difference when considering longer input, linear regression was better for volatile stocks, while MLP regression shows less error for static stocks. Future research should focus on enhancing data quality, incorporating additional features like intraday prices and sentiment analysis, and exploring more advanced models to further improve prediction accuracy. These advancements can better equip investors to make informed decisions, manage risks, and optimize their trading strategies.

Acknowledgements:

Odysseas Drosis, my mentor, deserves my deepest appreciation. His helpful ideas and advice were critical in blooming this research endeavor.

References

DeepLearning.AI. (2023). Natural Language Processing (NLP) - A Complete Guide. *Deeplearning AI*. https://www.deeplearning.ai/resources/natural-language-processing/.

Examples. *scikit-learn*. https://scikit-learn.org/stable/auto_examples/index.html#model-selection.

Kumar, A. (2023). Sklearn Neural Network Example - MLPRegressor. *Data Analytics*. https://vitalflux.com/sklearn-neural-network-regression-example-mlpregressor/.

Sinchana-SH. (2024). Sinchana-SH/Stock-Predictions-using-Multilayer-Perceptron-Regression. *GitHub*. https://github.com/Sinchana-SH/Stock-Predictions-using-Multilayer-Perceptron-Regression.

Thelin, R. (2020). Scikit-learn Tutorial: How to Implement Linear Regression. *Educative: Interactive Courses for Software Developers*. https://www.educative.io/blog/scikit-learn-tutorial-linear-regression.

UncomplicatingTech. (2023). Feed Forward NN Working Explained! Deep Learning | Neural Networks | Machine Learning. *Youtube*. https://youtube.com/shorts/1Zwv5PbAbIk?si=q_I5aUVPX2SxVGWc.

Zach, B. (2020). How to Calculate Mean Squared Error (MSE) in Python. *Statology.* https://www.statology.org/mean-squared-error-python/.

Zach, B. (2021). How to Interpret MAPE Values. *Statology.* https://www.statology.org/how-to-interpret-mape/.

Zhou, Y. (2024). Stock Forecasting Based on Linear Regression Model and Nonlinear Machine Learning Regression Model. *Advances in Economics Management and Political Sciences,* 57(1), 7–13. https://doi.org/10.54254/2754-1169/57/20230364.