Abstract:

This research paper delves into the realm of stock price prediction, aiming to empower traders and investors with valuable insights into financial decision-making. It compares the efficiency and utility of the performance of three predictive models, linear regression, neural networks, and random forest regression. Historical stock price data, collected from Yahoo Finance, serves as the basis for generating predictions and evaluating model performance. Results indicate which model outperforms the other. These results will allow us to realize the significance of a specific type of relationship in stock price dynamics. The research also introduces a trading bot that translates predictive insights into actionable decisions. While these models offer promising capabilities, the inherently dynamic and uncertain nature of financial markets must be acknowledged. Nonetheless, this research invites further exploration into refining predictive models and enhancing their reliability in the intricate world of stock price prediction.

Introduction:

In the fluctuating landscape of financial markets, the significance of stock price prediction in financial markets lies in its potential to assist investors and traders in making informed decisions. The potential rewards of accurate stock price predictions are to allow market participants to make well-informed decisions, strategically time their trades, optimize the allocation of their portfolios, and manage risks adeptly.

This research delves into the realm of stock price prediction, undertaking a comparative exploration of linear regression and neural networks. By determining their predictive capabilities, I aim to shed light on the potential insights that these models can provide to guide

traders and investors toward more effective decision-making and empower market participants. The report includes data collection and preprocessing, feature selection and engineering, detailed model analysis, training, optimization, results and analysis, discussions of implications, and a comprehensive conclusion, culminating in a call for further exploration in the domain of stock price prediction.

Related Work:

Several studies have been made that explore similar topics. A majority of these studies found that AI models can be effective in capturing complex patterns in financial data, improving prediction accuracy. However, they often face challenges related to data availability, fine-tuning model settings, and the ever-changing nature of financial markets. In contrast, my research focuses on comparing the performance of two simpler models, linear regression and neural networks, to see how well they predict stock prices. This approach provides real insights into how these models can be used in real-world trading and investment decisions, making it easier to bridge the gap between research and application in financial markets.

Data Collection and Preprocessing:

The foundation of any predictive model lies in the quality of data it relies upon. In this study, historical daily stock price data over a span of five years was collected from Yahoo Finance using the "yfinance" Python library. The chosen ticker 'AAPL,' representative of Apple Inc., served as a canvas upon which to paint the dynamics of stock price movements. However, this ticket can be easily changed, causing the prediction to be made for any stock. This temporal scope was selected to capture a comprehensive range of market conditions, including periods of

stability, volatility, and transformative events. The data collection process, facilitated by the 'history' method, furnished a repository of stock price records. However, the raw data was far from being readily usable. Rigorous preprocessing was necessary to ensure data integrity, consistency, and relevance.

The challenge of aligning input features and output labels emerged as a concern during data preprocessing. Many trials were run to develop the best combination of “Open” and “Closed” prices. The output label (Y) was understood to be the set as the "Open" price of the third, or succeeding day. This alignment was essential to construct a temporal context that could potentially capture evolving patterns and trends. Proper data preprocessing laid the groundwork for subsequent predictive modeling.

Feature Selection and Engineering:

Within the realm of stock price prediction, feature selection is a decisive step that governs the predictive power of the models. The choice of features hinges on their capacity to encapsulate relevant market dynamics and anticipate price movements. In this study, both the "Open" and “Closed” prices emerged as a focal feature. The rationale behind this selection is rooted in the intrinsic significance of the both prices in shaping market sentiment. It encapsulates an immense amount of information, including news, investor sentiment, and market expectations, all of which influence the opening prices of subsequent trading days.

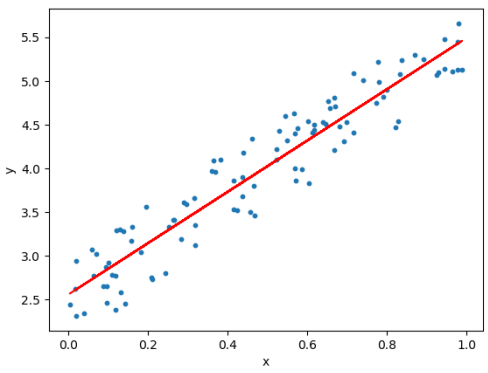
For feature engineering, the "Open" and “Closed” prices of the previous day were chosen to construct the input features (X). Then the current day's “Open” price was used as a third parameter. The next day's "Open" price was earmarked as the output label (Y). This approach aimed to exploit the temporal dimension inherent in stock price data. By examining how the

prices of prior days relate to the "Open" price on the third day, the models could potentially discern patterns that underpin price movements.

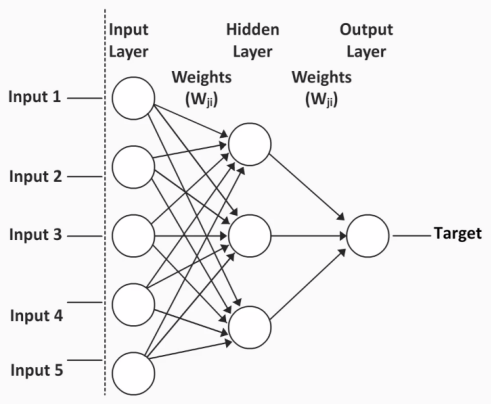
Model Selection and Evaluation:

In order to predict stock prices, two distinct models were enlisted as agents of prediction: linear regression, neural networks, and random forest regression. Linear regression seeks to establish a linear relationship between input features and output labels. Splitting the data is important as we want to have one set of data that is untouched and is unknown. It operates under the assumption that historical price data can be leveraged to infer future price trends. On the other hand, neural networks, inspired by the complex neural pathways of the human brain, offer a more intricate approach. The MLP architecture, known as MLPRegressor in the context of regression tasks, is a form of neural network. Unlike linear regression, neural networks possess the capability to capture nonlinear relationships, making them more adept at handling the intricate and often nonlinear dynamics that characterize financial markets. Similarly to neural networks, random forest is able to handle complex relationships in data. It operates by constructing a multitude of decision trees during training and outputs the average prediction of those trees.

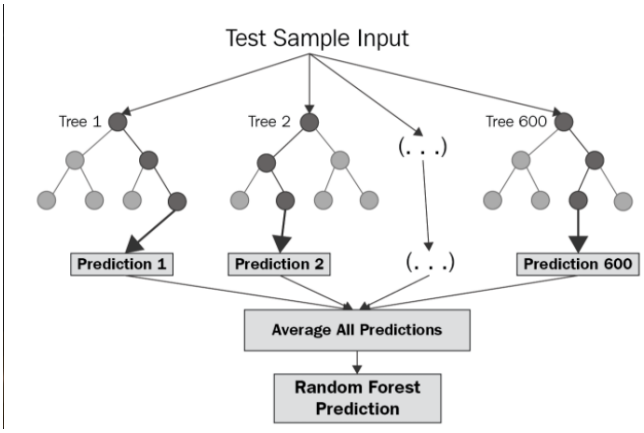
Example of a Linear Regression Model



Example of a Neural Networks



Example of Random Forest Regression



The effectiveness of these models was evaluated using two pivotal metrics: mean squared error (MSE) and R-squared (R2). MSE quantifies the average squared difference between predicted and actual values, providing insights into the models' predictive precision. R2, conversely, offers a measure of the variance captured by the model, thus illuminating its ability to explain the underlying trends in the data.

Model Training and Optimization:

Model training represents the crucible where algorithms imbibe the essence of historical data to refine their predictive capabilities. Linear regression, endowed with a closed-form solution, embarked on a process of parameter learning to optimize its predictive performance. Neural networks, characterized by their intricate architecture, undertook a more iterative process, iteratively adjusting weights through backpropagation. Random Forest, distinguished by its unique approach which allows it to effectively capture complex relationships and patterns in data.

Throughout the training phase, the models underwent calibration to optimize their performance. The guiding lights of this calibration were the MSE and R2 metrics. A model that achieved lower MSE values and higher R2 scores exhibited superior predictions. For example, using the Apple stock (‘AAPL’), linear regression produced a MSE of 5.77, neural network produced 6.46, and random forest regression produced 6.49. As a lower MSE in MLPRegressor signifies higher prediction accuracy, these values are extremely beneficial in creating valuable stock price predictions.

Results and Analysis:

The outcomes of the predictive models unveiled many new insights into stock price movements. Both linear regression and neural networks showcased promising capabilities in forecasting stock prices based on historical data. When measured against the metrics of MSE and R2, the linear regression model frequently outshone its neural network counterpart.

The performance divergence between the two models underscores a fundamental idea of financial markets: they are dynamic, nonlinear systems shaped by many influences. While linear regression can capture some trends, it may also often fall short in capturing the complex interactions that can lead to sudden shifts and fluctuations in stock prices. Therefore, further improving the prediction capabilities was necessary.

Combination of Models:

Although stock price predictions were obtained, creating a system in which they became more accurate was possible. By combining the predictions from linear regression, neural networks, and random forest regression, the accuracy of the prediction would be enhanced. This

would be done by giving different weights to the predictions from each model based on their individual performance on the test data.

First, the testing data was utilized. The error if each model was calculated by computing the absolute difference between the actual test values and each model's prediction. The weights are initialized equally at 0.33 for each model. Then, these weights are adjusted based on the magnitude of the prediction errors, so models with lower errors obtain higher weights. Then, a weighted average of the three model predictions is calculated, which results in a new prediction for each data point in the test set. This process is repeated for all test data points.

The same process is repeated for the training data and adjusted the weights based on training data errors.

Finally, the outcome is a set of predictions for both the test and training datasets. By combining the models based on their individual performance and assigning weights, the overall accuracy of these predictions are tremendously improved. Once acquiring the new predictions, Mean Squared Error (MSE) is re-evaluated to determine the extent to which the quality of the predictions were improved. Through the testing of countless stocks on this method of combining the models, the MSE was regularly less than each model’s individual performance, which signifies a large advancement in the capabilities of this stock price prediction AI.

MSE Results:

Results of Testing (10/15/2023)

	Linear Regression	**Neural Network**	**Random Forest**	**Combination of Models**
Apple (AAPL)	6.12	7.06	7.99	6.09
Tesla (TSLA)	57.70	63.64	76.46	57.75
Amazon (AMZN)	8.86	8.98	10.07	8.70
Johnson & Johnson (JNJ)	3.57	3.90	3.77	3.53
The Boeing Company (BA)	38.63	43.39	52.04	38.10
Exxon Mobil Corporation (XOM)	2.02	2.27	2.73	2.02
Average:	19.483	21.54	25.51	19.365

Results of Training (10/15/2023)

	Linear Regression	**Neural Network**	**Random Forest**	**Combination of Models**
Apple (AAPL)	5.63	6.40	0.97	0.98
Tesla (TSLA)	64.74	70.33	11.25	11.26
Amazon (AMZN)	8.60	10.05	1.47	1.47
Johnson & Johnson (JNJ)	2.54	2.96	0.41	0.41
The Boeing Company (BA)	47.29	52.35	8.40	8.38
Exxon Mobil Corporation (XOM)	1.80	2.09	0.30	0.27
Average:	21.76	24.03	3.8	3.795

Trading Bot:

Upon obtaining the predictions, a trading bot simulates real world application of the models was added. This code transforms the predictive insights made earlier into actionable decisions. The code simulates a trading scenario with an initial budget of $5000 and zero stocks owned. It then iterates through the data points, which represent different trading days, and assesses whether the actual opening prices (Y) for those days are lower or higher than the predicted opening prices for the following day. If the prediction indicates a price increase, and the budget is sufficient, the code purchases stocks. Conversely, if the prediction suggests a price decrease, the code sells stocks if any are held. The budget and stock count are updated accordingly. The results of the bot depend on how good its predictions are and whether its buying and selling rules work well. Such automated trading tools are becoming more popular in financial markets, offering a systematic way to make decisions based on data, with the goal of making money while managing risks.

Conclusion and Discussion:

Attempting to predict stock prices can lead to lots of potential, as well as challenges. The models, be they linear regression, neural networks, or random forest regression, offer valuable insights into potential price movements, equipping investors and traders with tools to refine their strategies. However, the ensemble model, combining insights from multiple models, contributed to a much improved AI than any of the individual models, with exponentially more potential to output accurate stock price predictions. On average, the MSE of the combination of models was lower than the average of any of the models on their own. Additionally, the trading bot takes all

the layers of information into account and translates them into actionable decisions in a systematic manner.

This research contributes a distinct aid to financial forecasting, inviting traders and investors to explore the power of machine learning in their decision-making processes. Yet, future stock prices remain uncertain, and predictive models, while helpful, are not infallible. These predictions are unable to consider external events, like political issues or changes in investor sentiment, which can impact stock prices. Also, a trading bot is far from a human trader, who can adapt and reason to information that might not be reflected in historical data. The potential for further exploration abounds, from the integration of more features to the exploration of advanced modeling techniques, each promising to unlock new layers of understanding in the intricate world of stock price prediction. Through the continuous research in this area, the potential to provide investors with great tools for navigating the stock market is exponential.