Engineering and Technology Journal e-ISSN: 2456-3358

Volume 09 Issue 08 August-2024, Page No.- 4978-4986

DOI: 10.47191/etj/v9i08.50, I.F. – 8.227

A Comparative Analysis of LSTM, ARIMA, XGBoost Algorithms in Predicting Stock Price Direction

Aiyegbeni Gifty[1], Dr. Yang Li[2]

[1,2]University of East London, Department of Engineering and Computing, London, England, United Kingdom

ABSTRACT: This research report presents a comprehensive investigation into the prediction of Google's stock prices using advanced machine-learning techniques. The study focuses on assessing the predictive capabilities of three distinct algorithms: XGBoost, LSTM, and ARIMA, applied to historical stock price data with a specific emphasis on close prices. The primary goal is to develop accurate univariate models to forecast the closing stock price for the next day, a crucial aspect of financial decision-making. The evaluation of model performance utilizes a range of metrics including R-squared, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) to provide insights into predictive accuracy. Furthermore, the study explores the effectiveness of hyperparameter tuning and ensemble methods in optimizing model performance. The findings highlight the strong performance of the XGBoost model, which achieves a notable R-squared value and effectively minimizes error metrics. While ensemble techniques exhibit potential, they do not consistently outperform all individual models. The subsequent hyperparameter tuning of the XGBoost algorithm achieves a higher R-squared value of 99.47%, accompanied by an MAE of 15.98 and an RMSE of 27.34. This research contributes valuable insights into the potential of machine learning for stock price prediction, emphasizing the importance of thoughtful model selection and parameter optimization.

KEYWORDS: Stock prediction; Stock prices; Machine learning; Ensemble models; Hyperparameter tuning.

1. INTRODUCTION

The stock prices of a company can be used as a metric for evaluating the financial performance of the business. Stock prices can draw in investors, or churn off investors from a company [1]. This is because large fluctuations in stock prices could cause adverse impacts on companies, investors, and economies. If stock prices can be properly forecasted, investors and company owners will be able to make targeted actions to set a balance in the financial market [2].

Predicting the prices of stocks has been a very interesting topic in finance for decades. After the epidemic, stock prices began to fluctuate even more widely [3]. These predicted stock prices have been a crucial topic for several areas in recent years. Researchers, practitioners, academia, and businesses have been looking to explore and figure out the trend of stock prices based on existing data [4].

Stock prices have been seen to be influenced by several macro-economic factors like; financial news, interest rates, company policies, inflation rates, epidemics, commodity price index, investors' expectations, political events, social factors, institutional investors' choices, and even economic conditions [5], [6]. All these are put into consideration when trying to predict stock prices in the finance industry. Hence, understanding and dealing with the stock market and stock prices requires expertise, resources, as well as up-to-date information [7].

For a very long time, researchers have been on the case of predicting how stock prices will move to maximize profits. Some researchers have looked at machine learning techniques and tried to explore them to a great extent [5].

This work aims to collect the stock price of Google LLC over a reasonably long period of five years and develop a robust forecasting framework for forecasting the Google closing stock price using different machine learning algorithms.

2. LITERATURE REVIEW

2.1 Overview

Stock, also known as a share or equity, represents ownership in a company. When a company decides to raise capital by going public, it divides its ownership into shares and offers them to the public in the form of stocks [8]. By purchasing stocks, investors become partial owners of the company and are entitled to a portion of its profits and assets [9]. Stock prices, on the other hand, refer to the value at which these shares are bought and sold on a stock exchange. They represent the current market price of a particular stock at any given time. Stock prices are determined by the interaction of supply and demand in the stock market, influenced by various factors hence stock prices can be very volatile [10], [11].

Stock prediction is the process of using various techniques, models, and data to estimate future movements and trends in stock prices [12]. Stock prediction is essential because it has the potential to provide valuable insights and information to

investors, traders, financial institutions, and the overall market. Amongst several other reasons, stock prediction is key for investment decision-making, risk management, market timing, trading strategies, portfolio optimization, market efficiency, and economic indicators [6], [7].

2.2 Google Stock Performance

Google is a division of Alphabet Inc. which is an American multinational technology conglomerate holding company that specializes in artificial intelligence, online advertising, search engine technology, cloud computing, computer software, quantum computing, e-commerce, and consumer electronics. Along with Amazon, Apple, Microsoft, Meta, and its parent firm Alphabet, these companies are known as the "Big Five" in the United States. Google's stock price has been increasing overall since its initial public offering though there have been a few fluctuations here and there [13]. Alphabet Inc. was created through a restructuring of Google on October 2, 2015, and became the parent company of Google and several former Google subsidiaries.

As of June 14th, 2023, the closing price for Google stock price was $124.38 per share with about 12,781,000,000 shares outstanding [14].

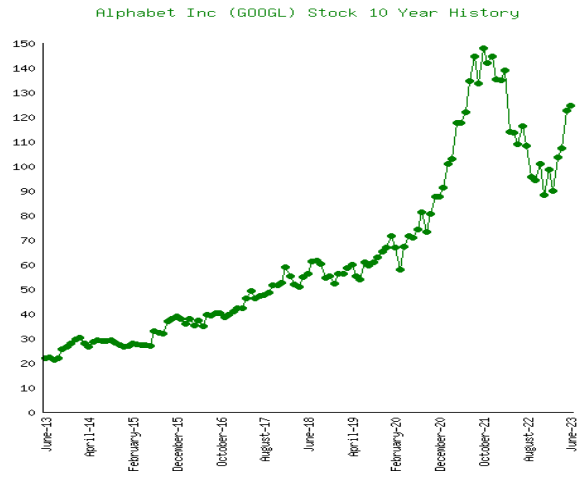

Figure 1 GOOGL 10-Year Adjusted close Price Chart

The graph above displays the closing prices for Alphabet Inc. (GOOGL) over the previous ten years. In general, the price of Google's stock has usually increased over time, with sporadic spikes in volatility which are dependent on several variables, such as market conditions, economic developments, and corporate performance [15].

2.3 Related Works

Due to the rising need for the prediction of stock prices, in recent times, several kinds of research have been carried out to develop more suitable and efficient machine models to make stock price predictions [1], [16].

Some previous researchers have looked into predicting Google stock prices. [3] for instance, worked on predicting Google stock price with the use of linear regression and random forest algorithms. The aim was to use the YouTube platform to study the Google stock price trend. Predictions were also made to investigate if there are some traces of factors affecting the stop price. The price was then predicted by the author using the techniques of linear regression and random forest regression. The linear regression prediction results' inaccuracy was less than 5%, which is within the usual range, but the random forest regression's accuracy for the next five days' predictions is substantially lower (65%).

In [17], the researchers worked on understanding the stock market price trend properly. They worked with the Stock Market Turnover Ratio which was gotten from the Federal Reserve Bank of St. Louis. The dataset includes information on the total share price exchanged throughout various periods in comparison to the average market capitalisation for certain timespans. The research employed the Time Series Linear Model (TSLM) with Support vector machines and also an Autoregressive integrated moving average (ARIMA). Based on the results obtained, it was confirmed that the TSLM model was robust and could predict stock prices to a reliable degree.

Similarly, some research has compared several models in predicting the stocks of more than one company from historical data. [18] evaluated three models in predicting four different stocks from the Yahoo finance database. The machine learning algorithms considered were Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Support Vector Regression (SVR), and the companies used for the study were Apple, Mastercard, Ford, and ExxonMobil. Except for Mastercard, for which data are available starting in 2006, the data for these stocks span from January 1st, 2002, to March 11th, 2020. The SVR model was able to predict the stock prices of these companies with the highest accuracy compared to the CNN and LSTM models.

Some work has also been done on optimizing the hyperparameters of models to yield better predictions. [19] focused on hyperparameter optimization for LSTM models. It involved the implementation and comparison of three metaheuristic algorithms: Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Symbiotic Organism Search (SOS) in predicting stock.

The use of machine learning algorithms and predictive models has gained attention in stock prediction research. These related works discussed above have attempted to identify patterns and relationships in historical data to predict future stock prices. These studies have reported promising results using machine learning approaches while a few have explored hyperparameter tuning and combining multiple predictive models to enhance accuracy and mitigate individual model weaknesses [20]. Ensemble methods and hyperparameter optimization have shown the potential to improve forecasting performance.

3. METHODOLOGY

This section presents an overview of the steps undertaken in this project aimed at predicting stock price movements. The process involves data acquisition and pre-processing, algorithm modeling, and performance evaluation.

3.1 Dataset Description

The dataset for this project was obtained from a Kaggle website and can be accessed at https://www.kaggle.com/datasets/shreenidhihipparagi/google-stock-prediction. The dataset is named Google Stock Price Data Set and it contains data for 5 years from 14th June 2016 to 11th June 2021. There are 1258 instances of Google stock price information with no duplicates or missing values. The dataset has 14 variables with 13 independent variables and one dependent (target) variable which will be the close price for the next trading day.

3.2 Machine Learning Algorithms

3.2.1 LSTM (Long Short-Term Memory):

LSTM is a type of recurrent neural network (RNN) that is particularly effective in modeling and predicting sequences of data. It overcomes the limitations of traditional RNNs by incorporating memory cells and gating mechanisms [20], [21],[22]. The gating mechanisms control the flow of information within the network, allowing it to selectively retain or forget information at each time step.

3.2.2 ARIMA (AutoRegressive Integrated Moving Average):

ARIMA is a widely used time series forecasting method that combines auto-regression, differencing, and moving average components. ARIMA models are valuable for modeling and predicting time-dependent data in various fields, including finance and economics [23], [24]. While ARIMA is a powerful tool, it assumes linearity and stationary data, making it less suitable for handling certain non-linear and volatile time series patterns.

3.2.3 XGBoost (Extreme Gradient Boosting):

XGBoost is a powerful machine learning algorithm known for its effectiveness in predictive modeling and for winning numerous data science competitions. It is an optimized implementation of the gradient boosting algorithm, which combines multiple weak predictive models (decision trees) to create a strong ensemble model [20]. XGBoost excels in various tasks, including classification, and ranking problems. It is known for its speed, scalability, and ability to handle large datasets with high-dimensional features [25].

3.2.4 Ensemble Modeling

Ensemble modeling is a technique in machine learning where multiple diverse models are combined to improve prediction accuracy. It works by training different models on the same data and then aggregating their predictions using methods like averaging or voting [18]. This approach helps overcome individual model weaknesses and provides more reliable and robust forecasts [20]. But while ensemble models often exhibit improved predictive performance compared to individual models, their effectiveness can vary based on factors such as the diversity of base models, the quality of predictions from individual models, and the nature of the dataset [26].

3.2.5 Hyperparameter Tuning

Hyperparameter tuning refers to the process of finding the optimal values for the hyperparameters of a machine-learning model [19], [27]. Hyperparameters are parameters that are set before the learning process begins, and they affect the behavior and performance of the model. Tuning involves systematically searching through a predefined range of values for these hyperparameters to identify the combination that results in the best performance on a specific task or dataset [27]. The goal is to improve the model's accuracy, generalization, and overall effectiveness by fine-tuning these parameters [28].

3.3 Performance Metrics for Evaluation

Model evaluation is the process of assessing the performance and effectiveness of a predictive model [11]. During model evaluation for this project, the metrics that will be analysed are;

3.3.1 Root Mean Square Error (RMSE):

RMSE is a commonly used metric to measure the average magnitude of prediction errors. It calculates the square root of the mean of the squared differences between predicted values and actual values [11], [21]. RMSE is expressed in the same units as the target variable and provides a measure of the overall prediction accuracy.

Formula [20]:

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i-O_i)^2}$$

Equation 1 Equation for Root Mean Square Error (RMSE)

Where:

n is the total number of samples in the dataset.

y_pred (E_i) represents the predicted values.

y_actual (O_i) represents the actual (observed) values.

3.3.2 Mean Absolute Error (MAE):

MAE is another metric used to assess the average magnitude of prediction errors. It calculates the mean of the absolute differences between predicted values and actual values, disregarding the direction of the errors [1], [21].

Formula [20]:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|E_i-O_i|$$

Equation 2 Equation for MAE (MAE)

Where:

n is the total number of samples in the dataset.

y_pred (E_i) represents the predicted values.

y_actual (O_i) represents the actual (observed) values.

3.3.3 Coefficient of Determination (R^2):

The Coefficient of Determination often denoted as R^2, measures the proportion of the variance in the dependent variable that can be explained by the independent variables in the model [2], [21]. It indicates how well the model fits the observed data. R^2 ranges from 0 to 1, where 1 indicates a perfect fit and 0 suggests that the model does not explain the variance in the dependent variable.

Formula [20]:

$$R^2=1-(SSR/SST)$$

Equation 3 Equation for Coefficient of Determination (R^2)

Where:

SSR (Sum of Squared Residuals) represents the sum of the squared differences between the predicted values and the mean of the dependent variable.

SST (Total Sum of Squares) represents the sum of the squared differences between the actual values and the mean of the dependent variable.

4. IMPLEMENTATION

4.1 Data Preparation and Preprocessing

The project begins with data preparation and preprocessing to ensure the dataset's suitability for analysis.

Data profiling is done to examine and summarize key characteristics of the dataset to assess its quality, understand its content, and identify potential issues or anomalies.

Furthermore, data cleaning is done to identify and rectify errors, inconsistencies, and inaccuracies in the dataset to improve its quality and reliability for analysis and decision-making. The dataset is carefully examined for any missing values and duplicated values.

For feature selection, categorical variables need to be in numerical form hence the dataset is inspected to check if all columns are in the right data types.

The Date variable is converted to date-time format. This is because the DateTime format provides more information about the date, such as the day, month, year, and hour. This information can be used to create more informative visualisations.

Data visualization allows for a better understanding of the spread, skewness, and presence of extreme values in the data.

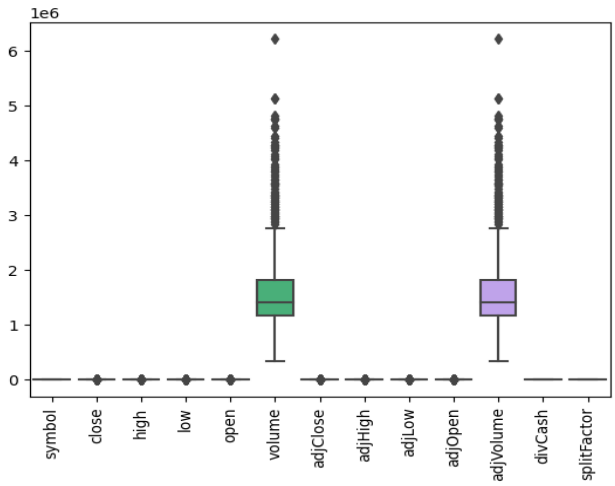

Figure 2 Box plots of all the variables

The boxplot shows that some of the ranges for volume and adjvolume are way more than others, which means that the volume distributions have larger variability or higher dispersion. To resolve this, the dataset needs to be scaled to ensure that all variables are on a comparable scale, helping to mitigate the impact of differences in magnitudes in feature selection.

4.2 Exploration of Target Variable

It is also important that the target is explored to gain insights into its distribution. This allows for a better understanding of its spread, skewness, and presence of extreme values in the data. Line graphs are also employed to see the progression of the dependent variable over the years to observe trends and patterns in the data over time.

4.3 Correlation Matrix

Correlation is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.

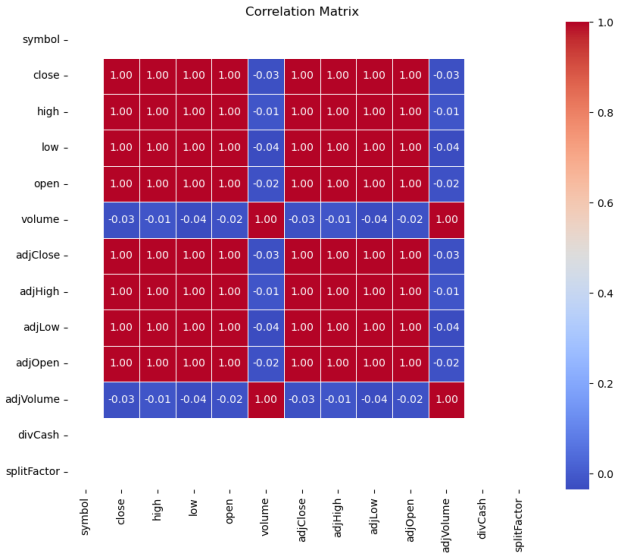

Figure 3 Correlation matrix of the variables

In the heatmap generated, red indicates a positive correlation between two variables. When two variables are positively correlated, it means that as one variable increases, the other variable also tends to increase. Blue indicates a negative correlation between two variables. A negative correlation implies that as one variable increases, the other variable tends to decrease. As seen, most of the variables have perfectly positive correlations with other variables save for volume and adjusted volume which appear to have a negative correlation with other variables. Meaning that as volume increases, the other variable tends to decrease.

4.4 Feature Scaling

Feature scaling is a preprocessing technique used in machine learning to standardize or normalize the range of independent variables or features in a dataset. The goal of feature scaling is to bring all features to a similar scale, ensuring that they contribute equally to the feature selection process and preventing certain features from dominating others due to their larger magnitude during feature selection. The 'MinMaxScaler()' function is used to scale the data and all independent features are brought to a range between 0 and 1 ready for feature selection.

4.5 Dataset Splitting

The 'Close' price variable of the stock is the only feature utilized for modeling purposes. To begin, the data is manipulated by shifting the 'Close' price values upwards by one position. This creates a new variable named 'Next_Day_Close', where each entry corresponds to the 'Close' price of the subsequent day. Essentially, this transformation aligns the data such that each day's 'Close' price becomes the predictor for the

'Next_Day_Close' value, facilitating the development of a predictive model to forecast future stock prices based on the previous day's closing price.

Then the closing price variable is split into training and testing sets to accurately evaluate model performance. The training set is used to teach the model the underlying patterns and relationships within the data, while the test set evaluates its ability to make accurate predictions on new instances. This process helps to detect overfitting, where the model memorizes the training data but fails to perform well on new data.

The data is split with a ratio of 80:20 (80% training data and 20% testing data) to have a sufficient amount of data for both training and testing to ensure a well-performing and reliable machine learning model. With 1006 instances for training and 252 instances for testing the models.

5. MODEL BUILDING

After splitting, the study utilizes three individual machine learning algorithms to make predictions. Following this, the top-performing model undergoes optimization through hyperparameter tuning. This refined model is then compared to two ensemble models, one which is constructed using all three models and the second with the two best-performing individual models.

5.1 Individual Algorithms

5.1.1 XGBoost Algorithm

An XGBoost regression model is created, trained using the training data, and then used to predict stock prices for the test data. The Root Mean Squared Error value of 30.24 and MAE value of 17.63 provide insight into the model's prediction accuracy. Lower values for both RMSE and MAE are generally indicative of better model performance. The R-squared value of 0.99 signifies that the model captures a high proportion of the variance in the data, indicating a strong goodness of fit between predicted and observed values. These metrics demonstrate the model's ability to approximate the actual outcomes effectively.

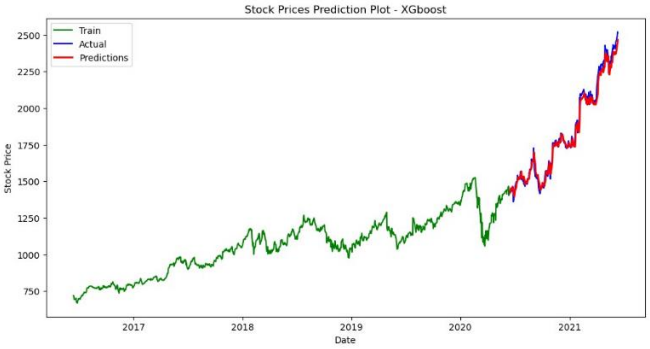

Figure 4 Plot of the XGBoost forecast over time

The plot above visualizes the train, test, and predicted values of the XGBoost model over time.

5.1.2 LSTM Algorithm

For LSTM the training data is preprocessed using a sliding window approach to create input sequences (x_train) and corresponding target values (y_train). The architecture of the model is summarized, and it is configured for training with the mean squared error loss function and the Adam optimizer.

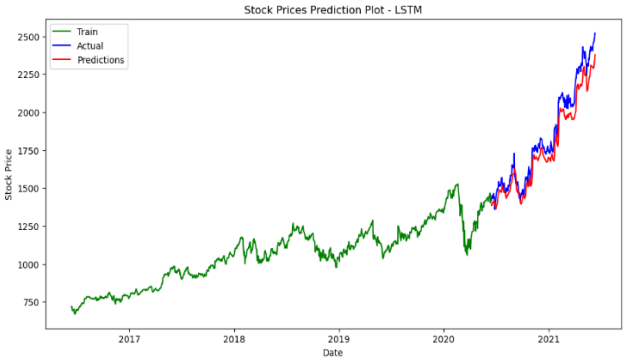

Figure 5 Plot of the LSTM forecast over time

The model gives an RMSE value of 57.28 and an MAE value of 49.35 illustrating that the model's overall accuracy is quite good. The R-squared value of 0.968 indicates that approximately 97% of the variability in the data is captured by the model, reflecting its strong ability to explain observed fluctuations. The plot above visualizes the train, test, and predicted values of the LSTM model over time.

5.1.3 ARIMA Algorithm

The ARIMA model is fitted for time series forecasting by setting the model's order parameters, autoregressive (p), differencing (d), and moving average (q) orders set as 1, 2, and 5, respectively. The ARIMA model is trained and subsequently, future stock prices are forecasted for a specified number of steps using the trained ARIMA model.

The RMSE for this model stands at 188.11, indicating the average magnitude of prediction errors. The MAE measures 140.12, signifying the average absolute difference between predicted and actual values. The R-squared value is 0.66, suggesting that the model captures approximately 66% of the data's variance. This is a good model. The plot below visualizes the train, test, and predicted values of the ARIMA model over time.

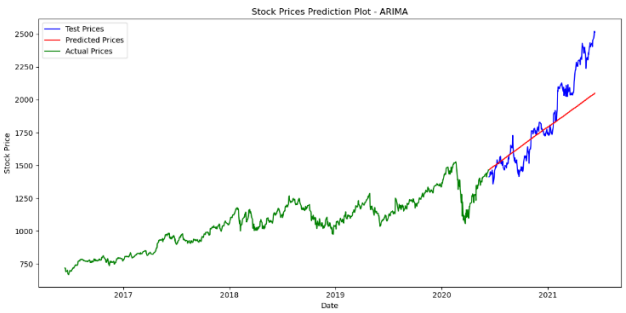

Figure 6 Plot of the ARIMA forecast over time

5.1.4 Individual Model Evaluation

Here in the individual model evaluation, we assess the performance and quality of each predictive model's predictions against actual outcomes to determine the best model.

Table I Model Evaluation for Individual Models

	R-Squared	MAE	RMSE
XGBOOST	99.35%	17.63	30.24
LSTM	96.87%	49.35	57.28
ARIMA	66.37%	140.12	188.11

Upon thorough evaluation, the XGBoost model stands out as the most favorable option due to its exceptional performance across multiple key metrics. With a remarkable R-squared value of 99.35%, the XGBoost model adeptly captures the underlying variability in the dataset, indicating a robust representation of the actual trends and behaviors. This high R-squared value demonstrates the model's proficiency in explaining the observed outcomes with precision.

Furthermore, the XGBoost model achieves the lowest MAE of 17.63 and RMSE of 30.24 in comparison to the alternative models, namely LSTM and ARIMA. This signifies that the XGBoost model's predictions exhibit substantially smaller discrepancies from the actual values, indicating superior predictive accuracy and a closer fit to real-world observations.

Taking a comprehensive view, the XGBoost model's outstanding combination of a high R-squared value and minimal prediction errors, as reflected by the lower MAE and RMSE scores, establishes it as a robust and reliable choice. Its accurate forecasts positions the XGBoost model as the superior option for predicting stock prices.

The second-best model among the evaluated algorithms is the LSTM model. While it demonstrates a slightly lower R-squared value compared to the XGBoost model, it still offers a substantial level of accuracy. Additionally, the LSTM model yields an MAE of 49.35 and an RMSE of 57.28. Although these values are higher than those of the XGBoost model, they indicate that the LSTM model's predictions maintain relatively close alignment with the actual values. Despite not achieving the top spot, the LSTM model proves to be a solid contender for predicting stock prices, offering reasonable performance and predictive capabilities compared to the ARIMA model, which exhibits comparatively lower performance among the evaluated algorithms.

5.2 Optimization and Ensembling

In this section a new model is created by optimizing the best model, XGBoost will be optimized by hyperparameter tuning, an ensemble model will be created by combining the ARIMA, LSTM, and XGBoost algorithms, and finally, another model with the two best-performing algorithms (XGBoost and LSTM) since the ARIMA model's accuracy didn't match up.

5.2.1 XGBoost Model Optimization

In this step, hyperparameter tuning is performed using GridSearchCV to optimize the XGBoost model's performance. The parameter grid is defined with variations in the number of estimators, learning rates, and maximum depth of the model. The process involves fitting the model with different combinations of these hyperparameters and evaluating their performance using cross-validation with three folds. The scoring metric used is the negative mean squared error, aiming to minimize prediction errors. After the grid search, the best-performing model is selected based on the optimized hyperparameters. Subsequently, this optimized model is used to predict stock prices on the test data, resulting in improved prediction accuracy compared to the initial model configuration. The hyperparameter tuning process using GridSearchCV reveals the best parameters for the XGBoost model: a learning rate of 0.1, a maximum depth of 3, and 100 estimators. This fine-tuned model is showcasing improved performance metrics on the test data, achieving an RMSE of 27.34 and an MAE of 15.98. Furthermore, the R-squared value is standing at 0.995, signifying that the model is capturing a substantial portion of the data's variance and generating accurate predictions with a strong alignment to the actual values. This optimal parameter combination results in heightened predictive capabilities than that of the previous XGBoost model, rendering the model highly suitable for accurate stock price forecasting.

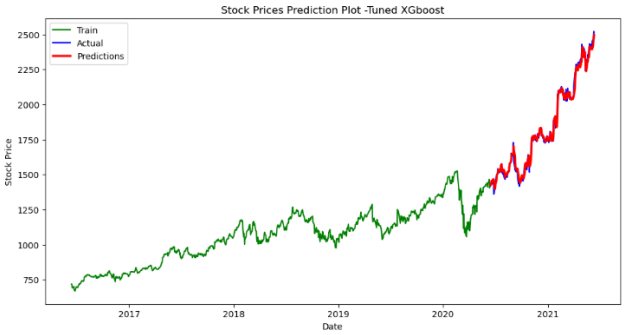

Figure 7 Plot of the Tuned XGBoost forecast over time The plot above visualizes the train, test, and predicted values of the Tuned XGBoost model over time.

5.2.2 Ensemble 1 – 3 Algorithms

In the current phase of the analysis, three distinct models are being employed to predict stock prices: ARIMA, XGBoost, and LSTM. For the ARIMA model, an order of (1, 2, 5) is utilized, and it is fitted to the training data. Similarly, the XGBoost model is constructed with specific hyperparameters, including 100 estimators and a learning rate of 0.1, and it's trained using the scaled input and output data. On the other hand, the LSTM model, defined with an input shape of (number of features, 1), consists of an LSTM layer followed by a dense layer and is trained over 50 epochs. After individual forecasts are obtained from each model, an ensemble forecast is generated by averaging these predictions. Finally, the ensemble forecast is inverse-transformed to revert to the original data scale, yielding the anticipated stock prices. The RMSE measures the average magnitude of prediction errors, with a value of 149.94 indicating the model's typical prediction error. The MAE quantifies the average absolute difference between predicted and actual values, with a score of 110.64. The Ensemble R-squared, at 0.79, indicates the proportion of variance in the target variable that the model captures.

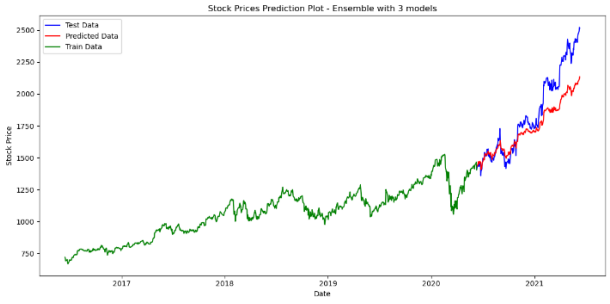

Figure 8 Plot of the Ensemble 1 forecast over time

The plot above visualizes the train, test, and predicted values of the Ensemble 1 model over time.

5.2.3 Ensemble 2 – 2 Best Algorithms

For the second ensemble model, two algorithms, XGBoost and LSTM, are created and fitted for forecasting. The XGBoost model utilizes 100 estimators with a learning rate of 0.1 and is trained on the scaled input of the first feature. It generates predictions for the target variable using the scaled input data. On the other hand, the LSTM model is designed with a sequential architecture, comprising a 50-unit LSTM layer activated by ReLU, followed by a dense layer. This model is trained for 50 epochs using batches of size 16 and predicts the target variable based on the scaled input. The ensemble forecast is then computed by averaging predictions from both the XGBoost and LSTM models. This combined forecast is subsequently transformed back to actual stock prices using an inverse scaling process, producing a more refined prediction that leverages the strengths of both models.

This ensemble model achieves an RMSE of 141.03, indicating the average error between the predicted and actual values. The MAE is 107.24, representing the average absolute difference between predictions and actual values. The ensemble model's R-squared value is 0.81, indicating the proportion of variance in the target variable that can be explained by the combined predictions. These metrics collectively demonstrate the ensemble model's performance in accurately forecasting stock prices. The plot below visualizes the train, test, and predicted values of the Ensemble 2 model over time.

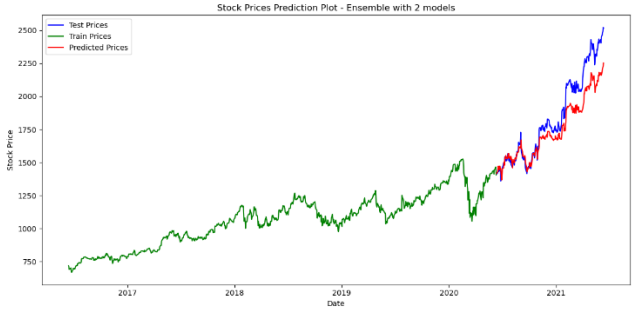

Figure 9 Plot of the Ensemble 2 forecast over time

5.2.4 Model Evaluation

	R-Squared	MAE	RMSE
Tuned XGBOOST	99.47%	15.98	27.34
Ensemble 1	78.63%	149.94	104.64
Ensemble 2	81.10%	107.2	141.03

Table II Model Evaluation for the Tuned and Ensemble Models

Upon thorough examination of the model performance metrics, the Tuned XGBoost model emerges as the most robust contender for predicting stock prices. With an exceptional R-squared value of 99.47%, the Tuned XGBoost model achieves the lowest MAE of 15.98 and RMSE of 27.34, signifying its capability to make accurate predictions with minimal deviations from the true values.

In contrast, Ensemble 1 (built with 3 algorithms), exhibits an R-squared value of 78.63%, indicating a relatively weaker correlation between its forecasts and actual prices, and it also yields a higher MAE of 149.94 and RMSE of 104.64. Ensemble 2 (built with the 2 best algorithms), on the other hand, demonstrates improved performance compared to Ensemble 1, with an R-squared value of 81.10%, an MAE of 107.20, and an RMSE of 141.03. Despite this improvement, the Tuned XGBoost model maintains its superiority by offering the most accurate and reliable stock price predictions based on the considered evaluation metrics.

Based on the results, we conclude that in this specific scenario, hyperparameter tuning of the XGBoost model is the more effective approach for accurate stock price prediction.

5.2.5 Comparison of All Model Performance: Tuned XGBoost Vs. Two Ensembles Vs. XGBoost, LSTM, and ARIMA Models

Table III Evaluation of all the model's performance

	R-Squared	MAE	RMSE
Tuned XGBOOST	99.47%	15.98	27.34
Ensemble 1	78.63%	149.94	104.64
Ensemble 2	81.10%	107.2	141.03
LSTM	96.87%	49.35	57.28
ARIMA	66.37%	140.12	188.11
XGBOOST	99.35%	17.63	30.24

Comparing the models all together, the Tuned XGBoost model shows the best overall performance with the highest R-squared value of 99.47% and the lowest MAE and RMSE values, indicating accurate predictions and a strong fit to the data. The XGBoost model exhibits the next highest overall performance, closely followed by the LSTM model. Both ensemble 1 and ensemble 2 models show intermediate performance, as they have slightly lower accuracy and slightly higher errors compared to the Tuned XGBoost, XGBoost, and LSTM models. The ARIMA model has the lowest R-squared value and higher errors, suggesting that it may not capture the underlying patterns in the data as effectively as the other models.

6. DISCUSSION

In the study's findings, XGBoost demonstrates exceptional performance, highlighting its ability to capture intricate data relationships and make accurate predictions. The subsequent hyperparameter tuning further refines XGBoost's predictive capability, underlining the importance of optimizing

hyperparameters to enhance accuracy. This achieves a high R-squared value of 99.47%, accompanied by a MAE of 15.98 and an RMSE of 27.34.

The first ensemble model, comprising ARIMA, XGBoost, and LSTM, yields an R-squared value of 78.63%, an MAE of 149.94, and an RMSE of 104.64. This suggests potential, yet it does not surpass the individually fine-tuned XGBoost model in predictive accuracy. The second ensemble model, incorporating XGBoost and LSTM, achieves an R-squared value of 81.10%, an MAE of 107.2, and an RMSE of 141.03. This indicates that blending the strengths of XGBoost and LSTM may lead to improved predictive performance compared to the first ensemble model. Nonetheless, it still falls short of the individually fine-tuned XGBoost model's accuracy. The study underscores the significance of algorithm selection and hyperparameter refinement for precise stock price predictions.

While this study contributes valuable insights into stock price prediction, it is important to acknowledge certain limitations that may impact the interpretation and generalization of the findings. One limitation lies in the reliance on historical price data as the primary input feature. This approach overlooks the potential influence of external factors, such as geopolitical events or market sentiment, which could significantly impact stock prices but are not considered in the current model. Additionally, the study focuses on a single stock and does not account for potential variations in predictive performance across different stocks or market sectors.

7. CONCLUSION

In conclusion, this project addresses its objectives by demonstrating that univariate algorithms can generate predictions close to actual values, minimizing errors and providing reliable insights. Additionally, the project develops computationally efficient models while considering the identification of key variables for multivariate modeling.

This study explores stock price prediction using a diverse range of machine learning algorithms and ensemble techniques. Through meticulous evaluation and analysis, it becomes evident that the XGBoost algorithm emerges as a standout performer, showcasing remarkable predictive capabilities. Hyperparameter tuning further enhances the accuracy of the XGBoost model, underscoring the importance of careful parameter optimization in machine learning models to enhance their performance and predictive capabilities in financial forecasting tasks. Although the study introduces two ensemble configurations, each with its distinct combination of algorithms, they were not as superior as the tuned model.

While ensemble approaches hold promise, their superiority over meticulously tuned individual algorithms is not universal. As such, the study encourages practitioners to meticulously assess the trade-offs between complexity and performance, making informed decisions based on specific forecasting objectives and constraints.

FUTURE OUTLOOK

Future research endeavors could explore a deeper exploration of ensemble methods could be pursued, focusing on novel combinations of algorithms and innovative techniques to capitalize on their collective predictive power. Investigating the integration of deep learning architectures, such as attention mechanisms or transformer-based models, may uncover new dimensions of accuracy and interpretability. Additionally, a more comprehensive analysis of feature engineering could be undertaken to identify and incorporate relevant variables, potentially enriching the input space for more precise predictions. Moreover, the study could extend its scope to encompass the prediction of stock price volatility or explore the dynamics of intraday trading patterns. The incorporation of alternative data sources, such as sentiment analysis of news articles or macroeconomic indicators, could offer valuable insights for refining predictive models.

ACKNOLEGEMENTS

First and foremost, I humbly acknowledge the divine guidance and blessings of God, which have been an unwavering source of strength and inspiration throughout this journey.

I extend my deepest appreciation to my supervisor, Dr. Yang Li, for his invaluable guidance, constant support, and mentorship throughout this dissertation journey. Your expertise and guidance have played a critical role in shaping this work.

I am also thankful to my other instructors, Mr. Joseph Annan, Mr. Kevin, and Miss Candace, for providing constant guidance and assistance.

To my family and friends, I owe a profound debt of gratitude for their constant encouragement, firm support, and boundless motivation. This dissertation would not have been possible without the collective assistance and encouragement of these remarkable individuals.

REFERENCES

1. B. Ma, Y. Yang, J. Zhang, and K. Zhang, 'A Comparison of Stock Price Prediction with ANN and ARIMA', 2023.
2. J. Zhang, L. Ye, and Y. Lai, 'Stock Price Prediction Using CNN-BiLSTM-Attention Model', *Mathematics*, vol. 11, no. 9, May 2023, doi: 10.3390/math11091985.
3. L. Peng, 'Stock Price Prediction of "Google" based on Machine Learning', 2022.
4. Q. A. Al-Radaideh, E. Alnagi, and A. A. Assaf, 'Predicting Stock Prices Using Data Mining Techniques Individual Project View project Reduct Computation View project The International Arab Conference on Information Technology (ACIT'2013) PREDICTING STOCK PRICES USING DATA MINING TECHNIQUES', 2013. [Online]. Available: https://www.researchgate.net/publication/281865047
5. Q. Yan, 'The Stock Price Analysis of Netflix Prediction', 2022.

6. X. Li, Y. Li, H. Yang, L. Yang, and X.-Y. Liu, 'DP-LSTM: Differential Privacy-inspired LSTM for Stock Prediction Using Financial News', Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.10806
7. Y.-G. Song, Y.-L. Zhou, and R.-J. Han, 'Neural networks for stock price prediction', May 2018, [Online]. Available: http://arxiv.org/abs/1805.11317
8. S. R. Riady, 'Stock Price Prediction using Prophet Facebook Algorithm for BBCA and TLKM', *International Journal of Advances in Data and Information Systems*, vol. 4, no. 2, Apr. 2023, doi: 10.25008/ijadis.v4i2.1258.
9. K. Hiba Sadia, A. Sharma, A. Paul, and S. Sanyal, 'Stock Market Prediction Using Machine Learning Algorithms', *Int J Eng Adv Technol*, 2019, [Online]. Available: www.ijeat.org
10. B. Wamkaya, 'ANN Model to Predict Stock Prices at Stock Exchange Markets', 2017.
11. M. Roondiwala, H. Patel, and S. Varma, 'Predicting Stock Prices Using LSTM', *Article in International Journal of Science and Research*, vol. 6, 2017, doi: 10.21275/ART20172755.
12. Y. Deshmukh, D. Saratkar, and Y. Tiwari, 'Stock Market Prediction Using Machine Learning', *IJARCCE*, vol. 8, no. 1, pp. 31–35, Jan. 2019, doi: 10.17148/IJARCCE.2019.8107.
13. Wikipedia, 'Google'.
14. Yahoo Finance, 'Google current stock price'.
15. Netcials, 'Alphabet Inc Stock 10 Year History'.
16. Prof. Sulochana Sonkamble, Vaibhav Vyas, Prathamesh Shimpi, Aniket Mule, and Mihir Sonawane, 'Stock Price Prediction System', *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 273–277, Apr. 2023, doi: 10.32628/CSEIT2390229.
17. A. Jeevan Kumar, O. Gangadhar Reddy, B. Sai Kumar Reddy, B. Sandeep Kumar, and A. Sri Hari, 'EasyChair Preprint Analysis and Forecasting of Stock Market Using Time Series Algorithm ANALYSIS AND FORECASTING OF STOCK MARKET USING TIME SERIES ALGORITHM ABSTRACT', 2023.
18. L.-P. Chen, 'Using Machine Learning Algorithms on Prediction of Stock Price', *Journal of Modeling and Optimization*, vol. 12, no. 2, pp. 84–99, Dec. 2020, doi: 10.32732/jmo.2020.12.2.84.
19. D. S. N. Ulum and A. S. Girsang, 'Hyperparameter Optimization of Long-Short Term Memory using Symbiotic Organism Search for Stock Prediction', *International Journal of Innovative Research and Scientific Studies*, vol. 5, no. 2, pp. 121–133, 2022, doi: 10.53894/ijirss.v5i2.415.
20. L. Han, 'Analysis of Stock Price and Price Movement Prediction based on Machine Learning Models for E-Hualu', 2023.
21. Ishitva Upadhyay, Devashish Katiyar, and Dr. Vasudha Vashisht, 'TECHNICAL ANALYSIS OF STOCK MARKET & PREDICTION USING DATA SCIENCE', *International Research Journal of Modernization in Engineering Technology and Science*, May 2023, doi: 10.56726/IRJMETS39976.
22. H. C. Lin, C. Chen, G. F. Huang, and A. Jafari, 'Stock Price Prediction using Generative Adversarial Networks', *Journal of Computer Science*, vol. 17, no. 3, pp. 188–196, 2021, doi: 10.3844/JCSSP.2021.188.196.
23. A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, 'Stock price prediction using the ARIMA model', in *Proceedings - UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014*, Institute of Electrical and Electronics Engineers Inc., 2014, pp. 106–112. doi: 10.1109/UKSim.2014.67.
24. Y. Shao, 'Prediction of Moderna Adjusted Closing Stock Price Trend Using ARIMA Model', 2023.
25. Y. Zhang, 'Stock Price Prediction Method Based on XGboost Algorithm', 2023, pp. 595–603. doi: 10.2991/978-94-6463-030-5_60.
26. R. Ruhal and E. Vandana Prashar, 'A Comparative Study Of Statistical Methods And Machine Learning Approaches For Stock Price Prediction', vol. 10, 2023, doi: 10.13140/RG.2.2.19210.44483.
27. E. Ismanto, 'LSTM Network Hyperparameter Optimization for Stock Price Prediction Using the Optuna Framework', *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 1, pp. 22–35, 2023, doi: 10.26555/jiteki.v9i1.24944.
28. G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, and S. K. Bhat, 'Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications', *International Journal of Financial Studies*, vol. 11, no. 3, p. 94, Jul. 2023, doi: 10.3390/ijfs11030094.

APPENDIX

A. Dataset

Google.csv

B. Code

Click here