Predicting Stock Closing Price

Liew Kangze, Ryan Chong Junyi, Tai Eugene, Te Ming Xian

Introduction

Stock market prices are often volatile and hard to predict with many factors affecting different stocks' closing prices. However, through this project, we aim to gain insights on potential factors that can influence the stock closing prices and analyze different AI and machine learning techniques to compare their prediction results. Gaining a better understanding of the influences of different factors on stock market prices can allow companies to better plan growth strategies and for economists to prevent a financial crisis or spot potential signs of it. Time series data refer to data that are collected over different periods of time, which is a characteristic of the stock price data. Hence, we decided to use Multiple Linear Regression (MLR), Random Forest (RF) and Long Short-term Memory (LSTM) methodologies in analyzing the stock closing price data.

Our research took ideas from the work of Antionio Rafael Sabino Parmezan, Vinicius M.A. Souza, and Gustavo E.A.P.A Batista in their paper titled "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model." A limitation in their research is that they confined their analysis to univariate data predictors, and expressed an interest in exploring the multivariate domain which is still insufficiently addressed in the literature. To make up for this identified gap, our study uses multivariate data for model training to overcome this limitation. Specifically, we utilize available dataset features, such as trading volume, and augment our model with lagged features to enrich the dataset and enhance the predictive capability of our model.

Our research was also influenced by the study "Applying machine learning algorithms to predict the stock price trend in the stock market - The case of Vietnam" conducted by Tran Phuoc, Pham Thi Kim Anh, Phan Huy Tam, and Chien V. Nguyen. This study acknowledges the efficacy of the LSTM model in making accurate predictions, yet it also expresses a desire to explore additional machine learning algorithms that have shown considerable development and are widely used in the finance sector, such as Random Forest and Support Vector Machine (Phuoc et al., 2024). Drawing inspiration from this paper, we incorporated the Random Forest algorithm into our research to complement the LSTM model and address the mentioned limitation.

Another recent study titled 'Stock closing price using machine learning techniques' by Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal and Arun Kumar published in Procedia Computer Science made use of Artificial Neural Network (ANN) and RF to do its analysis. (Vijh et al., 2020) While ANN produced better prediction results than RF, their analysis focused mainly on analyzing how the moving average of closing prices affected the stock price and made use of 6 inputs only whereas our dataset has 22 features, significantly more than theirs.

The recent works done on predicting stock market prices have used LSTM, RF, ANN and Principal Component Analysis (PCA). PCA was not suitable for our data set due to the sparse input variables, hence the covariance matrix computed is dominated by zero values, rendering the method useless for our data. While ANN has shown positive prediction results, the high number of input variables for our dataset propels us to use LSTM instead to utilize the forget gate for filtering out variables of less importance. The selection of MLR serves as a base prediction method to compare LSTM and RF prediction results to, assuming a linear relation between stock prices and the input variables

Dataset

For this project, the main dataset consisting of stock price, trade volume, news events and news sentiment for S&P 500 companies was used to train and test the model. The dataset contains 217811 samples in total, with 22 features per sample from the time period October 2020 to July 2022. The dataset contains categorical, discrete and continuous attributes, alongside missing values. An additional dataset containing a list of S&P 500 companies and different metadata such as headquarter location, date of entry into S&P 500 index, central index key, year company was founded and wikidata id was also provided. This additional dataset of 524 rows and 10 columns was given to supplement the main dataset as an assisting feature.

One of the main issues with the dataset was the presence of missing values. To address this, missing values were imputed with zeros using the fillna() function. Additionally, categorical variables such as 'Symbol', 'GICS Sector', and 'GICS Sub-Industry' were encoded using label encoding to convert them into numerical representations suitable for machine learning models.

To capture the temporal dependencies and trends in the stock prices, lagged features were generated for the past 5 days (for the first data frame) for the 'Close', 'Volume' and news sentiment columns that showed some correlation with our target feature “Close”. The dataset starts from 30/9/2020, however we use 7/10/2020 as the training dataset starting date as it is the sixth entry of earliest date. By excluding the first five days of the dataset, we eliminate the predictions that have NaN values for their lag values.

For the second data frame moving averages, such as the 5-day simple moving average (SMA) and additional moving averages (7-day, 14-day, and 21-day) and the 7-day standard deviation were computed to smooth out short-term fluctuations and identify trends. Reason for separating lagged features and moving averages into separate data frames is to prevent information leakage and look-ahead bias when the two types of features are trained together in a model. Data points with NaN values were then excluded (first 5 dates in dataframe 1 and first 21 days in dataframe 2) due to their low proportion to the entire dataset.

The dataset was also analyzed for seasonality by extracting the year, month, and day from the date column. This information can be useful for capturing any seasonal patterns or variations in the stock prices

To evaluate the model's performance and generalization ability, the dataset was split into training and testing sets based on a specific date. The split was performed such that the training set consisted of data before the split date

The testing set contained data from the split date onwards. The proportion of training and testing data is approximately 80:20.

Dataset	Training Dataset 1	Test Dataset 1
30/09/2020 - 30/06/2022	7/10/2020 - 23/02/2022	24/02/2022 - 30/06/2022

Table 1: Statistic of dataset 1

Dataset	Training Dataset 2	Test Dataset 2
30/09/2020 - 30/06/2022	28/10/2020 - 28/02/2022	01/03/2022 - 30/06/2022

Table 2: Statistic of dataset 2

This ensures that the model is trained on historical data and tested on unseen future data, mimicking a real-world scenario.

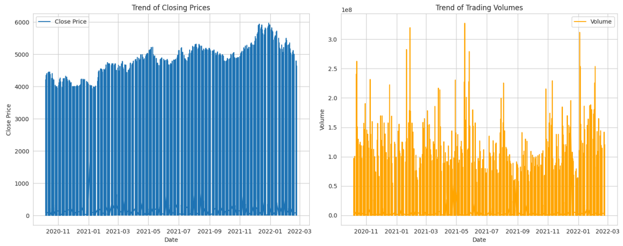

Figure 1: Trend of closing prices and trading volumes

Close price and trading volume was then plotted against date to determine the general trend of close prices between the specific time period.

Methodology

Predicting closing stock prices is a time series forecasting task where historical data is used to predict future price movements. Various algorithms can be employed for this purpose, each with its own strengths and weaknesses.

MLR is a linear model used to establish multiple input features and the target variable, estimating the relationship between independent variables and the dependent variable by fitting a linear equation to the data points. RF is an ensemble learning algorithm that builds multiple decision trees on different random samples of the dataset and aggregates their predictions. Decision trees are highly sensitive to training data which can result in high variance and fail to generalize the data. RF uses random sampling of the data known as bootstrapping and generates decision trees from each bootstrapped data with a random selection of features. Predictions are then made through aggregating the predictions of all the random decision trees formed. RF involves the process of bootstrapping and aggregating is also known as bagging, where bootstrapping results in decreased sensitivity to data and aggregating different decision trees results in reduced correlation between features, decreasing variance. LSTM networks are a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data like time series. LSTMs have memory cells that allow them to remember information over extended time periods, making them well-suited for modeling complex temporal patterns and forecasting future values in time series data.

Feature Selection Process

First we plotted the correlation matrix between all the features and identified inputs with higher correlation to the 'Close' feature. From there, we chose features that had a correlation of at least 0.06 as inputs for our MLR model.

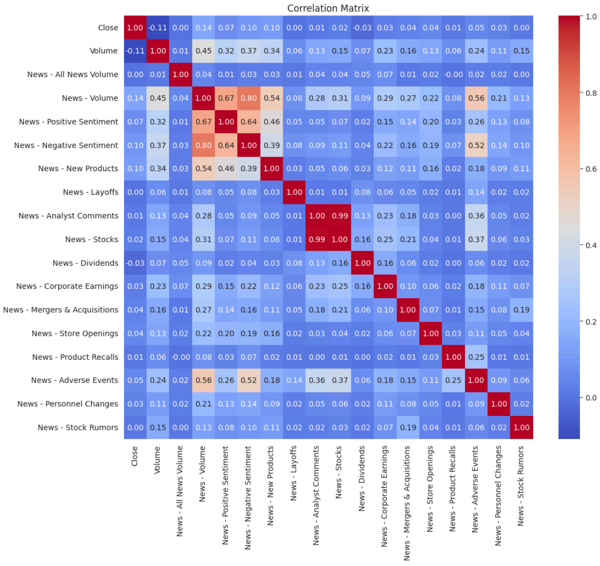

Figure 2: Correlation matrix of features

After including moving averages in our second model, we plotted another correlation matrix, showing the high correlation values between the moving averages and the closing prices.

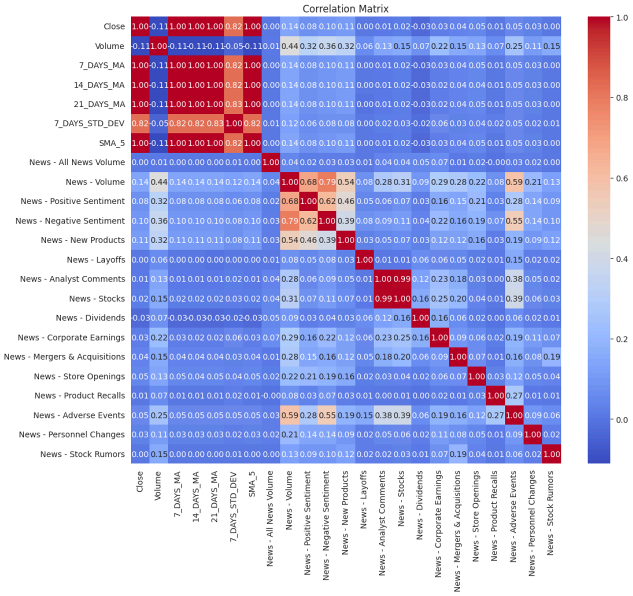

Figure 3: Correlation matrix with moving averages

Additionally, the decision trees in RF extracts out the most discriminative features in each node, hence plotting the RF feature importance plot helps to extract out the important features used in RF for the prediction of closing price.

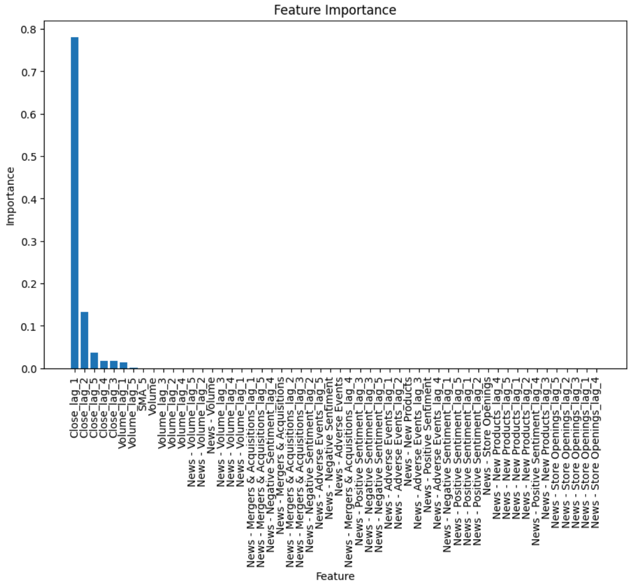

Figure 4: Feature Importance Plot

From figure 4, we can see that the closing price for the previous days have the highest importance in the prediction. Hence, we angled our predictions to use the different moving average of past closing prices for the prediction.

LSTM Model

To capture the temporal dependencies and trends in the stock price data, 2 LSTM models were employed. Multiple experiments were ran using different numbers of LSTM and dropout layers, resulting in the final decision of using the following model infrastructure:

1. LSTM Layer: The first layer is an LSTM layer with 32 units, which learns the long-term dependencies in the input sequence. The return_sequences parameter is set to True to output a sequence of hidden states.
2. Dropout Layer: A dropout layer with a rate of 0.25 is added to prevent overfitting by randomly dropping out a fraction of the units during training.
3. Dense Layer: The final layer is a dense layer with a single unit, which produces the predicted closing price.

The LSTM model takes a sequence of input features (lagged features for first model, moving averages for second model) for a specified number of time steps (n_input = 1) and predicts the closing price for the next time step. The models are trained using the Adam optimizer and mean squared error (MSE) as the loss function.

Results and Discussions

Prediction results of Disney

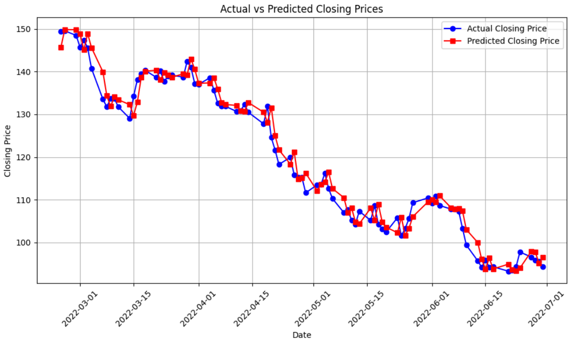

Figure 5a: MLR(1) Disney prediction results

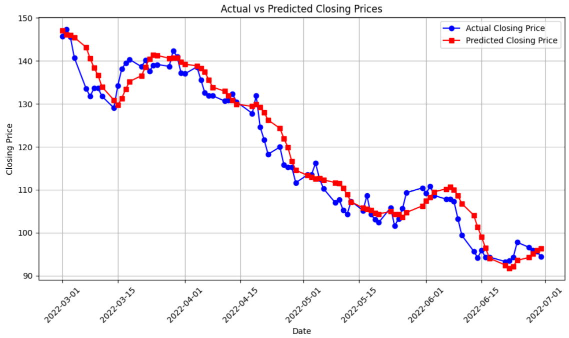

Figure 5b: MLR(2) Disney prediction results

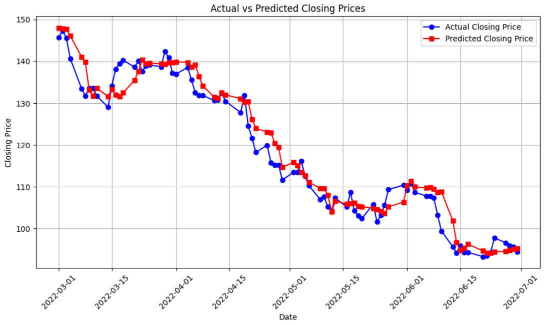

Figure 5c: RF Disney prediction results

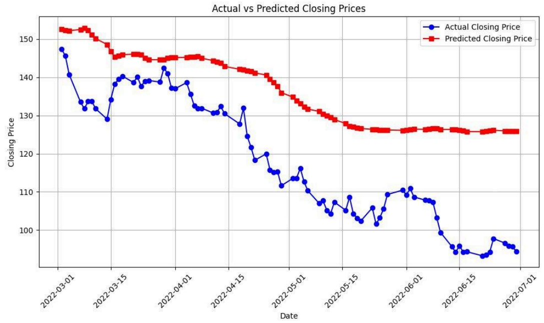

Figure 5d: LSTM Disney prediction results

	MLR(1)	MLR(2)	RF	LSTM
RSME	2.709	3.666	3.435	22.268

Table 3: RSME of disney predictions

Evaluation of ML/AI methods

To evaluate the different models, we make use of root mean-square error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE). RMSE provides the difference between actual and predicted closing prices, MAE measures the average absolute difference between actual and predicted closing prices and MAPE measures the average absolute percentage difference between actual and predicted closing prices.

MAPE $=1/n\ \Sigma((y_{actual}-y_{pred})/y_{actual})$ x100%

MAE $= 1/n\ \Sigma(y_{actual}-y_{pred})$

Performance Metrics	RMSE	MAPE	MAE
MLR(1)	6.7315	1.3880%	2.5197
MLR(2)	8.6997	1.7216%	3.2560
RF	15.4263	1.9892%	4.5588
LSTM	328.1649	146.2556%	137.2667

Table 4: Performance of different ML/AI methods

Comparing our results for the different methods used, the first MLR produced better prediction results than the second. This can be attributed to the fact that the first MLR makes use of only past 5 days closing prices for its prediction, hence a closer date range as compared to the second MLR which uses inputs of 7, 14 and 21 days moving average. With a relatively small time frame of only 2 years, using a larger date range for prediction might not be able to capture the trends accurately and the changes in the moving average might not be as significant. Since our aim is to predict the next day's closing price, the first MLR which uses shorter time lags can provide a better insight and more accurate prediction.

Comparing RF to MLR, RF produced worse prediction results than MLR with higher RMSE, MAPE and MAE. Given our relatively small dataset, it is expected for RF to perform poorer since RF requires large datasets to extract deep underlying patterns in the data to make predictions. Additionally, RF performs poorly when there are many categorical features with many levels since it can lead to 'noise' in model results (Samarth, 2023), which applies to our data set as well.

LSTM produced the worst prediction results amongst the methods we used. One explanation could be the small dataset which results in poor generalization of the model. Alternatively, it is suggested to use at least 10 epochs (Rokhsatyazdi et al., 2020), but due to long computational time per epoch, it was not feasible for us to keep testing and varying multiple epoch numbers, which could explain the poor model prediction.

Conclusion

Stock market closing prices are highly volatile and hard to produce accurate predictions. While our ML and AI methods might not have produced accurate prediction results, it gives us an insight into what are the possible factors that can contribute to a stock's closing price and possible relations between certain features. From our correlation matrix, it can be seen that not all news results have correlations to the closing prices and from the feature importance plot produced by RF, the impact of the different news on closing prices are negligible in comparison to the past days closing prices. However, most studies published on stock market prediction have used 10 year data trends to analyze and train their models. Given that our data only has a 2 year time frame, the results produced from this study might pale in comparison since a longer time frame and larger data set allows for deeper analysis of long term trends which can produce more accurate predictions. Hence, if provided a longer term data set to do the analysis, the LSTM and RF models are likely to produce better prediction results. While our models cannot predict accurate closing prices, it gives a good indication of the general trend of prices and predicting the direction of movement of the price.

References

Parmezan, A. R. S., Souza, V. M. A., & Batista, G. E. A. P. A. (2019, January 30). Evaluation of statistical and Machine Learning Models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. Information Sciences. https://www.sciencedirect.com/science/article/abs/pii/S0020025519300945

Phuoc, T., Thi Kim Anh, P., Huy Tam, P., & V. Nguyen, C. (2024, March 12). Applying machine learning algorithms to predict the stock price trend in the stock market – the case of Vietnam. Nature. https://www.nature.com/articles/s41599-024-02807-x

Rokhsatyazdi, E., Rahnamayan, S., Amirinia, H., & Ahmed, S. (2020, September 3). Optimizing LSTM Based Network For Forecasting Stock Market. https://ieeexplore-ieee-org.libproxy1.nus.edu.sg/document/9185545.

Samarth, V. (2023, December 14). *What is Random Forest In Data Science and How Does it Work?* Emeritus India. https://emeritus.org/in/learn/data-science-random-forest/

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020, April 16). Stock closing price prediction using Machine Learning Techniques. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050920307924