

Stock Price Prediction Via Improved Machine Learning Techniques*

Wickramasinghe J.A.D.L.
239187E

Dilhan M.W.S.
239155F

Wijesena C.M.C.H.
239188H

Samarasinghe T. A. A.
239173H

I. INTRODUCTION

The world's economy is currently facing a financial crisis due to the Covid-19 pandemic. This situation has made investors lose money because they can't predict what will happen to stocks. Financial markets have a big impact on economies, and many companies rely on them to make money. Therefore, it's important to study how financial markets behave and perform. This research involves predicting things like stock prices, exchange rates, and trade volumes. In this study, we use machine learning to predict stock prices, which can help investors feel more confident.

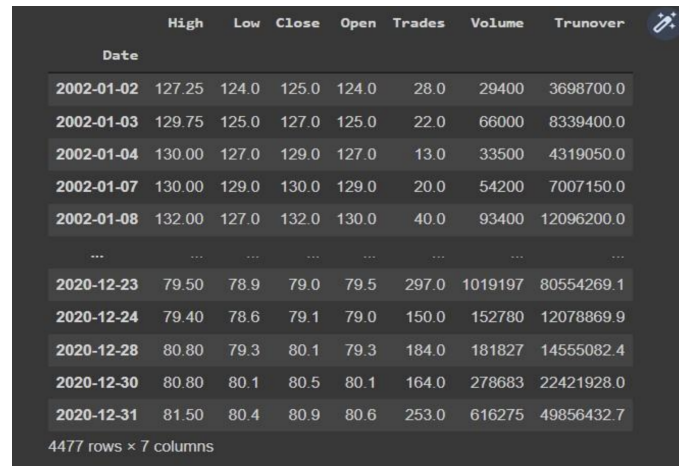
When the stock market goes up, it usually means that companies are doing well, and when it goes down, it usually means the economy is not doing well. Predicting the stock market is important because it can help investors make better decisions. However, predicting the stock market is challenging because it is affected by many things like politics, natural disasters, and the stock market itself. To predict stock prices, we need to use statistical methods that analyze past data. This can help companies make better decisions about their products.

This project aims to create a model that can predict the closing price of stocks in the Sri Lankan stock market. This model will use past stock price data and the All Share Price Index to predict future stock prices. We will use a machine learning technique called LSTM Recurrent Neural Network to make predictions. The model can help investors feel more confident about investing and attract more local and foreign investors to the stock market. This could have a positive impact on the Sri Lankan economy. The main goal of the project is to create a machine learning model that can predict stock prices and offer several benefits, such as reducing analyzing time and contributing to the country's economy.

II. DATA DESCRIPTION

To do this research project, we collected information from the Colombo stock exchange (CSE). We used different data like the closing, opening, high, low, trades, volume, and turnover for each company as a historical dataset. There were other data like All Share Price Index (ASPI), Sector indices, foreign holdings, foreign buying and dividends payments. We collected data on the Commercial bank (COMB.N) from 2002 to 2020 to build a model that can predict the stock market closing price in the Sri Lankan stock market. We also plan to

use tweets throughout the project to improve our predictions as a next step.



Date	High	Low	Close	Open	Trades	Volume	Turnover
2002-01-02	127.25	124.0	125.0	124.0	28.0	29400	3698700.0
2002-01-03	129.75	125.0	127.0	125.0	22.0	66000	8339400.0
2002-01-04	130.00	127.0	129.0	127.0	13.0	33500	4319050.0
2002-01-07	130.00	129.0	130.0	129.0	20.0	54200	7007150.0
2002-01-08	132.00	127.0	132.0	130.0	40.0	93400	12096200.0
...
2020-12-23	79.50	78.9	79.0	79.5	297.0	1019197	80554269.1
2020-12-24	79.40	78.6	79.1	79.0	150.0	152780	12078869.9
2020-12-28	80.80	79.3	80.1	79.3	184.0	181827	14555082.4
2020-12-30	80.80	80.1	80.5	80.1	164.0	278683	22421928.0
2020-12-31	81.50	80.4	80.9	80.6	253.0	616275	49856432.7

Fig. 1. Dataset

The following table describe the dataset description of the Commercial Bank dataset despite above.

TABLE I
DATASET DESCRIPTION

Field name	Description
Date	The date and time when the stock market data was collected.
High	The highest price at which the stock was traded on that day.
Low	The lowest price at which the stock was traded on that day.
Close	The closing price of the stock on that day.
Open	The opening price of the stock on that day.
Trades	The total number of trades executed on that day.
Volume	The total volume of shares traded on that day.
Turnover	The total value of shares traded on that day.

A. Candlestick chart

This chart is used to represent the daily price movements of the stock. It shows the high, low, open and closing prices for each day in a rectangular bar, with the color of the bar indicating whether the stock price increased or decreased that day.



Fig. 2. Candlestick Plot

B. Time series plot of Closing Prices

This graph shows the trend of the closing prices of the bank stock over time. The graph displays data points in chronological order. In this case, the X-axis represent the date or time period, while the Y-axis represent the changing price of the commercial bank stock.

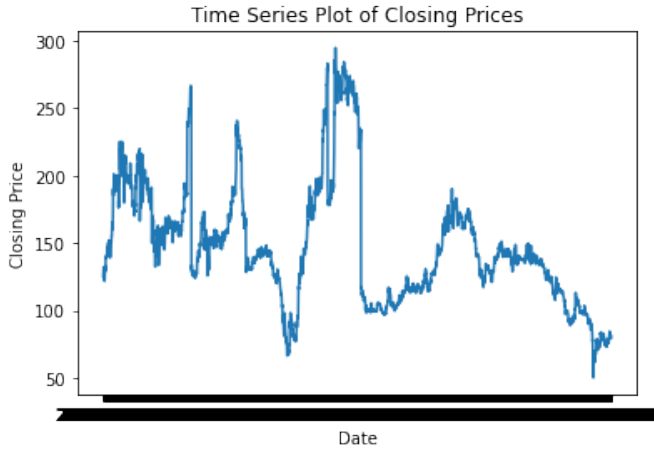


Fig. 3. Time Series Plot

C. Box plot

This box plot is a graphical representation of a dataset that shows the median, quartiles, and outliers of the data. In this case, the X-axis represent the year, while the Y-axis represent the changing price of the bank stock. This plot shows the distribution of closing prices for each year.

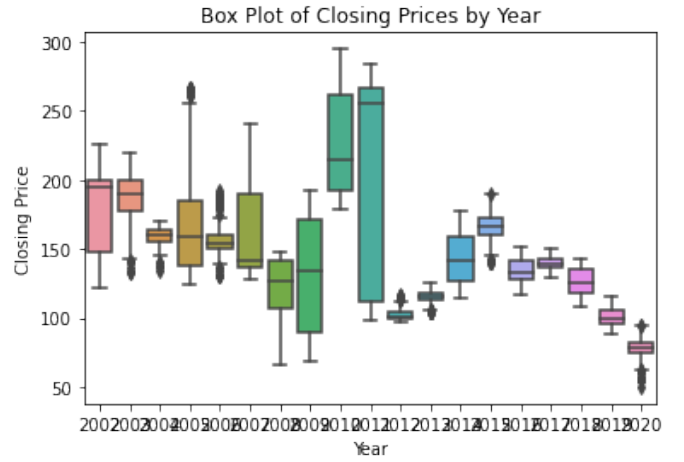


Fig. 4. Box Plot

III. PROPOSED METHOD

Investors who invest in the Colombo Stock Exchange and other stock markets struggle to predict the future prices of shares. This problem has been made worse by the global financial crisis caused by COVID-19, which has led to many investors losing confidence in their investments. To analyze the behavior and performance of stock markets, researchers predict the prices of securities such as stocks, bonds, exchange rates, market indicators, securities benefits, trade volumes, and inventory classification.

Predicting stock prices is challenging because the stock market is unpredictable, and there is a lot of data to analyze. Factors such as economic indicators, news events, and market sentiment affect stock prices and make it hard to forecast future values. Moreover, some machine learning algorithms cannot model non-linear relationships found in stock market data. To tackle these challenges, researchers need to preprocess the data carefully and choose and evaluate the most suitable machine learning techniques.

This project aims to create a machine learning model that predicts stock closing prices and boosts investor confidence by using the right algorithm and considering market sentiments. The model's accuracy will be assessed using metrics such as R Square/Adjusted R Square, Mean Square Error (MSE)/Root Mean Square Error (RMSE), and Mean Absolute Error (MAE).

IV. EXPERIMENTS

To predict stock prices, we tested different algorithms to see which one worked best. We looked at Linear Regression, Support Vector Regression (SVR), Random Forest Regression, XGBoost Regression, Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), and Gradient Boosting Regression.

Linear Regression is a straightforward algorithm that fits a line to the data. Support Vector Regression (SVR) is similar, but can handle more complex relationships. Random Forest Regression combines many decision trees to make a

prediction, which is helpful for data with lots of features. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) are types of neural networks that are good at handling time-series data. Gradient Boosting Regression combines weak models to make a strong one.

We used different metrics to evaluate the algorithms, such as R Square/Adjusted R Square, Mean Square Error (MSE)/Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The results are shown in a table that compares how well the six algorithms worked using the four metrics for the regression task.

TABLE II
MODEL COMPARISON

Algorithm	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	R-squared (R2) score
Linear Regression	576.36	24.01	19.37	0.30
Support Vector Regression	23543.74	153.44	147.15	-10.92
Random Forest Regression	2897.34	53.83	46.26	-2.51
XGBoost Regression	1583.13	39.79	31.91	-1.72
Recurrent Neural Networks	13189.95	114.85	112.60	-1182878.85
Long Short Term Memory	14.19	3.77	2.92	0.98

The MSE measures the average squared differences between the predicted and actual values. The RMSE is the square root of the MSE, which provides the average distance between the predicted and actual values. The MAE measures the average absolute differences between the predicted and actual values. The R2 score measures how well the model fits the data, with a score of 1 indicating a perfect fit and a score of 0 indicating no fit at all.

From the table, it can be observed that Long Short Term Memory (LSTM) performs the best among all the models as it has the lowest MSE, RMSE, and MAE and the highest R2 score. It indicates that the LSTM model can predict the target variable more accurately and fit the data better than the other models.

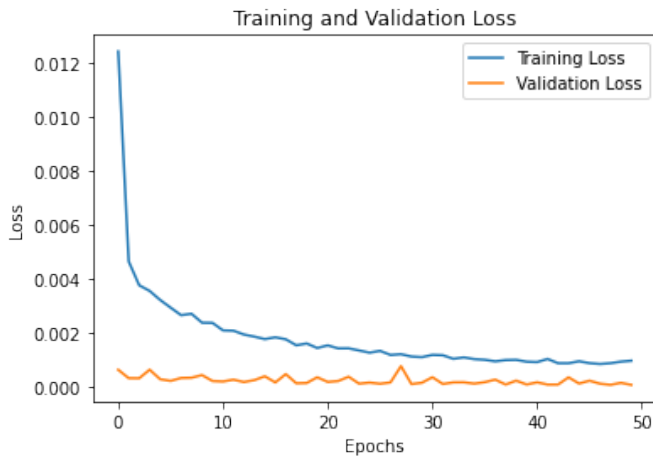


Fig. 5. RNN - Training and Validation Loss

On the other hand, Support Vector Regression (SVR), Random Forest Regression, and XGBoost Regression have higher errors and lower R2 scores compared to the other models,

indicating poor performance in predicting the target variable. Among these three models, XGBoost Regression performs relatively better than the others.

Linear Regression and Recurrent Neural Networks (RNN) perform moderately well, with RNN having slightly lower errors and higher R2 score than Linear Regression. However, RNN has a significantly negative R2 score, indicating a poor fit of the model to the data.

In conclusion, the comparison table provides a quick overview of the performance of different algorithms using multiple evaluation metrics. Based on the results, it is evident that LSTM outperforms all the other models, while SVR, Random Forest Regression, and XGBoost Regression perform poorly. Linear Regression and RNN have moderate performance, with RNN performing slightly better than Linear Regression.

V. PROGRESS AND NEXT STEPS

Next, we plan to find the best model for stock prediction by adjusting hyperparameters. Additionally, we want to examine the impact of Twitter news on stock prices. We will use Natural Language Processing (NLP) techniques to analyze the sentiment of tweets and use the data to predict future stock prices. This approach can capture the sentiment of investors towards news and events that may affect the stock market, like an expected rise in interest rates.

Using Twitter data for stock prediction has potential benefits, as it can provide unique information that may not be available from other sources. However, there are also challenges associated with using Twitter data, such as noise, bias, and the need for careful data preprocessing and feature engineering.

In conclusion, incorporating sentiment data from Twitter news into the stock prediction model is a promising and interesting direction for future research. This approach could provide valuable insights into the complex relationships between news, sentiment, and stock prices.

VI. DIVISION OF WORK

The table below outlines the division of work among the members of our group.

TABLE III
CONTRIBUTION

Index No	Name	Contribution
239187E	Wickramasinghe J.A.D.L.	Dataset Description, Experiments
239155F	Dilhan M.W.S.	Progress and Next Step
239188H	Wijesena C.M.C.H.	Proposed Method
239173H	Samarasinghe T. A. A.	Introduction

REFERENCES

- [1] K. J. Kesavan M, "Stock Market Prediction with Historical Time Series Data and Sentimental Analysis of Social Media Data," 2020.
- [2] T. F. A. Samarawickrama, "A Recurrent Neural Network Approach in Predicting Daily Stock Prices, University of Sri Jayewardenepura, 2018," 2018.
- [3] M. H. M. H. Adil MOGHARa, "Stock Market Prediction Using LSTM Recurrent Neural Network," 2020.

- [4] Z. D. Zhao J, "Prediction model for stock price trend based on recurrent neural network," *Journal of Ambient Intelligence and Humanized Computing*, 2021.
- [5] V. B. Priyank Sonkiya, "Stock price prediction using BERT and GAN," 2020.
- [6] G. KangZhang, "Stock Market Prediction Based on Generative Adversarial Network," 2019.
- [7] R. A. Md.Tanvir Rahman, "Forecasting Stock Market Price Using Multiple," 2018.