**STOCK PRICE PREDICTION USING MACHINE LEARNING ALGORITHMS AND STREAMLIT**

**COM 624(MACHINE LEARNING)**

ASSESSMENT REPORT

BY

DAVID AGBOLADE

STUDENT NO.: Q14872382

BSC(Hons) SOFTWARE ENGINEERING

**INTRODUCTION**

Stock price prediction is the technique of predicting future stock prices using historical data and certain mathematical models. Numerous methods, including fundamental analysis, technical analysis, and machine learning algorithms, can be used to do this.

In fundamental analysis, stock prices are forecasted by assessing a company's financial performance and state. This entails taking a close look at elements like earnings, dividends, and the status of the economy in general.

On the other side, the technical analysis makes predictions by using previous price and volume data to spot patterns and trends. Moving averages, patterns on charts, and other indicators are examples of this.

Making predictions about stock prices can also be done using machine learning techniques like neural networks. These algorithms may learn to recognise patterns and make predictions based on past data after being trained on vast volumes of historical data.

In this project, technical analysis and machine learning algorithms like linear regression and Long Short-Term Memory (LSTM) would be used to predict the trends of the stock market. The objective is to predict the stock prices to make more informed and accurate investment decisions.

**PROBLEM STATEMENT**

SOLFINTECH customers are interested in learning about the market's future state to make profitable investments. By giving traders enabling information like the direction of the market in the future, effective prediction systems indirectly assist traders.

In this project, we'll investigate the possibility of developing a machine-learning model that can forecast stock prices with a lower percentage of inaccuracy. We will also examine how to get the next day’s price and tell if it is a rise or fall in the stock market.

For this project, I will use both the Linear regression model and Long Short-Term Memory Model.

Goals:

* Explore stock prices and observe the trends.
* Implement a Linear regression model with an accuracy of at least 90%
* Implement the LSTM model by using the required libraries with an accuracy of at least 90%.
* Get the next-day price.
* Create a web application using streamlit for easy access.

**COLLECTION OF DATA**

From January 1, 2010, to date, Apple Inc. data was used for this assessment. This is a time series, or collection of data points, indexed by time. After training, our objective was to predict the closing price.

All data used were obtained in real time via the Yahoo Finance Python data-reader package to facilitate replication and reuse. For the closing (adjusted closing) price of the data, a prediction must be made. Considering that Yahoo Finance has already adjusted the closing pricing for us, all we need to do is estimate the "CLOSE" price.

Running the code below will load the data into a new Data-frame object.

Text

Description automatically generated

So I had a bit of an issue with the code above after I ran it again after some weeks, the error was string indices must be integers in data.DataReader, then I fixed it by using yf.pdr\_override() from the pandas DataReader library. Updated code below

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

This dataset reveals that the date, high, and low values are not significant characteristics of the data. Because it is irrelevant at what levels the stock traded at its highest or lowest prices in each day. The stock's opening and closing prices are what are significant.

At the conclusion of the day, if the closing prices are higher than the opening prices, we will have made money; if not, we will have lost money. In contrast, an increasing price and a decreasing volume suggest a lack of interest and serve as a signal of a likely reversal. A market that is expanding should also see rising share volume. The stock's fundamentals are more strongly indicated by a price decrease (or gain) in high volume.

**EXPLORING THE DATA**

Before exploring the data, had to make sure the dataset has no missing values by using the ISNA () function and then get the info and the shape of the dataset using the code below.

Graphical user interface, text

Description automatically generated

This shows that there is no missing data in our dataset and the data types are all in float which is very good for the data. Using a describe function to generate the summary statistics of the datasets.

Graphical user interface, text, application

Description automatically generated

Observations:

* we could tell the minimum closing price is 6.85 in the dataset and the maximum closing price is 182.09.
* it has 3258 rows/data with 6 columns.
* The mean average is 50.96.

Applying the multivariate analysis to explore the data is a fantastic option for getting a rapid overview of the dataset. Running the code below creates line graphs for each column in our Data-Frame.Text

Description automatically generated

Chart, histogram

Description automatically generated

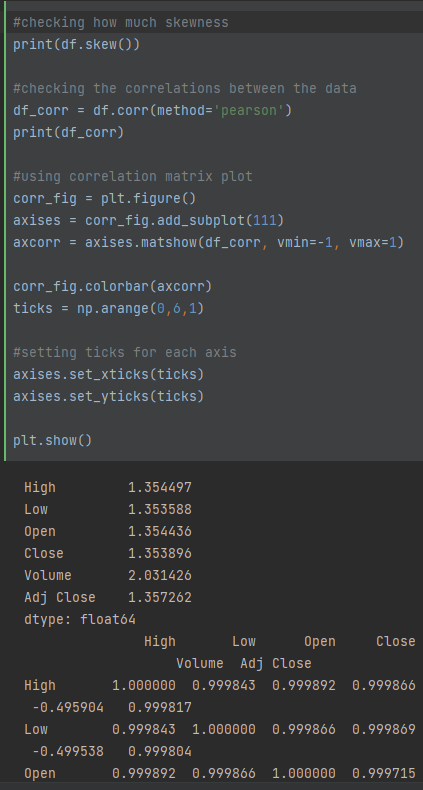
In this analysis for the volume variable, we could tell that a lot of shares of apple stock was traded between the year 2010 – 2014 and it started declining afterwards.

Chart

Description automatically generated

For the open variable, we can tell that at the beginning of 2022, the stock price was over 175 which was the highest at that specific period and we could see the trends.

Also checked the skewness of the datasets using the df. skew ()



**DATA PREPROCESSING**

For this project, data acquisition and pre-processing take place in the order listed below.

• Request and save the data from the Python Data reader library for Yahoo Finance.

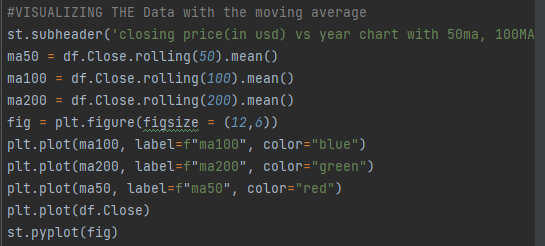
• Using the ScikitLearn MinMaxScaler helper function, normalised the data so that they fit in between 0 and 1.

• Reshaped the data.

• The rolling function was used to find the mean of each 100ma and 200ma data set.

Text

Description automatically generated



As a technical analysis tool, moving averages can be used to identify long-term trends and reduce short-term data fluctuations. A moving average is derived by averaging a certain set of data over a specific period, and it is updated as new data become available. Trends and patterns that may not be immediately evident in the raw data can be discovered with the aid of a set of averages created in this manner.

Chart, histogram

Description automatically generated

**Model development**

I used the linear regression model and the LSTM model for this assessment, and I would compare the two models’ evaluated performances side by side.

**Linear Regression Model**

**Data Processing:**

* As it serves the same purpose as the close column in this model—to decrease the dimension before resetting the index—I removed the adj columns and the date column.

Graphical user interface, text

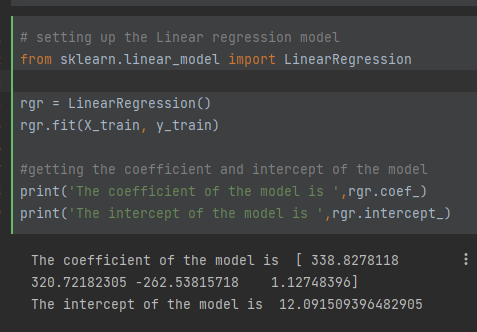
Description automatically generated

* Separated the data into train and test data.
* Using the MinMaxScaler to scale the values of the train data
* Split the dataset into training (70%) and test (30%)

Linear regression is used to identify the relationship between the historical stock data which we would use in our model. Based on historical data, including prior prices, earnings, and other variables that can have an impact on the stock's performance, linear regression can be used to predict the future value of a stock price. To be able to predict the stock prices using linear regression, we would need to get the historical data which is what we have done by collecting data in real-time via yahoo finance and then pre-process the data making sure the scales are between 0 and 1 then fit the regression model using the code below.

Text

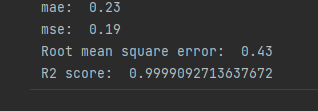
Description automatically generated



From the code above, the coefficient of the linear regression model and the intercept were gotten using the coef\_ and intercept\_ functions, then we check the evaluation of the model.

There are several evaluation metrics that can be used to evaluate the performance of a linear regression model. Some of the most used metrics include the following: Mean squared error (MSE) Root mean squared error (RMSE) Mean absolute error (MAE) R-squared (R^2)Text

Description automatically generated



From the result above:

* The mean absolute error was 0.23 which indicates that on average the model prediction is off by 0.23 units.
* The mean square error was 0.19 which indicates that, on average, the predicted values are 0.19 units off from the actual values. This might be seen as having a low error rate.
* The root mean square error means that on average, the predicted values are 0.43 units off from the actual values. This number can be understood as the model's average error. A lower RMSE value indicates a more accurate model.
* The R2 score was 0.99 which indicates that the model is very good at predicting the outcome of the stock This is because the R2 score is a measure of how well a model fits the data, and a value of 0.999 is close to 1, which is the maximum possible R2 score. This means that the model can explain almost all the variations in the data and is likely to make very accurate predictions.

Predicting the model, we would use the model. predict function on the test data.

Text

Description automatically generated

Created a data frame which will include the actual price against the predicted price and see the similarities.

Graphical user interface, text, application

Description automatically generated

From the result above, the actual price and the predicted price are very similar to one another, plotted a bar graph of the actual price against the predicted price.

Graphical user interface, application

Description automatically generated

From the graph above, we could tell that the model predicted the stock prices well.

**LONG SHORT-TERM MEMORY(LSTM) MODEL**

**THEORETICAL BACKGROUND**

Long Short-Term Memory (LSTM) is one of several varieties of Recurrent Neural Networks (RNN), and it can capture data from earlier stages and utilise it to make predictions about the future. Typically, an artificial neural network consists of three layers: the input layer, the hidden layer, and the output layer.

The number of input layer nodes in a Neural Network with a single hidden layer is constantly influenced by the dimension of the data. The input layer nodes are connected to the hidden layer through connections known as "synapses." Every two-node relationship from the input to the hidden layer contains a coefficient called weight that determines how signals are processed. After the learning process is complete, the Artificial NN will have the best weights for each synapse. The process of learning is naturally a continuous modification of weights. The sum of weights from the input layer is applied by the hidden layer nodes to a sigmoid or tangent hyperbolic (tanh) function, known as the activation function, which produces values with a reduced error rate between the train and test data using the SoftMax function.

The values obtained after this transformation make up the output layer of our NN. If the values obtained are not the best, a back propagation process will be used to target the ideal value of error. The backpropagation process connects the output layer to the hidden layer, sending a signal conforming to the best weight with the ideal error for the chosen number of epochs. Repeating this process will help us improve our forecasts and reduce prediction errors. This process will be finished with the model being trained. Recurrent Neural Networks (RNN) are a class of NN that use previous stages of learning to predict future value based on sequences of data.

LSTM MODEL:

For the lstm model, scaling and reshaping our train dataset so that it works with the neural network, then instantiate the lstm model called Sequential and lstm has some key hyperparameters which can help improve accuracy and performance of an LSTM model, which is shown using the code below:

Text

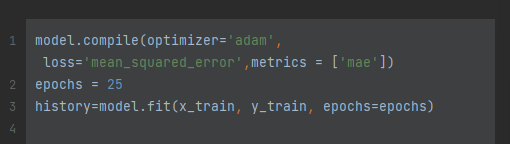
Description automatically generated

From the above code:

* Units is the number of units present in the LSTM layers
* Relu (Rectified Linear Unit) is the type of activation used in the LSTM layer and is very efficient.
* Return sequence means returning either the full sequence of outputs or just the last output.
* A dense layer, on the other hand, is a type of fully connected neural network layer in which all neurons in the layer are connected to all neurons in the preceding and subsequent layers.
* A final dense layer that outputs the predicted value

**Training the Model**:

Once the data is ready and prepared, we may train our model.



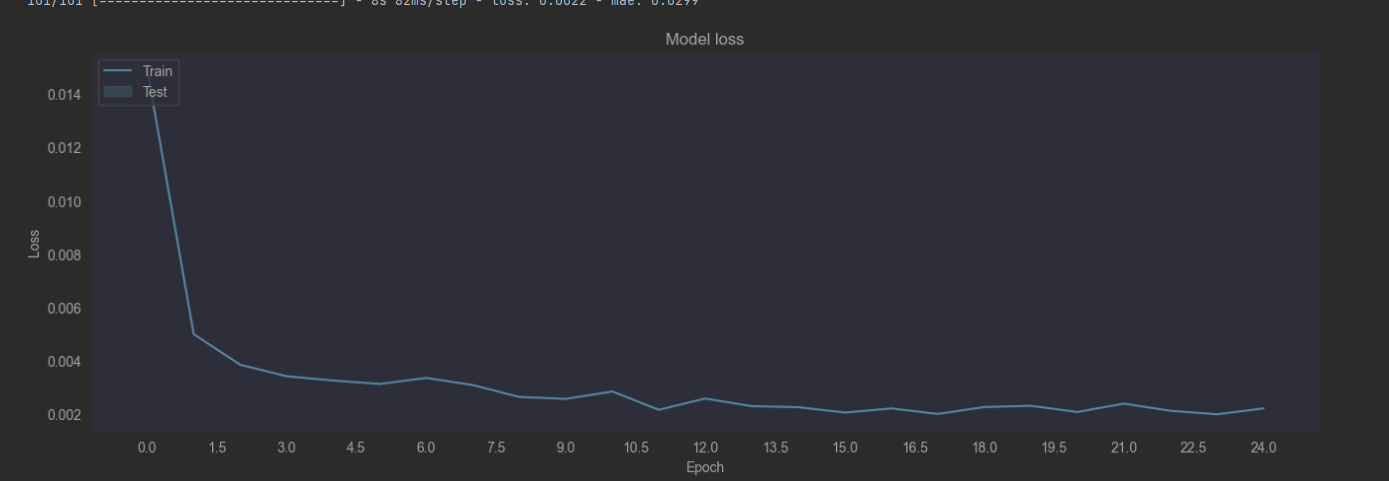
From the code above:

* Adam optimizer is the type of optimizer and learning rate used to train the LSTM model
* Epochs are single iterations of the training dataset. In each epoch, the model will therefore look at each example in the training dataset once. When one epoch is finished, the model will start a new one, and so on until the desired number of epochs has been attained.

Plotting a loss curve to determine the loss in the training process.

Text

Description automatically generated



The loss rapidly decreases to a lower threshold, indicating that the model has improved during the period of training.

**MAKING PREDICTIONS ON TEST DATA**

Text

Description automatically generated

Chart, histogram

Description automatically generated

The graph shows the actual price against the predicted price.

**MODEL PERFORMANCE**

Text

Description automatically generated

From the results above:

* A MAE of 7.56 means that, on average, the model's predictions are off by 7.56.
* A RMSE of 10.93 means that, on average, the model's predictions are off by 10.93.
* An R2 score of 0.946 means that 94.6% of the variance in the dependent variable can be explained by the model's predictions. This is considered to be a very good result, indicating that the model can make accurate predictions.

**MAKING PREDICTIONS FOR THE NEXT DAY**

To make predictions for the next day

Text

Description automatically generated

With this model, we were able to predict the apple stock price for tomorrow.

**Web-App Using Streamlit**

I used streamlit to design a web application after successfully developing the model and testing it on numerous stocks. The user can enter any stock ticker of his choosing along with the starting date, and the web application will then predict the stock price.

Graphical user interface, application

Description automatically generated

Chart, line chart, histogram

Description automatically generated

The web app displays the actual price vs the predicted price as a graph and then displays the next day’s price with the dollar sign next to it.

Graphical user interface, text, application, email

Description automatically generated

It also shows if there could be a rise or fall for tomorrow price.

Chart

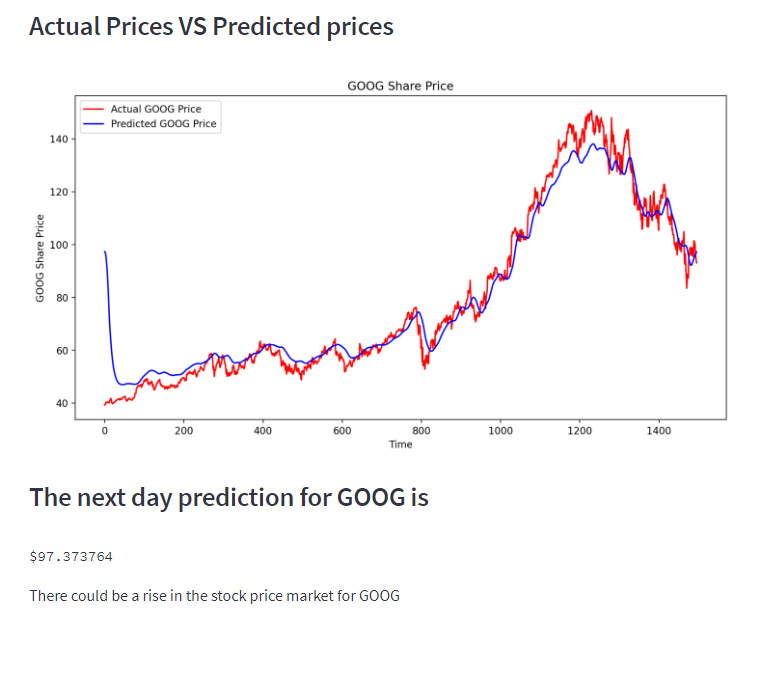
Description automatically generated

The above results show the stock price trend for Microsoft and the Next day’s price.

Chart, histogram

Description automatically generated

The above results show the stock price trend from January 2020-01-01 till Date for Tesla and the Next day’s price.



The above results show the stock price trend from January 2017 -01-01 till Date for Google and the Next day’s price.

I also tried seeing if it could predict and show some crypto data since it the data would be gotten from yahoo finance and it did.

Graphical user interface, table

Description automatically generated

Chart

Description automatically generated

I could remember the moment doge went to an all-time high in 2021 as I was also trading then.

Chart

Description automatically generated

**CONCLUSION**

Both LSTM and linear regression can be used for stock prediction, but LSTM may be better suited for this task in some cases. LSTM is a type of recurrent neural network that can process sequences of data, such as time series data. This makes it well-suited for tasks such as stock prediction, where the order of the data points is important. In contrast, linear regression is a more basic machine learning model that is used for predicting a continuous value based on a set of input features. It does not have the ability to process sequences of data, so it may not be as effective for stock prediction as an LSTM model.

I developed a web application for predicting close stock prices using Linear Regression and LSTM algorithms for prediction. Datasets belonging to Google, Tesla, Apple, and Microsoft have been applied and achieved above 90% accuracy for these datasets.

**References**

* LINDEMANN, B. et al., 2021. A survey on long short-term memory networks for time series prediction. Procedia CIRP, 99, 650–655
* LI, Y. et al., 2019. EA-LSTM: Evolutionary attention-based LSTM for time series prediction. Knowledge-Based Systems, 181, 104785
* SAXENA, S., 2021. *LSTM | Introduction to LSTM | Long Short Term Memor* [online] [viewed 14 Dec 2022]. Available from: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/