New York Stock Value Prediction using Linear Regression and LSTM - A comparison

Roshni Balasubramanian

Gavin Rufus Arul Samraj

Saket Sharma

University at Buffalo The State University of New York

Introduction - What is NYSE?

- **New York Stock Exchange (NYSE)** is the largest stock exchange in the world in terms of market capitalization, with a value of over $30 trillion as of 2021. Located on Wall Street in New York City.
- The NYSE is also referred to as the "Big Board" because, in the early days of trading, stock quotes and trade activity were updated manually on a big board.
- The NYSE employs a hybrid trading methodology that includes computerized trading and conventional floor trading.
- **The data used in this project is from the New York Stock Exchange, dated from 2010 to 2016, sourced from Kaggle.**

Research Questions

- Can we predict the stock behavior?
- If yes, how accurately can we predict the “Closing” stock price of a company based on its “Open”, “Low”, and “High” prices?
- How can we leverage this model?

Preprocessing

- To ensure the quality and suitability of the dataset for building effective models.
- Filter out the stock price records of Amazon.
- Duplicate entries are removed, if any, and NULL values are checked for.
- Imputation used to address missing data by replacing them with estimated values calculated from the data that is available, or deletion can be used to delete entire rows or columns of missing data from the dataset.

	open	close	low	high	volume
count	1762.000000	1762.000000	1762.000000	1762.000000	1.762000e+03
mean	337.875664	337.899058	333.969688	341.464438	4.607596e+06
std	189.294231	189.109339	187.654696	190.525796	3.091557e+06
min	105.930000	108.610001	105.800003	111.290001	9.844000e+05
25%	192.962494	193.377506	190.284997	195.532501	2.741550e+06
50%	282.500000	282.915008	279.869995	285.074997	3.890700e+06
75%	398.425003	398.014999	393.799988	402.082496	5.384450e+06
max	845.789978	844.359985	840.599976	847.210022	4.242110e+07

Fig. 1: Descriptive Statistics of Dataset

Correlation

- Check for strength of relationship between variables in the dataset.
- The figure shows a **strong positive correlation** between "Open," "High," "Low," and "Volume" prices with the "Close" price and amongst each other.
- To reduce complexity, "Open," "High," and "Low" prices are chosen for analysis.
- "Volume" is dropped off from the analysis.

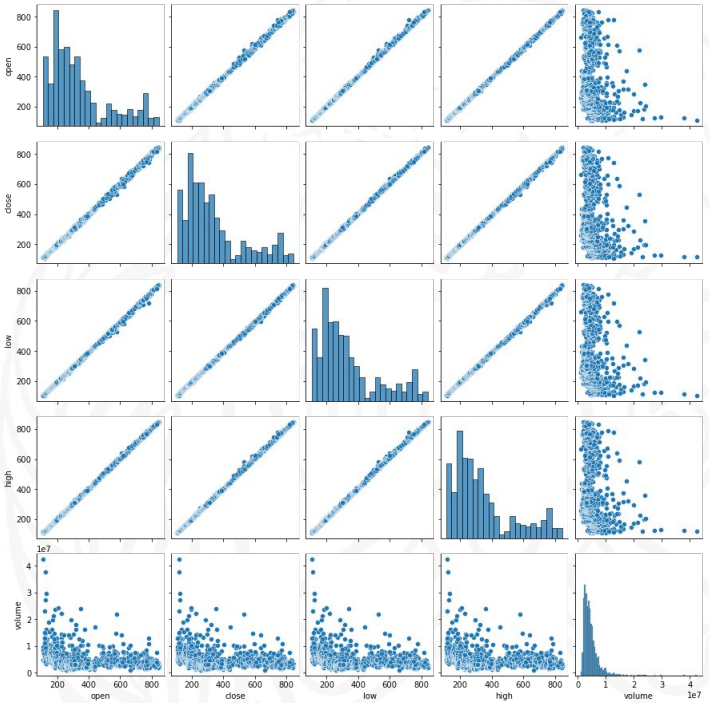
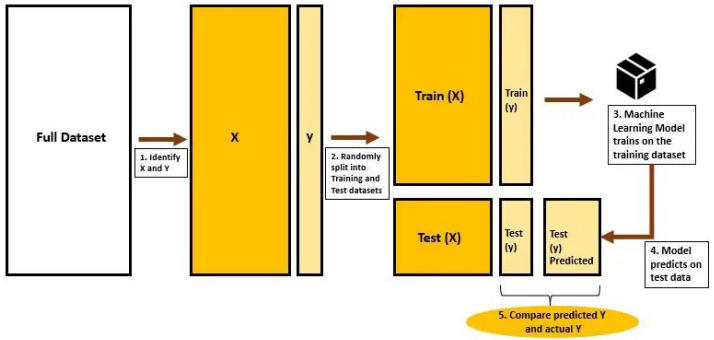


Fig. 1: Correlation Plots

Splitting Dataset

- The dataset is split into training data and testing data to evaluate the performance of the models.

- The training data consists of 80% of the records, and the testing data consists of the remaining 20%.



Model and Performance Overview

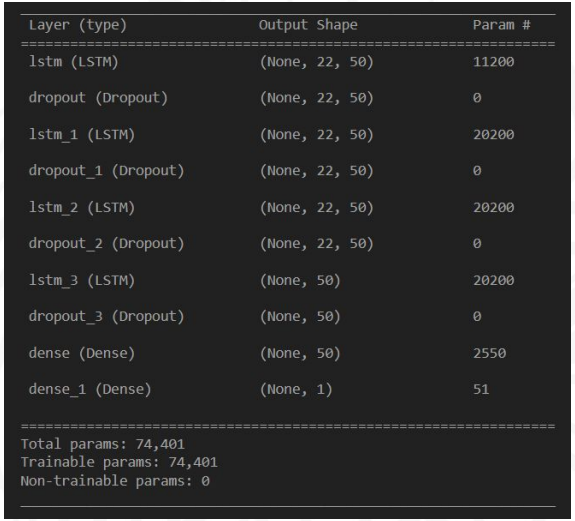
- Two models were used: a simple **Linear Regression** and a more complex **Long-Short Term Memory (LSTM)** deep learning model.
- The performance of each model is evaluated based on metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- We can determine which model is more appropriate for data analysis and prediction by comparing the performance of the two models.

LINEAR REGRESSION

- The training data is fed into the Linear Regression model to fit the data along a line that extends in a linear fashion.
- The testing data is used to validate the training by passing test variables through the model.
- Predicted values are obtained on the fitted line based on the testing data.
- The predicted values are compared to the actual test values to calculate accuracy and error.

LONG SHORT-TERM MEMORY (LSTM)

- **LSTM** is a type of **Recurrent Neural Networks**
- Each record of linear data can be represented as a linear expression with coefficients as weights of an LSTM node.
- We adjust the weights through multiple epochs to fit the linear relationship and get accurate output.
- LSTM has a memory cell to remember the cell state for efficient processing.
- In this instance, 100 epochs are given for processing.
- The activation function used is the Linear activation function.
- Dropout Layers are used to deactivate a few weights to avoid overfitting and prevent the model from memorizing the values.



Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 22, 50)	11200
dropout (Dropout)	(None, 22, 50)	0
lstm_1 (LSTM)	(None, 22, 50)	20200
dropout_1 (Dropout)	(None, 22, 50)	0
lstm_2 (LSTM)	(None, 22, 50)	20200
dropout_2 (Dropout)	(None, 22, 50)	0
lstm_3 (LSTM)	(None, 50)	20200
dropout_3 (Dropout)	(None, 50)	0
dense (Dense)	(None, 50)	2550
dense_1 (Dense)	(None, 1)	51

Total params: 74,401
Trainable params: 74,401
Non-trainable params: 0

Fig. 1: Layers of the LSTM model

Results

How does each model perform?

LINEAR REGRESSION

- **Train score: 0.99 & Test score: 0.99**
- **Mean Absolute Error (MAE): 1.56**
- Initial impression of overfitting due to high train accuracy score
- The Test accuracy score proves model efficiency in predicting new data.
- The figure shows significant overlap between actual and predicted test values with no overfitting.

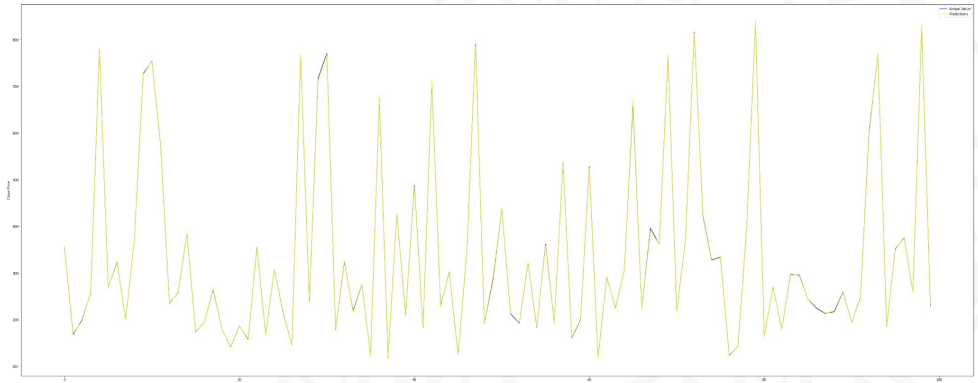


Fig.1: Actual Test values and Predicted Test values comparison (LR)

LONG SHORT-TERM MEMORY (LSTM)

- **Train score: 0.86 & Test score: 0.32**
- **RMSE for Train data: 17.09**
- **RMSE for Test data: 67.27**
- The LSTM model is overfitting with poor accuracy scores, indicated by the test accuracy being much lower than train accuracy.
- The losses from Figure 1 confirm the **overfitting**.
- The predictions in Figure 2 are centered around a certain value, which may be due to a bias from overfitting.

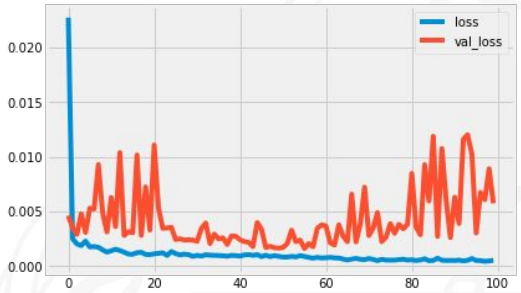


Fig. 1: Train and Validation

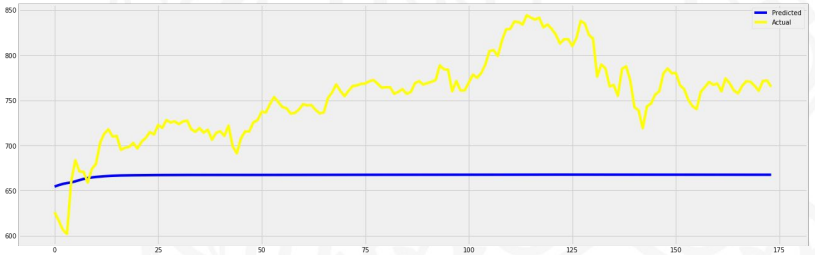


Fig. 2: Actual Test values and Predicted Test values comparison (LSTM)

Conclusion

Which approach is better? Naive or Complex?

Which model is better?

- Naive approach (**Linear Regression**) is good at predicting Closing Stock Prices compared to complex LSTM method.

How can we improve the accuracy?

- LSTM efficiency can be improved by adding more Dropout Layers to reduce overfitting
- Additional LSTM layers can be added to arrive at more accurate weight values
- Increasing the epochs can increase the training time and improve accuracy

Benefits of accurate stock prediction

- Accurate prediction of stock market helps investors make better decisions, manage risks, and reduce vulnerability to market turbulence
- Companies can use it to understand their own value better and know how their financial performance is perceived by the public

Why this model?

- Linear Regression model is practical to use due to its inevitable benefits
- It can be deployed as a mobile/web application to predict the best closing price when input variables are provided
- Building such an app is cost-effective considering the benefits of the model's efficiency.

Thank you

Ask your questions!