LSTM vs ARIMA on Google Stock Price Data

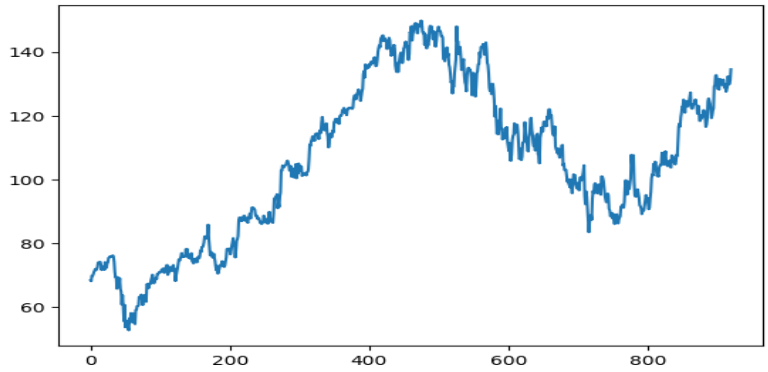
Name: Tanmoy Paul Date:1st Sept 2023

Abstract:

This study compares the results of two completely different models: statistical (ARIMA) and deep learning (LSTM) based on Google stock price data. Both models are used to predict daily or monthly average prices for Stocks. This project shows which model performs better in terms of the chosen input data, parameters. The chosen models were compared using the relative metric mean square error (MSE) and mean absolute percentage error (MAPE). The performed analysis shows which model achieves better results by comparing the chosen metrics in different models. It is concluded that the ARIMA model performs better than the LSTM model

Methodology:

1. **INPUT DATA:** we have used Google Stock Price Data to compare the two models. Collected the Stock data-GOOGL from 2020-01-01 to 2023-08-30 using the yfinance package



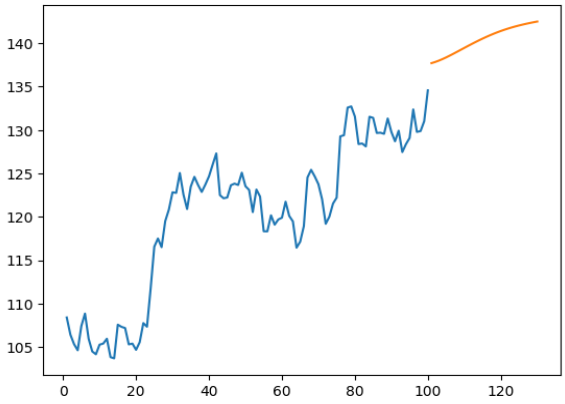
2. **ARIMA model:** The basic model in the time series analysis is the ARIMA model. It is a combination of two processes – autoregressive (AR) and moving average (MA). The structure of ARIMA is based on the phenomenon of autocorrelation. ARIMA can be used for modeling stationary time series or non-stationary time series that can become stationary through differentiation. Stationary series are those whose expected value and variance do not change over time. The universal notation ARIMA (p, d, q) is used to describe the form of the ARIMA model. The letter p is the order of the regression, d is the order of differentiation and q is the order of the moving average. A decision is made about the need for data to transform and differentiate the series to stabilize its mean, variance, and covariance. This is done by both examining the autocorrelation function (ACF) and partial autocorrelation (PACF), and by performing statistical Dickey-Fuller and Augmented-Dickey-Fuller tests. Next, the parameters of the selected models are estimated. The final model selection is usually based on the analysis of several criteria – error metric and information criterion (Akaike's Information Criterion, Bayesian Information Criterion). Here I have used the auto_arima module in Python. Auto_arima works by conducting differencing tests (i.e., Augmented Dickey-Fuller) to determine the order of differencing. It also automatically selects the optimal p and q using Model selection criteria.

3. **LSTM model:** Long Short-Term Memory Networks is a deep learning, sequential neural network that allows information to persist. It is a special type of Recurrent Neural Network which is capable of handling the vanishing gradient problem faced by RNN. LSTMs offer greater expressiveness in modeling sequential data. They can capture and retain information over longer sequences. Here I have used stacked LSTM. The original LSTM model is comprised of a single hidden LSTM layer followed by a standard feedforward output layer. The stacked LSTM is an extension to this model that has multiple hidden LSTM layers where each layer contains multiple memory cells.

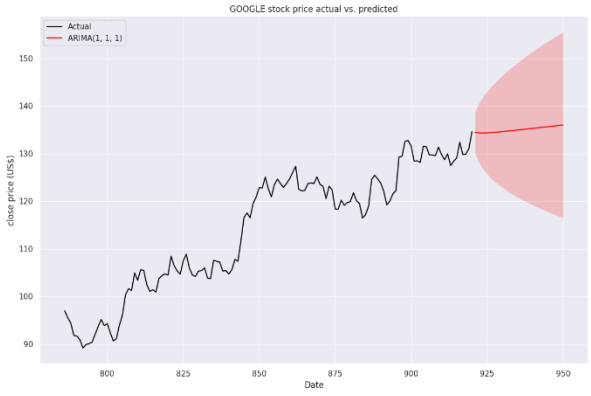
Advantages of Stacked LSTM over Simple LSTM:

1. Increased Model Capacity: to capture complex patterns.
2. Hierarchical Feature Learning: The lower layers can capture short-term dependencies, while the higher layers can focus on capturing longer-term dependencies in the data.
3. Increased Expressiveness: Stacked LSTMs offer greater expressiveness in modeling sequential data. They can capture and retain information over longer sequences
4. Parallelism in Training: The LSTM units within a layer are processed sequentially at each time step, different layers can be trained in parallel, which can result in faster convergence during training.

Predicting stock price for 30 days in the future using both LSTM and ARIMA.



LSTM Prediction



ARIMA Prediction

Result & Finding after comparing LSTM and ARIMA

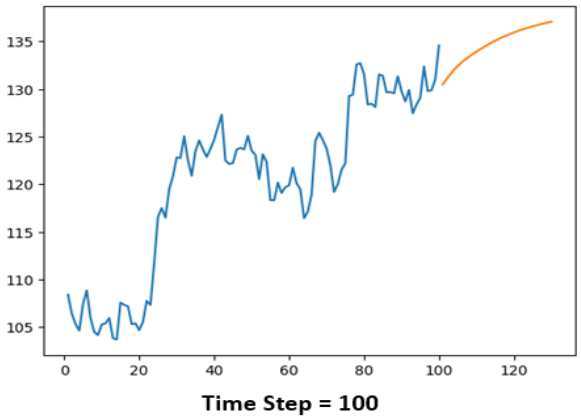
I have considered different Time windows for predicting stock prices using the two different models. Here I have used two metrics (MAPE & RMSE) to measure the performance of the two models for different predicting time windows (1, 3, and 6 months)

Prediction Time window	MAPE		RMSE	
	ARIMA	LSTM	ARIMA	LSTM
1 month(30days)	0.119	0.175	11.37	16.49
3 months	0.08	0.55	8.97	53.24
6 months	0.09	0.48	12.46	51.04

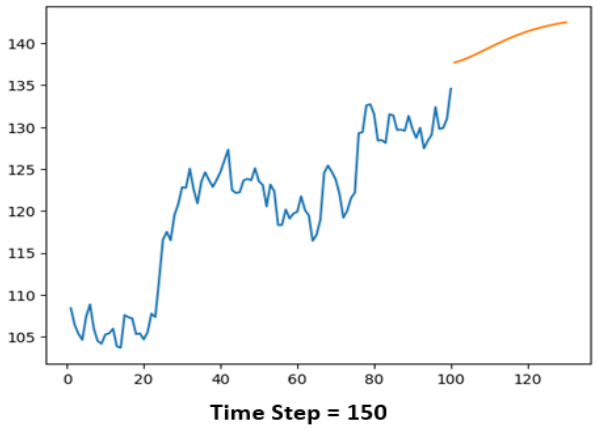
Note: MAPE(Mean Absolute Percentage Error), RMSE(Root Mean Square Error)

Conclusion:

- We see that ARIMA performs better than LSTM in prediction for all the prediction time windows. Thus ARIMA is better than LSTM for Stock Price Prediction.
- We also see that the performance of the ARIMA decreased (rather RMSE, MAPE increased) while we were predicting for 6 months compared to 3 months. This indicates that ARIMA models tend to be better suited for short-term forecasting rather than long-term predictions.
- Whereas the performance of the LSTM increased while we were predicting for 6 months compared to 3 months. This indicates that LSTM is good for long-term forecasting rather than short-term forecasting
- We also have experimented with different time step =100, 150 for LSTM.



Time Step = 100



Time Step = 150

. While experimenting we observed that for time step=150, the model expects to have a higher closing price for the next 30 days with respect to the model with time step =100.

Possible reason: ***The model with the smaller time step (100) might be overfitting to short-term noise or random fluctuations in the data***

My Opinion: ***We should take time step =150 because when we take a small time step models can be less reliable for long-term predictions.***