STOCK PREDICTION USING SENTIMENT ANALYSIS AND MACHINE LEARNING

Rohan Sharma

UB ID : 50545225

Motivation

Investing in the right assets has become an important financial objective for companies and individuals. Thus, stock market prediction is an essential tool for maximizing gains on capital invested.

Traditional stock market forecasting strategies have primarily relied on quantitative methods, such as statistical and mathematical models that examine historical price data and financial indicators.

In this research, I have incorporated Sentiment Analysis of tweets about the relevant stock company to see how they impact the prediction of the stock prices.

Introduction

- I have used Apple stock price data for my project

- I have used 2 batches of Apple data

- From 2016 to 2020

- And from 2021 to 2023

- I have predicted the 'adjusted closing' price of Apple stock in my LSTM models.

Methodology

- Data Collection and Pre-Processing

- Splitting into Training and Testing Sets

- Applying the model

- Model performance evaluation on the test set using R-squared and actual vs predicted plots.

- Insights generated from the results.

1ˢᵗ Model: Auto-Regressive Integrated Moving Averages (ARIMA)

- After data preprocessing steps such as removing NA values and setting date column as index, and splitting the data into train and test sets, I applied the first model which was ARIMA.

- The ARIMA model was only able to capture the basic trend of our data and did not yield very good prediction.

- I have considered the above ARIMA model as a baseline model.

ARIMA Forecast vs Actual for AAPL
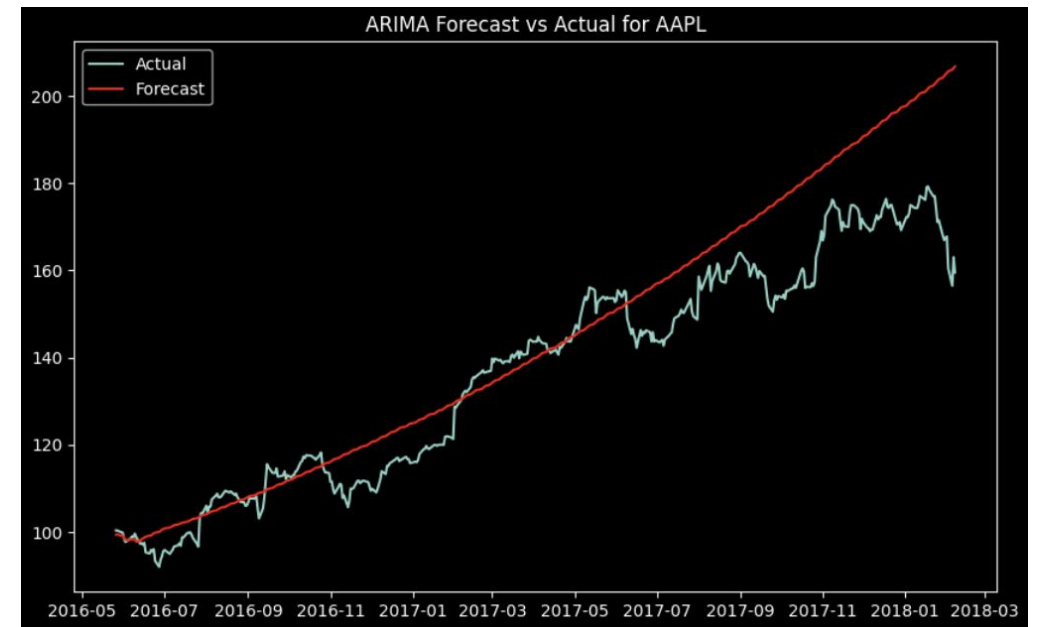
# Long Short Term Memory Network (LSTM) without Sentiment Analysis

For the next phase, I compared using LSTM model with and without Sentiment Analysis to see what impact Sentiment Analysis has on our prediction and whether it can significantly improve over our LSTM model.

I used LSTM-RNN model as they are very robust for time-series data and generally give good results for it.

I used the LSTM model on Apple stock data from 2016 to 2020 and the LSTM model gave me extremely good results by using the quantitative data alone.

R squared on testing : **0.9755**
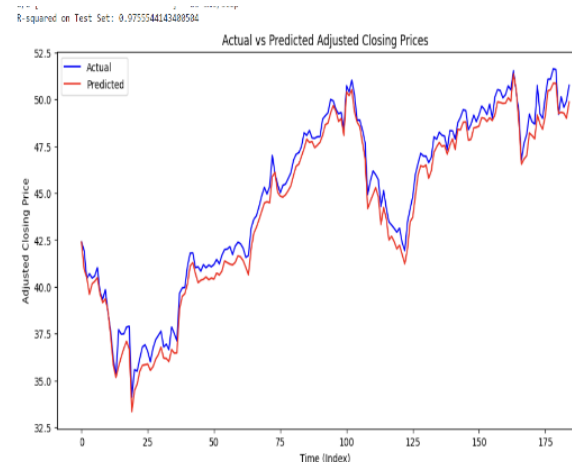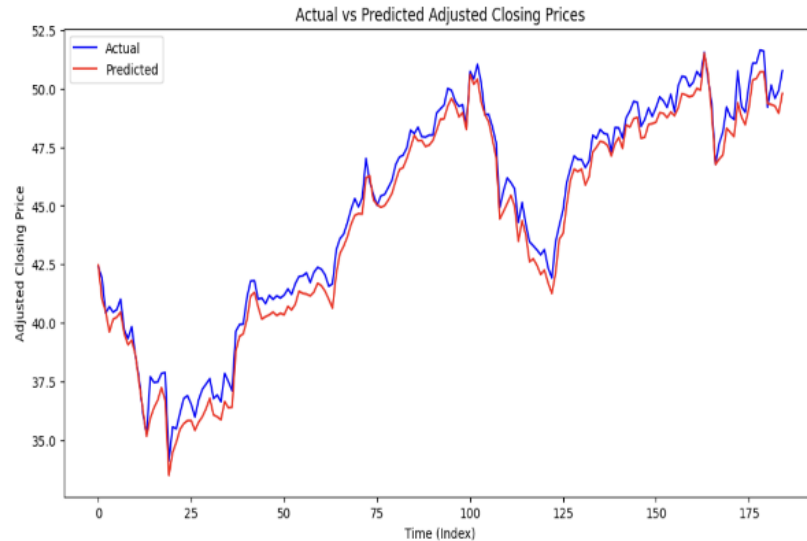
LSTM model with Sentiment Analysis

Actual vs Predicted Adjusted Closing Prices

- For Sentiment Analysis, I took data of Apple stock for the same time period with Twitter Sentiment polarity scores recorded for each day the market was open.

- By using LSTM model on the dataset including the polarity scores, I got a slightly better result for prediction.

- R-squared on Test Set: **0.97787**

LSTM model on same stock but different time period

Actual vs Predicted Values (Testing Data)
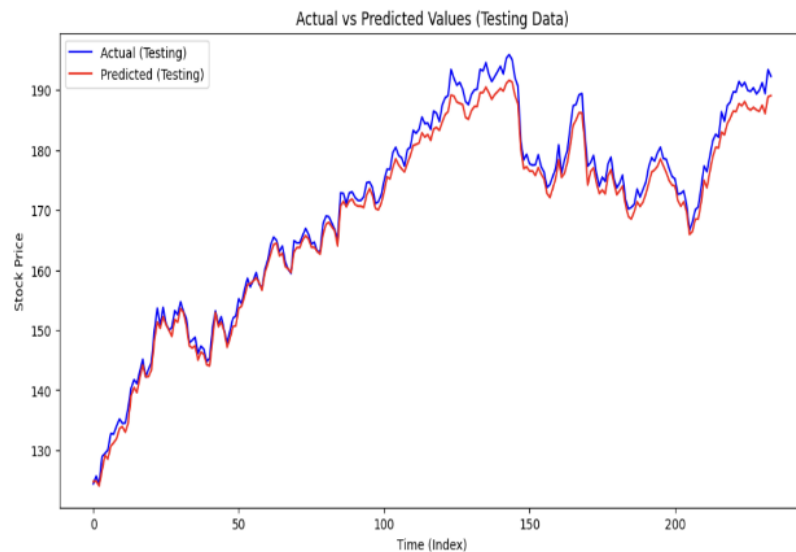
- I applied the same LSTM model on Apple stock data using data from 2021 and 2022 as training and data of 2023 as testing.

- On applying the LSTM model without any sentiment analysis on this data, I again achieved very good predictive performance on the stock prices.

- **R-squared testing : 0.98566**

Insights

- After analyzing the results of the LSTM models, I concluded that the models were performing extremely well on the quantitative data alone without the sentiment scores, thus it was difficult to see the impact of sentiment analysis on stock price prediction.

- When we used sentiment analysis, the prediction improved by a very small amount(0.23% more variance was explained).

- I think future work is needed in this domain by using varied companies' stocks and also trying different sources of sentiment data other than twitter such as news, relevant articles etc. and see what impact they have.
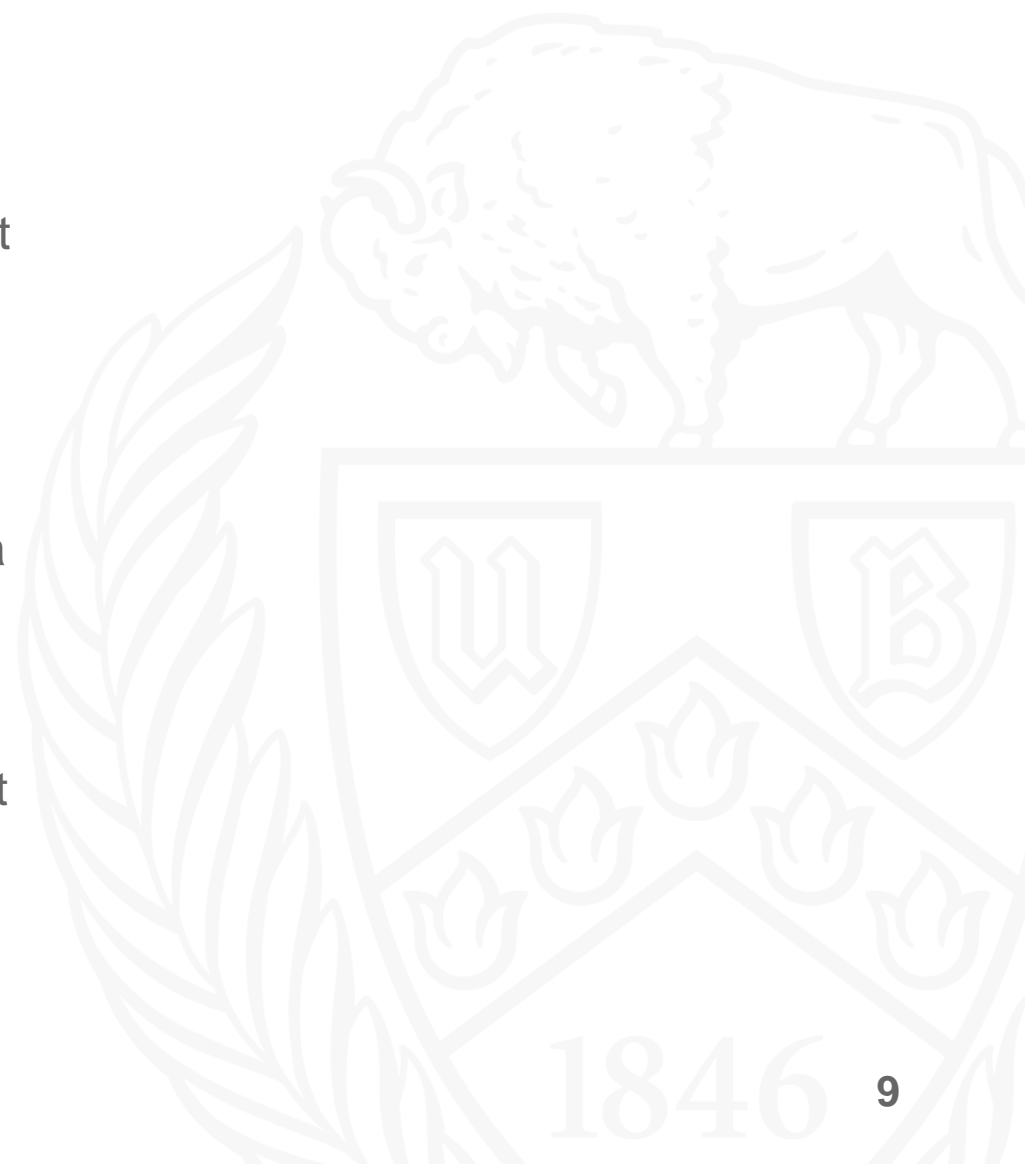
Contribution

- Thus, this research work has contributed to future researchers' work by showing how effective LSTM Neural Network models are for predicting time series data such as stock prices using the basic indicators such as opening price and volume traded.

- It also guides them to include better sources of sentiment data for doing sentiment analysis using NLP techniques to convert the sentiments to polarity.

- They can also try to experiment with the architecture of the LSTM model itself to see how they can improve it for different types of stock data.