PROJECT REPORT

Stock Price Prediction Using LSTM, ARIMA, and SVR

- **By Neeharika Yeluri**
- **016680508**

ISE 244 – AI Tools and Practice for Systems Engineering

Dr. Shilpa Gupta

shilpa.gupta@sjsu.edu

SJSU SAN JOSÉ STATE UNIVERSITY

Table of Contents

1. Problem Definition:

The project aims to predict the future stock prices of publicly traded companies using machine learning and deep learning models. The study proposes the use of a novel approach Long Short-Term Memory (LSTM) model for stock price prediction and compares their performance with traditional methods such as Autoregressive Integrated Moving Average (ARIMA) and Support Vector Regression (SVR). The project involves collecting real-world stock price data from various sources, preprocessing the data, training and evaluating the models, and comparing the results with traditional methods. The goal is to develop a model that accurately predicts future stock prices and outperforms traditional methods in terms of RMSE and MAE.

2. Project Objectives:

- Develop and implement machine learning and deep learning models for predicting future stock prices of publicly traded companies.
- Compare the performance of Long Short-Term Memory (LSTM) model with traditional methods such as Autoregressive Integrated Moving Average (ARIMA) and Support Vector Regression (SVR) in terms of evaluation metrics.
- Collect and preprocess real-world stock price data from various sources to ensure consistent and reliable analysis.
- Train and evaluate the models using the collected data to accurately forecast stock prices and identify potential investment opportunities.
- Assess the effectiveness of LSTM in outperforming traditional methods, specifically in terms of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics.
- Conduct exploratory data analysis to gain insights into the factors influencing stock prices and the performance of different companies.
- Provide valuable information and insights to investors and traders to aid in making informed decisions about buying and selling stocks.
- Explore the limitations and challenges associated with the LSTM model, such as data size requirements and hyperparameter tuning.
- Consider the possibility of expanding the analysis to include additional factors like news sentiment, macroeconomic indicators, and geopolitical events that may impact stock prices.
- Investigate the potential of more advanced deep learning models, such as convolutional neural networks and transformers, for stock price prediction.

3. Analysis:

3.1 Dataset: Webscraped from https://in.finance.yahoo.com using selenium and BeautifulSoup.

Datasets contain stock market data for each company from the period starting from years like 1987, 2004 to 2021-02-26.

These datasets contain 7 columns, namely:

Date: the date of the stock market data

Open: the opening price of the stock on that day

High: the highest price of the stock on that day

Low: the lowest price of the stock on that day

Close: the closing price of the stock on that day

Adj. Close: the adjusted closing price of the stock on that day

Volume: the volume of the stock traded on that day

The four datasets represent stock prices of four technology giants, namely Apple, Google, Microsoft, and Amazon. These datasets have been collected and compiled from different sources and provide information about the stock prices of these companies.

The Apple dataset has a total of 9800 rows and 7 columns.

The Google dataset has 4162 rows and 7 columns.

The Microsoft dataset has 8890 rows and 7 columns.

The Amazon dataset has 5989 rows and 7 columns.

3.2 Data preprocessing:

Data preprocessing steps on stock data of four companies, namely Apple, Google, Microsoft, and Amazon are as follows:

- Drop NULL rows: The dataset has missing values that are removed from the dataset using the dropna() function.

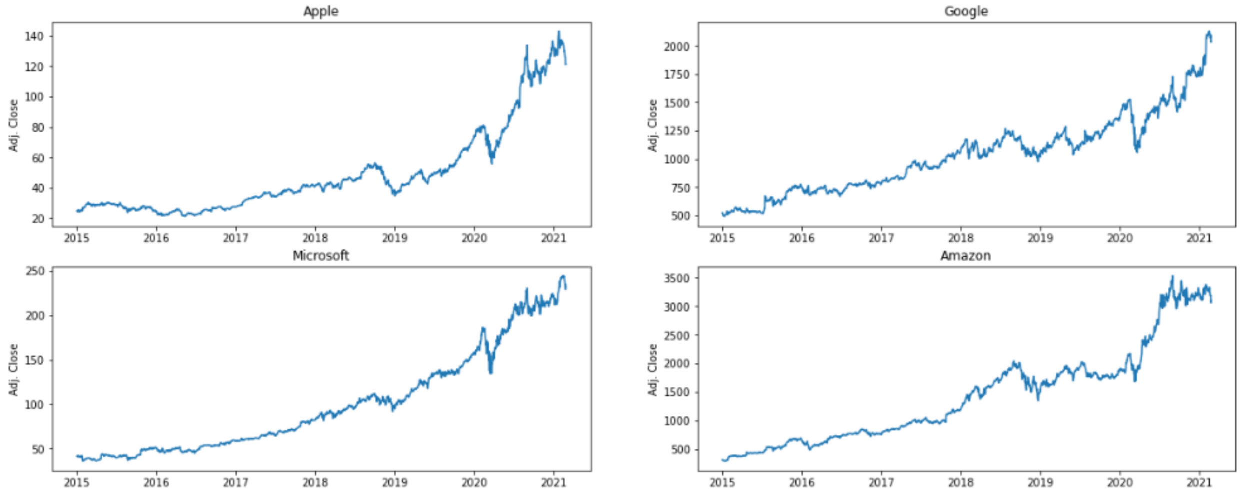
- Change Dtype of Columns: The columns, "Open" and "Volume" columns are cleaned by removing commas. After cleaning, the columns are converted to the float data type using the astype() method.

- Sort the Database by Date: The dataset is sorted by the "Date" column using the sort_values() function.

- Drop rows having Date < '2015-01-01': drops any rows that have a date earlier than January 1, 2015. This is done to ensure that the data used for analysis is consistent across all four datasets.

After the data is pre-processed, exploratory data analysis is performed. Visualizations are plotted for closing prices and trading volume of each company using the matplotlib library. The moving average of the stock is calculated using the rolling() function of pandas and is plotted along with its respective stock prices.

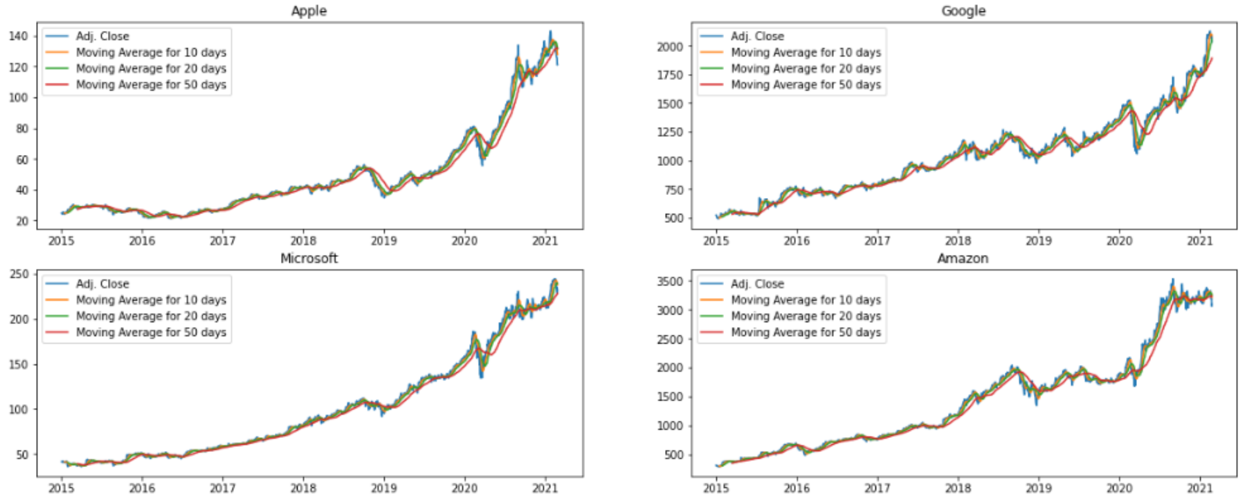
3.3 Exploratory Data Analysis:



Historical view of the closing prices

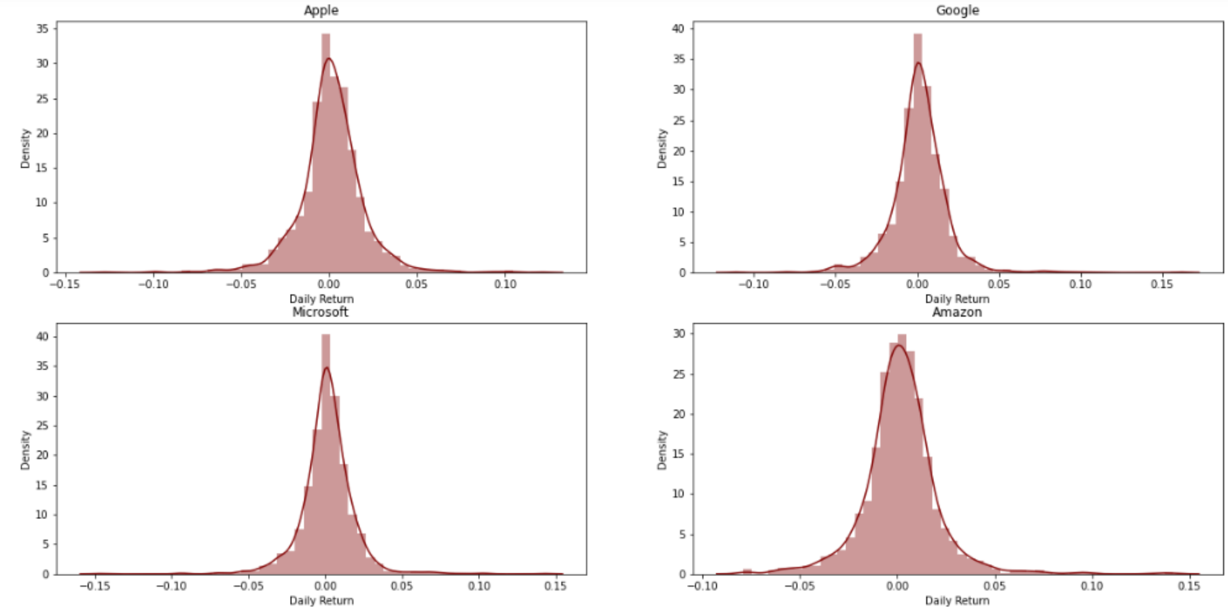
We can see from the above graph that Apple shares have tremendous growth in the 2020-2021 period.

We can assume that COVID-19 is the primary factor affecting the 2020-2021 period.



Moving average of the stock

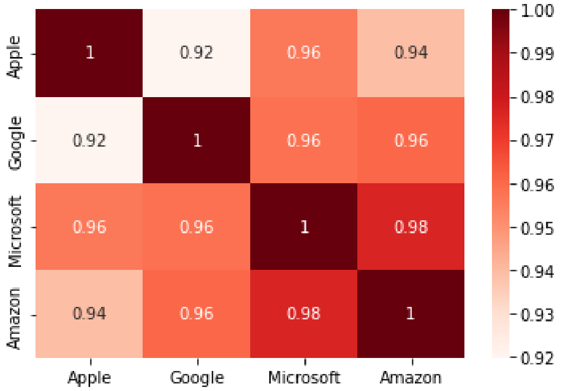
We can see that there is a drastic increase in the closing price of the stock of all four datasets from 2019 – 2021.

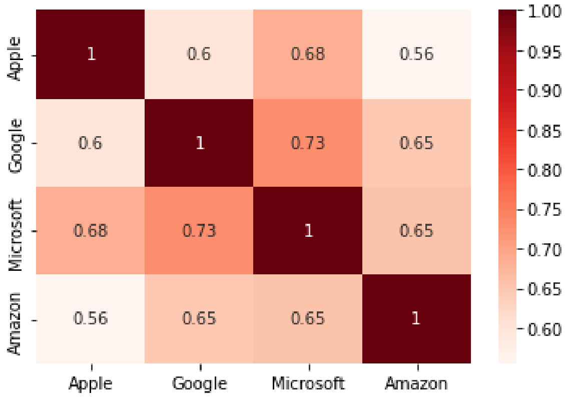


Distplots on Daily Returns

We can say that from the above plots, almost all datasets have the same daily returns of the stock price on each day but a little higher for the amazon dataset.

Correlation Plots

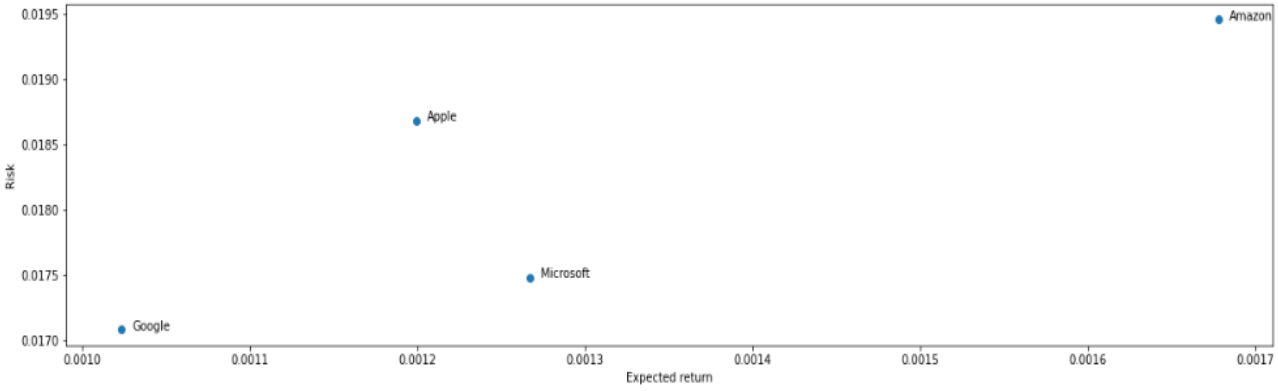




correlation between stocks closing price **correlation between stocks daily returns**

From the above plot, we can see that Microsoft and Google had the strongest correlation in stocks daily returns.

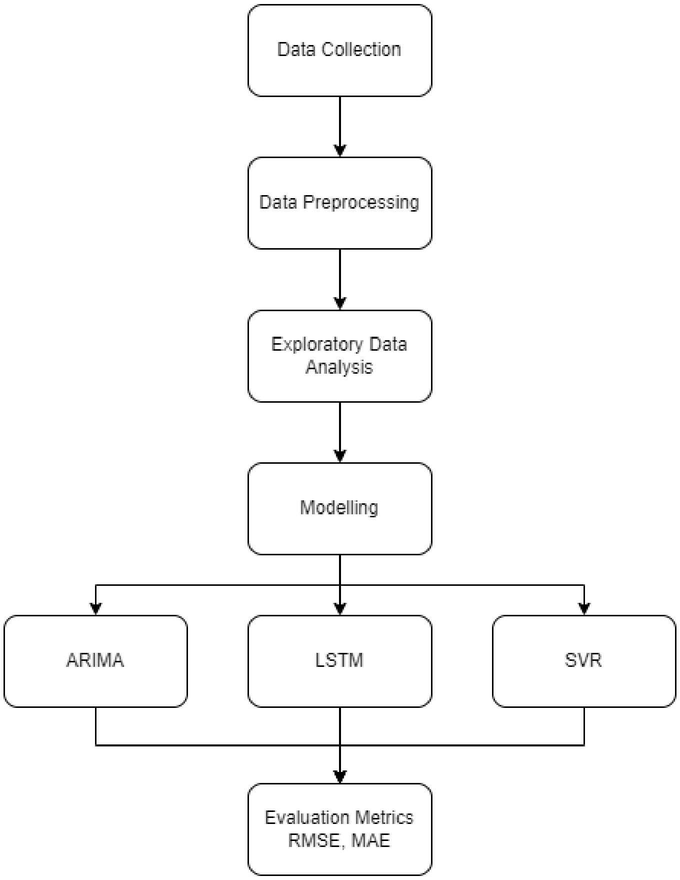
Also, Microsoft and Amazon have the highest correlation in stocks closing price.



Risk v/s Expected Returns

From the above graph, we can see that Amazon has the highest expected returns and the highest risk factor. Google has the lowest expected returns and the lowest risk factor.

3.4 Modeling:

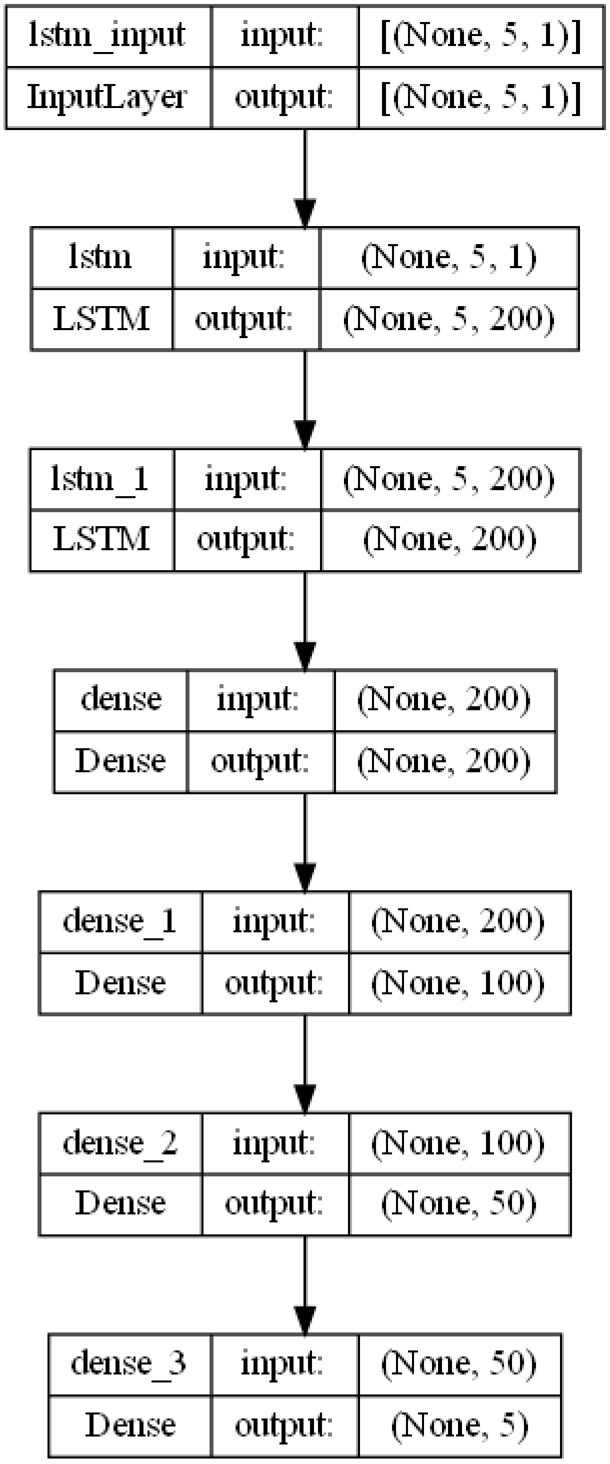


Pipeline

3.4.1 LSTM:

The data of four tech companies are split into training data from 2015-01-02 to 2020-09-30 and testing data from 2020-10-01 to 2021-02-26. The model is trained using the training data and is fit to predict the adjusted closing price of the stock. The neural network model is compiled using the Adam optimizer and mean squared error loss function with the root mean squared error used as the evaluation metric. The model is trained using 200 epochs for Amazon, 1000 epochs for Apple, Microsoft, and Google, and also early stopping is applied using the callback function. The history of the model fit is plotted in two graphs, one displaying training and validation loss and the other displaying training and validation root mean squared error. The model is then used to predict the adjusted closing price of the stock for the testing period. The predictions are plotted along with the actual stock price, and the root mean squared error and mean absolute percentage error of the model are calculated and printed.

As a week consists of five working days, we utilize the training dataset records to train the model, which we then use to estimate the closing values for the upcoming week. A walk-forward validation mode is used throughout the multi-step forecasting process.

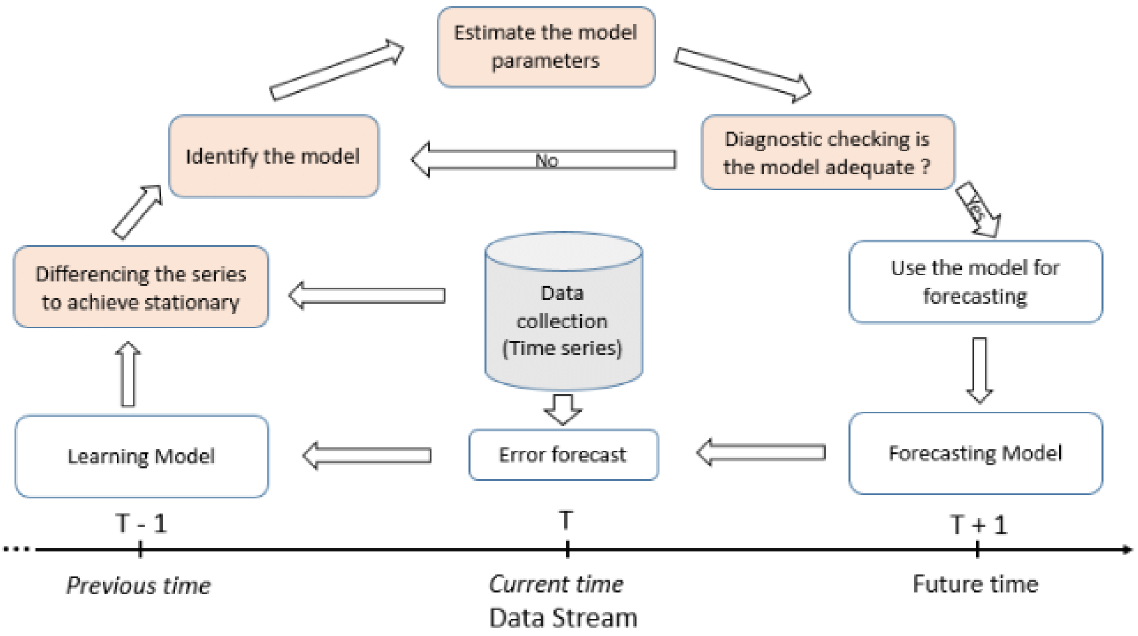


Architecture

The following describes the model's overall architecture and the specifics of each layer's design:

The input data's shape to the network's input layer is (5, 1), meaning that the time series' last five values (i.e, data from one week) are utilized as the input. The closing value is the only attribute of the data that is taken into account. The LSTM layer, which has 200 nodes at the output and uses the Leaky ReLU activation function, receives the data from the input layer and passes it on. The output of the first LSTM layer is sent to a second LSTM layer that has 200 nodes and uses the Leaky ReLU as its activation function. The output of this layer is then sent to a dense layer with 200 nodes at its input and output with the Leaky ReLU activation function. The output of this layer is sent to another dense layer that has 100 nodes with a Leaky ReLU activation function at the output and 200 nodes at its input. The output of this layer is sent to another dense layer that has 50 nodes with a Leaky ReLU activation function at the output and 100 nodes at its input. The output layer, which is also fully connected, is at last connected to the dense layer. There are 50 nodes at the output layer's input and 5 nodes at the output. The anticipated values for the five days of the upcoming week are produced by the five nodes at the output. Leaky ReLU is once again used as an activation function in the output layer. The model employs ADAM as the optimizer with a unique learning rate, and MSE as the loss function.

3.4.2 ARIMA:



Estimate the model parameters
Identify the model
No
Diagnostic checking is the model adequate ?
Yes
Differencing the series to achieve stationary
Data collection (Time series)
Use the model for forecasting
Learning Model
Error forecast
Forecasting Model
T - 1
T
T + 1
Previous time
Current time
Future time
Data Stream

In this project, we built an ARIMA model to forecast the stock prices of Apple, Google, Microsoft, and Amazon.

We used historical stock prices of datasets from 2010 to 2021, obtained from Yahoo Finance. The dataset contains daily closing prices, adjusted for splits and dividends, along with the date.

We first changed the data type of the date column to datetime. Then, we checked the stationarity of the time series using the Dickey-Fuller test. The test indicated that the data is not stationary, and hence we applied the log transformation to reduce the trend and stabilize the variance.

We then split the data into a training set from January 2015 to September 2020 and a testing set from October 2020 to February 2021.

Then we used the Auto ARIMA algorithm to find the best set of parameters for the ARIMA model. The algorithm identified the best parameters as (1, 1, 0). We then fit an ARIMA model with the same parameters to the training data.

The performance of the model is evaluated by comparing the predicted values to the actual values in the testing set. We used the mean squared error (MSE) and mean absolute error (MAE) as evaluation metrics.

The ARIMA model was able to capture the general trend and seasonal fluctuations in the data, as seen in the plot of predicted versus actual values. However, it failed to predict the sudden drop in prices in September 2020, indicating that the model might not be able to predict extreme events.

3.4.3 SVR:

Support Vector Machines (SVM) is used for classification tasks and the goal of SVR is to find the best-fit line (or hyperplane) that has the maximum margin from the actual data points. The SVR algorithm tries to minimize the error between the predicted values and actual values while also trying to maximize the margin. It is a powerful tool for regression problems where there is a non-linear relationship between the input variables and the target variable. It can handle both linear and non-linear data and can work well in high-dimensional space. SVR is a useful algorithm for regression tasks, and its flexibility in handling non-linear data makes it a popular choice for various applications.

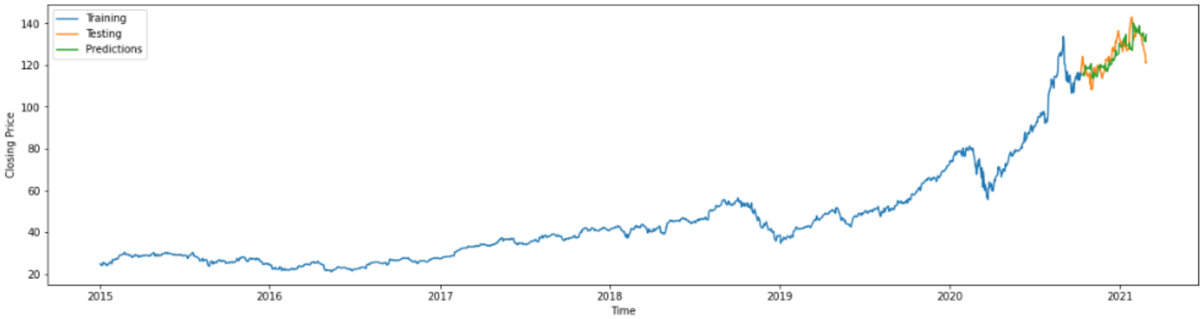
Here, we used the Support Vector Regression (SVR) algorithm to predict the stock prices of four different companies, namely Apple, Google, Microsoft, and Amazon. We first import the required libraries and the data for each company, which is split into training and testing data.

We then create an instance of the SVR model and reshape the training and testing data to fit the model's requirements. The training data's input and output are used to train the model, and the trained model is then used to predict the stock prices for the testing data. We plot the actual stock prices and the predicted stock prices on a graph.

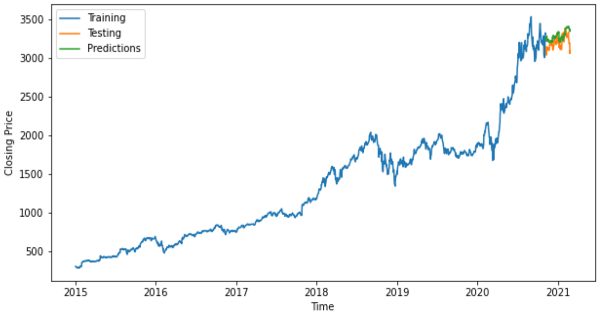
We can see that the predicted stock prices follow the actual stock prices relatively closely for each company. The RMSE and MAE values indicate that the SVR algorithm's performance is reasonable in predicting stock prices. However, we should note that stock price prediction is a challenging task, and prediction accuracy depends on many factors, including the complexity of the market, the data used, and other external factors. Therefore, we should not rely solely on these results to make financial decisions.

4 Results:

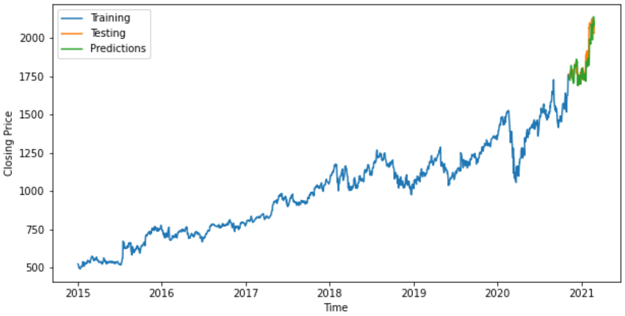
LSTM



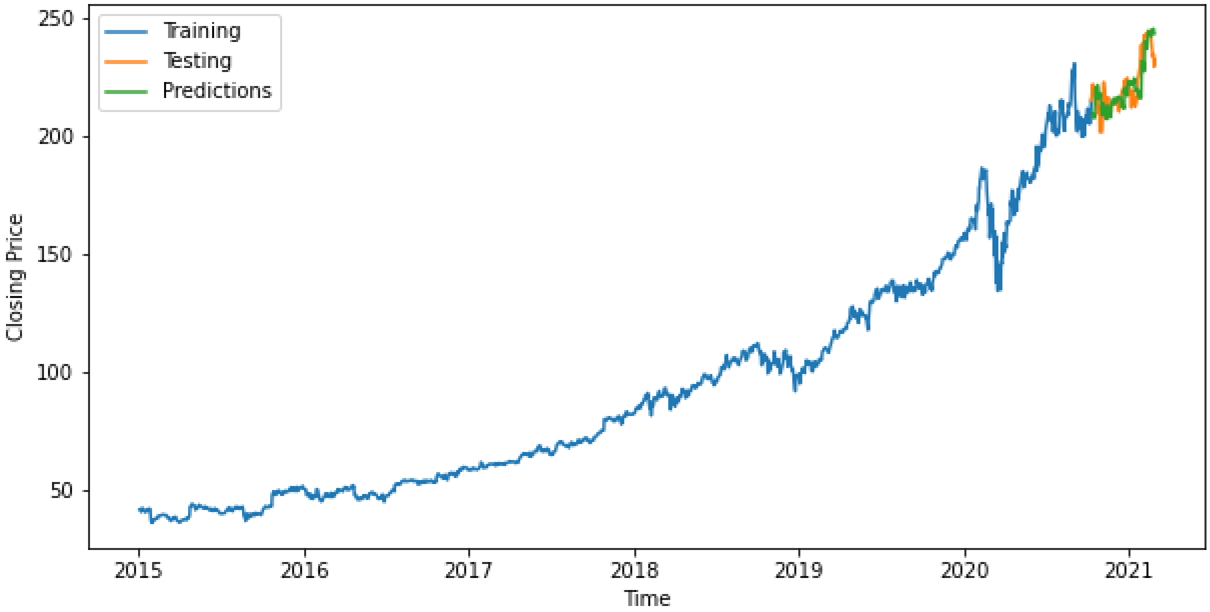
Apple



Amazon

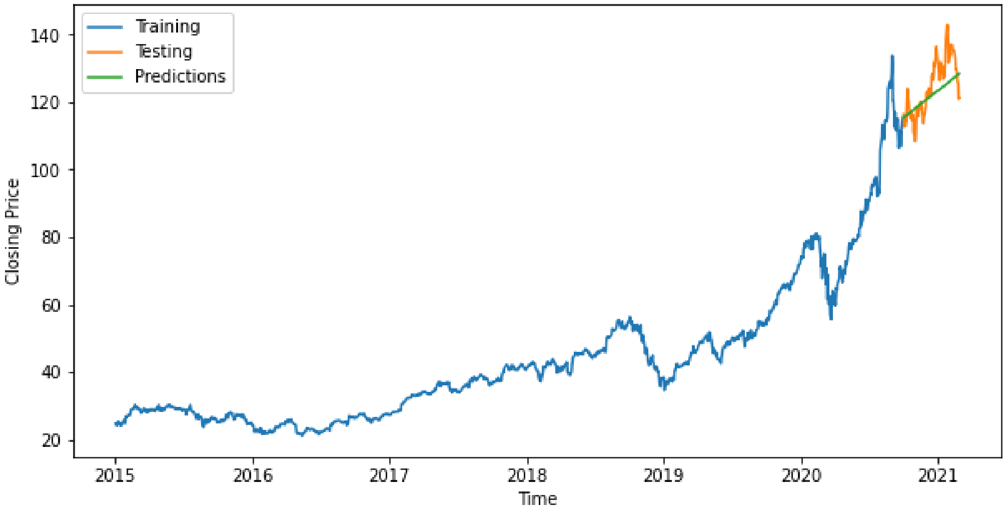


Google

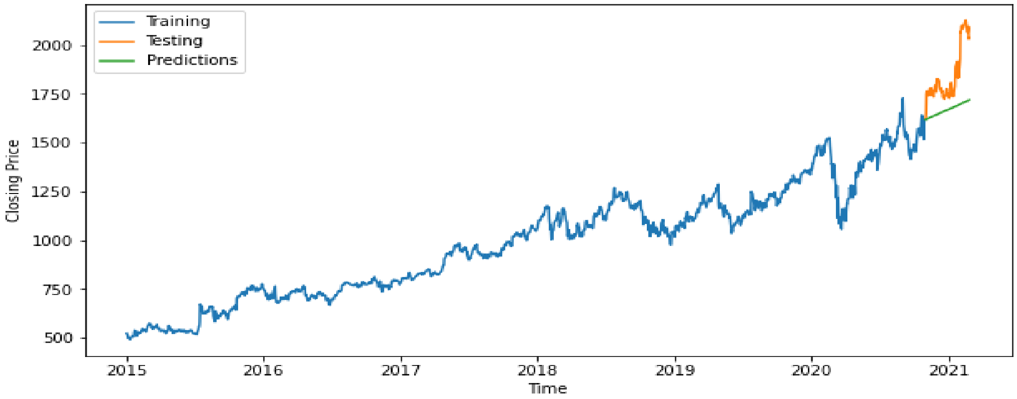


Microsoft

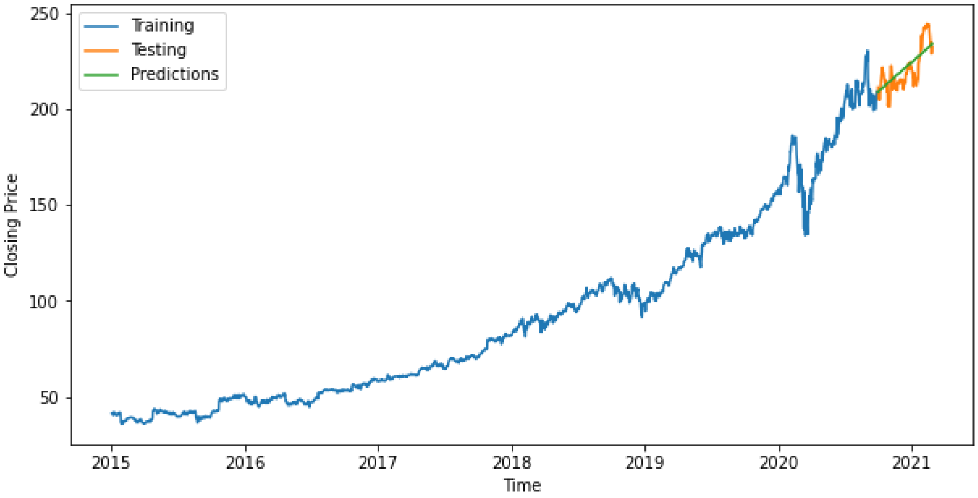
ARIMA



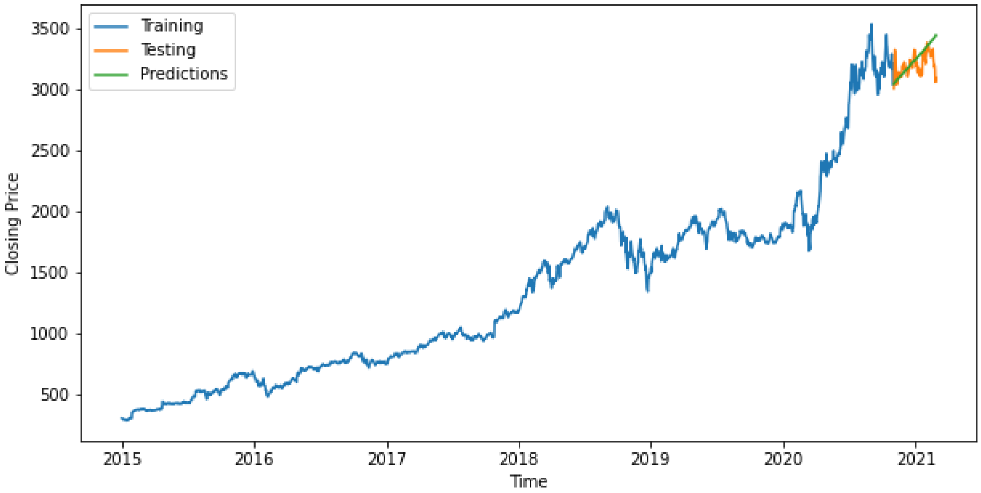
Microsoft



Google

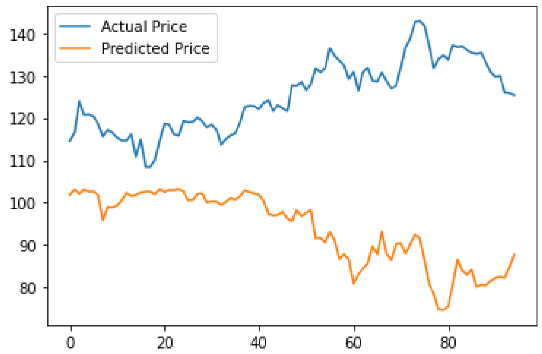


Microsoft

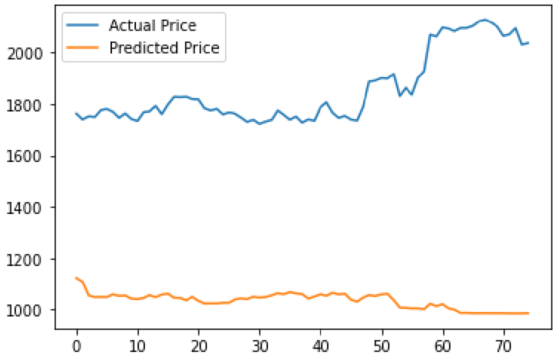


Amazon

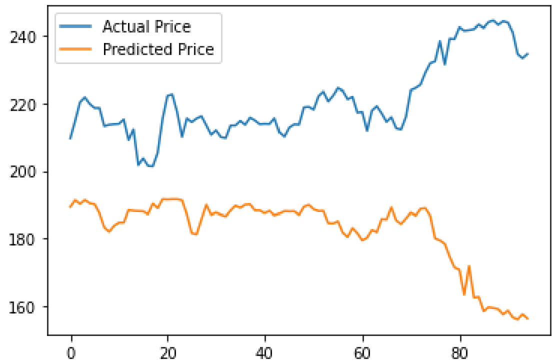
SVR



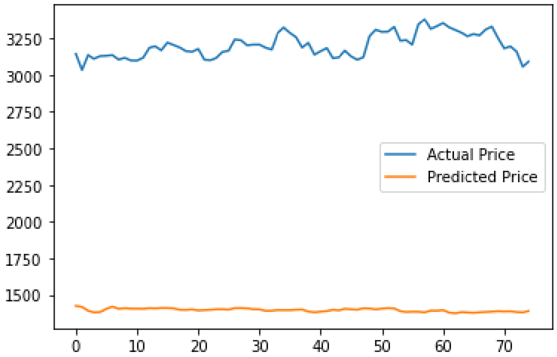
Apple



Google



Microsoft



Amazon

RMSE

	ARIMA Model (RMSE)	Deep Learning Model LSTM (RMSE)	SVR Model (RMSE)
Apple	6.40	4.45	35.18
Google	210.40	71.53	834.48
Microsoft	7.27	6.02	43.58
Amazon	113.18	111.42	1805.78

MAE

	ARIMA Model (MAE)	Deep Learning Model LSTM (MAE)	SVR Model (MAE)
Apple	0.04	0.02	31.41
Google	0.09	0.03	819.21
Microsoft	0.03	0.02	38.49
Amazon	0.02	0.03	1803.78

5 Discussion:

The study's findings have significant implications for investors and traders looking to make informed decisions about buying and selling stocks. The LSTM model's ability to accurately predict future stock prices can help investors make better investment decisions and maximize their returns. Furthermore, the study's exploratory data analysis provided insights into the factors affecting stock prices and the companies' performance, which can help investors identify potential opportunities and risks in the market.

The project explored the performance of different machine learning models for stock price prediction, and we found that LSTM outperformed traditional methods such as ARIMA and SVR. However, there are some limitations to the study. Firstly, we only used data from four technology giants, and the performance of the models may differ for other companies. Secondly, the project only considered the stock prices, and other factors such as news sentiment, macroeconomic factors, and geopolitical events may also affect the stock prices.

In addition, the LSTM model has a complex architecture and requires a large amount of data and computing resources. Therefore, the model may not be suitable for small datasets or low-powered devices. Furthermore, the model may overfit the training data if not tuned properly, and the hyperparameters need to be carefully selected.

Future work can extend the study to include other factors that may affect stock prices, such as news sentiment, macroeconomic factors, and geopolitical events. The study can also be extended to include more companies and compare the performance of different models on a larger dataset. Additionally, more sophisticated deep learning models can be explored, such as convolutional neural networks and transformers, for stock price prediction.

6 Evaluation and Reflection:

The project has successfully achieved its objectives of developing and comparing the performance of different machine learning models in predicting stock prices. The project has utilized various methods such as web scraping, data preprocessing, exploratory data analysis, and machine learning modeling to predict the stock prices of four technology giants.

The LSTM model has outperformed traditional methods such as ARIMA and SVR in terms of RMSE and MAE. The results indicate that the LSTM model is more accurate in predicting future stock prices, making it a promising model for stock price prediction.

However, the project has some limitations. The datasets used in the project are limited to four technology giants, and the analysis and modeling are based only on the information contained in these datasets. Further research could be conducted on other sectors to determine the effectiveness of the models in different industries. Additionally, the project has not considered external factors such as news, social media, and political events that can affect stock prices. Incorporating these external factors into the models can improve their accuracy.

In terms of ethical considerations, the project does not raise any significant ethical concerns. The project has used publicly available data and has not collected any personal data. However, the project's results could potentially be used by investors to make investment decisions, which may have ethical implications if these decisions are not based on ethical considerations. Therefore, it is essential to use the results of the project responsibly and consider the ethical implications of using them.

7 References:

[1] Mehtab, S. (2020, September 20). Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models. ArXiv.Org. https://arxiv.org/abs/2009.10819

[2] Chauhan, N. S. (2020, January). Stock Market Forecasting Using Time Series Analysis. KDnuggets.https://www.kdnuggets.com/2020/01/stock-market-forecasting-time-series-analysis.html

[3] Dev, U. (2020, June 21). EDA of Stock Market using Time Series - Usharbudha Dev. Medium.https://usharbudha-dev09.medium.com/eda-of-stock-market-using-time-series-9662fd18bfc5