**CMT307 Coursework 2 Group Project**

|  |  |
| --- | --- |
| **Group number** | **21** |
| **Project title** | **Stock price prediction** |
| **Supervisor** | Yingying Wu |
| [Student Names]  HuaZhen Chen  Gadigenahalli Umashankar Punith Babu  Ghorpade Shreyas  Karopoulos Charalampos  Nuriyev Solomon  Saini Vaishali  Tang Chun Yuen  Tseloni Eleni Niki | [student IDs]  21041593  21103765  21102662  21087709  21127463  21019303  21042131  21109494 |

1. **Introduction**

**Main goal**: Developing machine learning models to create a profitable trading strategy by predicting a stock price in the future based on historical daily OHLC (Open-high-low-close) price data.

**Contributions**: This paper improves on the current methodology found online for algorithmic trading.

We intend to add to the literature on stock market returns prediction by demonstrating that it is possible to predict stock market returns with high accuracy, even though we recognize the challenges that occur due to the high volatility of the stock market.

The conclusions of the study can be used to guide asset allocation, buy-sell decisions, and the formulation of optimal portfolios, that best fulfil the needs of the investors.

Furthermore, machine learning approaches continue to be useful in determining which predictors, such as those utilized in this research, are essential in determining monthly stock market returns.

For companies studies:

* Stable stocks: Google and Microsoft
* Unstable stocks: Tesla and Tullow oil PLC

1. **Literature Review**

In Machine Learning Techniques for Stock Prediction by Vatsal H.Shah ([1),](#One) he writes how there are stock prediction mythologies. The first is Fundamental Analysis which is the method of analysts making decisions based on the past performance of the company. The second is Technical Analysis which is the method of analysts using time-series analysis.

Time series forecasting is often used to predict stocks. This paper [(2](#Two)) looks at the effectiveness of time series modelling. In this paper, they used an ARIMA model compared to the standard AMRA model. Their findings showed an accuracy of above 85% in predicting stock prices suggesting ARIMA modelling is accurate at predicting stock prices.

As to which model is the best to use comes down to trial and error, there is no one size fits all solution. Currently, deep learning neural networks like CNN, RNN and LTSM are used due to their increased capacity and efficiency compared to linear models. As computers get more powerful it is likely we will see more powerful algorithms being produced.

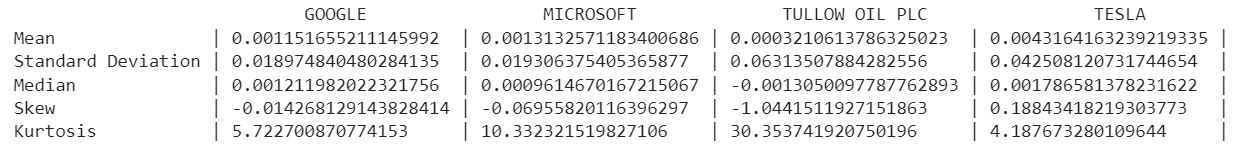
Most models use either data of year (too small) or or more years (too much) to train the model. Considering abnormal stock prices during n-coronavirus pandemic peak, we take data of years to nullify the abnormal trend during mid-. We imported limited data from yfinance only, so the model does not struggle with over fitting. For LSTM, the train-test splitting isas compared to the usual ratio, to optimize the model for under and over fitting.

**3.Descriptive Analysis**

Daily returns are the percentage change of closing price between two days:

|  |  |
| --- | --- |
|  |  |

Statistical observations:



Q-Q Plots:

|  |  |
| --- | --- |
|  |  |
|  |  |

As seen from all the above statistics and plots, the returns are not normally distributed as the quantiles of our dataset (blue curve) are not comparable with the quantiles of a normal distribution (red line) with the same mean and standard deviation.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

For the daily percentage change in volatility of the stocks, we filled in the missing values with 0:

Chart, histogram

Description automatically generated

**Correlation** between stocks:

|  |  |
| --- | --- |
| Correlation between **closing prices** | GOOGL & MSFT were highly correlated. TSLA had a high correlation with GOOGL & MSFT while TLW had a negative price correlation with the other three stocks. |
| Correlation between **volatility of daily returns** | GOOGL & MSFT were significantly positively correlated. TLW & TSLA did not have much correlation with other stocks. |
| Correlation between **daily returns** | GOOGL and MSFT are very positively correlated. TSLA had little positive correlation with GOOGL & MSFT. TLW had a minor positive correlation with the others. |

We investigated the **autocorrelation** of each stock in the same time series, autocorrelation function graphs and partial autocorrelation function graphs were created for:

**A) Daily Closing Prices:**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

The partial autocorrelation graphs imply there were statistically significant closing stock prices for around two hundred days. These graphs showing the closing prices for each stock were highly autocorrelated.

**B) Daily Return**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

For GOOGL & MSFT, there were few significant lags. Which implied daily returns were slightly autocorrelated. For the other stocks, no statistically significant lags in daily return were observed.

**C) Daily volatility**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

The stocks had high positive correlations showing daily volatilities were statistically significant. Partial autocorrelation graphs revealed statistically significant lags before lag one hundred. These graphs showed the daily returns were slightly autocorrelated while the stock daily volatilities were highly autocorrelated for short periods of lags.

**Technical analysis**

Volatile stock prices can be analyzed as a sequence of discrete-time data as their observations are taken daily. Hence why time series forecasting applies well to stock forecasting. We only did technical analysis for google to stay within the word limit.

**Time series analysis**

The adjacent closing price adjusts a stock's closing price to represent its value after price modifications. We chose this strategy to examine past returns and conducted an analysis of previous performance.

Chart

Description automatically generated

The main trend is upward with fluctuations. The stock’s lowest price was and highest price approximately. No patterns were observed for cyclicality of the time series. A slight seasonality trend is spotted as the closing price rises, falls, and rises again in first, second, and third quarters. External factors cause stock prices to fluctuate.

Simple moving average (SMA) to determine the stock’s general trend:

Chart, line chart, histogram

Description automatically generated

The -day SMA crosses below the -day SMA for the first two quarters of indicating that more losses are to follow. Likewise, for , however the trend indicates gains rather than losses.

SMA vs EWMA: EWMA method of assessing volatility focuses more on recent returns. Since it is thought recent times are the strongest predictor of future price movement. Calculating the EWMA with a span value of 36, it is noted the EWMA’s trend line is less volatile and crosses below the Adj close value by more than the SMA does.

Chart, line chart

Description automatically generated

Time Series decomposition:

Chart, line chart

Description automatically generated

With some fluctuations, we observed an upward trend with a clear seasonal component. It is important that a time series is free of trend and seasonality, to be stationary and provide an accurate prediction of the future stock’s price.

We used the Dickey-Fuller test to determine stationarity. The test’s null hypothesis (Ho) states that the series is non-stationary, and the alternative hypothesis (Ha) is that the series is stationary.

**Time-series Error Analysis**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Stock** | **GOOG** | | **ADF Statistic** |  | | **P-value** |  | | -value is , we do not reject the null hypothesis. Therefore, the series is not stationary. |

Consequently, we need to transform the non-stationary time series into stationary time series. Using the method of differencing and computing the differences between successive observations, we made time series stationary by calculating the first difference.

To improve, we also calculated the second difference as shown by graphs:

Chart, arrow

Description automatically generated

Non-Stationary Time Series

Chart, line chart

Description automatically generated

Stationary Time Series

Chart, waterfall chart

Description automatically generated

**Risk Analysis**

Unpredictability and high volatility of the stock market influences the investor to decide accepting the risk.

Daily returns: highest returns were around in the last quarter of and the smallest reached in the first quarter of .

Chart, scatter chart

Description automatically generated

**Calculating Value at Risk:**

As observed, the quantile of daily returns is at . With a confidence interval, Google’s worst daily loss should not exceed .

Chart, histogram

Description automatically generated

**Monte Carlo Analysi**Chart, scatter chart

Description automatically generated**s**

We calculated portfolio losses for each trial and used the aggregation of all simulations, to establish stock risk.

As stock price followed a random walk, we used Markov process which is compatible with the efficient market hypothesis.

The change in stock price is equal to the current stock price multiplied by two terms. The term "drift" is calculated by multiplying the average daily return by the change of time. In each period, the stock will "drift" before being ‘shocked’, causing the stock price to arbitrarily move. We used the Monte Carlo method to repeat this process thousands of times and obtain an approximation of Google’s expected price.

Chart, histogram

Description automatically generatedWe observed the prices range from to and VaR for every investment of . This means, for every initial GOOGL stock, the investor is putting about at risk of the time, according to our Monte Carlo Simulation.

Monte Carlo Simulation, bar chart runs

**4.Methodology Description**

Knowing how risky the stocks are in our data paved our way to create models that predict trends for sound investment.

**4.1 Random Forest Model**

The mean response of the training observations that belong to the same terminal node determines the anticipated response for an observation in a regression tree. Each observation in a classification tree is projected to correspond to the most prevalent class of training observations in the region to which it belongs.

Diagram

Description automatically generated

**4.1.1** **Assumptions for our Random Forest**

Assumptions for a RF classifier that combines all trees for prediction:

* There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
* The predictions from each tree must have exceptionally low correlations.

**Parameters:** For training our random forest regression model, we will use the sklearn package, specifically the RandomForestRegressor function. Many possible parameters for our model are listed in the RandomForestRegressor specification. The following are some of the most critical parameters:  **estimators** - is the number of decision trees to be used in your model. **criterion** - this variable selects the criterion (loss function) that will help to decide the model outputs. Loss function is mean squared error (MSE) by default but can also be mean absolute error (MAE). **max depth** — determines the maximum depth of each tree max features — determines the maximum amount of features the model will evaluate when deciding a split bootstrap — the default value is True, indicating that the model follows bootstrapping principles. **max** **samples** — implies that bootstrapping is set to True; otherwise, it has no effect. This value, in the event of True, determines the largest sample size for each tree.

**4.1.2 Training and implementation:**

Before training all the models, it was taken into consideration the volatility during the covid year where the stocks fell more than and then shot up suddenly after a few months with more than overall.

**4.1.2.1 Feature selection**

To obtain the data set,  and features are selected and declared for the training and testing sets where consists of 'Volume', 'High', 'Low', 'day MA', 'day MA', 'day MA', 'Std\_dev', 'Daily Return in percentage' and consists of closing price.

We chose these features since stock prices were highly auto correlated and volatility and daily return in percentage showed significant relations with closing prices as well. Therefore, we tried to use a few moving averages (MA) with narrower time frames.

**4.1.2.2 Hyperparameter tuning**

The hyperparameters of the random forest model are the model's settings used to either improve the model's prediction capability or make it faster:

Graphical user interface, text, application

Description automatically generated

To minimise the errors, we used random search to find which combination of above hyperparameters gives the best result. We would apply these hyperparameters to the RF model.

# 4.1.2.3 Applying the model

# We chose a random state value to construct the tree based on the number of random states. Overfitting is avoided by producing random subsets of features and using these subsets to build smaller trees. The data was trained with hyperparameter tuning. The following is for Google Stock:

Text

Description automatically generated

# 4.1.3 Statistical metrics and performance evaluation

# Statistical metrics are regression error measures, which we utilised to calculate risks. Model evaluation is crucial, and it must be done to minimise risks and improve model performance.

Since the data is continuous and in time-series, F1, precision and recall are not suitable for measuring errors since they are used for classification models. Therefore, below error measures are used for evaluating the performance of the RF model.

For Google stock:

Graphical user interface, text

Description automatically generated

**4.1.4 Observations:**

We automated the process of filling in missing values in data. Because we employed a rule-based approach, no data normalisation was needed.

However, it necessitated a significant amount of computer power as well as resources because the process constructed several trees and combines their outcomes. It took a long time to train because it uses several decision trees to select the class. It lacked interpretability due to the ensemble of decision trees and failed to evaluate the significance of each variable.

Summarising, the RF model gives good results for stable stocks like Google and MSFT. But even in the MSFT, due to sudden fall and later sudden shooting up of the stock price during the covid year made it difficult to learn for the model and so the predicted price of MSFT is not very accurate. The same trends were observed in the Tesla stock price training and testing model.

**4.2 ARIMA and Auto-ARIMA**

We use ARIMA and Auto-ARIMA models to analyze time series data of the stock market price of a specific company and predict its future price.

AR-I-MA (Auto-Regressive Integrated Moving Average) will relate the present value of series to past values and past prediction errors. It has three hyperparameters – (auto regressive lags), (order of differentiation), (moving avg.) which respectively comes from the AR, I & MA (Moving Average) components. The AR is the correlation between previous & current time periods. “MA” is used to smooth the noise, and the “I” binds the AR & MA together.

Auto-Arima model is just the automated version of ARIMA where it takes the data and fits many models in a different order before comparing the characteristics and returns the best ARIMA model according to either AIC, AICc or BIC value. So, we apply all the steps simultaneously to both, to get the optimum hyperparameters.

**4.2.1 Data Pre-processing**

After importing required library functions, we used .head(), .info() functions to find any missing values in the dataset.

Text

Description automatically generated

Table

Description automatically generated

Here, the data is collected for last years with respect to OHLC, Volume and Adj Close. Date as its index. We plotted the close price to understand the trend of the stock.

Chart, scatter chart

Description automatically generated

The series must be stationary to apply any ML model on a time series. To check the Stationarity of the data we have used following methods:

1. Plotting Rolling mean and standard deviation to check whether the data varies with time.
2. ADCF Test (Augmented Dickey–Fuller test) it gives us the Test Statistics and Critical values. If the test statistics are less than the critical values, the null hypothesis can be rejected, and the series is stationary. The ADCF test also provided the -value.
3. ACF (Auto Correlation Function) & PACF (Partial Auto Correlation Function) graphs are used to find values of and for ARIMA

Chart, bar chart

Description automatically generatedChart

Description automatically generated with low confidence

Chart, line chart

Description automatically generated

From the graph above and augmented Dickey–Fuller (ADF) statistic, used in the test. We obtained a negative number. The smaller it is, the stronger the chances of rejecting the hypothesis.

The value should be as low as possible. Critical values at different confidence intervals should be close to the test statistic value.

As the data is prepared, we will the split the data into a training and testing set which is used to train the model and then evaluate it.

Here we have taken of the data to train and of the data to test the model. Text

Description automatically generated

**4.2.2 Pre-processing**

Before applying any model, a logarithmic transformation was implemented for auto ARIMA and ARIMA models. Logarithmic transformation will help to normalize the data and reduce the variation among them. Descriptive statistics’ observations reveal high variation for Google’s closing price.The code below presents the logarithmic transformation of the data.

Text

Description automatically generated

**4.2.3 Implementation**

ARIMA and auto-ARIMA A basic ARIMA model with and was used to start the search for the best ARIMA model. A greedy method for the hyperparameter tuning was implemented and we found the model with , and returned the best results for the prediction of closing price for Google stock.

We wanted to see if we can improve our prediction and so an auto-Arima model was trained. The Akaike information criterion (AIC) was used as a metric of error for the stepwise search of the auto Arima model. Unfortunately, the results did not improve much after the implementation of the auto Arima model.

**4.2.4 Error Analysis**

So, after employing logarithmic transformation to normalize the data, greedy method for hyperparameter tuning, and Akaike information criterion as a metric of error for search of auto Arima mode, we obtained the following results of the two models:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | MAE | MSE | RMSE |
| ARIMA |  |  |  |
| Auto-ARIMA |  |  |  |

Overall, the ARIMA model with and achieved the best results and predicted the closing prices of google stock with the highest accuracy. However, the auto ARIMA model did not improve the MAE, MSE and RMSE as it can be observed from table above that it performed very poorly for forecasting.

**4.2.5** **Result** **Visualization and Insights**

The following graphs show the results of the ARIMA and auto ARIMA models:

Auto Arima

Chart, line chart

Description automatically generated

|  |  |
| --- | --- |
| Graphical user interface, text, table  Description automatically generated |  |

ARIMA:

Chart, line chart, scatter chart

Description automatically generated

|  |  |
| --- | --- |
| **Graphical user interface, text, application  Description automatically generated with medium confidence** |  |

This model may not be the best fit for long term investments. If the investors target is to make a lot of transactions in brief period, ARIMA is not the best model for this as it is computationally heavy and it will take a long time to produce results.

To conclude, ARIMA model with and did very well in the predictions and is recommended for use of short-term investments.

**4.3 LSTM model**

As the time series analysis is already done, we moved straight to the model:

**4.3.1 Long Short-Term Memory model (LSTM)**

**Long short-term memory** (**LSTM**) is an artificial recurrent neural network (RNN) architecture used for deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (e.g., images), but entire sequences of data such as speech or video inputs.

LSTM cannot process a single data point. It needs a sequence of data for processing and to be able to store historical information. LSTM is an appropriate algorithm to make predictions and process based-on time-series data. It is better used on regression problems. The stock market has enormously historical data that varies with trade date, which is time-series data, so LSTM can be a valuable tool for stock price prediction.

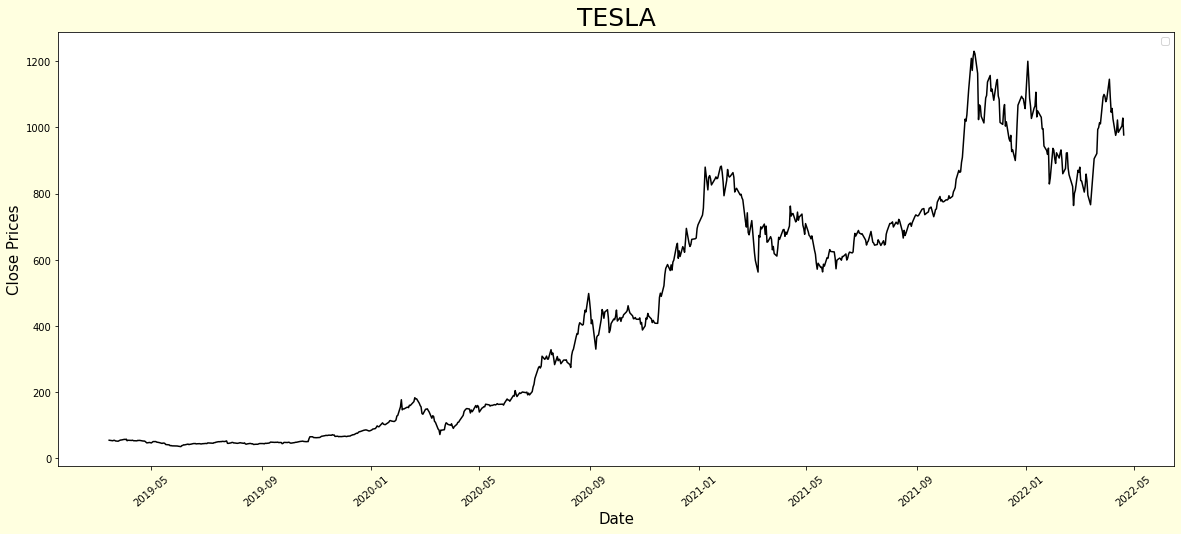
**4.3.2 Data Pre-processing**

The first stage we need to import all necessary libraries. The gathered data set was read using the panda library in python and displayed the records for understanding the data set for pre-processing. Then, we identified the behaviors and characteristics of the data set.

Text

Description automatically generated

We visualized the given data into a graph showing the change of stock prices.



The selected features of distinct types of data needs to be scaled so none dominates the others. Feature scaling and normalizing data are the best way to reduce error rate and improve the accuracy of the model:

Graphical user interface, text, application

Description automatically generated

Normalization is changing the values of numeric columns in the dataset to a common scale, which helps the performance of our model. To scale the training dataset, we use Scikit-Learn’s MinMaxScaler with numbers between zero and one.

Then, we divided the dataset into training and testing set. of the data is used for the training while the remaining of the data is used for testing. Below figure shows the code of splitting data for training.

Graphical user interface, text

Description automatically generated

**4.3.3 Implementation of Algorithm**

Once we finished the pre-processing stage, we moved on to the implementation of LSTM algorithm. We split the training data set into two parts Xtrain and Ytrain. The Xtrain dataset is an independent dataset and Ytrain data set is a dependent data set.

We converted the Xtrain and Ytrain data set into a NumPy array for training the LSTM model. Below is the code used.

A screenshot of a computer

Description automatically generated with medium confidence

LSTM model is a -Dimensional data set [number of samples, time steps, features]. Therefore, we reshaped the Numpy arrays from -Dimensional to -Dimensional.

We created the LSTM model which has two LSTM layers that contain fifty neurons. It also has two dense layers; one layer contains neurons and the other has one neuron. To create a model that has sequential input of the LSTM model, we used Keras library on DNN (Deep Neural Network). The compiled LSTM model used MSE (Mean Squared Error) for the loss function and ‘adam’ for the optimizer:

Graphical user interface, text, application

Description automatically generated

We trained the LSTM model and fit it with a single batch-size containing several training examples. Epochs are another parameter which means the number of iterations in the training model, if the Epoch value is increasing then the accuracy of the model is improving.

Graphical user interface, text, application

Description automatically generated

Here loss refers to the RMS error. We set the value of Epoch to fifty since after testing, the accuracy of this value was reasonable. In future, we plan to change and compare the value of the Epoch to improve this model.

**4.3.4 Error Analysis**

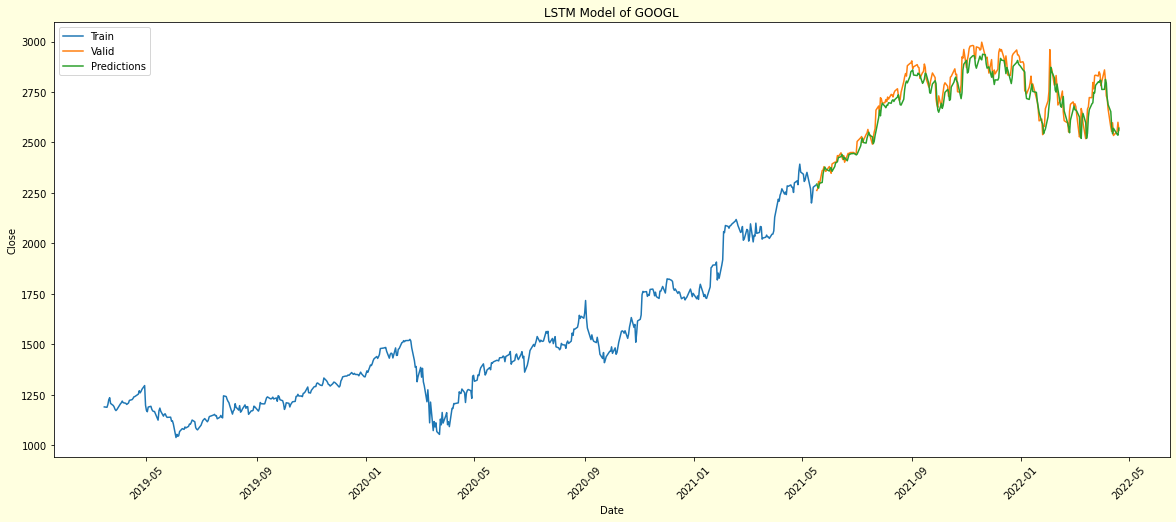
After Feature scaling and normalizing data, and setting epoch parameter to fifty, we achieved the following accuracy of our results is high which is what we expected.

Table

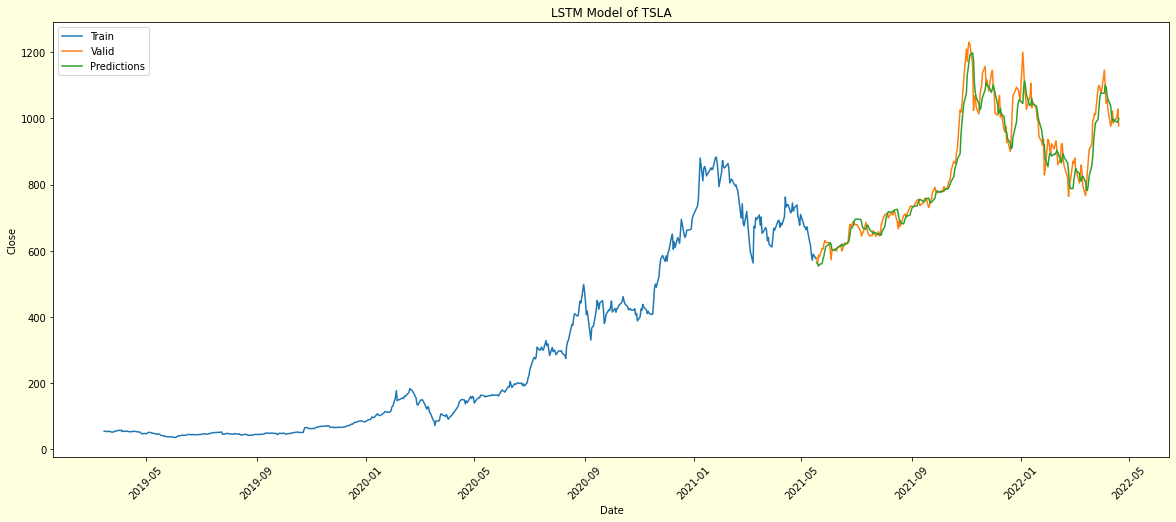
Description automatically generated

**4.3.5 Visualization of Results**

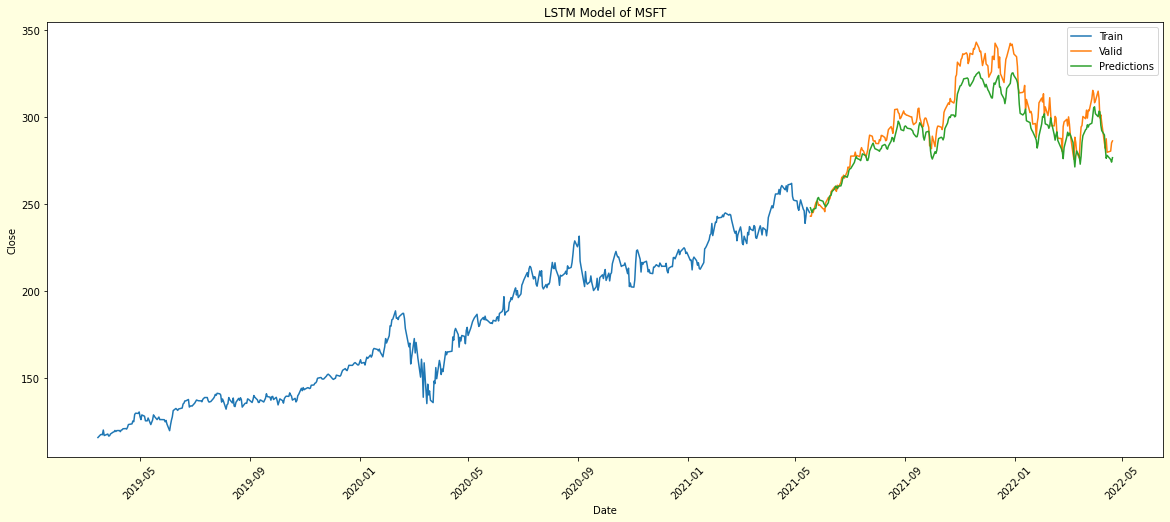
Text

Description automatically generated****

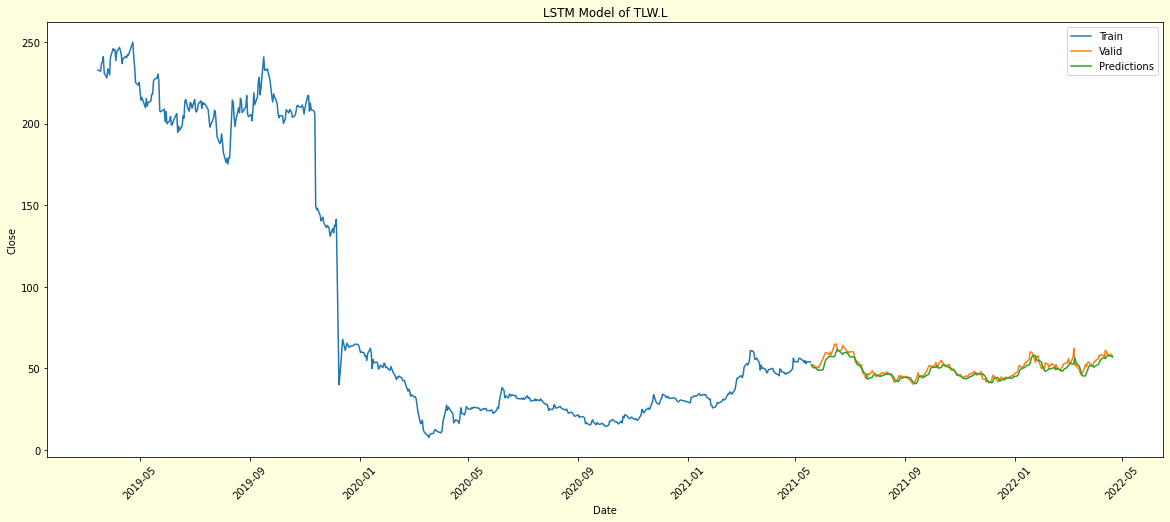
Text, table

Description automatically generated with medium confidence

Text

Description automatically generated

Graphical user interface, text, table

Description automatically generated

**4.3.6 Analysis of results**

From the chart above, we can see the predicted stock prices follow the trend of the real stock prices. This shows the effectiveness of the LSTM to work with the time series or sequential data like the stock prices.

LSTM can be a great tool for stock price prediction. However, this is important to note that the predicted stock prices cannot be used as a solely definitive guide to make an investment decision without further analysis. This is because the prediction is only based on the historical prices movement which is not the only factor that will affect future prices.

While the exact price points from our predicted price was not always close to the actual price, our model still indicated overall trends fairly.

**5.Conclusion**

The current project focused on the prediction of the stock market and more **specifically of Googles stock.** Many different approaches from divergent backgrounds such as machine learning, deep learning, and financial mathematics were used to identify the best model for this forecasting task. Although the models have their strengths and weaknesses, all of them performed well and they returned accurate results.

The Arima model trained in the current dataset achieved a high accuracy on its prediction with low errors and catches the trend and the seasonality of the data. Furthermore, it can be observed that the model can predict accurately the high variability of the data and its predictions are accurate on the turning points which makes the model trustworthy for low-risk investments. On the other hand, the ARIMA model required a set of parameters which must be decided before the training which makes the model hard to set up and time consuming. In a fast-paced environment like the stock market this is a disadvantage because the investors are looking for fast and accurate results. Finally, it is important to mention that the Arima model can be used for short-term investments as it did not do so well with the long-term trends as it did with the short-term ones.

In the random forest model, after observation of the results, it can be said that the model can provide better results mostly for stable stocks providing the long-term trend is sideways or most of the time in the same direction. Whereas, in the LSTM model, the predictions are with better accuracy overall with any stock and trend.

After observation, we can say that the models can be made more accurate if we delete the unusual stock activity, the model could be trained better but at the expense of loss of stock history if it is desired.

In the random forest model, the predictions are sideways for sudden uptrend or downtrend in any stock. This is not the case in the LSTM model. Due to the deep learning and neural network working in the LSTM model, there is an advantage over random forest models. Both models give statistically good accuracy but if we observe the trend graphically, we can conclude that LSTM provides better results.

**Future work**

The future stock price prediction can be effective if we consider predicting future monthly stock prices (short term).  No model can give a perfect prediction for the immediate next day. As prices are directly influential on the company news, the model training cannot have that characteristic in anyways. So human intervention will always be necessary from time to time to train any Machine Learning model.

To conclude, even though the models are not enough for the prediction of the stock prices, all the models performed very well and had accurate results and low errors. The investors should not buy/sell stocks based only on forecasting models. The main reason is that stocks are influenced by numerous factors such as political, economic, environmental, world events and pandemics. For this reason, our suggestion for future work is the creation of a model which can combine time series forecasting and other forms of machine learning algorithms. For example, the analysis of the financial news and twitter related to stocks can give valuable information about the closing prices of a stock. A combination of NLTK (Natural Language Toolkit) and Time series forecasting can predict the closing price of a stock by considering the historical data and extra factors which may influence future prices.

**6. Ethical implications and societal risks (associated with the deployment of machine learning methods):** (3)

1. Privacy invasion: increasing amount of health-related digital data creates challenges for protecting individual privacy
2. Intelligent neurotechnological systems may adversely affect human agency, autonomy, and personal identity
3. Inherent biases in the data structures and ontologies may be replicated or amplified by algorithmic decision-support systems.
4. information asymmetry
5. Automation leads to unemployment.
6. Machine learning algorithms may be “black boxes” where it may become impossible to know why a machine learning algorithm made a decision. (4)
7. AI and machine learning programs could develop behaviour that people would not want them to, such as stopping people from turning them off.

Seeing these tensions, multistakeholder discourse and deliberation are required to ensure effective and responsible development and implementation of these emerging technologies.

Considering the available technical solutions of critical importance to achieve a human-centric AI, we highlight the criticality and urgency to engage multi-disciplinary teams of researchers, practitioners, policy makers, and citizens to co-develop and evaluate in the real-world algorithmic decision-making processes designed to maximize

1. Fairness
2. Accountability and transparency
3. Privacy and data ownership (5)

**References:**

**Reference for Literature review:**

(1) Shah, V. H. (no date) ‘Machine Learning Techniques for Stock Prediction’, in *1library.net.* Available at: [Machine Learning Techniques for Stock Prediction. Vatsal H. Shah (1library.net)](https://1library.net/document/zlgwenx6-machine-learning-techniques-stock-prediction-vatsal-h-shah.html) (Accessed: 22nd April 2022).

(2) Mondal, P., Shit, L. and Goswami, S. (2014) ‘Study of Effectiveness of Time Series Modeling (Arima) in Forecasting Stock Prices’, *International Journal of Computer Science Engineering and Applications,* 4(2), pp. 13-29. DOI: 10.5121/ijcsea.2014.4202.

**Reference for Descriptive Analysis:**

Chan Phooi M’ng, J. and Zainudin, R., 2016. Assessing the Efficacy of Adjustable Moving Averages Using ASEAN-5 Currencies. *PLOS ONE*, 11(8), p.e0160931.

Dupernex, S., 2017. WHY MIGHT SHARE PRICES FOLLOW A RANDOM WALK?. *Student Economic Review, Vol. 21, 2007*,.

Fama, E., 1965. The Behavior of Stock-Market Prices. *The Journal of Business*, 38(1), p.34.

Idrees, S., Alam, M. and Agarwal, P., 2019. A Prediction Approach for Stock Market Volatility Based on Time Series Data. *IEEE Access*, 7, pp.17287-17298.

Ramyar, R., 2006. *City Research Online - Essays on technical analysis in financial markets*. [online] Openaccess.city.ac.uk. Available at: <https://openaccess.city.ac.uk/id/eprint/18945/>

Olivier Guéant, O., 2012. Computing the Value at Risk of a Portfolio: Academic literature and Partitioners’ response. [online] Available at: <https://www.semanticscholar.org/paper/Computing-the-Value-at-Risk-of-a-Portfolio-%3A-and-%E2%80%99-Gu%C3%A9ant/2f11f0371820fb595b8f3394140e5ee4c5ee0ce3>

**ARIMA model references:**

Auhl, P. 2021 “What is an ARIMA Model?”, *Towards Data Science*. Available at [What is an ARIMA Model?. Taking a quick peek into ARIMA modeling | by Miranda Auhl | Towards Data Science](https://towardsdatascience.com/what-is-an-arima-model-9e200f06f9eb) (Accessed 23rd April 2022).

*Penn State Eberly College of Science*, 2022 “Lesson 3: Identifying and Estimating ARIMA models; Using ARIMA models to forecast future values”, *Penn State Eberly College of Science*. Available at [Lesson 3: Identifying and Estimating ARIMA models; Using ARIMA models to forecast future values | STAT 510 (psu.edu)](https://online.stat.psu.edu/stat510/lesson/3) (Accessed 23rd April 2022)

Prasad,E. 2021 “What is Auto-ARIMA?”, *[featurepreneur](https://medium.com/featurepreneur?source=post_page-----b8025c6d732d--------------------------------)*[. Available at](https://medium.com/featurepreneur?source=post_page-----b8025c6d732d--------------------------------) [[What is Auto-ARIMA?. Auto\_arima, a routine from IMSL… | by Eswara Prasad | featurepreneur | Medium](https://medium.com/featurepreneur?source=post_page-----b8025c6d732d--------------------------------)](https://medium.com/featurepreneur/what-is-auto-arima-b8025c6d732d) [(Accessed 23](https://medium.com/featurepreneur?source=post_page-----b8025c6d732d--------------------------------)[rd](https://medium.com/featurepreneur?source=post_page-----b8025c6d732d--------------------------------) [April 2022)](https://medium.com/featurepreneur?source=post_page-----b8025c6d732d--------------------------------)

**References for Random Forest:**

Meltzer, R. 2021 “What Is Random Forest?”, *CAREERFOUNDRY.* Available at [What is Random Forest? [Beginner's Guide + Examples] (careerfoundry.com)](https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/#:~:text=When%20using%20Random%20Forest%20for%20regression%2C%20the%20forest,that%20make%20up%20the%20larger%20Random%20Forest%20model) (Accessed 23rd April 2022)

Beheshti, N. 2022 “Random Forest Regression”, *Towards Data Science.* Available at [Random Forest Regression. A basic explanation and use case in 7… | by Nima Beheshti | Mar, 2022 | Towards Data Science](https://towardsdatascience.com/random-forest-regression-5f605132d19d) (Accessed 23rd April 2022)

Zakariya, M. 2021 “Predict Stock Prices Using Random Forest Regression Model in Python”, *Medium.* Available at " [Predicting Stock Prices Using Random Forest Model | Medium](https://medium.com/@maryamuzakariya/project-predict-stock-prices-using-random-forest-regression-model-in-python-fbe4edf01664#:~:text=The%20random%20forest%20regression%20model,based%20on%20the%20predicted%20values)” (Accessed 23rd April 2022)

Singh, A. 2018 “Stock Prices Prediction Using Machine Learning and Deep Learning Techniques (with Python codes)”, *Analytics* *Vidhya.* Available at [Stock Price Prediction Using Machine Learning | Deep Learning (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/) (Accessed 23rd April 2022)

**Reference for LSTM:**

Loukas, S. 2020 “Time-Series Forecasting: Predicting Stock Prices Using an LSTM Model”, *Towards Data Science.* Available at [Time-Series Forecasting: Predicting Stock Prices Using An LSTM Model | by Serafeim Loukas | Towards Data Science](https://towardsdatascience.com/time-series-forecasting-predicting-stock-prices-using-an-lstm-model-d0056cd5e055) (Accessed 23rd April 2022)

**References for Ethical implications:**

Kellyer, P. 2020 “Chapter 18 - Ethical issues in the application of machine learning to brain disorders”, *ScienceDirect.* Available at [Ethical issues in the application of machine learning to brain disorders - ScienceDirect](https://www.sciencedirect.com/science/article/pii/B9780128157398000183) (Accessed 23rd April 2022(

Delony, D. 2022 “What are some ethical issues regarding machine learning?”, *Techopedia*. Available at [What are some ethical issues regarding machine learning? (techopedia.com)](https://www.techopedia.com/what-are-some-ethical-issues-regarding-machine-learning/7/33376) (Accessed on 23rd April)

Pentland, A., Oliver, N., and Lepri, O. 2021 “Ethical machines: The human-centric use of artificial intelligence”, *iScience*, 24(3). Available at [Ethical machines: The human-centric use of artificial intelligence - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7973859/) (Accessed 23rd April 2022)