**HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY  
SCHOOL OF ELECTRICAL & ELECTRONIC ENGINEERING**

**Ảnh có chứa biểu tượng

Mô tả được tạo tự động**

**AC4110E – Data Analysis and Visualization**

**FINAL PROJECT**

|  |  |
| --- | --- |
| **Instructor** | Assoc. Prof. Le Thi Lan |
| **Course** | AC4110E |
| **Class** | 149177 |
| **Group** | 3 |
|  |  |

**HÀ NỘI – 2024**

**final project’s report of DATA ANALYSIS AND VISUALIZATION**

Contents

[1 Introduction 3](#_Toc171189953)

[1.1 What is the stock market 3](#_Toc171189954)

[1.2 Stock market analysis 3](#_Toc171189955)

[2 Goal and objectives 3](#_Toc171189956)

[3 Relevant works 4](#_Toc171189957)

[4 Proposed solution 4](#_Toc171189958)

[4.1 LSTM model 4](#_Toc171189959)

[4.2 ARIMA model 5](#_Toc171189960)

[4.3 Support Vector Machine 6](#_Toc171189961)

[5 Dataset 6](#_Toc171189962)

[6 Data preprocessing 7](#_Toc171189963)

[7 Data visualization 8](#_Toc171189964)

[8 Analysis of results and discussion 11](#_Toc171189965)

[9 Conclusion 13](#_Toc171189966)

|  |  |  |
| --- | --- | --- |
| **Name** | **Student ID** | **Roles** |
| Ma Khoa Hoc | 20210388 | Using LSTM |
| Ngo Quang Tu | 20213593 | Using SVM regression |
| Phung Minh Chien | 20213565 | Using ARIMA |

**MEMBERS AND ROLES**

# Introduction

## What is the stock market

Stock market is the aggregation of buyers and sellers of stocks (also called shares), which represent ownership claims on businesses. These may include *securities* listed on a public stock exchange as well as stock that is only traded privately, such as shares of private companies that are sold to investors through equity crowdfunding platforms.

Understanding the stock market is a key component of financial literacy. It empowers individuals to make informed decisions about their investments, leading to financial independence and the ability to build wealth over time.

## Stock market analysis

Stock Market is considered to be one of the most highly complex financial systems which consists of various components or stocks, the price of which fluctuates greatly with respect to time. Stock markets being highly sensitive and susceptible to quick changes, thus the main aim of stock trend prediction is to develop new innovative approaches to foresee the stocks that result in high profits.

Predicting stock market trends is a critical and highly regarded task, as accurately forecasting stock prices can result in significant profits through well-informed decisions. However, stock market prediction poses a substantial challenge due to the non-stationary, noisy and chaotic of the data, making it difficult for investors to make profitable investments.

# Goal and objectives

The primary goal of this project is to identify opportunities that offer the potential for high returns on investment. The application of machine learning techniques in stock market analysis has garnered significant attention due to its potential to predict stock prices with considerable accuracy. It highlights the importance of selecting appropriate features and model architectures in financial time series prediction and suggests that machine learning can be a valuable tool for investors and financial analysts aiming to forecast stock market movements. Utilizing all the features that related to the stock of each company to predict the stock market trend, however, seems to be out of reach. Thus, our approach to the stock market trend prediction is to use the close price of the stock of previous days for predicting the close price of the future, which can be use to analyze the trend of stock market.

# Relevant works

Some relevant works to the subject are:

1. Sheikh Mohammad Idrees, M. Afshar Alam, Parul Agarwal. A Prediction Approach for Stock Market Volatility Based on Time Series Data. DOI: 10.1109/ACCESS.2019.2895252
2. Hirotaka Mizuno, Michitaka Kosaka, Hiroshi Yajima. Application Of Neural Network To Technical Analysis Of Stock Market Prediction
3. Pramod B S, Mallikarjuna Shastry P. M., Stock Price Prediction Using LSTM

These researches use different methods with different types of dataset in predicting the stock market trend in the future.  In each paper, they focus on applying machine learning models to get the best accuracy in each method they choose. In article 2, the results showed that the ARIMA model had quite good results in predicting Nokia Stock Price and Zenith Bank Stock Price. In study 3, the author predicted TATAMOTORS stock price in future based on its historical data using the LSTM model. This article used the LSTM model with 60 timestamps and 1 output to predict the open price for 300 days and achieved very minimal loss. The use of time series data in each research has shown that utilizing advanced model such as LSTM, ARIMA, … can give us an effective way to handle time series data of the stocks to predict the fluctuations of stock market in some specific regions.

# Proposed solution

In this project, we perform the following tasks:

* We will predict stock prices and compare model performance using 3 machine learning models: LSTM, SVM and ARIMA.
* For LSTM, we will compare the model's performance when performing training with and without scaled data, compare the model's performance when the number of epochs increases, and consider whether changing time steps(60 steps,90 steps,120 steps) has affected the accuracy of prediction or not.

## LSTM model

Long Short-Term Memory (LSTM) networks are a special type of Recurrent Neural Networks (RNN) capable of learning long-term dependencies. They are designed to avoid the long-term dependency problem that is often encountered in standard RNNs, which makes them particularly useful for sequential data tasks such as time series prediction, language modeling, and speech recognition.

Memory Cell: The core idea behind LSTM is the memory cell, which maintains its state over time. This allows LSTMs to retain information for long periods, making them suitable for tasks where the context or the sequence of information is crucial.

Gates:

* Forget Gate: Decides what information should be discarded from the cell state.
* Input Gate: Determines which new information is to be added to the cell state.
* Output Gate: Controls the output based on the cell state and the input.

In this project, we use a sequential model with two LSTM layers and two Dense layers for time series prediction. The first LSTM layer has 128 units and returns sequences, while the second LSTM layer has 64 units and does not return sequences. This is followed by a Dense layer with 25 units and a final Dense output layer with 1 unit. The model is compiled using the Adam optimizer and Mean Squared Error (MSE) as the loss function.

## ARIMA model

ARIMA model, stands for AutoRegressive Integrated Moving Average, is a popular statistical method used for time series forecasting, which is totally suitable for the stock market data.

ARIMA model is the combination of 3 key components**: AutoRegressive (AR), Integrated (I) Part and Moving Average (MA) Part.** AR involves regressing the variable on its own lagged (prior) values. It specifies how the past values influence the current value. I part involves differencing the data to make it stationary, which means the mean and variance are constant over time. The order of differencing required to achieve stationarity is represented by 'I'. MA part models the error of the model as a linear combination of error terms occurring at various times in the past.

An ARIMA model is usually denoted as ARIMA (p, d, q), where:

* p is the number of lag observations included in the model (order of the AR part),
* d is the number of times the raw observations are differenced (order of differencing),
* q is the size of the moving average window (order of the MA part).

For evaluating an optimized order for the model, we use the Akaike Information Criterion (AIC). It is a widely used metric in the field of statistical modeling and model selection. Developed by the renowned Japanese statistician Hirotugu Akaike in 1973, AIC provides a method for comparing different statistical models and selecting the one that best balances the trade-off between model complexity and goodness of fit to the observed data.

The mathematical formulation of AIC is given by:

AIC = 2k − 2ln(L)

where k represents the number of parameters in the model, and L denotes the maximum likelihood estimate of the model.

## Support Vector Machine

**SVM** is a supervised learning model that analyzes data for classification and regression tasks. In stock market analysis, SVMs are used for predicting stock prices based on historical data.

SVM searches for an optimal super flat to separate data into different layers so that the distance from super flat to the nearest points of each class (margin) is the largest. In case the data cannot be linearly separated, SVM uses kernel functions to map data into higher space, where they can be linearly separated.

* Margin: The distance between super flat and the nearest data points.
* Support Vectors: The nearest data points to super flat, deciding the position of super flat.

Kernel function is a technique that helps SVM process nonlinear data by mapping the original data into a higher specific space.

We choose the radial basis function (RBF),the mathematical formulation of this function is given by :



# Dataset

The StockNet dataset (https://github.com/yumoxu/stocknet-dataset.git) is crafted for stock movement prediction using a combination of tweet sentiment and historical stock prices. It spans from January 1, 2014, to January 1, 2016, covering 88 stocks across various sectors.

In this project, we use 5 historical stock price data from 5 different companies: Apple Inc. (AAPL), ABB Ltd (ABB), AbbVie Inc. (ABBV), Amgen Inc. (AMGN), American Electric Power Company Inc. (AEP) from the StockNet dataset.

For each company’s stock we have these entries:

A screenshot of a computer

Description automatically generated

* **Date**: the date of the trading day (usually formatted as YYYY-MM-DD).
* **Open**: the price of the stock at the beginning of the trading day.
* **High**: the highest price of the stock during the trading day.
* **Low**: the lowest price of the stock during the trading day.
* **Volume**: the total number of shares traded during the trading day.
* **Adj Close**: the adjusted closing price, which accounts for corporate actions like dividends, stock splits, and new stock offerings.
* **Close**: the last price at which the stock is traded during the regular trading day. A stock’s closing price is the standard benchmark used by investors to track its performance over time.

# Data preprocessing

Firstly, because our data is a time series data, we cannot just drop the missing data in any row. On the other hand, we only use previous day’s close price to predict close price of the current day, so dropping the column that contains missing data is also not suitable. Thus, the best way to handle missing data in our project is to replace the missing value by another value, in particular our choice is the means of all close price values.

Secondly, the data does seem to have some outliers due to our check function that has been implemented in Project\_data\_visualize.ipynb file. However, because the influence on the close price comes not only from the factors that are listed in the table but also from many other factors in the real stock market and outside, we still retain the outliers to represent those external influences.

# Data visualization

For each company, we plot the historical view of the closing price in *Figure 1*

A graph of stock market growth

Description automatically generated with medium confidence

*Figure 1: Historical view of the closing price*

A group of graphs showing different types of graphs

Description automatically generated with medium confidence

*Figure 2: Moving average of closing price for 20, 50, 100 days*

By adding moving average of closing price for 20, 50, 100 days we have the plot in figure 2. We see in the graph that the best values to measure the moving average are 10 and 20 days because we still capture trends in the data without noise.

Next, we analyze the daily changes of the stock by drawing scatter plot (*Figure 3*).

A group of blue dots

Description automatically generated

*Figure 3: Average daily return of the stock in scatter plot*

We get an overall look at the average daily return using a histogram *(Figure 4).*

A group of blue graphs

Description automatically generated

*Figure 4: Average daily return of the stock in histogram*

For visualizing the correlation between different stocks closing prices, we combine the histogram with scatter and KDE plot *(Figure 5)*

A group of blue dots

Description automatically generated A screenshot of a graph

Description automatically generated

*Figure 5: The correlation between different stocks closing prices*

Finally, we plot the correlation heat map to get actual numerical values for the correlation between the stocks' daily return values *(Figure 6)*

A chart of stock closing price

Description automatically generated

*Figure 6: Correlation between the stocks' closing prices in heatmap*

# Analysis of results and discussion

After performing training and testing the LSTM model, we obtained the results of RMSE as shown in the table below

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Scale or not | Time steps | AAPL | ABB | ABBV | AMGN | AEP |
| 5 epochs | Scale data | 60 | 2.09 | 0.24 | 0.89 | 7.68 | 0.57 |
| 90 | 3.14 | 0.48 | 0.68 | 2.45 | 0.53 |
| 120 | 2.97 | 0.49 | 1.16 | 2.69 | 0.48 |
| Not scale data | 60 | 2.96 | 0.26 | 2.21 | 3.63 | 1.26 |
| 90 | 3.43 | 0.26 | 2.48 | 2.81 | 0.89 |
| 120 | 3.66 | 0.31 | 2.42 | 2.25 | 2.69 |
| 10 epochs | Scale data | 60 | 2.32 | 0.26 | 0.84 | 2.30 | 0.93 |
| 90 | 2.22 | 0.22 | 1.02 | 4.75 | 1.13 |
| 120 | 2.03 | 0.40 | 0.82 | 1.95 | 0.55 |
| Range of price |  |  | Min:55.79  Max:164.05 | Min:16.06  Max:27.09 | Min:33.71  Max:75.42 | Min:81.36  Max:182.60 | Min:40.96  Max:74.10 |

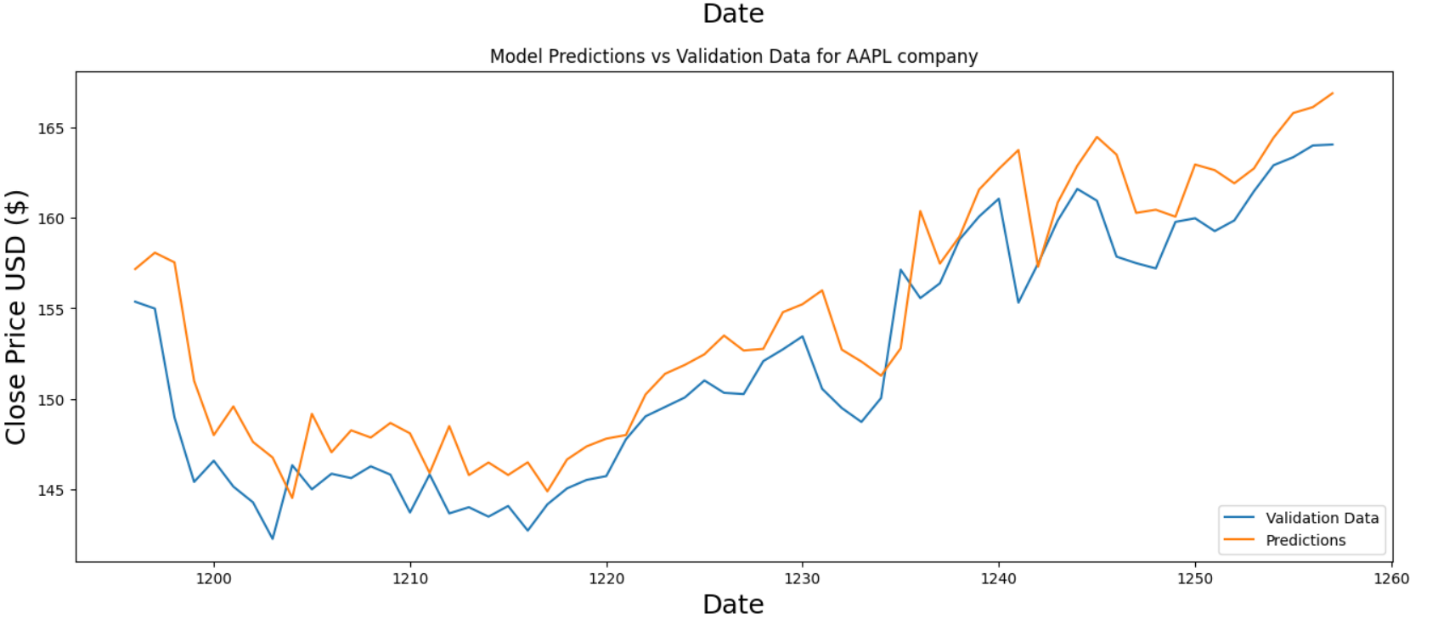
*Table 1: RMSE of LSTM model*

* **About epochs:** when increasing the number of epochs, RMSE also improves quite significantly(*Table 1*).
* **About scaling data:** when data is scaled, the results are clearly improved in terms of both RMSE and the fit of the graph between the predicted value and the actual value (*Figure 7).*

A graph with blue and orange lines

Description automatically generated

*Figure 7.a: Prediction values and actual values of AAPL when not scaling*



*Figure 7.b: Prediction values and actual values of AAPL when scaling*

* **About time steps (60,90,120):** When time step increases or decreases, we do not see an improvement in RMSE. Therefore, time step cannot be used to evaluate the performance quality of the LSTM model.

**Discussion all 3 models:** The two models, SVM and ARIMA, predict quite accurately with a fairly small RMSE(*Table 2*), and the graph of the predicted values relatively matches the actual value line(*Figure 8*). This demonstrates that these two models are performing well in the problem of predicting stock prices. As for the LSTM model, the RMSE of LSTM is not too high and is within an acceptable value threshold. However, in terms of the shape of the graph, the predicted values line of LSTM do not closely match the real values line(*Figure 9*). Therefore, we see that the effectiveness of this model in predicting the stock prices of these five companies is not very high.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | AAPL | ABB | ABBV | AMGN | AEP |
| SVM | 2.21 | 0.25 | 1.00 | 2.14 | 0.57 |
| ARIMA | 2.02 | 0.23 | 0.64 | 1.90 | 0.45 |

*Table 2: RMSE of SVM and ARIMA model*

A graph with lines and numbers

Description automatically generated with medium confidence

*Figure 8.a: Prediction values and actual values of AAPL of SVM*

A graph showing the growth of a stock market

Description automatically generated

*Figure 8.b: Prediction values and actual values of ABB of SVM*

A graph with blue and orange lines

Description automatically generated

*Figure 8.c: Prediction values and actual values of AAPL of ARIMA*

A graph with blue and orange lines

Description automatically generated

*Figure 8.d: Prediction values and actual values of ABB of ARIMA*

A graph of a line

Description automatically generated with medium confidence

*Figure 9.a: Prediction values and actual values of AAPL of LSTM(120 time steps)*

A graph of a line

Description automatically generated with medium confidence

*Figure 9.a: Prediction values and actual values of ABB of LSTM(120 time steps)*

# Conclusion

In conclusion, our research into stock price prediction models highlights three main points. Firstly, both SVM and ARIMA models demonstrate high accuracy and reliability, as evidenced by their low RMSE and close alignment with actual stock prices. This indicates their strong potential for effective use in financial forecasting. Secondly, while the LSTM model's RMSE remains within an acceptable range, its predictive performance is less reliable, as shown by the greater distance between predicted and actual values. Finally, these findings suggest that traditional models like SVM and ARIMA may currently offer more dependable results for stock price prediction compared to more complex deep learning models like LSTM.

To sum up, these results support the long-term goal of improving prediction models to increase their accuracy and applicability in financial markets. By understanding the strengths and limitations of each model, we can better tailor our approaches to leverage the most effective techniques for reliable stock price forecasting, contributing to making more informed decisions in financial investment.