|  |
| --- |
| STOCK MARKET PREDICTION |

|  |
| --- |
| (DSCI 6007-02 Distributed and Scalable Data Engineering)  Group - 9 |

|  |  |
| --- | --- |
| SPRING 23 |  |



Contents

[Project Name 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Technical Report 3](#_Toc96341551)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

[Submitted on: 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

[Abstract 4](#_Toc96341554)

**Descriptive Static Dataset………………..5**

[Methodology 9](#_Toc96341555)

visualization [10](#_Toc96341556)

Result ……………………………………14

[Discussion 13](#_Toc96341557)

[Conclusion 15](#_Toc96341558)

[Contributions/References 16](#_Toc96341559)

|  |
| --- |
| STOCK MARKET PREDICTION |

|  |  |  |
| --- | --- | --- |
| Executive Summary  * The fluctuation of the stock market is violent and there are many complicated financial indicators. * However, the advancement in technology provides an opportunity to gain steady fortune from the stock market and can help experts to find out the most informative indicators to make better predictions. * The prediction of the market value is of paramount importance to help in maximizing the profit of stock option purchases while keeping the risk low.   The next section of the paper will be methodology where we will explain about each process in detail. After that we will have pictorial representations of the analysis that we have made and we will also reason about the results achieved. Finally, we will define the scope of the project. We will talk about how to extend the paper to achieve better results. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **BAHAREH ARGHAVANI NOBAR**  **DEVNATH REDDY MOTATI**  **VIJITHA PYETA GOUDA KAKE**  **NAGA VENKATA SAI TEJA KEERTY** | **Questions?**  Contact: dmota1@unh.newhaven.edu |  |

|  |  |
| --- | --- |
| **Title of Project**  **Stock Market Prediction**  *Github link :*[*Https://GitHub.com/orgs/DSCI-6007-02-group-09*](https://github.com/orgs/DSCI-6007-02-group-09)  *Python notebook:* [*https://colab.research.google.com/drive/1MWrzg5yY1B5guLK0VthcZAN1NqNE0lRr?usp=sharing*](https://colab.research.google.com/drive/1MWrzg5yY1B5guLK0VthcZAN1NqNE0lRr?usp=sharing) |  |
| Highlights of Project this project is a basic stock analysis using Python and can be extended to more complex and detailed analysis.  The project reads stock data from yahoo and calculates daily returns for the selected company's stock. It then plots the daily returns on average and closing price of the selected company during the specified date range. Additionally, it also plots the total volume of stock trade for the selected company. Submitted on: 30 april 2023 |

|  |
| --- |
| Technical Report |

## Abstract

Will see how you can use a time-series model known as Long Short-Term Memory. LSTM models are powerful, especially for retaining a long-term memory, by design, as you will see later. You'll tackle the following topics in this report.

* Understand why you would need to be able to predict stock price movements.
* Download the data – You will be using stock market data gathered from Yahoo finance.
* Split train-test data and perform some data normalization.
* Go over and apply a few averaging techniques that can be used for one-step ahead predictions.
* Motivate and briefly discuss an LSTM model as it allows us to predict more than one step ahead.
* Predict and visualize future stock markets with current data.

**Descriptive Static Dataset**

The first step is to get the data and load it to memory. We will get our stock data from the Yahoo Finance website. Yahoo Finance is a rich resource of financial market data and tools to find compelling investments. To get the data from Yahoo Finance, we will be using yfinance library which offers a threaded and Pythonic way to download market data from Yahoo.

## Reading data:

final\_dateset = pd.read\_csv('data.csv')

final\_dataset.head()

Table

Description automatically generated

## Dataset Exploration:

## The dataset contains many columns associated with time series like the date and the different variables like close, high, low and volume. We will use opening and closing values for our experimentation of time series with LSTM.

## Table Description automatically generated

After this will filter the DataFrame based on the selected company name from requested date range in yyyy-mm-dd format.

Data Encoding

* Data encoding refers to the process of converting data from one form to another for the purpose of transmission or storage. This is necessary because different systems and devices may use different methods to represent and store data.
* The most common forms of data encoding include:

1. **ASCII Encoding:** This involves encoding text characters using a 7-bit binary code. It is widely used for representing text in computer systems.
2. **Unicode Encoding:** This is a standard for representing characters from different languages and scripts using a single encoding system. It supports more than 100,000 characters from over 100 scripts, including Latin, Greek, Cyrillic, Chinese, Japanese, and Korean.
3. **Binary Encoding**: This involves representing data using a sequence of 0's and 1's. It is used extensively in computer systems and is the basis for all digital communication.
4. **Base64 Encoding:** This is a method of encoding binary data using only printable ASCII characters. It is often used to encode data for transmission over email or other text-based communication channels.
5. **Huffman Encoding:** This is a form of variable-length data compression that assigns shorter codes to more frequently occurring data values. It is used to compress data for efficient storage and transmission.

Descriptive Statistics about the data:

.describe() generates descriptive statistics. Descriptive statistics include those that summarize the central tendency, dispersion, and shape of a dataset’s distribution, excluding NaN values.

Analyzes both numeric and object series, as well as DataFrame column sets of mixed data types. The output will vary depending on what is provided. Refer to the notes below for more detail.

A picture containing graphical user interface

Description automatically generated

Information about the data:

.info() method prints information about a DataFrame including the index dtype and columns, non-null values, and memory usage.

Graphical user interface, text

Description automatically generated with medium confidence

What was the daily return of the stock on average?

Here we're now going to analyze the risk of the stock. In order to do so we'll need to take a closer look at the daily changes of the stock, and not just its absolute value. Let's go ahead and use pandas

Graphical user interface, text

Description automatically generated

Plot Closing Price

The closing price is 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.

Chart, scatter chart

Description automatically generated

Methodology:

Long Short-Term Memory models are extremely powerful time-series models. They can predict an arbitrary number of steps into the future. An LSTM module (or cell) has 5 essential components which allows it to model both long-term and short-term data.

Cell state (ct) - This represents the internal memory of the cell which stores both short term memory and long-term memories

Hidden state (ht) - This is output state information calculated w.r.t. current input, previous hidden state and current cell input which you eventually use to predict the future stock market prices. Additionally, the hidden state can decide to only retrive the short or long-term or both types of memory stored in the cell state to make the next prediction.

Input gate (it) - Decides how much information from current input flows to the cell state

Forget gate (ft) - Decides how much information from the current input and the previous cell state flows into the current cell state

Output gate (ot) - Decides how much information from the current cell state flows into the hidden state, so that if needed LSTM can only pick the long-term memories or short-term memories and long-term memories.

Data Visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

****Data Preprocessing**:**

Data Preprocessing is the process of transforming or encoding data so that it may be easily parsed by the machine. In other words, the algorithm can now easily interpret the data's features.

Scaling the data

Graphical user interface, text, application

Description automatically generated

Now split the data into training and test for building a model.

Text

Description automatically generated

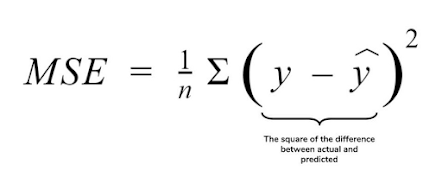
****Modeling**:**

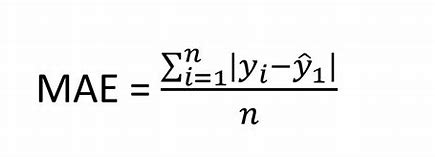
We used an LSTM neural network architecture, which may consist of one or more LSTM layers, followed by Dense layers and a final output layer. The manipulation of time series data is crucial for any data analyst or data scientist as it involves a series of data points ordered in time. This notebook focuses on exploring and analyzing data from the stock market, specifically technology stocks like Apple, Amazon, Google, and Microsoft. The yfinance library is used to obtain stock information, while Seaborn and Matplotlib are utilized to visualize different aspects of the data. In addition, we examine several methods of analyzing a stock's risk based on its past performance. Finally, we apply a Long Short-Term Memory (LSTM) model to forecast future stock prices.

****Evaluation**:**

We evaluate the model's performance using appropriate metrics, such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE), to measure how well the model is predicting the stock prices on

the validation set.





****Deployment**:**

**To deploy the above program, we used following**

**Choose a cloud service provider, such as Amazon Web Services (AWS), Microsoft Azure and created an account.**

**Created a virtual machine (VM) instance on the cloud provider's platform, and ensured that it meets the requirements for running the program (e.g., sufficient RAM, CPU, and disk space).**

**Installed the necessary software and packages, such as Python, Flask, and the required libraries for data manipulation and visualization (e.g., pandas, numpy, seaborn, matplotlib).**

**Uploaded the program files to the VM instance, and set up the environment variables and configuration files as needed.**

**Ran the Flask application, and verified that it is accessible and functional by navigating to the assigned IP address and port.**

****Result**:**

Graphical user interface, text

Description automatically generated

Chart

Description automatically generated

## Conclusion

Based on the LSTM model's predictions and the evaluation metrics, we can conclude the following about the stock price prediction project:

1. The LSTM model can be used to predict future stock prices with reasonable accuracy.
2. The model's accuracy can be improved by tweaking the hyperparameters, adding more data, or using a different model architecture.
3. The performance of the model may vary depending on the stock being predicted, as some stocks may be more volatile or influenced by external factors than others.
4. The predictions can be used to make informed investment decisions, but it is important to keep in mind that stock prices are subject to fluctuations and unexpected events that can affect their performance.

Overall, the project demonstrates the potential of using deep learning techniques like LSTM models for stock price prediction and highlights the importance of continuous monitoring and adjustment of the model to improve its accuracy over time.

## Contributions/References

<https://www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm>

[Python LSTM (Long Short-Term Memory Network) for Stock Predictions | DataCamp](https://www.datacamp.com/tutorial/lstm-python-stock-market)

[Stock market outlook for 2023: Here’s what experts predict amid recession fears (yahoo.com)](https://finance.yahoo.com/news/stock-market-outlook-2023-experts-141100877.html)