Visualizing and Prediction of Stocks using Machine Learning

A Project Report

Submitted in partial fulfillment of the

Requirements for the award of the Degree of

BACHELOR OF SCIENCE (Computer Science)

By

Kumawat Vicky Subhash Chand (As per 12th Mark sheet)

Seat Number:- 2019016401103745

Under the esteemed guidance of

Asst. Prof. Rutika Malik

Course Instructor

DEPARTMENT OF COMPUTER SCIENCE

SHREE L.R. TIWARI DEGREE COLLEGE

(Affiliated to University of Mumbai)

THANE, 401107

MAHARASHTRA

2021-2022

SHREE L.R. TIWARI DEGREE COLLEGE

(Affiliated to University of Mumbai)

THANE-MAHARASHTRA-401107

DEPARTMENT OF COMPUTER SCIENCE

PIONEERING EDUCATION
SINCE 1992

CERTIFICATE

This is to certify that the project entitled, **" Visualizing and Prediction of Stocks using Machine Learning"**, is bonafide work of **Kumawat Vicky Subhash Chand** bearing Seat. No: 2019016401103745 submitted in partial fulfillment of the requirements for the award of degree of BACHELOR OF SCIENCE in COMPTER SCIENCE from University of Mumbai.

Internal Guide
Asst. Prof. Rashid Mehmood

Coordinator
Asst. Prof. Rutika Deoo Malik

Principal
Dr. Sanjay Mishra

External Examiner

Date:

College Seal

ACKNOWLEDGEMENT

This project could not have been accomplished if not for the direct or indirect contribution from many known and unknown individuals. I wish to take this opportunity to express our sincere gratitude to all of them. I express our gratitude towards our internal guide **Asst. Prof. Rashid Mehmood**

who gave us unending support from the stage the project was initiated. A source of inspiration, given by them constantly kept our spirits high, whenever I was dispirited.

The foundation that I have been able to develop today owes much credit to them. Always ready to cooperate, they have been very kind in guiding us how to go about developing the project.

I would even like to thank all our friends for their cooperation and help in our project. Above all I would like to thank first, the almighty who have given me inspiration & courage to accept it as a course of life.

DECLARATION

I hereby declare that the project entitled, “Visualizing and Prediction of Stocks using Machine Learning” done at **place where the project is done**, has not been in any case duplicated to submit to any other university for the award of any degree. To the best of my knowledge other than me, no one has submitted to any other university.

The project is done in partial fulfillment of the requirements for the award of degree of **BACHELOR OF SCIENCE (COMPUTER SCIENCE)** to be submitted as final semester project as part of our curriculum.

Vicky Kumawat

PLAGARISM REPORT

SmallSEOTools

PLAGIARISM SCAN REPORT

Words		Date	
Characters	4989	Excluded URL	

10%	90%	3
Plagiarism	Unique	Plagiarized Sentences

Content Checked For Plagiarism

□ ABSTRACT
The stock market is highly volatile and complex in nature. Technical analysts often apply Technical Analysis (TA) on historical price data, which i
and might produce incorrect predictions. Machine learning coupled with fundamental and Technical Analysis also yields satisfactory results for s
prediction. In this work, I have made an effort to predict the price and price trend of stocks by applying Machine Learning technique's and adapti
Indicators (STIs)

□ INTRODUCTION
The stock market is a trading platform where different investors sell and purchase shares according to stock availability. Stock market ups and d
of stakeholders. If market prices going up with available stock then stakeholders get profit with their purchased stocks. In another case, if the ma
available stock prices then stakeholders have to face losses. Buyers buy stocks at low prices and sell stocks at high prices and try to get huge p
sellers sell their products at high prices for for-profit purposes.
Data analysis (DA) in machine learning (ML) is a process of applying technical skills (ML Algorithms) to historical data to obtain statistically as w
about predictions. It is also considered a technical process of data illustration and evaluation. Two authors (Adil E.Shamoo , 2012; David B.Resn
about DA, according to their theory DA is the process of distinguishing signals for decision making with statistical fluctuation of results.
In this research, I used several data resources in form of datasets and financial resources of data presentation. Yahoo Finance, Quandle, Kaggl
similar platforms provided data that is used in stock market predictions. I obtained data from these platforms for different stock exchange compa
applying ML algorithms I presented stocks predictions results statistically.
□ LITERATURE SURVEY
Stock Technical Indicators (STIs) STIs are statistical calculations based on the price, volume, or significance of a share, security, or contract. Th
on the fundamentals of a business, like earnings, revenue, or profit margins. The active stock traders and technical analysts commonly use STIs
term and long-term price movements and to identify entry and exit points.
To enhance the predictability of the daily stock price trends, Yuzheng Zhai et al presented a system based on an SVM (Support Vector Machine)
combines the technical indicators and related news releases. For each trading day, seven technical indicators are computed from the prices in th
Two groups of news releases are used. Two class categories, indicating the higher or equal price and lower price than the close price, are taken
next day's price movement. The system achieved higher accuracy than achieved using single-source i.e. news or technical indicators.
□ TOPIC
My Project concentrates on visualizing or representing and predicting stocks using Machine learning methods.
□ OBJECTIVE
My goal is to make a Machine Learning model that will predict the price and price trend of stocks by applying Long Short Term Memory (LSTM)
Stock Technical Indicators (STIs).

INDEX

Introduction

1.1 INTRODUCTION

The stock market is a trading platform where different investors sell and purchase shares according to stock availability. Stock market ups and downs affect the profit of stakeholders. If market prices going up with available stock then stakeholders get profit with their purchased stocks. In another case, if the market going down with available stock prices then stakeholders have to face losses. Buyers buy stocks at low prices and sell stocks at high prices and try to get huge profits. Similarly, sellers sell their products at high prices for for-profit purposes.

Data analysis (DA) in machine learning (ML) is a process of applying technical skills (ML Algorithms) to historical data to obtain statistically as well as tabular results about predictions. It is also considered a technical process of data illustration and evaluation. Two authors (Adil E.Shamoo , 2012; David B.Resnik , 2013) explained about DA, according to their theory DA is the process of distinguishing signals for decision making with statistical fluctuation of results.

In this research, I used several data resources in form of datasets and financial resources of data presentation. Yahoo Finance, Quandle, Kaggle and several other similar platforms provided data that is used in stock market predictions. I obtained data from these platforms for different stock exchange companies and after applying ML algorithms I presented stocks predictions results statistically.

1.2 LITERATURE SURVEY

Stock Technical Indicators (STIs) STIs are statistical calculations based on the price, volume, or significance of a share, security, or contract. These do not depend on the fundamentals of a business, like earnings, revenue, or profit margins. The active stock traders and technical analysts commonly use STIs to analyze short-term and long-term price movements and to identify entry and exit points.

To enhance the predictability of the daily stock price trends, Yuzheng Zhai et al presented a system based on an SVM (Support Vector Machine) algorithm that combines the technical indicators and related news releases. For each trading day, seven technical indicators are computed from the prices in the past five days. Two groups of news releases are used. Two class categories, indicating the higher or equal price and lower price than the close price, are taken up for indicating next day's price movement. The system achieved higher accuracy than achieved using single-source i.e. news or technical indicators.

1.3 OBJECTIVE

My goal is to make a Machine Learning model that will predict the price and price trend of stocks by applying Long Short Term Memory (LSTM) deep learning and Stock Technical Indicators (STIs).

1.4 PROPOSED METHOD

In an attempt to predict stock market trends and future stock prices, market researchers, investors, and scholars regularly propose a range of models. These models are based on various methods including the following.

- Single and dual sources of information. Single information source methods either utilize numerical or semantic information extracted from news and reviews while dual-source methods utilize historical price and volume data as well.

I am performing the dual-source of an information model for my topic which predicts the outcome on the basis of historical price and volume data and is using machine learning algorithms.

1.5 TECHNOLOGIES USED

i. Python 3.9
ii. Pandas 1.15.0
iii. Matplotlib 3.4.2
iv. Plotly 5.1.0
v. Machine Learning Algorithm
vi. NumPy
vii. TensorFlow
viii. sklearn

Requirement Specification

- **2.1 Purposed Method**:
  - In an attempt to predict stock market trends and future stock prices, market researchers, investors, and scholars regularly propose a range of models. These models are based on various methods including the following.
  - Single and dual sources of information. Single information source methods either utilize numerical or semantic information extracted from news and reviews while dual-source methods utilize historical price and volume data as well.
  - I am performing the dual-source of an information model for my topic which predicts the outcome on the basis of historical price and volume data and is using machine learning algorithms.

- **2.2 Software Requirement**
  - Python 3.6.x
  - Environment:
    - Visual Studio Code
    - Jupyter Notebook
  - Framework:
    - TensorFlow

- **2.3 Hardware Requirement**
  - Modern Operating System:
    - Windows 7 or 10
    - Mac OS X 10.11 or higher, 64-bit
    - Linux: RHEL 6/7, 64-bit (almost all libraries also work in Ubuntu)
  - x86 64-bit CPU (Intel / AMD architecture)
  - 8 GB RAM
  - 5 GB free disk space

System Design Details

3.1 Design Considerations:

Software Requirements:

A. Python:

- Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed.

- Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

Features :

1. Easy Language
2. Python is an easy language. It is easy to read, write, learn and understand.
3. Python has a smooth learning curve. It is easy to learn.
4. Python has a simple syntax and Python code is easy to understand.
5. Since it’s easy to understand, you can easily read and understand someone else’s code.
6. Python is also easy to write because of its simple syntax.
7. Because it is an easy language, it is used in schools and universities to introduce students to programming. Python is for both startups and big companies.
8. The Python language is designed to make developers life easy. Reading a Python code is like reading an English sentence. This is one of the key reason that makes Python best for beginners.
9. Python uses indentation instead of curly braces, unlike other programming languages. This makes the code look clean and easier to understand.
10. Python is an interpreted language. It comes with the IDLE (Interactive Development Environment). This

is an interpreter and follows the REPL structure (Read-Evaluate-Print-Loop).

B. Jupyter Notebook:

- The notebook extends the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results. The Jupyter notebook combines two components
- A web application: a browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output.
- Because you use Jupyter in a web browser, some people are understandably concerned about using it with sensitive data. However, if you followed the standard install instructions, Jupyter is actually running on your own computer. If the URL in the address bar starts with http://localhost: or http://127.0.0.1:, it's your computer acting as the server. Jupyter doesn't send your data anywhere else—and as it's open source, other people can check that we're being honest about this.
- You can also use Jupyter remotely: your company or university might run the server for you, for instance. If you want to work with sensitive data in those cases, talk to your IT or data protection staff about it.

- Main Features of Jupyter Notebook:

1. Notebook documents contains the inputs and outputs of a interactive session as well as additional text that accompanies the code but is not meant for execution. In this way, notebook files can serve as a complete computational record of a session, interleaving executable code with explanatory text, mathematics, and rich representations of resulting objects. These documents are internally JSON files and are saved with the .ipynb extension. Since JSON is a plain text format, they can be version-controlled and shared with colleagues.

2. Notebooks may be exported to a range of static formats, including HTML (for example, for blog posts), reStructuredText, LaTeX, PDF, and slide shows, via the nbconvert command.

3. Furthermore, any .ipynb notebook document available from a public URL can be shared via the Jupyter Notebook Viewer <nbviewer>. This service loads the notebook document from the URL and renders it as a static web page. The results may thus be shared with a colleague, or as a public blog post, without other users needing to install the Jupyter notebook themselves. In effect, nbviewer is simply nbconvert as a web service, so you can do your own static conversions with nbconvert, without relying on nbviewer.

C. TensorFlow:

- TensorFlow is an open source library for fast numerical computing.

- It was created and is maintained by Google and released under the Apache 2.0 open source license. The API is nominally for the Python programming language, although there is access to the underlying C++ API.

- Features of TensorFlow:

1.Open-source Library
   i. It is an open-source library that allows rapid and easier calculations in machine learning. It eases the switching of algorithms from one tool to another TensorFlow tool.
   ii. With the help of python, it provides the front-end API for the development of various machines and deep learning algorithms.

2. Easy to run
   i. We can execute TensorFlow applications on various platforms such as Android, Cloud, IOS and various architectures such as CPUs and GPUs. This allows it to be executed on various embedded platforms.

3.TensorFlow has its own designed hardware to train the neural models known as Cloud TPUs (TensorFlow Processing unit).

4.Fast Debugging
   i. It allows you to reflect each node, i.e., operation individually concerning its evaluation. Tensor Board works with the graph to visualize its working using its dashboard. It provides computational graphing methods that support an easy to execute paradigm.

5.Effective
   i. It works with multi-dimensional arrays with the help of data structure tensor which represents the edges in the flow graph. Tensor identifies each structure using three criteria: rank, type, shape

D. Sklearn (scikit-learn):

- Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

- Features of Sklearn:
  1.Simple and efficient tools for predictive data analysis
  2.Accessible to everybody, and reusable in various contexts
  3.Built on NumPy, SciPy, and matplotlib
  4.Open source, commercially usable - BSD license

E. Matplotlib:

- Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-

purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

- Features of Matplotlib:
    1.Semantic way to generate complex, subplot grids.
    2.Colored labels in legends
    3.Ticks and labels: new alignments of labels for axes and color bars

F. Pandas:

- Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution, from those that come with your operating system to commercial vendor distributions like ActiveState's ActivePython.

- Features of Pandas :

    1.Data cleansing
    2.Data fill
    3.Data normalization
    4.Merges and joins
    5.Data visualization
    6.Statistical analysis
    7.Data inspection
    8.Loading and saving data

G. Long short-term memory(LSTM):

- Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems).
- A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.
- LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional

RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications.

H. RNN(Recurrent Neural Network):

- RNNs are a powerful and robust type of neural network and belong to the most promising algorithms in use because it is the only one with an internal memory.

- Like many other deep learning algorithms, recurrent neural networks are relatively old. They were initially created in the 1980's, but only in recent years have we seen their true potential. An increase in computational power along with the the massive amounts of data that we now have to work with, and the invention of long short-term memory (LSTM) in the 1990s, has really brought RNNs to the foreground.

- Because of their internal memory, RNN's can remember important things about the input they received, which allows them to be very precise in predicting what's coming next. This is why they're the preferred algorithm for sequential data like time series, speech, text, financial data, audio, video, weather and much more. Recurrent neural networks can form a much deeper understanding of a sequence and its context compared to other algorithms.

3.2 **System Module** :

The module involved in this project is

- Main Application

❖ Main Application:

- Importing required Library's
- Loading the Data
- Preparing the data
- Building the Model
- Testing the model accuracy on existing data
- Loading the Test Data
- Making predictions on test data
- Plotting the tested predictions
- Predicting next day

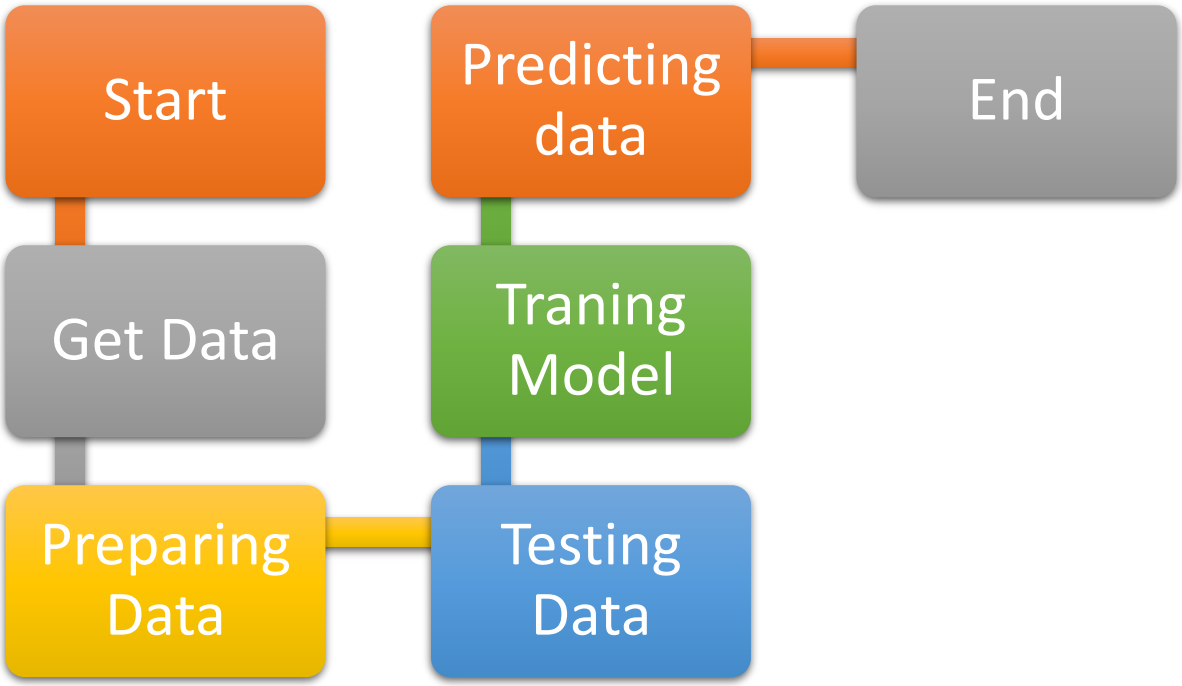
3.3 **System Design**:

3.2.1 Data Flow Diagram :

- A data flow diagram (DFD) maps out the flow of information for any process or system. It uses defined symbols like rectangles, circles and arrows, plus short text labels, to show data inputs, outputs, storage points and the routes between each destination.

- Notations :
  - Process

  - Dataflow



System Implementation

```
#Importing required Library's

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from pandas.core.algorithms import mode
import pandas_datareader as web
import datetime as dt

from sklearn.preprocessing import MinMaxScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout, LSTM
```

```
#Loading Data
company = 'GOOG' #"GOOG" is a Ticker Symbol more can be found on
https://www.nasdaq.com/market-activity/stocks/screener

start = dt.datetime(2012, 1, 1)
end = dt.datetime(2020, 1, 1)

df = web.DataReader(company, 'yahoo', start, end)
```

```
#Preparing the dataa
scaler = MinMaxScaler(feature_range=(0,1))
scaled_data = scaler.fit_transform(df['Close'].values.reshape(-1,1))

prediction_days = 60

x_train = []
y_train = []

for x in range(prediction_days, len(scaled_data)):
    x_train.append(scaled_data[x- prediction_days:x, 0])
    y_train.append(scaled_data[x, 0])

x_train, y_train = np.array(x_train), np.array(y_train)
x_train = np.reshape(x_train, (x_train.shape[0], x_train.shape[1], 1))
```

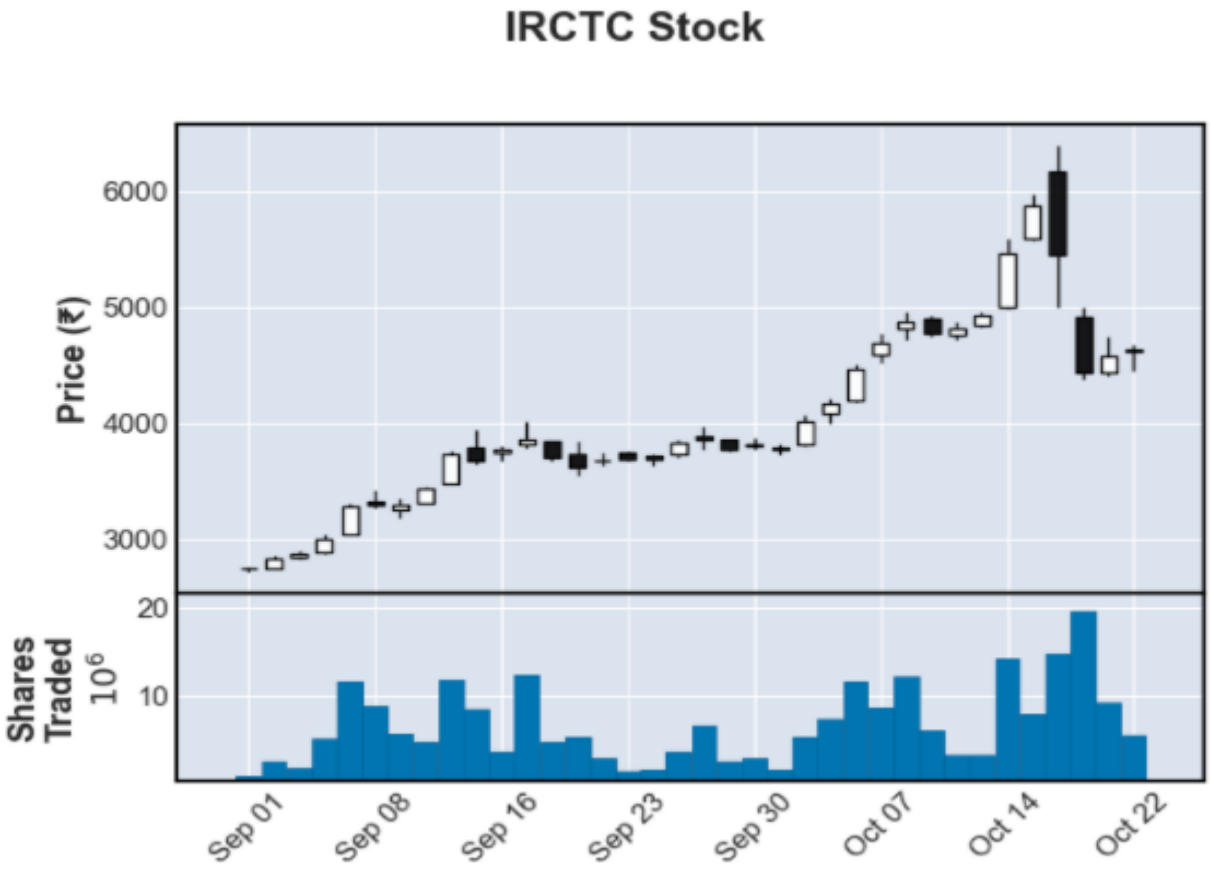
```
#Building  the Model
model = Sequential()

model.add(LSTM(units = 50, return_sequences=True, input_shape = (x_train.shape[1], 1)))
model.add(Dropout(0.2))
model.add(LSTM(units =50, return_sequences=True))
model.add(Dropout(0.2))
model.add(LSTM (units =50))
model.add(Dropout(0.2))
model.add(Dense(units =1))  #prediciton of next closing value

model.compile(optimizer='adam', loss = 'mean_squared_error')
model.fit(x_train, y_train, epochs=25, batch_size= 32)  #a number of epochs means the number
of times the model go thorugh the training set
```

```
#Testing the model accuracy on existing data

#Loading the Test Dataa that the model has never seen before
test_start = dt.datetime(2020, 1, 1)
test_end = dt.datetime.now()

test_data = web.DataReader(company, 'yahoo', test_start, test_end)
actual_prices = test_data['Close'].values

total_dataset = pd.concat((df['Close'], test_data['Close']), axis=0)

model_inputs = total_dataset[len(total_dataset) - len (test_data) - prediction_days:].values
model_inputs = model_inputs.reshape(-1, 1)
model_inputs = scaler.transform(model_inputs)
```

```
#making predection on test dataa

x_test = []

for x in range (prediction_days, len(model_inputs)):
    x_test.append(model_inputs[x - prediction_days:x, 0])

x_test = np.array(x_test)
x_test = np.reshape(x_test, (x_test.shape[0], x_test.shape[1], 1))

prediced_prices = model.predict(x_test)
prediced_prices = scaler.inverse_transform(prediced_prices)
```

```
#ploting the tested predictions

plt.plot(actual_prices, color = 'black', label =  f"Actual {company} Price")
plt.plot(actual_prices, color = 'yellow', label = f"Predicted {company} Price")
plt.title(f"{company} Share Price")
plt.xlabel('Time')
plt.ylabel(f'{company} Share Price')
plt.legend()
plt.show()

#OUTPUT
Prediction: [[283.6123]]
```

```
#predicting next day

real_data = [model_inputs[len(model_inputs) +1 - prediction_days:len(model_inputs+1), 0]]
real_data = np.array(real_data)
real_data = np.reshape(real_data, (real_data.shape[0], real_data.shape[1], 1))

prediction = model.predict(real_data)
prediction = scaler.inverse_transform(prediction)
print(f"Prediction: {prediction}")

#Visualizing the Stock
import pandas as pd
import matplotlib.pyplot as plt
import pandas_datareader.data as web #----- used when data is to be called from yahoo server
import mplfinance as fplt
import datetime as dt

start = dt.datetime(2021, 9, 1)
end = dt.datetime.today()


df = web.DataReader('IRCTC.NS', 'yahoo', start, end)
fplt.plot(

    df,
    type = 'candle',
    title = 'IRCTC Stock',
    ylabel = 'Price (₹)',
    volume = True,
    ylabel_lower = 'Shares\nTraded'
    )

print(df.info())
```


IRCTC Stock
Price (₹)
6000
5000
4000
3000
Shares Traded
10^6
20
10
Sep 01
Sep 08
Sep 16
Sep 23
Sep 30
Oct 07
Oct 14
Oct 22

Conclusion and Future Scope

This Program is built using Python and Machine Learning algorithm that receives and analyzes input data to predict output values based on historical data.

5.1 System maintenance:

- No Maintenance is required for this program the program train itself with the data and predicts the future on the basis of the past data

5.2 Future Enhancement & Work:

- The prediction accuracy can be increased by taking the model and training it with more data.
- Creating a new optimized algorithm to optimize and increase the accuracy of the model.

References

➔ **Full URL of online references**

- https://finance.yahoo.com/quote/FB?p=FB
- https://finance.yahoo.com/quote/GOOG?p=GOOG
- https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/
- https://en.wikipedia.org/wiki/Long_short-term_memory#:~:text=Long%20short-term%20memory%20(LSTM)%20is%20an%20artificial%20recurrent,the%20field%20of%20deep%20learning.&text=Relative%20insensitivity%20to%20gap%20length,learning%20methods%20in%20numerous%20applications.
- https://towardsdatascience.com/sequence-models-and-recurrent-neural-networks-rnns-62cadeb4f1e1#:~:text=Sequence%20models%20are%20the%20machine,Applications%20of%20Sequence%20Models
- https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/#:~:text=The%20batch%20size%20is%20a%20hyperparameter%20that%20defines%20the%20number,more%20samples%20and%20making%20predictions.
- https://www.w3schools.com/python/matplotlib_pyplot.asp
- https://www.nasdaq.com/market-activity/stocks/screener
- https://in.finance.yahoo.com/
- https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
- https://www.youtube.com/watch?v=5dMXyiWddYs&ab_channel=MATLAB
- https://www.youtube.com/watch?v=vmEHCJofslg&ab_channel=KeithGalli
- https://www.youtube.com/watch?v=hSPmj7mK6ng&ab_channel=CharmingData
- https://www.youtube.com/watch?v=GwIo3gDZCVQ&ab_channel=edureka%21
- https://nbviewer.jupyter.org/urls/bitbucket.org/hrojas/learn-pandas/raw/master/lessons/01%20-%20Lesson.ipynb
- https://economictimes.indiatimes.com/marketstats/pid-51,pageno-1,year-2012,sortorder-asc,sortby-CurrentYearRank.cms