B. P. Poddar Institute of Management & Technology

Department of Computer Applications

Major Project

Paper Code: MCAN 482

Project Title: STOCK PRIZE PREDICTION.

Mentor's Name: ABHIJIT BHATTACHARYA

Project Member's Details:

Name	Roll Number	Registration No	Dept.
Saheb Mukherjee	11571020039	201150571010038 Of (2020-2022)	MCA 2nd Year

Certificate of Originality

This is to certify that project report entitled **"STOCK PRIZE PREDICTION"** is a bona-fide record of the project done by:

Saheb Mukherjee (Roll No:**11571020039**, Registration No: **201150571010038** of 2020-2022),

During the project, there were found sincere, regular and hardworking and successfully completed the project.

We wish them success in future.

Project Co-Ordinator	**Project Guide**	**Head of dept.**	**External Examine**

ACKNOWLEDGEMENT

We extend our sincere and heart filled thanks to our esteemed guide, Mr. Abhijit Bhattacharya asst professor for providing us with the right guidance and advice at the crucial junctures and for showing us the right way.

We are especially grateful to our respected HOD/Dept. In-Charge, Dr. Inadyuti Dutt for allowing us to use the facility available. We would like to thank the other faculty members also.

TABLE OF CONTENT

<u>Abstract</u>

The main objective is to collect the streaming Data of 10 years through web scrapping technique using a python library `pandas_datareader` The Data that would be collected will be the records of user desired stock which can be enlisted in any stock exchange company. Which will be then taken as a base for our exploratory data analysis (EDA). After that we would be manipulating the data and will only use the `Close` Column from the dataset for further pre-processing. Now using python data science module like pandas and NumPy we will prepare our target data set and feature dataset. Feature dataset will have the record of previous 60 days closing price and target set will have the 61st day's closing price. In these ways will going to prepare all the 10-year record. Then the LSTM algorithms will find patterns in the data and use regression method to predict the future ups and downs of the closing stock price of that particular company stock.

Introduction

A stock market is a place where people buy/sell shares or stocks of publicly listed companies

It is a widely used source for people to invest money in companies with high growth potential.

Stock market is a very risky place where it is most likely that you will face some amount of loss at a point of time.

So our aim is to minimize this risk and provide support to the investors

The paper focuses on the use of Regression technique using LSTM based Machine learning technology to predict the closing stock values.

Technology used

Machine learning techniques have been used wherein we are primarily going to exercise machine learning algorithm which comes under artificial intelligence. The purpose of AI or ML techniques is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Basically, Here we used the dynamic python programming language and it's properties of data science where we used famous data science modules like NumPy, Pandas, Matplotlib, Scikit-Learn, TensorFlow, etc. To predict a time series data which is our stock data.

Related Concept

Machine learning:

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Artificial Intelligence:

Artificial intelligence is a technology that enables a machine to simulate human behavior. Machine learning is a subset of AI which allows a machine to automatically learn from past data without programming explicitly. The goal of AI is to make a smart computer system like humans to solve complex problems.

Neural Network:

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature.

LSTM (Long-Short Term Memory):

The proposed framework that learns online anticipating the close costs of the stock with the assistance of Long Short-Term Memory (LSTM). The Long Short term Memory (LSTM) is a counterfeit intermittent neural system (RNN) used in the field of deep / machine learning, Unlike standard feed forward neural systems, LSTM has input associations. Not only does the procedure does not focus on single information (e.g. pictures) but also on full information arrangements, (For example, a speech or a video). For example, LSTM is material for undertakings, such as un partitioned, associated penmanship recognition, speech recognition and recognition of peculiarities in arranged traffic or IDS (Interruption location frameworks).

System Descriptions

a) Existing System Description

While doing the literature survey Stock Market Prediction Using Machine Learning Algorithms, IJEAT, 2019
the data about Stock market prediction systems that are as of now being utilized are considered. Over the most recent two decades determining of stock returns has become a significant field of research. In most of the cases the scientists had endeavoured to build up a straight connection between the information macroeconomic factors what's more, the stock returns, be that as it may, with the revelation of nonlinear slants in the financial exchange record returns, there has been an incredible move in the focal point of the scientists towards the nonlinear expectation of the stock returns.
Even though, there after numerous writings have come up in nonlinear measurable displaying of the stock returns, the majority of them required that the nonlinear model be indicated before the estimation is done.

### b) Disadvantages of the existing system

While this technique may have had good success in other fields like speech recognition, it does not perform as well when applied to financial data. They are in fact characterized by high noise-to-signal ratio, which makes it difficult for a machine learning model to find patterns and predict future prices.

c. Overview of the proposed system

As represented in the previous section getting the Historical data from market is mandatory step. Then there is a need to extract the feature which is required for data analysis, then divide it as testing and training data, training the algorithm to predict the price and the final step it to visualize the data.

The typical LSTM unit consists of a cell, an info door, an entrance door and a door with a view. The cell collects values over discretionary time intervals, and the three inputs manage the progress of data into and out of the cell. The main advantage of the LSTM is its ability to learn context-specific temporal dependence. Each LSTM unit collects information for either a long or short period of time (hence the name) without explicitly using the activation function within the recurrent components.

d. **Data Flow-Diagram:**
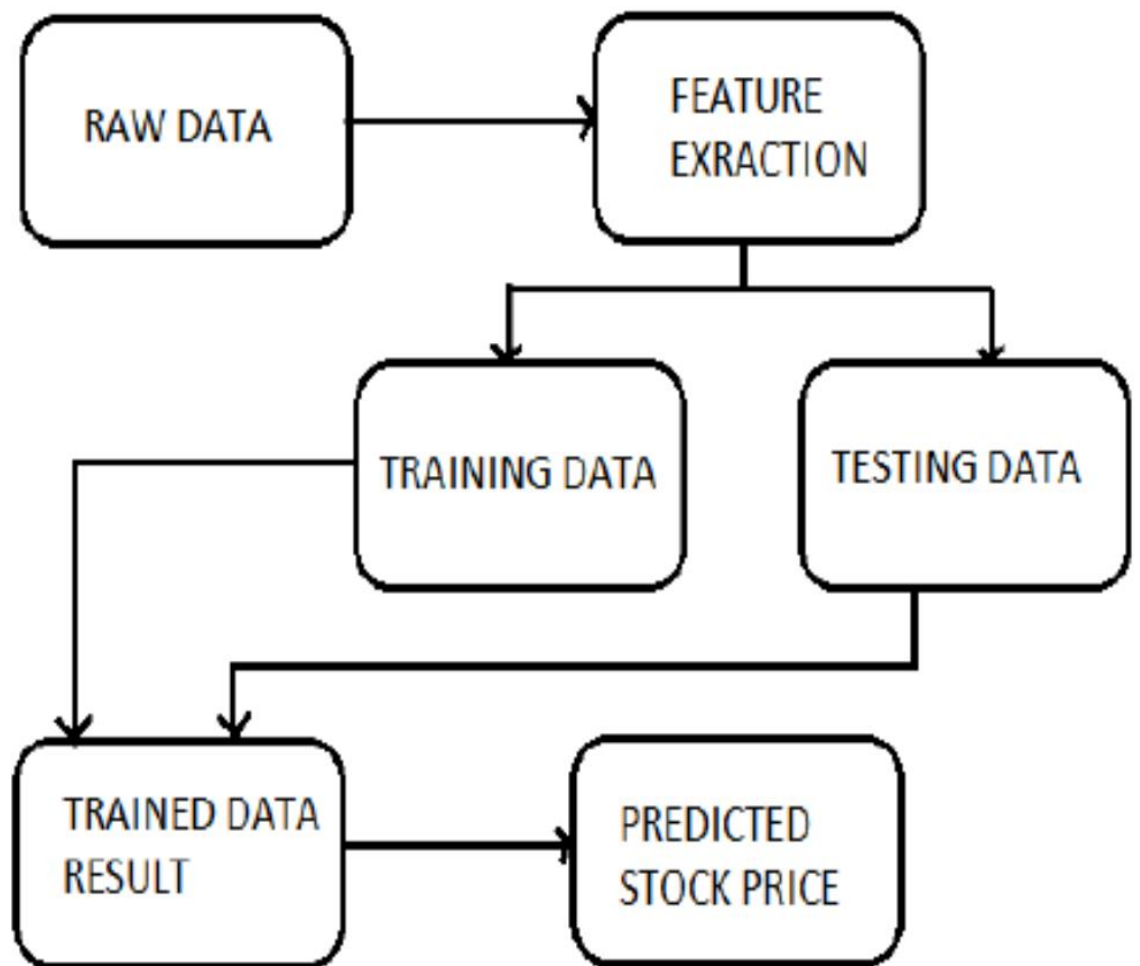
Resource requirement

Hardware requirement

- Processer: AMD A6-6400K APU with Radeon(tm) HD Graphics3.90 GHz
- RAM: 8 GB
- Good Internet Connection
- Hard Disk: 1 TB

Software requirement

- Python (IDE) (version 3.10)
- Windows 10 Home Single Language
- Google Collaboratory
- Anaconda Prompt
- Jupyter notebook
- Scikit Learn
- TensorFlow
- keras
- Pandas
- Pandas data-reader
- NumPy
- Matplotlib
- Seaborn

Functions of modules

Description of the modules:
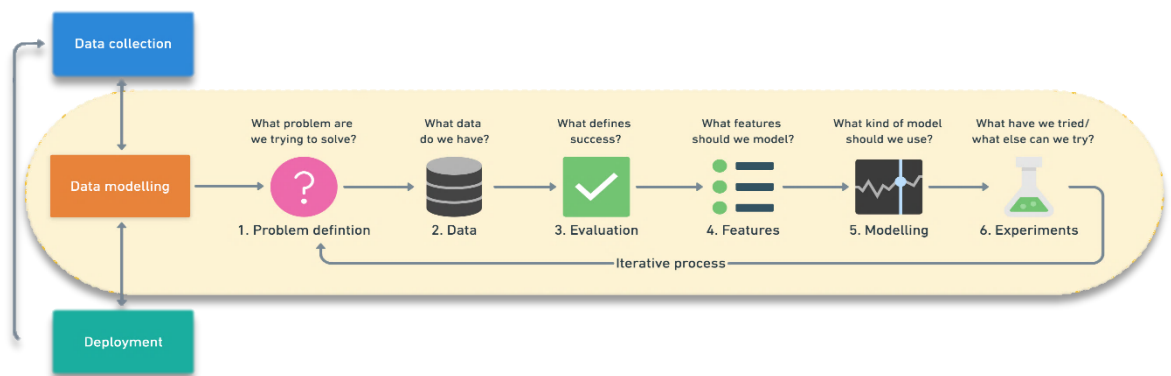
I.   Data collection:

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.
Data collection is the process of gathering and measuring information from countless different sources. In order to use the data, we collect to develop practical artificial intelligence (AI) and machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand.

II.   Data Modelling:

Data modeling is the process of creating a simplified diagram of a software system and the data elements it contains, using text and symbols to represent the data and how it flows. Data models provide

a blueprint for designing a new database or reengineering a legacy application.

Sub-topics of Data Modelling are –

- *Problem Definition-*
  The first significant step/process in a deep learning project is the problem definition. Problem Definition: A clear statement describing the initial state of a problem that's to be solved.
  A problem statement is a concise description of the problem or issues a project seeks to address. The problem statement identifies the current state, the desired future state and any gaps between the two.

- *Data-*
  It can be any unprocessed fact, value, text, sound, or picture that is not being interpreted and analyzed. Data is the most important part of all Data Analytics, Machine Learning, Artificial Intelligence. Without data, we can't train any model and all modern research and automation will go in vain.
  In machine learning, one of the things that should be taken care of is the type of data given to the model. If we have more data, there is a higher chance for a machine learning algorithm to understand it and give accurate predictions to the unseen data respectively.

- *Evaluation-*
  Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.

- *Features-*
  A feature is an attribute used to describe each example. ML models are effective because they learn to generate predictions for new examples; this process is called inference. The examples used during the learning process are commonly referred to as training data.

A feature store is an ML-specific data system that: Runs data pipelines that transform raw data into feature values. Stores and manages the feature data itself, and. Serves feature data consistently for training and inference purposes.

- *Modelling-*
  The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data.

The six steps to building a machine learning model include:

1. Contextualise machine learning in your organisation.
2. Explore the data and choose the type of algorithm.
3. Prepare and clean the dataset.
4. Split the prepared dataset and perform cross validation.
5. Perform machine learning optimisation.
6. Deploy the model.

- **Experiments-**
  AI Experiments is a showcase for simple experiments that make it easier for anyone to start exploring machine learning, through pictures, drawings, language, music, and more.

What is ML experiment tracking?

- Scripts used for running the experiment.
- Environment configuration files.
- Versions of the data used for training and evaluation.
- Parameter configurations.
- Evaluation metrics.
- Model weights.
- Performance visualizations (confusion matrix, ROC curve)

III. Development-

The development set is a significant dataset in the process of developing a ML model and it forms the basis of the whole model evaluation procedure. A machine learning algorithm has two parameters - model parameters that define individual models and hyperparameters define high-level structural settings for algorithms.

The 7 Key Steps To Build Your Machine Learning Model

1. Step 1: Collect Data. ...
2. Step 2: Prepare the data. ...
3. Step 3: Choose the model. ...
4. Step 4 Train your machine model. ...
5. Step 5: Evaluation. ...
6. Step 6: Parameter Tuning. ...
7. Step 7: Prediction or Inference.

<u>Some other important Libraries:</u>

Pandas-

Pandas is one of the tools in Machine Learning which is used for data cleaning and analysis. It has features which are used for exploring, cleaning, transforming and visualizing from data. Pandas is an open-source python package built on top of NumPy developed.

Numpy-

Numpy-ml is a growing collection of machine learning models, algorithms, and tools written exclusively in NumPy and the Python standard library. The purpose of the project is to provide reference implementations of common machine learning components for rapid prototyping and experimentation.

Sklearn preprocessing –

The sklearn. preprocessing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators. In general, learning algorithms benefit from standardization of the data set.

MinMaxscaler-

In machine learning, MinMaxscaler and StandardScaler are two scaling algorithms for continuous variables. The MinMaxscaler is a type of scaler that scales the minimum and maximum values to be 0 and 1 respectively.

Matplotlib.pyplot-

Matplotlib is one of the plotting library in python which is however widely in use for machine learning application with its numerical mathematics extension- Numpy to create static, animated and interactive visualizations.

Sequential-

Machine learning models that input or output data sequences are known as sequence models. Text streams, audio clips, video clips, time-series data, and other types of sequential data are examples of sequential data. Recurrent Neural Networks (RNNs) are a well-known method in sequence models.

Keras.Model-

Keras is a Machine Learning (ML) library built on top of TensorFlow, making it extremely easy to create and fit Deep Learning model architectures. As they like to say, "Keras API is designed for human beings, not machines." Keras provides three main ways to solve your ML tasks: The sequential API.

Dense-

Dense Layer is simple layer of neurons in which each neuron receives input from all the neurons of previous layer, thus called as dense. Dense Layer is used to classify image based on output from convolutional layers. Working of single neuron. A layer contains multiple number of such neurons.

Pandas-DataReader-

Pandas Datareader is a Python package that allows us to create a pandas DataFrame object by using various data sources from the internet. It is popularly used for working with Realtime stock price datasets.

Design of the system:

Algorithm:

Step 1: Start.

Step 2: Data Pre-processing after getting the historic data from the market for a particular share.

Step 3: import the dataset to the data structure and read the open price.

Step 4: do a feature scaling on the data so that the data values will vary from 0 and 1.

Step 5: Creating a data structure with 60 timestamps 0 and 1 output.

Step 6: Building the RNN (Recurrent neural network) for Step 5 data set and Initialize the RNN by using sequential repressor.

Step 7: Adding the first LSTM layer and some Dropout regularization for removing unwanted values.

Step 8: Adding the output layer.

Step 9: Compiling the RNN by adding adam optimization and the loss as mean_squared_error.

Step 10: Making the predictions and visualizing the results using plotting techniques.

Parameter used:

Parameter Used	Meaning
Date	Date of stock price
Open	Open price of a share
Close	Closing price of a share
Volume/ trade quantity	Number of shares traded
High	Highest share value for the day
Low	Lowest share value for the day
Turnover	Total Turnover of the share

Input, Processing & Outputs

Input: we are fetching the data from internet collect the data and stored in dataset.

Date	High	Low	Open	Close	Volume	Adj Close
2012-01-02	351.542725	340.348846	345.128540	349.957764	8679938.0	323.036774
2012-01-03	360.037201	351.839905	352.780975	358.922760	9455771.0	331.312134
2012-01-04	362.043182	353.325836	360.284851	354.712677	8557084.0	327.425934
2012-01-05	359.071350	343.791199	354.143066	346.465851	13364666.0	319.813507
2012-01-06	358.600830	345.054230	345.252350	355.406097	9495456.0	328.065979
...	...	...	...	...	...	...
2022-03-28	2629.750000	2586.500000	2610.000000	2621.949951	4564891.0	2621.949951
2022-03-29	2638.000000	2607.399902	2638.000000	2622.550049	4007695.0	2622.550049
2022-03-30	2688.000000	2617.100098	2639.899902	2672.949951	7297028.0	2672.949951
2022-03-31	2669.699951	2628.600098	2664.949951	2634.750000	6102744.0	2634.750000
2022-04-01	2665.149902	2622.000000	2636.000000	2655.850098	3656408.0	2655.850098

2527 rows × 6 columns

Processing:

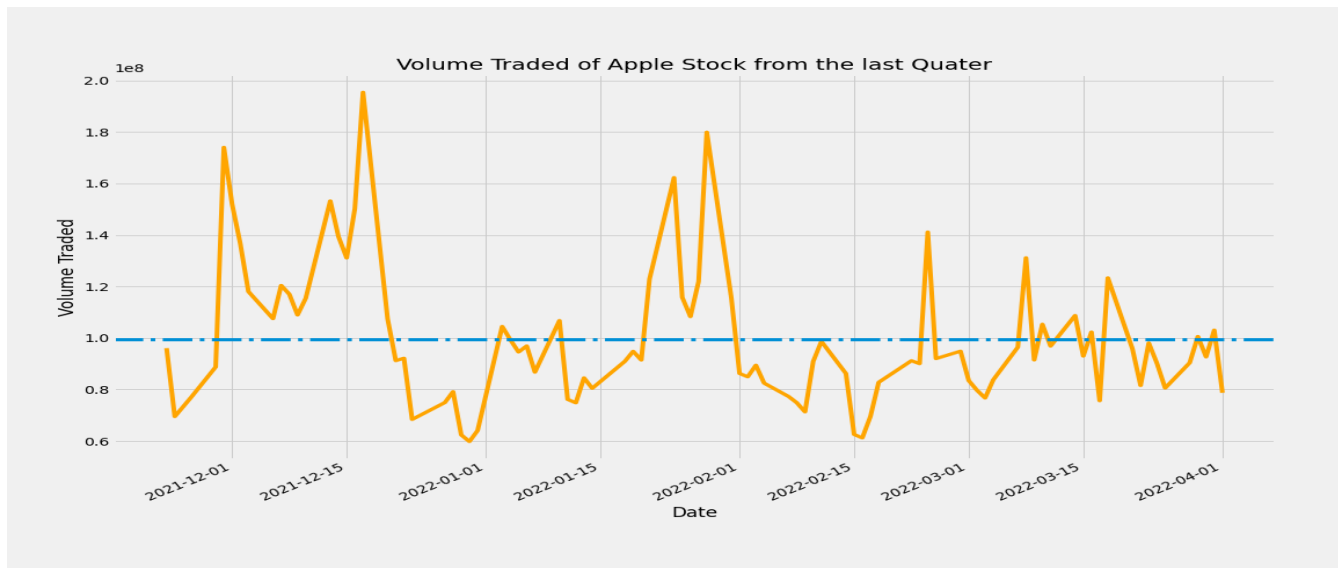From that collect dataset we can train the data and as user choice we can test them.

Here we are taking two example and read the 10years performance observing in the graph.
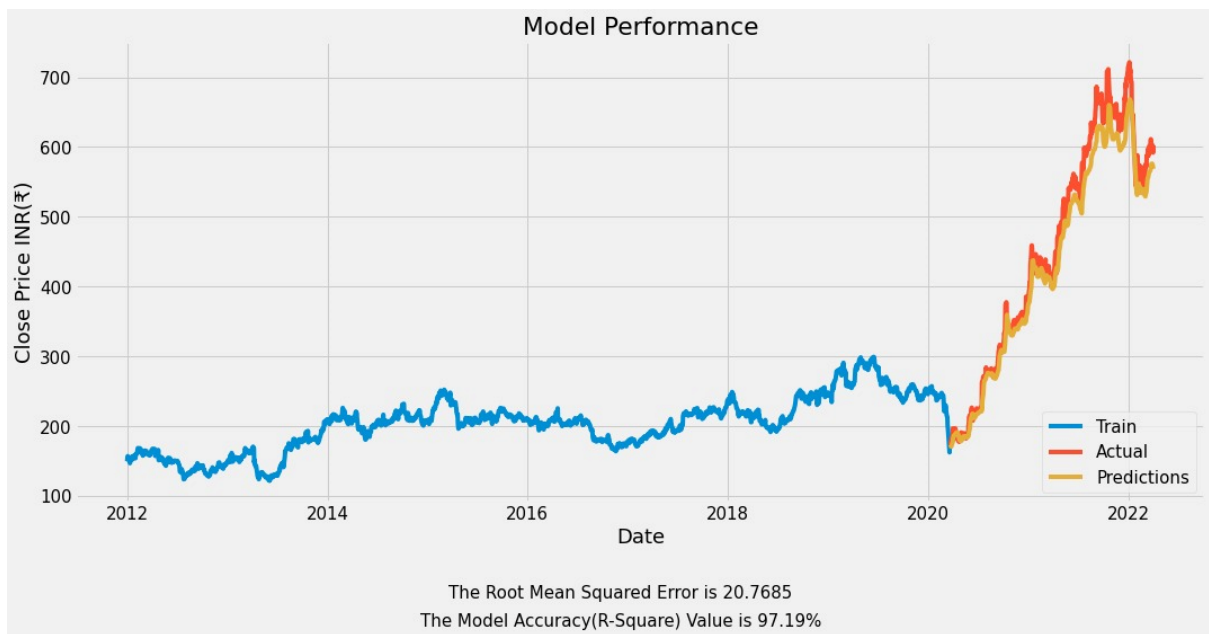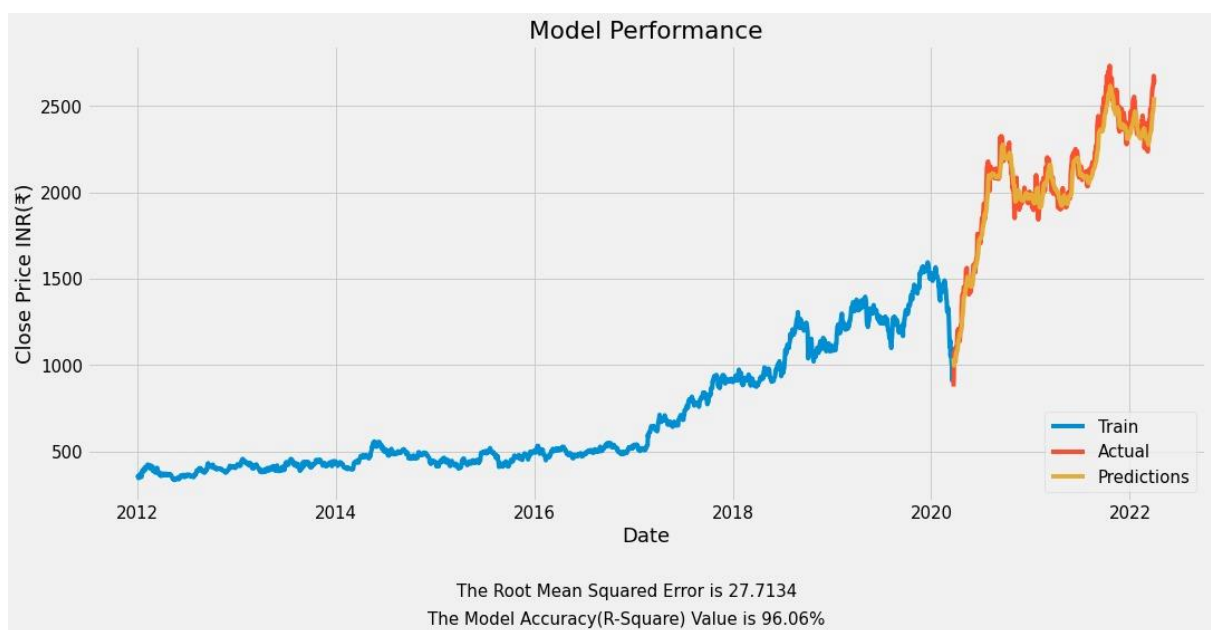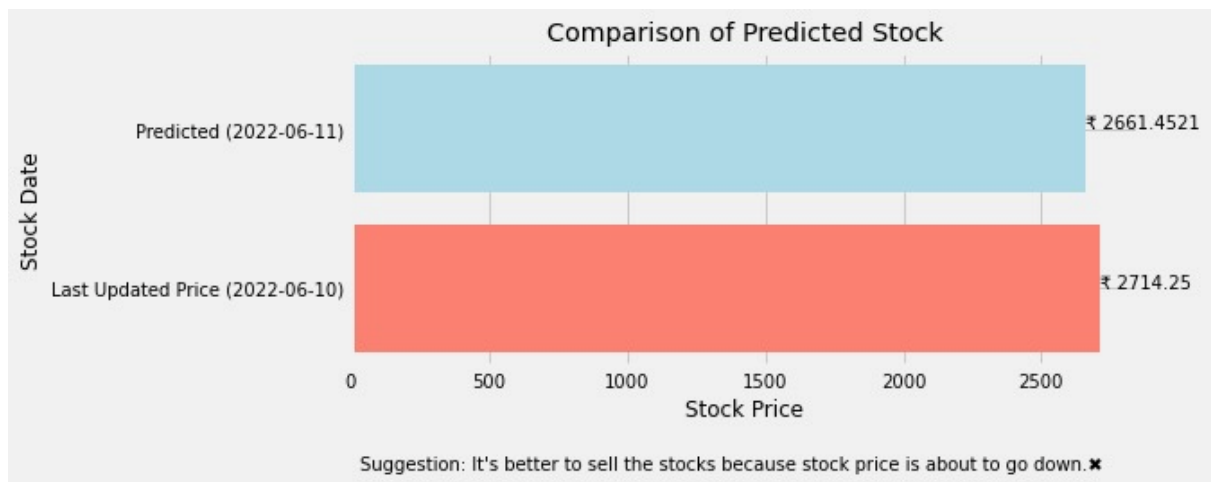
Output:

We are taking last 60 days data then we he implementation of LSTM model using python which predicts the future price of Wipro and Reliance share based on its historical data. The below visualization figure shows the visualization of SHARE prediction. In our paper the implementation of an algorithm which predicts the stock price of a share for given period, the below graph from our algorithm will show the predicted price of Wipro and Reliance share. In the result shown in the below graph is the plotted form our algorithm outcome by applying 96 LSTM units for achieving the accuracy.

We also finding Root Mean Square Error and R-square value.

Figure: Volume Traded of Apple Stock from the last Quater

These figure is graphical representation volume trade of Apple Stock.

Figure: Model Performance

The Root Mean Squared Error is 20.7685
The Model Accuracy(R-Square) Value is 97.19%

Comparison of Predicted Stock

Predicted (2022-06-11) — ₹ 2661.4521

Last Updated Price (2022-06-10) — ₹ 2714.25

Suggestion: It's better to sell the stocks because stock price is about to go down. ✖

Model Performance

The Root Mean Squared Error is 27.7134

The Model Accuracy(R-Square) Value is 96.06%

Comparison of Predicted Stock

Predicted (2022-06-11) — ₹ 454.3594

Last Updated Price (2022-06-10) — ₹ 460.95001220703125

Suggestion: It's better to sell the stocks because stock price is about to go down. ✖

Testing :

his program is scrapping data from a trusted website that is Yahoo finance to get the stock data by using a web api. So, the data in which we are working has been verified.

The data in which we are working can be classified as time series data, so to predict a time series data we are using a RNN algorithm that is LSTM which is tested by various developers and has been their first choice for predicting time series data.

The project itself test the predicted results through some famous matrixs like RMSE(Root Mean Square Error), and R-square metrics and is performing well enough with less error and high accuracy.

The overall project has been tested many times and the predicted result is very close to accurate to the actual data and some graphs has also been generated to visualise the overall performance of the project.

Conclusion:

The study of the share is carried out in this paper and it can be carried out for several shares in the future. Prediction could be more reliable if the model trains a greater number of data sets using higher computing capacities, an increased number of layers, and LSTM modules. The expected outcome of this project is to develop a system which will help us to predict the stock market future 60 day's outcome. Also predict the stock assurance and predict the names where we can invest our stock price or not.

Future Scope:

For future improvement we have thought that we will be going to apply even more advance algorithm then LSTM Like ARIMA, Bi-LSTM model, etc.

We are also going to add even more evaluation system which will be better than previous evaluation system and also we will be improving the model accuracy and will reduce the training time by saving those models weights which are most used. For example RELIANCE.NS, WIPRO.NS, etc.

We are also working on deploying this system in a web application form using flask/Django.

Issues that we are facing:

Any type of burnable stock i.e., having too much up-down fluctuations in the market, that kind of burnable stock becomes hard for LSTM to predict accurately and the accuracy scores and RSME score might get effected.

References:

- YouTube/KrishNaik

- YouTube/codebasic

- SK-Learn Documentation

- TensorFlow Documentation

- LSTM Documentation

- Stackoverflow.com

- Udemy

- Wikipedia