# Predictions and Trend Analysis for Stock Market Using Machine Learning Algorithms

**V. Manas Advaith, J. Jeshwanth Reddy, V. P. Srinidhi, Prerana Umakant Bandekar, and Ananthanagu**

**Abstract** This paper exploits some of the Industry standard applications for trading and gives new insights on few ML models, namely LSTM, ARIMA, SVR. Techniques like Bollinger Bands and Support and Resistance are implemented to help in understanding the stock data better. The usage of Optimal Portfolio has been demonstrated to aid in making buy/sell decisions. It also talks clearly about the 20 min window used by certain trading application when a new user tries to learn trading, and how the new users are being manipulated. It gives an alternate approach to paper trading with no real money involved and the most appropriate ML models that can be used for short term prediction through commodity hardware.

## 1 Introduction

(A) Most trading applications provide a platform for users to test their strategy or to learn trading but they are being manipulated by what is called a 20-min trading window (Fig. 1).

As you can see from the image that only the peak in each 20 minutes time frame is taken and plotted in the pseudo platform giving users a wrong idea that the stocks are always increasing or decreasing at a slower rate and thereby persuading them to start the actual trading so that the platform can make money from brokerage.

To counter this we have come up with an approach to provide a platform just for pseudo trading which fetches real-time data from Yahoo Finance API. There is little lag in the data that is being displayed so the users can study the stock more clearly. We also provide them with certain understandings using techniques like Support and Resistance, Bollinger bands, Reinforcement learning, Optimal portfolio for price prediction for the next day [3].

V. Manas Advaith · J. Jeshwanth Reddy (✉) · V. P. Srinidhi · P. U. Bandekar · Ananthanagu
Department of CSE, PES University, Bengaluru, India
e-mail: jeshwanth7899@gmail.com

N. R. Shetty et al. (eds.), *Advances in Computing and Information*, Lecture Notes in Electrical Engineering 1104, https://doi.org/10.1007/978-981-99-7622-5_22

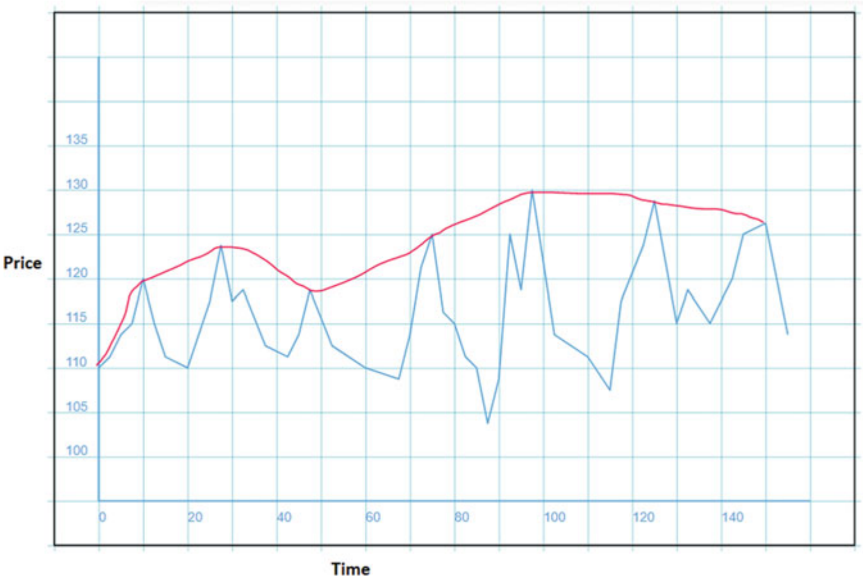

**Fig. 1** 20 min trading window

(B) Some of the ML models have been practiced to see their acceptance for prediction. Namely LSTM, SVR, Reinforcement Learning and ARIMA [1–3]. These models have been tweaked to give acceptable results using less resources for short term prediction. We've compiled all of those results to come up with a model that gives RMSE less than 3. The RMSE can be set by the users to give better results but would require a lot of computation. In our research we found RMSE $\leq 3$ to be a good standard [1].

## 2 Literature Survey

A deep neural network stacked with bidirectional long short-term memory is suggested for stock market prediction in earlier work [1]. The model is stacked in both the forward and backward directions, the model is stacked. It is made up of two layers: a prediction layer and a dropout layer. The temporal properties are captured in both directions using this. It comprised six datasets of well-known companies throughout a five-year period. The model's ability to learn temporal information improves its capacity for prediction. Additionally, according to the study, artificial neural networks are particularly effective at understanding intricate time series data patterns. The average RMSE value across all forecasts from the 6 datasets utilised has been calculated to be approximately 0.001, which is quite accurate. This demonstrates the suitability of SBLSTM as a model for such predictions.

In [2], we examine a neural network model once more that was trained using data from more than 300 stocks over a 25-year span. The model is a multi-layer perceptron, with 59 neurons in the first layer. Instead of making price predictions, the focus is primarily on two factors that affect stock prices. These factors are classified as either authentic or fake golden crosses, which are again separated into two categories. In contrast to other classifiers, the MLP model fails to distinguish between a fake and a real golden cross. Consequently, the investor will not suffer any loss even if a golden Cross is thought to be a fake golden cross. This investing strategy model is highly helpful in identifying genuine and phoney golden crosses to generate large positive returns. The volume also matters since a bigger volume indicates widespread participation, which supports the market participants' choice. But because this strategy often only turns a profit after 10 years and because this model is unable to account for any volatility throughout that time, it is clear that no model can completely forecast how the trend will develop.

In [3], a portfolio is built by utilising LSTM with some added characteristics to forecast stock prices. Financial analytics have benefited greatly from machine learning algorithms' capacity to manage the nonlinear structure of the data. The sharpie ratio and the 1/N rule are mostly used to build the portfolio. The estimated outcomes of the LSTM model link these together. To estimate the returns for creating this new portfolio, characteristics from candlestick charts are used, together with feature vector-based LSTM to predict closing prices. According to [1], LSTM is a very effective model for predicting stock prices. It is therefore the model that may be utilised to build a lucrative portfolio in the most practical way. When the decoder reconstructs the same data from the encoder vector, the encoder is shrunk in relation to the input data to a smaller dimension, and the data is then sent to the network, which has four completely linked layers starting at 4096 units and going down to 512 units. The same arrangement also applies to the decoder network. This technique was very helpful in building a portfolio with big profits.

Multilayer Perceptron (MLP), Long Short-Term Memory, and Auto-Regressive Integrated Moving Average (ARIMA) were the models utilised in [5]. (LSTM). After training the models for eight years, SBI stock was selected for training and forecasting values for 50 market days. Appropriate P, D, and Q values were then specified for the ARIMA model. These are the outcomes: ARIMA fared the best by displaying the lowest RMSE number, MLP displayed a low RMSE value after some minor tuning, and LSTM came in last. Thus, this finding demonstrates that ARIMA is effective for short-term prediction. It was also noted that altering the value of Q caused a sizable regular shift in both the RMSE and the projected values. This leads us to the conclusion that a smaller window size produced ARIMA forecasts that were more accurate. Additionally, it was shown that when the trend changed, LSTM could not reliably forecast the values. MLP more easily recognises the trend and produces the forecasts properly with greater accuracy, but it fails to do so over time. A crucial component of [6] is the detection of abnormalities using stock price data since this reveals whether there has been intentional market manipulation. The deliberate manipulation of markets that results in price changes may also be the source of failed predictions. A deep learning algorithm that can recognise abnormalities using historical patterns from previously traded stock prices was suggested for application. It is crucial to distinguish between a true manipulation and a fake manipulation because it is also possible to experience volume and price increases that are not necessarily the result of manipulation. A total of 13 stocks, including businesses that joined the Shanghai and Shenzhen stock exchanges, were examined. The study's time frame was 1219 days, with the first 1,000 days being training data and the final 219 days considered verifying data. The Conv1D layer received the uni-variant and multi-variant data so that it could slice them into readable lengths for the LSTM inputs. After this layer, the organised data is divided into several LSTM and dense neural layer types. When the MAE's forecast on validation reflects the trend of open price, it is discovered to be less than 4.5. 13 equities were chosen at random to test it statistically given historical performance data. The model should forecast new information based on earlier data sets. The projected period is labelled as problematic if the MAE on the training set is determined to be greater than 30%. Using this technique, it was discovered that on 2019-02-13, there may have been an anomaly that indicated market manipulation.

For predicting e-commerce sales, it was suggested in [9] to use a prediction model based on the Granger casualty test and XGboost algorithm, where the Granger casualty test is used to extract balance information and the XGBoost algorithm is used to create predictions. It was noted that the hidden Markov prediction model was employed to obtain the prediction results because the data was dispersed. Given the likelihood of the outcome, which can increase the prediction's accuracy. The future sales volume of e-commerce is examined, and it is discovered that the ARIMA-BP combined forecasting model may be successfully utilised to estimate the future sales volume of e-commerce.

## 3 Proposed Methodology

The idea is to take the existing models and change the number of hidden layers and increase/decrease the parameters to give approximate prediction using less resources. Multiple combinations have been tried to finally come up with a suitable, easy to execute, accurate model. Some work on Bollinger bands and Support and Resistance has also been done which eases the buy/sell decision.

### *3.1 Data-Set*

Data is extracted with the help of Yahoo Finance API, with a lag less than 1 second, to display the graph in real-time. Data from 2018—previous day is taken in the form of data-frame to test and train the models. Data in the form of .CSV file is also used which was downloaded from Yahoo Finance for some models (Fig. 2).

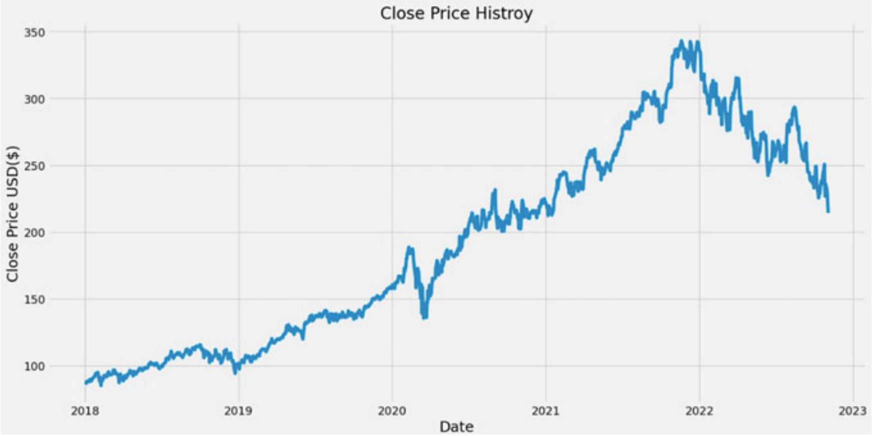

**Fig. 2** Microsoft data from 2018-01-01 to 2022-11-06

## 3.2 Support and Resistance Levels

A support or resistance level is an important technical indicator that tells the user whether a stock will go up or down at that specific level. This level is derived from the history of the stock prices. The price of a stock always varies by either being in an uptrend or a downtrend and the reason why a stock always bounces from a specific price is due to these levels. The stock market prices are somewhat a culmination of human emotions. Where the culmination of these emotions decide how the prices move. If people think that a stock can not go below a specific level, this level is know as support, because the price always recovers when it reaches this point vice versa. These levels tell a user whether he should sell or buy a stock in order to make a profit. There are parameters that decide how sensitive these values must be and the user can select the preferred version of these levels (Fig. 3).

Support and resistance levels are shown using blue line. As you can see a price either drastically goes up or down when the price reaches these levels (Fig. 4).

Multiple Trading platforms do not include this feature because it causes loss of orders for the platform. This is because traders wait for a move that either results in the price going above or below the level. Hence during this period the traders avoid placing orders.

The model has been developed with the option of setting parameters that decide how sensitive the level(Lines) should be. Ex : A level of 60,000 rupees is consider to be an important support for MRF stocks as the price has rarely gone below it. Thus this level is considered very sensitive. The higher the magnitude of the sensitivity parameter, higher is the sensitivity of the displayed levels.

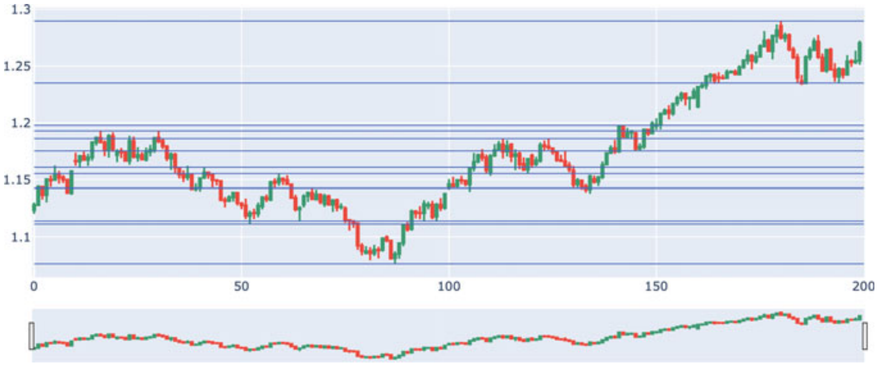

**Fig. 3** Support and resistance levels with less sensitivity

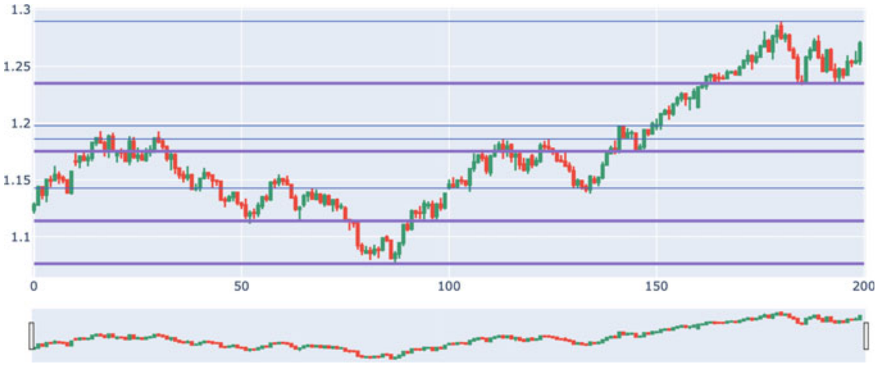

**Fig. 4** Support and resistance levels with more sensitivity

### *3.3 Bollinger Bands*

Original Bollinger bands give the lower and upper bounds for a stock to aid buy/sell decision. The upper bound is MA (Moving Average) + 2 × SD (Standard Deviation) and the lover bound is MA (Moving Average)—2 × SD (Standard Deviation). A stock is meant to be bought when it goes below the lower band and meant to be sold when it goes above the upper band. However it always bears a risk of losing money if it has been executed in a wrong way (Figs. 5 and 6).

Our Strategy: To reduce the risk the formula has been slightly altered to MA + (1.5 × SD)/MA—(2.3 × SD). By doing this the stock to be sold before reaching the upper bound of the actual Bollinger band strategy and the stock to be bought only if it touches a point which is less than the original lower bound strategy. There by reducing the risk and increasing the profit. However by doing this the number of buy/sell transaction will increase considerably which in turn will increase the brokerage.

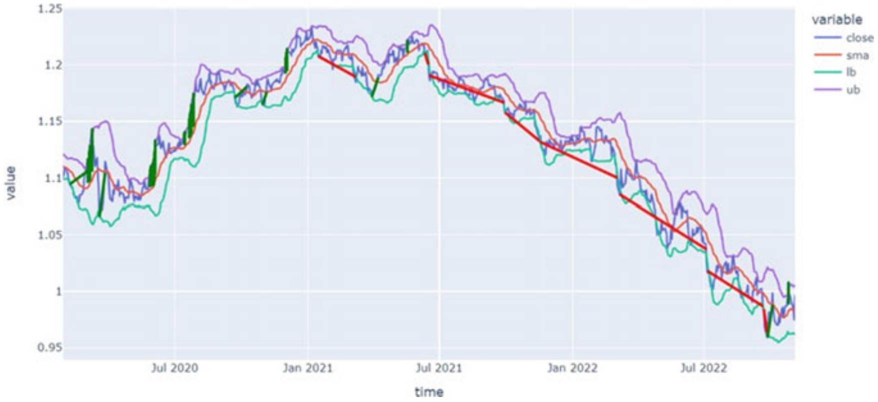

**Fig. 5** Original number of buys/sells

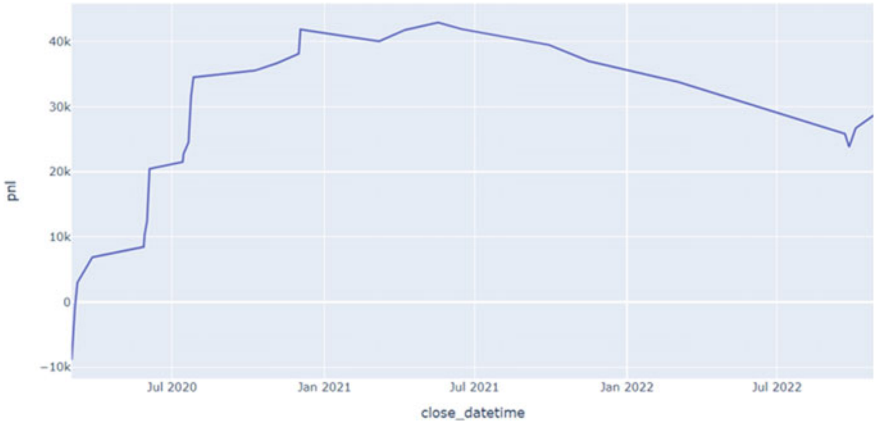

**Fig. 6** The profit graph by following the original strategy

This whole process of buying and selling can be automated in most cases (Figs. 7 and 8).

Profit is evident from the above graphs that the number of buys/sells increases by a large amount but so does the profit. It goes from 28k USD to 110k USD, all while reducing the risk of bollinger bands.

## 3.4 *Support Vector Resistance*

Support vector regression (SVR) is a machine learning model which is used to solve regression problems. Unlike classification problems, regression problems are those

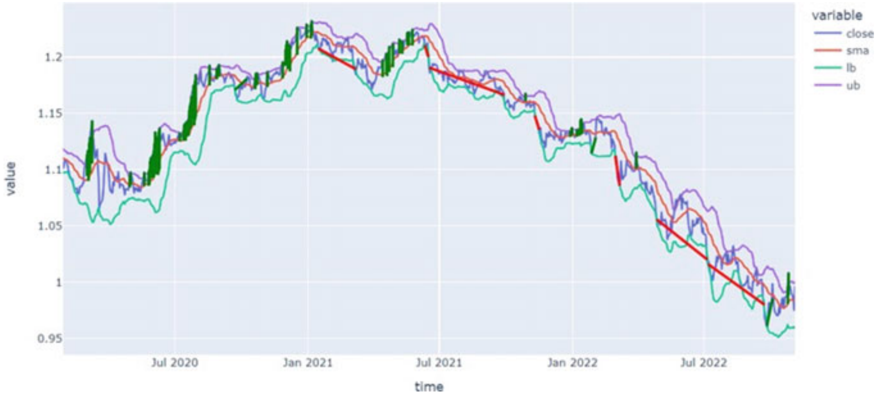

**Fig. 7** Number of buys/sells according to new strategy

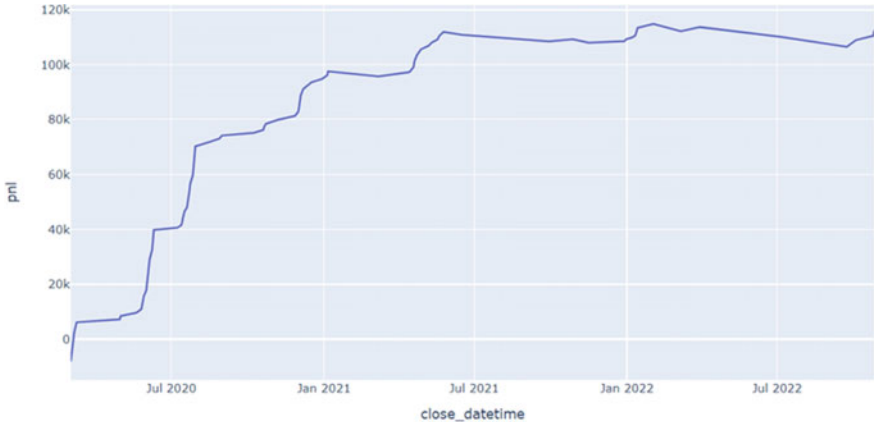

**Fig. 8** The profit graph by following the changed strategy

where a value for one variable—the dependent variable is derived from the value of another variable—the independent variable based on the relationship between them. Regression is usually used in cases where time series data is involved and tasks like forecasting, modelling and predictions need to be performed. In linear regression a linear equation is tried to be fit for the given data and cut back on the sum of squared errors as much as possible. This technique does not work with stock market data due to the nonlinear nature (seasonality) and volatile nature of the market. SVR aims at adjusting the general error bound rather than the observed error. The error margin is specified by epsilon, and we are given the ability to adjust it to control the accuracy of the model. We find 2 hyperplane margins and try to fit our data within this margin and reduce the points outside them. If a point is found outside a penalty (slack) is added to the objective function which helps the model perform better. Slack denotes the deviation of the data point from the already defined error margin. To adjust the number of points we tolerate outside the margins we add another hyper-parameter called C. C denotes the tolerance of the model for points outside the margins. The larger this value, the higher the tolerance. To equate to the non-linearity we make use of kernels which maps data from low dimensions to higher dimensions. With the gamma parameter we decide how much influence each training point have in the whole training process.

Implementation: We start with data generation. For the data we use historical data since 2018 to current day. We obtain the data using the yahoo API for python. To visualize the movement of the stocks we plot the data values along with the moving average curve Next we build our model. We are using the SVR available in the sklearn module. The kernel we are using is the rbf kernel. This kernel is the most general type of kernel and is widely. It works by transforming the data using the L2 norm between 2 points. We adjust the C and gamma values so that the model will be able to perform with a decent accuracy. We fit the SVR model on the dates and prices obtained from the yahoo finance API and predict the closing price for the next day.

The hyper-parameters used are tuned. The C value is set to a value of 1e4, and the gamma value is adjusted according to the stock for which price is calculated (Fig. 9).

The graph shows the actual prices of Microsoft stocks along with the moving average of the stocks. Moving average smooths out the data taking averages regularly with a moving window. The window chose here is for 60days. Moving average helps you understand the movement of stock prices and analyse them to seek profits.

The value predicted for Microsoft stock is 220.1927 while the actual price for the day was 221.39.

The problem with this model is that for each stock the gamma value needs to be calculated. This process is cumbersome. Also, over a period of time it is observed that these gamma values don't perform as expected. We observed that LSTM has better results compared to SVR and hence decided to use that for the price prediction in our applications.

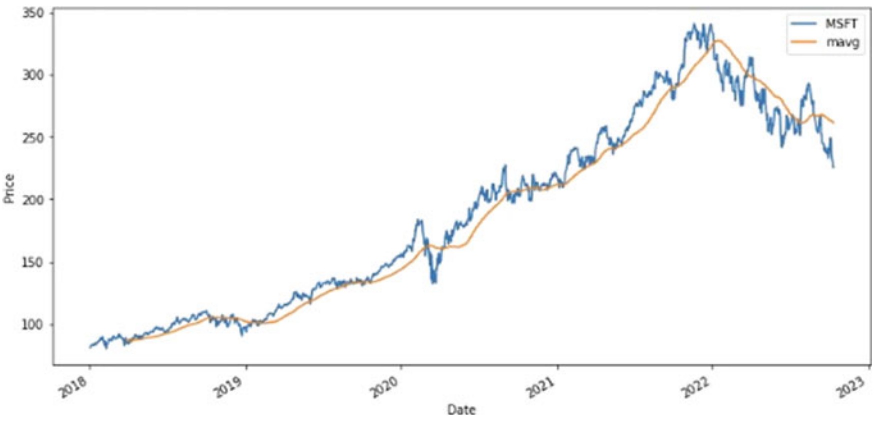

**Fig. 9** Graph of microsoft stock versus moving average

### 3.5 Portfolio Optimization

A portfolio is an aggregate of assets that include securities that are invested like stocks, mutual funds, bond, pension plans, real estate. It consists of assets that provide profited returns. Optimization is the process of adjusting the variables of the analysis such that returns are maximized. This can be done by reducing the risks in each of the transactions such that the minimal risk for each asset is maintained while profit is maximized. Portfolio optimization is a technique of choosing an optimal distribution of assets that could lead to a maximal profit with the minimal risk factors out of all options of distributions being considered in accordance with an objective function defined beforehand. In order to provide suggestions to the new users optimized portfolio would be a great help. Therefore, implementation of an optimized portfolio for the end users would help them make better decision. Its implementation can be done in many ways. One of the ways is Reinforcement Learning. It is the concept concerned with the developing of an AI that consists of the action of intelligent agents that decisions in a well-defined environment that maximize the cumulative reward. In RL problem, the agent is associated with the environment considering the observations of the input and taking actions based on it with the motive of maximizing the cumulative rewards or outputs of the problem. This comprises of many actor-critic algorithms ensemble together to solve the problem of optimized portfolio. The main constituents of reinforcement learning are Agent Environment Reward Agent: The aim of the reinforcement learning is making the agent learn how to complete the defined task within an undefined environment based on the observations it makes at given time intervals. Based on the action it takes, the agent receives rewards, either positive or negative based on how successful the action is with respect to completing the given task. It consists of two components i.e., policy and a learning algorithm. A policy is the set of decisions that the agent can take based on the current observation of the environment with a probability distribution of how important each of the action would be in order to complete the task given. The policy is made up of various approximators with tune-able parameters and a model for approximation like a deep learning model. Approximators used here are actor and critics. The actor returns the optimal action that can be taken at any interval of time that will maximize the reward over a long term whereas the critic returns the predicted cumulative value of the reward in order to accomplish the given task. The learning model here constantly updates the policy parameters based on the environmental variables with its main aim being to find a policy that can maximize the long-term reward received in order to achieve the goal. Environment: It is the domain where the agent resides and interacts. The agent can perform any actions based on its observation at different time intervals, but it cannot impact or control the rules that govern the environment. The agent is sent into a new state of the environment once it acts into the previous state based on its actions and the rewards and feedback of those actions. Rewards: It describes the expected behaviour of an agent in the environment in order to fulfil the task assigned. Positive rewards are given when the agent performs an action that is needed to reach the goal and negative rewards to discourage any action that does not reach the goal.

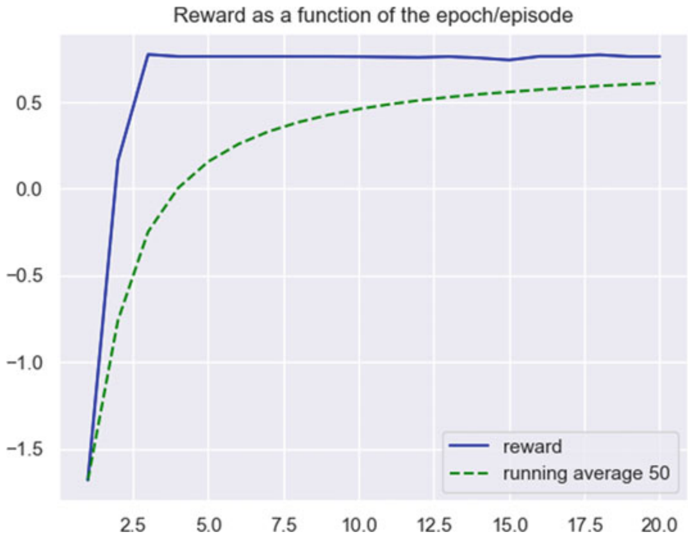

**Fig. 10** Training the agent over 20 epochs with initial portfolio value as 0

A precise reward function directs the agent in order to maximize the expectation of long-term reward (Fig. 10).

Thus, solving the problem of portfolio optimization is solved using reinforcement learning. The AI designed gives the number of stocks of each ticker that needs to be owned and can lead to maximal profits based on the stock values which helps the end user have a clear picture of how much he needs to invest in each of the stock in order to yield maximal profits.

### *3.6 LSTM*

Artificial RNN called LSTM Is a combination of both “long term” and “short term” memory. This is one of the best models for short-term prediction. We have used past 60 day closing data points to make prediction for the next day. LSTM can work with just 60 data points however it holds into account the previous data as well and predicts accordingly. To make computations faster a modulated version of LSTM has been used with 2 dense layer and 50 hidden neurons, unlike the existing LSTM models this model can be executed much faster and gives good approximation.

The RMSE of this model can be set by the manually but for demonstration purpose the RMSE value is set to less than 3. Reducing the RMSE increases computation and through many trial and error efforts it is found RMSE 3 to be a optimized approximation.

The data has been fetched from Yahoo Finance API and the model works for all US stocks, with lag less than 1 second. The closing price of the stock is pulled

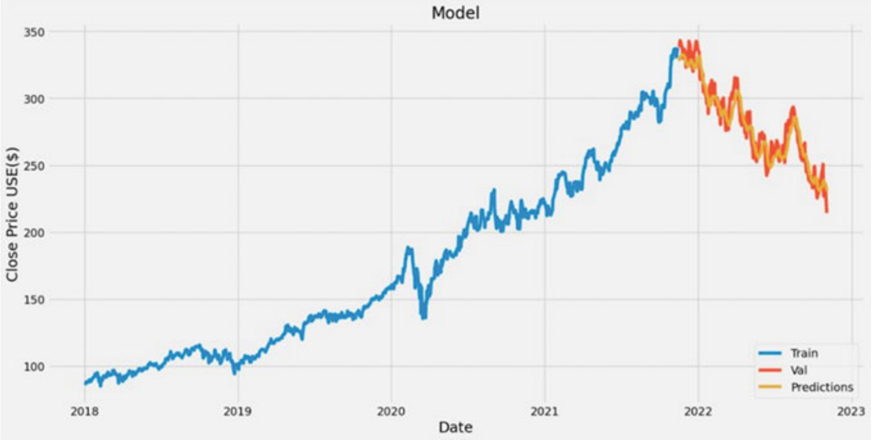

**Fig. 11** Predicted values versus actual values for microsoft

for each day starting from 2018-01-01 to the current previous day to predict for the upcoming day. The data is divided into 2 parts, 80% for training the parameters and 20% for testing (Fig. 11).

If RMSE is found to be greater than or equal to 3 while testing the model is re-trained automatically. The model can be easily extend for not only stocks but also crypto currency given a good API and ticker. Since only the prediction for the next day is given, running the model once a day is enough.

The predicted result for 2022-11-05 was $ 227.457 and the actual closing price was $ 225.89.

## 4 Auto-ARIMA

ARIMA is a statistical model that mainly uses time series data to predict future values. It uses stationary data in order to achieve this but sometimes data collected may not be stationary therefore ARIMA uses a parameter called d that equals to the number of times differencing transformation that has been done. AUTO ARIMA model from pmdarima has a reliable reputation and therefore it has been used. AUTO ARIMA automatically identifies the most optimal P,D,Q values and predicts using the same.

To assess how well the model has fit, few statistical tests and indicators have been used:

- Dickey-Fuller test—This test states if a unit root is present or not. If a unit root is present then the data is non stationary. For our data, it is found that we did have a unit root and hence there has to be a differencing (d) parameter in order to convert the data to stationary data.

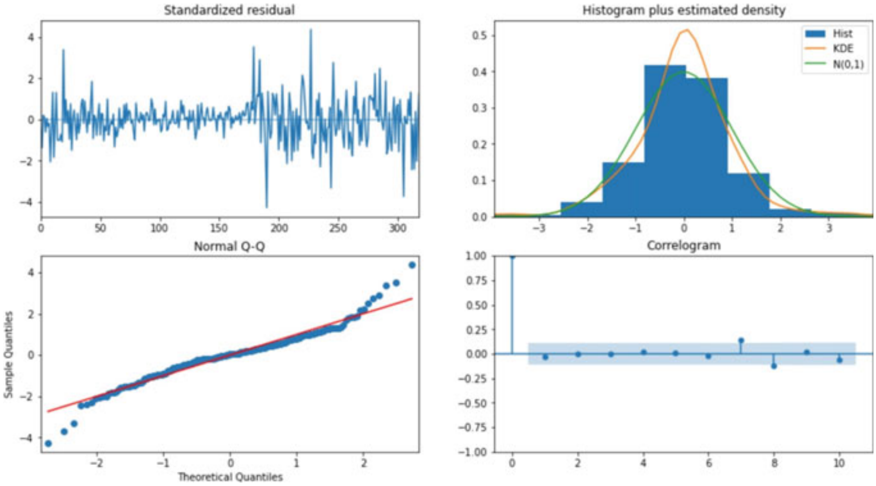

**Fig. 12** A diagnosis of the model show that all the plots show satisfactory results

- The ACF plot has been used to confirm the order of AR and MA decided by the AUTO ARIMA model.
- Breaking down of the data into seasonality, trends and noise show us the presence of seasonality and trends in the data and hence once again confirms the presence of non stationary data.
- To check the result and fitting of the model the inbuilt feature for diagnostics was used. The following plots displayed are the results (Fig. 12).

A look at these plots show the model has fit quite well.

To be confident that the model works well with majority of the stocks, SP 500 Exchange traded funds closing price has been used. The reason for this is that the SPY index(SP 500 Exchange traded funds) is the culmination of 500 biggest stocks. Even though the tests show that the model has fit well, it performs bad at predicting closing values. It has been unreliable at identifying trends as well.

An RMSE value of 24.939 tells that the model is not performing very well and is hence unreliable. This maybe due to the way it selects P,D,Q values. The values it selects are not the most optimal ones and are hence giving this results. The model does not work well with short-term-prediction.

## 5 Result and Analysis

- Support and Resistance is a good strategy that should be implemented by every user while trading as it reduces the risk of losing cash.

- Bollinger bands with the new formula is preferable over the original formula as it yields more profits with lower risk. However it increases the number of buys and sells which can increase brokerage charges. The whole buy/sell process can be automated.
- SVR and ARIMA did not perform as expected for short-term-prediction. The problem with this model is that for each stock the gamma value needs to be calculated. This process is cumbersome. Also, over a period of time it is observed that these gamma values don't perform as expected.
- SVR and ARIMA also took more time and resources to produce results and is not a good model to be followed by laymen. These models however perform exceptionally good with big brokering firms who have heavy resources and prefer long-term-prediction.
- In our study LSTM was found to be the best model for short-term-prediction as the RMSE value could be manually set. Computation was much faster as it used only the previous 60 data points to predict once the model was trained.
- Optimal portfolio was found to be a good add-on to the trading world.

## 6 Conclusion

The old platforms for pseudo-trading are outdated. They only provide users with an option to buy/sell stocks without a brokerage system, which is far from reality. Their aim is to get the users into actual trading to benefit from the brokerage.

In this paper we have used modern ways to inference information from the graph through models like Bollinger Bands and Support and Resistance. We have also provided the user with an interface to buy/sell shares using some limited points along with predictions for the stocks. Optimal portfolio is another smart functionality we have implemented to aid users in their buy/sell decisions.

For price prediction layman should avoid using ARIMA and SVR as their results are off-putting and usually takes longer time in training. Out of all the models, tuned-LSTM with 2 dense layers and 50 neurons gave the best prediction in a shorter time. It was easier to train the data. The model is automatically trained to a point where the RMSE is less than 3 hence this is the most suited ML model for short-term-prediction using commodity hardware.

## References

1. Lim JY, Lim KM, Lee CP (2021) Stacked bidirectional long short- term memory for stock market analysis. In: 2021 IEEE international conference on artificial intelligence in engineering and technology (IICAIET)
2. Shi M, Zhao Q (2020) Stock market trend prediction and investment strategy by deep neural networks. In: 2020 11th international conference on awareness science and technology (iCAST), pp 1–6. https://doi.org/10.1109/iCAST51195.2020.9319488

3. Ozbilen ML, Yaslan Y (2021) Portfolio construction with stock prices¨ predicted by LSTM using enhanced features. In: 2021 6th international conference on computer science and engineering (UBMK), pp 639–643. https://doi.org/10.1109/UBMK52708.2021.9558889
4. Vij A, Saxena K, Rana A (2021) Prediction in stock price using of python and machine learning. In: 2021 9th international conference on reliability, infocom technologies and optimization (trends and future directions) (ICRITO), pp 1–4. https://doi.org/10.1109/ICRITO51393.2021.9596513
5. Isha S, Dixit MK, Ahirwar D, Sakethnath S, Rakha M (2021) Stock prediction by analyzing the past market trend. In: 2021 9th international conference on reliability, infocom technologies and optimization (trends and future directions) (ICRITO), pp 1–4. https://doi.org/10.1109/ICRITO51393.2021.9596263
6. Yang W, Wang R, Wang B (2020) Detection of anomaly stock price based on time series deep learning models. Manage Sci Inform Econ Innov Develop Conferen (MSIEID) 2020:110–114. https://doi.org/10.1109/MSIEID52046.2020.00029
7. Wang Y, Guo Y (2020) Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost in China. Communications 17(3):205–221.https://doi.org/10.23919/JCC.2020.03.017
8. Zhou T, Zhang W, Ma S (2021) Tidal forecasting based on ARIMA-LSTM neural network. In: 2021 33rd Chinese control and decision conference (CCDC), pp 4028–4032. https://doi.org/10.1109/CCDC52312.2021.9601933
9. Bowen T, Zhe Z, Yulin Z (2020) Forecasting method of e-commerce cargo sales based on ARIMA-BP model. IEEE Int Conferen Artific Intell Comput Applicat (ICAICA) 2020:133–136. https://doi.org/10.1109/ICAICA50127.2020.9181926
10. Qin J, Tao Z, Huang S, Gupta G (2021) Stock price forecast based on ARIMA model and BP neural network model. In: 2021 IEEE 2nd international conference on big data, artificial intelligence and internet of things engineering (ICBAIE), pp 426–430. https://doi.org/10.1109/ICBAIE52039.2021.9389917
11. Gasˇperov B, Sˇaric´ F, Begusˇic´ S, Kostanjcˇar Z (2020) Adaptive rolling window selection for minimum variance portfolio estimation based on reinforcement learning. In: 2020 43rd international convention on information, communication and electronic technology (MIPRO), pp 1098–1102.https://doi.org/10.23919/MIPRO48935.2020.9245435
12. Seshu V et al (2020) Performance analysis of bollinger bands and long short-term memory (LSTM) models based strategies on NIFTY50 companies. In: performance analysis of bollinger bands and long short-term memory(LSTM) models based strategies on NIFTY50 companies, IEEexplore