LSTM for Return Prediction and Portfolio Optimization in America Stock Market

Kailiang Chen[1, †], Zheng Guo[2, †], Xin Huang[3, †], Yulan Jin[4, *, †]
201900820197@mail.sdu.edu.cn, Zheng.Guo20@student.xjtlu.edu.cn, huangxin888@stu.xjtu.edu.cn,
[*]Corresponding author: 19030102018@st.zisu.edu.cn

[1]School of Statistics, Shandong University, Weihai, China

[2]International Business School Suzhou, Xi'an Jiaotong-liverpool University Suzhou, China

[3]School of Mathematics and Statistics, Xi'an Jiaotong University Xi'an, China

[4]International Business School, Zhejiang International Studies University, Hangzhou, China

[†]These authors contributed equally.

Abstract—With the development of the financial sector, asset portfolios have become an important part of the financial industry but using the traditional tools to complete asset portfolios is inefficient. In the paper, an asset portfolios system which uses machine learning as main analysis method was designed, aiming to get asset portfolios quickly and precisely. More specifically, firstly, this study used random forest model to process data to get variables importance, observing every variable' s contribution and importance in prediction part. Then, based on known data, Long short term memory (LSTM) model was employed to predict future stock returns which was used as data base of asset portfolios. After that, this study combined prediction result with real data and put it into equal-weight allocation model to obtain annual return and annual volatility. Finally, efficient frontier was carried out using Monte Carlo stimulation, calculating the sharp ratio and the best weight over chosen stocks. By this process, this paper finally completed the asset portfolios and built the asset portfolios system. The result shows that the asset portfolios system, combination of random-forest model and LSTM model, can do the asset portfolios efficiently and precisely.

Keywords; Return prediction; Portfolio optimization; Machine learning; LSTM; Random forest

1 INTRODUCTION

Asset allocation is one of the most important procedures in investment strategy and will be the principal determinants of investment results. There are three main asset classes—equities, fixed-income, and cash and equivalents [1]. The opportunity to gain high profits has led investors to favor stock trading [2]. Energy sector is an attractive field in this market, especially new energy,

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328431

owing to the increasing awareness of protection of environment. Stock prediction and portfolio optimization are critical issues in stock market. Investors try to predict the stock's future price to select from a huge volume of options so that optimum profit can be earned [3]. Portfolio optimizing helps investors to improve return while minimizing risk levels [4]. However, prediction analytics and portfolio optimization models are dynamic because of the nature of stock data such as non-linear, non-stationary and volatile [5]. During the 2020, U.S. stocks triggered 4 circuit breakers with a decline of more than 30% due to convid-19 [6]. Also, energy sector saw a continually rise during the Russian-Ukrainian War after great downfall in pandemic [7]. It can be observed that these two events have changed the patterns and breached previous regulation of stock prices. Therefore, new predictive model and new advice on prediction-based portfolio are required [8].

As an important part of asset allocation, the area of stock prediction and portfolio optimization has been developed over decades years. Began at 1930s, the key to successful investment is valid asset allocation [9]. In the past, a lot of research revolve round the investor. For instance, the authors study an investor's optimal consumption and portfolio choice problem when they are confronted with two possibly mis specified sub models of stock returns: one with IID returns and the other with predictability by using a generalized recursive ambiguity model to accommodate the investor's aversion to model uncertainty [10]. In recent several years, area of stock prediction and portfolio optimization turn its research directions from past theory innovation to model micro-innovations and combining with machine learning. In [11], the author proposed that financial network indicators could be applied for global stock market investment strategies, and author achieved model via several machine learning approaches (i.e. logistic regression, support vector machine, and random forest).

In [12], authors used random-forest to predict the stock returns mainly based on the various consumption-based and income-based inequality data. However, results suggest that the random-forest does not have a good prediction performance. Arjun Singh Saud et al. used three deep learning models to conduct price prediction performance and found that Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) had better performances than Variational RNN (VRNN) did in forecasting stock price [13]. They also researched on the values of the look-back period used with each model. Their work provided our research with a reliable choice of predictive model as well as a suggested parameter of look back period on LSTM. However, the research work stopped at the predicting stock price. Besides, this research only focused on the stock price which is less intuitive than stock return in stock market. AA Adebiyi et al. compared the predictive performance of Auto Regression Integrate Moving Average (ARIMA) and artificial neural network (ANN) model and revealed a superiority of neural networks model over ARIMA model [14]. However, according to [7], ANN is not the best choice for stock prediction.

In this regard, this paper decided to use 5-year direct data of stocks (i.e., open price, close price and so on) in predicting rather than inequality data to cover the pre-epidemic period and post-epidemic period. Random forest was proposed to find variables importance (VIMP) which allows users to identify which variable plays a key role in prediction [15]. This paper used LSTM as predictive model which is good at handling long-term dependencies as well as remembering immediate past values that ANN is not able to deal with [1], and then conducted a portfolio allocation based on the predicted stock returns and gained two set of portfolios with efficient frontier, which shows more guidance to investor to optimal the portfolio of five assets.

Table 1 The original sample data OF PLUG

Date	Plug stock					
	Open	High	Low	Close	Adj Close	Volume
2017-01-03 00:00:00	1.200000048	1.230000019	1.190000057	1.210000038	1.210000038	2075600
2017-01-04 00:00:00	1.210000038	1.299999952	1.200000048	1.299999952	1.299999952	2930500
2017-01-05 00:00:00	1.299999952	1.320000052	1.25999999	1.25999999	1.25999999	3506400
2017-01-06 00:00:00	1.279999971	1.289999962	1.25	1.25	1.25	1861100
2017-01-09 00:00:00	1.25999999	1.269999981	1.220000029	1.25	1.25	1386100

Table 2 The Original Sample Data OF DVN

Date	DVN stock					
	Open	High	Low	Close	Adj Close	Volume
2017-01-03 00:00:00	46.81999969	47.56000137	45.61999893	47.08000183	39.80078125	5828600
2017-01-04 00:00:00	46.93999863	47.52999878	46.49000168	47.5	40.15584564	3706200
2017-01-05 00:00:00	47.72999954	49.45000076	47.66999817	49.00999832	41.43238068	7513000
2017-01-06 00:00:00	49.02999878	49.25	48.43000031	48.66999817	41.14495087	3193900
2017-01-09 00:00:00	48.06999969	48.18000031	46.50999832	46.58000183	39.3780899	5736100

2 METHODOLOGY

2.1 Dataset prepration and preprocessing

In this study, we picked 5 stocks through a package of python called yfinance (TSLA, DVN, OXY, PLUG, MRO), which are related to energy resource market. Then we got features of the data (with time series) in all directions. Table 1 and Table 2 provides the original sample data of stock of DVN and PLUG.

The preprocessing is consisted of four parts. First, we did the outlier test and deleted the null value since we have to ensure the dataset is valid. After that we drew simple sequence diagram for all five stocks to describe data directly. Figure 1 and Figure 2 provides the example called DVN and PLUG.

Then the data were proportioned into the training and test sets at a ratio of 7:3. Finally, the data of training and test set were normalized by the method of z-score respectively which used mean and standard deviation of the data.

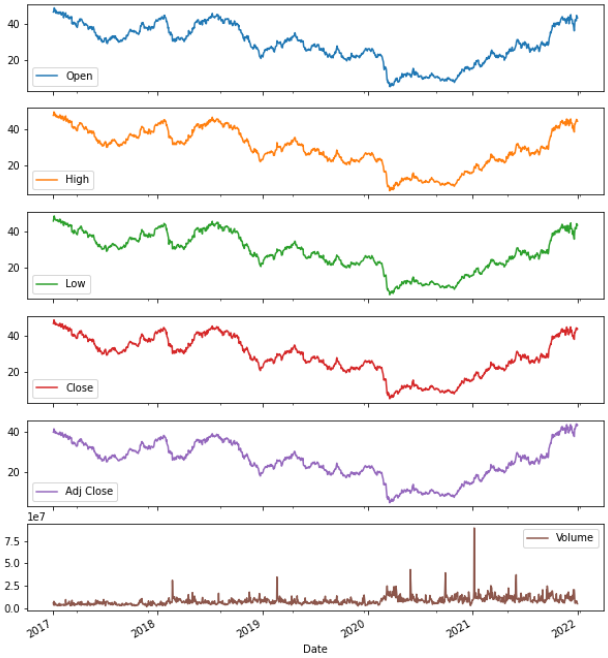

Figure 1. The data distribution of DVN.

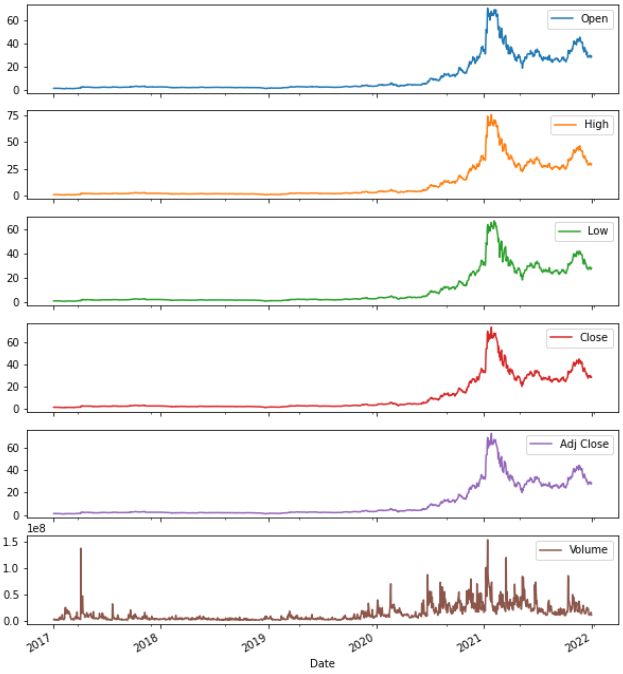

Figure 2. The data distribution of PLUG.

2.2 Artificial nerual networks

Artificial neural networks (ANNs) are biologically inspired computational networks which simulates the electricity activity of the brain and nervous system [16]. Multilayer perceptron (MLP), the ANNs most commonly used for a wide variety of problems is shown in Figure 3.

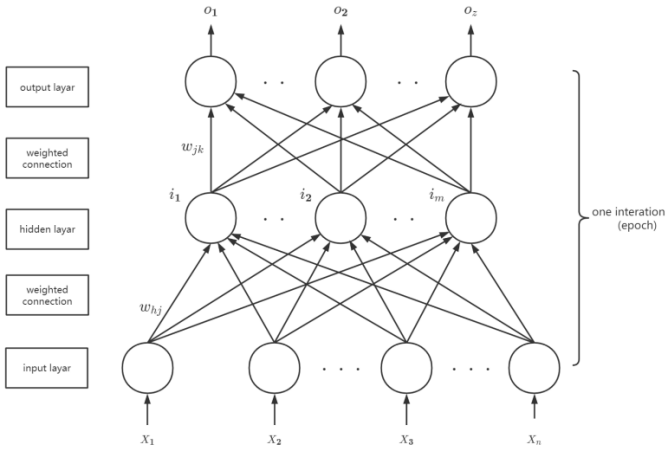

Figure 3. The structure of multilayer.

The bottom layer is called input layer where values are input (x_1, …, x_n). The values of data participate in a certain nonlinear activation function like sigmoid, hyperbolic-tangent when getting through the hidden layer with m number of neurons (i_1, …, i_m). The z number of output (o_1, …, o_z) which are stood by output neurons come out from the output layer. w_{hj} is a connection weights which connects the input layer with hidden layer and w_{jk} is a transfer function that links the hidden layer to output layer. These stages will get updated while the whole interaction(epoch)is repeated thousands of times until the output value meet the standard level of accuracy.

2.3 Recurrent Neural Network and Long Short-Term Memory

RNN is a type of ANN which has connection between hidden neurons addition to that between layers as depicted in Figure 4. RNN outperforms in remembering immediate past values. By contrast, LSTM is a special RNN which is advanced in learning long-term dependencies by some connected memory blocks and each block consists of memory cells and three gates - input gate, output gate and forget gate that enable the machine to protect and control information. This work proposed LSTM to predict stock returns of five assets which is shown in Figure 5.

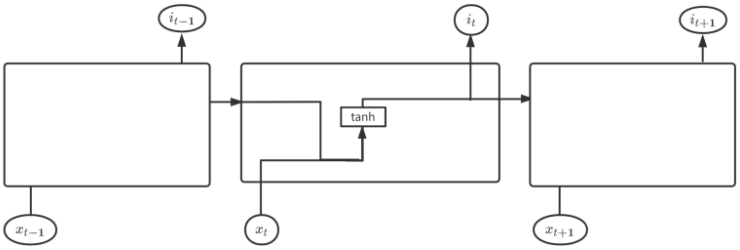

Figure 4. The structure between hidden neurons of RNN.

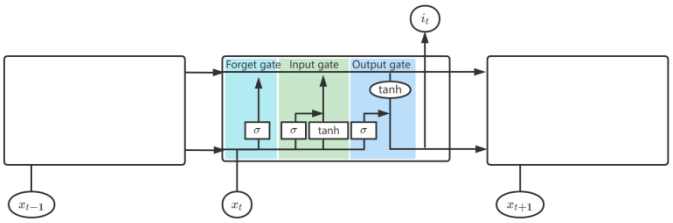

Figure 5. The structure between hidden neurons of ANN.

2.4 Implementation details for LSTM

1) Parameters identifying. The number of neurons in input layer is6 according to 6 features of stock being put into the machine. The only output is return so the output size is 1. The number of hidden layer and hidden neuron are 2 and 13, respectively.

2) Model designing. A python package called TensorFlow was used to construct the model conveniently. With the help of this package two connection weights were chosen randomly. The value of Forget bias which influenced the level of relative error is 0.5. Some operations were conducted to change the dimensionality of the values in order to let outcome be carried to another further layers and flow of the algorithm of the machine. A dropout function has also been made to avoid overfitting.

3) Model training. As for training model, the batch size was 60 and time step was 20. Learning rate is 0.0006. A loss function that was designed based on the squared error is shown in Eq.1. $\hat{\theta}$ is the predicted value and θ means the real value.

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \tag{1}$$

4) Predicting. The prediction is conducted based on the 30% of the data set and the error between real values and predictive values was measured by the quotation in Eq.2. X stands for the predicted value, μ means the real value, $X - \mu$ is the absolute error.

2.5 Implementation details for Random Forest

We designed and trained the random-forest (rf) model. The rf model is designed to get the importance of variables which measures the contribution of each variable to the predicted outcome.

$$\frac{X-\mu}{\mu} \times 100\% \quad (2)$$

The rf model is a popular model in both prediction area and classification area, which is widely used in financial markets e.g., stocks and bonds and scientific fields e.g., atmospheric sciences.

The principle of rf model is ensemble learning. Firstly, the system will create a kind of estimators called decision-tree, and each decision-tree is consisted of nodes (used for processing data) and edges (representing outputs of different conditions like yes or no, raise or fall). Once input data is received, the data will be processed by the first node and decided by node's specific conditions, and then data is outputted from the edge corresponding to the specific condition and transported to the next node connected with this edge. The process will not stop until the data reaches the final non-split node (called leaf nodes), and the final node represent the final decision result of this decision tree. Then, many estimators (i.e., decision-tree) will combine forming a 'forest', which is rf model. Every estimator in 'forest' is equal and mutually independent. After processing data, every estimator will obtain its own result, and every estimator will publish the result and give one vote to this result. Finally, the rf model will count the number of votes for each result, and the result with highest vote number will be final result of rf model.

In this paper, we build a rf model with parameters, 700 n_estimators, None max_depth, 6 min_samples_split and 3 min_samples_leaf. And we split stock data into two parts, train-data and test-data, and use train-data as input data to train rf model. After training, the rf model had a general grasp of data rules. Then we input the test-data into rf model to do prediction. Finally, the rf model got the variables importance from the prediction result.

3 RESULTS AND DISCUSSION

3.1 Random forest

The rf model get the importance of every variable in input data. There we use two stocks, PLUG and DVN, as example shown in Figure 6 and Figure 7. In the result, variable 'Volume' has the largest importance. Then, importance of variable 'Open' is second. And importance of variables 'High', 'Low', 'Close' and 'Close_adj' are roughly the same. The importance of variables 'Year', 'Month' and 'Day' are small enough to ignore.

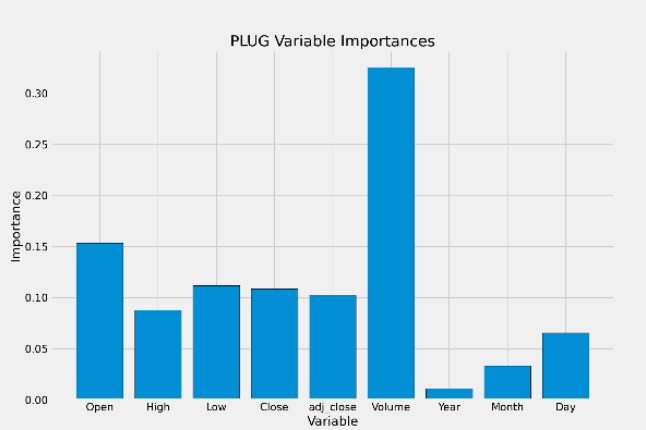

Figure 6. PLUG variable importance.

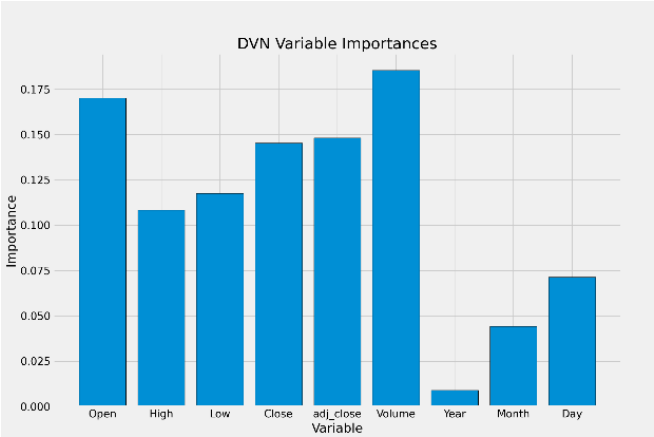

Figure 7. DVN variable importance

The variable importance shows the contribution of each variable to the prediction. Firstly, the 'Volume' is the most important variable, and it is an important indicator measuring future stock prices trend and market attitude to this stock. In general, if volume and price of one stock raise, it means investors are bullish on this stock, and vice versa. In the prediction, there are lots of huge fluctuations forming steep crests and troughs. The model is hard to predict these fluctuations precisely unless using 'Volume' as the key determinants. Then, the 'Open' is second important variable and it is always using with volume together in prediction process. The main significance is that unnormal 'Open', like Parity bulk opening price or jump high opening price and jump low opening price. This information can help model analyses and predict large capital's intentions to later development of stock. Finally, the remaining four variables ('High', 'Low', 'Close' and 'Close_adj') provide basic data support to do continuous prediction and help model predict more precisely.

3.2 Long Short-Term Memory (LSTM)

After performing LSTM model with adjusted parameters, we got the following Figure8, Figure 9, Figure 10 and Figure 11 as results. And we choose two stocks DVN and PLUG to illustrate. Figure 10 and Figure 11 show that the loss value decreases with higher iteration numbers. And it can finally reach a relative low loss value around 0.5 with fewer than 200 times of iteration. And it can be seen from Figure 8 and Figure 9 that the predicted value coincides with the real value to a large extent. Though the relative error is small, we found that the prediction result is better if using price directly as the training data, with smaller relative error. However, we consider return more significant, and the prediction of return value is more meaningful. So, we still used return as training data set and the error rate reflects the reality that there is possibility for difference between prediction and real value.

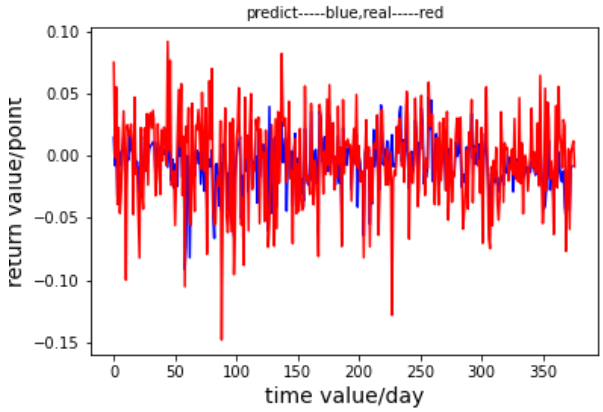

Figure 8. The predicted values against actual values for DVN stock.

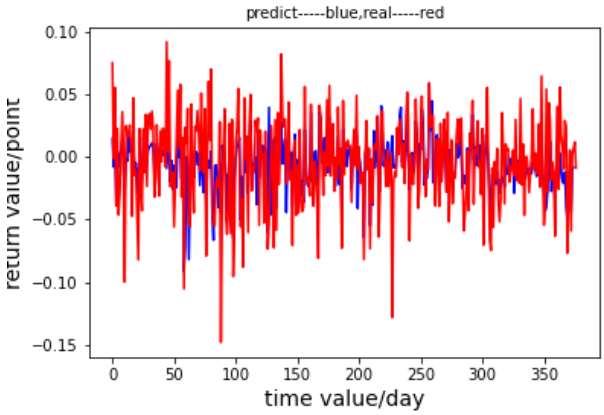

Figure 9. The predicted values against actual values for PLUG stock

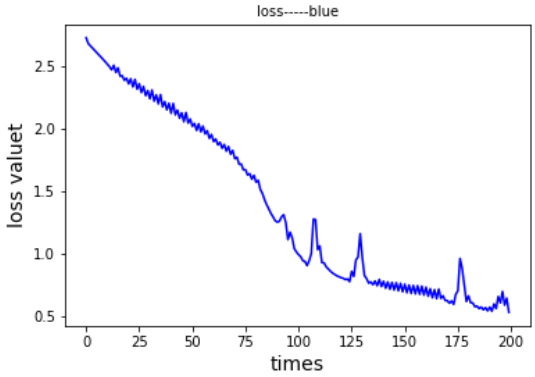

Figure 10. The loss value and the number of iterations of DVN stock

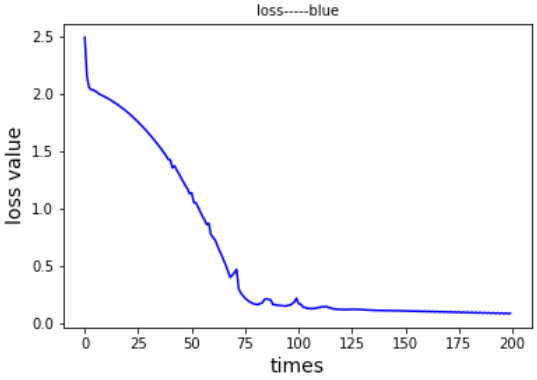

Figure 11. The loss value and the number of iterations of PLUG stock

3.3 1/n portfolio allocation

After getting the predicted data, we used real returns of last 5 years plus predicted returns of the following 6 months to be the input in the equal-weight allocation model. And then we get relative data showed in Table 3. The annual return is 51.11% and annual volatility is 42.392%, which is to be compared to the result of efficient frontier method.

Table 3. Back test data of 1/n portfolio allocation

Start date	2017-01-04
End date	2022-09-12
Total months	68
Back test	
Annual return	51.11%

Cumulative returns	956.391%
Annual volatility	42.392%
Sharp ratio	1.19
Calmar ratio	0.79
Stability	0.70
Max drawdown	-65.0915
Omega ratio	1.27
Sortino ratio	1.74
Skew	-0.83
Kurtosis	18.60
Tail ratio	1.13
Daily value at risk	-5.14%

3.4 Efficient Frontier

We used Monte Carlo stimulation to calculate the sharp ratio and find the best weight over the 5 stocks as presented in Figure 12.

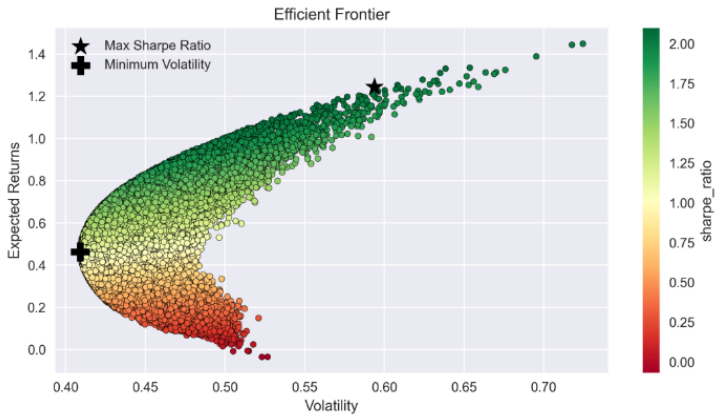

Figure 12. The efficient frontier of predicted portfolio

Table 4. Performance of Maximum Sharpe Ratio portfolio

returns	**124.52%**
volatility	**59.37%**
Sharpe Ratio	**209.76%**
Weight of DVN	**2.40%**

Weight of MRO	**2.46%**
Weight of OXY	**32.01%**
Weight of PLUG	**63.00%**
Weight of TSLA	**0.12%**

4 CONCLUSION

In this work, an advanced deep learning model called LSTM was proposed for prediction and constructed a portfolio optimization by using predicted results. Additionally, random forest was used to obtain the importance of variables. It is found found that LSTM shows a good predictive capability in energy sector of America stock market and enables the value of relative error to be reduced to an acceptable level. Besides, it is believed that there were significant changes happening in stock market after the outbreak of covid-19, so this study chose a 5-year period data set for experiments which is sufficient enough to cover the point when pandemic happened in America. It is found that two brand new types of portfolios which can be matched to different preference of investors by comparing the outcomes of two weight allocation models. In the future, further study plans to adapt the proposed method to more stocks of all kinds and improve the prediction by adopting some hybrid predictive models to get better prediction and predicted portfolio.

REFERENCES

[1] Investopedia, “Asset Allocation,” https://www.investopedia.com/terms/a/assetallocation.asp, 2022

[2] S. Arjun, S. Subarna, “Analysis of look back period for stock price prediction with RNN variants: A case study on banking sector of NEPSE,” Procedia Computer Science, Volume 167, 2020, pp. 788-798, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.419.(https://www.sciencedirect.com/science/article/pii/S1877050920308851)

[3] L. Ge, M. Wenping, “A quantum artificial neural network for stock closing price prediction”, Information Sciences, Volume 598, 2022, pp. 75-85, ISSN 0020-0255.

[4] R. Bhuyan, et al. "Optimal portfolio allocation among REITs, stocks, and long-term bonds: An empirical analysis of US financial markets." Journal of Mathematical Finance 2014, 2014

[5] S. Kumar, “Hybrid models for intraday stock price forecasting based on artificial neural networks and metaheuristic algorithms”, Pattern Recognition Letters, Volume 147, 2021, pp.124-133, ISSN 0167-8655.

[6] Sohu, “A review of Chinese and American stock markets in 2020”, https://www.sohu.com/a/442513132_120585360, 2021.

[7] ChinaNews, “The energy sector has led the S&P 500 in gains this year”, https://www.chinanews.com.cn/cj/2022/03-25/9711423.shtml

[8] M. Akhter, “LSTM-based Deep Learning Model for Stock Prediction and Predictive Optimization Model, EURO Journal on Decision Processes”, Volume 9, 2021, 100001, ISSN 2193-9438.

[9] J. Zhao, M. Zhou. “A Review of the Research on Asset Allocation Theory”, 2017 International Conference on Sports, Arts, Education and Management Engineering (SAEME 2017). Atlantis Press, 2017, pp. 398-401.
[10] H. Chen, N. Ju, J. Miao. “Dynamic asset allocation with ambiguous return predictability”. Review of Economic Dynamics, 2014, 17, pp. 799-823.
[11] T. Lee, et al. "Global stock market investment strategies based on financial network indicators using machine learning techniques." Expert Systems with Applications, vol. 117, 2019, pp. 228-242.
[12] R. Gupta, et al. "The predictive value of inequality measures for stock returns: An analysis of long-span UK data using quantile random forests." Finance Research Letters, vol. 29, 2019, pp. 315-322.
[13] Y. Ma, H. Ruizhu, and W. Weizhong "Portfolio optimization with return prediction using deep learning and machine learning." Expert Systems with Applications, vol. 165, 2021.
[14] A. Adebiyi, O. Aderemi, and K. Charles, "Comparison of ARIMA and artificial neural networks models for stock price prediction." Journal of Applied Mathematics, 2014.
[15] H. Ishwaran, and L. Min"Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival." Statistics in medicine, vol. 38, 2019, pp. 558-582.
[16] W. Steven, C. Narciso, “Artificial Neural Networks,Editor(s): Robert A. Meyers, Encyclopedia of Physical Science and Technology (Third Edition), Academic Press, 2003, pp. 631-645.