NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY
[IJIERT] ISSN: 2394-3696 Website: ijiert.org
VOLUME 8, ISSUE 6, June. -2021

STOCK PRICE PREDICTION USING STATISTICAL, MACHINE LEARNING AND DEEP LEARNING MODELS

Neda Yousefi
Faculty of Computer Science, Allameh Tabataba'i University, Tehran, Iran.
neda.yousefii@gmail.com

ABSTRACT

Forecasting stock price is a challenging topic for the researchers by the way of statistics or in newer version by the way of Machine Learning and Deep learning. There are researches that prove that the direction of time series for a stock price can be predicted with a good accuracy.

Design of this kind of predictive models requires choice of appropriate variables, right models and methods, and tuning the parameters. In this research, the goal is applying different algorithms and approaches for stock price prediction then compare and evaluate them together. This research also aims to apply these models for short-term stock price prediction.

The daily stock price data is used, from tsetmc for five biggest Iranian companies active in the Tehran stock Exchange. For each stock, the data is gathered from January 1, 2018 to January 1, 2021 to train, test and evaluate the algorithms.

For evaluating performance of all models, three measures are used for performance evaluation including The Mean Absolute Error (MAE), The Mean Squared Error (MSE), and The Root Mean Squared Error (RMSE). Statistical, Machine learning and deep learning approaches that used are include Auto-Regressive Integrated Moving Average (ARIMA), Random Forest (RF), Support Vector Machines (SVM), and Long Short-Term Memory (LSTM) Networks.

Results for applying those models are tested and evaluated and the LSTM model has the best performance concerning other approaches and prove outperformance.

Keywords: Machine Learning, Deep learning, Stock price prediction; Auto-Regressive Integrated Moving Average (ARIMA), Support Vector Machines (SVM), Long Short-Term Memory (LSTM), Random forest (RF)

INTRODUCTION

Forecasting Stock price and prediction of stock price direction have been very interesting topics for many analyses, researches and projects. Stock price prediction is a complex problem because of the number of different factors that can affect in this process. It is possible to group stock market analysis as Fundamental (qualitative) analysis and Technical (quantitative) analysis. Fundamental methods mainly relies on the public information about the industry or companies such as market news and corporate information. The quantitative analysis uses mathematical models and historical price and is concerned with price action. In This research, focus is on technical analysis (quantitative analysis). This research considers stock analysis and prediction under three categories: statistical, machine learning (ML), and Deep Learning approaches.

One popular theory provide by [1][2] is the Efficient Market Hypothesis (EMH) that declares that at anywhere in the time, the stock market price is included all information about that stock. As per EMH, price changes cannot be predicted and forecasting a stock or financial market is only like a random action. There are also number of studies that do not accept EMH and applied different and prosperous approaches that used technical analysis, statistics, data mining, and artificial intelligence for stock price prediction [3]. Group of statistical approaches are the Auto-Regressive Moving Average (ARMA), the Auto-Regressive Integrated Moving

Average (ARIMA), the Generalized Autoregressive Conditional Heteroskedastic (GARCH) volatility [4] [5]. ARMA is a mixture of Auto-Regressive (AR) models and Moving Average (MA) models. One of the problems with the ARMA model is about not considering volatility clustering. ARIMA is an expanded ARMA model and it is able to turn a non-stationary series to a stationary series [5].

Machine learning algorithms can be considered into supervised and unsupervised learning. In supervised learning, a set of labelled input and output data are available, while, in unsupervised learning, all data is unlabelled. Supervised learning try to train an algorithm to automatically map the input data to the given output data and unsupervised learning try to train an algorithm to find a pattern or cluster in the given dataset [5]. Several machine-learning algorithms have been applied for stock price prediction matter same as Support vector Machines, Decision tree, Random Forest, logistic regression, and neural networks.

The stock price prediction is a kind of complex sequential decision-making problem and deep learning has achieved successes in solving complex sequential decision-making problems. Some deep learning models are applied for stock price prediction same as recurrent neural networks, Gated recurrent units (GRUs) and Long Short Term Memory (LSTM).

In this research, ARIMA model as the statistical model; Support vector Machines and Random Forest as Machine Learning models and LSTM as the deep learning model applied for stock price prediction. Moreover, Tehran stock market data is used for comparing, evaluating, testing and verification of the proposed models. The effectiveness of the models is tested by three performance measures: The Mean Absolute Error (MAE), The Mean Squared Error (MSE), and The Root Mean Squared Error (RMSE).

The contribution of this research is applying statistical, machine learning and deep learning algorithms for stock price prediction on the biggest companies of Tehran Stock Exchange, then evaluating, and comparing the performance of the above-mentioned algorithms. The rest of this research paper is organized as follows. Literature review is provided in Section 2, and research methodology is presented in Section 3. Results and discussion are in Section 4. Finally, conclusion is presented in Section 5.

LITERATURE REVIEW

One of the most methods that usually used for time series models is the Autoregressive Integrated Moving Average (ARIMA) [6]. ARIMA model is more popular than the other statistical methods for stock price prediction [7][8]. [8] Explore the extensive process of building ARIMA models for New York Stock Exchange (NYSE) and Nigeria Stock Exchange (NSE). In addition, their Results show that the ARIMA model is a good predictor in compare to other common techniques for stock price prediction.

Support vector machine (SVM) concept was a great success for time series prediction. SVM is used for pattern recognition and classification and also it has the ability for a better generalization performance. It supports linear and nonlinear regression that called as Support Vector Regression (SVR). SVM methodology has become one of the popular techniques for time series prediction problem [9][10]. In SVM, the solution only depends on a subset of the training data points, which are called support vectors and the solution found by SVM method is always unique and globally optimal. SVM algorithms are capable for operating with kernels, which enable them to operate in a high-dimensional feature space [9][10]. [11] Used Benchmark ensemble methods (Random Forest, AdaBoost and Kernel Factory) and single classifier models (Neural Networks, Logistic Regression, Support Vector Machines and K-Nearest Neighbor) for stock price prediction. They used data from European companies and their results show that Random Forest is the top algorithm followed by Support Vector Machines for stock price prediction.

Deep Learning is a type of artificial neural network that consists of multiple layers and is able to extract the good features of time series. Deep learning methods have also better performances to learn non-linear relationships in various financial features. [12] Applied different RNNs architectures for google stock price

movement. They used multi-layer RNN, long-short term memory (LSTM) and gated recurrent unit (GRU) for forecasting Google stock price movements. Their results show that LSTM outperformed other variants with a 72% accuracy on a five-day horizon. [13] Propose Long Short-Term Memory (LSTM) neural network to capture the complex features such as non-linearity, non-stationary and sequence correlation of financial time series for stock price prediction. They applied the model to predict the daily closing price of the Shanghai Composite Index and their results show that the LSTM performs a good prediction for financial time series.
Overall, regarding introduction and background literature, this research aims to compare best performance different statistical (ARIMA), machine learning (SVM and RF) and deep learning (LSTM) algorithms for short-term stock price prediction on a seven-day horizon.

RESEARCH METHODOLOGY

The stock price prediction has done by using Autoregressive Integrated Moving Average Model (ARIMA), Support Vector Machines, Random forest, and Long Short-Term Memory (LSTM) on a seven-day horizon. In this section, the algorithms are described and then data and metrics for evaluation, namely MSE, RMSE and MAE error, are explained.

3-1- Autoregressive Integrated Moving Average Model (ARIMA)

Concerning previous researches and background, ARIMA has better performance in regard to other statistical approaches. Before applying ARIMA to a time series, the time series needs to be stationary. Stationary means that the statistical features are all constant all the timestep. Once stationarity has been reached by a certain number of differences (parameter d is the number of non-seasonal differences needed for stationarity), it is time to determine the number of autoregressive terms (parameter p) and the number of lagged forecast errors in the Prediction equation (parameter q). The number of lagged forecast errors is also often referred to as the "moving average". A ARIMA model can classified as an "ARIMA(p,d,q)" model. The forecasting equation is constructed as follows:

$$\hat{Y}_t = \mu + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} \quad (1)$$

Where; moving average parameters (θ's), Autoregressive (φ's).

For this research, at first step, being stationary of each stock time series is checked and then trend and Seasonality from the time series are separated. Then the best p,q,d for ARIMA model is calculated and the best ARIMA model is created with provided optimal parameters (p,q,d). Forecasting via ARIMA model is used for 7 days and results are evaluated by MSE, RMSE and MAE metrics.

3-2- Support Vector Machines (SVM)

Support vector machine (SVM) algorithm is used for classification and regression problems. For linearly separable data the optimization problem is as follows: [10]

Minimize $J(W) = \frac{1}{2}\, W^T W = \frac{1}{2}\, \|W\|^2$ (2)

Subject to $y_i(W^T x_i + b\,) \geq 1\,;\ \forall i = 1,2,\dots,N$ (3)

For Non-linearly Separable Data the optimization problem is as follows: [10]

Minimize $J(W,\varepsilon) = \frac{1}{2}\, \|W\|^2 + \mathrm{C}\,(\,\sum_{i=1}^{N} \varepsilon_i\,)$ (4)

Subject to $y_i(W^T x_i + b\,) \geq 1 - \varepsilon_i\,;\ \forall i = 1,2,\dots,N\ \ ;\ \varepsilon_i \geq 0$ (5)

For Non-linear mapping of input space to the feature space: [10]

The data of the input space X$\subseteq \Re^n$ are mapped to feature space H by nonlinear mapping $\Phi: X \rightarrow H$. The linear decision function in H is as follows:

$$y(X) = sgn\,(\,\sum_{i=1}^{N} y_i\, \alpha_i\, \left(\varphi(X)^T\, Q(X_i)\right) + b_{opt})\ \ ) \quad (6)$$

By finding the function K $(x_i\,, x_j) = \varphi(X_i)^T\, Q\left(X_j\right)$, it can be used in SVM equation and there is no need to know $\varphi(X)$. Function K is called Kernel. Some Kernel functions that are used in this research are as follows:

Linear Kernel $K(x,y) = x^T y$ (7)

Polynomial Kernel $K(x,y) = (1 + x^T y)^d$ (8)

Radial Basis Function (RBF) Kernel $K(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$ (9)

3-3- Random Forest (RF)

Decision trees used for machine learning applications, but trees have potential to overfit the training sets. Random Forest can solve this problem by training different decision trees on different subspace of the feature space by increasing bias. In the other word, the trees exist in the forest cannot see the all-training dataset. The data is recursively split into partitions. The choice for the splitting criterion is based on some impurity measures such as Shannon Entropy or Gini impurity. [14]

3-4- Long Short-Term Memory (LSTM)

The main idea of Recurrent Neural Network is to apply the sequential observations learned from the earlier stages to forecast future trends. Long-Short Term Memory (LSTM) model can overcome the drawback of RNN in capturing long-term influences. LSTM introduces the memory cell that enables long-term dependency. The memory cells filter the information through the gate structure to maintain and update the state of memory cells. The gate structure includes input gate, forget gate and output gate. Input gate consists of the input. Cell State Runs through the entire network and has the ability to add or remove information with the help of gates. Forget gate layer decides the fraction of the information to be allowed. Output gate consists of the output generated by the LSTM. [15][16]

$f_t = \sigma\,(W_f.[\,h_{t-1}, x_t\,] + b_f)$ (10)

$i_t = \sigma\,(W_i.[\,h_{t-1}, x_t\,] + b_i)$ (11)

$C_t = tanh\,(W_c.[\,h_{t-1}, x_t\,] + b_c)$ (12)

$O_t = \sigma\,(W_o.[\,h_{t-1}, x_t\,] + b_o)$ (13)

$h_t = O_t * \tanh(C_t)$ (14)

x_t= Input Vector; h_t = output Vector; C_t= Cell State Vector; f_t=Forget gate Vector;
i_t= Input gate Vector; O_t= Output gate Vector.

3-5- Technical Indicators [17]

3-5-1- Simple moving average (SMA): Daily Financial data can be noisy. We can use the moving average to get more signal rather than noise from the data. The basic concept here is to provide a window of a set time and then use that to calculate aggregate statistic (Here, we have considered a simple 5-day moving average and a 10-day moving average as our features).

3-5-2- Exponentially Weighted Moving Average (EWMA): assigns more weight to more recent values in the time series and is a much better option than Simple Moving Averages.

$\mathrm{EWMA_t} = \alpha * \mathrm{r_t} + (1 - \alpha) * \mathrm{EWMA_{t-1}}$ (15)

3-5-3- Relative Strength Index (RSI): is an oscillator indicator, with a range between 0 to 100.
It measures the value of recent price changes to evaluate overbought or oversold conditions in the price of a stock.

3-5-4-Williams Percent Range: Williams Percent Range is a Momentum Indicator as well as an Oscillator, between -100 and 0. It measures overbought and oversold levels. The Williams Percent Range can used to find entry and exit points in the market.

3-6- Evaluation Metrics

The Mean Absolute Error (MAE): $\frac{1}{n}\sum_{i=1}^{N} |e_t|$ (16)

MAE measures the average value of the errors. It is the average of the absolute differences between the prediction and the actual observation where all individual differences have an equal weight.

The Mean Squared Error (MSE): $\frac{1}{n}\sum_{i=1}^{N} {e_t}^2$ (17)

MSE measures the average of the squares of the errors.

The Root Mean Squared Error (RMSE): $\sqrt{\frac{1}{n}\sum_{i=1}^{N} {e_t}^2}$ (18)

Root Mean Squared Error (RMSE). It is just the square root of the mean square error.

3-7- Dataset

In this paper, the above-mentioned algorithms are applied for short time stock prediction (7-day horizon). Dataset of five biggest companies of Tehran stock market is used for evaluating, testing and verifying the implementations: Esfahan's Mobarakeh Steel Company (MSC), National Iranian Copper Industries Company (NICICO), Isfahan Oil Refinery (IORC), Telecommunication Company of Iran (TCI), Ghadir Investment Company (GHADIR) . Dataset includes: Open, Close,High, Low, Adj Close, Volume and Technical indicators.

RESULT AND DISCUSSION

The implementation has done by using Python. All data for five companies downloaded from tsetmc.com. All data consider from the first of 2018 up to January 2021. The dataset is divided to train and test. Data from 2018 up to 2020 is used for training and data from 2020 up to 2021 is used as a test data.

The first thing has done for each dataset is pre-processing and normalizing data and all models are evaluated for accuracy metric with Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE).

First, in the following figures (1, 2, 3, 4, and 5) the close price history for all five Tehran stock companies are showed.

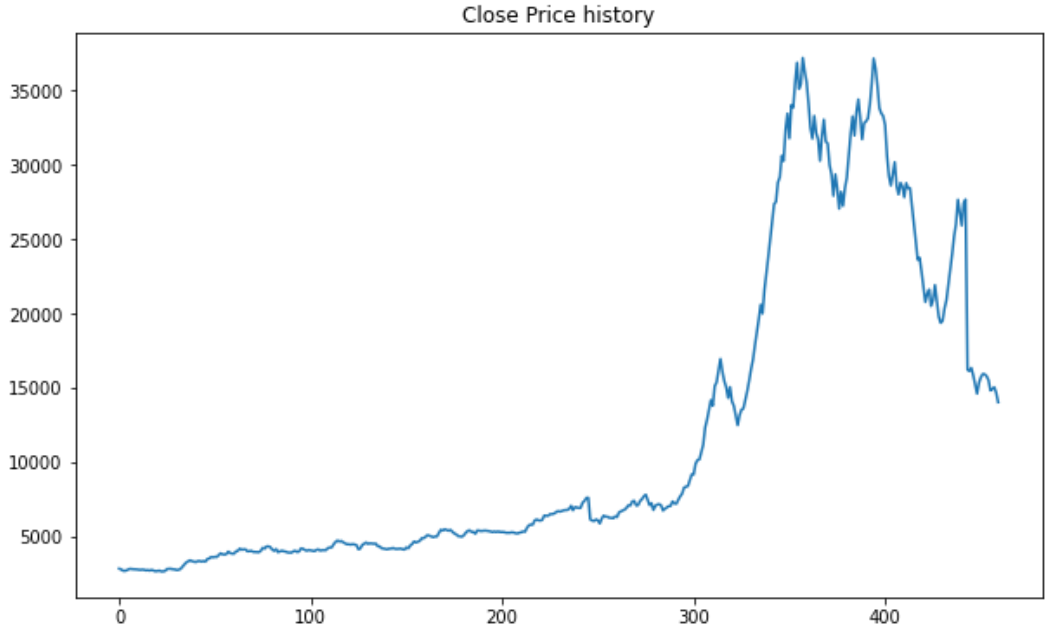

Figure1: Stock price history as per close price National Iranian Copper Industries Company (NICICO)

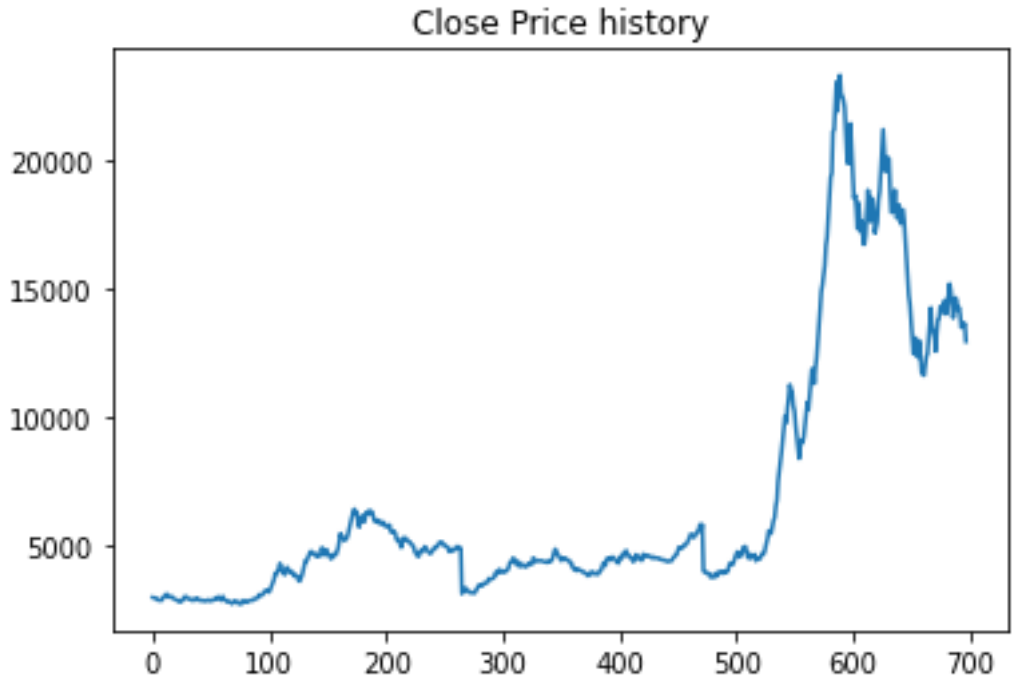

Figure2: Stock price history as per close price Esfahan's Mobarakeh Steel Company (MSC)

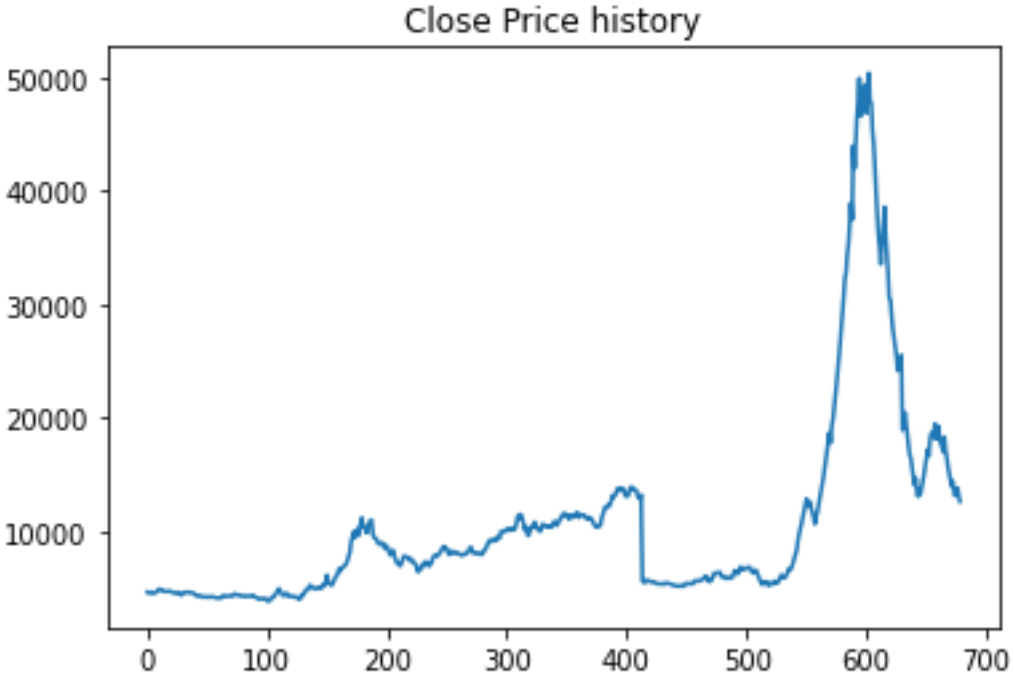

Figure3: Stock price history as per close price Isfahan Oil Refinery (IORC)

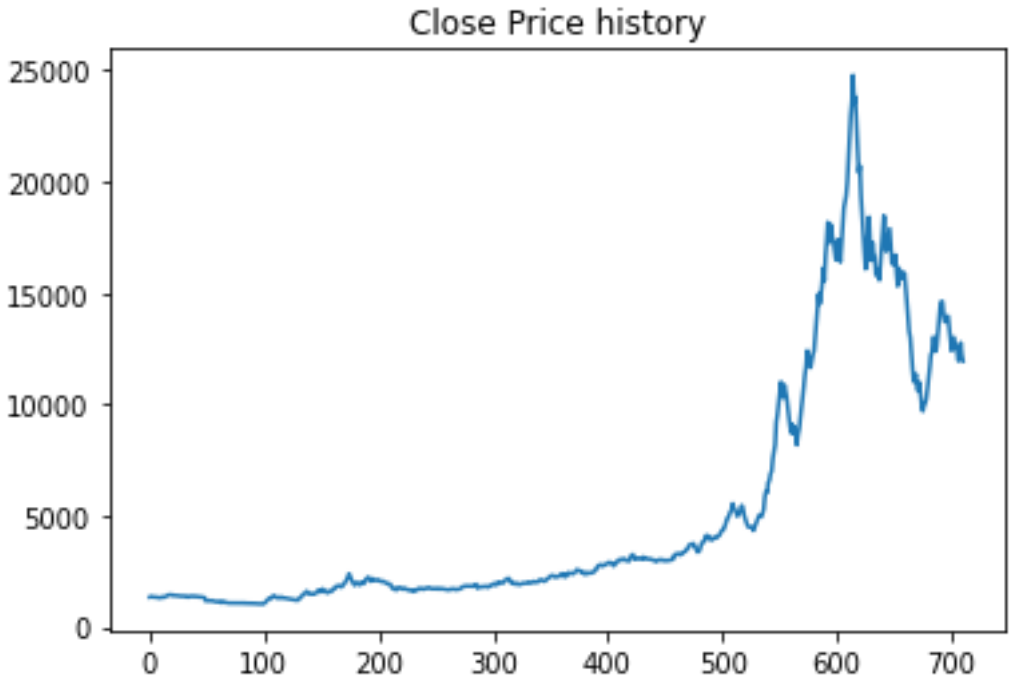

Figure4: Stock price history as per close price Ghadir Investment Company (GHADIR)

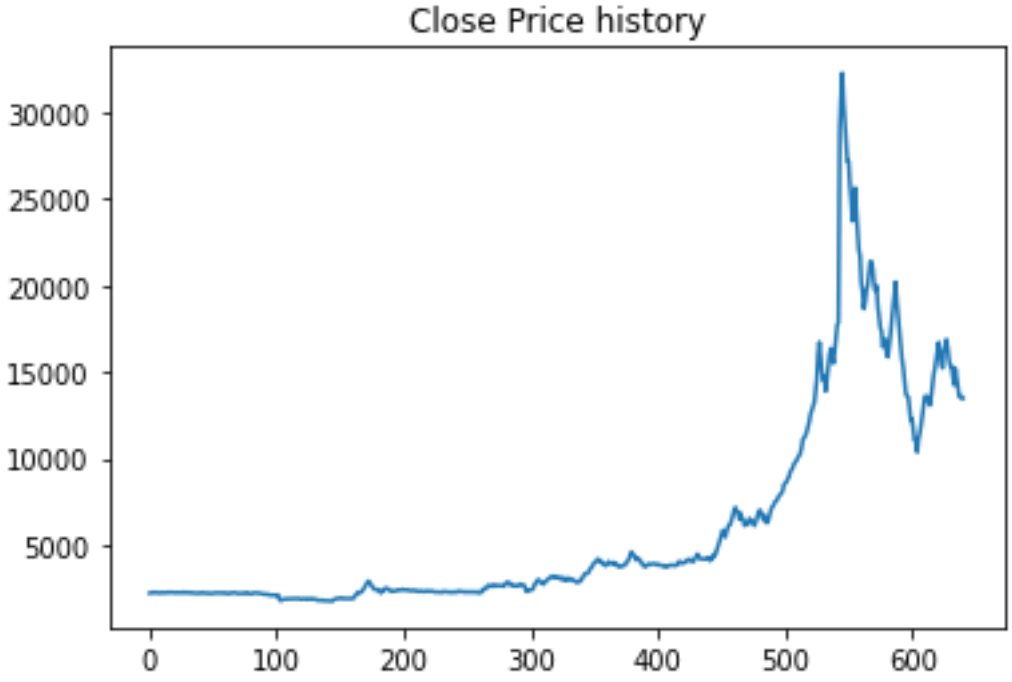

Figure5: Stock price history as per close price Telecommunication Company of Iran (TCI)

The results in tables (1,2,3,4,5) show that the LSTM has the best performance for short term prediction on all stocks. Moreover, SVM with linear kernel has the best performance between SVM and Random forest. In addition, ARIMA model concerning SVM and Random Forest has a better performance for short-term stock price prediction.

Table (1): Results for Test datasets
National Iranian Copper Industries Company (NICICO)

National Iranian Copper Industries Company (NICICO)			
	MSE	MAE	RMSE
ARIMA	0.059	0.21	0.24
SVM Kernel= linear	0.35	0.35	0.59
SVM Kernel= RBF	0.56	0.56	0.75
RF	0.38	0.38	0.61
LSTM	0.0032	0.024	0.036

Table (2): Results for Test datasets
Esfahan's Mobarakeh Steel Company (MSC)

Esfahan's Mobarakeh Steel Company (MSC)			
	MSE	MAE	RMSE
ARIMA	0.045	0.19	0.21
SVM Kernel= linear	0.42	0.42	0.65
SVM Kernel= RBF	0.58	0.58	0.76
RF	0.45	0.45	0.67
LSTM	0.0007	0.023	0.035

Table (3): Results for Test datasets
Isfahan Oil Refinery (IORC)

Isfahan Oil Refinery (IORC)			
	MSE	MAE	RMSE
ARIMA	0.085	0.25	0.29
SVM Kernel= linear	0.34	0.34	0.59
SVM Kernel= RBF	0.50	0.50	0.71
RF	0.30	0.30	0.55
LSTM	0.0027	0.029	0.045

Table (4): Results for Test datasets
Ghadir Investment Company (GHADIR)

Ghadir Investment Company (GHADIR)			
	MSE	MAE	RMSE
ARIMA	0.032	0.15	0.18
SVM Kernel= linear	0.44	0.44	0.66
SVM Kernel= RBF	0.52	0.52	0.72
RF	0.41	0.41	0.64
LSTM	0.002	0.021	0.028

Table (5): Results for Test datasets
Telecommunication Company of Iran (TCI)

Telecommunication Company of Iran (TCI)			
	MSE	MAE	RMSE
ARIMA	0.020	0.12	0.14
SVM Kernel= linear	0.47	0.47	0.68
SVM Kernel= RBF	0.51	0.51	0.71
RF	0.52	0.52	0.72
LSTM	0.004	0.024	0.040

CONCLUSION

As described above in this paper, different algorithms is applied for stock price prediction concerning to short-term (7 day-horizon) forecasting. The five biggest stock companies of tehran stock market used for evaluating the algorithms and according to the results, LSTM approach has the best performance in regard to other algorithms. SVM and Random forest is not good in compare to LSTM but SVM with linear kernel has a better performance in regard to other kernels and random forest algorithm. In addition, ARIMA model shows the better performance in regard to SVM and random forest for stock price prediction.

REFERENCES

1) Fama, E.F., 1970. Efficient market hypothesis: A review of theory and empirical work. Journal of Finance, 25(2), pp.28-30.
2) Fama, E.F., 1995. Random walks in stock market prices. Financial analysts journal, 51(1), pp.75-80.
3) Arévalo, R., García, J., Guijarro, F. and Peris, A., 2017. A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting. Expert Systems with Applications, 81, pp.177-192.
4) Fu, T.C., 2011. A review on time series data mining. Engineering Applications of Artificial Intelligence, 24(1), pp.164-181.
5) Shah, D., Isah, H. and Zulkernine, F., 2019. Stock market analysis: A review and taxonomy of prediction techniques. International Journal of Financial Studies, 7(2), p.26.
6) Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, *50*, pp.159-175.
7) Meyler, A., Kenny, G. and Quinn, T., 1998. Forecasting Irish inflation using ARIMA models.
8) Ariyo, A.A., Adewumi, A.O. and Ayo, C.K., 2014, March. Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation (pp. 106-112). IEEE.
9) Cao, L.J. and Tay, F.E.H., 2003. Support vector machine with adaptive parameters in financial time series forecasting. IEEE Transactions on neural networks, 14(6), pp.1506-1518.
10) Adhikari, R. and Agrawal, R.K., 2013. An introductory study on time series modeling and forecasting. arXiv preprint arXiv:1302.6613.
11) Ballings, M., Van den Poel, D., Hespeels, N. and Gryp, R., 2015. Evaluating multiple classifiers for stock price direction prediction. Expert systems with Applications, 42(20), pp.7046-7056.
12) Di Persio, L. and Honchar, O., 2017. Recurrent neural networks approach to the financial forecast of Google assets. International journal of Mathematics and Computers in simulation, 11, pp.7-13.

13) Yan, H. and Ouyang, H., 2018. Financial time series prediction based on deep learning. Wireless Personal Communications, 102(2), pp.683-700.
14) Khaidem, L., Saha, S. and Dey, S.R., 2016. Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.
15) Zou, Z. and Qu, Z., 2020. Using LSTM in Stock prediction and Quantitative Trading.
16) Selvin, S., Vinayakumar, R., Gopalakrishnan, E.A., Menon, V.K. and Soman, K.P., 2017, September. Stock price prediction using LSTM, RNN and CNN-sliding window model. In 2017 international conference on advances in computing, communications and informatics (icacci) (pp. 1643-1647). IEEE.
17) https://www.investopedia.com/financial-term-dictionary-4769738.