INDIAN STOCK MARKET PRICE PREDICTION USING MACHINE LEARNING AND DEEP LEARNING

Name of Authors and Affiliation

Anandharaja K and Vijaya Kalavakonda
Department of Computing Technologies
School of Computing
Faculty of Engineering and Technlogy
SRM Institute of Science and Technology, Kattankulathur

Abstract: The nature of stock market movement has always been ambiguous for investors because of various influential factors. This study aims to significantly reduce the risk of trend prediction with machine learning and deep learning algorithms. Four stock market groups, namely diversified financials, petroleum, non-metallic minerals and basic metals from Tehran stock exchange, are chosen for experimental evaluations. This study compares nine machine learning models (Decision Tree, Random Forest, Adaptive Boosting (Adaboost), eXtreme Gradient Boosting (XGBoost), Support Vector Classifier (SVC), Naïve Bayes, K-Nearest Neighbors (KNN), Logistic Regression and Artificial Neural Network (ANN)) and two powerful deep learning methods (Recurrent Neural Network (RNN) and Long short-term memory (LSTM). Ten technical indicators from ten years of historical data are our input values, and two ways are supposed for employing them. Firstly, calculating the indicators by stock trading values as continuous data, and secondly converting indicators to binary data before using. Each prediction model is evaluated by three metrics based on the input ways. The evaluation results indicate that for the continuous data, RNN and LSTM outperform other prediction models with a considerable difference. Also, results show that in the binary data evaluation, those deep learning methods are the best; however, the difference becomes less because of the noticeable improvement of models' performance in the second way. As extension we are added LSTM, Linear Regression , Lasso Regression, Ridge Regressor, Xgboost, Voting Regression, Decision Tree, Random forest, SVM, Stacking Regression, adaboost, SGDregressor, Adaboost, Catboost, LightBoost, Voting Regression-[catboost , lightboost], Stacking Regressor - [catboost , lightboost] are used and algorithms all are giving high prediction accuracies compare to existing ones.

Index Terms – *Machine learning, Deep learning, Stock prediction.*

1. INTRODUCTION

The task of stock prediction has always been a challenging problem for statistics experts and finance. The main reason behind this prediction is buying stocks that are likely to increase in price and then selling stocks that are probably to fall. Generally, there are two ways for stock market prediction. Fundamental analysis is one of them and relies on a company's technique and fundamental information like mar ket position, expenses and annual growth rates. The second one is the technical analysis method, which concentrates on previous stock prices and values. This analysis uses historical charts and patterns to predict future prices [1], [2]. Stock markets were normally predicted by financial experts in the past time. However, data scientists have started solving prediction problems with the

progress of learning techniques. Also, computer scientists have begun using machine learning methods to improve the performance of prediction models and enhance the accuracy of predictions.

Employing deep learning was the next phase in improving prediction models with better performance [3], [4]. Stock market prediction is full of challenges, and data scientists usually confront some problems when they try to develop a predictive model. Complexity and nonlinearity are two main challenges caused by the instability of stock market and the correlation between investment psychology and market behavior [5]. It is clear that there are always unpredictable factors such as the public image of companies or political situation of countries, which affect stock markets trend. Therefore, if the data gained from stock values are efficiently preprocessed and suitable algorithms are employed, the trend of stock values and index can be predicted.

In stock market prediction systems, machine learning and deep learning approaches can help investors and traders through their decisions. These methods intend to automatically recognize and learn patterns among big amounts of information. The algorithms can be effectively self-learning, and can tackle the predicting task of price fluctuations in order to improve trading strategies [6].

2. RELATED WORK

Stock price modeling and prediction have been challenging objectives for researchers and speculators because of noisy and non-stationary characteristics of samples. With the growth in deep learning, the task of feature learning can be performed more effectively by purposely designed network. In this paper [1], propose a novel end-to-end model named multi-filters neural network (MFNN) specifically for feature extraction on financial time series samples and price movement prediction task. Both convolutional and recurrent neurons are integrated to build the multi-filters structure, so that the information from different feature spaces and market views can be obtained. We apply our MFNN for extreme market prediction and signal-based trading simulation tasks on Chinese stock market index CSI 300. Experimental results show that our network outperforms traditional machine learning models, statistical models, and single-structure(convolutional, recurrent, and LSTM) networks in terms of the accuracy, profitability, and stability.

In this paper [2] propose and implement a fusion model by combining the Hidden Markov Model (HMM), Artificial Neural Networks (ANN) and Genetic Algorithms (GA) to forecast financial market behaviour. The developed tool can be used for in depth analysis of the stock market. Using ANN, the daily stock prices are transformed to independent sets of values that become input to HMM. We draw on GA to optimize the initial parameters of HMM. The trained HMM is used to identify and locate similar patterns in the historical data. The price differences between the matched days and the respective next day are calculated. Finally, a weighted average of the price differences of similar patterns is obtained to prepare a forecast for the required next day. Forecasts are obtained for a number of securities in the IT sector and are compared with a conventional forecast method.

Investor sentiment plays an important role on the stock market. User-generated textual content on the Internet provides a precious source to reflect investor psychology and predicts stock prices as a complement to stock market data. This paper [3] integrates sentiment analysis into a machine learning method based on support vector machine. Furthermore, we take the day-of-week effect into

consideration and construct more reliable and realistic sentiment indexes. Empirical results illustrate that the accuracy of forecasting the movement direction of the SSE 50 Index can be as high as 89.93% with a rise of 18.6% after introducing sentiment variables. And, meanwhile, our model helps investors make wiser decisions. These findings also imply that sentiment probably contains precious information about the asset fundamental values and can be regarded as one of the leading indicators of the stock market.

Ability to predict direction of stock/index price accurately is crucial for market dealers or investors to maximize their profits. Data mining techniques have been successfully shown to generate high forecasting accuracy of stock price movement. Nowadays, in stead of a single method, traders need to use various forecasting techniques to gain multiple signals and more information about the future of the markets. In this paper [4], ten different techniques of data mining are discussed and applied to predict price movement of Hang Seng index of Hong Kong stock market. The approaches include Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), K-nearest neighbor classification, Naive Bayes based on kernel estimation, Logit model, Tree based classification, neural network, Bayesian classification with Gaussian process, Support vector machine (SVM) and Least squares support vector machine (LS-SVM). Experimental results show that the SVM and LS-SVM generate superior predictive performances among the other models. Specifically, SVM is better than LS-SVM for in-sample prediction but LS-SVM is, in turn, better than the SVM for the out-of-sample forecasts in term of hit rate and error rate criteria.

The problem of predicting stock returns has been an important issue for many years. Advancement in computer technology has allowed many recent studies to utilize machine learning techniques such as neural networks and decision trees to predict stock returns. In the area of machine learning, classifier ensembles (i.e. combining multiple classifiers) have proven to be a method superior to single classifiers. In order to build a better model for predicting stock returns effectively and efficiently, this study aims at investigating the prediction performance that utilizes the classifier ensembles method to analyze stock returns. In particular, the hybrid methods of majority voting and bagging are considered. Moreover, performance using two types of classifier ensembles is compared with those using single baseline classifiers (i.e. neural networks, decision trees, and logistic regression) [5]. These two types of ensembles are 'homogeneous' classifier ensembles (e.g. an ensemble of neural networks) and 'heterogeneous' classifier ensembles (e.g. an ensemble of neural networks, decision trees and logistic regression). Average prediction accuracy, Type I and II errors, and return on investment of these models are also examined. Our results indicate that multiple classifiers outperform single classifiers in terms of prediction accuracy and returns on investment. In addition, heterogeneous classifier ensembles offer slightly better performance than the homogeneous ones. However, there is no significant difference between majority voting and bagging in prediction accuracy, but the former has better stock returns prediction accuracy than the latter. Finally, the homogeneous multiple classifiers using neural networks by majority voting perform best when predicting stock returns [5].

3. MATERIALS AND METHODS

In a local and global event sentiment based efficient stock exchange forecasting using deep learning. they consider four countries- US, Hong Kong, Turkey, and Pakistan from developed, emerging and underdeveloped economies' list. We have explored

the effect of different major events occurred during 2012–2016 on stock markets. We use the Twitter dataset to calculate the sentiment analysis for each of these events. The dataset consists of 11.42 million tweets that were used to determine the event sentiment. We have used linear regression, support vector regression and deep learning for stock exchange forecasting.

Drawbacks:

1. This process is time-consuming which makes deep learning model training slow
2. The existing system fails when there are rare outcomes or predictors.
3. The previous results indicate that the stock price is unpredictable when the traditional classifier is used.
4. The existence system reported highly predictive values, by selecting an appropriate time period for their experiment to obtain highly predictive scores.

We propose nine machine learning and two powerful deep learning methods (Recurrent Neural Network (RNN) and Long short-term memory (LSTM). Ten technical indicators from ten years of historical data are our input values, and two ways are supposed for employing them. Firstly, calculating the indicators by stock trading values as continues data, and secondly converting indicators to binary data before using. Each prediction model is evaluated by three metrics based on the input ways.

Benefits:

1. The binary data evaluation, those deep learning methods are the best. Which can make this process fast and efficient
2. Stock markets help companies to raise capital.
3. It helps generate personal wealth.
4. Stock markets serve as an indicator of the state of the economy.
5. It is a widely used source for people to invest money in companies with high growth potential.

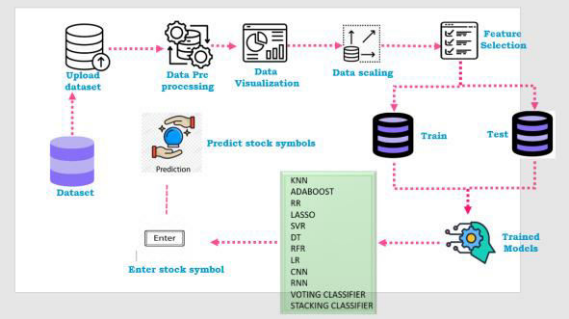


Fig.1 Proposed Architecture

The project is to predict the stock prices in order to make more informed and accurate investment decisions. A stock market prediction is described as an action of attempting to classify the future value of the company stock or other financial investment traded on the stock exchange. The forthcoming price of a stock of the successful estimation is called the Yield significant profit. This helps you to invest wisely for making good profits.

- Have a very significant effect on an organization performance. This paper proposed a technique to reveal the performance of a company.
- The technique deployed in the paper is used to find the relationships between the frequencies of email exchange of the key employees and the performance of the company reflected in stock values.

• In order to detect association and non-association relationships, this paper proposed to use a machine learning and deep learning algorithms on a publicly available datasets.

i) Dataset Collection:

UCI Machine Learning Repository:

The hypothyroid dataset, sourced from the UCI Machine Learning Repository, is used in this study to evaluate the effectiveness of various missing data imputation methods. The dataset contains 3,772 instances and 30 features, with categorical attributes representing clinical and diagnostic data related to thyroid disease. These features include patient age, sex, and medical history, alongside diagnostic measures such as TSH (Thyroid Stimulating Hormone) levels, thyroxine levels, and other thyroid-related biomarkers. Missing values are present in several categorical features, representing a common challenge in healthcare data analysis. The dataset is imbalanced, with a higher prevalence of negative hypothyroidism cases. This imbalance poses additional challenges in imputation and model evaluation. The hypothyroid dataset is commonly used in machine learning studies for classification and imputation tasks, making it an ideal choice for testing the imputation techniques proposed in this study, especially in real-world healthcare scenarios [1][9].

	age	sex	on thyroxine	query on thyroxine	on antithyroid medication	sick	pregnant	thyroid surgery	I131 treatment	query hypothyroid	...	TT4 measured	TT4
0	41.0	F	f	f	f	f	f	f	f	f	...	t	125.0
1	23.0	F	f	f	f	f	f	f	f	f	...	t	102.0
2	46.0	M	f	f	f	f	f	f	f	f	...	t	109.0
3	70.0	F	t	f	f	f	f	f	f	f	...	t	175.0
4	70.0	F	f	f	f	f	f	f	f	f	...	t	61.0

Fig 2 Dataset Collections

ii) Pre-Processing:

Preprocessing involves visualizing data, handling missing values, label encoding categorical variables, and extracting relevant features to prepare the dataset for machine learning analysis and model training.

a) Visualization: Visualization is a crucial step in the preprocessing phase, involving the use of graphs, charts, and plots to understand the data distribution, detect patterns, identify outliers, and observe missing values. This process aids in gaining valuable insights into the dataset before applying machine learning techniques [9][12]. Visualization helps in making informed decisions about how to handle missing data and other irregularities, ensuring a more robust analysis.

b) Data Processing: Data processing involves preparing raw data for analysis by addressing missing values, cleaning noise, and transforming data formats. Effective data processing ensures the dataset is suitable for model training, enhancing the overall performance of the machine learning algorithms [8][17]. Handling missing values, especially in healthcare datasets, is a key challenge in ensuring the reliability of the analysis [5][6].

c) Label Scaling: Label scaling refers to the process of increasing the capacity or performance of a system to handle more data or traffic. There are two primary approaches to scaling a system: horizontal scaling and vertical scaling.

d) Feature Extraction: Feature extraction involves selecting and transforming raw data into meaningful features that highlight important patterns in the dataset. This step helps reduce dimensionality, improve model performance, and create relevant, informative input features for training [13][15]. By identifying key features, feature extraction improves the efficiency and accuracy of machine learning models used in data imputation and analysis.

iii) Training & Testing:

In the proposed system, the dataset is split into training and testing sets to evaluate the performance of imputation models. The training set is used to train the machine learning algorithms, such as Random Forest, SVM Bagging-MIX, KNN Bagging-MIX, and XGBoost, on the known data, including features with missing values. The models learn to predict and impute missing values based on the patterns observed. The testing set is then used to evaluate the models' accuracy in predicting missing values. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to compare the performance of each model.

iv) Algorithms:

SVM Algorithm:

Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyper plane which separates the data into classes.

Random Forest Algorithm:

it's an ensemble algorithm which means internally it will use multiple classifier algorithms to build accurate classifier model. Internally this algorithm will use decision tree algorithm to generate it train model for classification.

Decision Tree Algorithm:

This algorithm will build training model by arranging all similar records in the same branch of tree and continue till all records arrange in entire tree. The complete tree will be referred as classification train model.

Gradient Boosting Algorithm:

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets, and have recently been used to win many Kaggle data science competitions.

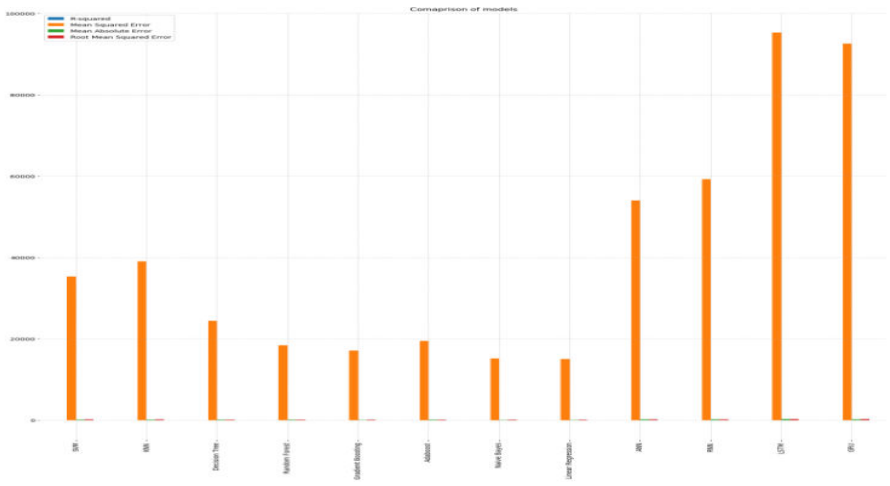
Long short-term memory (LSTM):

LSTM is an artificial recurrent neural network (RNN) architecture[1] used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition,[2] speech recognition[3][4] and anomaly detection in network traffic or IDSs (intrusion detection systems).

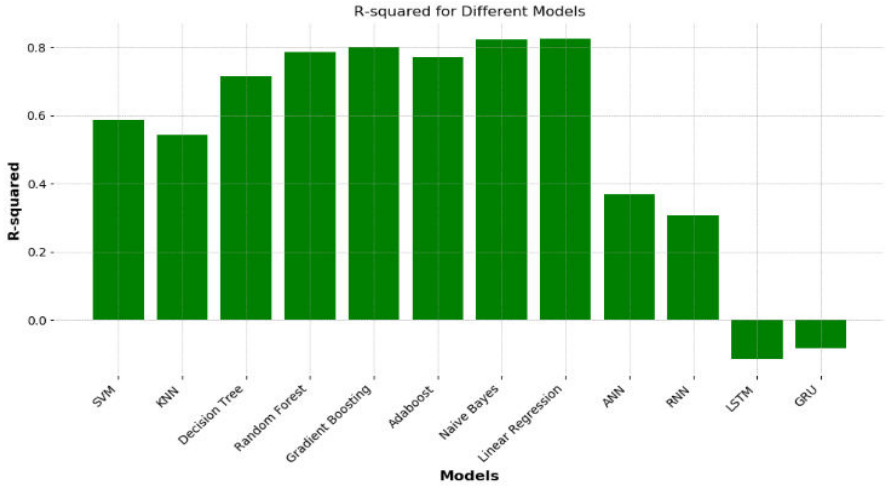
Deep Learning ANN Algorithm:

An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes - or learns, in a sense - based on that input and output.

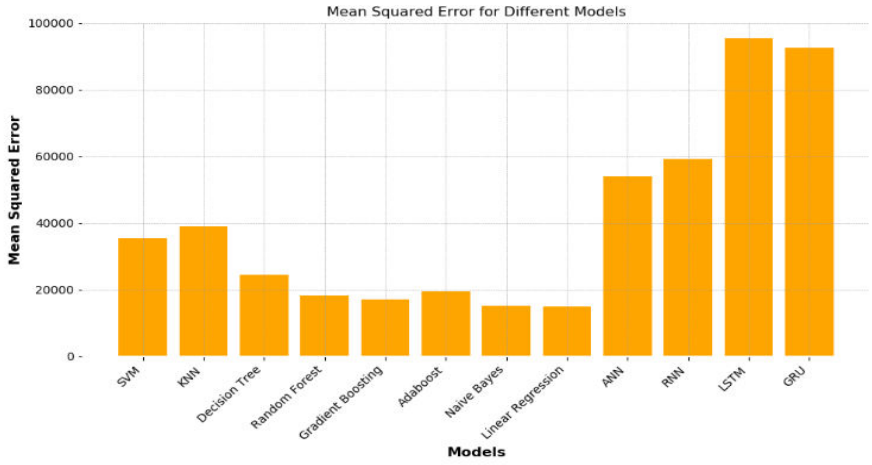
4. RESULTS & DISCUSSION



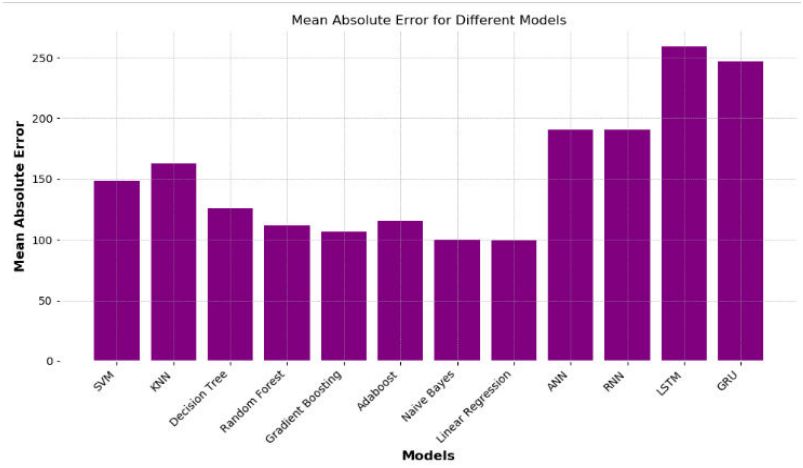
Graph.1 Performance Evaluation for algorithms



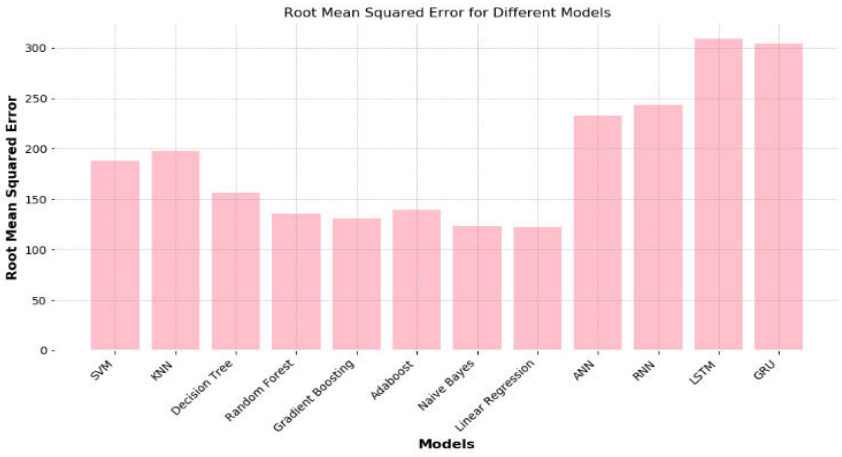
Graph.2 R-Squared for different models



Graph.3 Mean-Squared error for different models



Graph.4 Mean absolute error for different models



Graph.4 Root mean squared error for different models

Results

	R-squared	Mean Squared Error	Mean Absolute Error	Root Mean Squared Error
SVM	0.705873	63669.437448	171.712683	252.328035
KNN	0.768408	50132.480360	156.675078	223.902837
Decision Tree	0.884728	24952.843262	121.044758	157.964690
Random Forest	0.827340	37375.635649	159.814064	193.327793
Gradient Boosting	0.894132	22917.147613	114.410066	151.384106
Adaboost	0.872840	27526.280714	125.080202	165.910460
Naive Bayes	0.923407	16580.060556	97.030634	128.763584
Linear Regression	0.924192	16410.091854	96.602012	128.101881
ANN	0.652502	75222.512429	204.624370	274.267228
Voting Classifier	0.910542	19364.801332	105.823006	139.157470
Stacking Classifier	0.925030	16228.744210	95.857658	127.392088
RNN	0.643827	77100.518801	208.352852	277.669802
LSTM	-0.389089	300694.950061	455.957187	548.356590
GRU	0.642992	77281.106816	221.448633	277.994796
				244.731847
				244.731847

In graphs, x-axis represents algorithms and y-axis represents performance metrics. In all graph we are find outing the measures about all algorithms. The graphs above visually illustrate these findings.

Fig.3 Home Page

In the above figure 1, this is a user interface dashboard for Darkweb, it is a welcome message for navigating page.

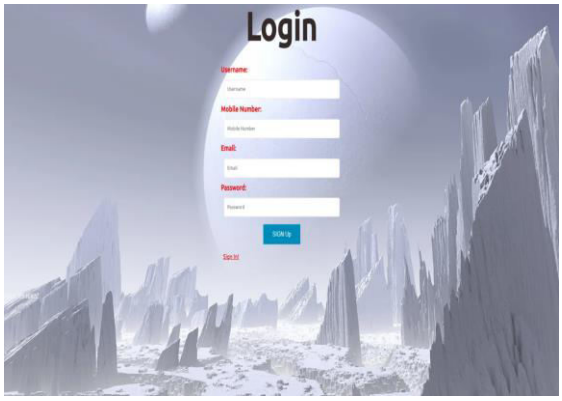


Fig.4 User signup Page

In the above figure 2, this is a user registration page, using this user can get registartion.

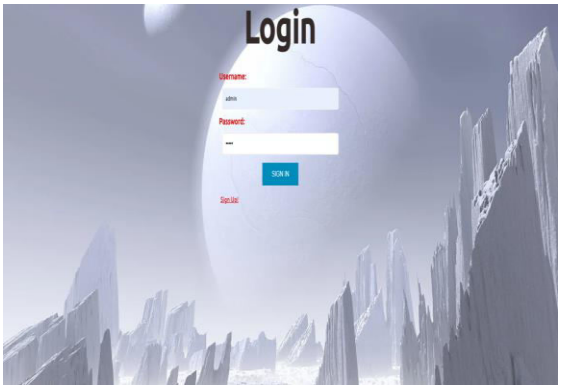


Fig.5 User login Page

In the above figure 3, this is a user login page, using this user can get login into the application.

Fig.6 User input Page

In the above figure 4, this is a user input page, using this user can enter stock symbol for prediction.

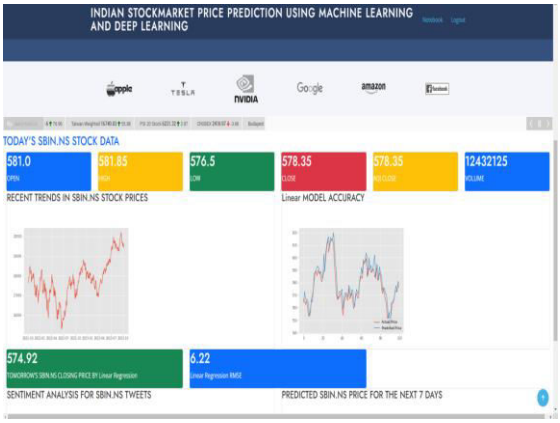


Fig.7 Prediction result

In the above figure 5, this is a result screen, in this user will get output for loaded input data.

5. CONCLUSION

In conclusion, the purpose of this study was the prediction task of stock market movement by machine learning and deep learning algorithms. Four stock market groups, namely diversified financials, petroleum, non-metallic minerals and basic metals, from Tehran stock exchange were chosen, and the dataset was based on ten years of historical records with ten technical features. Also, nine machine learning models (Decision Tree, Random Forest, Adaboost, XGBoost, SVC, Naïve Bayes, KNN, Logistic Regression and ANN) and two deep learning methods (RNN and LSTM) were employed as predictors. We supposed two approaches for input values to models, continuous data and binary data, and we employed three classification metrics for evaluations. Our experimental works showed that there was a significant improvement in the performance of models when they use binary data instead of continuous one. Indeed, deep learning algorithms (RNN and LSTM) were our superior models in both approaches. As extension we are added LSTM, Linear Regression , Lasso Regression, Ridge Regressor, Xgboost, Voting Regression, Decision Tree, Random forest, SVM, Stacking Regression, adaboost, SGDregressor, Adaboost, Catboost, LightBoost, Voting Regression-[catboost , lightboost], Stacking Regressor - [catboost , lightboost] are used and algorithms all are giving high prediction accuracies compare to existing ones.

FUTURE SCOPE:

In this work, we proposed machine learning and deep learning based classification model architectures that predicts the stock trends based on the linguistic features extracted from stock symbols. The overall results suggest that the proposed models successfully extract the relevant trends from the report that help with the stock dynamic predictions. The analysis highlights the strengths and the limitations of the models and how important to combine the linguistic features and non-linguistic feature to the performance of the model predictions.

REFERENCES

[1] M. R. Hassan, B. Nath, and M. Kirley, ‘‘A fusion model of HMM, ANN and GA for stock market forecasting,’’ Expert Syst. Appl., vol. 33, no. 1, pp. 171–180, Jul. 2007.

[2] W. Huang, Y. Nakamori, and S.-Y. Wang, ‘‘Forecasting stock market movement direction with support vector machine,’’ Comput. Oper. Res., vol. 32, no. 10, pp. 2513–2522, Oct. 2005.

[3] J. Sun and H. Li, ‘‘Financial distress prediction using support vector machines: Ensemble vs. Individual,’’ Appl. Soft Comput., vol. 12, no. 8, pp. 2254–2265, Aug. 2012.

[4] P. Ou and H. Wang, ‘‘Prediction of stock market index movement by ten data mining techniques,’’ Modern Appl. Sci., vol. 3, no. 12, pp. 28–42, Nov. 2009.

[5] F. Liu and J. Wang, ‘‘Fluctuation prediction of stock market index by legendre neural network with random time strength function,’’ Neurocomputing, vol. 83, pp. 12–21, Apr. 2012

[6] J. Sun, K. Xiao, C. Liu, W. Zhou, and H. Xiong, “Exploiting intra-day patterns for market shock prediction: a machine learning approach,” Expert Systems With Applications, vol. 127, pp. 272–281, 2019.

[7] Z. Lin, “Modelling and forecasting the stock market volatility of sse composite index using garch

models," Future Generation Computer Systems, vol. 79, pp. 960–972, 2018.

[8] Y. Shynkevich, T. M. McGinnity, S. A. Coleman, A. Belatreche, and Y. Li, "Forecasting price movements using technical indicators: investigating the impact of varying input window length," Neurocomputing, vol. 264, pp. 71–88, 2017.

[9] J. Patel, S. Shah, P. /akkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," Expert Systems with Applications, vol. 42, no. 1, pp. 259–268, 2015.

[10] O. B. Sezer and A. M. Ozbayoglu, "Algorithmic financial trading with deep convolutional neural networks: time series to image conversion approach," Applied Soft Computing, vol. 70, pp. 525–538, 2018.

[11] E. Hoseinzade and S. Haratizadeh, "Cnnpred: cnn-based stock market prediction using a diverse set of variables," Expert Systems with Applications, vol. 129, pp. 273–285, 2019.

[12] Y. Chen, W. Lin, and J. Z. Wang, "A dual-attention-based stock price trend prediction model with dual features," IEEE Access, vol. 7, pp. 148047–148058, 2019.

[13] L. Chen, Z. Qiao, M. Wang, C. Wang, R. Du, and H. E. Stanley, "Which artificial intelligence algorithm better predicts the Chinese stock market?" IEEE Access, vol. 6, pp. 48625–48633, 2018.

[14] P. Yu and X. Yan, "Stock price prediction based on deep neural networks," Neural Computing and Applications, vol. 132, pp. 1–20, 2019.

[15] H. M, G. E. A., V. K. Menon, and S.K. P., "Nse stock market prediction using deep-learning models," Procedia Computer Science, vol. 132, pp. 1351–1362, 2018.