Predicting Stock Closing Price using Supervised Learning Techniques

Kunal Sharma
College of Computing &
Informatics
Drexel University
Philadelphia, PA 19104, USA
kos26@drexel.edu

Abstract--Stock prediction aims to predict the future trends of a stock in order to help investors make good investment decisions. Traditional solutions for stock prediction are based on time-series models. The main objective of this paper is to compare different regression models for predicting the closing price of the stock market. During the process of considering various techniques and variables that must be taken into account, we found out that techniques like Long short-term memory, Extra Tree Regressor, Random Forrest and Gaussian Boosting Regressor doesn't necessarily perform better than simple models such as Linear Regression in every case. In this paper comparision is made between most of the popular regression models and root mean squared error were presented of each model to find one with higher accuracy. Real data set of last five years have been gained from yahoo finance for real analysis and project. The data will then be cleaned if necessary so that there are no null values in the data for better predictions. The first 80% of data is been used for training statistical models and the last 20% for testing and making predictions. Regression models such as Linear Regression, Support Vector Regressor, Random Forrest Regressor, Extra Tree Regressor, Gradient Boosting Regressor, Bayesian Ridge and Long short-term memory are trained in the project and predictions were made on the closing price of the stock of the testing set. Root mean square error was calculated for each model and later compared to other models. The project concludes simple regression model Linear regression from sklearn have outperformed complex deeplearning model Long short-term memory from tensorflow-keras in this case.

I. INTRODUCTION

A stock is a share or ownership of a part of a publicly listed company. These shares are issued by the company to exchange for traders to trade. These stocks are sold by the owners of the company to raise money/funding for further development of the company. When the company is first listed on an exchange, it is called an Initial Public Offering where the initial selling price of that stock is set by the owners. The price of the stock after the initial public offering is decided by the equilibrium of buy and sell orders, which can also be thought of as demand and supply equilibrium [4].

It is easy to understand the demand and supply is the root cause of price change, however, demand and supply are based on several variables and factors like inflation, positive or negative news, market sentiment, socio-economic factors, trends, and many more.

Since the beginning of the stock market, the goal of the speculators/investors has been to predict the price of the stock as to buy low and sell high, thus earning a profit. For the purpose of this paper, stock prices have been assumed to be a regression problem and different models have been compared to forecast stock's closing price [11].

Chongyang Zhang, Zhi Ji, Jixiang Zhang, Yanqing Wang, Xingzhi Zhao, and Yiping Yang (2018), states traditional models used in stock prediction involved statistical methods such as time series model and multivariate analysis, which are often from a mathematical point of view. The financial value was considered as a function of time series and solved as a regression problem. The applications of machine learning approach in the stock market solve the problem in a new way. When viewing the problem as a classification problem, the result performance could be better [1]. A variety of algorithms such as Linear Regression, Support Vector Regressor, Random Forrest Regressor, Extra Tree Regressor, Gradient Boosting Regressor, Bayesian Ridge and Long short-term memory is explored in this paper. The basic approaches which is been used to predict stock closing price in this paper are fetching data from finance.yahoo, stock analysis, using regression models from sklearn and keras for predicting the stock prices, calculating the root mean square error of each model and comparing the models.

The objective of this research is to explore different regression models and get a sense of which models is a good fit with low error in performing regression. The paper aims to use simple regression models such as Linear Regression and compare their predicted results with predictions made by complex models such as Random Forrest Regressor, Long Short-term Memory etc.

In sum, this paper makes the following contributions:

- It enumerates importance, related work and different methodologies for predicting stocks.
- It presents the analysis and implementation of a Stock Prediction System using simplest to complex deep learning models for regression and comparing their predictions.
- It provides insights about the effects of using different models while making stock predictions.

Collectively, the paper will demonstrate the analysis of stocks, different supervised learning techniques used to make predictions and their effects.

II. RELATED WORK

A. Supervised Learning

Masud Rana, Md. Mohsin Uddin, and Md. Mohaimnul Hoque [7] have stated that compared to other stock market forecast techniques, machine learning methods are proven to be more reliable and secure. At present, predicting the stock market price using machine learning is one of the most significant concerns of shareholders as it allows them to maximize profits. Along with time, machine learning methods are also developed to predict stock market precisely by the researchers.

Chun-Hung Cho, Guan-Yi Lee, Yueh-Lin Tsai, and Kun-Chan Lan [2] gave their prediction based on the papers they have surveyed on artificial intelligence-based stock market prediction. And, they evaluated the papers to see which method worked best and provided greater accuracy to the prediction of the stock price.

Asad Masood Khattak, Habib Ullah, Hassan Ali Khalid, Ammara Habib, Muhammad Zubair Asghar, and Fazal Masud Kundi [1] predicted the stock price using Linear Regression and LSTM. By using Deep Learning Model [11], a great job has been conducted on time series analysis to forecast stock market.

Pratik Patil, Ching-Seh Mike Wu, Katerina Potika, and Marjan Orang [11]. have used multiple source as input features, especially financial news and articles. They show that investor opinions through news and social media has significant effect on the market volatility. Sentiment analysis was done on financial news and fed as an input to the model. They concluded that accurate consistent news information significantly increases the accuracy of the model. [11] have used LSTM with news sentiment analysis and achieved an accuracy of 88%.

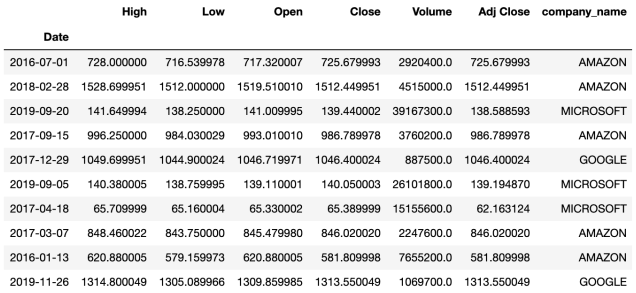
B. Reinforcement Learning

Recently an emerging branch of deep learning known as reinforcement learning (RL) have showed promising opportunities in the field of stock market prediction. Reinforcement learning agent is based on a certain set of actions and a goal which in this case is to maximize the profit [11]. A deep recurrent Q-learning was employed in this study and the results of the experiment were positive profit. This is the first positive results by a pure deep reinforcement learning algorithm under transactional costs and therefore RL provides promising opportunities for researchers in this field.

III. METHODOLOGY

A. Initial Phase: Data Collection and Analysis

For the project, past 5 years data was obtained from Yahoo Finance using Pandas-datareader library in python. The companies targeted were Apple, Google, Amazon and Microsoft.



	High	Low	Open	Close	Volume	Adj Close	company_name
Date							
2016-07-01	728.000000	716.539978	717.320007	725.679993	2920400.0	725.679993	AMAZON
2018-02-28	1528.699951	1512.000000	1519.510010	1512.449951	4515000.0	1512.449951	AMAZON
2019-09-20	141.649994	138.250000	141.009995	139.440002	39167300.0	138.588593	MICROSOFT
2017-09-15	996.250000	984.030029	993.010010	986.789978	3760200.0	986.789978	AMAZON
2017-12-29	1049.699951	1044.900024	1046.719971	1046.400024	887500.0	1046.400024	GOOGLE
2019-09-05	140.380005	138.759995	139.110001	140.050003	26101800.0	139.194870	MICROSOFT
2017-04-18	65.709999	65.160004	65.330002	65.389999	15155600.0	62.163124	MICROSOFT
2017-03-07	848.460022	843.750000	845.479980	846.020020	2247600.0	846.020020	AMAZON
2016-01-13	620.880005	579.159973	620.880005	581.809998	7655200.0	581.809998	AMAZON
2019-11-26	1314.800049	1305.089966	1309.859985	1313.550049	1069700.0	1313.550049	GOOGLE

Fig. 1. This table shows collected data from Yahoo, where close column is the target column and open, high, low are the independent features of the statistical models.

The first column "Date" denotes the date, "Open", "High" and "Low" columns indicate the starting value, peak value and lowest value of an equity, and close represents the final worth of an equity at the time of closing stock market. Similarly, "Volume" column shows the number of stocks traded on a specific day.

Matplotlib library in python have been utilized for analyzing the stock data. The library was able to generate historical stock data as a graph. The implementation of matplotlib is very simple. After getting the data from pandas-datareader, assign a variable to each stock's dataframe, for example, google for google's dataframe. For displaying the graph, only the close column is used from each stock to get a sense of the closing price during the last 5 years. Finally, the different stocks is displayed by plotting the graph.

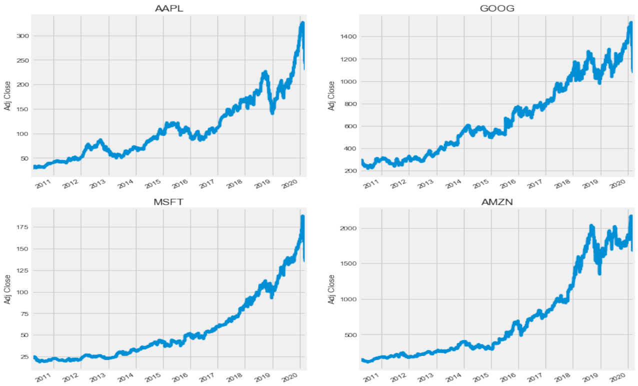


Fig. 2. This figure shows closing price of four different stocks i.e. Apple, Google, Microsoft, Amazon.

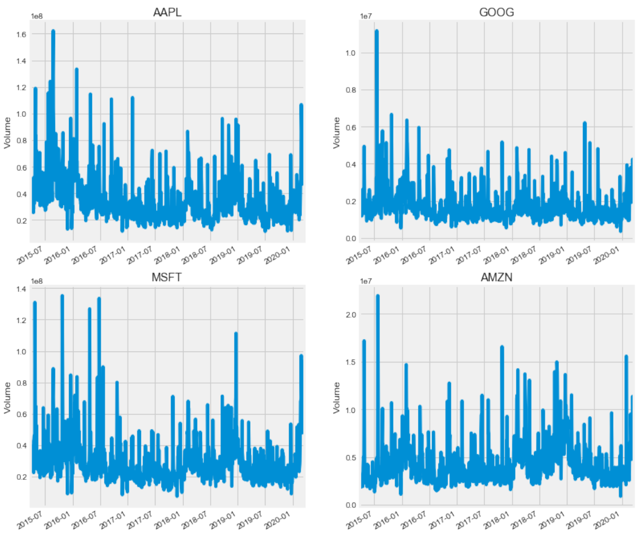


Fig. 3. This figure shows volume of stocks traded each day of four different stocks i.e. Apple, Google, Microsoft, Amazon.

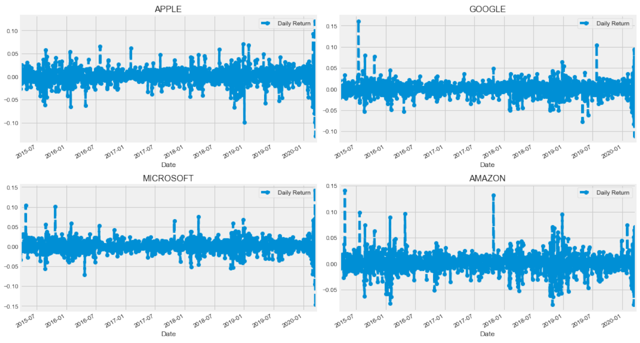


Fig. 4. This figure shows percent change of stocks traded each day of four different stocks i.e. Apple, Google, Microsoft, Amazon.

B. Second Phase: Splitting the data and predicting closing values of test set with different statistical models

The project was performed using Jupyter Notebook with Python 3.7 and classes like numpy, pandas, pandas-datareader, matplotlib-plt, sklearn and keras was heavily used for creating graphs and giving functions to the class. The stock data gain from yahoo was spitted in four parts, X_train, X_test, y_train and y_test, where X_train and X_test were the independent features used for training and predicting dependent features, y_train and y_test. The training set consisted 80% of the data and the remaing were in the testing set. sklearn and keras library is used to import statistical models like Linear Regression, Support Vector Regressor, Random Forrest Regressor, Extra Tree Regressor, Gradient Boosting Regressor, Bayesian Ridge and Long short-term memory.

1. Linear Regression

A linear regression is a simple statistical algorithm where the targeted value is a linear combination of the independent features. A linear regression line has an equation of the form Y = a + bX, where x is the explanatory variable and y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

2. Support Vector Regressor

The SVR model [7] is created using the parameters gamma, C and Radial Basis Function (RBF) kernel. Gamma component influences the model a lot. For a large value of gamma, the radius of the region of impact of the support vectors involves only the support vector itself and no amount of regulation with C will be prepared to avoid over-fitting from happening. On the other hand, for a small value of gamma, the model becomes too finite and cannot able to capture the shape of the data. In this model, gamma's value is 0.1.

3. Random Forrest Regressor

Random forest is an ensemble learning method by constructing multiple trees at training time and outputting the final class label. Decision trees have very low bias and high variance. Small noise in the data could lead the tree grow in a completely different manner. This weakness is avoided in random forest by training multiple decision trees on different subspace of the feature space at the cost of slightly increased bias. None of the trees in the forest could see the entire training data. The data is recursively split into partitions. A particular node is built according to a certain attribute. The choice of the separating criterion is based on some impurity measures such as Shannon Entropy or Gini impurity.

4. Extra Tree Regressor

The Extra-Tree method (standing for extremely randomized trees. With respect to random forests, the method drops the idea of using bootstrap copies of the learning sample, and instead of trying to find an optimal cut-point for each one of the K randomly chosen features at each node, it selects a cut-point at random.

5. Gradient Boosting Regressor

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

6. Bayesian Ridge

Ridge regression uses regularization with L2 norm, while Bayesian regression, is a regression model defined in probabilistic terms, with explicit priors on the parameters. The choice of priors can have the regularizing effect, e.g. using Laplace priors for coefficients is equivalent to L1 regularization.

7. Long short-term memory

RNN's (LSTM's) are pretty good at extracting patterns in input feature space, where the input data spans over long sequences. Given the gated architecture of LSTM's that has this ability to manipulate its memory state, they are ideal for such problems. LSTMs can almost seamlessly model problems with multiple input variables. All we need is a 3D input vector that needs to be fed into the input shape of the LSTM. So long as we figure out a way to convert all our input variables to be represented in a 3D vector form, we are good use LSTM. This adds a great benefit in time series forecasting, where classical linear methods can be difficult to adapt to multivariate or multiple input forecasting problems.

IV. EXPERIMENTS AND RESULTS

The predictions were made on four different stocks using the statistical models and a deep learning model. The predictions were made on four of the top companies that is Google, Apple, Amazon and Microsoft. The results for every type of model were merged to get a representative value for that model. This is done for the sake of simplicity in comparing different models. Below are the results for all the models.

Table 1: Calculated Root mean Squared error of each model's Prediction

	Google	**Apple**	**Amazon**	**Microsoft**
Linear R	3.78	0.55	6.07	0.34
SVR	196.44	33.30	443.60	21.05
Random Forrest R	22.80	1.05	10.49	0.41
Extra Tree R	19.97	0.99	9.88	0.44
Gradient Boosting	26.40	1.14	10.36	0.22
Bayesian Ridge R	3.78	0.55	6.07	0.34
LSTM	44.71	19.52	61.09	13.01

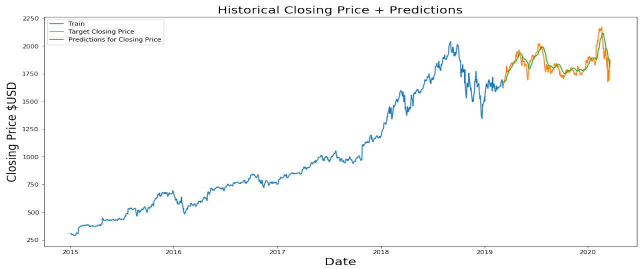


Fig. 5. This figure shows the historical value, Targeted closing price and predictions of closing price of Amazon's stock made by LSTM model.

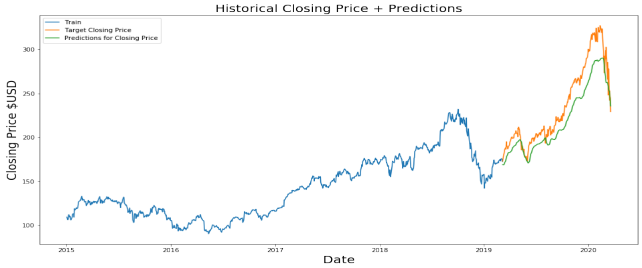


Fig. 6. This figure shows the historical value, Targeted closing price and predictions of closing price of Apple's stock made by LSTM model.

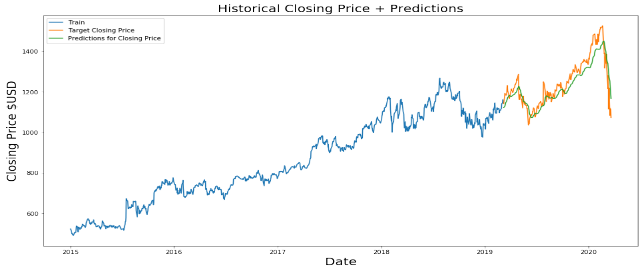


Fig. 7. This figure shows the historical value, Targeted closing price and predictions of closing price of Google's stock made by LSTM model.

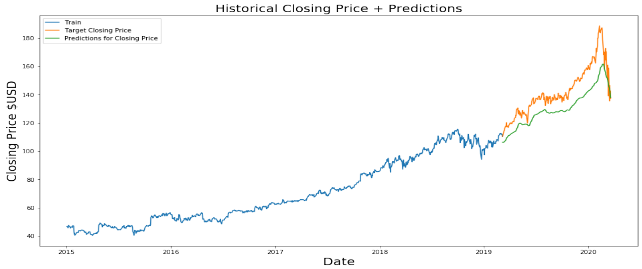


Fig. 8. This figure shows the historical value, Targeted closing price and predictions of closing price of Microsoft's stock made by LSTM model.

Figure 5,6,7,8 are the graphs of predictions made by the long short-term memory model, which was initially assumed to outperform the less complex regression models from sklearn. The results were not as I expected, linear regression outperforms most of the regression models and LSTM with lowest RMSE.

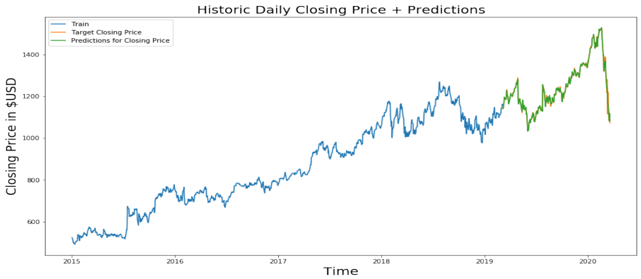


Fig. 9. This figure shows the historical value, Targeted closing price and predictions of closing price of Google's stock made by LinearR model.

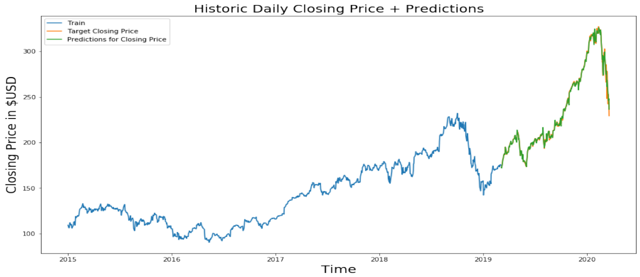


Fig. 10. This figure shows the historical value, Targeted closing price and predictions of closing price of Apple's stock made by LinearR model.

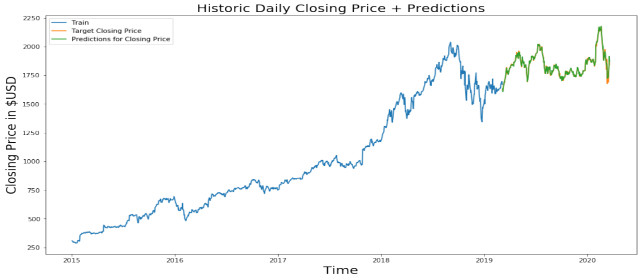


Fig. 11. This figure shows the historical value, Targeted closing price and predictions of closing price of Amazon's stock made by LinearR model.

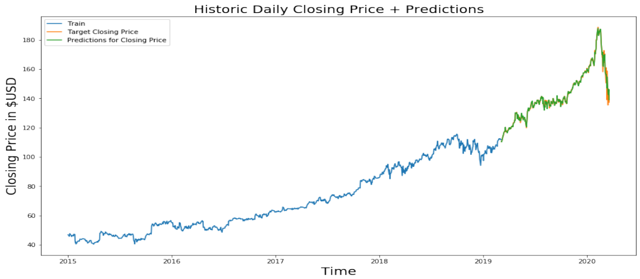


Fig. 12. This figure shows the historical value, Targeted closing price and predictions of closing price of Microsoft's stock made by LinearR model.

V. CONCLUSION AND FUTURE WORKS

Based in this paper, I applied Linear Regression, Support Vector Regressor, Random Forrest Regressor, Extra Tree Regressor, Gradient Boosting Regressor, Bayesian Ridge and Long short-term memory to predict the rise and fall of a given stock. The stock's considered for the project were Google, Amazon, Microsoft and Apple. The data observed was for the duration of 5 years. The data was first gained from yahoo finance and then analyzed using matlotlib library in python. The data was split for training and testing, where the independent features were opening, high, and low price of a given day and the dependent value was the closing price of a given day. Every statistical model was first trained with first 80% of the data and predictions were made for the last 20 %. Overall every model did good in the predictions, Linear Regression and Bayesian Ridge model outperformed every other model with the lowest root mean squared error. For this case a simple regression model outperformed complex model such as Long short-term memory and Extra Tree regressor. In the project, predictions were made on the closing price given the opening, high and low prices. Future scope of this project will involve adding more parameters and factors like the financial ratios, multiple instances, etc. The more the parameters are taken into account more will be the accuracy. The algorithms can also be applied for analyzing the contents of public comments and thus determine patterns/relationships between the customer and the corporate employee. The use of traditional algorithms and data mining techniques can also help predict the corporation's performance structure as a whole.

VI. BIBLOGRAPHY

[1] Asad Masood Khattak, Habib Ullah, Hassan Ali Khalid, Ammara Habib, Muhammad Zubair Asghar, and Fazal Masud Kundi. 2019. Stock Market Trend Prediction using Supervised Learning. In Proceedings of the Tenth International Symposium on Information and Communication Technology (SoICT 2019). Association for Computing Machinery, New York, NY, USA, 85–91. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3368926.3369680

[2] Chun-Hung Cho, Guan-Yi Lee, Yueh-Lin Tsai, and Kun-Chan Lan. 2019. Toward Stock Price Prediction using Deep Learning. In Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC '19 Companion). Association for Computing Machinery, New York, NY, USA, 133–135. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3368235.3369367

[3] Arijit Chatterjee and Kendall Nygard. 2017. Predicting Stock Close Price Using Microsoft Azure. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM '17). Association for Computing Machinery, New York, NY, USA, 749–757. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3110025.3120983

[4] Biao Huang, Qiao Ding, Guozi Sun, and Huakang Li. 2018. Stock Prediction based on Bayesian-LSTM. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing (ICMLC 2018). Association for Computing Machinery, New York, NY, USA, 128–133. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3195106.3195170

[5] Yung-Keun Kwon, Sung-Soon Choi, and Byung-Ro Moon. 2005. Stock prediction based on financial correlation. In Proceedings of the 7th annual conference on Genetic and evolutionary computation (GECCO '05). Association for Computing Machinery, New York, NY, USA, 2061–2066. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/1068009.1068351

[6] Huy D. Huynh, L. Minh Dang, and Duc Duong. 2017. A New Model for Stock Price Movements Prediction Using Deep Neural Network. In Proceedings of the Eighth International Symposium on Information and Communication Technology (SoICT 2017). Association for Computing Machinery, New York, NY, USA, 57–62. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3155133.3155202

[7] Masud Rana, Md. Mohsin Uddin, and Md. Mohaimnul Hoque. 2019. Effects of Activation Functions and Optimizers on Stock Price Prediction using LSTM Recurrent Networks. In Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence (CSAI2019). Association for Computing Machinery, New York, NY, USA, 354–358. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3374587.3374622

DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3364335.3364401

[8] Chun-Hung Cho, Guan-Yi Lee, Yueh-Lin Tsai, and Kun-Chan Lan. 2019. Toward Stock Price Prediction using Deep Learning. In Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC '19 Companion). Association for Computing Machinery, New York, NY, USA, 133–135. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3368235.3369367

[9] Yoojeong Song and Jongwoo Lee. 2019. Design of stock price prediction model with various configuration of input features. In Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing(AIIPCC '19). Association for Computing Machinery, New York, NY, USA, Article 3, 1–5. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3371425.3371432

[10] Chongyang Zhang, Zhi Ji, Jixiang Zhang, Yanqing Wang, Xingzhi Zhao, and Yiping Yang. 2018. Predicting Chinese Stock Market Price Trend Using Machine Learning Approach. In Proceedings of the 2nd International Conference on Computer Science and Application Engineering (CSAE '18). Association for Computing Machinery, New York, NY, USA, Article 83, 1–5. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3207677.3277966

[11] Pratik Patil, Ching-Seh Mike Wu, Katerina Potika, and Marjan Orang. 2020. Stock Market Prediction Using Ensemble of Graph Theory, Machine Learning and Deep Learning Models. In Proceedings of the 3rd International Conference on Software Engineering and Information Management (ICSIM '20). Association for Computing Machinery, New York, NY, USA, 85–92. DOI:https://doi-org.ezproxy2.library.drexel.edu/10.1145/3378936.3378972