Comparative Analysis of Stock Price Prediction

Jonghyun Yoo
Computer Science and Engineering Department
University of Nebraska - Lincoln
jyoo3@unl.edu

Jay Suekang Chae
Mechanical & Materials Engineering Department
University of Nebraska - Lincoln
jschae@huskers.unl.edu

***Abstract*— This project aims to predict the adjusted closing prices of Samsung, Tesla, and Facebook using linear, logistic, and artificial neural network we have learned from the class CSCE 478/878, and compare the results of each methods.**

I. Introduction

From old times, the stock market has become one of the indispensable factors to influence individual assets, corporate finance, and even global market economy, and human desires have been projected to predicting stock prices with machine learning technological advancement for the past decades. Before machine learning technologies were applied to the stock trend prediction, the prediction of stock market fluctuation was the most significant tasks of investors, but accurate prediction of future stock price is considered a challenging task due to its high level of difficulty. Because the stock market is impacted by numerous volatile and chaotic factors, such as worldwide political and economical issues, supply-demand connection of stock, and so on. Therefore, machine learning is applied to area of stock market investment.

The leading motivation based on this project is our curiosity that how much accurate our machine learning skills we have learned are possible to predict the stock price.

This paper starts with the problem definition and an information of the dataset which was used. Detail data preprocessing, exploratory data analysis, and evaluation metric are depicted. The methods which were treated at this problem, results, conclusion, and future work are at the last part of this paper.

II. Problem Definition

This project aims to predict the adjusted closing prices of Samsung, Tesla, and Facebook using linear, logistic, and artificial neural network we have learned from the class CSCE 478/878, and compare the results of each methods. For logistic and artificial neural network methods, this problem is classification type, and the regression type is applied to linear methods.

Almost every human being and all creation of people are inseparable relations from capital, and the capital assets are able to make massive profits with investment at stock market. For the past decades, machine learning has been treated at the investment area for predicting the stock market trend, and we would like to know how well or bad we can predict the stock market trend with the machine learning knowledge we have acquired from the class, CSCE 478/878 up this point.

Research Question is which method is going to show best prediction of future stock price.

Hypothesis is that Artificial Neural Network model will have best accuracy of stock price prediction, because the stock market data is easily affected by volatile and chaotic characteristics of stock market factors.

III. Dataset

The dataset, which is used in this project, stock prices of Samsung, Tesla, and Facebook from January 1st, 2017 to September 30th, 2018 for 637 days. However, Yahoo finance does not keep track of all daily data of previous years, but for all dates, Yahoo finance has recorded, yahoo finance has got highest and lowest values on each day as below Fig.1. The 424 samples for Samsung, the 439 samples for Tesla, and 439 samples for Facebook were utilized.

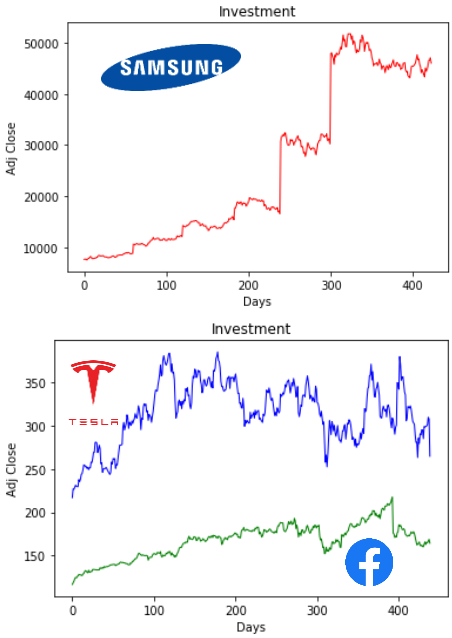

Fig. 1. Shapes of the Cost Function [Feature 1(on the left, same scale) & Feature 2(on the right)]

Date	Open	High	Low	Close	Volume	Adj Close
2015-11-02	1385000	1393000	1374000	1383000	386500	1383000
2015-11-03	1381000	1381000	1350000	1352000	301800	1352000
2015-11-04	1352000	1361000	1326000	1330000	281000	1330000
2015-11-05	1330000	1354000	1330000	1342000	173000	1342000
2015-11-06	1343000	1348000	1330000	1338000	164300	1338000
2015-11-09	1338000	1344000	1321000	1344000	185600	1344000
2015-11-10	1336000	1341000	1314000	1321000	197500	1321000
2015-11-11	1321000	1345000	1321000	1333000	140400	1333000
2015-11-12	1333000	1334000	1317000	1317000	157400	1317000
2015-11-13	1317000	1317000	1300000	1300000	177600	1300000
2015-11-16	1291000	1291000	1263000	1263000	275700	1263000
2015-11-17	1275000	1290000	1270000	1270000	186100	1270000
2015-11-18	1272000	1290000	1272000	1281000	167700	1281000
2015-11-19	1290000	1290000	1271000	1289000	192800	1289000

Fig. 2. Shapes of the Cost Function [Feature 1(on the left, same scale) & Feature 2(on the right)]

Total six features, Open, High, Low, Close, Volume, and Adj Close are used for training from the dataset, and all the features are real valued type as Fig.2. 'Open' represents the price at which the stock started trading on a particular date. 'Close' represents the price at which the stock closed on a particular date. 'High' represents the highest price stock encounter on a particular data. 'Low' represents the lowest price stock encounter on a particular date[1]. 'Volume' is the number of shares or contracts traded in a security or an entire market during a particular date[2]. 'Adj Close' is the closing price after adjustments for all applicable splits and dividend distributions[3]. The target variable is Adj Close.

IV. Data Preprocessing

We loaded the datasets using pandas's read_csv function, and instead of splitting the datasets, we downloaded separate files for training and evaluation to make it more clear which gets trained and evaluated. Training datasets are the ones from 01/01/2017 to 09/30/2018, and evaluation datasets are stock marker information for the companies from 10/01/2018 to 12/31/2018.

If there is an undefined or null value, mostly NaN in python, we dropped the those values in the dataset using dropna() function from pandas. We decided to drop these values because NaN values can cause trouble in computing.

There was only one NaN value in evaluation set for Samsung, and we removed it and then used reset_index() function so that the sample after the NaN value can take over NaN value's row number.

Since the dataset is time based, we decided not to shuffle the data. Shuffling the dataset would affect training trends.

V. Exploratory Data Analysis

We have used info() to see what features are there and using describe() we showed mean, standard deviation, min, max and quartiles of each feature of each company. Also we used isnull() to check if there exists a null value (= NaN) in the samples or not.

VI. Evaluation Metric

Since we are comparing three different models on Algorithm Trading, we have several different Evaluation Metrics.

For linear regression, we used RMSE as its metric to see the validity of the model.

For Logistic Regression and Artificial Neural Network(MLP), we used confusion matrix and f1 score to see how valid the prediction the model is producing on the datasets.

VII. Methods

A. Linear Regression

Linear regression is the most basic and the first type of regression algorithm in machine learning, and has been widely utilized in practical application, such as finance area[4]. Linear regression is also a statistical algorithm to predict the linear connection of dependent and independent variables as depicted as $Y = AX + B$ where Y is the dependent variable and X is the independent variable, B is a constant and A is the slope of the linear regression. The reason is that linear regression models rely linearly on their parameters are simpler to adapt, because the resulting estimators are statistically easily to determine[5].

B. Logistic Regression

Logistic regression is effective when forecast of the presence or absence of a characteristic or consequence based on values of a predictor variable set is required. Therefore, logistic regression could be close to a linear regression. However Logistic regression has better performance when the dependent variable is binary. Logistic regression coefficients could be utilized to calculate estimation of odd ratio for each independent variable. Logistic regression is useful to build a multivariate regression between one dependent variable and a number of independent variables. When the independent variables are continuous or categorical, and the dependent variable is binary, logistic regression estimates the multivariate explanatory model parameters[6].

C. Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is the most prospective machine learning algorithm. ANN is the non-linear model, which is able to predict by using stock market data, without former learning of the connection of input and output variables[7], so ANN have been developed and used by academia to predict the finance trend.

D. Training and Hyperparameter Tuning

There are three different models we are comparing: Linear Regression, Logistic Regression, and Artificial Neural Network by Multi-Layer Perceptron classifier. Since out intuition of the project was to compare validity of popularly used machine learning models in algorithm trading that we learned through our class, we decided to pick those three.

For training, we decided to use GridsearchCV to find the optimal parameters of each model and use the hyperparameter derived by it as the optimal parameter. We decided to use it

because we were given a lot of time to tune the parameters and it is very brute force way to check all combinations of hyperparameters.

Undoubtedly, the hyperparameters chosen are different for each model. For Linear Regression, only considered parameters that are set by sklearn and to produce best best curve we have found to turn on the.

For Logistic Regression, Maximum iteration was in range of 'max_iter' :[500, 1000, 2000]. Since the early stopping is on, the process will terminate as needed, so we kept various iteration sizes to be sure. As we have it to take solver, tol, and C value as hyperparameters.

Tol is set to be in one of [1e-3, 1e-4, 1e-5] because our algorithm's training is likely not going to be perfect, and the accuracy will not be as high. Since predicting investment is very difficult and Machine Learning approach is said to have limits. So we tried to set the tol to be not too small, but small enough that it tightens the validity.

C is in range [1,10,50] because we thought the inverse of regularization does not need to be too strong, we kept C little big.

Solver algorithm we took everything because it each function could lead to better solution in a specific combination.

For MLPclassifier, tol, alpha, and maximum iterations, are determined in a same manner as Logistic Regression, and the hidden layer sizes were limited to up to 2 layers and 200 nodes at each layer due to the limitations of our computers having trouble handling more values.

VIII. Summary of Results

A. *Linear Regression*

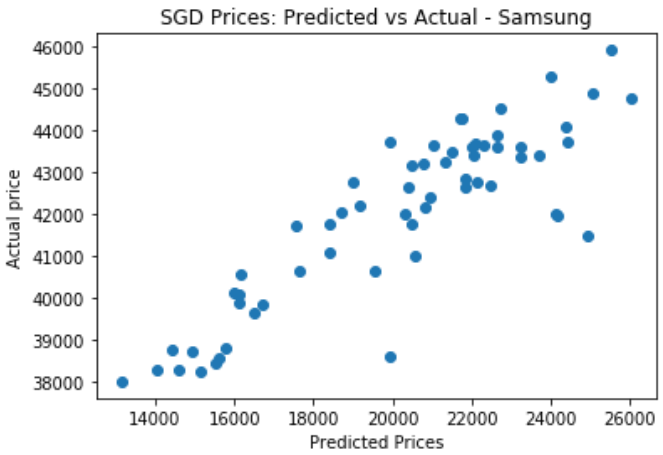

Fig. 3. Linearity Figure - Samsung

The SGDregression on Samsung's stock market prices show some what linearity in the fact that the graph shows linear trends, however the values predicted are far off, and the accuracy is too low. Even the RMSE values are too high to be considered

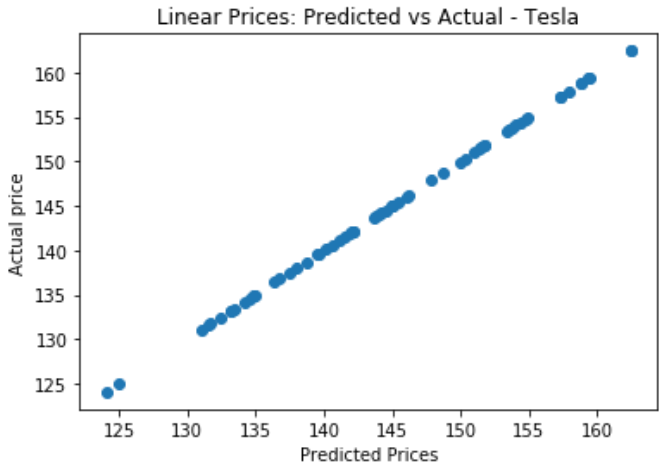

Fig. 4. Linearity Figure - Tesla

Above plot shows the graph of predicted values over actual values. The graph shows perfect linear line which means the prediction was almost 100 percent correct. However, the regression itself cannot reduce itself down to decimal levels, and in decimal levels there are quite a bit a difference. This difference in cents make huge difference when it is traded billions of times. The RMSE value was found to be 1.0128028479525118e-14.

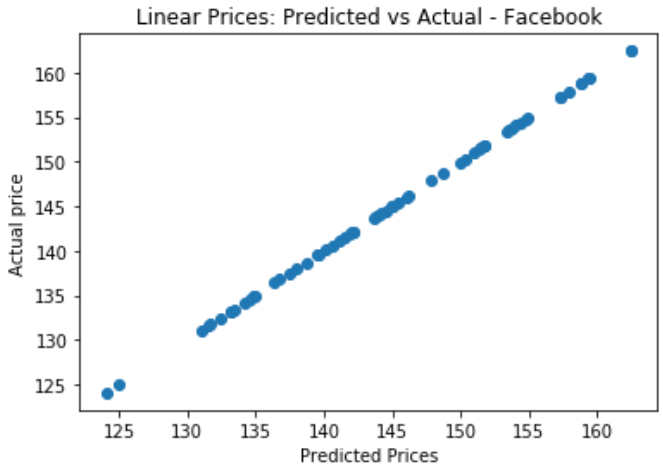

Fig. 5. Linearity Figure - Facebook

Facebook's results shows pretty much same results shown by Tesla's regression. It is making little bit of difference which will be gigantic and disastrous when such huge trades are made. The RMSE 1.0128028479525118e-14.

Since we were not using hyperparameter tuning, Samsung's data showing such low accuracy and high RMSE is thought to be caused by differences in currencies. Because USD is worth about 1,200 Korean Won, just a little bit of a difference in a dollar can be shown as a huge number in Korean won.

This problem could be avoided if we knew all the exchange rate from Korean Won to USD from Jan/01 of 2017 to end of

2018, but it is such a time consuming job, so we decided to neglect it.

Linear Regression showed reasonably good data, but it is found that prediction of actual values could be harmful with expecting that the volume of trade is over million a day.

B. Logistic Regression

TABLE I. EVALUATION OF LOGISTIC REGRESSION

	Test Accuracy	Confusion Matrix
Samsung	0.7049	$\begin{bmatrix} 25 & 11 \\ 7 & 18 \end{bmatrix}$
Tesla	0.8254	$\begin{bmatrix} 28 & 5 \\ 6 & 24 \end{bmatrix}$
Facebook	0.7302	$\begin{bmatrix} 26 & 7 \\ 10 & 20 \end{bmatrix}$

Logistic Regression on stock market data show moderate well prediction. Mean accuracy is above 0.75 which is third quartile (0.7535 specifically). We found this to be reasonably well trained data by comparing our results to those of others done by Dutta, Bandopadhyay and Sengupta showing around 0.75 of accuracy[6].

The confusion matrix shows different conclusions depending on the dataset. For Samsung, as the matrix implies, it has higher recall over precision, meaning that the model is correctly classifying correct answers, but at the same time having issues with numerous false positives. On the hand, Facebook's confusion matrix shows higher precision over recall, meaning that the data classified true were mostly indeed true, yet there were numerous false negative. Tesla's confusion matrix shows the most balanced of all, having highest F1 score, indeed Logistic Regression was the most suitable for Tesla.

One thing we came up as a problem making aspect is the fact that we were not concerned about the regularization techniques for Logistic Regression. Due to our lack of understanding in sklearn's Logistic Regression class, we only considered l1 regularization. It might not have been necessary since the dataset is considerably big.

C. Artificial Neural Network (ANN)

TABLE II. EVALUATION OF ANN

	Test Accuracy	Confusion Matrix
Samsung	0.5902	$\begin{bmatrix} 36 & 0 \\ 25 & 25 \end{bmatrix}$
Tesla	0.4762	$\begin{bmatrix} 0 & 33 \\ 0 & 30 \end{bmatrix}$
Facebook	0	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

ANN's Accuracy was in around 0.55 which was reasonable comparing the results to that of R. Sitte and J. Sitte[8] having accuracy around 0.6, and they were utilizing 22 years of data set.

However the confusion made the dataset meaningless because the derived precision, recall, F1 score are all zeros.

We have trouble training ANN with multiple layers and nodes, our computers were limited in performance so that we were only able to train the model in limited layers and nodes. Kwong[9] had similar problem in his project.

IX. CONCLUSION & FUTURE WORK

The hypothesis is wrong. Logistic regression model shows best prediction for future stock price. For future work, we can try to apply and compare other machine learning models, such as KNN, Naïve Bayes, etc.

REFERENCES

[1] P. Shinde, "How to Use Basic Machine Learning Models for Stock Market Prediction," *medium.com*, para. 10, Dec. 18, 2018. [Online]. Available: https://medium.com/@palashshinde6/how-to-use-basic-machine-learning-models-for-stock-market-prediction-6090ceb46ca5. [Accessed: Dec. 8, 2019].

[2] A. Hayes, "Volum Definition," investopedia.com, para. 1, Feb. 4, 2019. [Online]. Available: https://www.investopedia.com/terms/v/volume.asp. [Accessed: Dec. 8, 2019].

[3] "What Is the Adjusted Close?," help.yahoo.com, para. 1, [Online]. Available: https://help.yahoo.com/kb/SLN28256.html. [Accessed: Dec. 8, 2019].

[4] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, Hoboken, NJ: A John Wiley & Sons, Inc, 2005.

[5] A. Sharma, D. Bhuriya, U. Singh, "Survey of Stock Market Prediction Using Machine Learning Approach," presented at the 2017 International Conference of Electronics, Communication and Aerospace Technology, Coimbatore, India, April 20-22, 2017.

[6] S. S. Ali, M. Mubeen, I. Lal, A. Hussain, "Prediction of Stock Performance by Using Logistic Regression Model: Evidence from Pakistan Stock Exchange (PSX)," *Asian Journal of Empirical Research*, vol. 8, no. 7, pp. 247-258, 2018.

[7] L. Maciel, R. Ballini, "Design a Neural Network for Time Series Financial Forecasting: Accuracy and Rbustness Analysis," *Anales do 9º Encontro Brasileiro de Finanças*, Sao Pablo, Brazil, 2008.

[8] R.Sitte and J.Sitte, "Analysis of the predictive ability of time delay neural networks applied to the S&P 500 time series", *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol.30, no. 4, pp.568-572, 2000

[9] C. Kwong, ""Financial Forecasting Using Neural Network or Machince Learning Techniques", Bachelor. Thesis, Dept. Elect. Eng., University of Queensland, 2001