Comparison of Price Prediction Based on LSTM, GRU, Random Forest, LSSVM and Linear Regression

Jiali Liang[1, *]

[1]Department of Economics and Finance, City University of Hong Kong, Hong Kong, China

*Corresponding author: jialliang7-c@ad.cityu.edu.hk

Abstract: Investment decision-making involves numerous factors to yield significant profit. Contemporarily, various models are proposed to be used in stock price prediction. However, the traditional linear model lacks the ability to mine the implicit information of the data, resulting in difficulties in deliver satisfactory performance on nonlinear data with large fluctuations and strong noises. LSTM, GRU, Optimized Random Forest, and LSSVR were employed as training methodologies to study the effectiveness in predicting directional movements of close stock prices of TESLA from July 2018 till July 2022 in comparison with the Linear regression model. This study adopted a combination of technical and fundamental analysis to reflect various sources of influence factors in the movement of stock prices. According to the analysis, the proposed models demonstrate a better accuracy score and excelled in avoiding a severe overfitting issue found in the benchmark algorism. These results shed light on guiding further exploration on machine learning techniques.

Keywords: Stock price prediction, LSTM, GRU, Random Forest, LSSVM

1. Introduction

Nowadays, artificial intelligence and machine learning (ML) technologies are in the ascendant and quantitative investment technology based on machine learning models has been widely put into practice. Not only the process time in analyzing massive data has been shortened, but machine learning techniques are often accompanied with better generalization capabilities. Rasekhschaffe, Keywan, and Robert discussed how the inherent nature of machine learning enables itself to help discover patterns that are difficult to find with traditional linear regression and other statistical tools and possesses unparalleled advantages in solving non-linear relationships and multicollinearity problems [1]. Notwithstanding, controversies lies in the feature of noisy data in the financial market; for instance, an irrational rise and fall of certain equity in the market within a certain period interferes with the entire data set. It explains the reason that applications of quantitative strategy of machine learning models presents an extreme efficiency in the sample yet performs poorly out-of-sample. Low signal-to-noise ratio imposes issue of over-fit, by which according to a simulation experiment, model error outside sample data increases after 50 iterations.

Thanks to recent enhancement in various ML algorisms, desirable results have been demonstrated possibilities both in stable return generation and improvements of technical flaws. According to a recent empirical study [2], the proposed setting of LSTM obtained a 0.23% higher daily return prior to transaction costs compared to the standard setting in Fischer and Krauss's paper [3]. LSSVM, as an implemented algorithm over standard Support vector regression (SVR) [4] to achieve fast and simple computation in large-scale regression problem [5]. Previous study in IBM stock prediction has shown the superiority of its variant LSSVR with accuracy scored 99.41% [6].

Most researchers solely deal with technical data in model training owing to its availability and being easy to retrieve [7]. Technical analysis, however, is mainly targeting at shorter periods that does not exceed one year [8]. When it comes to the new energy automobile industry in U.S., EV registrations have surged 60% in the first quarter of 2022 [9] with forecast showing the sales volume is expected to cross the chasm in 2028 [10]. In this case, it means that electric vehicles would leap from a relatively new concept into a well-known status and it symbolizes an opportunity for the industry to achieve hypergrowth and market success. In terms of competition landscape, Tesla continues to dominate EV sales amid continuing global microchip crises owing to its efficient vertical

integration business model, productive partnership, and industry-leading products. Overall, the uptake would usher into broader space with continuous government support packages and growth of charging infrastructure growth. Technology growth drivers includes advances in battery that cut costs, as well as further innovations (e.g., autonomous driving) would stimulate consumption. More heavily expansion of charging infrastructure, together with favorable policies (e.g., free parking, low registration fee and implementation of ZEV supported) by the Biden government, enable easier path of market penetration. Even the market is promising, the severe supply chain environment has made production capacity struggled to keep pace with higher-than-anticipated demands. Manufacturers are also facing higher borrowing costs and increasingly stingy lenders owing to interest rate hikes.

Faced with such a dynamic and volatile market, this research aims to also combine fundamental data which concentrates on the influence of demand-supply economic forces on long-term trend movement. Based on above literature review, this paper will conduct comparative study on application of Linear regression model and advanced algorithms: Long-short term memory (LSTM), Gated recurrent unit networks (GRU), Random Forest combined with k-fold cross validation, and Least squares support vector regressor (LSSVR), in the leading U.S. new energy automaker TESLA. The rest part of the paper is organized as follows. The section 2 will discuss information of dataset, model details, and research process. Interpretations on experiment results will be given in section 3. Eventually, a brief summary is given in section 4.

2. Methodology

2.1 Description of data

The historical data of Tesla stock were scripted from Yahoo finance, ranging from July 9th 2018 till July 9th 2022 which in total 1007 rows. The raw dataset contains features including the open, high, low, close, adjusted close price, and transaction volumes where only the relevant factors, i.e., closing prices and volumes were extracted for use as illustrated in Figure 1. Four new variables were added to be used for model training and testing prediction accuracy as listed in Table. 1.

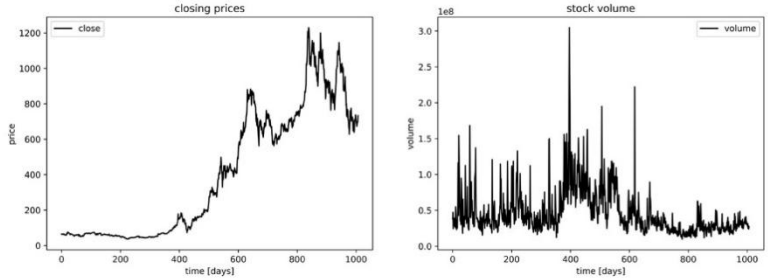

Fig. 1 Relevant features of TSLA from 09/07/2018 to 09/07/2022.

Table 1. Description of variables

Dependent Variable	Y: Daily close price of TSLA (close_tsla)
Independent Variables	X_1: Daily volumes of TSLA shift back by 1 (volume_pre)
	X_2: Monthly federal funds rate (FFR)
	X_3: Daily West Texas Intermediate index (Oil)
	X_4: Combined daily close price of CATL and Panasonic (Battery)
	X_5: Combined daily close price of Ford and GM (Competitors)

2.2 Models

Linear regression is selected to be the benchmark model for the purpose of performance comparison with the more advanced machine learning algorisms and studying relations between the 5 predictors and desired target parameters to form a pattern. Coefficients and intercepts were obtained from the regression:

$$y = 0.16 + 0.02x_1 - 0.20x_2 - 0.24\,x_3 + 0.42x_4 + 0.45x_5 \quad (1)$$

LSTM, as a special form of recurrent neural network (RNN) mainly tackles exploding and vanishing gradient problem during long sequence training process by introducing gating mechanisms

to control flows of information. Horizon needed to be determined first since a too-short or too-large window would either cause the system to merely capture compulsive reactions or letting data easily slip out of investor's memory [11]. After multiple iterations aiming at accuracy maximization, 60 days were chosen to be the time step. Similar trials were carried out on other hyperparameters such as number of layers, dropout rate, epochs, and batch size aiming at alleviating potential over-fitting situation. GRU is another variant of RNN and a simplified version of LSTM concerning only the reset gate and update gate with less parameters to largely run the risk of overfitting. Proposed model shares a similar architecture as the LSTM model with less severe dropout rate as well as smaller epochs and batch size.

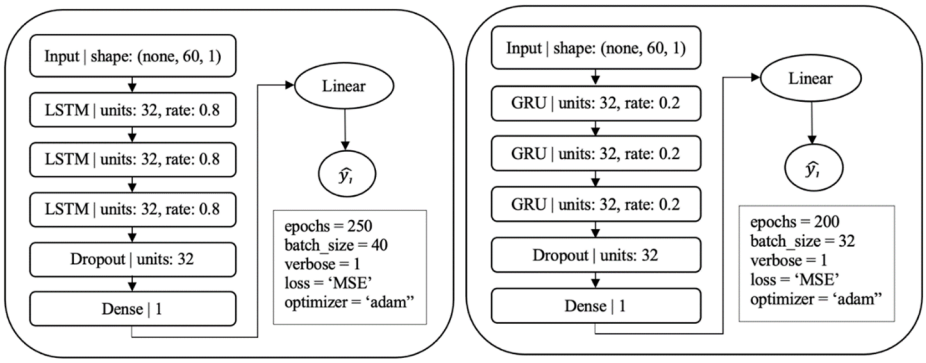

Fig. 2 The LSTM (left panel) and GRU (right panel) Architecture.

Optimization of the conventional random forest model follows the approach of hyperparameter tuning via K-fold cross validation. A parameter grid was created before instantiating the RandomizedSearchCV to search for optimized values of hyperparameters as shown in Table 2. Three-fold cross validation was adopted and the value of "n_iter" was set 100 which determines the capacity of search with a greater value suggests higher result accuracy yet longer training time. Training data were then fitted to the grid search.

Table 2. Best parameters through random grid search

'bootstrap' : True
'max_depth' : 80
'max_features' : 3
'min_samples_leaf' : 5
'min_samples_split' : 12
'n_estimators' : 100

LSSVM extends the framework of SVM, transforming the quadratic programming problem in standard SVM into solving linear equations system through considering equality instead of inequality constraints. Model efficacy depends highly on tuning hyper-parameter; therefore, the regressor was initialized through setting Radial basis function (RBF) to be the kernel function $K(x, x_i)$ with penalty parameter (Gamma) equals 0.01. The kernel function is given as:

$$K(x, x_i) = \left(1 + \frac{x_i^T x}{c}\right)^d \tag{2}$$

2.3 Workflow

As missing values and inadequate sampling are expected in the raw datasets, preprocessing is applied to improve data's quality and for ease of later preparation and transformation. Interpolation is used to convert monthly FFR into identical daily data, as well as filling in missing values with the average closing prices of the day before and after. All data was merged into one data matrix and sorted by time. To minimize potential bias and shorten training time, each feature in the merged dataset were mapped into a scale between 0 and 1 using Min-Max normalizing method by converting values x into x':

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3}$$

Finally, train-test split was performed with a ratio of 80% for training (805 rows) and 20% for evaluation (202 rows) before identifying "X_train", "X_test", "y_train", and "y_test". General workflow follows a sequence of model initialization, feeding training dataset, prediction using testing datasets, performance measurement. LSTM and GRU models require additional inputs reshaping into a structure [samples, time steps, features] before model fitting. Lastly, the predicted stock prices were inversely transformed back to its original form.

Regression metrics based on calculating distance between "X_train", "X_test", and "train_predict", "test_predict". The metrics includes Mean absolute error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2). A comparison between the scores between training and testing dataset was conducted to identify possible existence of a significant overfitting issue in this relatively small data size. The formulae are given as follows

$$MAE = \frac{1}{N}\sum_{j=1}^{N} |y_j - \hat{y}_j| \tag{4}$$

$$MSE = \frac{1}{N}\sum_{j=1}^{N} \left(y_j - \hat{y}_j\right)^2 \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N} (y_j - \hat{y}_j)^2} \tag{6}$$

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2}{\frac{1}{N}\sum_{j=1}^{N}(y_j - \bar{y}_j)^2} \tag{7}$$

Here, y_j is y_train or y_test and $\hat{y}_j$ corresponds to train_predict or test_predict while $\bar{y}_j$ represents for mean of all values.

3. Results & Discussion

3.1 Correlation analysis

The value 0.00 of Prob (F-statistic) given by the Ordinary least squares (OLS) test implies an overall meaningful regression. Individual P > |t| values of all independent variables as listed in table 3, are less than the confidence level suggests statistical significance and their association to changes in the dependent variable "close_tsla". Noteworthily, Durbin-Watson value (0.057) is less than 2, indicating a positive autocorrelation in the residuals. In consistency, Jarque-Bera test detects non-normally distributed errors.

Table 3. OLS test results

Prob (F-statistic): 0.00	Durbin-Watson: 0.057	Jarque-Bera (JB): 113.070	
	P > |t|	[0.025	0.975]
volume_pre	0.005	0.019	0.104
FFR	0.000	-0.254	-0.205
oil	0.000	0.073	0.162
battery	0.000	0.348	0.462
competitors	0.000	0.313	0.421

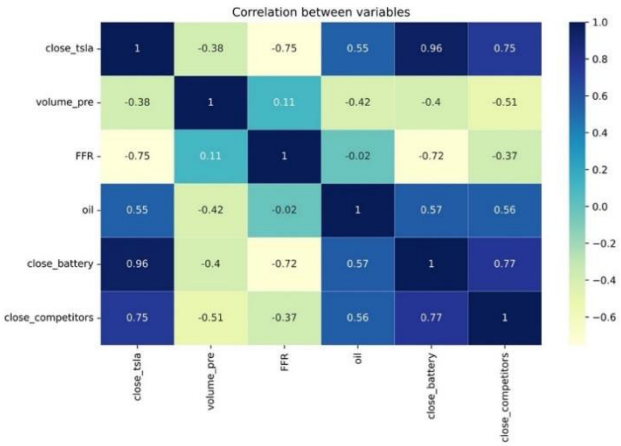

Fig. 3 The Heatmap of correlation matrix.

Correlation matrix shows a strong positive correlation of the X_4 (battery), X_5 (competitors), X_3 (oil) to the depend variable. Strong and moderate negative correlations were found in FFR and volume_pre to the closing prices of TSLA as demonstrated in Figure 3.

3.2 Comparison of regression results

To inspect model accuracy, Table 4 shows that all solution models outperformed benchmark which received a negative R-squared in testing dataset despite positive training accuracy score, implying an alleviation of over-fit. GRU excelled among all models by giving a score of 87.39%, which, given Adusumilli's claim, indicating that major profits could be made for such predicting model achieving an accuracy rate above 60% [12]. Consistency could be found in the negatively-oriented error measurements, i.e., RMSE, MSE and MAE scores, suggesting GRU outperforms all models in both training and testing datasets. In addition, Fig. 4 displays a good fit of GRU model and validation loss maintained relatively static after 50 iterations. On the contrary, Random Forest presents high error in both training and testing dataset, suggesting existence of bias and underfit.

Table 4. The Regression metrics

Models	RMSE		MSE		MAE		R^2/Accuracy	
	Train	Test	Train	Test	Train	Test	Train	Test
LR	0.06	0.22	0.00	0.05	0.04	0.18	0.93	-2.19
LSTM	0.03	0.08	0.00	0.01	0.02	0.06	0.98	0.55
GRU	0.01	0.04	0.00	0.00	0.01	0.03	0.99	0.87
RF	0.01	0.25	7.33	0.06	0.00	0.19	0.76	
LSSVM	0.10	0.10	0.00	0.06	0.01	0.23	0.91	0.52

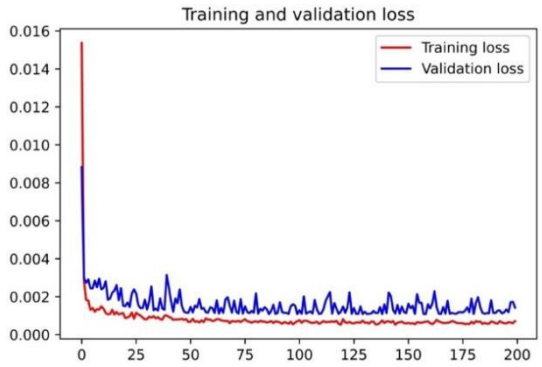

Fig. 4 Training and validation loss of GRU.

As demonstrated in Fig. 5, GRU was most successful in capturing movement patterns, following with LSTM model. It could be observed that LSSVM also roughly seized the movement and almost identified the trend towards the end whereas Random Forest performed worst and failed to properly fulfil predicting task.

3.3 Implication

According to the results, a considerable relevancy of macroeconomic factors and other impact factors synthesized with market changes in stock price prediction. To be specific, it is observed that increasing EV orders fail to boost stock price given the shares fall seen in not only Tesla but other major players in the industry. Reason behind lies within how rising inflation contributes to consumers' concern in higher costs which hurts the profit margin of automakers. Meanwhile, federal funds rate, an important instrument used to balance the high inflation, was targeted at 0.75% in July 2022, i.e., the biggest hike since 1994. The signal tells investors to remain aggressive towards cooling the rising stock prices. Moreover, the cascading effects of the 2022 Soviet-Russian war and COVID-19 lockdown, e.g., the global supply chain disruption and high gas and oil prices, could either speed up EV adoption or force a production cut. Potential threaten in TESLA's market share is also received from competitors both from vigorous growth of emerging firms (e.g., NIO) and traditional automobile makers launching successful EV models. A high positive correlation coefficient figure proposed in the research does not contradict such relationship of influence but demonstrate the way a booming market push all the players in parallel.

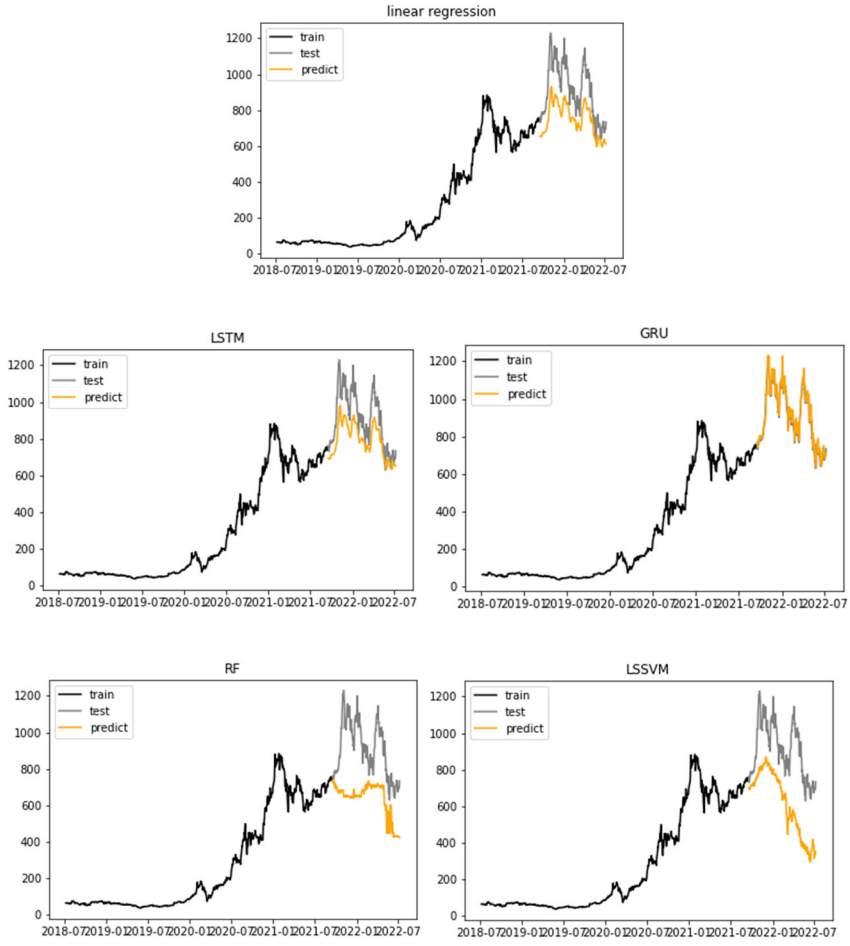

Fig. 5 Fitted graph of LR, LSTM, GRU, RF, and LSSVM.

4. Summary

In conclusion, this study carries out comparison between LSTM, GRU, optimized Random Forest based on hyperparameter tuning, and LSSVM model with Linear regression through close price forecasting. Based on experimental result, GRU has the ability to improve prediction accuracy and alleviate issue of lacking generalization capacity. Comprehensive analysis shows that GRU performed best in terms of accuracy and trend forecasting ability, followed by LSTM, LSSVR and

random forest algorithm. Yet, certain candidate variables that are believed to be significant and highly relevant to the depended variables were filtered due to unmatching data length, which thus leave rooms for future investigation. In the future, ML models could be further evaluated with their application in other contexts, e.g., Asian stock markets within the same period, smaller firms, and stock price during economic recession. Overall, these results offer a guideline for stock prediction based on different models.

References

[1] Rasekhschaffe Keywan Christian, and Robert C. Jones. Machine learning for stock selection. Financial Analysts Journal, 2019, vol. 75.3, pp. 70-88.

[2] Ghosh Pushpendu, Ariel Neufeld, and Jajati Keshari Sahoo. Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. Finance Research Letters 2022, vol. 46, 102280.

[3] Fischer Thomas, and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 2018.vol. 270.2 , pp. 654-669.

[4] Cortes Vapnik, Cortes C., Vapnik V. Support-vector networks, Machine learning 1995, vol. 20.3, pp. 273-297.

[5] Gestel T. V. et al. Benchmarking least squares support vector machine classifiers. Machine Learning, 2004, 54(1), pp. 5–32.

[6] Chandana C., and K. Vijitha. Stock market prediction using machine learning techniques. Int J Comput Sci Mob Comput, 2019, vol. 8.2, pp. 44-48.

[7] Jiang Weiwei. Applications of deep learning in stock market prediction: recent progress. Expert Systems with Applications, 2021, vol. 184, 115537.

[8] Wilder J. Welles. New concepts in technical trading systems. Trend Research, 1978.

[9] LaChance Dave. Report: EV Registrations Surge 60% in First Quarter of 2022." Repairer Driven News, 31 May 2022, Retrieved from https://www.repairerdrivennews.com/2022/05/12/report-ev-registrations-surge-60-in-first-quarter-of-2022/.

[10] Ev Sales Forecasts. EVAdoption, 26 Apr. 2022, Retrieved from: https://evadoption.com/ev-sales/ev-sales-forecasts/.

[11] Crain, Dylan M. Stock Movement Prediction using Technical and Data. Standford University Online.

[12] Adusumilli R. Predicting stock prices using a keras LSTM model. Medium, 2019.