Stock Price Prediction of Walmart Based on Combination of SVM and LS-SVM Models

Chenguang Yan[1, *]

[1] Department of Economics, University of Toronto, Toronto, Canada

*Corresponding author: chenguang.yan@mail.utoronto.ca

Abstract. One of the most significant operations in the finance sector is stock trading. The stock market is an essential part in the economy of a country and serves as the indicators of the situation of a country's economy as the stock prices go up or down. Therefore, stock price prediction, the behavior of attempting to predict the potential worth of a corporation or any financial instruments successfully, will maximize investor's gain, enhance market's confidence, and help government policymakers to make economic decisions. In order to forecast the price of a stock, a machine learning approach is constructed in this study. The suggested algorithm includes random forest, support vector machine (SVM), and least square support vector machine (LS-SVM). In particular, the random forest is employed to select the most important features from the technical indicators calculated for stock price prediction. The SVM and the LS-SVM models are employed to predict the daily stock prices. Besides, R-Squared (R^2), mean squared error (MSE) and mean absolute error (MAE) are used for model evaluation. According to the results, both SVM and LS-SVM models can predict stock price well, but both algorithms are not suitable for large datasets, and overfitting problem exists. These results shed light on guiding further exploration of stock price predictions.

Keywords: Stock Price Prediction; Support Vector Machine; Least Square Support Vector Machine; Technical Indicators; Machine Learning.

1. Introduction

For a long time, stock price forecasting has been in the spotlight. In 1602, the Dutch East India Company became the first joint stock company. In 1606, it offered shares to public investors. Since then, people invest in stocks for potential return. However, forecasting the stock market accurately is a difficult task, mostly as a result of a stock time series' behavior that resembles a random walk [1]. Moreover, due to the underlying structure of the financial sector and in part owing to the combination of known criteria and unknown factors, stock prices are thought to be highly dynamic and subject to sudden adjustments [2]. Mainly, there are two types of methods to analyze stock performance before investors investing in a stock, i.e., fundamental analysis and technical analysis [3].

The main idea behind the fundamental analysis is to evaluate the intrinsic value of any financial instrument by analyzing macroeconomic factors (Gross Domestic Product, economic growth rate etc.) and microeconomic factors (company's profit, demand of capital etc.), as well as compare the stock's value to its intrinsic value to determine whether the stock is overpriced or underpriced. Fundamental analysis assumes efficient market hypothesis holds, and the value of all financial instruments will eventually equal to its intrinsic value. On the other hand, technical analysis is the study of overall market sentiment as it manifests in asset purchases and sales. It is predicated on the notion that supply and demand interact to set pricing. Price and volume reflect the collective behavior of buyers and sellers. Investors using technical analysis believe the patterns and trends tend to repeat and can be identified and used for forecasting prices. Different from fundamental analysis, technical analysis makes the assumption that the efficient markets hypothesis is false and that both rational and irrational investor behavior are reflected in market prices. This assumption is more realistic since not all investors are rational when they making investment decisions. Therefore, this paper focuses on technical indicators instead of fundamental indicators.

With the development of technology, the importance of machine learning in stock market forecasting is rising. Artificial Neural Networks (ANN) is the most popular technique used in stock price prediction. However, since there are so many parameters to adjust and the user has no prior

knowledge regarding the importance of the inputs in the investigated situation, it has an overfitting problem [4]. In addition, SVM model is one of the most useful algorithms for data classification [5], regression [6], and prediction [7], which avoids such limitations suffered from ANN models [8]. The success of using SVM to predict stock prices is supported by the strong theoretical underpinnings based on VC-theory [9]. The least-squares support vector machine (LS-SVM) model is the least-squares version of the support vector machine (SVM) model, which is a collection of related supervised learning techniques used for both classification and regression analysis. The user does not need to know in advance how the free parameter values will affect the problem being examined in the LS-SVM model [10].

In this paper, both SVM and LS-SVM models are selected to forecast Walmart future stock close price based on technical indicators and previous close price. Random forest algorithm is employed to determine the feature importance. Besides, Grid Search Cross Validation technique is employed for hyperparameter tuning to find out the optimal parameters (sigma, epsilon etc.) for both SVM and LS-SVM models. Finally, using MES, MAE and R^2 score for model evaluation. The aim of this paper is to compare the performance and efficiency of SVM and LS-SVM models in a single company's stock price prediction, and to determine the best conditions for improving the predicted result when using the both models. The remaining parts of the essay is set up as follows. The Sec. 2 will present the support vector machine, least-squares support vector machine algorithms, and the data used for Walmart's stock price predictions. The results, model evaluations, empirical analysis, and limitations are discussed in Sec. 3. Eventually, Sec. 4 provides a brief summary.

2. Methodology

In this paper, the task of stock price prediction of Walmart is mainly carried out by the Support Vector Machine (SVM) and Least-Squares Support Vector Machine (LS-SVM). The flow chart in Fig. 1 shows the process of this project.

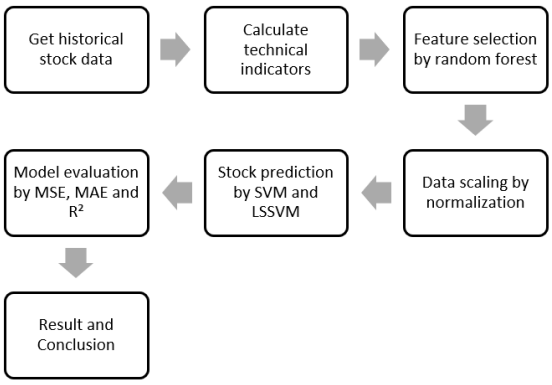

Fig. 1 Flow Chart for Process of Stock Prediction.

2.1 Support Vector Machine

The classification and regression analysis tasks of the support vector machines, which are supervised learning models with accompanying learning algorithms, both need data analysis. This paper will use support vector regression (SVR) for stock price prediction. The main idea behinds SVR is to find the best line that is the hyperplane and has the maximum number of points. SVR aims to maximize the margin while minimizing error by modifying the hyperplane.

Since the goal of SVR is to find the best line in the hyperplane, the SVR involves in plotting of datapoints in the n-dimensional space. These dimensions are attributes that are plotted on particular coordinates. It draws a boundary over the datasets that separates the data into different classes as shown in Fig. 2.

Given a dataset in n dimension $\{(x_1, y_1), (x_2, y_2),\ldots, (x_i, y_i),\ldots,t(x_n, y_n) \mid x_i \in \mathrm{R}^n, y \in \mathrm{R}, i = 1,2,\ldots,n\}$, the following is the output space to input space regression function: $f(x) = \omega \cdot \Phi(x) + \alpha, x \in R^n, \alpha \in$

R, ω and α are weight and threshold, respectively. Finding the function of f(x) that makes the difference between the true and the predicted values is less than or equal to a specified error term *e* is the objective of SVR. Afterwards, the issue is changed into a conditional restricted optimization problem and can be stated as follows:

$$min\frac{1}{2}\|\omega\|^2 + C\cdot\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$
$$\text{s.t.}\begin{cases} y - f(x_i) \le \varepsilon + \xi_i, \xi_i \ge 0 \\ y - f(x_i) \ge \varepsilon + \xi_i^*, \xi_i^* \ge 0 \end{cases} \quad (1)$$

where the penalty parameter C denotes the accuracy of the sample fitting and the complexity of the SVR model. The insensitive loss coefficient is the parameter ε; the slack variables are only significant for outliers. The pairwise issue with the aforementioned equation can be represented as follows:

$$max -\frac{1}{2}\sum_{i,j=1}^{n}(\beta_i - \beta_i^*)(\beta_j - \beta_j^*)K(x_i, y_j) - \varepsilon\sum_{l}^{i=1}(\beta_i + \beta_i^*) + \sum_{l}^{i=1} y_i(\alpha_i - \alpha_i^*)$$
$$\text{s.t.}\begin{cases} \sum_{j=1}^{n}(\beta_i - \beta_i^*) = 0 \\ C \ge \beta_i \ge 0, C \ge \beta_i^* \ge 0 \end{cases} \quad (2)$$

In order to achieve the dimensionality reduction effect, the function of kernel *K(xi, yi)* is used to make the input space of low dimension equal to the inner product of the space of high dimension. To be specific, there are three kinds of kernel function: polynomial, linear, and Gaussian kernel functions. Three kernel functions are listed in Table. 1. In this work, linear kernel is used for SVR based on the best parameters selected by Grid Search Cross Validation technique.

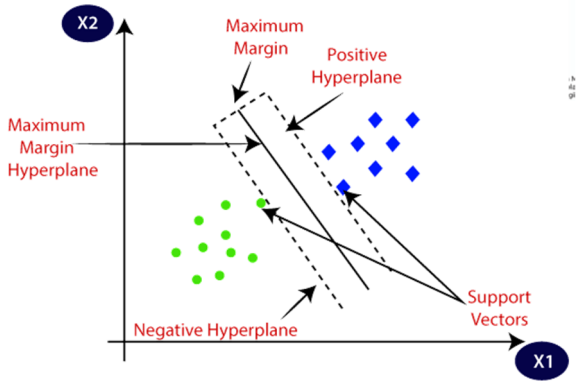

Fig. 2 SVM Decision Making Boundary [11].

Table 1. Three Kernel functions for SVM.

Kernel Name	Kernel Function
Linear Kernel	$G(x_j, x_k) = x_j' x_k$
Polynomial Kernel	$G(x_j, x_k) = (1 + x_j' x_k)^q$
Gaussian Kernel	$G(x_j, x_k) = \exp(-\|x_j - x_k\|^2)$

2.2 Least Squares Support Vector Machine

Least squares support vector machines (LS-SVM) model are least squares versions of support vector machines (SVM). The minimization problem of LS-SVM algorithm is similar to conventional SVM model. However, the primary distinction between these two methods is that the LS-SVM model, which is based on a least square cost function, contains equality requirements rather than inequalities. In addition, the SVM model focuses on solving a quadratic problem while LS-SVM model focuses on solving a linear problem. Given the input data matrix X as *n* times *p*, and an output vector y as *n* times 1, the training dataset can be expressed as $\{x_i, y_i\}^n{}_{i=1}$, where $x_i \in R^p$, and $y_i \in R$, the main task for LS-SVM algorithm is to build the function $f(x) = y$, where the output y_i is dependent on the independent variable x_i. The function is expressed as:

$$f(x) = W^T\varphi(x) + b \quad (3)$$

where W and $\varphi(x)$: $R^p \rightarrow R^n$ are vectors shaped as n times 1, and $b \in R$. Subsequently, the problem of optimization and the constraints of equality are expressed as below:

$$f(x) = W^T\varphi(x) + b \min_{w,e,b} j(w,e,b) = \frac{1}{2}w^T w + C\frac{1}{2}e^T \quad (4)$$

$$y_i = w^T\varphi(x_i) + b + e \quad (5)$$

In this case, the parameter for balancing the solution size and training errors is $C \in R^+$, and e is the n times 1 vector with all elements set to 1. When a Lagrangegian is generated in Equation (4) and differentiated with regard to *w, b, e*, and *a* (where *a* is the Lagrange multipliers), the following results are obtained:

$$\begin{bmatrix} I & 0 & 0 & -Z^T \\ 0 & 0 & 0 & -1^T \\ 0 & 0 & CI & -I \\ Z & 1 & I & 0 \end{bmatrix} \begin{bmatrix} W \\ b \\ e \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ y \end{bmatrix} \quad (6)$$

$$Z = [\varphi(x1), \varphi(x2), \ldots, \varphi(x_n)]^T \quad (7)$$

where I in equation (6) represents the identity matrix. The kernel matrix will then be defined as $K = ZZ^T$, and the parameter $\lambda = C^{-1}$, the following general solution results from the requirements for optimality:

$$\begin{bmatrix} 0 & 1^T \\ 1 & K + \lambda I \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (8)$$

The kernel functions for LS-SVM algorithm is similar to SVM algorithm, it also can be divided into four types: linear, polynomial, radial basis function (RBF), and multilayer perceptron (MLP). These four kernel functions can be expressed as shown in Table. 2. In this work, RBF kernel is used for LS-SVM based on the best parameters selected by Grid Search Cross Validation technique.

Table 2. Four Kernel functions for LS-SVM.

Kernel Name	Kernel Function
Linear Kernel	$K(x, x_i) = x_i^T x$
Polynomial Kernel	$K(x, x_i) = (1 + x_i^T x/c)^d$
RBF Kernel	$K(x, x_i) = \exp(-\|\|x - x_i\|\|^2/\sigma^2)$
MLP Kernel	$K(x, x_i) = \tanh(kx_i^T x + \theta)$

2.3 Data

All datasets are polled from yahoo finance [12]. This paper focus on Walmart's daily stock-related data for three years (from July 2019 to July 2022) and five years (from July 2017 to July 2022). The main reason to investigate Walmart's stock price is, Walmart is one of the biggest brick-and-mortar retailers in the world. As of September 2021, the company has more than 10500 locations across the globe [13]. As the Walmart's stock price is closely related to people's daily life, it is worth to predict Walmart's stock price because it reflects people's purchasing behavior, consumer's demand of goods, and residents' wealth. Thus, predicting Walmart's stock price can be useful for forecasting economic environment in the future.

The independent variable is tomorrow's stock daily close price, and dependent variables are today's close price, and some most important calculated technical indicators selected by random forest algorithm. Since the algorithm is only used for feature selections, the details will not be discussed in this paper. Recursive feature elimination (RFE), which starts with every feature in the training dataset and effectively eliminates features one at a time until the required number of features is left, is the key concept. In this work, top three most important features are selected from total of thirteen features. The features used for predictions based on SVM and LS-SVM are: today's close price, convergence or divergence moving average (MACD), and the difference between the MACD and the 9-day exponential moving average (EMA) of the MACD for the trigger line that trigger for the convergence or divergence value (MACDH). The explanations of these technical indicators are shown as below:

- *Moving Average* (*MA*): a technical indication that is used to identify a trend's direction. To determine an average, it adds up the data points for the financial instruments over a given time period and divides the sum by the total.
- *Exponential Moving Average* (*EMA*): a type of moving average, the difference between EMA and MA is that EMA focuses on recent price changes and gives more weights to the most recent price points. The formula is given in Equation (9):

$$\text{EMA} = [\alpha * T \text{ Close }] + [1 - \alpha^{*}\text{YEMA}] \quad (9)$$

Where T means today's close price and Y means yesterday's close price.

- *Convergence or divergence moving average* (*MACD*): a momentum trend-following indicator that depicts the relationship between two stock price moving averages. The formula is given in Equation (10):

$$\text{MACD} = [0.075 \times \text{EMA } of\ Closeprices] - [0.15 \times EMAofcloseprices] \quad (10)$$

The three-years and five-years Walmart's stock data then is both split into training datasets and test datasets. For three-year dataset, there are 689 and 68 datapoints for traning and test datasets respectively. For five-year dataset, there are 1144 and 114 datapoints for training and test datasets respectively. All datasets include three features. In addition, distances are very sensitive to the order of magnitude of the features in both support vector machines and least-squares support vector machines. Therefore, it is necessary to apply scaling transformations on training datasets before using models to predict stock price. In this paper, normalization technique is used. Besides, the estimator parameters are exhaustively searched and optimized by grid search cross-validation method.

3. Emiprical analysis

In this section, the results and limitations of proposed support vector machine (SVM) and least-squares support vector machine (LS-SVM) algorithms will be presented and discussed.

3.1 Correlation analysis

The correlation analysis is introduced to check the reliability of the results given by these two models. The correlation matrices of three years and five years are shown in the Tables. 3 and 4, respectively.

Table 3. Three Years Correlation Matrix

	Today's Close	MACD	MACDH	Tomorrow's Close
Today's Close	1	0.3	0.12	0.99
MACD	0.3	1	0.38	0.29
MACDH	0.12	0.38	1	0.12
Tomorrow's Close	0.99	0.29	0.12	1

Table 4. Five Years Correlation Matrix

	Today's Close	MACD	MACDH	Tomorrow's Close
Today's Close	1	0.09	0.02	1
MACD	0.09	1	0.37	0.09
MACDH	0.02	0.37	1	0.02
Tomorrow's Close	1	0.09	0.02	1

Correlation measures the linear relationship between two variables, it has no unit and ranges from negative one to positive one. As the correlation coefficient approaches to absolute, which indicate two variables are strongly correlated. Besides, correlation coefficient approaches to zero indicating no linear relationship exists among two variables. The two tables above indicate the correlation between predictor variables (today's stock close price, MACD, and MACDH) are quite low for both three- and five-years datasets, which means the features used in the price predictions for both SVM and LS-SVM models are not strongly correlated. Therefore, there is a low chance that the performances of the model will be impacted by the problem of multicollinearity that may cause misleading results. In addition, since the best parameters are exhaustively filtered by grid search cross

validation that avoid subjective judgement, it is believed that the predicted results also are the best represented both in three- and five-year datasets of Walmart's stock price information.

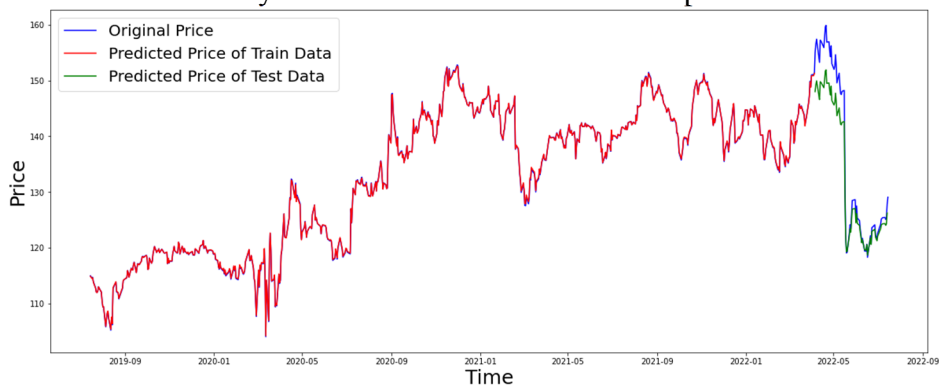

Fig. 3 SVM Results for Three Years.

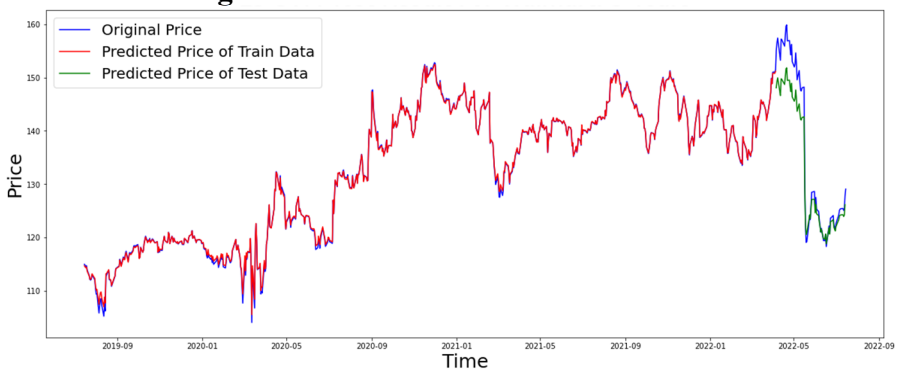

Fig. 4 LS-SVM Results for Three Years.

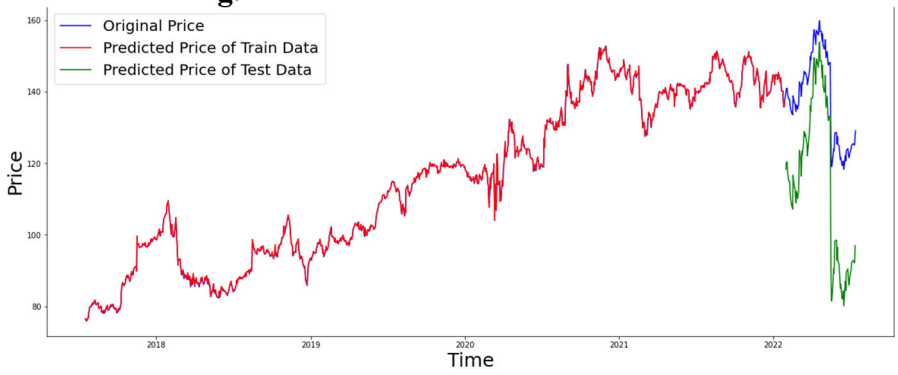

Fig. 5 SVM Results for Five Years.

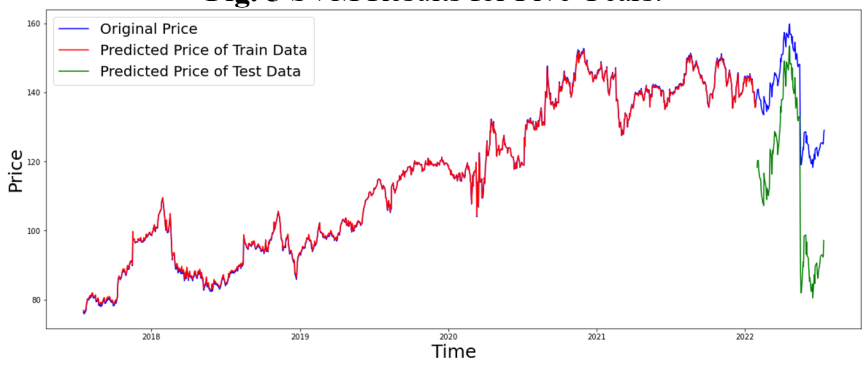

Fig. 6 LS-SVM Results for Five Years

3.2 Regression models

Figures. 3 and 4 outline the application of SVM and LS-SVM models on three-years Walmart's stock datasets (from July 2019 to July 2022). Figures. 5 and 6 present the results for five years (from July 2017 to July 2022). The blue line represents original historical stock's daily closing prices, the

red line represents the predicted prices on training datasets, and green line represents the predicted results on test datasets. The summary of model evaluations is shown in the tables below, where Table. 5 shows the model performances on three years datasets, and Table. 6 gives the model performances on five years datasets.

Table 5. Model Evaluations on three-year Datasets

Model Performances on three-year Datasets		
	Support Vector Machine	**Least-Squares Support Vector Machine**
Main Tuning Parameters	Kernel: linear, C = 10, Epsilon = 0.01	Kernel: RBF, Gamma = 100.0, Sigma = 10.0
MSE on Train Data	3.3340	3.2170
MES on Test Data	23.0232	22.9883
MAE on Train Data	1.2117	1.2102
MAE on Test Data	3.7351	3.7421
R^2 on Train Data	0.9781	0.9790
R^2 on Test	0.8996	0.8998

Table 6. Model Evaluations on five-year Datasets

Model Performances on five-year Datasets		
	Support Vector Machine	**Least-Squares Support Vector Machine**
Main Tuning Parameters	Kernel: linear, C = 100, Epsilon = 0.1	Kernel: RBF, Gamma = 1000.0, Sigma = 100.0
MSE on Train Data	2.5610	2.5952
MES on Test Data	621.1760	617.2734
MAE on Train Data	1.0227	1.0445
MAE on Test Data	22.7761	22.7801
R^2 on Train Data	0.9948	0.9947
R^2 on Test Data	-2.9795	-2.9545

3.3 Explanations

Figure. 3 and Figure 4 indicate that the performances on both support vector regression (SVR) and least-squares support vector machine (LS-SVM) algorithms on three-year datasets. It is observable that the predicted curves using the proposed models are equally closing to the real trend of Walmart's daily stock price, especially in fluctuation cases. Moreover, it is remarkable that the incredible performance of both models can predict the price directions (up or down) accurately. Beisdes, the model evaluations given by Table. 5 supports this result, both models achieve the low mean squared error (MSE), mean absolute error (MAE), and high R^2 scores. However, the similar performances on both models might suggest that there is no significant difference using these two models in prediction tasks.

On the other hand, both SVM and LS-SVM models have worse performances on five-year datasets to compare with the results given on three-year datasets. Seen from Figure.5 and Figure. 6, the predicted curves using both models are poorly fitted to the actual trend of Walmart's daily stock price with larger-fluctuated patterns, and the model evaluations given by Table. 6 supports this result. In particular, the negative coefficient determinations (R^2 scores) on both models indicate that the prediction results of both models approximately equivalent to random guesses, even worse than just calculating the average value of target variable. This implies that SVM and LS-SVM models are not suitable for large datasets. The main reasons are given as follows. Firstly, the nature of complexity of large dataset makes the predicted results of both models is difficult to fit the actual stock market trends. Second, both algorithms are not suitable for large datasets because their training complexity is highly depended on the size of datasets. In addition, this paper defines that the suitable data sizes should be within three years (approximately no more than 700 data points for training purpose). Moreover, both models can predict the stock daily prices' directions (up or down) well on both three- and five-years datasets. As illustrated in Figure. 3 to Figure. 6, the actual and predicted curves go up or down simultaneously, indicating although SVM and LS-SVM models are not suitable for large datasets, they still can be used for classification tasks regardless of the size of dataset.

3.4 Limitations

The limitation of this paper are discussed as follows. Primarily, the fewer features are selected for training purpose that leads some inefficient features, the features have low correlation relationship, are selected based on limited total number of features by grid search cross validation, which causes the imbalanced weights for features. For instance, the "today's close price" is relatively overweighted compared with other features. Additionally, overfitting problems exist in both three- and five-years datasets based on Table. 5 and Table. 6 as the MSE, MAE are lower on the training datasets compared with the scores on test datasets, and the coefficient determinations (R^2 scores) are higher on training datasets compared with test datasets. This suggests that both models should be used with combination of other models or methods (e.g., particle swarm optimization) to overcome the overfitting problems that found in support vector machines and least-squares vector machines.

4. Summary

In conclusion, the support vector machine (SVM) and the least-squares support vector machine (LS-SVM) models are applied in comparison of Walmart's stock price prediction tasks using today's close price and two technical indicators: MACD and the EMA of it. The random forest algorithm is used for determination of feature importance, and the grid search cross validation technique is applied for searching best parameters in SVM and LS-SVM models. Both SVM and LS-SVM models give the similar prediction performance on three- and five-years dataset.

However, both models are only giving the highly accurate predictions on three-year datasets rather than the predictions on five-year datasets indicating they are only suitable for stock price predictions on relative smaller datasets. Additionally, the overfitting problems found in both models suggesting that these two models should be used in combination with other algorithms to reduce this problem. In the future, when applying SVM and LS-SVM models, feature selections, dataset sizes, and other techniques should be taken into consideration for more accurate prediction outcomes. Overall, these results offer a guideline for stock price predictions using both SVM and LS-SVM models, they achieve better results on smaller datasets than larger datasets.

References

[1] Hegazy O, Soliman Omar S, and Salam Mustafa A. A Machine Learning Model for Stock Market Prediction. International Journal of Computer Science and Telecommunications, vol. 4, no. 12, December 2013, pp. 17-23.

[2] Shah V H. Machine learning techniques for stock prediction. Foundations of Machine Learning| Spring, 2007, 1(1): 6-12.

[3] Reddy V K S. Stock market prediction using machine learning. International Research Journal of Engineering and Technology (IRJET), 2018, 5(10): 1033-1035.

[4] Tao X, Renmu H, Peng W, et al. Input dimension reduction for load forecasting based on support vector machines. 2004 IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies. Proceedings. IEEE, 2004, 2: 510-514.

[5] Burges C J C. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 1998, 2(2): 121-167.

[6] Smola A J, Schölkopf B. A tutorial on support vector regression. Statistics and computing, 2004, 14(3): 199-222.

[7] Müller K R, Smola A J, Rätsch G, et al. Predicting time series with support vector machines. International conference on artificial neural networks. Springer, Berlin, Heidelberg, 1997: 999-1004..

[8] Luca Gabriele De. Advantages and Disadvantages of Neural Networks against SVMs. Retreieved from: Baeldung on Computer Science, 20 June 2022, https://www.baeldung.com/cs/ml-ann-vs-svm

[9] Vapnik V. The nature of statistical learning second edition, Springer, 1999.

[10] Alvarez Meza A M, Daza Santacoloma G, Acosta Medina C D, et al. Parameter selection in least squares-support vector machines regression oriented, using generalized cross-validation. Dyna, 2012, 79(171): 23-30.
[11] Support Vector Machine (SVM) Algorithm – Javatpoint, *www.javatpoint.com*, Retreived from: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.
[12] Data of Walmart on yahoo finance. Retreived from: https://finance.yahoo.com/quote/WMT?p=WMT&.tsrc=fin-srch
[13] Tarver Evan. Benefits of Investing in Walmart. Investopedia, Investopedia, 13 July 2022, Retreived from: https://www.investopedia.com/articles/markets/092115/these-are-benefits-investing-walmart.asp