International Conference on Business Research
University of Moratuwa, Sri Lanka
December 3, 2021, 1-8

BUSINESS RESEARCH UNIT
FACULTY OF BUSINESS
UNIVERSITY OF MORATUWA

SELECTING A SUITABLE VARIABLE COMBINATION TO PREDICT STOCK PRICES USING SUPPORT VECTOR MACHINES

[1]Pasindu Indikadulle, [2]Dasni Hendahewa, [3]Nelusha Perera, [4]Dimithri De Meraal, [5]Sulanie Perera and [6]Samadhi Rathnayake

[1, 2, 3,4,5,6]Department of Decision Sciences, University of Moratuwa, Sri Lanka

Emails: [1]*1176033p@uom.lk*, [2]*dasni.17@business.mrt.ac.lk*, [3]*nelusha.17@business.mrt.ac.lk*, [4]*dimithri.17@business.mrt.ac.lk*, [5]*sulaniep@uom.lk*, [6]*samadhic@uom.lk*

ABSTRACT

With technological development, trading in stock markets has become more accessible to the general public. However, owing to the highly volatile nature of stock prices, stock price predictions remain a challenging task. Literature shows Support Vector Machines as a promising technique. This paper aims at identifying the best variable combination to predict the stock prices using Support Vector Machines along with the application of forward filling and linear interpolation as data filling methods and random search and grid search as hyper parameter optimization methods. After the individual evaluation of all models, data filling method of linear interpolation, hyper parameter optimization method of grid search and independent variable combinations with adjusted close price are found to give better results for prediction of stock prices.

Key Words: Support Vector Machines, Forward filling, Linear interpolation, Random search, Grid search

1. Introduction

Markets where stocks are traded publicly between stockbrokers and companies, are commonly referred to as stock markets (Chou & Nguyen, 2018). Investing in such markets involves risk due to the highly volatile and nonlinear nature of stock prices that are dependent on various factors such as political and economic conditions, traders' expectations, etc. (Vijh et al., 2020) Thereby, making precise stock price predictions, rather a difficult task. Yet, as rational investors are motivated to follow effective strategies in order to generate positive returns on their stock market investments, stock price prediction is considered to be an interesting area of research (Pagolu et al., 2016).

Pertinent literature shows Support Vector Machines (SVM) producing promising results in predicting stock prices when compared to other contemporary machine learning models. There are further studies that tune SVM models through optimizing the hyper parameters, in order to achieve more accurate predictions.

While the focus of most previous work relies on the development of the SVM algorithm itself, this paper further extends the use of SVM in predicting stock prices using different variable combinations and data filling methods. This study aims at identifying a suitable variable combination to predict stock prices using SVM by comparing their performances based on the forecasted accuracies and it has been organized as follows: Section 3 provides a detailed data description and introduces different variable combinations, data filling and hyper parameter optimization techniques which are used to carry out the data analysis. Section 4, compares these models as per the results obtained. Finally, section 5 concludes the research with a discussion.

2. Literature Review

Over the past few decades, researchers have done various studies related to forecasting the behavior of stock market prices using Support Vector Machines (SVM).

Recent research (Kim, 2003) has examined the feasibility of applying Support Vector Machines to forecast stock price prediction of Korea composite stock price index by finding optimal values for the parameters used. In addition, the results were compared with back- propagation neural networks and case-based reasoning. The experimental results obtained showed that SVM provides a promising alternative to stock prediction because it implements the structural risk minimization principle which leads to better generalization.

A study (Kumar & M., 2006) was carried out to predict daily closing prices for the S&P CNX NIFTY Index by using classification techniques such as linear discriminant analysis, logit model, artificial neural network, random forest and SVM. In order to obtain the best SVM model, a variety of trials were done using different types of kernel parameters and constants. Then they focused on comparing the SVM model and random forest to forecast the daily movement of the index and compared the results obtained with the traditional models, logit models and artificial neural networks. The results obtained showed that SVM outperformed random forest, neural network and other traditional models.

A study (Samsudin et al., 2010) has been conducted to examine the flexibility of SVM in forecasting time series by comparing it with Artificial Neural Networks. The parameters

of SVM were determined by the grid search technique using tenfold cross validation. The experimental results obtained using the Root Mean Squared Error and Mean Absolute Error indicated that Support Vector Machines outperformed Artificial Neural Network.

3. Methodology

This section elaborates on the methodology used towards making stock price predictions using the libraries offered in Python programming language.

3.1. Data

The data used comprises of stock market data of Exxon Mobil Corporation from 01-01-2010 to 01-01-2020, extracted from Yahoo Finance (Yahoo Finance, 2021). Exxon Mobil, a leading petroleum and petrochemical enterprise in the world, is selected since it trades in fossil fuels that are important in determining operations for many companies and economies. The dataset has 3649 rows and 6 columns, where each row represents the trading information of a single day, and the columns represent the highest, lowest, opening, closing, and adjusted closing prices (high, low, open, close and adjusted close) and volume of the Exxon Mobil stock traded on that day. The date has been stored as an index.

3.2. Data Preprocessing

In a 365 days calendar, the stock market operates for around 252 trading days on average. However, as a conventional machine learning algorithm, SVM requires data for all 365 days for the full 10 years period of analysis. Thereby, two data filling methods have been used to populate the dataset.

One technique of missing value imputation, is forward filling, that uses the last known data point before the gap of missing values for the data filling task (Moahmed et al., 2014). The second technique is linear interpolation, that uses a straight line that passes through the last known data point before the gap of missing values and the first known data point after the gap, to impute the missing values (Lepot et al., 2017). After applying these techniques, the sliding window method was used on stock data which converts a time series to a supervised learning problem. In this mechanism, the independent variables comprise of data pertaining to previous time steps and the dependent variable comprises data pertaining to the next time step (Chou & Nguyen, 2018).

3.3. Variables

The overall dependent variable for all models is considered to be the adjusted closing price one day forward. The independent variables are the previous day's variables that are been used as inputs in seven different variable combinations, in order to capture the effect of such variables on the next day's stock price. The variable combinations used can be listed as follows;

- Univariate models
    - Adjusted close price
    - Volume
- Bivariate model
    - Adjusted close price and volume
- Multivariate model
    - Adjusted close, close, open prices and volume

- Open, close, high and low prices
- Adjusted close, high, low prices and volume
- Adjusted close, close, open, high, low prices and volume

3.4. Methods

When using functions or kernels to train the model, features having a large range will dictate the model by providing abnormal weights. This makes SVM unable to properly train the model. Using inbuilt packages (scikit-learn library in python), feature scaling is done to bring all features to a similar scale. In SVM, the regularization parameter C and the parameter γ that defines how far the influence of a single training example reaches, are important to reduce the misclassification pertaining to training examples. Hyper parameter optimization methods such as grid search and random search are used to determine the value for these parameters, γ and C. These methods use the cross-validation score to determine the suitability of the model with C parameter values ranging from 0.01 to 1000 and for γ, 0.001 to 10. When the cross-validation score nears 1 (best score), the parameters selected for the SVM model provides the best possible result. The grid search will create a grid with given C and γ values and then select the combination of parameters that will give the best score. In random search, the values for γ and C are generated randomly from the given range and then the parameter combination with the best score is selected.

The RBF or Gaussian kernel is regularly used due to its ability to nonlinearly map samples into a higher dimensional space. A linear kernel can be considered as a special case of RBF (Keerthi & Lin, 2003) since the linear kernel with a penalty parameter C, has the performance similar to the RBF kernel with some parameter. However, since the linear kernel is a heavily regularized RBF kernel, adding a penalty parameter can hinder the prediction accuracy. The sigmoid kernel also behaves like RBF for certain parameters (Lin & Lin, 2003). The polynomial kernel has a higher number of parameters that makes the model more complex (Hsu et al., 2003). As a result, the RBF kernel was selected in this comparative study.

3.5. Evaluation Measures

Root Mean Squared Error (RMSE) has been used for evaluating the models.

4. Results/Analysis and Discussion

Following table represents the RMSE calculated for all the models.

Table 1: Performance measures of all the variable selections, models, and data filling methods

Independent Variables	Hyper parameter optimization	Data filling method					
		Forward filling			Linear interpolation		
		C	γ	RMSE	C	γ	RMSE
Adjusted Close	Grid search	100.00	0.001	0.6055	100.00	0.001	0.5480
	Random search	52.36	5.265	0.6057	752.68	1.298	0.5504
Volume	Grid search	100.00	0.010	6.0671	100.00	1.000	5.6925
	Random search	681.59	0.648	5.9580	897.78	0.205	5.8340
Adjusted Close, Volume	Grid search	100.00	0.001	0.6075	100.00	0.001	0.5676
	Random search	208.75	0.715	0.6090	781.77	1.280	0.6086
Adjusted Close, Close, Open, Volume	Grid search	1000.00	0.001	0.6060	1000.00	0.001	0.5475
	Random search	516.08	0.622	0.6892	719.97	3.163	0.9035
	Grid search	1000.00	1.000	4.9894	1000.00	1.000	4.9431
Open, Close, High, Low	Random search	812.36	0.964	5.0064	605.93	0.374	4.9708
Adjusted Close, High, Low, Volume	Grid search	1000.00	0.001	0.6054	1000.00	0.001	0.5481
	Random search	397.93	1.970	0.8403	978.65	1.638	0.7432
Adjusted Close, Close, Open, High Low, Volume	Grid search	1000.00	0.001	0.6053	1000.00	0.001	0.5463
	Random search	692.27	0.591	0.7301	110.60	0.374	0.5637

Source: Author constructed

Above results indicate that all variable combinations provide a generally lower RMSE when the missing values are imputed through linear interpolation rather than forward filling. However, the model with random search for adjusted close, close, open and volume variables with interpolation, creates a RMSE higher than the forward filling and can be considered as an anomaly. This anomaly can be caused by how random search assigns C and γ parameters. It assigns values randomly and then selects the best score. This can lead to the parameters selected by random search not being the best parameter combination, so that the RMSE can be higher than expected.

Adjusted close is the closing price after adjusting for all relevant stock splits and dividends ("What is the adjusted close?" 2021). All the independent variables combinations with adjusted closing price are able to produce a RMSE less than one while the majority of such models maintains a value between 0.5 and 0.8. The variable combinations that do not take adjusted closing price as an independent variable, have RMSE values between 4.9 and 6, and this is observed even in the models that only uses a combination of stock prices such as open, close, high, and low prices. This is evident from Figure 1 that compares the actual and predicted prices from variable combination of open, high, low, close prices and combination of volume, adjusted close, high, low prices as shown in charts from left to right respectively. The highest RMSE is recorded in the univariate models that uses volume as the independent variable. Thereby, adjusted close and volume can be considered as the most significant and least significant predictors of future stock prices respectively. Since we are attempting to predict the adjusted closing price one day forward using the lagged adjusted closing price, the variables in the model are highly correlated and as a result, models with independent variable combinations including the adjusted close variable, will definitely record lower RMSE.

However, when considering the other variable combinations excluding the adjusted close, the RMSE values are found to be lower if more variables are added to the variable combination, regardless of the data filling method. This is proven by the variable combination of high, low, open, and close prices, having lower RMSE in all instances, when compared with the RMSE obtained by taking only volume as the independent variable. This lower RMSE is due to the additional features in the above independent variables. Thereby, as the features can be added without increasing the RMSE, the SVM model used can be considered to be not suffering from overfitting.

When comparing the hyper parameter optimization methods, random search seems to record higher RMSE values for the majority of the models, regardless of the variable combination and data filling method.

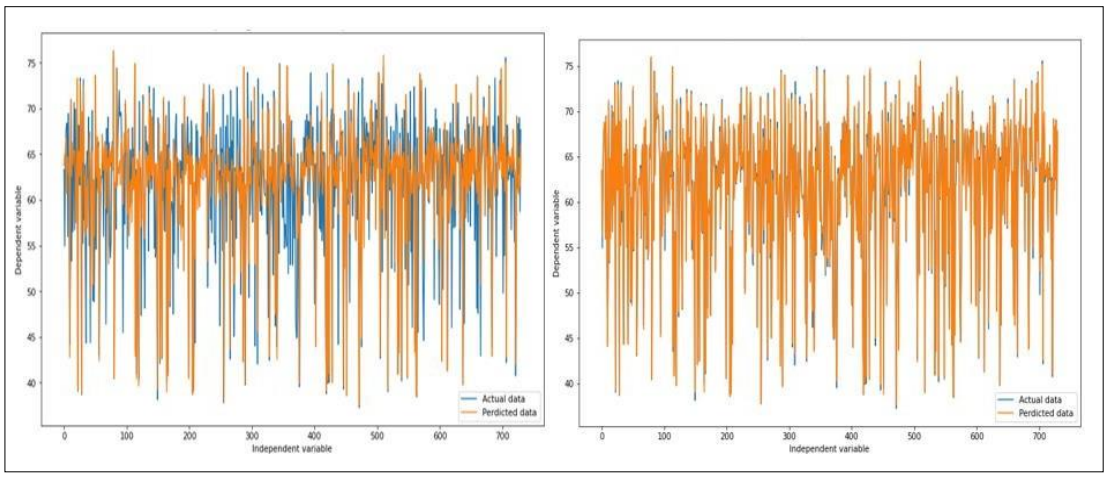

Figure 1. Open, high, low, close variables (left) and adjusted close, high, low, volume variables (right) for interpolation approach

The random search gives higher RMSE values because while the grid search selects the best parameter combinations from the given values aiming at minimizing the error, the random search only produces the parameters randomly and thus not proven to be the

best parameter combination like grid search. This results in higher RMSE values, even though random search is much faster than other optimization methods.

5. Conclusion and Implications

This paper aims at conducting a comparative analysis between different models that uses SVM to predict the stock prices, with the objective of studying how well SVM can be used to obtain accurate predictions. The paper discusses forward filling and linear interpolation as data filling methods, random search and grid search as hyper parameter optimization methods and various variable combinations, in order to identify the best data filling, hyper parameter optimization and variable combination that will give the most accurate stock price prediction using SVM.

As per the results, SVM performed well for predicting stock prices when the data filling method of linear interpolation, hyper parameter optimization method of grid search and independent variable combinations with adjusted close price, are used, due to its lower RMSE values recorded. Additionally, to improve the comparison between these models, certain steps can be taken in the future, where the analysis can be extended by adding more variables that impact stock prices, such as Gross Domestic Product (GDP) growth, corporate tax rates, interest rates, etc., without limiting to the data available on the Yahoo API. Further, hybrid machine learning models and neural networks such as Long Short Term Memory (LSTM) can be introduced to widen the analysis.

References

Chou, J., & Nguyen, T. (2018). Forward Forecast of Stock Price Using Sliding-Window Metaheuristic-Optimized Machine-Learning Regression. IEEE Transactions On Industrial Informatics, 14(7), 3132-3142. https://doi.org/10.1109/tii.2018.2794389

Hsu, C., Chang, C., & Lin, C. (2003). A Practical Guide to Support Vector Classification. Technical Report, Department of Computer Science and Information Engineering, University of National Taiwan, Taipei,1-12. https://www.scirp.org/referenceid=39118

Kim, K. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55(1-2), 307-319. https://doi.org/10.1016/s0925-2312(03)00372-2

Kumar, M., & M., T. (2006). Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.876544

Keerthi, S., & Lin, C. (2003). Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel.Neural Computation, 15(7),1667-1689. https://doi.org/10.1162/089976603321891855

Lepot, M., Aubin, J., & Clemens, F. (2017). Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment. Water, 9(10), 796. https://doi.org/10.3390/w9100796

Lin, H., & Lin, C. (2003). A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods. Technical report, Department of Computer Science, National Taiwan University. https://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf

Moahmed, T., El Gayar, N., & Atiya, A. (2014). Forward and Backward Forecasting Ensembles for the Estimation of Time Series Missing Data. Advanced Information Systems Engineering, 93-104. https://doi.org/10.1007/978-3-319-11656-3_9

Pagolu, V., Reddy, K., Panda, G., & Majhi, B. (2016). Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. International Conference on Signal Processing, Communication, Power and Embedded System, 1345–1350.

Samsudin, R., Shabri, A., & Saad, P. (2010). A Comparison of Time Series Forecasting using Support Vector Machine and Artificial Neural Network Model. Journal Of Applied Sciences, 10(11), 950-958. https://doi.org/10.3923/jas.2010.950.958

Vijh, M., Chandola, D., Tikkiwal, V., & Kumar, A. (2020). Stock Closing Price Prediction using Machine Learning Techniques. Procedia Computer Science, 167, 599-606. https://doi.org/10.1016/j.procs.2020.03.326

What is the adjusted close? Help.yahoo.com. (2021). https://help.yahoo.com/

Yahoo is now a part of Verizon Media. Finance.yahoo.com. (2021). https://finance.yahoo.com/.