TRADITIONAL AND MACHINE LEARNING-BASED METHODS FOR FINANCIAL INSTRUMENT PRICE FORECASTING: A THEORETICAL APPROACH

Justinas Lape
"UVS Group", Savanoriu ave. 123A, LT-03150, Vilnius, Lithuania,
justinas.lape@uvs.lt

Inga Zilinskiene
Mykolas Romeris university, Ateities st. 20, LT-08303 Vilnius, Lithuania,
inga.zilinskiene@mruni.eu

Saulius Preidys
Vilnius University, Saulėtekio ave. 9, I house, 10222 Vilnius, Lithuania,
saulius.preidys@kf.vu.lt

Abstract. Financial markets are one of the main components of the economy, and their growth and development is a crucial and significant factor in the world. Meanwhile, artificial intelligence is an exponentially developing field. The use of artificial intelligence in financial markets is a new and intensely developing phenomenon, requiring extensive research. The aim of this paper is to present a methodology of machine learning-based method effective application in financial instrument price forecasting in comparison with the traditional method, namely ARIMA. Consequently, the scientific literature on financial markets, traditional and machine learning-based methods were analyzed. Finally, a theoretical model for stock prices forecasting is presented. The main results of the analysis show the most extensively techniques applied in the stock markets and cash markets are methods of time series analysis, econometrics and machine learning. After analysis of methods of machine learning, it can be found most popular supervised learning algorithms are linear regression, decision trees/regression tree, random forest classification/ regression, and support vector machines. As a result of the research, a theoretical model for financial instrument price forecasting is presented and discussed. Based on the theoretical approach proposed several experiments will be performed using ARIMA and Support Vector Regression (SVR) methods.

Keywords: financial instruments, forecasting, ARIMA, artificial intelligence, machine learning, SVR.

Introduction.

Financial markets are one of the main components of the economy, and their growth and development is a crucial and significant factor in the world. Meanwhile, artificial intelligence is an exponentially developing field. The use of artificial intelligence in financial markets is a new and intensely developing phenomenon because of “<… both supply factors, such as technological advances and the availability of financial sector data and infrastructure, and by demand factors, such as profitability needs, competition with other firms, and the demands of financial regulation.” (Financial Stability Board, 2017).

Financial market is a broad term describing a marketplace where buyers and sellers are involved in exchange of such assets as shares, bonds, derivatives, currencies and other tradable financial instruments. Investors have access to many financial markets and exchange offices where different types of instruments can be traded. There are different sizes of financial markets across the globe, one of which has only a few participants, while other daily turnover amounts to trillions of euros. Financial markets differ not only in their size but in their functionalities and in the possibility of trading different risk financial instruments. Also, the financial market can be called a complex system, because it is influenced by many of the economic, political and psychological factors involved. Knowing when and how to invest in financial instruments is a complex decision to be taken by the investor. Trading financial instruments requires knowledge, deep market analysis and extensive experience.

Both investors and scientists are trying to predict the stock markets. This is an interesting and challenging area of research for the academic world, while the investor's main incentive is profit. Nevertheless, not everyone manages to cope with this task successfully. The successful pricing and trading of financial instruments is aggravated by factors such as the constant price volatility of financial instruments due to various economic, political factors and sentiments of participants operating in the markets. The price volatility of financial instruments means that the price may fall below the cost of the purchased instrument. A trivial solution that would lead to a successful investment would be the sale of a financial instrument before it was struck by its value or, ideally, by the price rise. For this reason, a variety of approaches are designed to create the most accurate model of the price forecasting of the financial instrument.

There is a wide range of studies found in the literature, which apply traditional forecasting techniques. In this regard, it can be noted that still a large part of the focus is given by applying traditional forecasting methods. On the other hand, research on the application of new types of methods is increasing significantly in scientific databases. A new trend is the development of artificial intelligence-based approaches, such as the application of machine learning algorithms, in developing predictive price models for financial instruments.

The purpose of the scholarly analysis – to present a methodology of machine learning-based method effective application in financial instrument price forecasting in comparison of the traditional method, namely ARIMA.

Research object – traditional and machine learning-based methods for financial instrument price forecasting.

Applied methods – scientific literature analysis, mathematical modeling.

The remainder of this paper is organized as follows: Section I reviews the concepts of financial markets, financial instruments; methods for financial instrument price forecasting; Section II describes the theoretical approach for stock prices forecasting; Section III presents conclusions.

Background

1.1. The Concepts of Financial Markets and Financial Instruments

The financial market is defined as the market in which the trading of financial instruments takes place (Fabozzi, Modigliani and Jones, 2014). Other researchers define the financial markets as a place where various types of financial instruments can be sold and/or exchanged by different entities based on their price, which is influenced by the supply and demand prevailing on the market for change. (Financial Stability Board, 2017)

Financial markets use different types of financial instruments. All financial instruments belong to a particular type of financial market (Kidwell, Blackwell and Whidbee (2011), Levinson (2014), Fabozzi, Modigiliani and Jones (2014), etc.). Different classifications of financial markets can be found as well: e.g., Kidwell, Blackwell and Whidbee (2011) differentiate financial markets to primary and secondary, organized and unorganized (over the counter), institutional and domestic markets. In the secondary financial market, financial instruments such as shares, bonds, futures and forwards, options, currencies, loan securities and other, newly derived instruments are traded. (Parameswaran, 2011) This article describes the methodology intended to forecast only the prices of financial instruments for which various mathematics-based analyses and forecasts are used, i.e. shares, options, futures, and currency pairs. In the stock and currency markets usually time series analysis, econometrics is used. Recently testing of the efficiency and benefits of machine learning algorithms has been researched (Dymova, Sevastianiov and Bartosiewicz, 2016).

It is important to underline that the simultaneous application of different approaches to the analysis can only enrich and improve the accuracy of the forecast – the latter factor can be identified as one of the limitations of this work since the created methodology of the forecast will not incorporate hybrid methods.

1.2. Methods for financial instrument price forecasting

Scientists Akansu and Torun (2015) distinguish three main trading strategies of financial instruments – fundamental, technical and quantitative analysis methods. Nazario, Silva, the Obobir and Kimura (2015) argue that the fundamental analysis relies on various economic factors to determine the real price of the securities, while technical analysis relies on the historical cost of the security and volume data. The fundamental analysis assesses the finances of businesses – cash flow, income, business performance, credit risk and other factors that allow to draw conclusions about the company's value. In contrast to the fundamental analysis, the technical analysis focuses on the changes in the price of the financial instrument. The latter method uses a wide range of visualization techniques (trends, resistance and support lines, etc.) to decide when to purchase and when to sell the financial instrument. Meanwhile, the most recent branch is called quantitative analysis is about studying and developing complex mathematical models and using statistical methods of analysis.

Both the technical and the quantitative analysis of financial instruments use mathematical and statistical methods to help investors detect the most optimal moment for opening or closing a position. In recent times, artificial intelligence methods as machine learning, deep learning, neural networks are applied more often with quantitative analysis.

The quantitative trading is any form of trading that uses sophisticated algorithms (programmable systems) to automate all or multiple trading cycles. It also includes encoding of the rules that the computer will have to perform and carry out backward testing or forward testing (Treleaven, Galas and Lalchand, 2013).

Quantitative trading can be divided into two parts: creating mathematical models for analysis and forecasting and developing an automated trading system by programming. Only the development of specifically applied math-based methods is retained in this work. The quantitative analysis relies to a large extent on statistics and the application of math-based methods.

Due to the fact that the financial data is a time-series data, a quantitative analysis can be called a time series analysis. Time series analysis is based on data analysis to find the optimal model that fits the given data in the best possible way. The primary purpose of the time series analysis is to find a model that can be successfully extrapolated to future data. Time series analysis is widely used for non-stationary data, namely this type of data and is generated in financial markets.

Time series forecasting methods are divided into two groups: based on statistical concepts and based on computer-intelligence techniques such as machine learning, neural networks, or genetic algorithms. The most popular statistical predicting methods for the time series are exponential alignment methods, regression methods, autoregressive integrated moving average methods (ARIMA), threshold methods, generalized autoregressive conditional heteroskedasticity or autoregressive conditional heteroskedasticity methods (GARCH/ARCH).

In the light of the studies carried out and on the basis of the reviews found in scientific sources, it can be concluded that, in most cases, research papers or studies on the analysis of data from financial instruments are seen in the context of machine learning algorithms (Dingli and Fournier, 2017): logistical regression; Random forests (RF); Support vector machines (SVM); K-Nearest Neighbor (kNN).

The traditional statistical ARIMA method generates the relevant predicted price values for the financial instrument. For this reason only the regression methods of supervised learning are suitable for verifying the effectiveness of the two methods of forecasting, since the classification methods distribute the predicted data into categories, which means that the latter methods are more appropriate to identify the tendency – the price trend of a financial instrument is at rising or at fall. Literature review reveals machine learning techniques commonly used to predict share prices, mainly: Neural networks, support vector machines are among the most commonly used methods. Both neural networks and support vector machines are standard machine learning techniques that can be used to predict time series data due to their specifics (Meesad and Rasel, 2013). Support vector machines is a machine learning algorithm falling into the category of supervised learning. The algorithm can be used to execute the price forecast of financial instruments. On the basis of the studies found in the scientific databases, it can be argued that the SVM algorithm is used in most cases, or the SVM algorithm is combined with other algorithms. Also, the SVM algorithm is applied not only in the financial markets but also in other areas due to its versatility. The popularity and applicability can be justified by the high aggregate performance of the method and by a mathematically well-prepared method of algorithm training (Liu and Duan, 2018).

In the context of the financial market, the application of the SVM algorithm is promising for two main reasons:

1. The algorithm does not apply any categorical assumptions to the data you are working with;
2. Characterized by the ability to minimize data over-fitting.

In conclusion, the algorithm is extremely sophistical, capable of working with non-linear data, while it also can be used as a regression method, with the help of which real values are predicted, so the methodology presented will be based on the application of SVR algorithm for financial instrument price forecasting.

II. A methodology to stock prices forecasting

In order to be able to determine if machine learning methods are effective in financial instrument price forecasting a methodology of research is needed. The two methods are selected on the basis of scientific analysis. First one is traditional statistical method ARIMA, the second one is a machine learning-based, support vector regression (SVR) algorithm. Each method has its own specific course of application, which is presented below.

2.1 The Methodology

The subsequent methodology (Fig. 1) will be used in the future to determine whether the results obtained by the SVR method are more accurate than the results using the ARIMA method.

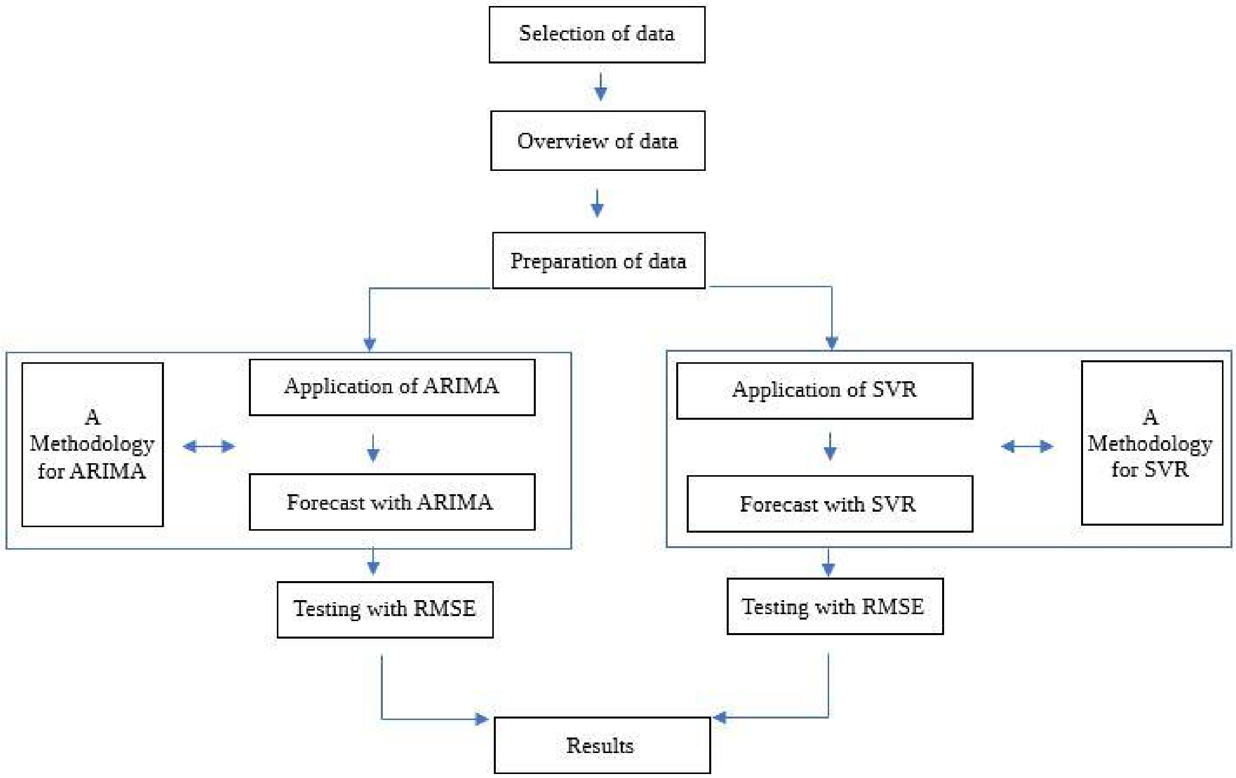

Fig. 1. A methodology for financial instrument price forecasting

Data selection.

Due to the availability of data and the specificity of the financial instrument, shares will be selected as a tool and the data provided by them will be used to construct and study the models. The same datasets will be used to compare ARIMA and SVR models.

Overview of selected data.

When forecasting models are created, it is necessary to divide the dataset into two groups – one of the groups being defined as a training set, another as a testing set (Nayak, Pai and Pai, 2016). The training data set is designed to train models and the testing data set is designed to evaluate the predictive capabilities of a trained model. In the future experiments, predictive models are verified by the principle of the back-testing. The non-return test will be carried out using historical price data. The purpose of the back-testing is to determine the effectiveness of the forecast with an assumption that if the prediction has been successful in the past, it is likely to continue to perform well in the future (Bailey et al., 2014).

Testing methods.

The RMSE (Root Mean Square Error) indicator will be used to measure the accuracy of the forecasts. The RMSE indicator is the standard deviation of residues. RMSE is used in situations where there is a comparison of forecasts of several different methods for the same datasets based on RMSE metering, the lower the error, the better the forecasting performance. In this case, the residue is a dimension that shows how far is the real value of the data from the regression line (Chai and Draxler, 2014).

Application of ARIMA.

ARIMA method shall be used for both the training and testing data sets to achieve the most accurate results. The forecast is for the fixed period, using a one-step approach.

Based on the Box-Jenkins methodology, the ARIMA process covers a total of three essential steps – identification, evaluation, diagnosis. The main goal is to define ARIMA (p, d, q) parameters because only the successful selection of parameters can lead to the reliability of the prediction of the method. Identification involves checking time series stationarity. During this phase, Dickey-Fuller test or/and ACF and PACF analysis are usually employed to identify stationarity. Estimation phase involves estimation of the parameters – information criteria such as AIC, AICc and BIC are used. Diagnosis phase involves determining whether the model is adequate – Ljung-Box test or ACF/PACF analysis are performed to check if there are any remaining correlation between residuals. If a correlation between the residual values is recorded, this means that the model can still be adjusted by the GARCH/ARCH methodology.

Application of SVR.

SVR method is categorized into several types and parameters. SVR method development includes three major steps: method type has to be selected, kernel function has to be chosen and parameters has to be set. SVR model predictions are heavily influenced not only by the type and kernel function but also by the choice of parameters. R software *e1071* package is limited to the following types of SVR: C-classification, NU-classification, ONE-classification and EPS-regression. Model type is usually selected considering the purpose of the modelling. EPS-regression type is the most relevant because this type is for forecasting real values. In addition radiant kernel function is found to be the most accurate for time series data. The parameters of "gamma", "epsilon", "cost" usually are determined in two ways:

1. The first method is error-testing – Multiple variants of all parameter values are tested and determined which combination of parameters provides the best results;

2. The second method is using tune.svm function from the *e1071* package to check the suitability of the selected parameters, which determines the most efficient combination of parameters in the 10-level cross-validation method. Cross-validation is a statistical method for determining the performance of a machine learning algorithm, in which the data set is randomly selected for group data on which the model is and is trained and adaptable.

The parameters selected in both ways will be checked and the most efficient model will be selected.

Comparison of forecasts using both methods.

In this part, the results obtained will be compared and conclusions drawn. After the forecast parameters have been completed, the RMSE indicators will be calculated and based on them a comparison of the effectiveness of the methods will be concluded.

Implementation environment

For the analysis, open source data analysis and programming software R will be used. *Forecast* package will be used for ARIMA modelling, *e1071* package will be used for SVR modelling.

2.2. Assumptions and limitations

1. The Efficient Market Hypothesis (EMH). According to the EMH, stocks always trade at their fair value on stock exchanges, making it impossible for investors to either purchase undervalued stocks or sell stocks for inflated prices.
2. Evaluation of the performance of the forecast is done through historical data. The model is not tested in real-time. It is assumed that if the forecast worked well on historical data, it will work well in real-time.
3. The study aims to predict the prices of financial instruments rather than develop a trading strategy and calculate profitability and/or returns.
4. Software specifics – it is likely that by developing models and forecasting using other software platforms such as Python, MATLAB, etc., there may be an error in the results.
5. The work-study is limited to the application of a traditional statistical method and a single machine learning approach. The application of multiple methods may affect the final results.
6. Forecasting methods are created solely on the basis of the historical cost of shares. The scope does not include additional data sources that are likely to increase the accuracy of the forecast.
7. Both models are adapted for a specified period. In the next period (longer/shorter), the results of models are likely to vary as well.
8. ARIMA and AVR methods can be parameterized differently. For this reason, the final forecast results may vary.

Conclusions

The forecasting of stock prices is highly researched, but most of the research is concentrated on the application of traditional methodologies, such as Box-Jenkins or technical analysis. Nevertheless, the application of artificial intelligence techniques are already being explored on the market. To analyze whether the application of machine learning-based approaches in the forecast of share prices can actually have tangible benefits, a literature analysis is done. A methodology based on the traditional ARIMA method, and the support vector regression method is proposed as well. The main findings are as follows:

1. Following the theoretical analysis of financial markets, it was found that the types of analysis in the financial markets are: fundamental analysis, technical analysis and quantitative analysis. The quantitative analysis consists of two parts: the development of mathematics-based methods and the automation of trade. Mathematics-based methods are mainly used for the analysis and forecasting of shares, options and futures, as well as currency pairs. Time-Series analysis, econometrics, and machine learning techniques have been found to be most intensely applied in the stock and currency markets.

2. As a result of the literature analysis, data generated by the financial markets were found to be time-series data. The most commonly used methods of analysis of the time series are the methods of ARIMA class. Machine learning is one of the main subgroups of artificial intelligence, the main purpose of which is to find the templates in the data, summarize them, learn without programming. Given that the financial timeline data is dynamic, non-linear and highly volatile, machine learning is the appropriate tool for analyzing the latter type of data. Machine learning algorithms are classified in the classes of supervised learning, unsupervised learning, enhanced learning and deep learning. However, the most commonly used algorithms for the price forecast of financial instruments are those falling under the class of supervised learning. Not limited to this, supervised learning algorithms are also categorized into classification and regression algorithms with different and specific requirements at the beginning and different results at the end of the application.

3. Based on the scientific analysis, a theoretical approach, how to determine the effectiveness of machine learning method for financial instruments prices forecasting were presented. Several experiments will be performed following the methodology proposed using ARIMA and SVR methods.

References

Akansu, A. N., & Torun, M. U. (2015). *A Primer for Financial Engineering: Financial Signal Processing and Electronic Trading*. Cambridge, MA: Academic Press.

Bailey, D. H., Ger, S., Lopez de Prado, M., Sim, A., & Wu, K. (2014). Statistical Overfitting and Backtest Performance. *SSRN Electronic Journal*. doi:10.2139/ssrn.2507040

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Online: https://www.geosci-model-dev-discuss.net/7/C473/2014/gmdd-7-C473-2014-supplement.pdf

Dingli, A., & Fournier, K. S. (2017). Financial Time Series Forecasting – A Deep Learning Approach. *International Journal of Machine Learning and Computing*, *7*(5), 118-122. doi:10.18178/ijmlc.2017.7.5.632

Dymova, L., Sevastianov, P., & Bartosiewicz, P. (2010). A new approach to the rule-base evidential reasoning: Stock trading expert system application. *Expert Systems with Applications*, *37*(8), 5564-5576. doi:10.1016/j.eswa.2010.02.056

Fabozzi, J. F., Modigliani, F. P., & Jones, F. J. (2014). *Foundations of Financial Markets and Institutions*. New York, NY: Pearson Higher Ed.

Farias Nazário, R. T., E Silva, J. L., Sobreiro, V. A., & Kimura, H. (2017). A literature review of technical analysis on stock markets. *The Quarterly Review of Economics and Finance*, *66*, 115-126. doi:10.1016/j.qref.2017.01.014

Financial Stability Board. (2017). Artificial Intelligence and machine learning in financial services. Market developments and financial stability implications. http://www.fsb.org/wp-content/uploads/P011117.pdf

Kidwell, D. S., Blackwell, D. W., & Whidbee, D. A. (2011). *Financial Institutions, Markets, and Money, 11th Edition*. Hoboken, NJ: Wiley Global Education

Levinson, M. (2014). *The Economist Guide To Financial Markets 6th Edition*. London, England: Profile Books.

Liu, M., & Duan, C. (2018). A Review of Using Support Vector Machine Theory to Do Stock Forecasting. *Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)*. doi:10.2991/ncce-18.2018.184

Meesad, P., & Rasel, R. I. (2013). Predicting stock market price using support vector regression. *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*. doi:10.1109/iciev.2013.6572570

Nayak, A., Pai, M. M., & Pai, R. M. (2016). Prediction Models for Indian Stock Market. *Procedia Computer Science*, *89*, 441-449. doi:10.1016/j.procs.2016.06.096

Parameswaran, S. (2011). *Fundamentals of Financial Instruments: An Introduction to Stocks, Bonds, Foreign Exchange, and Derivatives*. Hoboken, NJ: John Wiley & Sons.

Treleaven, P., Galas, M., & Lalchand, V. (2013). Algorithmic trading review. *Communications of the ACM*, *56*(11), 76-85. doi:10.1145/2500117