ASSESSMENT OF MOVEMENT OF THE BELEX15 INDEX USING LINEAR REGRESSION AND NEURAL NETWORK

Marijana Petrović
University in Novi Sad, Faculty of Economics in Subotica, Serbia
marijana.petrovic@ict.edu.rs

Original Scientific Paper
doi:10.5937/joupроman7-22690

Abstract: The stock market remains a focus of attention for researchers. So far, various technical and statistical methods and models have been developed and used to show different results, but none with great success. In this paper the author used two approaches to estimate price movement in stock market: simple linear regression and backpropagation neural network.The aim of this paper is to estimate the value of the Belex15 stock exchange index and to analyze its movement using two aforementioned approaches.

Keywords: *Belex15, Linear regression, Neural Network.*

1. Introduction

Considering the large number of elements affecting the price movements in the stock market, assessment of future trends is a very difficult task. However, since the possibility of their accurate appraisal could lead to enormous benefits for investors, a number of researchers are engaged in precisely these estimates. Machine learning has greatly facilitated this job, as researchers today have a wide range of tools, techniques, models and software at their disposal, which are for the greater part freely available on the Internet.

In this paper, two models were used to estimate the value of the Belex15 stock exchange index: linear regression, as one of the most commonly used techniques for data mining and future values prediction of the stock exchange price movements based on their linear association, as well as the nonlinear approach using the neural network. The aim of this paper is to evaluate the adequacy of the selected method application, to examine whether there are differences if the prediction is made with minor or multiple input parameters (5 and 10 days), to compare approaches in order to identify possible differences in performance, and to examine whether and how the trend affects the forecast of the time series. Further work in Chapter 2 discuses some previous works on this subject. Chapter 3 describes the analysis of time series including trend function, through linear regression and neural network approaches, while in Chapter 4 these approaches are applied to open price of the stock exchange index Belex15 in order to estimate and compare its future movement. Chapter 5 gives final considerations.

2. Literature review

For years, the authors have been looking for the most appropriate model that could accurately predict stock price movements. Regression and neural networks are commonly used methods in this field.

(Ahangar et al. 2010) used the linear regression method and the General Regression Neural Network (GRNN) method to estimate stock price movements of companies in Tehran (Iran). Researchers have come to the conclusion that neural networks are faster than regression, and that they are more suitable for estimating movements in a chaotic environment such as stock trading. (Cao et al. 2005) used artificial neural networks (ANN) to predict stock price movements in the Shanghai (China) stock market. They compared the results obtained with the linear models used in the financial forecasting literature and concluded that neural networks outperform linear models in the achieved forecasting results. (Faria and Gonzalez 2009) used ANN and AESM (Adaptive Exponential Smoothing Method) approach to determine the movement of indices on the Brazilian stock market and to evaluate their ability to estimate stock returns. The results of the study showed that both methods yield similar results in predicting stock returns, but that neural networks outperform the AESM approach in estimating stock price movements. (Wilton et al. 2008) used the linear regression approach and the SBPNN (Standard Back Propagation Neural Network) to predict index price movements on the stock exchanges of the USA, Europe, China and Hong Kong, based on two-year historical data. The results of the study showed higher accuracy of neural network prediction. For markets with high price fluctuations (such as the Chinese market), the authors have shown that up-to-date data can help reduce forecasting errors. (Pokhriyal et al. 2011) applied the MLR (Multiple Linear Regression) and ANN approach to investment decision making, based on data from the Asian Stock Exchange. They concluded that MLR can be used as a simple tool to study the linear relationship between variables because it offers understandable explanations for the effects of various factors, but the problem is that this method does not prove to be the best when using realistic data to estimate stock price movements. ANN gives more accurate results that are not sensitive to data errors, and the method improves its performance with the large number of examples. The problem with ANN approach is that in certain cases the method does not provide explanations for parameter estimation, so MLR can provide necessary information to decision makers when buying / selling stocks. (Enke and Thawornwong 2005) examined the effectiveness of neural network models used to predict and classify stock price movements. The results showed that classification neural network models generate higher profits than other strategies considering the same risk. The paper shows that the model with the highest percentage of accurate prediction of price movements does not necessarily produce the highest profit, and that trading results based on several neural networks forecast can improve profitability more than Buy-and-hold strategies. Some authors (King 1999) and (Desai and Bharati 1998) have come to the conclusion that neural networks do not always supersede the MLR approach. To predict high earnings on large stocks, neural networks outperform MLR access in times of high volatility, while MLR shows better performance for low variance problems. Also, neural networks proved superior over larger data samples for analysis, while analyzing smaller data samples showed better results with MLR approach.

(Desai and Bharati 1998) have found that neural network forecasts are conditionally efficient when compared to the linear regression model in the case of large company stocks, whereas in the case of small company stocks the findings are not statistically significant.

3. Time Series Analysis

Time series represent the movement of a phenomenon at different levels over a certain period of time. The main goal of time series modeling is to carefully collect and study a chronologically ordered set of data of a particular phenomenon in order to develop an appropriate mathematical model that reveals the fundamental process of generating serial data, that is, describes the inherent structure of the series. Time-series matching procedure in the appropriate model is called Time series analysis. Predictions of future values can be made based on the developed model, and in order for them to be successful it is essential that the model be appropriate. When preparing projections, it is advisable to select the simplest model with as fewparameters as possible, in order to avoid overfitting with training data, which would show good results during the learning phase, but not in the future forecast (Adhikari and Agrawal 2013).

Within the time series there is a trend that represents a slow and gradual change in some properties of serial data over the entire time interval. Detecting and understanding the trend can enable faster modeling and more efficient selection and evaluation of the model. The process of detrending in prediction model includes mathematical or statistical operations which remove the trend from the series of data as to show only the absolute changes in values and to enable the identification of potential short-term cyclical patterns. Also, by removing the trend, we can simplify modeling and provide additional information to the model so as to improve its performance. Detrending process is done by using regression and other statistical techniques, and issometimes also applied as a preprocessing step phase.

Time series analysis using Linear Regression

Regression analysis is a type of statistical evaluation that allowsus to describe the relationships between independent and dependent variables, to estimate dependent variable values from the observed independent variable values and to identify the outcomes and individual forecasts of their movements (Schneider 2010). Linear regression is used to study the linear relationship between two or more phenomena, that is, between the dependent variable y and the independent variable x.

The parameters determining the linear regression are: a spread diagram, an equation, and a standard error. The spread diagram is a graphical representation of the original data on a scatterplot diagram which helps us ascertain whether there is a linear correlation (positive or negative) in two dimensions or not. The relationships are modeled using linear function based on a set of coefficients and values of an independent variable, which are then used to predict the outcome of a dependent variable. Basic form of the aforementioned linear function is : $y = \beta_0 + \beta_1 x$. In order to evaluate the accuracy of the linear model, we can use the RMSE (Root Mean Squared Error) or some version of the coefficient of determination - R2.

The coefficient of determination measures the percentage of theexplained variations of the dependent variable by a linear relation, according to the formula: $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y})}$. Often, the MAE (Mean Absolute Error) is used to evaluate the accuracy of the model for continuous variables in addition to RMSE metrics. RMSE represents the standard deviation of the residual linear regression and is calculated according to the formula: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. MAE measures the average magnitude of the errors in the set of predictions and shows the average absolute difference between the observed and predicted values: $MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$.The smaller the model, the better.

Linear regression is one of the most commonly used data mining techniques for predicting the future values of variables based on their linear association. It bases its prediction on the assumption that there is a real line that approximates a certain set of data (Olaniyi 2011).

Time series analysis using Neural Network

Neural networks are used in machine learning. They simulate the human brain function, that is, they imitate the intelligence used by the brain to recognize the regularities and patterns in the input data and learn through previous experience, which further enables them to generalize the results. The neural network consists of a network of interconnected computer units called neurons, where each of the connections has certain weight coefficients that affect the strength at which the input values will determine the output. Neurons in the network are organized in layers. The first, input layer, is made of input neurons which process the data that is entered into the network further into the system. The output layer consists of the output neurons that make predictions based on data from the input layer, and the hidden layer /layers are located between the aforementioned layers. Their job is to transform the input data so that the output layer can use them (Pathak 2014). The input values (which must be in the 0 to 1, or -1 to 1 interval) pass through the connections with different weight coefficients that automatically amplify or reduce them and sum up within the neuron into the total output. The total output is passed through the mathematical function to obtain an analogue output in the interval from 0 to 1 (or -1 to 1). The most commonly used functions are sigmoid function (Sigmoid) or hyperbolic tangent (Tanh). The Sigmoid function equals the equation: $f(x;a) = \frac{1}{1+e^{-ax}}, -\infty < x < \infty$, and Tanh function equals the equation: $f(x) = \frac{e^x - e^{-x}}{e^x - e^{-x}}, -\infty < x < \infty$. (Heaton 2012)

If the output differs from the required one, then the weight connections are adjusted to yield an output similar to the one desired. This is called the learning phase. After the learning phase, the test phase follows. In this phase new data will be passed through the model, which will produce predictions based on the learned templates. Research has shown that neural networks havegreat ability to identify patterns and make predictive analysis and moreover, have proven to be very useful for classification and regression.

Primitive methods of predictive analysis included models based on linear and logarithmic regression whose efficiency was dependent on the type of model, while neural networks that appeared at a later datewere less dependent on the model, and more on data based training (Pathak 2014). Even when input data are incomplete or irrelevant, the network can learn from those essential data traits (Hajizadeh et al., 2010).

4. Using linear regression and neural networks to predict the value of the Belex15 index

Forex market forecasts are used to identify the market trends and can also be used to plan investment strategies as well as to discern the best moment to purchase or sell a particular action. (Hajizadeh et al., 2010). As the stock market is influenced by a number of factors, such as the historical movement of prices, the current situation in the country, global economies, etc., forecasting price movements is not easy, but it did not discourage experts from continuously seeking the best prediction model. So far, various techniques and statistical methods and models have been developed and used to show different results, but none with great success, and this field remains the subject of study by many researchers (Pathak 2014).

In this paper we used a linear and nonlinear method for analyzing the accuracy of the model and predicting the movement of the open price of the Belex15 index. Belex15 is the leading index of the Belgrade Stock Exchange that represents the movements of prices for most liquid Serbian shares and can be monitored in real time.The first tool used is linear regression as one of the basic and often used tools for predictive analysis in this field. As a non-linear approach to the problem, neural networks wereused due to the fact that they can learn complex non-linear mappings, enabling them to solve advanced problems in prediction.

Prediction of Belex15 index movement using Linear Regression

In order to create a model and estimate the future trends of the index, the author used open price of the leading index of the Belgrade Stock Exchange (Belex15) in the period from 10.01.2014. until 28.12.2018. Since the correlations obtained by linear regression between dependent and independent variables can be used as templates in the prediction of unknown values, the set of data used for linear regression is divided into two parts: 80% of data was used to define the model training data and 20% of data was used to test the built-in model, i.e. to evaluate predictions based on test data. During the creation of the model, data were taken in the time window of 5 and 10 working days of the Belgrade Stock Exchange.

Using the LINEST function in Excel, the following linear regression equation is applied to data from a five-day time window (Model LR1):

$$y = 3.10098 - 0.02751x_1 - 0.01451x_2 - 0.03220x_3 + 0.07031x_4 + 0.99939x_5$$

In the same way, the linear regression equation is applied to data from a ten-day time window (Model LR2):

$$y = 2.72551 + 0.01028x_1 - 0.02000x_2 + 0.06141x_3 - 0.08049x_4 + 0.02968x_5 - 0.02922x_6 - 0.01469x_7 - 0.02934x_8 + 0.06288x_9 + 1.00555x_{10}$$

Using the Excel add-in for the financial analysis and defining of the predictive templates (Data Analysis add-in), and ANOVA F statistics, linear regression was applied to the training data and the results obtained are shown in the table below.

Table 1. Results of Data analysis add-in and ANOVA statistics applied to linear regression models

Statistics		*Model 1 LR*	*Model 2 LR*
Data analysis add-in	**Multiple R**	0.996519765	0.996518364
	R Square	0.993051642	0.993048849
	Adjusted R Square	0.993016585	0.992977701
	Standard Error	4.515957217	4.497739601
	Observations	997	988
ANOVA statistic	**F**	28326.525	13957.527
	Significance F	0.00	0.00

The value of the Multiple R is very close to 1 for data from both models, which shows a strong linear connection. The determination coefficient (R Square) shows the percentage of the variation of y values around the average explained by the x value, which in our case is 99%. The standard error estimation (Standard Error) shows that the accuracy of the regression coefficient measurement is slightly higher in Model LR1. The last item in the table (Observations) represents the number of data used in the regression analysis for building the model. F value in regression is the result of a test in which the null hypothesis states that all coefficients of regression are equal to zero, that is, the model has no predictive ability. The F-test compares the obtained model with zero predictor variables, based on which it determines whether the coefficients we added improved the model. If the obtained result is significant, it means that the model has been improved. F value is calculated as the middle of the square of the regression divided by the middle of the square of the residual. Given that the value obtained for the significance of F statistics is less than the predicted significance level of 0.05 and that in cases of both models R2 is over 0.99 we can say that the regression model is significant.

The evaluation of linear regression models resulted in the following outcomes for mean absolute error (MAE) and the root mean square error (RMSE):

Table 2. MAE and RMSE evaluation of linear regression models

		Training Data Error	*Test Data Error*
Model LR1	**MAE**	3.301	3.480
	RMSE	4.487	4.878
Model LR2	**MAE**	3.269	3.547
	RMSE	4.446	4.943

MAE measures the magnitude of the error in the forecast series, and RMSE measures the average magnitude of the error. MAE and RMSE movements in the case of Model LR2 show slightly less overall error and error in training data, while the same indicators applied to the test data showed a minor error in the case of a Model LR1. If we take into account the MAE and RMSE evaluations, we can conclude that more precise prediction is the one with a smaller time window, i.e. Model LR1.

Prediction of Belex15 index movement using Neural Network

Data Engine is a professional data analysis tool. It is a software product that is used to analyze data using conventional statistics, neural networks and "fuzzy" technologies. It is successfully applied in the field of forecasting, database marketing, quality control, process analysis and diagnostics. In this paper, in order to predict the movement of the Belex15 indices using the Data Engine tool, author selected Multilayer Perception type of a neural network whose classifier uses the back propagation technique to classify instances. In Model NN1, the author used five input and one output neuron, so that on the basis of the five-day movement of open prices we could predict the price of the next day. In network configuration the author chose to implement one hidden layer with five neurons and a Tanh activation function. The original data are normalized in the range from -1 to 1 to which the Single Step (Delta) learning strategy is applied. The learning rate is set to 0.001 and the momentum of 0.3 is used. As the terminal criterion, the number of epochs in training is taken, with testing for every 100 epochs. Model NN2 used the same network configuration, taking into account the ten-day time window, so the number of input neurons were adjusted to 10.

During the training of the model, the data in the time window of 5 and 10 working days of the Belgrade Stock Exchange were identical to those data over which linear regression was applied, making a total of 1,255 examples for learning in the first case (Model NN1) and 1250 examples in the other model (Model NN2). Thereof, 80% of the data was used for training the network, and 20% for its testing. To evaluate the forecasting trend model, the Data Engine calculates the square root error (RMSE). From the table below, we can conclude that in the NN1 model used for predicting, RMSE is smaller compared to the Model NN2, and that the model with a five-day time window has shown a greater accuracy of the prediction in both cases (in case of using neural networks and also in the case of linear regression).

Table 3. RMSE of NN1 and NN2 Models with and without trend

	Trended		***Detrended***	
	Training Data Error	***Test Data Error***	***Training Data Error***	***Test Data Error***
Model NN1	3.3226	3.8895	2.8588	2.8037
Model NN2	5.2480	114.9744	3.0695	3.0763

From the table above, we can conclude that in Model NN1 used for forecasting, RMS errors are smaller compared to Model NN2, and also that detrending data plays an important role in improving accuracy of the model. With 1000 epochs used in Model NN2, detrended model showed more than 37 times smaller error, than model with trend.

In the graph shown in Figure 1, we can see the graphic representation of the forecasted prices of the Belex15 index in case the trend from data is removed and in case it is not. From the presented we can see that not removing the trend causes a slightly larger error in the prediction. Using the same data sample over which the prediction of linear regression was made, data was detrended and a prediction with backpropagation neural network was performed. For better display, the following graph in Figure 2 shows the Belex15 index open price forecast in February 2018 (randomly selected month from test data)using all four models, and the graph shown in Figure 3 represents visual display of RMSE.

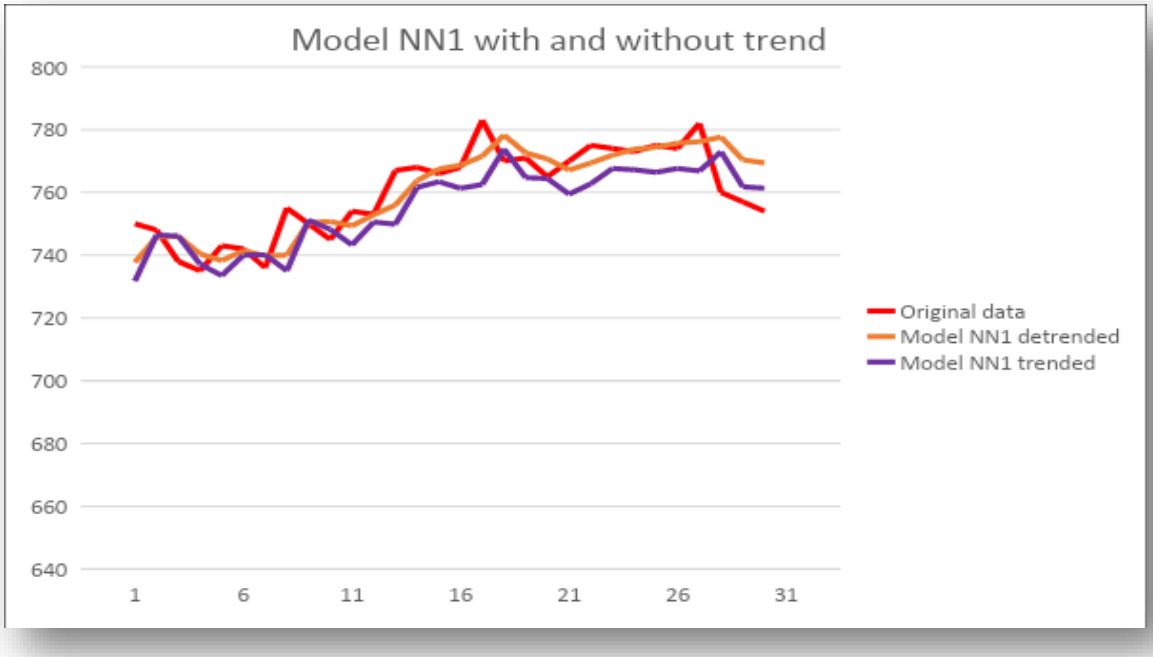

Figure 1. Model NN2 with and without trend

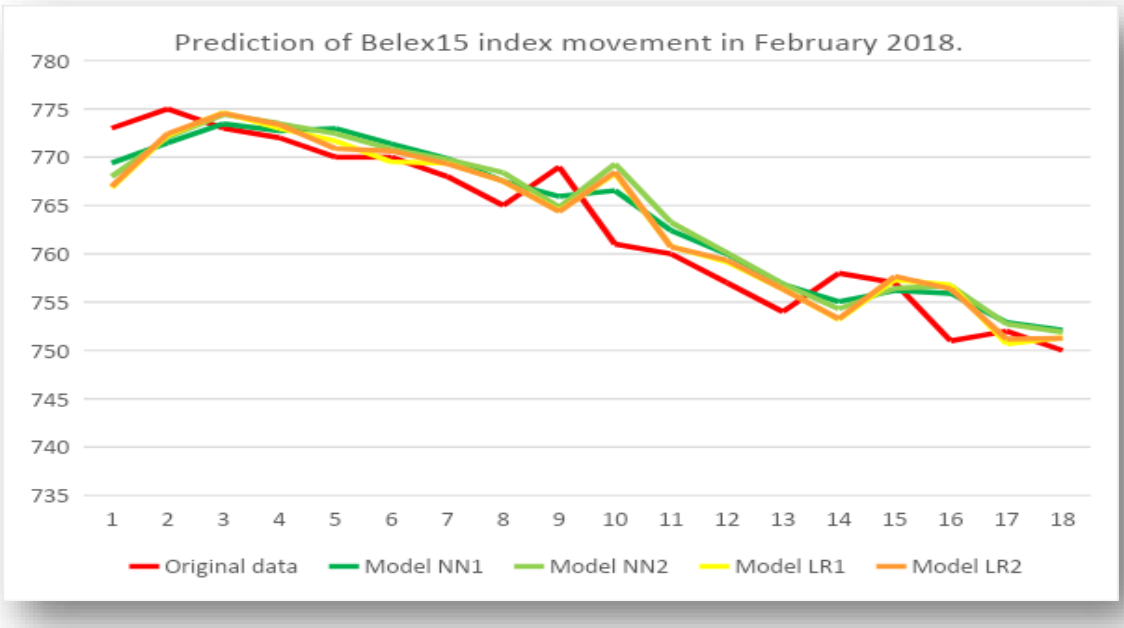

Figure 2. Prediction of Belex15 movement in February 2018. using all models

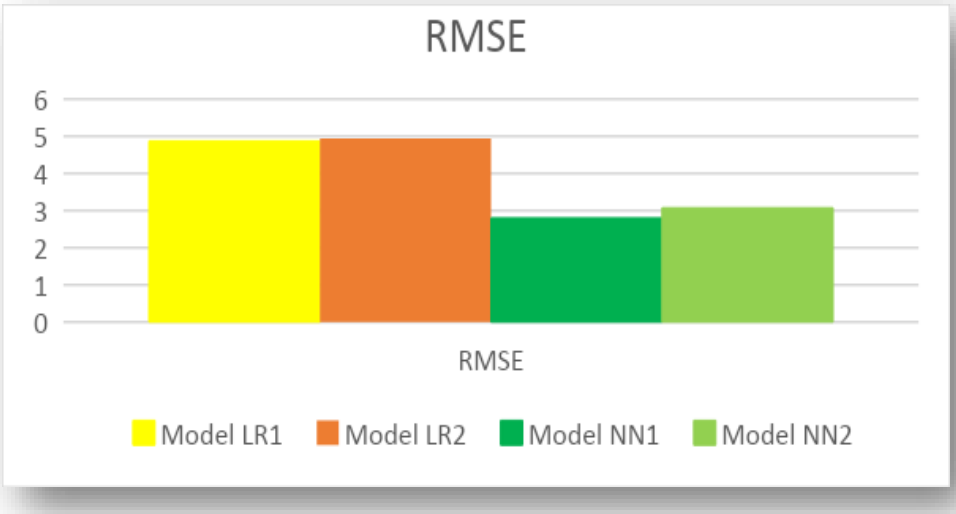

Figure 3. RMSE for all models

The graph shown in Figure 2 represents a segment of all data predictions in one month using different models. We can see that the prediction of open price Belex15 movement has slightly different movement when data predictions are made from data taken from different time windows. In case of this data segment, Model NN2 shows slightly better predictive results in Figure 2, but Model NN1 shows better overall result in predicting price movement, based on the RMSE of both models shown in Figure 3. Linear regression also shows good results in predicting Belex15 index movement, with Model LR1 showing slightly better results considering RMSE from Figure 3. Considering that the Model NN1 shows the smallest RMSE, we can say that it is most suitable for predicting the movement of the open price of the Belex15 index.

This paper shows generated results for the backpropagation neural network model chosen from offered forecast models in Data Engine. The network used stop parameter, which automatically calculates the number of epochs in the different cases. By optimizing the network, i.e. setting different parameters like: learning rate, momentum, terminal criterion, number of epochs, number of hidden layers and neurons etc., better results could be achieved. The further work of the author will be conducted in that direction.

5. Conclusion

The stock market remains a focus of attention for researchers. So far, various technical and statistical methods and models have been developed and used to show different results, but none with great success. In this paper we have used two approaches: simple linear regression and backpropagation neural network, and we have come to the conclusion that the model with the smallest RMSE, hence the best predicting power, is the backpropagation neural network model that used 5-day time window for training the model with detrended data.

Based on the evaluation, we can conclude that the mean squared error for predicted values is in a similar range for all four models, but due to the non-linearity of the approach, it can be assumed that the forecast of neural networks will be more accurate.

Research practice has shown that better results in value estimation are obtained by combining different models and methods as by doing this the shortcomings of some models can be overcome when estimating the future values. In addition to the mining of numerical data, text mining can also be used in supervised learning, in order to take into account economic changes recorded in the form of unstructured textual data in the financial statements and news. The author's future work will be centered around performing further research and combining different methods in order to achieve better prediction of stock prices movement.

Reference

Ahangar, R. G., Yahyazadehfar M, Pournaghshband H. (2010). The Comparison of Methods Artificial Neural Network with Linear Regression Using Specific Variables for Prediction Stock Price in Tehran Stock Exchange, *(IJCSIS) International Journal of Computer Science and Information Security*, Vol. 7, No. 2

Cao, Q., Leggio, K. B., Schniederjans, M. J., (2005) A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market, *Computers & Operations Research*, Volume32, pages 2499-2512

de Faria E.L., and Gonzalez, J.L., (2009). Predicting the Brazilian stock market through neural networks and adaptive exponential smoothing methods, *Expert Systems with Applications Article in Press*.

Wilton.W.T. Fok, IAENG, Vincent.W.L. Tam, Hon Ng, (2008) Computational Neural Network for Global Stock Indexes Prediction, *Proceedings of the World Congress on Engineering 2008* Vol II, WCE 2008, July 2 - 4, 2008, London, U.K.

Pokhriyal, A., Singh, L., Singh, S., (2011) Comparative Analysis of Impact of Various Global Stock Markets and Determinants on Indian Stock Market Performance - A Case Study Using Multiple Linear Regression and Neural Networks, *Conference Paper in Communications in Computer and Information Science*

Enke, D., Thawornwong S., (2005). The use of data mining and neural networks for forecasting stock market returns, *Expert Systems with Applications* 29 (2005) 927–940

Desai. V., Bharati, R., 1998, Annals of Operations, Research. 78, 1998t: 27-163.

King, S. L., Neural Networks Vs. Multiple Linear Regression for Estimating Previous Diameter,*12th central hardwood forest conference*;

Heaton, J., (2012). *Introduction to the Math of Neural Networks.* Heaton Research.

Heaton, J., (2015). *Artficial Intelligence for Humans. Vol. 3: Deep Learning and Neural Networks.* Heaton Research.

Maimon, O., Rokach, L. (2005). *Data mining and knowledge discovery handbook*. Springer.

Olaniyi, S. A., Adewole, S., Kayode, S., Jimoh, R.G. (2011). Stock Trend Prediction Using Regression Analysis – A Data Mining Approach. *ARPN Journal of Systems and Software*, I (4), 651-656.

Adhikari, R., Agrawal, R.K. (2013). An Introductory Study on Time Series Modeling and Forecasting. *ARPN Journal of Systems and Software*, 154-157.

Pathak, A. (2014). Predictive time series analysis of stock prices using neural network classifier. *International Journal of Computer Science & Engineering Technology (IJCSET)*, 5(3), 191-195.

Hajizadeh, E., Ardakani, H. D., Shahrabi, J. (2010). Application of data mining techniques in stock markets: A survey. *Journal of Economics and International Finance*, 2, 109-118.

Altay, E., Satman, M.H. (2005). Stock market forecasting: Artificial neural network and linear regression comparison in an emerging market. *Journal of Financial Management and Analysis*, 18-33.

Beogradska berza a.d. Retrieved April 12th, 2019,from: https://www.belex.rs/trgovanje/indeksi/belex15/istorijski

Kumar, GN., C. (2018) Machine Learning Types and Algorithms. Retrieved May 7th, 2019,from: https://towardsdatascience.com/machine-learning-types-and-algorithms-d8b79545a6ec