Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 44(2), 2017, Pages 292–298
ISSN: 1223-6934

Business Intelligence: Statistics in predicting stock market

Smaranda Belciug and Adrian Sandita

Abstract. Predicting stock market price over time is an important issue that needs to be addressed. The closing price for the stocks is strongly correlated to many variables, so, in order to make the prediction more accurate one must try to detect the most important ones. The main objective of this research is to predict the on day closing price of ten companies enlisted in the Romanian Stock Market using different machine learning (ML) techniques, such as multiple linear regression (MR), and multilayer perceptron neural networks (MLP) setup for regression.

2010 Mathematics Subject Classification. Primary 68T20; Secondary 68T99.
Key words and phrases. machine learning, neural networks, regression, stock market, statistics.

1. Introduction

In the last decade, Artificial Intelligence (AI) and ML have taken a great leap in detecting, diagnosing, predicting, and resolving certain issues that humankind has addressed. One important real life problem relates to trading. Can Machine learning predict stock market? The fluctuation of the on day closing price of a company is strongly influenced by certain predictors recorded over time, such as the number of transactions, the number of stocks, their price range, or the daily exchange rate [6].

Stock market prediction has been done with success using MR, fuzzy type-2 clustering, and neural networks (NNs) in [3], [7]. An evolutionary regressor selection in ARIMA model has been performed in forecasting stock price time series [8]. An optimized NN model has been used to predict the direction of stock market index movement in [9]

The statistical technique that expresses an outcome from different independent variables is the MR. The goal of this paper is two-fold: at first, to try to predict the on day closing price using this type of regression, and, secondly, to try to reduce the number of independent variables using a feature selection (FS) mechanism.

The paper is organized as follows. Section 2 presents the ML techniques which are used to predict and reduce the number of variables. Section 3 presents the real-world problem of predicting stock prices. Results are described in section 4, and conclusions and future work are given in section 5.

Received November 13, 2017.
This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI-UEFISCDI, project number PN-III-P2-2.1-CI-2017-0469, within PNCDI III..

2. Machine Learning techniques and Feature Selection for Stock Price Forecast

2.1. Multiple linear regression. MR represents a regression model in which the dependent variable is expressed as a linear combination of the predictor variables [4]. From a mathematical point of view, the dependent variable is computed from the multiple regression equation, given by:

$$Y = a + b_1X_1 + b_2X_2 + ... + b_nX_n,$$

where Y is the outcome and $X_1, X_2, ...X_n$ are the predictive variables. The constants $a, b_1, b_2, ..b_n$ are the intercept and the regression coefficients, respectively. In order to compute the regression coefficients, we used the *mean square error* (MSE) method [2]. The existing data satisfy the following equations:

$$y_i = b_0 + b_1x_1^i + b_1x_2^i + ... + b_mx_1^i + \epsilon_i, \quad 1 \leq i \leq n,$$

where x_j^i, $1 \leq j \leq m, 1 \leq i \leq n$ and y_i, $1 \leq i \leq n$ are taken from the training dataset, and ϵ_i are unknown random variables (i.e., the model's errors).

In a matrix form, our model becomes:

$$Y = Xb + \epsilon,$$

where:

$$Y = \begin{bmatrix} y_1 \\ ... \\ y_n \end{bmatrix}$$

is a n - dimentional vector, and

$$\begin{bmatrix} 1 & x_1^1...x_m^1 \\ ... & \\ 1 & x_1^n...x_m^n \end{bmatrix}$$

is a matrix, where the columns i are the m variables x_1^i, b is the regression coefficient that must be computed, and ϵ_i represents the error of the model [1].

The matrix X has the rank $m+1$, $X'X$ has an inverse, then:

$$b = (X'X)^{-1}X'Y.$$

2.2. Multilayer Perceptron Neural Network for linear regression. NNs mimic the way human brain works. MLP is a supervised learning algorithm that learns from a training dataset. Given a set of features and a target output, MLP can learn a non-linear function for regression. The difference between the classical MR and MLP for regression consists in the number of non-linear layers, also known as the hidden layers, thus providing a much powerful model.

2.3. Feature selection. FS is a key factor in designing an intelligent decision system, since even if the model is the best, it will perform poorly if the features are not chosen well. Thus, one must make an *a priori* use of specific methods of selecting the most relevant attributes, thus eliminating the redundant information.

Even if *"correlation does not always imply causation"*, we have used the analysis of the correlation matrix in order to establish possible connections between the values of the attributes and the predicted variable.

3. Materials

The data used in this research has been collected from the Romanian Stock Market (Bursa de Valori Bucuresti -BVB). For the on day closing price forecasting, there have been selected ten well-known companies (BIO, BRD, BVB, EBS, EL, FP, SNP, TEL, TGN, TLV), that are listed on the BVB. The datasets have eight independent attributes: the number of transactions ($X1$), total number of stocks ($X2$), total value of stocks ($X3$), minimum price ($X4$), medium price ($X5$), maximum price ($X6$), opening price ($X7$), the Romanian Leu-EUR exchange rate ($X8$), and one dependent variable, that is the on day closing price (Y). The data was recorded daily, except weekends from January, 4th till December, 5th.

4. Results

The models were implemented in Python. The scope is to forecast the on day closing price for each company in turn. For each company, another aim was targeted to determine which features should be used. This aim revealed that even if all the parameters where recorded for all the companies, different variables were used in the prognosis for each company.

Specifically, we studied the relationship between the attributes and the outcome on these datasets. The description of the relationship that might exist between these variables is done by examining the possible connections between the series of observations. Specifically, we examined whether there is an upward or downward trend, or no trend at all.

Thus, we computed the correlation coefficients for each group of variables. Because each one of the then dataset has over 230 observations, we can assume that the variables have approximately Normal distribution. We computed the *Pearson product-moment correlation coefficient* r, or *Pearson's* r, representing the strength of the linear relationship between two variables [5] given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The Pearson'r takes values between -1 and 1. If $r = 0$, then there it might be a nonlinear association between the two variables. The corresponding p-level, measuring the statistical significance, was computed to highlight the statistical significance of the strength of the relationship between variables.

In the tables below we presented the correlation coefficient and p-level for each attribute per each company.

Technically, the correlation coefficients and p-levels between the independent variables and outcome, for each company, are presented in tables. The highlighted variables are considered significantly correlated to the dependent variable (default p-level < 0.05). The regression equation for each company is displayed below the tables.

From the tables above, we can see that even if all eight variables were recorded for each company, the regression equation differs from case to case. It is noteworthy

TABLE 1. BIO - Correlation matrix (p-level)

X_1	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
-0.52 /	-0.21 /5	-0.18 /	**0.97** /	0.98 /	**0.97** /	**0.96** /
0.108	0.075	0.102	**0.000**	0.464	**0.000**	**0.000**

$$Y = 0.007184 + 0.505 * X4 + 0.682 * X6 - 0.27 * X7$$

TABLE 2. BRD - Correlation matrix (p-level)

X_1	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
0.04 /	-0.03 /	0.01 /	**0.98** /	**0.99** /	**0.08** /	**0.97** /
0.845	0.251	0.282	**0.000**	**0.000**	**0.000**	**0.000**

$$Y = -1228.1409 + 0.447 * X4 + 0.68 * X5 + 0.336 * X6 - 0.46 * X7$$

TABLE 3. BVB - Correlation matrix (p-level)

X_1	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
-0.13	0.02 /	0.05 /	**0.99** /	**0.99** /	**0.99** /	**0.98** /
0.012	0.172	0.15	**0.02**	**0.000**	**0.000**	**0.000**

$$Y = -883.3056 - 0.02 * X1 + 0.194 * X4 + 0.529 * X5 + 0.589 * X6 - 0.31 * X7$$

TABLE 4. EBS - Correlation matrix (p-level)

X_1	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
-0.42	-0.34	-0.31 /	**0.98** /	**0.99** /	**0.98** /	**0.97** /
0.06	0.285	0.23	**0.000**	**0.000**	**0.000**	**0.000**

$$Y = 9708.888 + 0.387 * X4 + 0.43 * X5 + 0.613 * X6 - 0.44 * X7$$

TABLE 5. EL - Correlation matrix (p-level)

X_1	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
-0.25 /	-0.17 /	-0.13 /	**0.99** /	**0.99** /	**0.99** /	**0.98** /
0.224	0.162	0.201	**0.000**	**0.000**	**0.000**	**0.000**

$$Y = 114.8568 + 0.434 * X4 + 0.385 * X5 + 0.661 * X6 - 0.48 * X7$$

TABLE 6. FP - Correlation matrix (p-level)

X_1	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
-0.25 /	0.06 /	0.09 /	**0.97** /	**0.99** /	**0.98** /	**0.96** /
0.711	0.913	0.903	**0.001**	**0.000**	**0.000**	**0.000**

$$Y = -0.0159 + 0.175 * X4 + 0.836 * X5 + 0.265 * X6 - 0.28 * X7$$

that only two variables are present in all equations, that is the minimum price and the opening price. Two other variables are present in nine from ten situations, that is the medium and maximum price. Besides the classical MR, we have also used the regression provided by MLP. We have used a 3-MLP architecture (that is one-hidden

Table 7. SNP - Correlation matrix (p-level)

X_1	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
-0.13 /	-0.14 /	-0.12 /	**0.98 /**	**0.99 /**	**0.98 /**	**0.96 /**
0.389	0.029	0.03	**0.000**	**0.000**	**0.000**	**0.000**

$$Y = -0.0015 + 0.495 * X2 - 0.47 * X3 + 0.38 * X4 + 0.501 * X5 + 0.574 * X6 - 0.45 * X7$$

Table 8. TEL - Correlation matrix (p-level)

X_1	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
-0.19 /	-0.04 /	-0.02 /	**0.96 /**	**0.98 /**	**0.97 /**	**0.94 /**
0.264	0.679	0.657	**0.000**	**0.000**	**0.000**	**0.000**

$$Y = -3865.0055 + 0.331 * X4 + 0.397 * X5 + 0.775 * X6 - 0.51 * X7$$

Table 9. TGN - Correlation matrix (p-level)

X_1	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
-0.22 /	-0.12 /	-0.08 /	**0.99 /**	**0.99 /**	**0.99 /**	**0.98 /**
0.143	0.521	0.492	**0.000**	**0.000**	**0.000**	**0.000**

$$Y = -22156.385 + 0.3 * X4 + 0.575 * X5 + 0.561 * X6 - 0.44 * X7$$

Table 10. TLV - Correlation matrix (p-level)

X_1	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
-0.06 /	-0.05 /	0.05 /	**0.99 /**	**0.99 /**	0.99 /	**0.98 /**
0.142	0.274	0.237	**0.000**	**0.000**	0.319	**0.000**

$$Y = 221.5868 + 0.321 * X4 + 0.926 * X5 - 0.32 * X7$$

layer) with 11 hidden nodes. The figures below show the targeted on closing day price, along with the output price provided by MLP.

We can see from the figures above that MLP can forecast the on day closing price very well for 6 out of 10 companies: BRD, BVB, FP, SNP, TGN and TLV. It is noteworthy to mention that we have used a standard architecture of MLP without tuning its parameters, because this was not our purpose in this study. It is expected that an optimization of the MLP architecture will lead to an increase in the percentage of companies for which the on day closing price can be predicted.

5. Conclusions and future work

In this study, we have applied MR and 3-layer MLP for regression, in order to predict the on day closing price for ten Romanian companies. For each company the strongly and significantly correlated independent variables were chosen in order to optimize the forecasting process.

However, the prediction can be further improved by applying the forward stepwise regression and the backward stepwise regression as well. Moreover, as we stated above, a tuning process of MLP is also needed. We may even use different NN types, such as

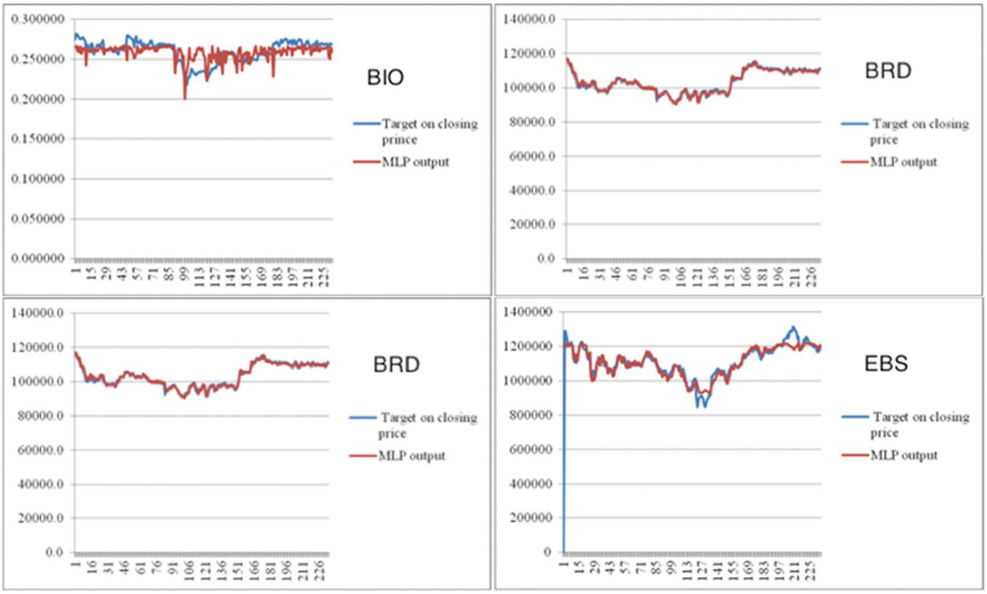

FIGURE 1. Forecasting on day closing price with MLP for regression: BIO, BRD, BVB, EBS.

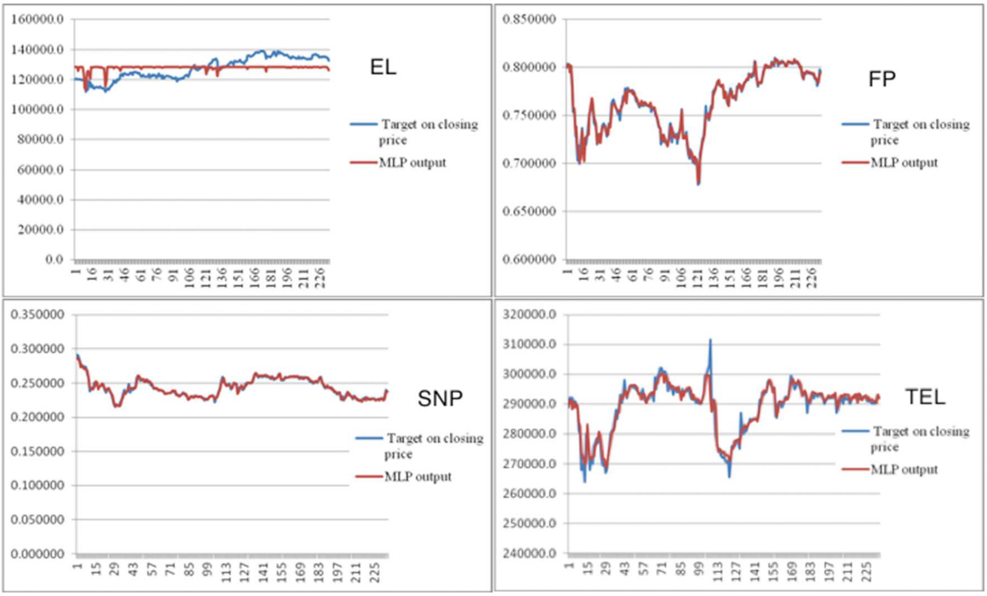

FIGURE 2. Forecasting on day closing price with MLP for regression: EL, FP, SNP, TEL.

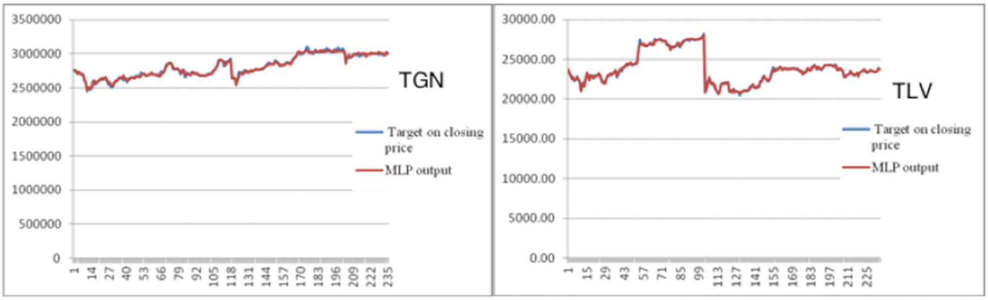

FIGURE 3. Forecasting on day closing price with MLP for regression: TGN, TLV.

radial basis functions (RBF) or probabilistic neural networks (PNN), to address this problem.

6. Acknowledgement

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI-UEFISCDI, project number PN-III-P2-2.1-CI-2017-0469, within PNCDI III.

References

[1] N. Chernov, *Circular and Linear Regression, Fitting circles and lines by least squares*, CRC Press, Taylor and Frances Group, 2011.

[2] R. Christenesen, *Analysis of Variance, Design and Regression, Linear Modeling for unbalanced data*, CRC Press, Taylor and Frances Group, 2016.

[3] D. Enke, M. Grauer, N. Mehdiyev, Stock Market Prediction with Multiple Regression, Fuzzy Type-2 Clustering and Neural Networks, *Procedia Computer Science* **6** (2011), 201–206.

[4] F. Gorunescu, *Data mining, Concepts, models and techniques*, Springer, 2011.

[5] S. Huet, A. Bouvier, M.-A. Poursat, E. Jolivet, *Statistical Tools for Nonlinear Regression*, Springer, 2004.

[6] I. Pardoe, *Applied Regression Modeling, A business approach*, Wiley, 2006.

[7] P.K. Sahoo, C. Krishna, Stock price prediction using regression analysis, *International Journal of Scientific & Engineering research* **6** (2015), no. 3, 1655–1659.

[8] R. Stoean, C. Stoean, A. Sandita, Evolutionary Regressor Selection in ARIMA Model for Stock Price Time Series Forecasting. In: Czarnowski I., Howlett R., Jain L. (eds) *Intelligent Decision Technologies 2017. IDT 2017. Smart Innovation, Systems and Technologies* **73** (2018), Springer, Cham.

[9] M. Qiu, Y. Song, Predicting the direction of Stock market index movement using an optimized artificial neural network model, https://doi.org/10.1371/journal.pone.0155133, *Plos one* (2016).

(Smaranda Belciug, Adrian Sandita) DEPARTMENT OF COMPUTER SCIENCE, FACULTY OF SCIENCES, UNIVERSITY OF CRAIOVA, 13 A.I. CUZA STREET, CRAIOVA, 200585, ROMANIA
E-mail address: sbelciug@inf.ucv.ro, asandita@inf.ucv.ro