APPRAOCH BASED ON LINEAR REGRESSION FOR STOCK EXCHANGE PREDICTION – CASE STUDY OF PETR4 PETROBRÁS, BRAZIL

Nadson S. Timbó, Sofiane Labidi, Thiago P. do Nascimento, Milson L. Lima, Gilberto Nunes Neto and Rodrigo C. Matos

Post-graduation Program in Electrical Engineering, Federal University of Maranhão, São Luís, Brazil 65080–805

ABSTRACT

The stock exchange is an important apparatus for economic growth as it is an opportunity for investors to acquire equity and, at the same time, provide resources for organizations expansions. On the other hand, a major concern regarding entering this market is related with the dynamic in which deals are made since the pricing of shares happens in a smart and oscillatory way. Due to this context, several researchers are studying techniques in order to predict the stock exchange, maximize profits and reduce risks. Thus, this study proposes a linear regression model for stock exchange prediction which, combined with financial indicators, provides support decision-making by investors.

KEYWORDS

Stock Exchange Prediction, Autoregressive Models & Linear Models

1. INTRODUCTION

According to Nascimento [1], the necessity for economic growth is a consensus in any country. A path for its achievement is through the capital markets, which directs the investments to alternatives that provides the highest returns. This is defined as a system of transferable securities distribution capable of providing liquidity to issuance of securities that at the same time makes the capitalization process viable. It occurs through negotiations between organizations and investors, up buying and selling shares in the stock exchange. The Bovespa is the main stock exchange in operation in Brazil. It is considered the largest stock market in Latin America and the eighth worldwide. In recent years, this stock exchange was the target of huge investments, closing in 2012, with over one trillion dollars. The main explanation for this is the increase of domestic markets, the reduction of interest rates and, specially, the currency appreciation, which encourages new companies and investors to seek, in this market, a new manner of growing through raising capital and increasing equity [2][3].

While it may be considered a very gainful way of investment, the stock exchange is also based upon losses, especially when transactions are made without the proper data analysis to support it. According to Rocha and Macêdo [4], once entered in the market, the goal is to always buy shares when their prices are the lowest and, consequently, selling at the highest price. This explains why predicting shifts in the market’s behavior means maximizing profits and reducing risks. Such anticipation, also called prediction, increase the profitability of investments.
Currently, there are several techniques used by financial experts that help with the stock market analysis. Two of them stand out, the Technical Indicators and Fundamental Indicators, that although are the most widespread, barely predict, with accuracy, the real market’s behaviour, as it is a complex task and it relies on variables capable of expressing from the investor’s intuition to natural disasters [1].

This study proposes the application of the autoregressive linear model of stock exchange prediction using, as parameters, data collection and technical indicators. In addition, in its conclusion, it is presented a comparative study between this study and other related work.

2. PROBLEM DESCRIPTION AND STATE-OF-THE-ART

The global trend is focused on the economy, aiming that investments be directed towards alternatives that generate higher profits. The Capital Markets allows such goals with superior flexibility and reliability in transactions, through the stock exchange.

According to Benjamin Graham [5], investing in the stock exchange requires intelligence. The investor desiring join the market, must obtain a safety margin on their investments. Obtain such margin is not a simple task, as a result of market volatility.

There are traditional methods that assist investors making the best decision about how much, how and when invest. These methods are commonly known as: Technical Analysis and Fundamental Analysis. The objective of the Fundamental Analysis is to predict the future behavior of a share, for instance, while taking into account the marketing situation of the company, that trades these shares or the sector in which this share belongs to [6]. The Technical Analysis seeks anticipate the market’s comportment or a share, through statistics and probabilities linked to the manners of the referred market or share. Is worth noting that, both the Fundamental Analysis and Technical Analysis, even being widely used by financial experts, are sometimes unable to predict the actual market state, owing to the comportment of the financial environment, which is dynamic and oscillatory. Pereira, Turrioni and Pamplona [7] argue that traditional methods of analysis (ROI, IRR, etc) have not evolved with information technology and no longer are as effective. If there is a demand for the technology, It is more present in the financial aspects [8][9].

Several studies discuss the predictive methods applied to financial markets. The field of Artificial Intelligence, for example, triggered the development of specialized computational tools for the stock market. These tools aim to use market variables to obtain predicted values increasingly closer to the actual values of these shares [10]. The Regressive Models, the Artificial Neural Networks, the Fuzzy Logic, Genetic Algorithms and Hybrid Models of prediction are emphasized.

Marques [11], in his work, used the Genetic Algorithm technique, with the Fuzzy Logic, to find optimal settings for the MACD indicator performing predictions about future trend of Bovespa shares. As a result, the aforementioned author, reached a precision rate of 53.75% and 49.5% for Petr4 and Vale5 shares, respectively.

In another research field, the work developed by Almeida [12] is available, entitled “A Market Prediction Model Stock Based on Fuzzy Logic”. In this study, the fuzzy control process was applied to assist the investor making the decision of keeping, buying or selling shares. Such as the model Almeida [12] used, as input data, was the quotation of Petr4 and Vale5 from years 2007 to 2014. Almeida [12] reported that the results obtained can propose coherent decisions, however, a
percentage of success/mistakes was not found, decreasing the credibility the decision-making supporting.

Nascimento [1], in his work “A System Based on Genetic Algorithms as a Decision Making Support for the Purchase and Sale of Assets at Bovespa”, used evolutionary techniques known as Genetic Algorithms to perform predictions on closing prices in shares traded in Bovespa. Therefore, he used, as parameters, stock market quotations from 01/01/13 to 12/31/13, totaling 258 samples for data comparison. In this study, 100 days of prior quotations were used to predict the 101th day. As the best results of this approach, it obtained 55.8% of predictions closest to real values. It is relevant noting that this study took into account the closest predictions which the difference to the actual closing price was inferior to 0.50 cents.

Finally, the work of Fagner, entitled “Modeling via artificial neural networks the Capital Market to the share value forecast and improve steering accuracy rate – Case Study Petr4 – Petrobras, Brazil” [13], objected the prediction of future trends regarding share prices of Petr4. Among the evaluated studies, this was the one that succeeded the best results, with a success rate of 91.48%. Fagner used the Artificial Neural Networks in his research, which consists of non-linear prediction models and, therefore, considered compelling methods of stock exchange predictions [13].

3. METHODOLOGY

Meaning prediction model of stock exchange assets, the opted methodology was the process of Knowledge Discovery in Databases (KDD), which is a AI field that looks for the knowledge extracting from larger amounts of data. To Fayyad [14] the KDD is a multistep process, non-trivial, interactive and iterative, for identification of comprehensive, valid, new and potentially useful standards from large data sets.

By identifying patterns, trends, and finally extracting the desired knowledge, the person responsible for the process is able to make a more informed and strategic decision about the transaction, then solving the problem caused by the “Information Age”. [14].

And possessing domain over its application and problem, one can attain knowledge through 5 stages (eg, Figure 1): Data selection, pre-processing, transformation, mining and, finally, Evaluation and Results Interpretation. These steps are described and exemplified by the proposed study.

![Figure 1. An Overview of the Steps That Compose the KDD Process](image)
At the first stage, there is a possibility to find a certain complexity if the data is not in a single repository or requires an expert to collect the data. It is worth mentioning the importance of this step, as we consider the data as key components in the full scope of the problem [16].

For data collection, Petr4 assets has been set as scope, in the timeframe ranging from 01/02/14 to 06/19/15. Located on the Yahoo Finance platform [17], through a web requisition and a few parameters, as depicted in the table (eg. Table 1).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>Action Stock Exchange</td>
</tr>
<tr>
<td>a</td>
<td>Initial Month</td>
</tr>
<tr>
<td>b</td>
<td>Initial Day</td>
</tr>
<tr>
<td>c</td>
<td>Initial Year</td>
</tr>
<tr>
<td>d</td>
<td>Final Month</td>
</tr>
<tr>
<td>e</td>
<td>Final Day</td>
</tr>
<tr>
<td>f</td>
<td>Final Year</td>
</tr>
<tr>
<td>g</td>
<td>Type of Grouping</td>
</tr>
</tbody>
</table>

Consequently, the quotation data was acquired: date, opening price, maximum, minimum, closing price, volume and the average closing.

On the second step, the Pre-processing and data cleansing process can take a larger amount of time [18], aiming integration of heterogeneous data, eliminating incompleteness, verified consistency of information, filling or eliminating null values, incomplete, duplicated or corrupted. The data was then analysed and was stated that there are no inconsistent, redundant, null or corrupted values. “Duplicated” values were found, however, was verified, and certified, that it was not the case of duplicity, but rather of days in which the closing price on the stock was the same as the day before.

A study was applied to the data and it was defined as variables only the opening, closing, minimum, maximum and volume values. Later, an exploratory analysis of variables was made and ranked them by measure levels, such as quantitative variables, and by manipulative levels, such as independent variables (i.e. their values affects others) and dependent (i.e. the measured or registered variables, which are influenced).

The third step is dependent on data mining, targeting the transformation of data, to present the ideal model for development and interpretation of the algorithm in the next step.

With the view to transform the data to meet the requirements of the mining step, was noticed the lack of variables making mining more based on the real stock exchange scenario. Then, the main indicators utilized by financial experts were collected by determining the behavior of the shares and, which to invest. Among those checked, it was found:

1. Relative Strength Index (RSI): it measures the weakening degree of a trend.
5. Force Index (FI): indicates force of an asset involving the shift change in price.
7. Moving Average Convergence Divergence (MACD): shows the relation between moving averages.
8. Direction Movement: it shows the strength of a positive and negative tendency.
9. Average Directional Index (ADX): it measures the strength of a positive and negative tendency.
11. Accumulation Distribution Line (ADL): determines the relation between demand and supply of an asset.
13. Average True Range (ATR): measures how much the price of a stock is subject to variation.

After gathering this data, a new analysis of the variables was realized and, subsequently, performed a two-dimensional analysis of all quantitative variables in relation to an independent variable, in other words, the closing price of an asset. This procedure had the requirement of listing which variables would be best applied in the mining algorithm. Thus, the Pearson correlation coefficient and the Coefficient of Determination were calculated, which determines if there is a correlation between variables, its force and direction. In the picture below, the variables are shown: opening price, maximum, minimum, moving average, MACD, OBV, ADL, pivot close and ATR obtained better results exceeding it by 0.5, given the time series chosen. Therefore, these were the ones chosen to be used in the next step.

Table 2. Coefficient of Indicators.

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Pearson Correlation Coefficient</th>
<th>Coefficient of Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>0.99487473</td>
<td>0.989005334</td>
</tr>
<tr>
<td>High</td>
<td>0.993211486</td>
<td>0.986469055</td>
</tr>
<tr>
<td>Low</td>
<td>0.9948341</td>
<td>0.989694887</td>
</tr>
<tr>
<td>Pivot Close</td>
<td>0.995855488</td>
<td>0.991728153</td>
</tr>
<tr>
<td>Moving Average</td>
<td>0.978291082</td>
<td>0.957053441</td>
</tr>
<tr>
<td>MACD</td>
<td>0.941127663</td>
<td>0.885712728</td>
</tr>
<tr>
<td>OBV</td>
<td>0.762275575</td>
<td>0.581064052</td>
</tr>
<tr>
<td>ATR</td>
<td>0.584487817</td>
<td>0.341626008</td>
</tr>
<tr>
<td>ADL</td>
<td>0.508326204</td>
<td>0.25839553</td>
</tr>
<tr>
<td>Directional Movement</td>
<td>0.452690778</td>
<td>0.204928941</td>
</tr>
<tr>
<td>FI</td>
<td>0.327514359</td>
<td>0.107265656</td>
</tr>
<tr>
<td>Volume Pivot Close</td>
<td>0.247212267</td>
<td>0.061113905</td>
</tr>
<tr>
<td>IFR</td>
<td>0.193201546</td>
<td>0.037326837</td>
</tr>
<tr>
<td>Stochastic</td>
<td>0.182864204</td>
<td>0.033439317</td>
</tr>
<tr>
<td>Williams</td>
<td>0.169468496</td>
<td>0.028719571</td>
</tr>
<tr>
<td>MFI</td>
<td>0.151817821</td>
<td>0.023048651</td>
</tr>
<tr>
<td>ADX</td>
<td>-0.029023545</td>
<td>0.000842366</td>
</tr>
<tr>
<td>Volume</td>
<td>-0.097613838</td>
<td>0.009528461</td>
</tr>
</tbody>
</table>

The mining stage is considered the center of the KDD, which is a technique to determine behavioral patterns, in larger databases, assisting decision-making [19].

Mining is composed by tasks and techniques (methods). The task consists of the specification of what do you want to search in the data, types of regularities and patterns. The technique consists of the specification of methods that show how to discover patterns.
The mining tasks are sorted by the diversified ability to extract knowledge, and its success relies exclusively on the application domain. The most common are: Classification, Association, Estimation or Regression, Clustering and Summarization [14]. The Regression is similar to the Classification however, is used to define a numerical value for an unknown continuous variable. It can be used to estimate future data, such as family income, lifetime-based in diagnosis, etc. Estimation is learning a function that maps a given item for a variable of estimated real prediction [14]. Given the main objective of this study the prediction of the closing price value of an asset, meaning, predict a continuous and quantitative variable, it can be asserted that this technique is best suited for this case study.

The mining techniques (or methods) can be applied to tasks, implemented by an algorithm elaborated to solve a task. For Harrison [20] there are different techniques are better suited to a determined task, having advantages and disadvantages.

The methods are divided into Supervised Learning and Unsupervised Learning. In this case study, the chosen was the former, also known as predictive, where there is a class for where each sample is assigned in training.

The algorithms belonging to these methods perform inferences in data, providing predictions, the classification and regression tasks nevertheless, use different methods, such as: neural networks, genetic algorithms, inductive logic, decision trees, Bayes classifier, rule based classification, SVM (Support Vector Machines), Fuzzy Set, linear regression, nonlinear regression (logistic, polynomial), etc.

It was verified that there are a wide range of works using neural networks, genetic algorithms and even fuzzy logic, so it was proposed using regression techniques, since it allows to find a function that estimates the manner based on a data set, through mathematical and statistics resources. The next step was the analysis of all the possible variables of the problem scope with objective to predict the closing price for the next day. Therefore, proving the existence of a univariate, linear and multiple model. As a result of only one dependent variable or answer (closing price) was found and, for all other variables that went through two-dimensional analysis only the ones with strong correlation with the dependent variable (as seen in the picture above) were selected, confirming linearity of the problem (for the selected variables), and multiple for presenting more than one predictor or independent variable. The model specification is about:

- Multiple regression (has more than one predictor variable);
- Univariate regression (has only one variable answer);
- Positive linear regression;

As a consequence, the selected method for applying data mining was the multiple linear regression, using the Least Square Method.

For the fifth and last stage of the KDD is the Evaluation and Interpretation of Results which seeks the knowledge, transforming it in a real model for decision-making. This step is intend to facilitate the interpretation and evaluation of the utility of the discovered knowledge, by the end user.

4. RESULTS

Only the preferential asset from the company Petrobras, or simply PETR4, was utilized in the scope of this study due to its history in Bovespa since it is the focus of other works related with the stock exchange of São Paulo, thereby, assisting in comparing results.
Data collection was conducted between 01/02/14 and 12/30/15, in a total of 342 days analyzed, willing to predict the next day. It was split 70% for training and discovery of the pattern and 30% for validation. Then it was set as predictor variables: opening price, maximum, minimum, 5-day moving average (MMS_5), 10-day moving average (MMS_10), 15-day moving average (MMS_15), exponential 20-day moving average (MME_20), MACD, OBV, ADL, Closing Point volume (Volume_PF), Closing Point (PF), ADX and ATR. Of the first data, 70% were applied to training, subsequently generating a pattern, therefore, a mathematical function, as shown in the figure (eg, Figure 2).

Through the exploratory and two-dimensional analysis of the variables already mentioned it is possible to fit into a linear model. Nonetheless, some variables did not present enough representation to be considered into account in linear regression. Therefore, obtaining the mathematical function coefficient equal to 0 (zero). This is seen through the variables of minimum value, OBV, ADL and Volume_PF, so, despite having 13 indicators, only 10 of those were used to predict values.

We also obtained a Correlation Coefficient of 0.9917, a Mean Absolute Error of 0.3699, a Root Mean Squared Error (identifies deviation from the best result) of 0.4716 and a Relative Absolute Error (indicates the proximity of the prediction compared to the real value) of 6.1555%. Therefore, claiming the current study proposes a positive impact and is an efficient model assisting decision-making by financial experts.

The equation found was applied to the remaining 30% of the data, so that is possible verify the percentage of correct answers, represented in a total of 66 instances. Discovering the prediction of the closing price and comparing it with the real closing price value, we obtained the following results.

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Pearson Correlation Coefficient</th>
<th>Coefficient of Determination</th>
<th>Hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>14,96</td>
<td>15,13929</td>
<td>0.179285</td>
<td>true</td>
</tr>
<tr>
<td>14,61</td>
<td>14,87246</td>
<td>0.262457</td>
<td>true</td>
</tr>
<tr>
<td>14,58</td>
<td>14,79304</td>
<td>0.213037</td>
<td>true</td>
</tr>
<tr>
<td>14,43</td>
<td>14,75516</td>
<td>0.325162</td>
<td>true</td>
</tr>
<tr>
<td>14,11</td>
<td>14,51267</td>
<td>0.402667</td>
<td>true</td>
</tr>
</tbody>
</table>
It can be verified how close the equations results are similar to the actual real closing, and considered as success a maximum deviation of 0.50 cents, which according to Bovespa is satisfactory.

Important to highlight that a $0.50 cents deviation represents 4.7169% of the value, which generates a success rate of 60.87% the cases. Proving that the main goal of the study was achieved by finding a regressive model to predict the value of an asset with a high success rate.

5. DISCUSSIONS AND COMPARISONS

The information systems possess applications in several areas, in which the study of the financial market was the focus in this paper.

Given the difficulty of time series predictions with countless variables and the complexity of specifying a consistent and efficient model, the use of computational techniques of Artificial Intelligence was selected to solve this problem. Consequently, the current study aimed to take benefit from KDD to specify a multiple linear regressive model for predicting the closing price of Petr4.

The application results prove the authenticity of the model, given the high accuracy rate, since the acceptable movement in Bovespa is $0.50, making it possible to accomplish an estimative of success of a movement even lower. This proposal can be applied to any asset, as long as it passes through all the KDD stages, mentioned before, and serving as a strong base for the financial expert in regards to decision-making. Confirming the proposed model as a contribution to the financial market and a new, complete and, efficient form of analysis. The comparative between the closing price and the prediction can be better observed in the picture (eg, Figure 3).
The necessity to insert several indicators as input variables is ratified, rendering a more robust model in a real situation of stock exchange manner, due to its complex environment and countless variables that can affect an asset.

What differentiates this study is the closing price prediction value, as most of others studies are concerned with discovering trends, while few focus on value predictions.

By analyzing and interpreting the results, it has been proven that trend indicators present better results when the database is composed of more than 100 entries. When the database is lower than that, it is recommended not to use these variables.

In literature, renowned models capable of prediction time series behavior can be found, such as the Box-Jenkins univariate method called ARIMA. However, applying this method using closing prices for predictions, it did not result in success, even though 12 differentiations were applied a stationary series that fulfilled the methodology was not found, leading towards believing in the method applied by this study.

Almeida [12] says, the fuzzy logic was used to predict the trend, comparing it to three consecutive days in the stock exchange. However, a percentage of success/mistakes was not enough to validate the model. It was also considered low the quantity of indicators used as variables input, given the stock exchange complexity. Nonetheless, in the work of Marques [11] a hybrid model was proposed for trend prediction, but its success rate was only considered average. The results produced by the current study the success rate was 60.87% and it can be considered an improvement over the results of Marques [11], which concludes that linear techniques can be considered in value predictions and its results are similar to other artificial intelligence techniques.

Regarding the work of Fagner [13], which obtained great results, there was a difference while choosing the data, as Fagner choose the end of the month as closing price values, while the current study opted to analyze the whole time series of the asset, aiming to stay truthful to the stock exchange complexity.
About Nascimento [1], the best predictions made involved the Petrobrás share Petróleo. In his model, the author conducted estimations based on 258 of quotations. Among these, the number of estimative that were closer to below 0.50 from the real values of shares were only 144, thus, 55.8% from the total sample. In a comparative study between the approach of Nascimento [1] and the current study being presented, is noted that the success rate of Nascimento [1] has been surpassed.

REFERENCES

AUTHORS

Nadson Silva Timbó
BS in Computer Science – Federal University of Maranhao (2013). Has experience in Computer Science Area.

Sofiane Labidi

Thiago Pinheiro do Nascimento

Gilberto Nunes Neto

Milson Louseiro Lima

Rodrigo Costa Matos
Has experience in Computer Science.