IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN:2278-487X, p-ISSN: 2319-7668. Volume 25, Issue 9. Ser. 5 (September. 2023), PP 25-31
www.iosrjournals.org

Multi-Layer Perceptron Prediction Model For Stock Market Based On Multi-Factor Data Analysis

Xiangjun Fu[1], Qiancheng Tan[1, *], Xianhao Deng[1]
[1]*Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin China*

Abstract: *Neural networks have been used in various fields, such as image recognition, cognitive science, and genomics. In this paper, we use a multi-layer perceptron model (MLP) to predict the stock market. Before making predictions, we need to process the collected data, use canonical correlation analysis to discuss the correlation between influencing factors, and use some of these factors as independent variables with the time axis. The trading volume in the stock market is input as labels into the MLP model. The results showed that the accuracy rate on the test set was 87%, and the effectiveness of the model was verified by comparing the predicted results with actual data.*

Key Word*: Keywords- Stock Market; Multi-layer Perceptron Prediction Model; Multi-Factor.*

Date of Submission: 16-09-2023 Date of Acceptance: 26-09-2023

I. Introduction

The characteristics of the stock market are often summed up as instability, variability, mutability, and so on, for example [1]. Thus, it is difficult to predict these characteristics of the stock market directly [1][2]. Nowadays, the study for the prediction of the stock market has been received attention and research [3][4] and references therein. Especially, the prediction method of the stock market is often depended on the multi-factors, such as earnings per share, inflation rate, etc[5][6]. Therefore, it is very important to analyse the data which effects the stock market. With the increasing number of stock market predictions, it is also scientifically significant to study deep learning models for stock market[7].

In recent years, the machine learning technology has important applications in many scientific fields, such as image recognition [8], cognitive science[9], and genomic. Thus, the kind of method has great potential for the prediction of the stock market[7]. Deep neural networks have great advantages in dealing with nonlinear problems and can better fit data. In stock market prediction, identifying key features that affect the performance of machine learning (ML) models is crucial for achieving accurate stock price prediction.

In this paper, we analyse the factors which affect the trend of the stock market, and focus on the analysis of the correlation between the digital economy sector with other sectors. Based on the correlation analysis results, it can be concluded that there is a strong correlation between the digital economy sector and others. Afterwards, we use the factors of the technology sector as the main influencing factors to explore the impact of multiple factors on the stock market trend. While Multilayer Perceptron (MLP) is an Man-made neural network based on forward feedback and contains multiple nodes, including input, hidden and output layers. The nodes in each layer are connected to the nodes in the next layer of the network in a fully connected manner. The input data is combined with the weight W and threshold b of the nodes in the layer, and the activation function is applied to obtain the corresponding output. The learning rule of MLP involves continuously adjusting the weight and threshold of the neural network through backpropagation. It utilizes the steepest descent method to minimize the mean squared error of the network.

II. Data Preprocessing

In order to well analyse and predict the price of the stock, the selection of data sets must come from many data sets, in order to facilitate us to divide these data sets into several Setion, first of all, data pre-processing, we take the technical indicators as an example of descriptive statistical analysis as shown in Table 1, followed by doing including the growth rate, variance, average, box plots and so on, in order to exclude the outliers and processing, as shown in Figure 1.

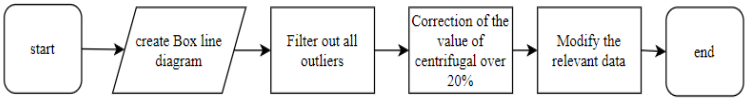

Figure 1 : Data cleaning flow chart.

Table 1: Descriptive statistical analysis under the technical indicators Section.

	VMA	VMACD	OBV	…	BOLL
Count	263000	263000	263000	…	263000
Mean	2524655573	-52003753.33	26729935.74	…	1570.061
Std	1004767501	207035479.5	963344.0452	…	72.1882
Min	468409381.2	-444700436.7	24761533.99	…	1472.261
25%	2036740395	-176721586.4	26058020	…	1509.658
50%	2530098667	-81971484.74	26674844.38	…	1557.022
75%	3026331529	40616615.61	27286554.77	…	1616.059
Max	5606072230	726077843	29541021.97	…	1756.395
Count	263000	263000	263000	…	263000

Step1: Referring to the relevant literature, it is clear that the normal distribution of the data can be determined by a box-and-line plot. If the data does not follow a normal distribution on the box plot, we must correct the data so that the overall data follows a normal distribution. If the data does not follow a normal distribution on the box plot, we must correct the data so that the overall data follows a normal distribution. The box-and-line plot after Python is shown in Figure 2. In this paper, the structural variables and six indicators of product performance are identified as outliers. Taking the technology segment as an example, this paper used the 13 indicators of the box-and-line diagram for outlier identification, and a total of 285 outliers in the 13 indicators were obtained. However, in the control experiment, the box-and-line diagram only screened out the discrete points of the same indicator. Combining the results of the previous screening and the existing literature review, two groups of outliers were identified among the six indicators in the data. Two sets of outliers in the data. In order to handle these outliers, the VMA and VMACD averages were calculated separately for the replacement process.

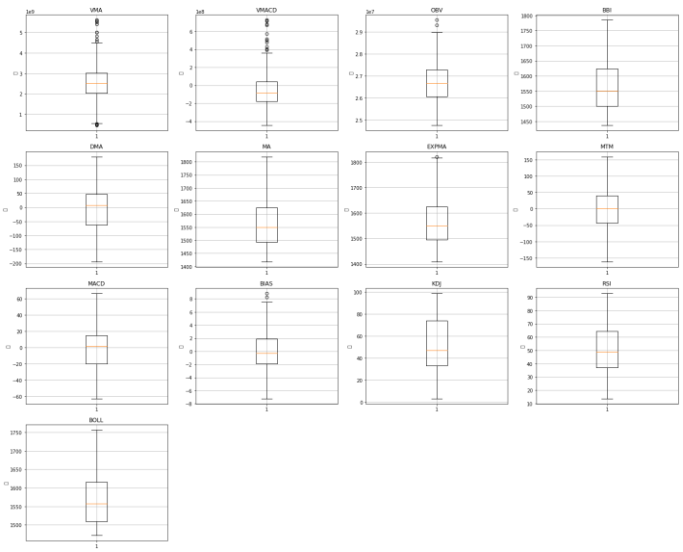

Figure 2： Standard deviation box plots for each indicator under the Technical Indicators Section.

Step2: Next, box-line plot analysis was performed for each variable, and the deviation data of the data were obtained by Python statistics. In this paper, the characteristics of the deviations to be treated are summarized by combining the mean values and relevant data queries, and the data are corrected to obtain Table 2.

Table 2： Corrected data

Nmber	Abnormal indicators	Abnormal value	Corrected Value
151	VMA	3080584565	2874655573
11	VMA	3364339276	3056489789
148	BBI	1610.4137	1599.3445
15874	BBI	1562.0318	1589.7811
95412	MA	1819.586	1765.54

5697	MA	1417.455	1556.64
1557	MTM	158.8808	134.65
9856	MTM	-161.865	-131.365

III. Canonical Correlation Analysis of Data

In mathematics, correlation refers to the degree of association between two variables. According to the context, we can study the correlation between different sectors and the digital economy sector to extract key indicators related to the digital economy sector. As for the analysis of correlation between groups of variables, typically Canonical Correlation Analysis is used, as shown in Figure 3. Canonical Correlation Analysis focuses on analysing the correlation between a subset of variables from one group and a subset of variables from another group. We can assume that these variables occur simultaneously at the same time, neglecting the influence of time on the variables. Let us represent variables from the digital economy sector and variables from other sectors, respectively, and

$$V(x) = \sum_{11} (\sum_{11} > 0),\ V(y) = \sum_{22} (\sum_{22} > 0),\ Cov(x, y) = \sum_{12} \quad , \tag{1}$$

Thus, we derive that

$$V\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}, \tag{2}$$

where $\sum_{12}^{'} = \sum_{21}^{'}$.

In this section of the article, the main focus of the research is on the correlation between technical indicators ("VMA", "VMACD", "BOLL","BIAS","ARBR"……), and six random variables (opening price every 5 minutes, closing price every 5 minutes, highest price every 5 minutes, lowest price every 5 minutes, trading volume every 5 minutes, and amount every 5 minutes) within the "digital economy" sector.

We employed canonical correlation analysis and obtained the analysis results using SAS software. The structural diagram of the correlation coefficient system is depicted in Figure 3, and the analysis results are presented in Table 3.

As shown in Table 3 of the summary table, the p-value under each statistic is <0.0001, and the typical correlation between the other segment indicators and the digital economy segment of the model is significantly valid.

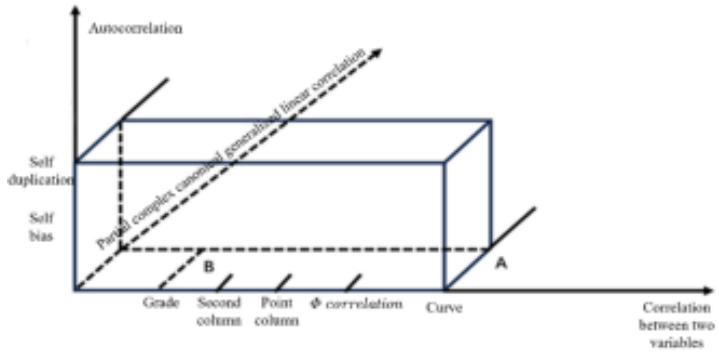

Figure 3: Structure diagram of correlation coefficient system.

Table 3: Typical correlation analysis between the five sectors and the digital economy sector.

Setion	Adjustments typically associated	likelihood ratio	p
Domestic stock market indicators	0.964767	19.39	<0.0001
International stock market indicators	0.999887	0.00000000003	<0.0001
Macro market indicators	0.812559	0.15046987	<0.0001
Technical indicators	0.754682	13.55	<0.0001
Other setion indicators	0.960788	26.94	<0.0001

IV. MLP Neural Network Based On TensorFlow

Without loss of generality, we first consider the three-layer neural network, which is given by

$$a^2 = f^2(w^2 f^1(w^1 p + b^1) + b^2), \tag{3}$$

where w^1 and w^2 are the weight from the input layer to the first hidden layer and the hidden layer to the output layer, respectively, b^1 and b^2 are the activation threshold of the hidden layer and output layer, respectively,

$f^1 = \frac{1}{1+e^{zi}}$ and $f^2 = \frac{e^{zi}}{\sum_{k=1}^{N} e^{zk}}$ are the sigmod and softmax type activation functions, respectively. The error of the neural network is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(T_i - Y_i)^2 , \tag{4}$$

where Y_i is the prediction output value of the neural network, T_i is the expected output one, n is number of the total sample. In this algorithm, the output layer error is obtained by using the gradient descent method. Thus, the weights and thresholds between the input and hidden layers can be updated by depending on this error result. Further, we repeat the above steps so that the error value defined in (4) is minimized or reaches the maximum number of training times. The network model of the three-layer MLP is shown in Fig 4.

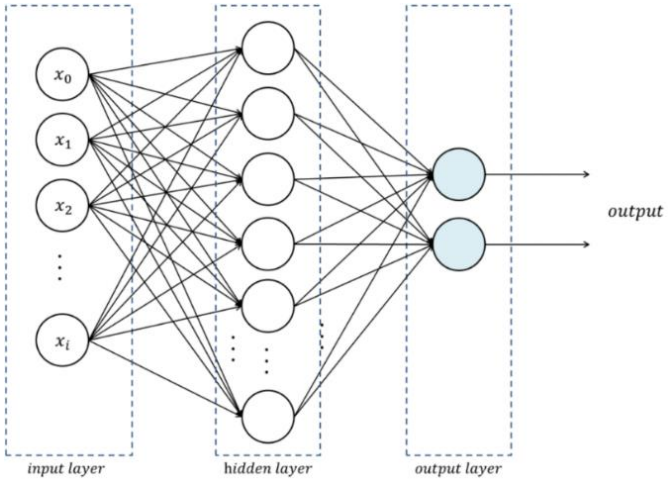

Figure 4: MLP Neural Network Diagram.

V. MLP Neural Network Based On TensorFlow

This article uses various technical indicators to expand the data of the digital economy section. The relationship matrix between the expanded data and transaction volume is shown in Fig 5. As shown in Fig 6, except for ARBR, all other indicators have an undeniable relationship with transaction volume. Therefore, using the expanded data and the timestamp corresponding to the predicted transaction volume as input, an MLP model is constructed. The MLP model includes 1 input layer, 4 hidden layers, and 1 output layer. The specific parameters are shown in Table 4.

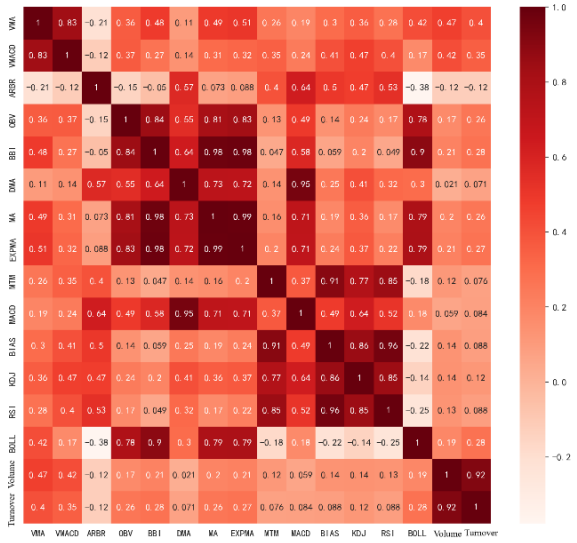

Figure 5: Expand the relationship matrix between data and transaction volume.

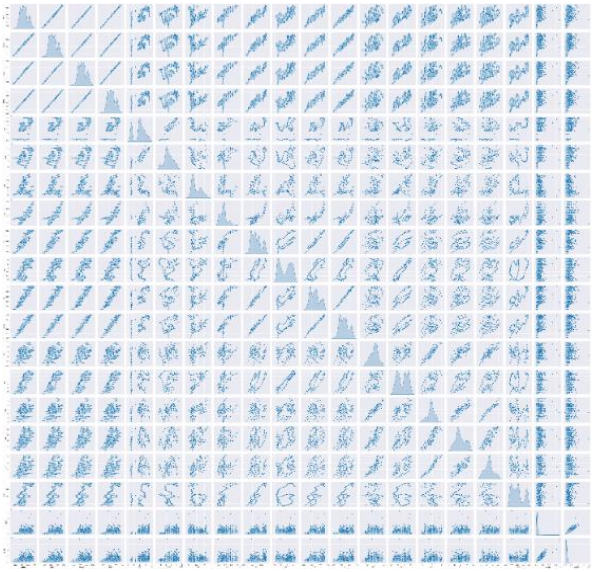

Figure 6: Relative distribution of data items in the input data.

Table 4: Description of MLP model parameters.

Dense of Neuron layer	Output Shape	Number of parameters
normalization_21	(0, 16)	33
dense_97	(0, 32)	544
dense_98	(0, 64)	2112
dense_99	(0, 64)	4160
dense_100	(0, 16)	1040
dense_101	(0, 1)	17

As shown in Figure 4, the relationship between input data and results contains a large number of nonlinear relationships. In order to better fit nonlinear relationships, nonlinear transformations are introduced into all hidden layers of the model:

$$LeakReLU = \max(\epsilon y, y), \tag{5}$$

where ϵ. It is a small negative gradient value, and all negative axis information will not be lost. This model uses $LeakyReLU$ to solve the problem of negative value loss in the nonlinear layer of $ReLU$. During the model training process, it was found that there were orders of magnitude differences between some parameters and other parameters, leading to a gradient explosion in the model. Therefore, this model first performed dimensionality normalization on the input data:

$$X' = \frac{X - mean(X)}{std(X)}, \tag{6}$$

where std is the internal standard deviation of a certain data item. This model uses various technical indicators such as "VMA", "VMACD", "BOLL","BIAS","ARBR" and so on, and timestamps as inputs, with transaction volume as labels and MSE loss as the objective function. After 500 iterations with a batch size of 32, the decrease in the objective function is minimized as depicted in Fig 7.

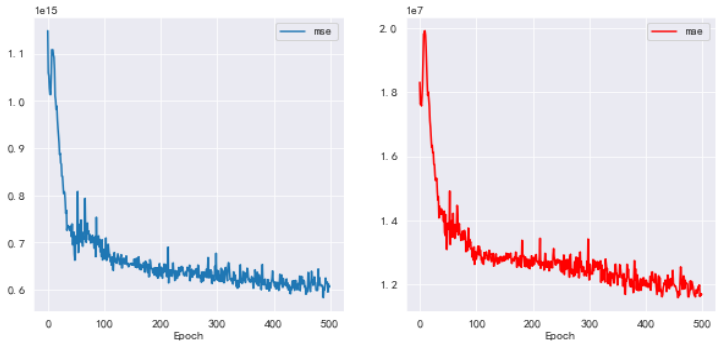

Figure 7: Decline of the objective function during model training.

Performing validation on the model using randomly selected test data: Expanding upon this statement, it involves assessing the model's performance and accuracy by applying it to a set of test data chosen at random. This process typically entails comparing the model's predictions to the actual labels (ground truth) provided by the test dataset to evaluate its effectiveness and reliability. The experimental results are as depicted in Fig 8.

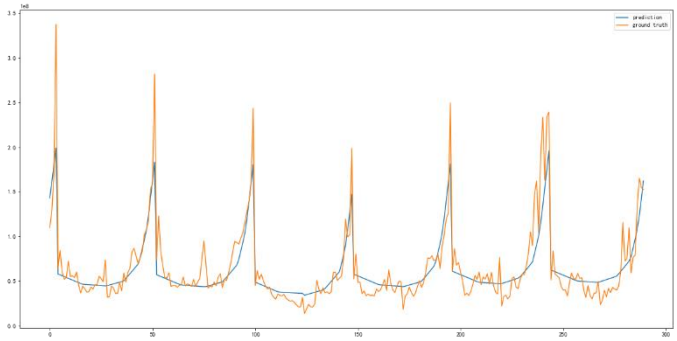

Figure 8: The model's predictions for randomly selected test data, where "prediction" refers to the predicted results and "ground truth" refers to the labels provided by the test dataset

Fig 9 illustrates the comparison between model predictions and actual outcomes. It is evident that the model can effectively utilize various metrics to make predictions about real-world scenarios. This indicates that the model demonstrates strong performance and accuracy on the test data, highlighting its reliability and accuracy.

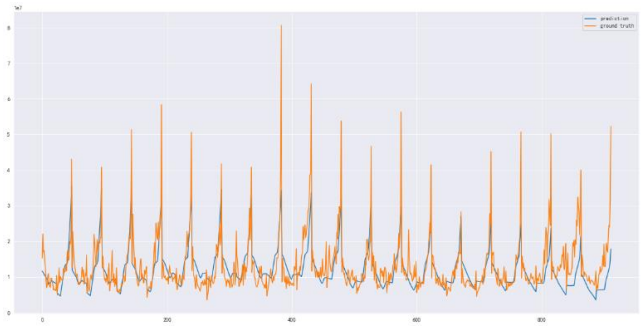

Figure 9: Comparison between model predictions and actual outcomes. "Prediction" refers to the predicted results, and "ground truth" refers to the labels provided in the test set.

VI. Conclusion

The comprehensive study findings suggest that stock prediction, as a complex and crucial task, is constrained by factors such as the instability, variability, and abrupt changes in the stock market, rendering direct forecasting of stock trends notably challenging. However, by considering objective factors related to the stock market, such as market conditions and macroeconomic development, valuable insights can be provided for stock prediction. In this study, through the analysis of different sectors' influence on the digital economy, we unveiled significant correlations among them, offering a fresh perspective for understanding market influencing factors. Utilizing a multi-layer perceptron neural network model built with TensorFlow, we successfully established an effective stock prediction model, which achieved an 87% prediction accuracy on the test dataset after extensive iterative training. Furthermore, the model demonstrated good versatility in predicting trading volumes and closing prices of stocks across various time periods, though further research is warranted to enhance its predictive performance. In summary, this study provides valuable insights and methodologies for the advancement of the stock prediction field, offering vital references for investment decisions and risk management.

Acknowledgments

This work is supported by the Guangxi Key Laboratory of Automatic Detecting Technology and Instruments (YQ22106) and the Innovation and Entrepreneurship Training Program for College Students of Guangxi (Project No. S202210595237, S202310595170).

References

[1]. Mintarya, L. N., Halim, J. N., Angie, C., Achmad, S., & Kurniawan, A 2023. Machine Learning Approaches In Stock Market Predic tion: A Systematic Literature Review. Procedia Computer Science, 216, 96-102.

[2]. Pahwa, K., & Agarwal, N. (2019), February. Stock Market Analysis Using Supervised Machine Learning. In 2019 International Conference On Machine Learning, Big Data, Cloud And Parallel Computing (Comitcon) (Pp. 197-200). IEEE.

[3]. Kumar, I., Dogra, K., Utreja, C., & Yadav, P. (2018, April). A Comparative Study Of Supervised Machine Learning Algorithms For Stock Market Trend Prediction. In 2018 Second International Conference On Inventive Communication And Computational Techn ologies (ICICCT) (Pp. 1003-1007). IEEE.

[4]. Htun, H. H., Biehl, M., & Petkov, N (2023). Survey Of Feature Selection And Extraction Techniques For Stock Market Prediction. Financial Innovation, 9(1), 26.

[5]. Al-Tamimi, H. A. H., Alwan, A. A., & Abdel Rahman, A. A. (2011). Factors Affecting Stock Prices In The UAE Financial Markets. Journal Of Transnational Management, 16(1), 3-19.

[6]. Allahawiah, S., & Al Amro, S. (2012). Factors Affecting Stock Market Prices In Amman Stock Exchange: A Survey Study. European Journal Of Business And Management, 4(8), 236-245.

[7]. Thakkar, A., & Chaudhari, K. (2021). A Comprehensive Survey On Deep Neural Networks For Stock Market: The Need, Challenges, And Future Directions. Expert Systems With Applications, 177, 114800.

[8]. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet Classification With Deep Convolutional Neural Networks. Advances In Neural Information Processing Systems, 25.

[9]. Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-Level Concept Learning Through Probabilistic Program Induct ion. Science, 350(6266), 1332-1338.