DOI: 10.53469/jtpes.2024.04(01).09

The Prediction Trend of Enterprise Financial Risk based on Machine Learning ARIMA Model

Xinqi Dong[1,*], Bo Dang[2], Hengyi Zang[3], Shaojie Li[4], Danqing Ma[5]

[1]Computer Science ,Independent research, Beijing, CN
[2]Computer Science ,Independent research,Fremont CA, US
[3]Big Data and Business Intelligence,Independent research, Shanghai, China
[4]Computer Technology ,Independent Reasearch, Beijing, CN
[5]Computer Science ,Independent Research, Beijing, CN
**Correspondence Author, dongxinqi@hotmail.com*

Abstract: ***The relevance of AI technology to abnormal prediction of corporate financial statements is reflected in its ability to identify and predict abnormal patterns in financial data through advanced algorithms. This predictive power is primarily based on machine learning and data mining techniques such as decision trees, neural networks, and deep learning. These techniques can analyze historical financial data and learn patterns and trends to effectively predict possible future anomalies, such as fraud, errors, or other irregularities. By identifying these anomalies in a timely manner, businesses can take preventive measures to reduce potential financial losses. The role of AI in the financial management of enterprises is reflected in its ability to process and analyze large amounts of complex data. AI technology can help companies automate cumbersome financial processes such as invoice processing and reimbursement management, increasing efficiency and accuracy. Based on AI deep learning algorithm, this paper uses ARIMA regression model to predict financial anomalies and financial development trends of enterprises, so as to help enterprises manage risk and make investment decisions, and better cope with financial risks and grasp investment opportunities.***

Keywords: Deep learning; ARIMA model; Abnormal financial statements; Risk prediction.

1. INTRODUCTION

Financial distress prediction is an important research direction in the field of financial management and investment management, because how or whether the financial situation of an enterprise will fall into financial distress is related to not only the strategy formulation and adjustment of the enterprise itself, but also the interests of its creditors or investors. The research on the judgment of enterprise's financial situation and the prediction of financial distress is especially of great theoretical and practical significance. The prediction of financial distress is to realize the accuracy and precision of the financial statements of enterprises through the analysis of the accounting statements publicly released by enterprises and the macroeconomic indicators released by the state, and to judge the overall financial situation of enterprises to predict the probability of financial distress in the future period of time[1]. The purpose of this paper is to predict the general financial anomalies and trends of enterprises through ARIMA model combined with artificial intelligence deep learning algorithm, aiming to propose a financial distress prediction method that can be widely used without the limitations of enterprise size, industry limitations, equity structure and other limitations. As a market entity, the evaluation of enterprise value is mainly reflected in its income and forecast. There are many indicators that can reflect corporate income, such as corporate free cash flow, EBIT and so on. However, different indicators have some drawbacks in some special cases, and using free cash flow as an indicator to forecast earnings can effectively avoid these drawbacks. Therefore, this paper chooses to evaluate the enterprise value by predicting the free cash flow of the enterprise.

2. MODEL AND ALGORITHM

With the rise of big data and artificial intelligence, machine learning has advantages in simulating specific characteristics of objects and processing complex and large amounts of data. By conducting multidimensional statistical analysis of big data and eliminating interfering information, high accuracy of corporate financial prediction can be obtained. Therefore, using machine learning method to study the factors of financial fraud has certain advantages

2.1 Support Vector Machine

Support vector machine (SVM) is a generalized linear classifier used to solve complex regression and classification problems. Based on the principle of maximizing the interval, the linear non-fractional data is extended to multidimensional space, and the hyperplane is divided to find the global optimal solution, and the generalization ability of the model is enhanced, so as to solve the statistical prediction of small samples and nonlinearity[2].

If the support vector machine is used to identify financial fraud, it is necessary to first find the point closest to the hyperplane in the sample points of fraud and non-fraud and maximize the distance from this point to the hyperplane, so as to distinguish the fraud sample from the non-fraud sample, and the determined hyperplane can be used as a classifier to judge whether the sample is fraudulent[3].

Given training sample (Xi1,Xi2.... ,Xik,Yi), constructs the objective function fx) so that it is as close to y as possible. Where Xi is the input vector and Yi is the output vector. The nonlinear mapping φ(x) needs to be introduced when selecting the optimal hyperplane:

$$f(x) = \omega^T \varphi(x) + b \tag{1}$$

In the formula, ω is the weight coefficient and b is the deviation. The optimal hyperplane problem is transformed into a quadratic problem:

$$\min Q = \frac{1}{2} ||W||^2 + C \sum_{i=1}^{n} (\beta_1 + \beta_2) \tag{2}$$

In the formula, Q is the objective value of optimization, w is the weight coefficient. Finally, the Lagrangian function is transformed into the dual form to obtain the optimal hyperplane. ai,ai*, is set as the Lagrange factor. K(xi,xj) is the kernel function, including linear kernel function, polynomial kernel function RBF kernel function, etc., the following regression function is obtained:

$$\begin{cases} f(x) = \sum_{i=1}^{n} \sum_{j=1}^{k} (\alpha_i - \alpha_i^*) K(x_i, x_j) \\ s.t. \begin{cases} \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 \\ 0 \leqslant \alpha_i \leqslant C \\ 0 \leqslant \alpha_i^* \leqslant C \end{cases} \end{cases} \tag{3}$$

2.2 Logistic Regression

Logistic regression, as a commonly used machine learning method, belongs to the factory regression model. The dependent variable of logistic regression model is binary classification variable[4]. The existing training set sample data is used to fit the model, and the obtained model is used to predict the test set. The formula is as follows:

$$f(x_i, \beta) = \frac{e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}}}{1 + e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}}}, i = 1, 2, \cdots, n \tag{4}$$

In the formula, Xi1, Xi2, Xik, represent the k characteristic attribute values of the i variable,β1,β2,βk, represent the regression coefficient of each characteristic attribute f(xi,β) represents the probability that a sample belongs to a positive class, y = 1 represents the financial abnormal state, then f(xi,β) represents the probability of predicting a financial anomaly.

2.3Neural Network

A neural network is a computing system that simulates the workings of the human brain and consists of a large number of interconnected nodes, or neurons[5]. These nodes process complex patterns of data by mimicking the way the human brain learns. In the prediction of financial anomalies, neural networks can identify and predict potential financial risks and anomalies. For example, by training a neural network to analyze historical financial data, such as profits, expenses, and balance sheets, it can learn to identify unusual patterns in financial statements, thereby helping companies to detect and respond to potential financial risks in a timely manner, such as fraud, financial distress, or declining profitability. In this way, neural networks become a powerful tool that enhances risk management and decision making.

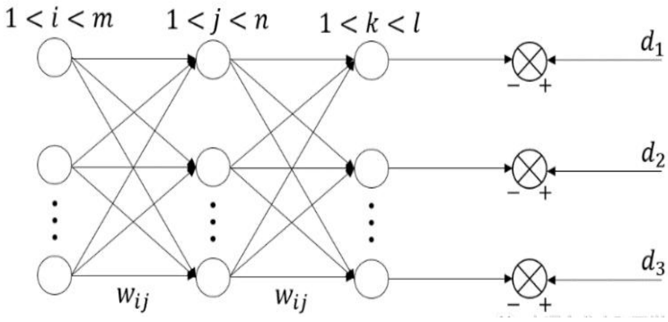

Figure 1: Three-layer neural network diagram of enterprise financial anomalies

2.4 ARIMA Model

The time series model (ARIMA) is built on the assumption that it is stable. Intuitively, if we think that a time series produces a particular behavior over time, there is a high probability that it will behave the same way in the future. Of course, without this feature, our quantitative analysis is useless, in ARIMA (p, d, q), AR is "autoregressive", p is the number of autoregressive terms[6]; MA is the "moving average", q is the number of terms of the moving average, and d is the number of differences (orders) made to make it a stationary sequence. These three parameters are statistical data set for seasonality, trend and noise. p is the autoregressive part of the model[7]. It allows us to incorporate the effects of past values into our models. Probability is called prior knowledge, popular point is to look at the past to know the future, if the past three days have been warm, tomorrow may be warm. Autoregressive term (p) : The AR condition is simply the lag of the dependent variable. For example, if P is equal to 5, then the prediction x (t) will be x (t-1)... (t-5).

d is the integrated part of the model. Including the amount of difference, intuitively speaking, if the temperature difference over the past three days is very small, there may be the same temperature tomorrow. Difference number (d) : The number of non-seasonal differences, i.e. in this case we use a first-order difference. Seasonal difference, first order, second order...

q is the moving average part of the model. This allows us to set the error of the model to a linear combination of the error values observed at previous time points in the past[8]. Moving average (q) : The MA condition is a delayed prediction error of the prediction equation. For example, if q is equal to 5, predict x (t) will be e (t-1)... e(t-5), e(i) is the difference between the moving average at the i th time and the actual value.

$$\begin{gathered} if\ d = 0,\ y_t = Y_t \\ if\ d = 1,\ y_t = Y_t - Y_{t-1} \\ if\ d = 2,\ y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ = Y_t - 2Y_{t-1} + Y_{t-2} \end{gathered} \tag{5}$$

Among them, ARIMA's prediction model can be expressed as: the predicted value of Y = the constant c and/or the weighted sum of Y for one or more recent times and/or the prediction error for one or more recent times.

Therefore, ARIMA model is a time series analysis method in the forecast of enterprise financial situation, which has significant advantages in the prediction of financial anomalies. The model predicts future values by considering the autoregressive (AR), difference (I), and moving average (MA) parts of the historical data. In the field of finance, the ARIMA model is particularly suitable for analyzing and forecasting time series data such as economic indicators, stock prices, and sales. Its advantage lies in its ability to efficiently process and analyze non-stationary time series data, thus providing more accurate predictions. For example, by applying the ARIMA model to a company's historical financial statement data, it is possible to predict the future trend of its revenue, expenses, or other financial indicators, thereby helping the company to identify possible financial anomalies, such as sudden revenue declines or cost increases, in advance, so that timely measures can be taken to avoid potential risks[9]. This makes the ARIMA model an important tool in enterprise financial analysis and risk management.

3. METHODOLOGY

3.1 Data Collection and Preprocessing

In order to test the financial situation of the enterprise predicted by the ARIMA model, the historical financial data of the enterprise, such as profit, revenue, expenditure, assets and liabilities, is first collected. Data cleaning is then carried out, including removal of missing values, outlier processing and data formatting to ensure data quality and consistency. Secondly, in order to make time series data conform to the requirements of ARIMA model (stationarity), it is necessary to conduct stationarity test of data, such as using unit root test[10]. If the data is not stationary, differential processing is required, i.e. the value of the previous point in time is subtracted from the current value to eliminate trend and seasonal effects.

This paper will conduct empirical research on 724 sample data in the recent financial statements of an enterprise, draw relevant conclusions and provide a judgment basis for investors and decision makers, and confirm the characteristics and performance of ARIMA model in deep learning algorithms.

Table 1: ADF test results of enterprise financial data difference

		t−Statistic	Prob.*
Augmented Dickey−Fuller test statistic		−10.04560	0.0000
Test critical values:	1% level	−3.439255	
	5% level	−2.865360	
	10% level	−2.568861	

It can be seen from Table 1 that the p-value of the test t statistic is 0, so the null hypothesis is rejected, that is, the first-order difference sequence is a white noise sequence and is stationary.

3.2 The Establishment of Financial Model

To sum up, since the first-order difference of the data in this paper is a stationary time series, the ARIMA(P,d,q) model is selected in this paper.

1) Determination of model parameters. First, we know from the above analysis that d=1. Secondly, according to the partial autocorrelation coefficient and autocorrelation coefficient, both are third-order truncated. So we know that P=3 and g=3. Finally, when P=3 and g=3, the AIC value is the smallest, which is 9.630019. According to the principle of information minimization, it can be determined that P=3 and g=3. Therefore, the ARIMA(3,1.3) model is established, and the parameter estimation results of the model are shown in Table 2:

Table 1: Financial data simulation results

Variable	Coefficient	Std.Error	t−Statistic	Prob.
C	0.050248	1.221591	0.041133	0.9672
AR(3)	−0.370233	0.199821	−1.852823	0.0643
MA(3)	0.509265	0.185167	2.750304	0.0061

To test the validity of the model. The white noise test is performed on the residual sequence of the estimated model. According to the residual correlation graph, the residual sequences of both ARIMA(3,1,3) models are autocorrelated. It's still a partial correlation. Both fall within two standard deviations. The P-values are all greater than 0.05, indicating that the random error term generated by ARIMA(3.1.3) model in this paper is white noise sequence. The model passed the test. It works.

3.3 Model Simulation Conclusion

This paper makes a hypothesis experiment on whether the ARIMA model can effectively predict the anomalies of enterprise financial data and statements, and verifies that the data results can effectively predict the risks of financial data. Therefore, the well-fitted ARIMA model can be used to predict future financial data, and the predicted results can be compared with the actual data to identify potential financial anomalies. For example, if actual revenue is much lower than forecast, it can indicate underlying financial problems.

However, in current financial forecasting algorithms, there are a variety of methods used to predict the trend of time series, each with its own advantages. But there are also some shortcomings[11]. The ARIMA model adopted in this empirical analysis can be used as a decision-making tool to evaluate financial investment. However, it should be noted that in the model prediction, the error between the predicted result and the actual occurrence value is relatively small in the first five trading days in the future, and the error ratio tends to increase after that. To some extent, it shows that this model is more suitable for short-term prediction of stock price, while other models need to be established for long-term prediction

3.4 Characteristics of ARIMA Model in Financial Forecasting

Advantages: The model is very simple, only requires endogenous variables and does not need to rely on other exogenous variables.

Cons:

1) The time series data is required to be stationary or stable through differencing.

2) In essence, it can only capture linear relationships, but not non-linear relationships.

Note that using ARIMA model to predict time series data must be stable, if the data is unstable, it will not be able to capture the rule. For example, the reason that stock data cannot be predicted with ARIMA is that stock data is volatile and often subject to policy and news.

2) Determine that time series data is stable.

A random variable of a time series is stable if and only if all of its statistical properties are time independent (are constants about time).

How to judge:

Stable data is no trend, no periodicity; That is, its mean has a constant amplitude on the time axis, and its variance tends to be the same stable value on the time axis. The Dickey-Fuller Test is usually used for hypothesis testing

4. CONCLUSION

Through the application of deep learning algorithm and ARIMA regression model, this paper successfully demonstrates the effectiveness and potential of AI technology in the field of enterprise financial anomaly prediction. Using these technologies, businesses can more accurately predict abnormal patterns in financial data, allowing them to detect and respond to potential financial risks, such as fraud, errors or other irregularities, in a timely manner[10-11]. In addition, the application of ARIMA model in financial data analysis and risk management shows that it can effectively process non-stationary time series data to provide more accurate forecasts for enterprises. With the continuous development of big data and artificial intelligence, it is expected that AI will further enhance its forecasting capabilities and application scope in the field of corporate financial risk forecasting. AI algorithms, particularly deep learning, will be able to process more complex data sets and provide deeper insights to predict and manage a variety of financial risks. In addition, as the algorithm continues to improve and optimize, it is expected to be able to improve the accuracy and efficiency of predictions while reducing the reliance on specialized knowledge.

Therefore, many enterprises can widely develop AI intelligent algorithms in analyzing corporate financial statements or realizing financial risk prediction, and should continue to invest in AI technology and data science capabilities to better use these tools for risk management and decision support, while ensuring high quality and secure data[12]. The performance of an AI model depends largely on the quality of the input data.

In this study, we have introduced a novel methodology tailored for the mechanical domain, specifically focused on predicting the position of a robot based on sensor data collected from the floor. The approach leveraged a Convolutional Neural Network (CNN) model, which proved highly effective in extracting and learning spatial features from the sensor data.

ACKNOWLEDGEMENT

In the process of compiling this paper, I am especially grateful to Qishuo Cheng for his research work, whose in-depth exploration in the field of financial data analysis and artificial intelligence has provided valuable reference and inspiration for my research. Cheng's research not only enhanced my theoretical foundation, but also provided a feasible methodology for me to explore the application of corporate financial risk prediction. I deeply respect and thank Cheng for his contributions in data processing, deep learning, and the application of ARIMA models. Cheng's work has been a major influence on the conception and realization of this article, for which I am deeply grateful.

REFERENCES

[1] Pan, Linying, et al. “Research Progress of Diabetic Disease Prediction Model in Deep Learning”. Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 15-21, doi:10.53469/jtpes.2023.03(12).03.
[2] Tianbo, Song, Hu Weijun, Cai Jiangfeng, Liu Weijia, Yuan Quan, and He Kun. "Bio-inspired Swarm Intelligence: a Flocking Project With Group Object Recognition." In 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 834-837. IEEE, 2023.DOI: 10.1109/mce.2022.3206678.
[3] Xinyu Zhao, et al. “Effective Combination of 3D-DenseNet’s Artificial Intelligence Technology and Gallbladder Cancer Diagnosis Model”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 81-84, https://doi.org/10.54097/iMKyFavE.
[4] Shulin Li, et al. “Application Analysis of AI Technology Combined With Spiral CT Scanning in Early Lung Cancer Screening”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 52-55, https://doi.org/10.54097/LAwfJzEA.
[5] Liu, Bo & Zhao, Xinyu & Hu, Hao & Lin, Qunwei & Huang, Jiaxin. . Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. Journal of Theory and Practice of Engineering Science. 2023,3: 36-42. 10.53469/jtpes.2023.03(12).06.
[6] Yu, Liqiang, et al. “Research on Machine Learning With Algorithms and Development”. Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 7-14, doi:10.53469/jtpes.2023.03(12).02.
[7] Xin, Q., He, Y., Pan, Y., Wang, Y., & Du, S. . The implementation of an AI-driven advertising push system based on a NLP algorithm. International Journal of Computer Science and Information Technology, 2023,1(1): 30-37.0
[8] Zhou, H., Lou, Y., Xiong, J., Wang, Y., & Liu, Y. . Improvement of Deep Learning Model for Gastrointestinal Tract Segmentation Surgery. Frontiers in Computing and Intelligent Systems, 2023,6(1): 103-106.6
[9] Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma. . Frontiers in Computing and Intelligent Systems,2024, 6(3): 127-131. https://doi.org/10.54097/zJ4MnbWW.
[10] Zhang, Q., Cai, G., Cai, M., Qian, J., & Song, T. . Deep Learning Model Aids Breast Cancer Detection. Frontiers in Computing and Intelligent Systems, 2023,6(1): 99-102.3
[11] Xu, J., Pan, L., Zeng, Q., Sun, W., & Wan, W. . Based on TPUGRAPHS Predicting Model Runtimes Using Graph Neural Networks. Frontiers in Computing and Intelligent Systems, 2023,6(1): 66-69.7
[12] Wan,Weixiang, et al. "Development and Evaluation of Intelligent Medical Decision Support Systems." Academic Journal of Science and Technology .2023,8(2): 22-25.
[13] Li Mei, GAN Xiaorong, LIU Xinle. Application of ARIMA Model to Stock price prediction and its Fourier Correction [J]. Journal of Yunnan Normal University (Natural Science Edition), 2011,31 (05) : 50-55.
[14] Tian, M., Shen, Z., Wu, X., Wei, K., & Liu, Y. . The Application of Artificial Intelligence in Medical Diagnostics: A New Frontier. Academic Journal of Science and Technology,2023, 8(2): 57-61.7
[15] Shen, Z., Wei, K., Zang, H., Li, L., & Wang, G. . The Application of Artificial Intelligence to The Bayesian Model Algorithm for Combining Genome Data. Academic Journal of Science and Technology, 2023,8(3): 132-135.2

[16] Zheng He, et al. “The Importance of AI Algorithm Combined With Tunable LCST Smart Polymers in Biomedical Applications”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 92-95, https://doi.org/10.54097/d30EoLHw.
[17] Prediction of Atmospheric Carbon Dioxide Radiative Transfer Model based on Machine Learning. . Frontiers in Computing and Intelligent Systems, 2024,6(3): 132-136. https://doi.org/10.54097/ObMPjw5n
[18] Liu, Y., Duan, S., Shen, Z., He, Z., & Li, L. . Grasp and Inspection of Mechanical Parts based on Visual Image Recognition Technology. Journal of Theory and Practice of Engineering Science, 2023,3(12): 22-28.1.
[19] Xinyu Zhao, et al. “Effective Combination of 3D-DenseNet’s Artificial Intelligence Technology and Gallbladder Cancer Diagnosis Model”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 81-84, https://doi.org/10.54097/iMKyFavE.
[20] Liu, B. . Based on intelligent advertising recommendation and abnormal advertising monitoring system in the field of machine learning. International Journal of Computer Science and Information Technology, 2023,1(1): 17-23.