Designing a profit and loss Prediction model for health companies using data mining

Ali Abdolahi^{1*}, Vali Nowzari², Ali Pirzad³, Seyed Ehsan Amirhosseini⁴

¹Sama Technical and Vocational Training College, Islamic Azad University, Bandar Mahshahr Branch, Bandar Mahshahr, Iran

²Department of Physical Education, Islamic Azad University, Arsanjan Branch, Arsanjan, Iran

³Department of Government Management, Islamic Azad University, Yasooj Branch, Yasooj, Iran

⁴Department of Sport Management, Islamic Azad University, Yasooj Branch, Yasooj, Iran

Article Info

Article type:
Research

Article History:

Received: 2021-05-10

Accepted: 2021-10-13

Published: 2021-10-16

* Corresponding author:

Ali Abdollahi

Sama Technical and Vocational
Training College, Islamic Azad
University, Bandar Mahshahr
Branch, Bandar Mahshahr, Iran

Email: abdolahi1113@gmail.com

Keywords:

Data Mining Forecast
Financial
Differential Analysis
Logistic Models

ABSTRACT

Introduction: Health companies need investment for development. Due to the high risk of their activities, it is very difficult to attract investment for this field, but this lack of financial resources leads to the failure of these companies, so providing a model for predicting profits and losses in companies is very important and functional.

Material and Methods: In this study, a combination of two logistic regression algorithms and differential analysis were used to design a profit and loss forecasting model. Also, the information of 20 companies in the field of health was used to evaluate the proposed model. 10 profitable companies and 10 loss-making companies were selected and for each company, nine variables independent of the financial information of these companies were collected.

Results: The designed prediction model was implemented on the data in this study. To do this, the data were divided into two sets: training and testing. The prediction model was implemented on training data and evaluated by test data and reached 99.65% sensitivity, 94.75% specificity and 96.28% accuracy. The proposed model was then compared with the methods of decision tree C4.5, Bayesian, support vector machine, nearest neighborhood and multilayer neural network and it was found to have a better output.

Conclusion: In this study, it was found that the risk in the field of health investment can be reduced, so the profit and loss situation of health companies can be predicted with appropriate accuracy. It was also found that the combination of logistic regression and differential analysis algorithms can increase the accuracy of the prediction model.

Cite this paper as:

Abdolahi A, Nowzari V, Pirzad A, Amirhosseini SE. Designing a profit and loss Prediction model for health companies using data mining. Front Health Inform. 2021; 10: 94. DOI: [10.30699/fhi.v10i1.305](https://doi.org/10.30699/fhi.v10i1.305)

INTRODUCTION

Investing means injecting financial resources with the expectation of profit from it. More precisely, investing is the commitment of money or investment to the financial purchase of other assets or assets for the benefit of beneficial returns in the form of interest, dividends, or appreciation of the value of the assets [1, 2]. The role and importance of investment in the process of economic growth and development of societies has been emphasized in most theories of economic growth and development [3, 4]. Among the most important issues raised in the field of financial management, investment and investment confidence. One of the issues that can help in investing decision

making is the existence of appropriate tools and models for assessing the financial condition and condition of organizations [5]. One of the most important issues in investing is the probability of profit and loss of companies [6, 7]. Given that today no venture investments specifically addresses the issue of digital health, so designing a forecasting model to calculate the amount of profits and losses can be very effective in expressing and investing with the help of a forecasting model that can be risky. Minimize yourself

Knowledge-based companies in the field of health and water startups need investment for their development. Due to the high risk of their activities, it

is very difficult to attract investment for this virtual class. Knowledge-based projects in the field of health and startups that this failure is lagging behind the growth of digital health in the country, so providing a model for predicting profits and losses in companies is very important.

Many studies have been done in the field of profit and loss forecasting models of companies. However, during the review of the literature in the field of bankruptcy prediction of health companies, no specific study was found

In 2008, Pindado et al. used a study of the 1992 to 2006 financial balances in a study of the G7 countries. They used econometric models and cross-sectional analysis in their research and concluded that variables such as profitability, financial expenditure and accumulated profit are of constant importance. The accuracy rate of the model for different years is about 87% [8].

Another study was conducted in 2013 by Bastes et al. in the field of predicting the bankruptcy of listed companies. In this study, regression was used to evaluate the stock market data of three countries, Australia, Brazil and Turkey, and it was shown that the regression algorithm can determine the variables that affect the bankruptcy forecast of companies. Before performing regression, correlation coefficient was used to determine the linearity of data and it was found that all variables defined in this study are directly related to the bankruptcy of companies [9].

In another study in 2017, Altman et al. developed a model called the Z-Score, which showed that the model has the power to evaluate financial companies. In this model, it was proved that five parameters of investment in production, retained earnings, pre-interest income and taxes, stock market value and sales are effective in the failure rate of an investment company [10].

In another study in 2019, Theo Wu et al. compared three different feature selection models using three-year data from listed companies, and found that the F-Score model is the best model for selecting these features. Then, the backup vector machine algorithm was implemented on the data and accurately predicted the bankruptcy rate of stock companies between 66 and 84% [11].

In 2013, Hosseini et al. used a combination of two logistic regression algorithms and a decision tree to predict the bankruptcy of companies listed on the stock exchange. The statistical population of this study was the financial statements of companies in Tehran Stock Exchange during the years 1378 to 1389. In this study, the algorithm with more than 95% accuracy was able to predict the bankruptcy of these companies [12].

Given that no study has been found on the

development of profit and loss forecasting models for health companies in Iran, researchers decided to use forecasting models to find a solution for forecasting profits and losses of companies active in health. In this study, a combination of logistic regression and linear differential analysis (LDA) algorithms was used for the forecasting model and the data of health companies are used for evaluation. The reason for choosing this field is that it therefore increases the investment risk.

MATERIAL AND METHODS

The study population in this study were accepted companies in the field of health that operated between 2011 and 2018; in this study, random sampling was performed, which includes 20 companies in the field of health information and medical equipment. 10 companies were loss-making and 10 companies were profitable, the criterion of loss was Article 141 of the Commercial Code. Therefore, using the financial statements of the companies, the ratio of accumulated profit (loss) to their capital in each of the mentioned years was calculated, then the companies whose accumulated loss was at least half of the company's capital were determined. And among the ten profitable companies, 10 companies were selected as a sample [13]. Meanwhile, the companies that were included in Article 141 more than once from 2011 to 2018, their first year of inclusion was considered. Nine independent variables have been selected to predict profit and loss. Table 1 shows the independent variables.

Table 1: Independent Selected Variables [4]

Row	Variables
1	Ratio of working capital to total assets
2	Ratio of accumulated profit to total assets
3	Ratio of profit before tax on total assets
4	Ratio of net profit to current debt
5	The ratio of total debt to total assets
6	Ratio of current debt to total assets
7	Ratio of current assets to current debt
8	Non-specific profit to sales ratio
9	Ratio of working capital to total debt

Suggested model

The proposed model is implemented on the data in this study.

Study data were divided into two sets: training and test. The model is designed by training data and evaluated by model test data. The structure of data division into two sets of training and testing uses the K-Fold method [14]. The structure of the K-Fold method is shown in Fig 1.

Fig 1: Structure of the K-Fold method with a value of K equal to five

The value of K is considered for all three methods 10 and the problem is executed 1000 times and the three parameters of sensitivity, specificity and accuracy are measured 1000 times, then the average of all three parameters is considered as output. Be. The reason why the algorithm is run 1000 times is to be able to observe different modes of training data selection and testing in the whole data. The structure of the proposed model is stated below.

The proposed prediction model is a combination of logistic regression algorithm and differential analysis algorithm. The structure of the model is that it is the basis of logistic regression algorithm and differential analysis algorithm is used to optimize the logistics algorithm. In logistic regression algorithm, equation (1) is used to calculate p.

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)} \quad (1)$$

The logistic regression algorithm is very strict. If the value of P is less than 0.5, the regression considers it to be zero. Its operation is shown in Equation (2). If it is greater than 0.5, it may be 0.49 percent of the output. This value is considered 0, which is probably a very high error. Its output is one. The points between 0.3 to 0.7 of differential analysis algorithm are used. In equation (3) the output of the hybrid algorithm is expressed. Fig 2 shows the graphical structure of the proposed model.

$$Y = \begin{cases} 0 & p < 0.5 \\ 1 & p \geq 0.5 \end{cases} \quad (2)$$

$$Y = \begin{cases} 0 & p \leq 0.3 \\ 1 & p \geq 0.7 \\ \text{LDA}(X) & 0.3 < p < 0.7 \end{cases} \quad (3)$$

Whenever the output value of the logistic regression is between 0.3 and 0.7, the output is not specified by the logistic regression and that value is differentiated by the algorithm. Points less than that are considered zero and values greater than that are considered one. Fig 3 is a flowchart of the proposed model.

Fig 2: Structure of the proposed algorithm

Fig 3: Flowchart of the proposed model

RESULTS

The most important task in a model is to evaluate that model and include a set of techniques for reviewing, recording and studying and improving that algorithm. Performance evaluation is an algorithm that will be one of the most important research tools for model performance. Calculations and results of job evaluation express the strengths, weaknesses and quality of the model. Evaluation is very important because it is an important task in model effectiveness and causes the properties of algorithms to be measured and their efficiency to be estimated. With the help of evaluation, it can be shown that the proposed model has done its job better than any of the other models. At this stage, the proposed model is evaluated.

Evaluation parameters

The evaluation parameters in the proposed algorithm can be calculated using the relationship between the actual value and the prediction value in the scatter matrix [15]. The structure of the dispersion matrix is shown in Table 2.

		Predictive value	
		Profit	Loss
Real amount	Profit	True Positive(TP)	False Positive(FP)
	Loss	False Negative(FN)	True Negative(TN)

Table 2: Scattering matrix structure [15]

TP: The number of companies in the healthcare sector that are correctly recognized as profitable.

TN: The number of loss-making health companies that are correctly identified as harmful.

FP: The number of for-profit companies that are incorrectly recognized as loss-making.

FN: The number of loss-making companies that are incorrectly recognized as profitable.

Equations (4), (5) and (6) are used to compare the proposed model with other proposed models.

$$\text{Sensitivity} = TP / (TP + FN) \quad (4)$$

$$\text{Specificity} = TN / (FP + TN) \quad (5)$$

$$\text{Accuracy} = (TP + TN) / (TP + FN + FP + TN) \quad (6)$$

Sensitivity, specificity and accuracy are important parameters for evaluating the proposed algorithms. Sensitivity means what percentage of the companies correctly recognized the profit, and specificity means what percentage of the companies correctly recognized the loss, and accuracy means correct diagnosis for both groups (Loss-making and profitable companies).

In this step, the combination algorithm, which includes a combination of two logistic regression algorithms and differential analysis algorithm, is evaluated. This part is repeated a thousand times and its mean and standard deviation are reported in Table 3 and Fig 4.

As can be seen in Table 3 of Fig 4, the algorithm has an accuracy of 96.28% and the standard deviation in the training data has a small value and the standard deviation in the test data has a value of less than 6%.

Table 3: Mean and standard deviation for three parameters of sensitivity, specificity and accuracy with the proposed algorithm

Evaluation parameter		Number of runs	Average number of runs	Standard deviation of execution number
Sensitivity	Training	1000	99.58	0.13
	Test	1000	99.65	0.95
Specificity	Training	1000	95.34	0.74
	Test	1000	94.75	5.32
Accuracy	Training	1000	98.01	0.29
	Test	1000	96.28	1.85

Fig 4: Mean and standard deviation for three parameters of sensitivity, specificity and accuracy with the proposed algorithm

Comparison of the proposed algorithm with other

algorithms

The performance of the proposed algorithm was compared with the methods of C4.5 decision tree, Bayesian, support vector machine, nearest neighborhood and multilayer neural network. Weka software was used to execute these five algorithms. Table 4 shows a comparison of the five algorithms with the logistic regression, differential analysis and proposed algorithms.

Table 4: Comparison of five algorithms with two logistic regression algorithms and the proposed algorithm

Algorithm	Sensitivity	Specificity	Accuracy
Decision Tree	92.93	90.76	91.81
Bayesian	95.4	89.91	93.25
Support Vector Machine	95.34	90.35	93.58
Nearest Neighbour	94.9	85.86	91.53
Artificial Neural Network	98.26	93.22	95.36
Logistic Regression	99.34	89.09	95.61
Differential Analysis	93.74	94.34	94.62
Proposed Algorithm	99.65	94.75	96.28

According to Table 4, the proposed algorithm is the best output for the bankruptcy detection problem. The proposed algorithm has better output in terms of three parameters of sensitivity, specificity and accuracy.

DISCUSSION

One of the concerns of investors and creditors, investing in healthcare companies, is that due to poor performance, they eventually go bankrupt and as a result, the principle of investment and their expected profits are lost. Using the model extracted in this research can help investors in choosing a company to invest.

The results obtained from this study show that the proposed model with the variables of the ratio of working investment to total assets, the ratio of accumulated profit to total assets, the ratio of pre-tax profit to total assets, the ratio of net profit to current debt, the ratio of total debt to Total assets, the ratio of current debt to total assets, the ratio of current assets to current debt, the ratio of non-specific profit to sales, the ratio of working investment to total debt have the power to predict the bankruptcy of companies in the field of health. Therefore, the first hypothesis of the research is confirmed. This hypothesis has also been confirmed by studies [3, 4]. In this research, the prediction model has reached 96.28% accuracy, which shows that the model has a high accuracy and the second hypothesis has been proven. The hybrid algorithm performed 0.67%

better than the logistic regression algorithm and the third hypothesis was proved.

The proposed model has a higher accuracy than [10-12] and has been able to predict the bankruptcy of companies active in the field of health with very good accuracy. Sensitivity is a very important parameter in the forecast model, which in this study has a line. It is less than 0.5%, which indicates that the proposed model is suitable for similar tasks [10-12].

One of the limitations of this study compared to similar studies [10, 11] was small sample size because companies in the field of technology in the country are very limited.

Bankruptcy rate is an important economic indicator in any society. In most developed countries, the central bank uses bankruptcy forecasting models to forecast the status of companies for lending or investing to prevent such crises from occurring in the country by implementing appropriate strategies. The results of this research can be used by investors in the field of health.

One of the important limitations of this study is the lack of independent variables in this study because there is little information about health companies, so it is suggested that the information of foreign companies be added to increase the efficiency of the model. The limitations of the questionnaire were designed and provided to these companies to design a more appropriate model.

In this research, it was found that the profit and loss situation of health companies can be predicted with appropriate accuracy and with the help of it, it is possible to plan for investment in these companies. It is also clear that the combination of logistic

regression and differential analysis algorithms can provide a predictive model with good accuracy.

CONCLUSION

Based on the predicting model used and the results obtained, the main hypothesis of this research is confirmed, in other words, profit or loss forecasting is possible by using the combination of financial ratios in health companies listed on the Tehran Stock Exchange. Based on this, first, the combination of financial ratios can be used to provide a profit forecasting model, which indicates the information content of financial ratios. Second, the research findings confirm the testability of the combination of financial ratios to predict profit or loss, although the assumption of linearity in the design of the model can be considered a fundamental limitation. Third, the combination of financial ratios has a higher profitability than individual financial ratios, so that it is also widely used in predicting bankruptcy models.

AUTHOR'S CONTRIBUTION

All authors contributed to the literature review, design, data collection and analysis, drafting the manuscript, read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE

This article is the result of a research project belonging to Sama Faculty of Islamic Azad University with the code 97/142/3499.

REFERENCES

- Investment [Internet]. 2018 [updated: 11 Jul 2019; cited: 20 Apr 2021]. Available from: <https://en.wikipedia.org/w/index.php?title=Investment&oldid=905855129>
- Alabi KO, Abdulsalam SO, Ogundokun RO, Arowolo MO. Credit risk prediction in commercial bank using chi-square with SVM-RBF. International Conference on Information and Communication Technology and Applications. Springer; 2020.
- Ismaili SA, Abraham M, Mirzaei K. Investigating the impact of investment and entrepreneurship infrastructures on increasing economic growth and national productivity of Iran. International Conference on Entrepreneurship, Creativity and Innovation. Kharazmi Institute of Science and Technology; 2016.
- Rehman SU, Baloch H, Qayyum A. Anticipating bankruptcy for non-financial sector firms: A case of Pakistan. Journal of Accounting and Finance in Emerging Economies. 2021; 7(2): 401-9.
- Liu J, Wu C. Hybridizing kernel-based fuzzy c-means with hierarchical selective neural network ensemble model for business failure prediction. Journal of Forecasting. 2019; 38(2): 92-105.
- Ghazanfari M, Rahimi Kia I, Askari A. Predicting bankruptcy of companies based on hybrid smart systems. Financial Accounting and Auditing Research. 2016; 10(37): 159-93.
- Scherger V, Terceño A, Vigier H. A systematic overview of the prediction of business failure. International Journal of Technology, Policy Management. 2019; 19(2): 196-211.
- Mangena M, Priego AM, Manzaneque M. Bank power, block ownership, boards and financial distress likelihood: An investigation of Spanish listed firms. Journal of Corporate Finance. 2020; 64: 101636.
- Pattnaik D, Hassan MK, Kumar S, Paul J. Trade credit research before and after the global financial crisis of 2008: A bibliometric overview. Research in International Business and Finance. 2020; 54:

- 101287.
10. Altman EI, Iwanicz-Drozdowska M, Laitinen EK, Suvas A. Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Journal of International Financial Management & Accounting*. 2017; 28(2): 131-71.
 11. Vu LT, Vu LT, Nguyen NT, Do PTT, Dao DP. Feature selection methods and sampling techniques to financial distress prediction for Vietnamese listed companies. *Investment Management and Financial Innovations*. 2019; 16(1): 276-90.
 12. Fan S, Liu G, Chen Z. Anomaly detection methods for bankruptcy prediction. *International Conference on Systems and Informatics*. IEEE; 2017.
 13. Nabavi A, Moses A, Mahdavi Farrokh Abadi S. Predicting bankruptcy of companies using logit model. *Journal of Financial Engineering and Portfolio Management*. 2010; 5: 55-81.
 14. Seref O, Kundakcioglu OE, Pardalos PM. *Data mining, systems analysis, and optimization in biomedicine*. American Institute of Physics; 2007.
 15. Han J, Pei J, Kamber M. *Data mining: Concepts and techniques*. Elsevier Science; 2011.