Constructing a Financial Risk Early Warning Model for Chinese Public Hospitals Based on Machine Learning

Xi Zhao¹ & Bing Lu¹

¹ Business School, Xianda College of Economics & Humanities Shanghai International Studies University, Shanghai, China

Correspondence: Xi Zhao, Business School, Xianda College of Economics & Humanities Shanghai International Studies University, Shanghai, China. Tel: 86-21-5127-8000. E-mail: 1410070@xdsisu.edu.cn

Received: January 2, 2024   Accepted: January 27, 2024   Online Published: February 5, 2024
doi:10.5539/ijef.v16n3p64 URL: https://doi.org/10.5539/ijef.v16n3p64

Abstract

In today’s increasingly complex healthcare environment, China’s public hospitals face enormous financial challenges. The high degree of uncertainty and suddenness of financial risks make public hospitals need more sophisticated and real-time financial risk early warning mechanisms. To address this challenge, machine learning algorithms are introduced as a powerful tool to construct more accurate and efficient financial risk early warning models. The purpose of this dissertation is to summarize the recent research progress in constructing financial risk early warning models for Chinese public hospitals based on machine learning algorithms. The establishment of financial risk early warning models can not only help hospital management better understand the financial situation, but also identify potential risks in advance, which can provide powerful support for timely adjustment of strategies and countermeasures.

Keywords: machine learning, risk early warning, predictive modeling, Chinese public hospitals

1. Introduction

1.1 Background

Against the backdrop of China’s evolving healthcare system and increasing demand for healthcare, public hospitals are responsible for a wide range of healthcare responsibilities as the primary healthcare providers. However, public hospitals are facing increasingly significant financial challenges, which are influenced by a variety of factors, including fluctuations in demand for healthcare services, adjustments in healthcare policies, and reforms in the healthcare insurance system.

Traditional financial management methods are often difficult to cope with these changing factors, and the financial situation of public hospitals has become more complex and uncertain. In the face of this situation, the establishment of an efficient financial risk early warning system has become an urgent need for public hospital management. This system not only needs to be sensitive to detect financial problems in a timely manner, but also needs to be predictive to gain insight into potential financial risks in advance.

Against this challenging background, machine learning algorithms have emerged to provide an innovative approach to financial management for public hospitals. By analyzing large-scale financial data, machine learning algorithms are able to identify potential patterns and trends, thus providing more accurate and intelligent support for financial decision-making. Therefore, the application of machine learning algorithms to construct financial risk early warning models for Chinese public hospitals has become a hotspot and focus of attention in current research.

The Nature of Hospital Financial Challenges: Public hospitals face multifaceted financial management challenges such as uneven medical services, changes in health insurance policies, and fluctuations in patients’ willingness to pay. The interaction of these factors leads to a high degree of complexity in financial decision-making, requiring more intelligent and flexible management tools.

The rise of machine learning: In recent years, the successful application of machine learning in various fields has attracted widespread attention. In the medical field, machine learning algorithms have achieved remarkable results in disease diagnosis and patient prediction. In order to better meet the challenges of hospital financial management, researchers have begun to introduce this powerful tool into the field of financial risk early warning,
with a view to constructing more intelligent and accurate models.

Motivation and significance of the study: Given the shortcomings of traditional financial management methods, this study aims to gain insight into the potential application of machine learning in financial risk early warning in public hospitals. By reviewing and analyzing the existing literature, we hope to gain a comprehensive understanding of the nature of financial risks faced by public hospitals, and to identify the advantages and challenges of machine learning algorithms in solving these problems.

1.2 Purpose and Application of Research

The purpose of this paper is to review and analyze past research on financial risk early warning in Chinese public hospitals, focusing on the practice and results of using machine learning algorithms to construct models. By synthesizing and analyzing the existing literature, we aim to provide public hospital management with more comprehensive financial information and make useful suggestions for future research and practice.

Machine learning is increasingly used in the medical field, including disease diagnosis, patient prediction, and drug development. In the field of financial management, the introduction of machine learning provides a new perspective for hospitals to improve the accuracy and efficiency of financial decision-making by analyzing complex financial data.

1.3 Domestic and International Research Status

In the modern healthcare system, the financial management of public hospitals has always been the focus of much attention. Dynamic changes in the demand for healthcare services, frequent adjustments in healthcare policies, and uncertainty in patients’ ability to pay are among the factors that make public hospitals face multilayered financial challenges. Traditional financial management methods rely on retrospective analysis of historical data, which is gradually becoming inadequate in today’s complex and fast-changing healthcare environment, and new intelligent means are needed to better adapt to the changing healthcare service needs and financial environment.

In recent years, the wide application of machine learning in the field of financial management has aroused great interest among researchers. By analyzing large-scale financial data, machine learning algorithms are able to identify patterns and trends hidden behind the data and improve sensitivity and accuracy to financial risks. In other industrial fields, such as finance and manufacturing, machine learning has been successfully applied to optimize decision-making, reduce risk and improve efficiency, which provides a strong theoretical basis for its application in the financial management of public hospitals. The successful application of machine learning algorithms in the healthcare sector provides ample support for their potential value in the financial management of public hospitals. In the medical field, machine learning has been widely used in disease prediction, patient classification, and drug development. These success stories show that machine learning has a promising future in processing medical data and improving the efficiency of medical services.

Established studies have achieved some results in financial risk early warning in public hospitals, with some focusing on constructing traditional financial models, while others have begun to explore the possibility of introducing machine learning algorithms. However, existing studies generally suffer from insufficient sample size, insufficiently systematic feature selection, and weak model interpretability. These limitations restrict the reliability and operability of the existing studies in practical applications. Despite the remarkable achievements of machine learning in the healthcare field, its application to financial risk early warning models in public hospitals still faces a series of challenges. These include the problem of feature selection, i.e., how to select important features related to financial risks from a large amount of data; data quality assurance, to ensure that the data input to the model is accurate, complete and reliable; and model interpretability, so that hospital managers can understand the decision-making process of the model.

1.4 Prospect

Future research can further deepen the application of machine learning algorithms in financial risk early warning in public hospitals. More advanced and applicable machine learning algorithms, such as deep learning and reinforcement learning, can be explored for the financial characteristics of hospitals in order to improve the accuracy and real-time financial prediction. Meanwhile, the integration of multi-source data, including financial data, patient consultation records, and health insurance data, can be considered in order to establish a more comprehensive and multi-dimensional financial risk early warning model. The integrated use of multi-source data will help capture potential financial risk signals more accurately and improve the comprehensiveness and reliability of the model.

Future research could move toward the development of actionable decision support tools that enable the model’s
predictions to provide practical advice directly to hospital management. This includes the design of user-friendly interfaces and the provision of real-time updated early warning information to help managers better formulate financial strategies. Customize the financial risk early warning model to take into account the size and geographical characteristics of different public hospitals. This will help the models to better adapt to the differences in different healthcare environments and improve the feasibility and effectiveness of their practical application. Considering the time-varying and dynamic nature of the healthcare service sector, future research can further enhance the real-time monitoring and feedback mechanism of the model. This means that the model should be able to flexibly adapt to new data and environmental changes, thus improving the adaptability and sustainability of the model.

Ultimately, the goal of future research should be to promote the integrated development of intelligent healthcare management, combining the financial risk early warning model with other healthcare information systems and hospital management platforms to form an efficient and intelligent healthcare service management system that provides comprehensive decision support for public hospitals.

2. Method

In this study, four machine learning algorithms, namely decision trees, support vector machines and random forests are used for the construction of early warning models.

2.1 Machine Learning

Decision Tree: A decision tree is a model for decision making based on a tree-graph structure, which is formed by dividing the data set into different subsets, each of which corresponds to a node of the tree, ultimately forming a tree structure. In a decision tree, each internal node represents a test for an attribute, each branch represents the result of the test, and each leaf node (also called a terminal node) stores a category label. A path from the root node to the leaf nodes allows new data to be categorized or predicted.

\[ H(T) = - \sum_{i=1}^{c} p_i \log(p_i) \]

Support Vector Machine: Support vector machine is a supervised learning algorithm for classification and regression. The main idea is to find an optimal hyperplane in the feature space that can separate samples of different classes and maximize the interval from the samples to the hyperplane. The key to a support vector machine is to find the support vectors, i.e., those sample points closest to the hyperplane that play a key role in determining the optimal hyperplane.

\[ f(x) = \text{sign}(w \cdot x + b) \]

Random Forest: Random Forest is an integrated learning method that performs classification and regression by combining multiple decision trees. It is a forest of multiple decision trees, each trained independently and averaged (for classification problems) or averaged (for regression problems) to improve the performance and robustness of the overall model.

\[ \hat{y}(x, \Theta_k), \quad k = 1, 2, ..., K \]

2.2 Data Including

Basic Financial Data: Total annual revenue, Cost of medical services, Cost-to-revenue ratio (cost of medical services/total annual revenue), Number of patient visits

Financial Health Status Label: Flags if financial risk is present as a target variable.

Patient Information: Number of patients, Number of visits/number of patients (average number of visits per patient), Patient age distribution

Medical service data: Type of medical service, Average service cost

Personnel costs and operational data: Number of medical staff, Cost per capita, Operating Costs

Management and decision-making data: Management Level, Decision-making response time

Geographic and demographic information: Economic status of the area where the hospital is located, Size of population served

Previous Financial Data: Trends in financial data over the past few years

2.3 Model Building Steps

Data Preparation: Collect financial data of the hospital, including total annual revenue, cost of medical services, number of patient visits, etc., as well as financial health status as the target variable.
Feature standardization: Standardize the features so that they have the same scale.

Data preprocessing: Handle missing values, outliers, and perform feature engineering to select appropriate features.

Data division: Divide the dataset into training and test sets.

Model Construction: Select appropriate index parameters to construct the model.

Model Training: Train the model using the training set.

Model Evaluation: Evaluate the model using a test set to calculate metrics such as accuracy, recall, etc.

3. Results

In this study, we explore the performance of three different machine learning models, Decision Trees, Random Forests and Support Vector Machines (SVMs), on financial risk prediction. The following is a detailed analysis of these three models in terms of accuracy, precision, recall, F1 score and AUC:

Decision tree modeling: Accuracy is 0.959, Precision (state 0 / state 1) is 0.967 / 0.500, Recall (state 0 / state 1) is 0.991 / 0.214, F1 score (state 0 / state 1) is 0.979 / 0.300, AUC is 0.603. The decision tree performs well in predicting a financial health state of 0, but has relatively low recall in state 1. The model may have some underreporting problems in the face of financial risk. See in Table 1 and Figure 1.

Random Forest Modeling is Accuracy is 0.966, Precision (state 0 / state 1) is 0.966 / 1.000, Recall (state 0 / state 1) is 1.000 / 0.042, F1 score (state 0 / state 1) is 0.983 / 0.080, AUC is 0.871. Random Forest is relatively good in terms of overall performance, but still faces the challenge of low recall in state 1, which means that the model may have some degree of underreporting in the face of financial risk. See in Table 2 and Figure 2.

<table>
<thead>
<tr>
<th>Table 1. Decision tree evaluation metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>Support</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Precision (Positive Predictive Value)</td>
</tr>
<tr>
<td>Recall (True Positive Rate)</td>
</tr>
<tr>
<td>False Positive Rate</td>
</tr>
<tr>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>F1 Score</td>
</tr>
<tr>
<td>Matthews Correlation Coefficient</td>
</tr>
<tr>
<td>Area Under Curve (AUC)</td>
</tr>
<tr>
<td>Negative Predictive Value</td>
</tr>
<tr>
<td>True Negative Rate</td>
</tr>
<tr>
<td>False Negative Rate</td>
</tr>
<tr>
<td>False Omission Rate</td>
</tr>
<tr>
<td>Threat Score</td>
</tr>
<tr>
<td>Statistical Parity</td>
</tr>
</tbody>
</table>

Note. All metrics are calculated for every class against all other classes.
Table 2. Random forest evaluation metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Class 0</th>
<th>Class 1</th>
<th>Average / Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>657</td>
<td>24</td>
<td>681</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.966</td>
<td>0.966</td>
<td>0.966</td>
</tr>
<tr>
<td>Precision (Positive Predictive Value)</td>
<td>0.966</td>
<td>1.000</td>
<td>0.967</td>
</tr>
<tr>
<td>Recall (True Positive Rate)</td>
<td>1.000</td>
<td>0.042</td>
<td>0.966</td>
</tr>
<tr>
<td>False Positive Rate</td>
<td>0.958</td>
<td>0.000</td>
<td>0.479</td>
</tr>
<tr>
<td>False Discovery Rate</td>
<td>0.034</td>
<td>0.000</td>
<td>0.017</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.983</td>
<td>0.080</td>
<td>0.951</td>
</tr>
<tr>
<td>Matthews Correlation Coefficient</td>
<td>0.201</td>
<td>0.201</td>
<td>0.201</td>
</tr>
<tr>
<td>Area Under Curve (AUC)</td>
<td>0.871</td>
<td>0.879</td>
<td>0.875</td>
</tr>
<tr>
<td>Negative Predictive Value</td>
<td>1.000</td>
<td>0.966</td>
<td>0.983</td>
</tr>
<tr>
<td>True Negative Rate</td>
<td>0.042</td>
<td>1.000</td>
<td>0.521</td>
</tr>
<tr>
<td>False Negative Rate</td>
<td>0.000</td>
<td>0.958</td>
<td>0.479</td>
</tr>
<tr>
<td>False Omission Rate</td>
<td>0.000</td>
<td>0.034</td>
<td>0.017</td>
</tr>
<tr>
<td>Threat Score</td>
<td>14.283</td>
<td>0.043</td>
<td>7.163</td>
</tr>
<tr>
<td>Statistical Parity</td>
<td>0.999</td>
<td>0.001</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Note. All metrics are calculated for every class against all other classes.

Support vector machine modeling is Accuracy is 0.962, Precision (state 0 / state 1) is 0.963 / 0.000, Recall (state 0 / state 1) is 0.998 / 0.000, F1 score (state 0 / state 1) is 0.981 / NaN (due to zero denominator), AUC is 0.499. The predictive performance of the Support Vector Machine in state 1 is very limited, and a recall of zero indicates that the model fails to correctly identify financial risks. See in Table 3 and Figure 3.
Table 3. Support vector machine evaluation metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>656</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.962</td>
</tr>
<tr>
<td>Precision (Positive Predictive Value)</td>
<td>0.963</td>
</tr>
<tr>
<td>Recall (True Positive Rate)</td>
<td>0.998</td>
</tr>
<tr>
<td>False Positive Rate</td>
<td>1.000</td>
</tr>
<tr>
<td>False Discovery Rate</td>
<td>0.037</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.981</td>
</tr>
<tr>
<td>Matthews Correlation Coefficient</td>
<td>-0.007</td>
</tr>
<tr>
<td>Area Under Curve (AUC)</td>
<td>0.499</td>
</tr>
<tr>
<td>Negative Predictive Value</td>
<td>0.000</td>
</tr>
<tr>
<td>True Negative Rate</td>
<td>0.000</td>
</tr>
<tr>
<td>False Negative Rate</td>
<td>0.002</td>
</tr>
<tr>
<td>False Omission Rate</td>
<td>1.000</td>
</tr>
<tr>
<td>Threat Score</td>
<td>12.843</td>
</tr>
<tr>
<td>Statistical Parity</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Note. All metrics are calculated for every class against all other classes.

From the accuracy point of view, the random forest model is slightly better than the other two models. At a financial health state of 0, the decision tree and random forest have relatively high recall, while the support vector machine is lower. In the case of a financial health status of 1, all models face the problem of low recall, which may result in financial risks not being adequately identified. The F1 score combines precision and recall, and Random Forest has a higher F1 score at state 0. In terms of AUC, the random forest model performs relatively well, while the AUC of the support vector machine is close to random classification.

4. Discussion

4.1 Model Advantages and Disadvantages

Strengths and weaknesses of decision tree modeling:

Strengths: Ease of Understanding and Interpretation: The decision rules of decision trees are intuitive and easy to understand, enabling lay people to understand and interpret how the model works.

Adapts to non-linear relationships: Decision trees are able to handle non-linear relationships effectively and are suitable for complex associations that may exist in healthcare data.

Insensitive to Missing Values: Decision trees are insensitive to missing values and are able to handle missing information present in the dataset.

Insufficient: Prone to overfitting: Decision trees are prone to overfitting when dealing with complex problems, especially when the depth of the tree is too large, which may lead to overfitting the training data and reduce the generalization performance on unseen data.

Sensitivity to data noise: Decision trees are sensitive to noise and outliers in the data, which may lead to unstable prediction results.

Instability: Small changes to the data may result in a completely different tree structure, making the model less stable.

Strengths and weaknesses of the Random Forest model:

Strengths: High accuracy: Random forests are generally highly accurate and reduce the risk of overfitting individual trees by integrating multiple decision trees.

Assessment of Feature Importance: Random Forests can provide information on feature importance to help identify the most critical factors in financial risk prediction.

Processing of large-scale data: Random forests are able to effectively process large-scale data with a high degree of parallelism.

Shortcomings: Higher computational cost: Random forests may require more computational resources for training and prediction, especially when the number of trees is large.

Lack of Explanation: The model structure of Random Forest is relatively complex and poorly explained, making
it difficult to explain the predictive principles of the model for non-specialists.

Strengths and weaknesses of support vector machine models:

Advantage: High-dimensional data processing ability: Support vector machines perform well in high-dimensional spaces and are suitable for a large number of features that may be present in medical data.

Modeling of nonlinear relationships: Support Vector Machines can handle nonlinear relationships well through kernel functions, which improves the expressive power of the model.

Effective for Small Sample Data: Support vector machines can still maintain good performance in small sample data situations.

Inadequate: Unfriendly to large-scale data: On large-scale datasets, Support Vector Machines can take longer to train and have higher memory requirements.

Model Parameter Sensitivity: The performance of Support Vector Machines is highly dependent on choosing the right parameters, and different problems may require different parameter settings.

Poor Interpretability: Decision boundaries of support vector machines are more complex and difficult to interpret, making the models relatively less interpretable in the healthcare domain.

4.2 Recommendations for Model Performance Improvement

After a comprehensive performance evaluation of Decision Trees, Random Forests, and Support Vector Machines, here are some suggestions to improve model performance and address possible challenges:

Decision trees and random forests: Try different parameters such as tree depth, minimum number of samples, etc., and select the optimal parameters through cross-validation to improve the generalization performance of the model and avoid overfitting.

Support Vector Machines: Adjust hyperparameters such as kernel function type, regularization parameters, etc., and find the optimal configuration through grid search or other tuning methods to optimize the performance of the model.

Collect as many financial risk cases as possible to increase the size of the training data to help the model better capture the complex relationships of financial risks and improve generalization capabilities.

Consider adjusting the feature set by analyzing the relevance and importance of features, possibly adding or subtracting features to improve the model’s sensitivity to key features.

Enhance ethical awareness in modeling applications to ensure that model predictions and applications comply with medical ethics. Privacy protection measures, such as data desensitization and encryption, are taken to safeguard sensitive patient information.

Especially for support vector machines, consider using models or techniques that are easier to interpret in order to improve the interpretability of the model and meet the need for transparency in medical decision making.

Consider using integrated learning methods, such as model fusion or stacking, to combine the strengths of different models and improve overall performance.

These suggestions for improvement aim to bring the models closer to real-world application requirements, overcome possible limitations of specific models, and ensure greater reliability and credibility in healthcare applications. Future research directions should aim to continue to optimize these models and conduct more in-depth validation and application in real-world healthcare environments.

4.3 Application of Modeling in Financial Risk Forecasting

Decision trees show great potential in healthcare financial risk prediction due to their intuitive interpretability and clear decision rules. Healthcare managers can develop specific and clear financial decision-making strategies based on the generated decision tree rules, making the application of the model more intuitive and easier to understand. However, when dealing with the complex hierarchical structure of healthcare data, the decision tree is susceptible to overfitting, so parameters such as the depth of the tree and the number of node samples need to be carefully selected to balance the generalization performance and interpretability of the model.

Random forest, as an integrated model of decision tree, shows higher accuracy and stability in financial risk prediction in the medical field. Its ability to adapt to the complex structure and nonlinear relationships of healthcare data provides a more comprehensive financial risk prediction capability. However, Random Forest has relatively high computational costs and may not perform as well as other models on large-scale healthcare datasets, requiring a trade-off between computational efficiency and model performance.
Support vector machines hold promise for a wide range of applications in financial risk prediction in healthcare with their ability to handle high dimensional and complex data. Their nonlinear modeling capabilities allow them to better capture the complex relationships present in healthcare data. However, support vector machines may be limited in practical applications due to their long training time for large-scale data and sensitivity to computational resources. Its interpretability is relatively poor, and the need for transparency and interpretability in medical decision-making scenarios needs to be considered comprehensively.

In the medical field, ethical issues and privacy protection are crucial. Decision trees are relatively easy to interpret and therefore better meet the requirements of transparency and interpretability in medical decision making. Random forests and support vector machines, on the other hand, need to deal with privacy and ethical issues more carefully, taking measures such as encryption and desensitization to ensure the privacy and security of patient data. Healthcare organizations need to consider factors such as model performance, computational efficiency, and ethical and privacy protection when choosing a model.

Acknowledgments

Thanks to Shanghai Education Development Foundation and Shanghai Municipal Education Commission. Also acknowledge colleagues who assisted in conducting the study or critiquing the manuscript.

Authors Contributions

Dr. Zhao was responsible for study design and revising. Dr. Lu was responsible for data collection. Dr. Zhao drafted the manuscript and revised it. All authors read and approved the final manuscript.

Funding

This work was supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission of Chenguang Program Foundation [project number 18CGB07].

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References


Qu, M., & Li, Y. (2021). Financial risk early-warning model based on kernel principal component analysis in


**Copyrights**

Copyright for this article is retained by the author(s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).