Financial Engineering and Risk Management (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2024.070424
ISSN 2523-2576 Vol. 7 Num. 4

# *Research on the application of data analysis in predicting financial risk*

**Jiangshan Wang[1,a], Tingting Deng[2,b], Shuochen Bi[3,c], Wenqing Bao[4,d], Jue Xiao[5,e]**

[1]*The Paul Merage School of Business, University of California, Irvine, Independent Researcher Salt Lake City, UT, 84121, USA*
[2]*Simon Business School, University of Rochester, Independent Researcher Chantilly, VA, 20151, USA*
[3]*D'Amore-McKim School of Business, Northeastern University, Independent Researcher Boston, MA, 02110, USA*
[4]*Americold Logistics, LLC Atlanta, GA, 30319, USA*
[5]*The School of Business, University of Connecticut, Independent Researcher Jersey City, NJ, 07302, USA*
[a]*Jiangshanwang6@gmail.com*, [b]*dengtin1@gmail.com*, [c]*bi.shu@northeastern.edu*, [d]*bao.234@osu.edu*, [e]*juexiaowork@gmail.com*

***Keywords:*** Data Analysis; Prediction of Financial Risk; Application; Exploration

***Abstract:*** This article mainly focuses on the application of data analysis in financial risk prediction, through the comprehensive use of a variety of data analysis methods and technologies, in-depth mining and analysis of relevant financial data. In particular, relevant cases in the US financial market and banking fields are introduced to reveal the effectiveness and importance of data analysis in identifying, evaluating and predicting financial risks. The research results provide strong theoretical support and practical guidance for financial institutions and enterprises to formulate more accurate risk management strategies.

## 1. Introduction

In today's complex and changeable economic environment, financial risk has become a major challenge facing enterprises and financial institutions. Accurate prediction of financial risks is of vital significance to ensuring economic stability and sustainable development. With the rapid development of information technology, data analysis has become a powerful tool to solve this problem. As one of the largest financial markets in the world, the financial institutions and banks of the United States have accumulated rich experience and lessons in the application of data analysis, which provides an important reference and reference for our research.

## 2. Theoretical basis of data analysis and financial risk

### 2.1 Concepts, methods, and techniques of data analysis

Data analysis refers to the analysis of large amounts of data collected with appropriate statistical

analysis methods, summarizing, understanding and digesting them, so as to maximize the function of data development and play the role of data [1].

There are various methods of data analysis, mainly including descriptive statistical analysis, exploratory data analysis, and confirmatory data analysis.

## 2.2 Definition and classification of financial risks

### 2.2.1 Definition of financial risk

Refers to the enterprise engaging in various financial activities, due to various unpredictable and uncontrollable factors, there is a possibility that the final financial results within a certain period and scope may deviate from the expected business objectives, thus leading to the potential for the enterprise to suffer economic losses or reap greater benefits.

### 2.2.2 Classification of financial risks

Financial risk mainly includes financing risk, investment risk, operational risk and liquidity risk. Among them, the financing risk involves the influence of the supply and demand market and the macroeconomic environment on the enterprise; the investment risk refers to the risk of the change of market demand.

## 2.3 The role mechanism of data analysis in financial risk prediction

By collecting and analyzing the financial data, market data and operating data of enterprises, the potential financial risks of enterprises can be revealed, providing a scientific basis for enterprises to formulate effective risk coping strategies.

Data analysis can help companies identify potential financial risks. Through the analysis of the historical data and real-time data, the outliers and changing trends in the data can be found, so as to identify the possible financial risks.

Risk assessment: After the potential risk is identified, the data analysis can assess the risk. By calculating the financial ratio, building the risk model and other methods, the size and possible impact degree of the risk can be quantified to provide a basis for the formulation of risk response strategies [2]. A financial risk early-warning mechanism can also be built. By setting early warning indicators and early warning thresholds, we can remind enterprises to take timely measures to deal with risks.

# 3. Overview of the financial risks in the US financial markets and banks

## 3.1 Structure and characteristics of the American financial markets

First, it is highly developed and mature. The financial market of the United States is composed of the Federal Reserve bank system, the commercial banking system and the non-bank financial institutions, forming a huge and complex financial system. This system not only has a huge market size, a wide variety of transactions, but also a high degree of market liberalization. Second, the currency market is well-developed. The money market in the United States includes the acceptance market, the commercial paper market, the short-term bank credit market, and so on. Third, the capital market is rich. The US capital market provides a rich financing channel for the long-term capital needs with a loan maturity of more than one year, including the bond market, stock market and loan market.

### 3.2 Major financial risks facing the US financial markets and banks

Liquidity risk: In a high interest rate environment, banks are faced with deposit loss and increased liquidity pressure, which may lead to banks being forced to "shrink their balance sheet", affecting the stability of their balance sheet.

Interest rate risk: The Fed rate hike leads to higher interest rates, increasing the interest rate risk on the asset side and the liability side of the banks. On the asset side, such as bond investments and loans, may suffer losses due to rising interest rates, on the liability side, deposits costs the pressure of deposits.

Credit risk: The deteriorating quality of credit assets, such as commercial real estate loans and mortgage loans for low-income groups, increases the credit risk of banks. Commercial real estate loans are weakening in terms of asset fundamentals, real estate value and debt risks, and the debt problem of low-income mortgage groups is also worth paying attention to.

## 4. The Specific application of data analysis method in predicting financial risk

### 4.1 Data collection and preprocessing

Data collection is the first and crucial step in predicting financial risk. At this stage, we need to obtain relevant data from various sources, including but not limited to the financial statements of enterprises, market transaction data, macroeconomic indicators, industry reports, etc. For US financial institutions and banks, it could also involve data from credit rating agencies, consumer credit records and public data from regulators. The collected data often has various problems, such as missing values, outliers, inconsistent data formats, etc. Therefore, data preprocessing is particularly important. Missing values can be handled by removing records containing missing values, filling in with a mean or median, or filling in by a predictive model based on other variables. The detection and processing of outliers need to be performed cautiously, sometimes the outliers may reflect real extreme cases, and sometimes they may be data errors that need to be judged by statistical methods or business knowledge. Unification and standardization of the data format facilitated the subsequent analysis [3]. For example, when a large US bank collected credit data from its customers, it was found that some customers' income information was missing. After analysis, if the other financial characteristics of these customers are complete and representative, the average income of such customers can be used for fill. At the same time, for some obviously unreasonable high or low income data, you need to check with the original data source to ensure the accuracy of the data.

### 4.2 Feature engineering and variable selection

The purpose of feature engineering is to extract meaningful features from the raw data so that the model can better learn and predict. This may include operations like transformation, combination, and encoding of the data. For example, we can divide continuous income variables into different intervals, or we can combine multiple related variables into a composite indicator.

Variable selection is to select the most valuable variables to predict financial risk among many characteristics. Common methods include statistical test-based methods (such as t-test, F-test), model-based methods (such as stepwise regression), and machine learning-based feature selection algorithms.

Take Finix, a leader in the fintech industry in the United States. When providing loan services to small and medium-sized enterprises, the company deeply realized the importance of accurately assessing the risk of loan default to ensuring the safety of funds and promoting financial inclusion. To this end, Finix has made full use of the power of big data and artificial intelligence technology to

build a set of efficient and accurate default risk assessment system. In the process of constructing the risk prediction model, the in-depth financial data analysis and characteristic engineering were first carried out. Through mining enterprise financial statements, bank water, transaction record multidimensional data sources, the data scientists team carefully designed to reflect the core financial situation, including but not limited to profitability indicators (such as net margin, rate of return on total assets), solvency indicators (such as current ratio, quick ratio, cash flow ratio), and operational efficiency index (such as inventory turnover, accounts receivable turnover). These new features not only capture the current operating conditions of enterprises, but also imply the potential of their future development and risks. Subsequently, the enterprise adopted the random forest algorithm, a powerful machine learning model, for variable selection and model training. By integrating multiple decision trees, random forest can demonstrate excellent performance in dealing with complex data relationships, especially in terms of variable selection, which automatically evaluates the importance of each feature to the predicted target (i. e., loan default risk). After random forest screening, the company found that the cash flow ratio, asset-liability ratio and revenue growth variables in predicting small and medium-sized enterprise loan default risk has a high significance, through the innovative risk assessment system, Finx not only greatly improve the efficiency and accuracy of loan examination and approval, also effectively reduce the non-performing loan ratio, for small and medium-sized enterprises provides more convenient, low cost financing channels, but also won a good market reputation for itself.

### 4.3 Model construction and selection

When predicting financial risk, many models can be selected, such as logistic regression, neural network, etc. Logical regression is a classical statistical model that assumes a linear relationship between the dependent and independent variables and transforms the linear combination into probability values through a logistic function [4]. The logistic regression model has strong interprepower and high computational efficiency, which is suitable for linear data characteristics and moderate sample size. The decision tree model forms a tree structure similar to a flowchart by recursively splitting the data. It is able to automatically handle nonlinear relationships with no strict requirements on the distribution of data, but is prone to overfitting. Neural network is a model that mimics the connection of neurons in the human brain. It has powerful fitting ability and can handle complex nonlinear relationships, but the model is more complex, with long training time and poor interpretability. In the banking industry of the United States, the logical regression model may be tried first to predict the personal credit card default risk, because the personal credit data usually has good linear characteristics. For predicting the bank portfolio risk caused by the complex financial market fluctuations, the neural network models may be selected to capture the complex nonlinear relationships in the market.

### 4.4 Model evaluation and optimization

After the model construction is completed, the model needs to be evaluated using appropriate indicators and methods. Common evaluation indicators include accuracy, recall rate, F1 value, and the area under the ROC curve (AUC), etc. A special focus on the predictive performance of a few categories is also needed for unbalanced data. After evaluating the problems of the model, it needs to be optimized. The optimized methods include adjusting the parameters of the model, increasing the amount of data, and introducing regularization terms to prevent overfitting.

For example, the credit risk prediction model constructed by the US Axos Bank found a low recall rate for high-risk customers in the initial evaluation. Through further analysis, it was found that it is due to the unreasonable weight allocation of the model to some features. By adjusting the model

parameters, we can increase the emphasis on these features. Furthermore, utilizing cross-validation to select the optimal combination of parameters will ultimately enhance the recall and overall performance of the model.

Moreover, optimization of the model is an continuous process. As the market environment changes and the new data accumulates, the models need to be regularly re-evaluated and optimised to ensure the accuracy and reliability of their predictive capabilities.

## 5. Challenges and countermeasures in the application of data analysis

### 5.1 Data quality and safety issues

Data quality and safety are the primary challenges when data analysis is used to predict financial risk. Poor quality of data may manifest as inaccurate, incomplete, inconsistent or outdated data. For example, in the financial market of the United States, the financial data reported by different institutions may vary, making it difficult to guarantee the consistency of the data. Incomplete data may missing key financial indicators or transaction records, affecting the comprehensiveness and accuracy of the analysis. Moreover, data obsolescence depresses predictions based on these data [5].

In order to deal with data quality problems, it is necessary to establish a strict data governance mechanism, including the standardized process of data collection, verification, cleaning and update. Data validation techniques such as cross-validation and data reasonableness checks are used to ensure the accuracy of the data. For incomplete data, they can be supplemented either by data interpolation or by using multiple data sources. At the same time, establish a data quality monitoring system, regularly evaluate the data quality and correct the problems in time. In terms of data security, encryption technology is used to encrypt, store and transfer sensitive data, and strict access control and authority management are implemented, and only authorized personnel are allowed to access specific data. We should strengthen network security protection, conduct regular security audit and vulnerability scanning, and timely find and repair potential security risks.

### 5.2 Model complexity and interpretability

Complex data analysis models may show high accuracy in predicting financial risk, but are often accompanied by reduced interpretability. For example, complex models such as deep neural networks may be able to capture subtle patterns and complex relationships in the data, but it is difficult to intuitively understand how the models make decisions and generate predictions. This creates difficulties for financial institutions to explain the risk assessment process to regulators, explain the decision basis to customers, and communicate their internal risk management decisions.

To balance the complexity and interpretability of the model, some highly interpretable models or techniques can be employed. For example, decision tree and rule induction models are relatively easy to understand and interpret. For complex models, such as neural networks, the output of the model can be interpreted using feature importance analysis, local explanatory model-interpretation of individual predictions (LIME), and SHAP values. Moreover, in the model development process, features and variables with practical significance and interpretability are selected by combining domain knowledge and business logic.

### 5.3 Regulatory environment and compliance requirements

The financial industry is strictly regulated, and the application of data analysis in predicting financial risks must meet the relevant regulatory requirements and compliance standards. In the United States, financial regulators such as the Federal Reserve and the Securities and Exchange

Commission (SEC) have clear regulations on the data use, model risk management and disclosure of financial institutions. For example, financial institutions need to verify and monitor the accuracy, stability and impartiality of models and report to regulators [6] .

Failure to comply with regulatory requirements can result in serious legal consequences and reputational risks. However, regulatory requirements are often dynamic, and financial institutions need to be timely informed and adapt to new regulations and policies. To meet the regulatory challenges, financial institutions should establish a sound model risk management framework, including the processes and standards for model development, validation, monitoring, and auditing.

Our organization should assign a dedicated team to monitor compliance matters, closely track regulatory dynamics, and timely adjust data analysis strategies and models to meet compliance requirements. We should also strengthen communication and cooperation with regulatory agencies, actively participate in the formulation and discussion of regulatory policies, and strive to obtain certain flexibility and innovation space under the premise of compliance.

At the same time, a sound documentation and disclosure mechanism is established to clearly record the methods of data analytics, the assumptions and parameters of the model, the verification results, and the process of risk assessment, in order to provide transparent and censored information to regulatory agencies and external auditors.

## 6. Conclusion

In short, data analysis shows great potential and value in predicting financial risk. This study profoundly reveals the core position of data analysis in the modern enterprise financial management. Through efficient and accurate data analysis tools and technologies, relevant enterprises can identify potential financial risk points earlier, provide strong support for decision-making, and then optimize the allocation of resources to ensure the stable operation of enterprises. Through the study of American financial markets and banks, we are deeply aware of its successful application experience and the challenges. In the future, with the continuous progress of technology and the increasingly rich data resources, the application of data analysis in financial risk prediction will continue to deepen and expand, which can provide stronger support for the innovation and development of the financial industry in the United States.

## References

*[1] Wang Yunyan, Gu Hua. Application of F-Score model in BYD company financial risk warning [J]. Shopping mall modernization, 2023 (18): 183-185.*

*[2] Ge Yanhua. Research on financial internal control system and financial risk prevention work [J]. Chinese Science and Technology Journal Database (full-text edition) Economic Management, 2023 (4): 4.*

*[3] Wang Jundan. Financial Risk analysis of united Bank of China Minsheng Bank [J]. Jilin University of Finance and Economics, 2017.*

*[4] Wan Zhongjie. Key analysis of the common financial risks and countermeasures encountered by enterprises [J]. Chinese Science and Technology Journal Database (full-text edition) Economic Management, 2023 (3): 4.*

*[5] Hua Chen. Research on the financial risk of the digital transformation of enterprises [J]. Chinese Science and Technology Journal Database (full-text edition) Economic Management, 2023 (4): 4.*

*[6] Manna. Analysis of the financial Risk and Buffering mechanism of the Federal Reserve after the financial crisis [J]. Financial Accounting, 2019.DOI: CNKI: SUN: JRKJ.0.2019-12-008.*