European Journal of Business, Economics & Management

Vol. 1 No. 3 2025

PINNACLE ACADEMIC PRESS
JOURNALS, BOOKS & PROCEEDINGS

Article Open Access

Data-Driven Credit Risk Assessment and Optimization Strategy Exploration

Lingyun Lai [1,*]

[1] Financial Department, BCG Glass Industry Inc., New York, NY, 10017, United States
* Correspondence: Lingyun Lai, Financial Department, BCG Glass Industry Inc., New York, NY, 10017, United States

Received: 27 June 2025
Revised: 04 July 2025
Accepted: 16 July 2025
Published: 22 July 2025

Abstract: With the rapid development of data-driven technology, the financial sector is increasingly reliant on data-driven approaches to credit risk assessment. This paper analyzes the application of decision tree, support vector machine, neural network and other models in credit risk assessment, discusses the current problems of data quality, bias, transparency and computing resources, and puts forward optimization strategies, such as strengthening data cleaning, reducing data bias, improving algorithm fairness, enhancing model transparency and optimizing computing resource allocation. The goal is to improve the accuracy and efficiency of assessments.

Keywords: credit risk assessment; data-driven; decision tree; support vector machine; algorithm fairness

1. Introduction

With the development of financial industry and the advancement of information technology, credit risk assessment has become the core tool of risk control for financial institutions. Traditional methods rely on manual experience and simple statistical models, which face challenges of accuracy and processing efficiency. In recent years, data-driven technologies have provided more precise and efficient solutions through machine learning models such as decision trees, support vector machines, and neural networks, which have demonstrated promising capabilities in risk prediction and classification tasks [1]. However, the effectiveness of these models is often hindered by issues related to data quality, algorithmic bias, and model transparency. These concerns have raised attention in both academia and industry regarding the reliability and fairness of automated credit evaluation systems [2]. Therefore, optimizing data processing workflows, enhancing algorithmic fairness, and improving model transparency are key to improving the evaluation effect and decision-making quality.

2. Data-Driven Credit Risk Assessment Model

2.1. Decision Tree Model: Classification and Decision in Credit Risk Assessment

The decision tree model is widely used in credit risk assessment. It classifies risks based on customer characteristics by constructing a tree structure. Each node represents a feature or decision point; the branch represents the direction of the decision, and the leaf node represents the conclusion of the credit evaluation. The advantage of decision trees is that they are clear and easy for people to intuitively grasp the evaluation process. In credit risk assessment, the decision tree selects the customer's credit level according to the customer's financial status (such as income level, debt status, credit history, etc.). However, decision trees are also susceptible to noisy data and are not efficient when dealing with more complex data relationships, which usually need to be optimized by pruning or integration methods such as random forests. Figure 1 below summarizes the process of credit risk assessment by decision tree:

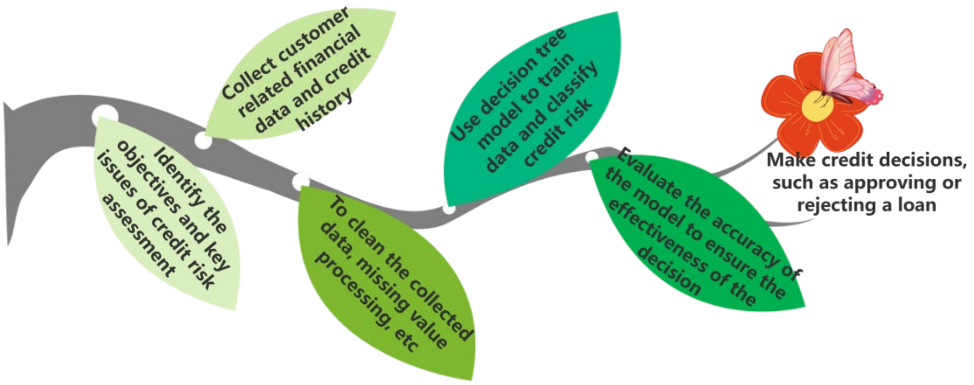
Identify the objectives and key issues of credit risk assessment
Collect customer related financial data and credit history
To clean the collected data, missing value processing, etc
Use decision tree model to train data and classify credit risk
Evaluate the accuracy of the model to ensure the effectiveness of the decision
Make credit decisions, such as approving or rejecting a loan

Figure 1. Process of credit risk assessment by decision tree.

2.2. Support Vector Machine Model: Risk Classification in High-Dimensional Space

Support Vector Machine (SVM) is a supervised learning algorithm widely used in credit risk assessment. By determining the best hyperplane in the high-dimensional feature space, it can segment different categories of data to achieve accurate classification. SVM improves the generalization ability of the model by maximizing the interval between classes, and shows strong efficiency in processing multi-dimensional complex data sets. With kernel functions such as polynomials or radial basis functions (RBF), SVM has the ability to project data into higher-dimensional Spaces, allowing efficient classification of otherwise linear, indivisible data sets. In the application of credit risk assessment, SVM uses financial indicators of customers (such as income level, debt status, credit history, etc.) to segment customer groups and accurately predict the likelihood of default, thereby improving the accuracy and effectiveness of assessments [3]. Table 1 below summarizes the application and advantages of support vector machines (SVM) in credit risk assessment:

Table 1. Application and advantages of support vector machine (SVM) in credit risk assessment.

gist	Content description
Algorithm type	Support vector Machine (SVM), a powerful supervised learning algorithm
core idea	By finding the optimal hyperplane in the high-dimensional space, the data of different categories can be segmented to achieve accurate classification
Optimization objective	Maximize the spacing between classes (the distance from the hyperplane to the data point) to improve the generalization ability of the model
Application field	Credit risk assessment, based on the customer's financial characteristics (such as income, liabilities, credit history) classification, forecast default risk

Advantages of high dimensional data processing	Kernel function techniques (such as polynomial kernel function and radial basis kernel function) are used to map data to higher dimensional Spaces, enhance the linear separability of data, and improve the adaptability of dealing with nonlinear problems
Common kernel function	Polynomial kernel function, Radial Basis kernel function (RBF)
Classification effect	Through support vector training, the classification boundary is automatically adjusted to improve the accuracy of credit risk assessment
Application situation	It is suitable for high-dimensional data sets, especially in the face of complex nonlinear problems
Value to financial institutions	Provide effective risk management tools to help accurately assess customers' credit risk

Table 1 shows that support vector machines (SVMs) offer significant advantages in credit risk assessment, particularly in processing high-dimensional data and enhancing classification accuracy, providing financial institutions with effective risk management support.

2.3. Neural Network Model: Application of Deep Learning in Credit Risk Assessment

Neural networks, especially deep learning, are widely used in credit risk assessment. Its unique multi-layer node structure can independently extract deep features from data, especially when analyzing large high-dimensional datasets, and it often achieves better predictive performance than traditional algorithms, though at a higher computational cost. This type of network can capture nonlinear relationships in customers' financial data and identify hidden patterns, thereby enhancing the accuracy of risk prediction. Different from traditional evaluation models, deep networks extract more recognizable features through their hidden structures, thus improving the accuracy of classification. Deep neural networks can learn important feature representations directly from raw data, thereby reducing the need for manual feature engineering, and they use backpropagation algorithms to optimize performance. Figure 2 below shows the process of the neural network from data input to prediction results:

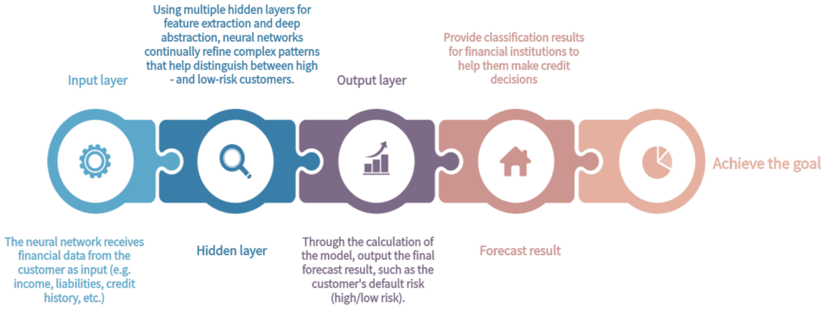
Using multiple hidden layers for feature extraction and deep abstraction, neural networks continually refine complex patterns that help distinguish between high - and low-risk customers.
Input layer
Output layer
Provide classification results for financial institutions to help them make credit decisions
Achieve the goal
The neural network receives financial data from the customer as input (e.g. income, liabilities, credit history, etc.)
Hidden layer
Through the calculation of the model, output the final forecast result, such as the customer's default risk (high/low risk).
Forecast result

Figure 2. The process of neural network from data input to prediction result.

3. Main Problems and Challenges in Data-Driven Credit Risk Assessment

3.1. Inaccurate Data Quality: Distorted Evaluation Results

In the data-driven credit risk assessment, the accuracy of data quality has a crucial impact on the assessment results. If the data is incomplete, wrong or contradictory, it will lead to the deviation in the training process of the model, and ultimately affect the accuracy of the evaluation results. Errors in data may stem from various sources, such as manual entry mistakes, system failures, or unstable data sources, all of which can hinder the predictive model's ability to accurately reflect the customer's true credit profile. When there is a problem with the data, the model will rely on inaccurate information to make predictions, resulting in a distortion of the credit risk assessment and may not reflect the actual default risk of the customer. This distortion will not only weaken the effectiveness of the model itself, but also may have a misleading effect on the decision-making of financial institutions. Inaccurate data quality can also lead to instability in the evaluation model, especially in the case of complex and volatile market conditions, which makes it difficult for the model to adjust quickly to new trends in the data, thereby amplifying the prediction bias. In addition, the timeliness and completeness of data are also indispensable, and outdated or one-sided data will not accurately reflect the credit status of customers, which will affect the correct assessment of customer risks [4].

3.2. Data Bias Injustice: Algorithm Discrimination

In data-driven credit risk assessment, data bias is a key factor leading to algorithmic unfairness. Training data often reflect historical inequities that can be unconsciously learned and perpetuated by algorithms, leading to bias in evaluation results. Algorithms can exert an unfair influence on specific groups, affecting their credit scores and risk assessments. The bias in the data arises not only from the unevenness of the sample distribution, but also from hidden features such as underlying socioeconomic factors and cultural attributes, which are not clearly identified but have a disproportionate impact on the model's predictions. Algorithmic discrimination is often difficult to detect because the model's decision-making process is complex and opaque, but its effects are far-reaching, especially in the financial sector, where unfair evaluation results can lead to unreasonable restrictions on credit opportunities for specific groups. As the application of artificial intelligence and machine learning in the financial industry continues to grow, the problem of data bias has become more prominent, and has become an important factor affecting fairness and credibility. If adequate attention is not paid to this problem and no measures are taken to address it, prejudice and discrimination could seriously undermine the fairness of financial decisions and erode public trust in financial institutions.

3.3. Lack of Model Transparency: Distrust of Decision-Making

In data-driven credit risk assessment, the opacity of the model poses a significant problem point. When the decision-making process of the model is not transparent, users and decision-makers cannot understand how the model arrives at specific decisions. This opacity can lead to reduced trust in the results of assessments, especially in the financial sector, where clients and regulators may have doubts about the judgment of the models. The "black box" nature of the model makes it difficult for outsiders to review and confirm its operational logic, thus exacerbating the uncertainty surrounding the decision. The lack of transparency not only weakens the interpretability of models, but also makes it difficult to trace problems back to their source and make effective adjustments and optimizations. Especially in credit risk assessment, the prediction of the model is directly related to the customer's financial opportunities and loan decisions. If the judgment process of the model cannot be clearly explained, the customer may question the evaluation results and think that the evaluation results are arbitrary or unfair [5].

3.4. Insufficient Computing Resources: Evaluation Delays and Inefficiencies

Insufficient computing resources are a significant challenge in data-driven credit risk assessment. When the computing power of the system cannot meet the needs of big data processing, there may be delays in the evaluation process, resulting in overall inefficiency. Credit risk assessment often requires processing large amounts of customer data and involves complex algorithms and model training, especially when using deep learning and other models with high computational requirements. Limited computing resources can lead to prolonged model training cycles, affecting the timeliness of decision making. This delay not only reduces the efficiency of risk assessment but may also compromise both the responsiveness and service quality of financial institutions. Especially in a rapidly changing environment in financial markets, the lag in assessment can miss the best time to make decisions. In addition, the lack of computing resources may also lead to the degradation of the performance of the model, which can not give full play to its prediction ability, and then affect the accuracy of the evaluation results. In some cases, to cope with computational limitations, institutions are compelled to compromise by simplifying model structures or reducing data size, which further undermines the comprehensiveness and credibility of the assessment.

4. Data-Driven Credit Risk Assessment Optimization Strategy

4.1. Strengthen Data Cleaning and Optimize the Pre-processing Process

In data-driven credit risk assessment, data cleaning and preprocessing are crucial. Financial data usually contains missing values, outliers and noise, which may cause errors in evaluation results if not processed. For example, in a credit risk assessment, a customer's revenue data may be missing, which affects the accuracy of the model. Faced with this problem, interpolation can be used to fill in the missing values, or data inference can be made based on the characteristics of similar customers to ensure the integrity of the data. When dealing with outliers, if the amount of customer debt is significantly beyond the normal level, we can identify and exclude these abnormal data by setting reasonable limit values or applying statistical techniques, so as to avoid adverse effects on model training. In addition, the standardization and normalization of data is also a key step. Assuming that the value ranges of the two features—income and liabilities—differ significantly, failure to standardize them may cause the model to overemphasize the income attribute, resulting in biased prediction results. By transforming the data to a unified scale, we can ensure that each feature gets a reasonable weight allocation in the model training.

This optimized path of data cleaning and preprocessing can not only improve the quality of data, but also enhance the efficiency of the model in identifying actual credit risks, and provide stronger support for the final evaluation results [6].

4.2. Reduce Data Bias and Improve Algorithm Fairness

In data-driven credit risk assessment, data bias can lead to algorithmic unfairness, resulting in unequal assessment results for certain groups. In order to improve the impartiality of the model, effective measures must be taken in the process of data acquisition, preprocessing and modeling to reduce the influence of bias. For example, in the data collection phase, over-reliance on one group (such as high-income groups or specific regions) can lead to an uneven sample, which can affect the accurate assessment of customers in low-income groups or other regions. In the face of this problem, sampling techniques (such as oversampling or undersampling) can be used to balance the data set and ensure the representation of various customer groups, thereby reducing the impact of bias. In the data preprocessing stage, the removal of sensitive features is also an effective method. By eliminating variables that may introduce discrimination (such as gender, race, etc.), the model avoids allowing these characteristics to unfairly influence its decisions. In addition, when training the model, fairness constraints can be introduced to ensure thatthe model's predictions are as balanced as possible across different groups. A simple fairness loss function can be expressed as:

$$L_{fair} = \alpha \cdot L_{accuracy} + \beta \cdot L_{fairness} \tag{1}$$

Among them, $L_{accuracy}$ Indicates loss of prediction accuracy, $L_{fairness}$ Loss for representation of equity, α and β Is the weight coefficient, Used to balance the trade-off between accuracy and fairness. By adjusting the loss function, the model not only optimizes prediction accuracy, It can also ensure fairness of evaluation between different groups.

4.3. Enhance the Transparency of the Model and Improve the Interpretability

In data-driven credit risk assessment, it is critical to enhance the transparency and interpretability of models. Many complex models, such as deep learning and random forests, while excellent at predictive accuracy, often make the decision process difficult to understand due to their "black box" nature, which can lead to distrust of the evaluation results by users and decision makers. Measures need to be taken to make the decision-making process of the model easier to understand and explain. For example, in the face of complex deep learning models, it is possible to reveal the specific impact of each feature in the model's decision making by using interpretability augmentation techniques such as LIME (locally interpretable model-independent model) or SHAP (Shapley value). These methods help analyze and explain how the model makes decisions based on customer attributes such as income and liabilities, enabling users to gain insight into the model's decision logic.

Additionally, integrating inherently interpretable components into the model design, such as rule-based systems or simplified tree structures, can further improve transparency during training. For example, in the decision tree model, the decision-making process can be intuitively displayed through the visual tree structure, and the model can clearly see how to gradually make risk assessment according to different financial characteristics. By enhancing the transparency and interpretability of the model, financial institutions can improve the trust of customers and regulators in the results of credit risk assessment, and also provide a basis for feedback and adjustment to improve the model, and further improve the accuracy and fairness of the assessment.

4.4. Optimize Computing Resource Configuration to Improve Real-Time Response Capability

As the amount of data increases and the complexity of the model increases, so does the demand for computing resources. If computing resources are not properly allocated, processing delays may occur, thereby compromising the timeliness of decision making. Therefore, it is necessary to allocate computing resources reasonably, use the cloud computing platform to flexibly adjust resource allocation according to the load, achieve dynamic expansion of computing capacity, thereby ensuring the model responds rapidly and updates evaluation results in real time during high concurrency or large-scale data processing. In addition, improving the computational efficiency of the algorithm is also a key measure to improve the responsiveness. For example, with distributed computing or parallel computing technology, computing tasks can be decomposed into multiple nodes for parallel processing, thus significantly reducing data processing time. In practical applications, when financial institutions need to process large amounts of customer data, distributed processing can speed up the credit risk assessment process and ensure that the customer's credit assessment can be completed in real time, thereby improving the overall efficiency of the service. To effectively optimize the allocation of computing resources, system performance can be evaluated using metrics such as response time and resource utilization efficiency, as represented by the following formula:

$$T_{total} = \frac{T_{task}}{N_{nodes}} + \alpha \cdot \frac{R_{load}}{C_{resources}} \tag{2}$$

Among them, T_{task} indicates the total processing time, T_{task} Is the processing time of a single task, N_{nodes} Is the number of nodes counted, α Is the adjustment factor, R_{load} Indicates the current load, $C_{resources}$ Is the total capacity of computing resources. By adjusting resource allocation and optimization algorithm, the real-time and accuracy of credit risk assessment can be significantly improved, the potential risks can be effectively reduced, and the overall service level can be improved.

5. Conclusion

In the financial sector, with the rapid development of data-driven technologies, credit risk assessment plays an increasingly critical role. Using algorithms such as decision trees, support vector machines and neural networks, financial institutions can more accurately determine the level of customer credit risk, but factors such as data volume, bias, transparency of assessment, and resource allocation continue to affect the accuracy and efficiency of assessment. This paper proposes optimization strategies, such as strengthening data cleaning, reducing data bias, improving algorithm fairness, enhancing model transparency, and optimizing computing resource allocation, to improve the quality of risk assessment, promote the development of more efficient and fair risk management in the financial industry, and contribute to the financial industry's sustainable development.

References

1. L. Yun, “Analyzing Credit Risk Management in the Digital Age: Challenges and Solutions,” *Econ. Manag. Innov.*, vol. 2, no. 2, pp. 81–92, Apr. 2025, doi: 10.71222/ps8sw070.
2. S. Yang, “The Impact of Continuous Integration and Continuous Delivery on Software Development Efficiency,” *J. Comput. Signal Syst. Res.*, vol. 2, no. 3, pp. 59–68, Apr. 2025, doi: 10.71222/pzvfqm21.
3. Y. Qin, M. Chen, Y. Liu, J. Zhang, L. Wang, X. Zhou, et al., “Data-driven optimisation of process parameters for reducing developed surface area ratio in laser powder bed fusion,” *Int. J. Adv. Manuf. Technol.*, vol. 136, no. 7, pp. 3821–3831, 2025, doi: 10.1007/s00170-025-15038-4.
4. G. Medio, A. Tarpani, F. Barboni, D. Laucelli, M. Berardi, L. Giustolisi, et al., “Sinkhole Risk-Based Sensor Placement for Leakage Localization in Water Distribution Networks with a Data-Driven Approach,” *Sustainability*, vol. 16, no. 12, p. 5246, 2024, doi: 10.3390/su16125246.
5. Z. Wang, “Artificial intelligence and machine learning in credit risk assessment: Enhancing accuracy and ensuring fairness,” *Open J. Social Sci.*, vol. 12, no. 11, pp. 19–34, 2024, doi: 10.4236/jss.2024.1211002.
6. M. K. Nallakaruppan, R. K. Gupta, A. P. Singh, M. Sharma, L. Thomas, H. Yadav, et al., “Credit risk assessment and financial decision support using explainable artificial intelligence,” *Risks*, vol. 12, no. 10, p. 164, 2024, doi: 10.3390/risks12100164.

Disclaimer/Publisher’s Note: The views, opinions, and data expressed in all publications are solely those of the individual author(s) and contributor(s) and do not necessarily reflect the views of PAP and/or the editor(s). PAP and/or the editor(s) disclaim any responsibility for any injury to individuals or damage to property arising from the ideas, methods, instructions, or products mentioned in the content.