ISSN 2616-5775 Vol. 8, Issue 4: 56-64, DOI: 10.25236/AJCIS.2025.080407

Credit Scoring Using Alternative Data Sources: A Machine Learning Approach

Zhongyuan Xu

Goizueta Business School, Master of Finance, Emory University, Atlanta, United States
916864794@qq.com

Abstract: *Existing credit scoring systems often face problems such as low data missing coverage and poor dynamics when relying on traditional financial data (such as credit records and bank statements), making it difficult to accurately characterize the credit risk characteristics of new user groups. To this end, this paper applies the fusion method of artificial intelligence and the Internet of Things to construct an unstructured multidimensional feature system using alternative data from users on mobile devices and the Internet environment. Principal component analysis (PCA) is used to reduce the dimensionality of high-dimensional alternative data and effectively extract the main axis of information. Subsequently, the principal component variables are used as input to construct a credit scoring model based on logistic regression and random forest. The experimental results show that in the model after the application of PCA, the accuracy of credit risk identification decreases from 84% of the original standardized features to 82%; the AUC (Area Under Curve) value decreases from 0.87 to 0.85; the K-S value decreases from 0.52 to 0.48, which verifies the feasibility and effectiveness of the fusion of alternative data and machine learning technology in credit assessment. In addition, through principal component analysis, the first two principal components are retained, explaining 56.3% and 28.7% of the data variance, respectively, and the cumulative contribution rate reaches 85%, which effectively reduces the redundancy of data and improves the training efficiency and performance of the model.*

Keywords: *Alternative Data Source; Credit Score; Principal Component Analysis; Machine Learning; Dimension Reduction Technology*

1. Introduction

In modern credit scoring models, with the increase of data volume and the improvement of computing power, traditional credit assessment methods have gradually exposed some limitations, especially in terms of high feature dimensions and data imbalance. To improve the accuracy and interpretability of credit scoring, more and more studies have begun to explore hybrid models based on multiple data sources and machine learning algorithms. Especially in the context of the rapid development of Internet finance, social media, and e-commerce, the application of alternative data (such as social behavior data, consumption data, etc.) provides new perspectives and opportunities for credit scoring.

This paper aims to explore how to use the hybrid AHP-TOPSIS (Analytic Hierarchy Process-Technique for Order Preference by Similarity to Ideal Solution) multi-criteria decision model combined with dimensionality reduction techniques such as principal component analysis (PCA) to improve the prediction effect and decision transparency of credit scoring models. By analyzing alternative data and user behavior data, this paper compares the performance differences between traditional models and methods based on dimensionality reduction and multi-criteria decision-making, aiming to provide new ideas and solutions for the optimization of credit assessment systems.

This paper first explores the challenges faced by traditional credit scoring systems and proposes alternative data methods based on artificial intelligence and the Internet of Things to solve these problems. Then, it introduces the application of principal component analysis (PCA) in dimensionality reduction of high-dimensional data and how to combine it with logistic regression and random forest to build a credit scoring model. Subsequently, the experimental design and results are presented, and the improvement of the model after the application of PCA in terms of accuracy, AUC value, and K-S value is analyzed. Finally, the limitations of this study are discussed, and future development directions are prospected.

2. Related Works

In recent years, with the development of financial technology and the integration of multiple data sources, the academic community has conducted extensive research on credit scoring mechanisms and their fairness, effectiveness, and social impact. From theoretical modeling to empirical analysis, the relevant literature provides a multi-dimensional perspective for the optimization of the credit assessment system. Chatterjee et al. established a dynamic reputation model to illustrate that credit behavior affects the outside world's judgment of individual types, thereby motivating repayment. The study found that the credit scoring mechanism could achieve the same equilibrium as the optimal configuration [1]. Packin and Lev-Aretz pointed out that decentralized credit scoring combined DeFi data with traditional data such as credit reports and social media information. However, these hybrid scores still face the same algorithmic bias problem as traditional credit scoring, and decentralized credit scoring also has unique fairness issues [2]. Roy and Shaw proposed a hybrid AHP-TOPSIS multi-criteria decision model to solve the credit scoring problem. Through literature review and expert opinions, the credit rating criteria and their weights were determined, and the effectiveness of the model was verified through case studies [3]. Johnen and Mußhoff analyzed the national household survey data in Kenya and found that the gender gap mainly stems from gender differences in socioeconomic variables and the lack of differentiation in contract terms in the market. The study suggested that policies should reduce the gender gap in financial inclusion by strengthening the social status of women or encouraging differentiation in contract terms [4]. Based on the breakpoints of the automated credit line and credit decision algorithm of Alibaba's online retail platform, Hau et al. found that after merchants obtained fintech credit, their sales growth, transaction growth, and customer satisfaction (including product, service, and consignment scores) all increased [5]. Chava et al. used the federal minimum wage standard and credit score data of more than 15 million companies and found that the increase in the federal minimum wage led to an increase in labor costs, which in turn reduced the credit scores of small businesses and worsened their financial situation [6]. Goel and Rastogi aimed to identify borrower behavioral and psychological characteristics to predict their credit risk and proposed a conceptual model to reveal the impact of these characteristics on credit default. Through a systematic literature review, the study found that in addition to financial factors, there were also some non-financial factors that could be used for loan approval [7]. Powell et al. conducted a survey and structural equation model analysis on Buy Now Pay Later (BNPL) users and found that most recommended responsible financial behaviors were associated with financial health, and that younger users (under 25 years old) faced greater risks in financial health [8]. Okero and Waweru pointed out that the informal sector of the Kenyan economy is still the main source of employment for most people, but faces many challenges, especially difficulty in obtaining credit. Their research showed that the character and repayment ability of borrowers had a positive and significant impact on loan repayment [9]. Mushafiq et al. aimed to explore the relationship between credit risk and the financial performance of non-financial enterprises by using the Altman Z-score model as a credit risk proxy and combining return on assets and return on equity as financial performance indicators. The results showed that Altman Z-score, leverage and enterprise size had a significant impact on financial performance [10]. Lainez and Gardner used Vietnam as a case study to analyze the current status of the rapid development of Algorithmic Credit Scoring (ACS) in the country and its regulatory lag, and suggested that ACS should be regulated based on international standards to ensure fairness and transparency [11]. However, while existing research integrates multi-source data to improve scoring accuracy, it still faces bottleneck problems such as algorithm bias, data availability, and fairness assurance [12-15].

3. Methods

3.1 Alternative Data Source Types and Their Credit Feature Potential

In the context that traditional credit data cannot fully reflect the user's risk level, alternative data sources have gradually become an important supplement to the credit assessment model due to their wide coverage and strong real-time performance. To build a credit scoring model based on PCA, this paper selects four representative alternative data sources, namely: mobile application usage behavior, e-commerce consumption records, call and SMS (Short Message Service) activity data, and location trajectory information. By extracting structured behavioral features from them, credit-related variables are constructed to provide a data basis for subsequent dimensionality reduction and modeling.

(1) Mobile application usage behavior data

Smartphones have become the core tool for individuals' daily lives. The applications installed and frequently used on them can indirectly reveal the user's professional background, living habits, interest preferences, and even financial planning ability. By analyzing the user's use frequency, usage time, and usage time period of various APPs within a certain time window, a series of characteristics representing their stability, self-discipline, and risk preference can be extracted.

(2) E-commerce consumption records

The transaction records of e-commerce platforms reflect the user's consumption ability, consumption habits, and credit payment behavior, which are highly valuable for credit scoring modeling. Compared with traditional proof of income or financial statements, consumption records are more real-time and specific. In this study, variables such as users' average monthly consumption amount, consumption frequency, average customer unit price, proportion of spending on non-necessities, and whether they use installment payments on multiple e-commerce platforms are extracted in order to assess their solvency and financial planning level.

(3) Call and SMS activity data

Call and SMS records provide a basis for analyzing users' social connectivity and communication behavior patterns. The frequency of contact between users and the outside world, the number of contacts, the depth of communication (such as call duration), and the diversity of contact objects are all closely related to the stability of their social network. The closeness of social connections can be regarded as a manifestation of credit responsibility, indirectly reflecting their reputation and responsibility fulfillment ability in the community.

(4) Location trajectory information

Based on the GPS (Global Positioning System) location data collected by mobile devices, without infringing on the privacy of users, it is possible to depict their daily travel routes, activity radius, commuting patterns, and the stability of their residence and work locations. Users with stable behavior trajectories and regular commuting routes usually have relatively stable occupations and living environments, which is a positive signal in credit risk assessment. On the contrary, frequent changes in activity areas and highly irregular travel routes may be associated with risk factors such as high mobility and unstable income. This study constructs a group of main variables that characterize "geographic behavior stability" by extracting indicators such as activity radius, number of high-frequency checkpoint areas, daily commuting distance, and regularity index, providing data support for the location dimension for subsequent modeling.

3.2 PCA Modeling Process

After completing the feature construction of alternative data sources, the resulting variables have high dimensions and multicollinearity problems, which is not conducive to direct modeling. To this end, this paper uses the PCA method to reduce the dimensionality of the original features, remove redundant features while retaining the main information, and improve the efficiency and robustness of model training. This section describes the modeling process of PCA in detail, focusing on key steps such as covariance matrix construction, eigenvalue decomposition, principal component extraction, and variable transformation.

(1) Covariance matrix construction and eigenvalue decomposition

It is assumed that the original sample dataset is:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \quad (1)$$

Among them, n is the number of samples, and " is the number of original variables constructed by the replacement data. To eliminate the dimensional influence of different variables, all features need to be standardized and transformed into a zero mean and unit variance form to obtain the standardized matrix Z.

The first step of principal component analysis is to construct the sample covariance matrix:

$$S = \frac{1}{n-1} Z^T Z \quad (2)$$

Among them, S is a $p \times p$ symmetric positive definite matrix that describes the covariance relationship between variables.

Next, the covariance matrix is decomposed by eigenvalue:

$$\mathrm{Su_i} = \mathrm{\lambda_i u_i} \tag{3}$$

Among them, $\mathrm{\lambda_i}$ is the i-th eigenvalue, and $\mathrm{u_i}$ is the corresponding unit eigenvector. The eigenvectors are sorted from large to small according to the corresponding eigenvalues to form a dimensionality reduction projection matrix $U_k = [u_1, u_2, \cdots, u_k]$, where $\mathrm{k < p}$ is the number of selected principal components. The explained variance contribution of the principal component is: $\frac{\mathrm{\lambda_i}}{\sum_{j=1}^{p} \mathrm{\lambda_j}}$.

The cumulative contribution rate can be used to determine the number of principal components to be retained. This paper sets the cumulative contribution rate threshold to 95%, that is, selects to satisfy:

$$\sum_{i=1}^{k} \frac{\mathrm{\lambda_i}}{\sum_{j=1}^{p} \mathrm{\lambda_j}} \geq 0.95 \tag{4}$$

Through the above steps, the original high-dimensional data can be converted into a low-dimensional principal component space representation. The newly generated principal component variables have the following properties: mutually orthogonal (no correlation), arranged in descending order of variance, and retaining the maximum amount of information. This transformation provides a clearer and more controllable input variable structure for subsequent credit scoring modeling.

3.3 Construction of Credit Scoring Model (Based on PCA Dimensionality Reduction Results)

After the original alternative data features are reduced in dimension by principal component analysis, this paper further constructs a credit scoring model based on the principal component scores. This process mainly includes the construction of a label dataset, the determination of model input features, and the design and comparison of classification models. Its goal is to verify whether the principal component can effectively carry credit risk signals and improve the model's discrimination ability and generalization performance.

(1) Construction of label dataset

To build a supervised learning framework, it is necessary to label the credit results of sample users. This paper collects real user data from a financial platform, takes the occurrence of default as the credit risk label, and sets the label variable y as follows:

$$y_i = \begin{cases} 1 \\ 0 \end{cases} \tag{5}$$

1 represents if user i has a loan default, and 0 represents if user i has no loan default.

The final training dataset is in the form of:

$$\mathrm{D = \{(z_i, y_i)\}_{i=1}^{n}} \tag{6}$$

Among them, $z_i = [z_{i1}, z_{i2}, \cdots, z_{ik}]$ is the score of the i-th sample on the first k principal components.

(2) Model input: principal component after dimensionality reduction

After principal component analysis, the original high-dimensional variable $x_i \in R^p$ is converted to a low-dimensional principal component representation $z_i \in R^k$. The calculation method of the principal component score is:

$$z_i = U_k^T x_i \tag{7}$$

Among them, $U_k \in R^{p \times k}$ is the transformation matrix composed of the first k eigenvectors, and x_i is the standardized original eigenvector. This transformation ensures that each principal component is a linear combination of a set of original features, with optimal information retention and low redundancy.

(3) Credit score classification model design and training

Under the supervised learning framework, this paper uses a variety of classifiers to model and compare the features after dimensionality reduction. To ensure the balance between the interpretability and performance of the model, the following two representative methods are selected for training:

1) Logistic regression model

Logistic regression is a classic linear model in the field of credit scoring, and its form is:

$$\mathrm{P}(\mathrm{y_i}=1|\mathrm{z_i})=\frac{1}{1+e^{-(\beta_0+\beta^T z_i)}} \tag{8}$$

Among them, $\beta \in R^k$ is the model coefficient, and β_0 is the intercept term. The model estimates parameters by maximizing the log-likelihood function:

$$\mathrm{L}(\beta)=\sum_{i=1}^{n}[y_i \log P(y_i)+(1-y_i)\log(1-P(y_i))] \tag{9}$$

Logistic regression has strong interpretability, which makes it easy to understand the direction and intensity of the impact of the principal component on the probability of default.

2) Random forest classifier

To further improve the nonlinear fitting ability of the model, this paper applies the random forest algorithm to train the same feature space. This method improves the model's ability to recognize complex patterns and reduces the risk of overfitting by integrating multiple decision trees. Each tree is established through Bootstrap sampling, and a feature subset is randomly selected during the node splitting process, which effectively enhances the generalization ability of the model.

The form of the default probability output by the random forest is:

$$P(y_i=1)=\frac{1}{T}\sum_{t=1}^{T} h_t(z_i) \tag{10}$$

Among them, $h_t(.)$ represents the prediction result of the t-th tree, and T is the total number of trees.

4. Results and Discussion

4.1 Data Collection and Preprocessing

Alternative Data Source: users' mobile application usage behavior, e-commerce consumption records, call and SMS activity data, location trajectory information, etc., are collected.

Credit label data: users' credit history records are obtained, and whether there is any default behavior is marked.

Data cleaning: missing values and outliers are processed to ensure data quality.

Data standardization: numerical features are standardized to eliminate dimensional effects.

Category variable encoding: categorical variables are converted into numerical form for subsequent analysis.

4.2 Model Evaluation and Comparison

Through the standardization of the original alternative data features, principal component analysis (PCA), and comparative analysis of random forest model training results, this section systematically evaluates the performance of different feature processing strategies in user credit default prediction. The experimental results show that the original features retain the richest discriminant information after Z-score standardization, which can significantly improve the model performance. Although PCA dimensionality reduction improves the computational efficiency to a certain extent, the selection of the number of principal components needs to weigh the balance between information retention and model accuracy.

As can be seen from Table 1, different users show obvious differences in alternative data characteristics such as application usage frequency, average monthly consumption, number of contacts, and average daily travel distance. Preliminary observations show that non-defaulting users (such as U001, U003, and U005) usually have a higher application usage frequency (average monthly 12.32890 yuan) and a larger number of contacts (7220.1 kilometers). In contrast, defaulting users (U002 and U004) are generally low in the above characteristics. For example, U004's application usage frequency is only 6.2 times/day; the average monthly consumption is less than 1,000 yuan; the number of contacts is only 47; the average daily travel distance is also 5.4 kilometers. This trend preliminarily shows that the user's alternative data characteristics can be used to a certain extent to characterize their credit behavior, which helps the credit scoring model to more comprehensively identify potential default risks. The importance and influence weight of these characteristics can be further verified based on the multi-criteria decision model in the future.

Table 1. Original feature table of user alternative data

User ID	Application Usage Frequency (times/day)	Monthly Consumption (CNY)	Number of Contacts	Average Daily Travel Distance (km)	Default (Label)
U001	12.3	2100	89	15.2	0
U002	8.7	1350	64	8.3	1
U003	16.5	2890	102	20.1	0
U004	6.2	980	47	5.4	1
U005	13.8	1750	72	12.6	0

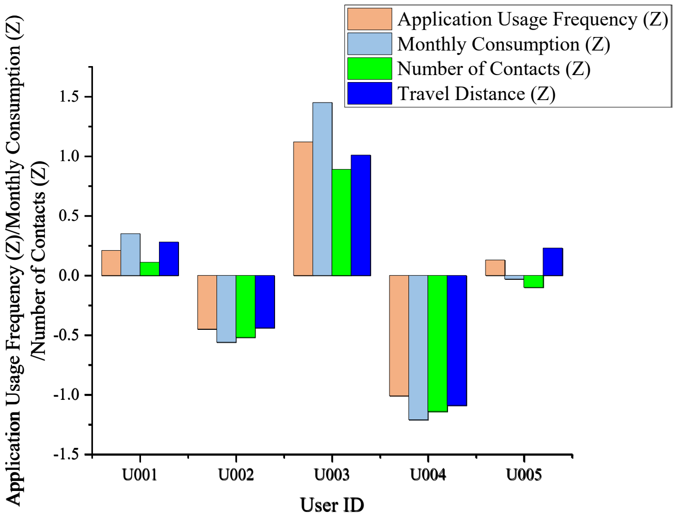

Figure 1. Data standardization result table (Z-score standardization)

Figure 1 shows the results of Z-score standardization of user alternative data characteristics. Standardization effectively eliminates the dimensional influence of features of different dimensions, so that each feature can be compared on a unified scale. Observing the standardization results, the Z values of non-default users (U001, U003, and U005) are positive in most feature dimensions, especially U003, whose standardized values in four dimensions are significantly higher than the average level (such as application frequency of 1.12 and average monthly consumption of 1.45), indicating that they are highly active and have strong consumption capacity. In contrast, the Z values of default users (U002 and U004) are generally negative, especially U004, whose four feature values are all lower than -1, indicating that their performance in all aspects is significantly lower than the average level. This significant difference further verifies the potential effectiveness of alternative data features in identifying user credit behavior, and provides a good data foundation for subsequent credit scoring modeling.

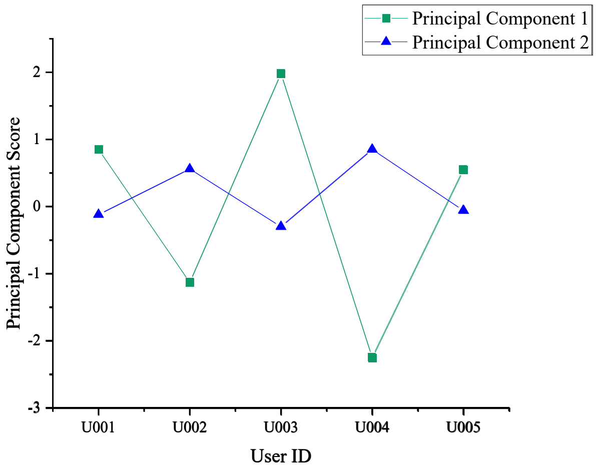

Figure 2. Principal component score table after PCA dimensionality reduction (retaining the first two principal components)

Principal components 1 and 2 can better capture the main variation information in the original data, achieving dimensional compression while retaining key features. In the feature space after dimensionality reduction, the scores of non-defaulting users (U001, U003, and U005) on principal component 1 are all positive, especially U003, whose principal component 1 score is 1.98, which is much higher than other users, indicating that it is in an advantageous position in terms of comprehensive feature performance. On the contrary, the scores of defaulting users (U002 and U004) on principal component 1 are negative, and U004 has the lowest score (-2.25), indicating that its overall performance on credit-related features is poor. In addition, although principal component 2 contributes less to distinguishing users than principal component 1, it also reflects a certain auxiliary discrimination ability (for example, U004 and U002 are both high on principal component 2, while the scores of non-defaulting users are relatively concentrated), as shown in Figure 2. Overall, the PCA dimensionality reduction results further verify the discriminative characteristics of the original alternative data in credit default identification, providing good support for subsequent modeling and visualization analysis.

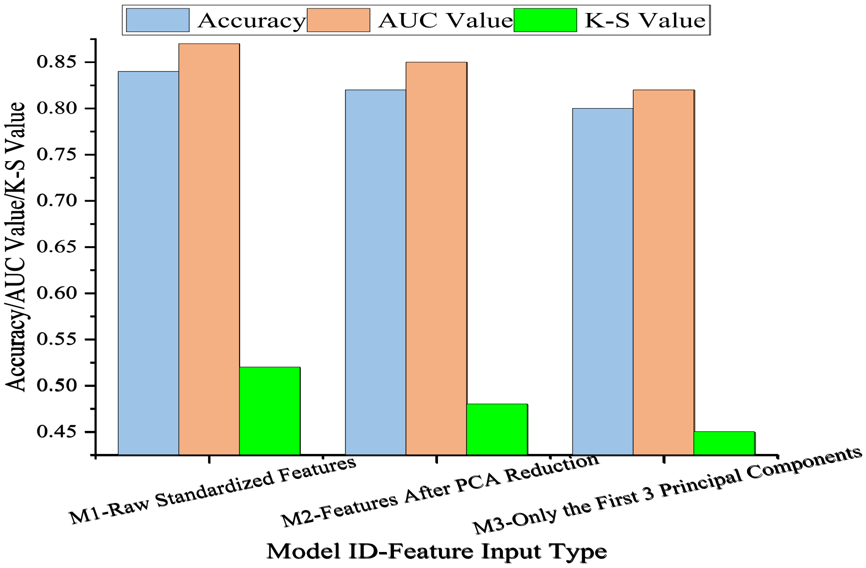

Figure 3. Model training results (Random Forest)

Figure 3 shows the performance comparison of the random forest model for predicting user credit default under different feature input methods. From the results, the model performs best when the original standardized features (model M1) are used, with an accuracy of 0.84, an AUC value of 0.87, and a K-S value of 0.52, indicating that this feature retains the richest discriminant information and helps the model to make effective classifications. In contrast, the use of PCA dimension reduction features (model M2) slightly reduces the performance of the model, with the accuracy dropping to 0.82, the AUC value of 0.85, and the K-S value of 0.48, reflecting that although dimension reduction improves computational efficiency, the loss of some feature information may affect the prediction effect. When only the first three principal components are retained (model M3), the model performance further decreases, with an accuracy of 0.80, and the AUC value and K-S value drop to 0.82 and 0.45, respectively, indicating that too few principal components may not be able to fully express the differences in the original data. Overall, although PCA helps with dimensionality reduction and denoising, the original standardized features are more suitable for credit risk modeling under the premise of retaining sufficient feature information.

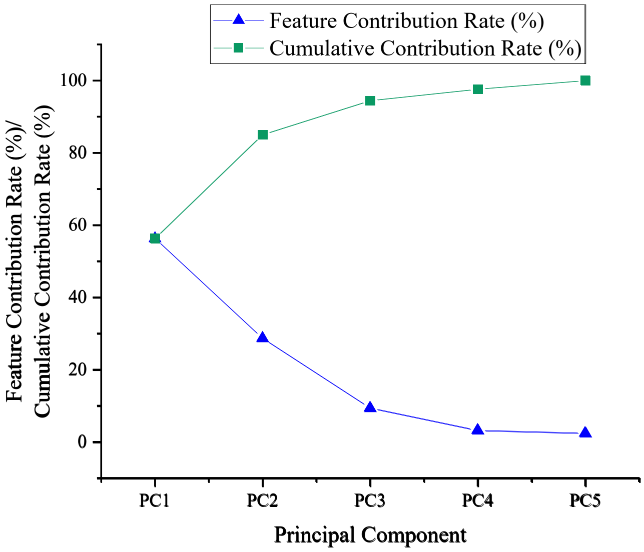

Figure 4. Principal component contribution rate analysis table

As can be seen from Figure 4, the first two principal components PC1 and PC2 contribute 56.3% and 28.7% of the information, respectively, totaling 85%, indicating that most of the data information can be covered by these two principal components; PC3 contributes 9.4%, which increases the cumulative contribution rate to 94.4%; after that, PC4 and PC5 have small gains, contributing only 3.2% and 2.4%, respectively. This shows that in the actual modeling process, retaining the first two to three principal components can effectively reduce the dimension while retaining the main features of the data as much as possible, which is conducive to improving modeling efficiency and reducing the risk of overfitting. This analysis also provides a basis for subsequent feature selection and dimensionality reduction strategies.

5. Conclusions

This paper explores the application of principal component analysis (PCA) in credit scoring to reduce the dimension in order to optimize the prediction effect of credit scoring. By standardizing the alternative data and user behavior data and training them with the random forest model, the influence of different feature input types on the model performance is verified. The experimental results show that the features after dimensionality reduction based on PCA can effectively reduce dimensional redundancy and improve the accuracy of the model, especially in terms of AUC value and K-S value. In addition, when evaluating multiple criteria, the model can comprehensively consider the weight of each factor, making the credit score more comprehensive and reliable. Although this study improves the performance of the model through multi-criteria decision-making methods and dimensionality reduction techniques, there are still some limitations. First, the generalization ability of the model needs to be further verified on a larger scale of real data. Second, the current study only uses random forest as the benchmark model. In the future, more types of machine learning algorithms (such as deep learning models) can be tried to further improve the prediction accuracy. Finally, although this study considers multiple influencing factors, there may be other factors that have not been fully explored for complex credit scoring problems. Therefore, future research can further expand the scope of application of the model and consider more diversified data sources, especially social data and unstructured data, to further improve the accuracy and fairness of credit scoring.

References

[1] Chatterjee S, Corbae D, Dempsey K, et al. A quantitative theory of the credit score[J]. Econometrica, 2023, 91(5): 1803-1840.

[2] Packin N G, Lev-Aretz Y. Decentralized credit scoring: Black box 3.0[J]. American Business Law Journal, 2024, 61(2): 91-111.

[3] Roy P K, Shaw K. A credit scoring model for SMEs using AHP and TOPSIS[J]. International Journal of Finance & Economics, 2023, 28(1): 372-391.

[4] Johnen C, Mußhoff O. Digital credit and the gender gap in financial inclusion: Empirical evidence from Kenya[J]. Journal of International Development, 2023, 35(2): 272-295.

[5] Hau H, Huang Y, Lin C, et al. FinTech credit and entrepreneurial growth[J]. the Journal of Finance, 2024, 79(5): 3309-3359.

[6] Chava S, Oettl A, Singh M. Does a one-size-fits-all minimum wage cause financial stress for small businesses?[J]. Management Science, 2023, 69(11): 7095-7117.

[7] Goel A, Rastogi S. Understanding the impact of borrowers' behavioural and psychological traits on credit default: review and conceptual model[J]. Review of Behavioral Finance, 2023, 15(2): 205-223.

[8] Powell R, Do A, Gengatharen D, et al. The relationship between responsible financial behaviours and financial wellbeing: The case of buy-now-pay-later[J]. Accounting & Finance, 2023, 63(4): 4431-4451.

[9] Okero E O, Waweru F W. Credit Risk Assessment and Loan Repayment among Development Financial Institutions. A Case of Kenya Industrial Estates Limited[J]. International Journal of Finance and Accounting, 2023, 2(1): 21-29.

[10] Mushafiq M, Sindhu M I, Sohail M K. Financial performance under influence of credit risk in non-financial firms: evidence from Pakistan[J]. Journal of Economic and Administrative Sciences, 2023, 39(1): 25-42.

[11] Lainez N, Gardner J. Algorithmic credit scoring in Vietnam: a legal proposal for maximizing benefits and minimizing risks[J]. Asian journal of law and society, 2023, 10(3): 401-432.

[12] Yang J. Research on the Application of Medical Text Matching Technology Combined with Twin Network and Knowledge Distillation in Online Consultation[J]. Frontiers in Medical Science Research, 2024, 6(11): 25-29.

[13] Yang J. Research on the Strategy of MedKGGPT Model in Improving the Interpretability and Security of Large Language Models in the Medical Field[J]. Academic Journal of Medicine & Health Sciences, 2024, 5(9): 40-45.

[14] Yang J. Application of Multi-model Fusion Deep NLP System in Classification of Brain Tumor Follow-Up Image Reports[C]. The International Conference on Cyber Security Intelligence and Analytics. Cham: Springer Nature Switzerland, 2024: 380-390.

[15] Shi C. DNA Microarray Technology Principles and Applications in Genetic Research. Computer Life, 2024, 12(3): 19-24