Journal of Service Science and Management, 2025, 18(6), 461-475
https://www.scirp.org/journal/jssm
ISSN Online: 1940-9907
ISSN Print: 1940-9893

Application of Precise Loan Qualification Identification Based on K-Means and Decision Tree Model from the Perspective of Consumer Behavior in Universities

Yan Fan*, Chaosheng Zhang, Xiao Ge

School of Mathematical Sciences, Jiangsu University, Zhenjiang, China
Email: *free1002@126.com

How to cite this paper: Fan, Y., Zhang, C. S., & Ge, X. (2025). Application of Precise Loan Qualification Identification Based on K-Means and Decision Tree Model from the Perspective of Consumer Behavior in Universities. *Journal of Service Science and Management, 18,* 461-475.
https://doi.org/10.4236/jssm.2025.186029

Received: October 20, 2025
Accepted: November 25, 2025
Published: November 28, 2025

Abstract

From the perspective of student consumption behavior, a data-driven framework for screening student loan eligibility was developed using K-means clustering analysis and decision tree models. A questionnaire survey was conducted on 829 students at colleges and universities to collect comprehensive data covering various dimensions such as economic background and consumption patterns. The K-means algorithm successfully predicted and identified the loan eligibility of the samples, with its predictive performance demonstrated by combining it with the decision tree model. Additionally, through in-depth discussions with credit departments, its practical value and reliability were confirmed. This study has enhanced the data-driven intelligent decision mechanism and provided strong support for precise loan disbursement in student loans, paving the way for the application of financial technology in credit areas.

Keywords

Targeted Loan Disbursement, Student Loans, Consumer Behavior, K-Means Algorithm, Decision Tree Model

1. Introduction

With the popularization of higher education, an increasing number of students require student loans to complete their studies. However, the screening of loan eligibility and risk prediction for student loans have always been significant challenges faced by financial institutions and educational departments. Traditional

loan models often rely on complex approval processes and subjective judgments, which are not only inefficient but also susceptible to human factors, leading to inaccurate lending decisions and insufficient risk control (Yao, 2021). Therefore, achieving precise loan disbursement for student aid has become an urgent issue to address.

In recent years, with the development of big data and artificial intelligence technologies, data-driven decision models have been widely applied in the financial sector. The K-means algorithm (Wang, Wang, Feng et al., 2012) and decision tree models (Fang, Wu, Zhu et al., 2011), as two classic data mining techniques, are extensively used in pattern recognition, cluster analysis, and predictive modeling. This paper aims to construct a precise loan qualification screening and risk prediction model using the K-means algorithm and decision tree model. Through in-depth analysis of university students' consumption behavior data, different student groups with varying economic statuses and consumption habits are identified, thereby providing a scientific basis for student loan disbursement.

The significance of this research lies in improving the accuracy and efficiency of student loan disbursement through data-driven methods and reducing lending risks for financial institutions. It also provides students with fairer and more transparent access to borrowing opportunities. Furthermore, the research findings offer valuable references for other financial institutions, promoting the application and development of financial technology in the credit field.

2. Research Design and Data Analysis

2.1. Research Design

This study selected two key universities in Zhenjiang City, Jiangsu Province—Jiangsu University and Jiangsu University of Science and Technology—as research sites. These two institutions respectively represent comprehensive universities and specialized technological universities, and their student composition, regional distribution, and consumption patterns exhibit a certain degree of representativeness within Jiangsu Province, providing a suitable sample base for exploring loan eligibility discrimination based on consumption behavior. However, this convenience sampling approach may limit the generalizability of the findings to other regions and different types of universities across China**.** Simultaneously, by comprehensively understanding students' consumption behavior patterns, including family income and fund flow, their eligibility for student loans was assessed. A questionnaire survey method (Zheng, 2014) was adopted, covering students' basic information, family income, consumption capacity, consumption structure, consumption concepts, consumption credit, and other aspects, striving to fully reflect students' economic levels and consumption habits.

2.2. Data Collection Process

To ensure data accuracy and timeliness, the team used an electronic questionnaire system for data collection. All participating students accessed the questionnaire

page by scanning a QR code or clicking a link. After completion, the system automatically saved the data, avoiding issues like data loss or damage associated with traditional paper questionnaires (Meng, 2017). Ultimately, 500 questionnaires were distributed at each university, totaling 1000 distributed questionnaires, with 829 questionnaires recovered.

Throughout the data collection and processing process, the team strictly adhered to research ethics and compliance requirements, respecting students' right to information and privacy, ensuring the legality and compliance of the research activities.

2.3. Data Quality and Error Control

Considering the accuracy and scientificity of the questionnaire, focus on the reliability and validity testing of the scale (Zeng & Huang, 2005). The team conducted a pre-survey with 50 students from each university.

Table 1. Reliability analysis of variables.

Variable	Number of Items	Cronbach's Alpha	Standardized Cronbach's Alpha
Consumption Capacity	5	0.738	0.769
Consumption Structure	3	0.803	0.815
Consumption Concept	3	0.887	0.889
Consumption Credit	3	0.794	0.796

Analysis of **Table 1** shows that the Cronbach's Alpha coefficients (Liu, Zhang, & Yang, 2018). for all subscales exceeded 0.7, the results indicate that the survey questionnaire used in this study has good internal consistency and good reliability.

Table 2. Validity analysis of the total scale.

Kaiser-Meyer-Olkin Measure		0.786
Bartlett's Test of Sphericity	Approx. Chi-Square	536.849
	Df	175
	Sig	0.002

From **Table 2**, it can be seen that the measurement items for each indicator in the questionnaire design have a good theoretical basis; meanwhile, the total scale's Cronbach's Alpha coefficient (Liu, Zhang, & Yang, 2018) is 0.892. In the validity analysis, the KMO value (Li & Bai, 2014) is 0.786, and Bartlett's test of sphericity value (Li & Bai, 2014) is 536.849, with a significance level of 0.002, indicating that the overall reliability and validity of the questionnaire are within a highly acceptable range, possessing good reliability and validity (Zeng & Huang, 2005).

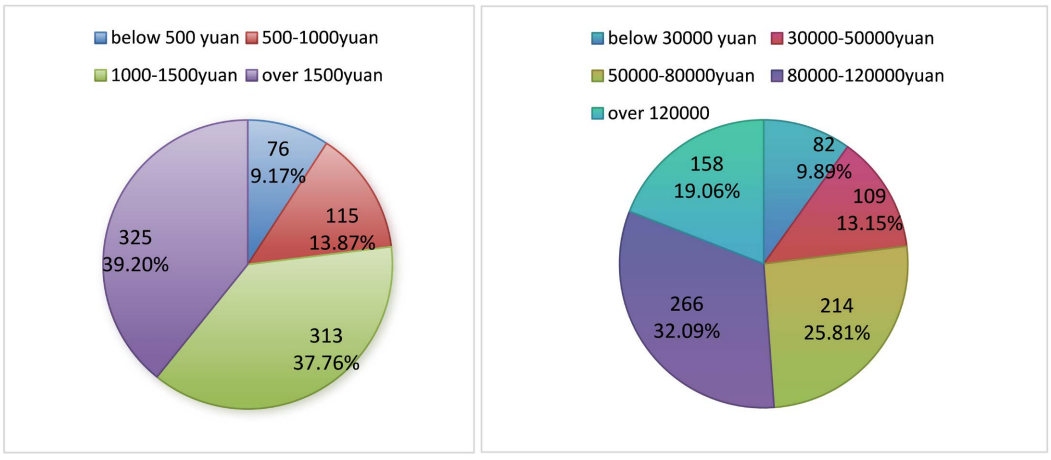
2.4. Data Information Analysis

This survey questionnaire is based on the analysis of the consumption behavior of students from the two universities (Zang & Li, 2012). A total of 1000 questionnaires were distributed, with 829 valid responses, resulting in an effective response rate of 82.9%. The basic information of the survey respondents is shown in Table 3.

Table 3. Basic information of respondents.

Item	Specific Content	Number of People	Percentage (%)
Gender	Male	401	48.4 4.4
	Female	428	51.6
Grade Level	Freshman	151	18.2
	Sophomore	297	35.8
	Junior	106	12.8
	Senior	63	7.6
	Graduate Student	212	25.6
Lives on Campus	Yes	736	88.8
	No	93	11.2

The distribution of annual family income showed that most respondents' family economic status was at a medium level, with families having an annual income of 50,000 - 120,000 RMB accounting for 62.7%. Meanwhile, 77.0% of students had a monthly living expense exceeding 1000 RMB. The distribution of annual family income and monthly living expenses is shown in Figure 1(a) and Figure 1(b), respectively.



(a) (b)

Figure 1. Distribution of respondents' annual household income.

The primary source of income for respondents was mainly full-time wages, ac-

counting for 59.1%. Part-time wages, business income, and other sources also constituted certain proportions, as detailed in Table 4. The study focused more on low-income families, defined as those with an annual family income below 30,000 RMB and relying on a single, relatively backward income methods, source of income such as part-time wages, relief funds, or pensions. Additionally, these samples had lower monthly living expenses, with 9.2% spending below 500 RMB and 13.9% spending between 500 - 1000 RMB. These samples became the key target group for student loans.

Table 4. Statistics on respondents' consumption capacity.

Item	Specific Content	Number of People	Percentage (%)
Primary Family Income Source	Full-time Wage	491	59.2
	Part-time Wage	76	9.2
	Business Income	197	23.7
	Investment Income	47	5.7
	Relief Funds/Pension	18	2.2
Family Income Stability	Very Stable	116	14.0
	Stable	351	42.3
	Average	182	22.0
	Somewhat Unstable	122	14.7
	Very Unstable	58	7.0
Has Part-time Job	Yes	183	22.1
	No	646	77.9

Family Income Stability was measured on a 5-point scale with the following anchors: "Very Stable", "Stable", "Average", "Somewhat Unstable", and "Very Unstable".

Details of respondents' consumption structure are shown in Table 5. Analysis shows that students' dining frequency in campus canteens is mainly one meal per day, accounting for 37.2%. In terms of dining consumption level, 7 - 15 yuan accounts for 48.5%, which is a normal consumption level. In contrast, meals for students from poorer backgrounds were typically resolved in the canteen, with most meal costs between 0 - 7 RMB, and their highest monthly expenditure category was usually food and daily necessities.

Average Cafeteria Spend was categorized into four brackets: "Below 7 RMB", "7 - 15 RMB", "15 - 20 RMB", and "Above 20 RMB".

Regarding consumption decision factors, cost-effectiveness was the most concerned element for students, accounting for 41.4%. Brand, quality, and design/appearance were also considerations for some students. In the self-assessment of consumption habits, the vast majority of students exhibited frugal consumption attitudes, with the combined proportion of "frugal" and "very frugal" reaching

83.7%. Furthermore, 96.3% of students hold a positive attitude towards the importance of consumption within their economic capacity. Statistics on respondents' consumption concepts are shown in Table 6.

Table 5. Statistics on respondents' consumption structure.

Item	Specific Content	Number of People	Percentage (%)
Cafeteria Meal Frequency	Almost 3 meals daily	156	18.8
	Mostly 2 meals daily	238	28.7
	Usually 1 meal daily	306	36.9
	Occasionally eats at cafeteria	96	11.6
	Almost never eats at cafeteria	33	4.0
Average Cafeteria Spend	Below 7 RMB	128	15.4
	7 - 15 RMB	393	47.4
	15 - 20 RMB	241	29.1
	Above 20 RMB	67	8.1
Highest Expenditure Category	Food & Daily Necessities	457	55.1
	Study Materials	101	12.2
	Entertainment & Social Activities	158	19.1
	Clothing & Personal Care	113	13.6

Table 6. Statistics on respondents' consumption concepts.

Item	Specific Content	Number of People	Percentage (%)
Primary Shopping Consideration	Cost-effectiveness	342	41.3
	Brand	178	21.5
	Quality	148	17.9
	Design/Appearance	161	19.4
Consumption Habit Type	Very Frugal	99	11.9
	Frugal	357	43.1
	Average	190	22.9
	Carefree/Spontaneous	183	22.1
Importance of Moderate Spending	Very Important	433	52.2
	Important	270	32.6
	Average	113	13.6
	Not Very Important	13	1.6

Consumption Habit Type was classified using a 4-point scale: "Very Frugal", "Frugal", "Average", and "Carefree/Spontaneous". The Importance of Moderate Spending was measured on a 4-point scale: "Very Important", "Important", "Average", and "Not Very Important".

Financial product usage, as seen from Table 7, shows that the vast majority of students had never used loan or credit services. The student group generally had high credit awareness, regarding repayment behavior, 75.2% of students stated they would repay on time. For factors affecting delayed repayment, emergency situations or unexpected events were considered the main factors, while personal factors had a smaller impact.

Table 7. Statistics on respondents' consumption credit situation.

Item	Specific Content	Number of People	Percentage (%)
Has Used Loan Services	Student Loan	58	7.0
	Consumer Loan	17	2.1
	Credit Card	31	3.7
	No	723	87.2
Repayment Attitude	Always On Time	618	74.5
	Mostly On Time	194	23.4
	Sometimes Overdue	17	2.1
	Often Overdue/Never Repays	0	0.0
Main Factor Affecting On-Time Repayment	Poor Spending Plan	103	12.4
	Lack of Financial Knowledge	31	3.7
	Emergency/Unexpected Event	695	83.9

3. Research Methods and Modeling Process

3.1. Cluster Analysis Model

Cluster analysis is a core method in unsupervised learning, aiming to partition a dataset into groups or clusters such that the similarity within clusters is maximized (Sun, Liu, & Zhao, 2008) and the difference between clusters is maximized. The greater the intra-cluster similarity and inter-cluster difference, the better the clustering effect (Sun, Liu, & Zhao, 2008). This paper primarily introduces the K-means algorithm, a classic unsupervised learning technique widely used in data mining and pattern recognition (Wang, Wang, Feng et al., 2012). The objective function of K-means is to minimize the Within-Cluster Sum of Squares (WCSS), iterating until convergence to find the optimal partition, making it suitable for discovering consumer behavior patterns (Wang, Wang, Feng et al., 2012). Its mathematical expression is:

$$J = \sum_{i=1}^{K} \sum_{x \in S_i} \left\| x - \mu_i \right\|^2,$$

where K is the number of clusters, S_i is the set of points in the i-th cluster, x is a point in the cluster, μ_i is the center of the i-th cluster, and $\left\| x - \mu_i \right\|^2$ is the Euclidean distance from point x to cluster center μ_i.

When using the K-means algorithm for cluster analysis, determining the optimal number of clusters K is a crucial step. The algorithm's effectiveness can vary significantly for different values of K. The Elbow Method (Long, Zhang, & Zhang,

2020) and the Silhouette Coefficient (Zhu, Ma, & Zhao, 2010) were used to evaluate the clustering effect and select the optimal number of clusters. For each point i, the Silhouette Coefficient $S(i)$ consists of two parts:

$a(i)$: The average distance between point i and all other points in the same cluster.

$b(i)$: The average distance between point i and all points in the nearest neighboring cluster.

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}.$$

The process was implemented using Python, including data preprocessing, feature selection, preprocessing transformation, determining the optimal number of clusters, and finally executing the K-Means clustering analysis. A Column Transformer was used for preprocessing: categorical features were One Hot Encoder (Liang, Chen, Zhang et al., 2019), converting all features into a format suitable for cluster analysis. To determine the optimal K, the Elbow Method was implemented, and the Silhouette Coefficient for different K values was calculated, selecting the K value corresponding to the maximum Silhouette Coefficient to ensure the best clustering effect. Finally, the clustering results were appended back to the original DataFrame, and a new dataset containing cluster labels was stored. The aim was to segment students' consumption behavior through cluster analysis to achieve precise loan disbursement.

3.2. Decision Tree Model

The decision tree algorithm recursively selects the optimal features for splitting, building a tree until stopping conditions are met. The core of the algorithm lies in feature selection and tree generation and pruning.

The decision tree based on Information Gain (Luan & Ji, 2004) (the splitting method used by ID3) for selecting the best split first requires calculating the data's information entropy and information gain. Calculate the dataset's Entropy (Luan & Ji, 2004):

$$H(S)=-\sum_{i=1}^{n}p_i\log_2 p_i,$$

where $H(S)$ represents the information entropy of dataset S, n is the number of classes in the dataset, p_i is the proportion of samples belonging to the i-th class in the dataset, calculated as $\frac{|S(i)|}{|S|}$, where $|S(i)|$ is the number of samples of the i-th class, and $|S|$ is the total number of samples in dataset S.

And calculate Information Gain (Luan & Ji, 2004):

$$IG(S,T)=H(S)-\sum_{j=1}^{m}\frac{|S(i)|}{|S|}H(S_j),$$

where: $IG(S,T)$ represents the information gain of feature T for dataset S.

The C4.5 algorithm was then used to calculate the Gain Ratio (Luan & Ji, 2004): for feature selection, combining information gain and the intrinsic value of the feature to address ID3's bias towards multi-value features:

$$GR(S,T)=\frac{IG(S,T)}{IV(T)},$$

where: $GR(S,T)$ represents the gain ratio of feature T for dataset S, $IV(T)$ is the intrinsic value of feature T, given by:

$$IV(T)=-\log_2\left(\sum_{j=1}^{m}\left(\frac{|S(i)|}{|S|}\right)^2\right).$$

Based on the results of cluster analysis, a decision tree model is constructed to validate the accuracy of the clustering results and further enhance the performance of the predictive model.

Based on the previous cluster labels and other features, a decision tree model was constructed with the target variable set as "Meets Loan Requirements". Combined with detailed consumption behavior data, the data was split into training and test sets (Zhang, Hu, & Huang, 2021), with the test set accounting for 30%, to evaluate the model's generalization ability.

In the construction and validation process of the decision tree model, a confusion matrix was used to calculate the model's accuracy, showing the comparison between actual and predicted classification results. It typically includes four basic categories: True Positives (TP); False Positives (FP); True Negatives (TN); False Negatives (FN). Based on the confusion matrix, the decision tree model's Accuracy can be calculated by the following formula:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}.$$

Additionally, to comprehensively evaluate model performance, Precision, Recall, and F1-score were calculated to ensure the model is not only accurate but also performs well and balanced in predicting both positive and negative classes.

4. Model Experimental Results and Validation

4.1. Cluster Analysis Model Results

Using the K-Means algorithm, a comprehensive evaluation was conducted on multiple participant variables including "family income", "cafeteria meal frequency and cost", "expenditure structure", and "credit concepts". The samples were successfully divided into four distinct consumption behavior clusters: 0, 1, 2, 3, with specific numbers shown in Table 8:

Table 8. Comprehensive clustering results of consumption behavior.

Comprehensive Cluster	0	1	2	3
Number of Samples	196	152	174	307
		Total		829

The research first identified that each cluster represented different economic statuses and consumption behavior characteristics, providing an important reference basis for student loan disbursement. Among them, Cluster 3 samples demonstrated weak loan qualifications, possessing stable family income sources and a relatively good standard of living. Conversely, Cluster 0 samples significantly exhibited optimal loan qualifications, becoming the primary target group for lending, characterized by unstable family income and economic hardship. For Clusters 1 and 2, which had ambiguous boundaries in the comprehensive clustering, separate one-dimensional cluster analyses were performed based on the four key dimensions: "Consumption Capacity", "Consumption Structure", "Consumption Concept", and "Consumption Credit". The resulting data is shown in **Table 9**:

Table 9. One-Dimensional clustering results of consumption behavior.

Comp. Cluster	0	1					2		3
Number	196	152					174		307
		Sub-Cluster Dim.	0	1	2	3	4	Total 336	
		Consumption Cap	31	96	37	84	88		
		Consumption Stru.	48	92	40	93	63		
		Consumption Con.	42	99	34	82	79		
		Consumption Cred.	39	91	25	99	82		
		Total							829

A Multi-Level Decision Algorithm (MLDA) (Niu & Cai, 2007) was applied to identify potential loan candidates from Clusters 1 and 2. Principal Component Analysis (PCA) (Lin & Zhang, 2005) was used to determine the weight of each feature in predicting loan eligibility, reflecting its impact on the final decision. A Random Forest classifier (RF) (Fang, Wu, Zhu et al., 2011) was utilized to build a comprehensive scoring model, calculating a loan eligibility score for each sample based on the assigned weights and quantified economic status. Ultimately, from the collected data of 829 samples, 281 samples were identified and judged as "Meeting Loan Qualifications".

Receiver Operating Characteristic (ROC) curve analysis (Cao, Li, Chen et al., 2021) was used to determine the optimal loan eligibility score threshold. Based on this threshold, samples were classified as "Meeting Loan Qualifications" and "Not Meeting Loan Qualifications". Through cross-validation and ROC analysis (Cao, Li, Chen et al., 2021), the optimal model parameters and threshold were determined. Experimental results showed that the MLDA algorithm performed excellently in identifying loan candidates, with high accuracy and a high AUC value (Cao, Li, Chen et al., 2021), as shown in **Figure 2**.

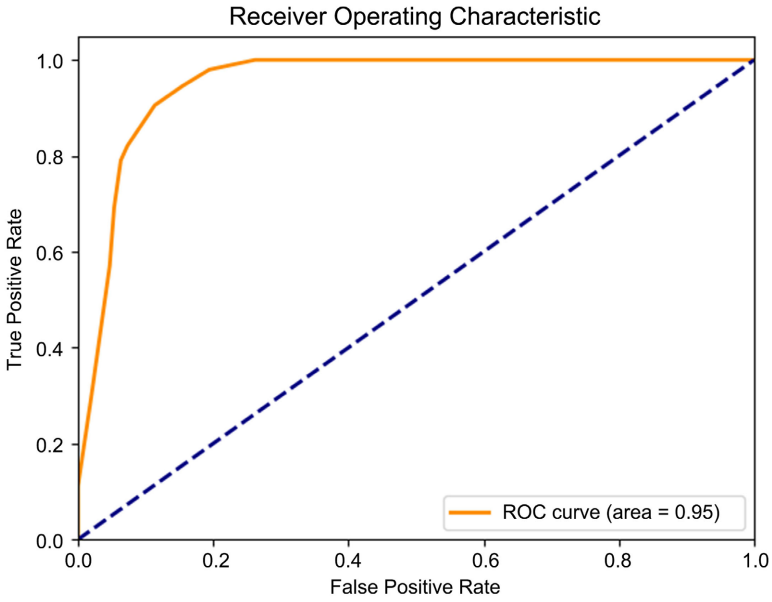


Figure 2. ROC curve for the screening model performance evaluation (AUC = 0.95).

The binary label "Meets Loan Requirements" was created through a multi-stage analytical process. First, K-means clustering was applied to comprehensive consumption behavior data—including family income, cafeteria meal frequency and spending, expenditure structure, and credit concepts—segmenting the student population into distinct groups (Clusters 0 - 3). Cluster 0, characterized by unstable family income and evident economic hardship, was directly identified as the primary group meeting loan requirements. For students in the ambiguously bounded Clusters 1 and 2, a Multi-Level Decision Algorithm (MLDA) was applied. This involved using Principal Component Analysis (PCA) to determine feature weights for predicting loan eligibility and a Random Forest (RF) classifier to build a comprehensive scoring model. An optimal score threshold was determined via Receiver Operating Characteristic (ROC) curve analysis, and students whose scores exceeded this threshold were classified as "Meets Loan Requirements." Ultimately, this multi-stage process identified 281 samples (from the total 829) as meeting the loan qualifications.

Using labels derived from unsupervised clustering as the ground truth for the supervised decision-tree test is methodologically valid. The initial clustering objectively revealed inherent group structures in the consumption behavior data without prior labels, identifying a core group (Cluster 0) with clear financial need. The subsequent MLDA refinement, based on PCA-weighted features and RF scoring, translated these nuanced consumption patterns into a robust, data-driven eligibility assessment. This combined approach ensures that the final binary label is not subjectively assigned but is grounded in the actual economic disparities and consumption characteristics manifested in the data. Consequently, it provides a reliable and objective foundation for training and evaluating the supervised decision tree model, effectively connecting the unsupervised discovery of patterns

with supervised predictive validation.

An optimal score threshold was determined via Receiver Operating Characteristic (ROC) curve analysis, and students whose scores exceeded this threshold were classified as "Meets Loan Requirements".

The MLDA pipeline implemented a sequential feature-to-score transformation: first, PCA was applied to reduce dimensionality and extract principal components from the multi-dimensional consumption behavior features, with component loadings serving as feature weights that reflect their relative importance in explaining variance in loan eligibility; these PCA-derived weights were then used to construct a weighted feature set, which served as input to a Random Forest classifier that computed a comprehensive loan eligibility score for each student by aggregating predictions across multiple decision trees (Niu & Cai, 2007). This PCA-to-RF scoring mechanism effectively transformed qualitative consumption patterns into quantifiable eligibility metrics, enabling precise discrimination of loan candidates from the ambiguous clusters.

4.2. Decision Tree Model Results

Running the established decision tree model on the 829 samples, splitting the data into training and testing sets, with the testing set accounting for 30%, yielded an identification accuracy of 97.992%. A classification report for the test set was generated, as detailed in Table 10:

Table 10. Decision tree model test report.

	precision	recall	F1-score	support
No	0.99	0.97	0.98	163
Yes	0.94	0.99	0.97	86
accuracy	\	\	0.98	249
macro avg	0.97	0.98	0.97	249
weighted avg	0.98	0.98	0.98	249

The confusion matrix (Confusion Matrix) (Kong & Jing, 2012) was also obtained, as detailed in Table 11:

Table 11. Confusion matrix of decision tree model.

Actual/Predicted	Not Meet Loan Requirements	Meet Loan Requirements
Not Meet Loan Requirements	True Negative (TN): 158	False Positive (FP): 5
Meet Loan Requirements	False Negative (FN): 1	True Positive (TP): 85

Regarding the 158 false negative cases (predicted as "Not Meet Loan Requirements" but actually meeting them), while the number is relatively high, it's important to note that these cases represent a conservative prediction approach. In the context of student loan approval, false negatives primarily result in missed

lending opportunities rather than actual financial losses, constituting a relatively low-risk error type. However, further optimization could focus on appropriately reducing false negatives to ensure eligible students are not unduly excluded from loan support.

The results indicate that the model produced only a very small number of erroneous predictions for both the "Yes" and "No" classes, demonstrating the model's high performance on the test set.

The experiment also simulated the test accuracy under different test set sizes, as detailed in Figure 3. The accuracy was found to be above 96% in all cases, further corroborating the accuracy of the cluster analysis model in identifying genuine loan eligibility.

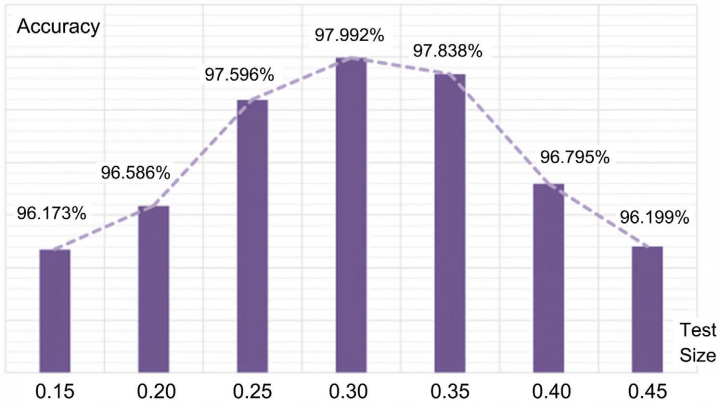


Figure 3. Test accuracy under different test set sizes.

4.3. Conclusion Verification

To verify the reliability and practical application value of the research results, the team conducted field visits to local bank business processes, focusing on the credit approval process. From the 281 samples identified as meeting loan qualifications, 80 cases were randomly selected for detailed analysis, and feedback was exchanged with credit approval experts. The study confirmed that 73 cases indeed met the bank's lending standards, resulting in a consistency rate of 91.25% with traditional manual screening results.

5. Conclusion

This study successfully constructed a precise loan qualification screening and risk prediction model for university student loans using the K-means algorithm and the decision tree model. The research results indicate that this model can effectively identify student groups with different economic statuses and consumption habits, providing a scientific basis for student loan disbursement. The K-means algorithm in the model effectively clustered student consumption behavior data into groups representing different economic statuses and consumption habits for loan eligibility prediction. Based on the clustering results, the decision tree model demonstrated high accuracy on the test set, validating the model's predictive per-

formance.

The model fully leverages the advantages of being data-driven. By deeply mining and analyzing consumer behavior data, it achieves precise screening of loan qualifications and risk assessment. Through exchanges and feedback with bank credit approval experts, the model achieved a consistency rate of 91.25% with traditional manual screening results. This high accuracy rate not only proves the model's reliability and practical application value but also provides powerful data support for banking business decisions, demonstrating the great potential of integrating technology with traditional finance. In the future, cooperation with financial institutions will continue to be deepened, and the algorithm model will be further optimized to improve the accuracy and efficiency of credit approval in broader scenarios, contributing to the development of financial technology. It is also looked forward to sharing these findings with more industry partners to jointly explore the boundless possibilities in financial technology.

Conflicts of Interest

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

References

Cao, G. G., Li, M. X., Chen, Y. et al. (2021). Improved Support Vector Machine Classification Method and Its Application in Primary Liver Cancer Screening. *Journal of Applied Sciences, 39,* 481-494. (In Chinese)

Fang, K. N., Wu, J. B., Zhu, J. P. et al. (2011). A Review of Random Forest Methodology Research. *Statistics & Information Forum, 26,* 32-38. (In Chinese)

Kong, Y. H., & Jing, M. L. (2012). Research on Classification Method Based on Confusion Matrix and Ensemble Learning. *Computer Engineering & Science, 34,* 111-117. (In Chinese)

Li, J., & Bai, J. (2014). Measurement of Regional Financial Agglomeration Level in China. *Seeking Truth, 41,* 52-58. (In Chinese)

Liang, J., Chen, J. H., Zhang, X. Q. et al. (2019). Anomaly Detection Based on One-Hot Encoding and Convolutional Neural Network. *Journal of Tsinghua University (Science and Technology), 59,* 523-529. (In Chinese)

Lin, H. M., & Zhang, W. L. (2005). Similarities and Differences between Principal Component Analysis and Factor Analysis and SPSS Software—Also Discussing with Comrades Liu Yumei, Lu Wendai, etc. *Statistical Research, No. 3,* 65-69. (In Chinese)

Liu, L. X., Zhang, T., & Yang, M. (2018). Calculation and Correction of Cronbach's α Coefficient Using Multilevel Models. *Chinese Journal of Health Statistics, 35,* 838-842. (In Chinese)

Long, W. J., Zhang, X. F., & Zhang, L. (2020). Business Process Clustering Method Based on K-Means and Elbow Rule. *Journal of Jianghan University (Natural Science Edition), 48,* 81-90. (In Chinese)

Luan, L. H., & Ji, G. L. (2004). Research on Decision Tree Classification Technology. *Computer Engineering, No. 9,* 94-96, 105. (In Chinese)

Meng, Z. J. (2017). *Research on the Spatiotemporal Behavior Characteristics of Linear Sports Activities in Beijing Central City Streets.* Ph.D. Thesis, Harbin Institute of Technology. (In Chinese)

Niu, Z. X., & Cai, K. Y. (2007). Design of a Multi-Level Distributed Intelligent Decision Support System Based on Regression Algorithm. *Computer Engineering and Design, No. 16,* 4004-4006. (In Chinese)

Sun, J. G., Liu, J., & Zhao, L. Y. (2008). Clustering Algorithm Research. *Journal of Software, No. 1,* 48-61. (In Chinese)

Wang, Q., Wang, C., Feng, Z. Y. et al. (2012). Survey of K-Means Clustering Algorithm Research. *Electronic Design Engineering, 20,* 21-24. (In Chinese)

Yao, Y. Y. (2021). *Research on Default Risk Scorecard for Shanghai Luohui Company's "Car Mortgage Loan" Business.* Ph.D. Thesis, Lanzhou University. (In Chinese)

Zang, X. H., & Li, Y. Q. (2012). Consumer Credit, Liquidity Constraints, and Consumption Behavior of Chinese Urban Residents—An Empirical Analysis Based on 2004-2009 Provincial Panel Data. *Economic Perspectives, No. 2,* 61-66. (In Chinese)

Zeng, W. Y., & Huang, B. Y. (2005). Analysis of Reliability and Validity of Survey Questionnaires. *Statistics and Information Forum, No. 6,* 13-17. (In Chinese)

Zhang, J. Y., Hu, J. H., & Huang, S. (2021). Application of CART Decision Tree Method in Energy Saving and Consumption Reduction of Coal-Fired Power Plants. *Control and Decision, 36,* 1232-1238. (In Chinese)

Zheng, J. J. (2014). Research Summary on Questionnaire Survey Method. *Theoretical Observation, No. 10,* 102-103. (In Chinese)

Zhu, L. J., Ma, B. X., & Zhao, X. Q. (2010). Cluster Validity Analysis Based on Silhouette Coefficient. *Computer Applications, 30,* 139-141, 198. (In Chinese)