ISSN 2959-6157

Credit Anomaly Detection Method based on Bayesian Networks

Xunwei Ran[1, *]

[1]Shenzhen College of International Education, Shenzhen, 518000, China
*Corresponding author: s22245.ran@stu.scie.com.cn

Abstract:

Due to the virtual nature of online transactions and the underdeveloped personal credit risk assessment system of Internet finance lending platforms in China, these platforms often struggle to verify the creditworthiness of applicants prone to fraud and default, leading to loan losses. As a result, risk control has become the core of Internet lending platforms. Modern risk control relies on statistical and data mining techniques to analyze and model data, uncover patterns that indicate loan defaults, and identify anomalies. This paper proposes a Bayesian network-based credit anomaly detection method, which detects anomalies by calculating and ranking the joint anomaly probabilities of sample instances. The technique operates under unsupervised learning and can effectively handle missing and imbalanced data. This method analyzes a loan credit dataset with 32,581 samples and 12 features of the customers. The result of the analysis demonstrates the feasibility and effectiveness of this method in detecting anomalies in online loans.

Keywords: Bayesian network; machine learning; default prediction.

1. Introduction

Online credit is a significant component of the internet finance sector, with many Internet financial companies launching online credit services. Online credit, as an efficient lending model, enhances the liquidity and accessibility of funds by effectively collecting idle capital and allocating it to small and medium-sized enterprises or individuals in need. However, this also imposes higher demands on the platform's credit risk management capabilities. According to data from Tianyancha, as of March 2022, there were 6,558 online lending platforms in China, with 1,621 providing personal credit services. As online lending platforms rapidly expand, the demand for robust credit risk control becomes increasingly critical.

Due to the virtual nature of online transactions and the underdeveloped credit risk assessment systems of Chinese internet financial platforms, it is challenging for platforms to verify applicants' credit status, making them prone to fraud and defaults. This leads to loan recoverability issues and economic losses for both platforms and clients. Currently, only 285 personal credit platforms are still operational. Apart from stringent regulatory oversight, platforms face severe credit risk issues.

On the one hand, these platforms have not integrated with credit reporting systems, lack access to credit information from other platforms, and thus have insufficient risk control capabilities. This is especially true for small internet financial platforms, which struggle with limited data collection capabilities and quality issues, making effective credit risk management difficult.

On the other hand, most small online credit platforms have not established comprehensive credit risk assessment systems and lack significant research investment in risk assessment models, resulting in poor performance in identifying low-credit clients. This leads to frequent defaults and, in severe cases, platform bankruptcy or suspension, causing financial losses for both platforms and customers. Therefore, internet financial credit platforms must develop credit risk assessment models tailored to their specific characteristics using available data, to monitor credit risk dynamics, detect anomalies, protect platform interests, and safeguard customer rights.

Internationally, researchers have long explored credit risk assessment for online lending. Fisher suggested that credit assessment should be based on practical experience to select features highly correlated with personal credit, which can classify the target group into different credit levels [1]. Durand applied similar ideas to classify loan risks into good and bad categories [2]. From the research on credit risk assessment methods in online lending, two main approaches can be identified: establishing evaluation indicator systems and developing predictive models. This paper focuses on developing accurate credit assessment models.

The "Peer Lending Risk Predictor" research team applied three machine learning algorithms-Random Forest, Logistic Regression, and Support Vector Machine (SVM)-and found that these algorithms provided better assessment re-

sults compared to traditional statistical methods [3]. Emekter et al. used the Lending Club data set and regression models to identify important variables, such as credit level, Fair Isaac Corporation (FICO) score, and loan interest rates, which were found to be closely related to credit risk assessment [4]. Malekipirbazari et al. used Random Forest algorithms for online credit risk modeling and prediction, showing higher accuracy in predicting non-default clients compared to FICO scores and Lending Club (LC) grades [5]. Subsequently, some artificial intelligence algorithms have been applied to the online credit field. Byanjankar used neural network models to classify data from the Bondora lending platform, finding higher accuracy compared to Logistic Regression [6]. Chee developed a Bayesian Network credit risk assessment model using data from a Singaporean online lending platform, demonstrating the effectiveness of Bayesian Networks in assessing credit risk [7]. Ding and Luo applied Stacking ensemble strategies to online lending default risk prediction models, significantly improving prediction accuracy compared to single learners [8]. Tan et al. used Gradient Boosting Decision Trees (GBDT) to build a credit assessment model for online lenders [9]. Li applied XGBoost for credit risk prediction, further enhancing accuracy and precision [10]. Chen used Easy Ensemble sampling methods to optimize Random Forest models, addressing data imbalance issues in online credit [11].

This paper explores a new credit anomaly detection method based on Bayesian Networks. This method identifies abnormal relationships within the Bayesian Network, calculates the joint probability values of sample instances, and detects anomalies in credit customer data by ranking these joint probability values. This approach helps alleviate funding issues for high-quality borrowers and provides financial safety and management assurance for internet financial credit platforms.

2. Methods

2.1 Data Source

This paper uses the dataset fetched from the Kaggle website (Credit Risk Dataset) which was updated in 2020 by Lao Tse. The dataset contains 32,581 entries with 12 features, related to personal and loan characteristics for credit risk assessment.

2.2 Variable Selection

The data includes variables such as personal income, home ownership status, employment length, loan purpose, interest rates, and loan status (Table 1).

Table 1. List of variables range

Variables	Range	Log
person_age	[20, 144]	X1
person_income	[4000, 6000000]	X2
person_home_ownership	(“RENT”,”OWN”,”MORTGAGE”,”OTHER”)	X3
person_emp_length	[0, 123]	X4
loan_intent	(“PERSONAL”,”MEDICAL”,”VENTURE”,”HOMEIMPROVEMENT”,”DEBTCONSOLIDATION”,”EDUCATION”)	X5
loan_grade	(A, B, C, D, E, F, G)	X6
loan_amnt	[500, 35000]	X7
loan_int_rate	[5.42, 23.22]	X8
loan_status	(0, 1)	X9
loan_percent_income	[0, 0.83]	X10
cb_person_default_on_file	(“Y”, “N”)	X11
cb_person_cred_hist_length	[2, 30]	X12

The dataset is representative of individuals seeking credit and includes financial indicators relevant to assessing credit risk. The primary commercial analysis indicators selected in Table 1 for this study include features relevant to the credit risk and the individual’s financial stability. These indicators were chosen based on their predictive power for loan default. These indicators help identify both high-risk loans and anomalies that may not conform to typical risk profiles.

2.3 Method Introduction

The study employs an unsupervised learning method

based on Bayesian networks for anomaly detection. The Bayesian network model identifies and quantifies anomalous relationships through causal logic, thereby detecting anomalies with practical significance.

This paper assumes that there are far more non-defaulting customers than defaulting customers in the dataset. This assumption aligns with real-world scenarios and allows the learning of causal relationships among various behaviors of non-defaulting customers to detect anomalies. This article uses unsupervised Bayesian Network-based learning method, which means that class labels can be ignored. The Bayesian network structure and parameter learning are performed on 11 variables, excluding the class label (defaulted or not).

The Bayesian network structure and parameter learning are obtained by GENIE software. The structure learning was performed using the K2 algorithm. After obtaining the Bayesian network structure, parameter learning can be conducted. The parameters in a Bayesian network include prior probabilities and conditional probabilities. The Expectation-Maximization (EM) algorithm was used for parameter learning in this study.

3. Results and Discussion

3.1 Bayesian Network Construction

The data is imported into the GENIE software and analyzed with its data distribution. For those continuous variables, manual binning is used to discretize the data. The Bayesian network structure (Fig.1) is built with K2 algorithm in GENIE which uses sample data to obtain a local Bayesian network structure score, and then use the greedy algorithm to obtain the Bayesian network structure with the highest score.

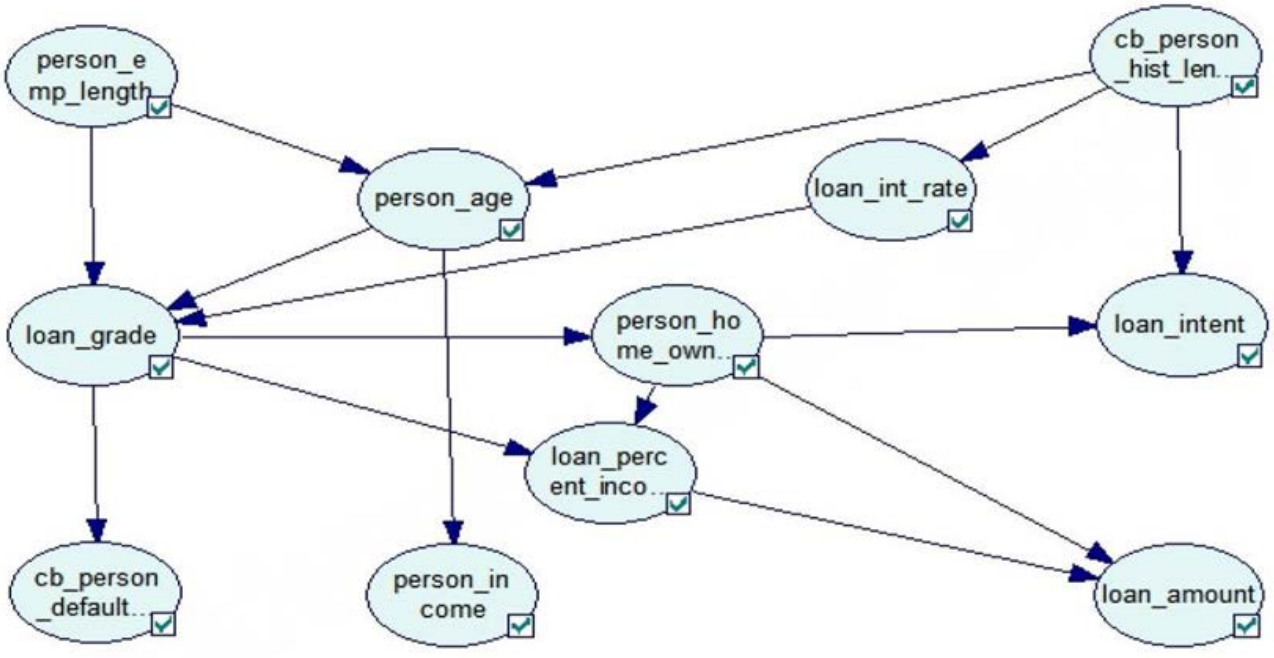

Fig. 1 Loan Customer Bayesian Network

This study uses the Expectation Maximization algorithm in GENIE to do the parameter learning. The EM algorithm can approximate the distribution of missing data to fill in the gaps and also complete the learning of the Bayesian network parameters (Fig. 2).

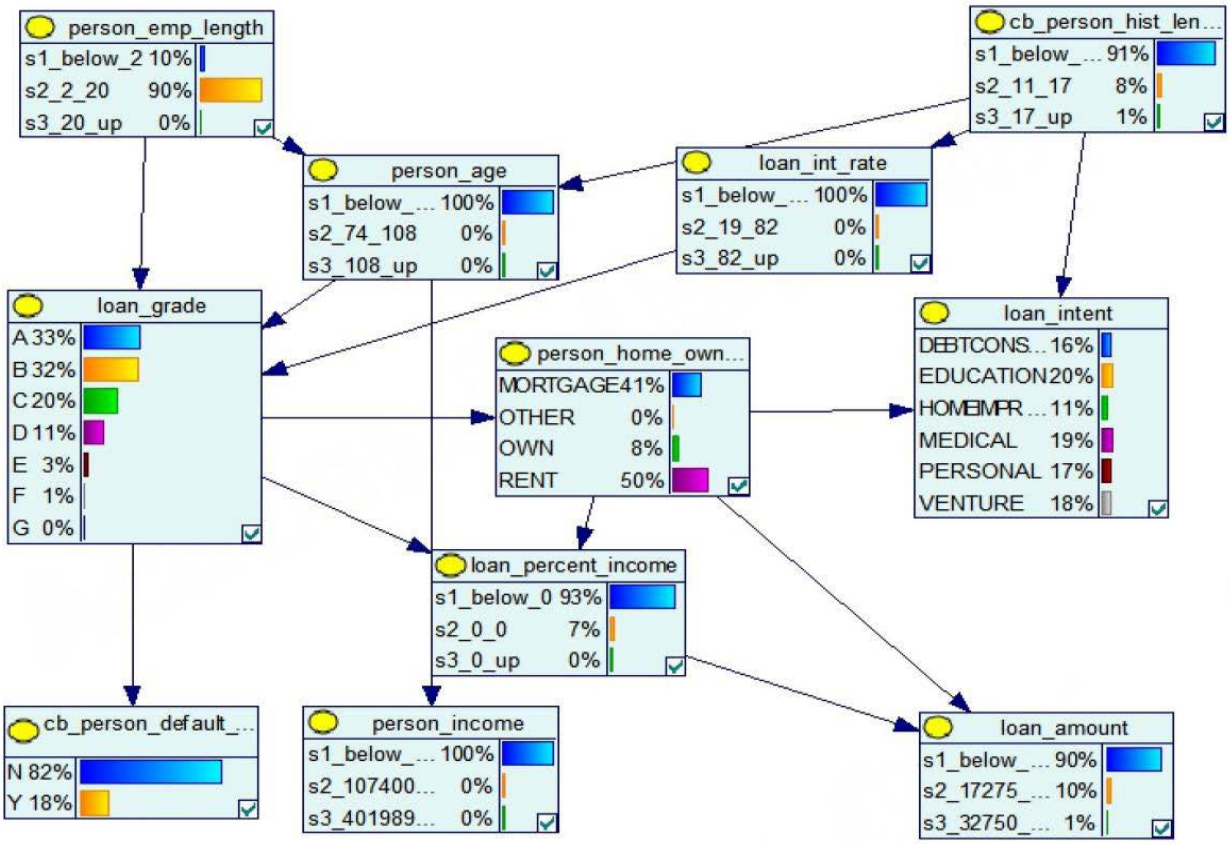

Fig. 2 Bayesian Network With prior Distribution

After parameter learning, the prior distribution and the conditional probability are obtained (some nodes' are displayed in the following table).

In Table 2, the prior distribution of the parent node, employment length, is displayed. The data is discretized into three states, where s1_below_2 means this person has an employment length shorter than 2 years, while s_2_20 states that a person's employment length is between 2 years and 20 years and s3_20_up indicates the working length longer than 20 years.

Table 2. Prior Distribution of Employment Length

Employment Length	Prior Distribution
s1_below_2	0.0956
s2_2_20	0.9029
s3_20_up	0.0014

As it shown in Table 3, the prior distribution of customers' credit history is classed into three states. Similarly, it denotes the past credit length below 11 years, between 11 and 17 years and longer than 17 years.

Table 3. Prior Distribution of Credit history length

Credit history length	Prior Distribution
s1_below_11	0.9079
s2_11_17	0.0833
s3_17_up	0.0087

The conditional probability of Age is demonstrated by Table 4. Take the first row of the table as an example, it

illustrates that the probability of a person aged below 74 years old, given that he has a working length shorter than 2 years and a credit history shorter than 11 years, is 0.9994.

Table 4. Conditional probability of Age based on Employment Length and Credit history length

Age	Employment length	Credit history length	Conditional probability
s1_below_74	s1_below_2	s1_below_11	0.9994
s1_below_74	s1_below_2	s2_11_17	0.9977
s1_below_74	s1_below_2	s3_17_up	0.9462
s1_below_74	s2_2_20	s1_below_11	0.9998
s1_below_74	s2_2_20	s2_11_17	0.9997
s1_below_74	s2_2_20	s3_17_up	0.9803
s1_below_74	s3_20_up	s1_below_11	0.9844
s1_below_74	s3_20_up	s2_11_17	0.8889
s1_below_74	s3_20_up	s3_17_up	0.3333
s2_74_108	s1_below_2	s1_below_11	0.0001
s2_74_108	s1_below_2	s2_11_17	0.0011
s2_74_108	s1_below_2	s3_17_up	0.0430
s2_74_108	s2_2_20	s1_below_11	0.0000
s2_74_108	s2_2_20	s2_11_17	0.0001
s2_74_108	s2_2_20	s3_17_up	0.0157
s2_74_108	s3_20_up	s1_below_11	0.0077
s2_74_108	s3_20_up	s2_11_17	0.0556
s2_74_108	s3_20_up	s3_17_up	0.3333
s3_108_up	s1_below_2	s1_below_11	0.0004
s3_108_up	s1_below_2	s2_11_17	0.0011
s3_108_up	s1_below_2	s3_17_up	0.0107
s3_108_up	s2_2_20	s1_below_11	0.0001
s3_108_up	s2_2_20	s2_11_17	0.0001
s3_108_up	s2_2_20	s3_17_up	0.0039
s3_108_up	s3_20_up	s1_below_11	0.0077
s3_108_up	s3_20_up	s2_11_17	0.0555
s3_108_up	s3_20_up	s3_17_up	0.3333

3.2 Anomaly Detection

In Bayesian networks, the joint probability of events can be broken down into the product of prior probability and conditional probability. When both probabilities are categorized as either “low” or “high,” four scenarios arise, which are low prior with high conditional probability, high prior with low conditional probability, high prior with high conditional probability, and low prior with low conditional probability

Typically, low joint probabilities indicate anomalies, especially when there’s a conflict between prior and conditional probabilities. Scenarios where the prior is low but the conditional is high, or vice versa, are key indicators of anomalies. Scenarios with both low probabilities are seen as noise, and those with both high indicate strong correlations rather than anomalies. Identifying anomalies in the data requires focusing on cases where these conflicts occur, using parameters such as minimum prior probability (minprior), minimum conditional probability (minconf),

and maximum conditional probability (maxconf).

Based on the Bayesian network model, whether the relationship between variables is abnormal is identified and judged in the following process. Firstly, output the prior probability and conditional probability of the Bayesian network and calculate the minprior of each node (Table 5). Then, Calculate the conditional probabilities of all nodes and compare them with minconf and maxconf. Based on the above two rules, the child nodes with abnormal relationships are screened for joint probability calculation, and then the logarithm is taken to calculate the abnormal union probability. Repeat these steps for each sample.

Table 5. Minimum prior probability of each node

Node Name	Node State	Minprior
X1	s2_74_108	0.0002
X2	s3_4019892_up	0.0001
X3	OTHER	0.0033
X4	s3_20_up	0.0014
X5	HOMEIMPROVEMENT	0.1105
X6	G	0.0020
X7	s3_32750_up	0.0062
X8	s3_82_up	0.0001
X10	s3_0_up	0.0031
X11	Y	0.1764
X12	s3_17_up	0.0087

3.3 Detailed Analysis

If the node, person_age, is taken as an example, the min prior of its two parent nodes, cb_person_cred_hist_length and person_emp_length, is 0.0087 and 0.0014 according to Table. 5. Set mincof=0.07, maxconf=0.8. If there is a sample that has a state of person_age (X1), person_emp_length (X4), and cb_person_cred_hist_length (X12) equals to (s2_74_108, s1_2_20, s1_below_11), the prior distribution and the conditional probability can be found, which is:

$$P(X4 = s1\_2\_20) = 0.9029, P = (X12 = s1\_below\_11) = 0.9080 \quad (1)$$

$$P(X1 = s2\_74\_108 \mid X4 = s1\_2\_20, X12 = s1\_below\_11) = 0.00001246 \quad (2)$$

Thus, it satisfies scenario 2, in which is prior probability is greater than min prior, and the conditional probability is less than minconf. The anomaly arises because, although each individual characteristic (employment length and credit history length) is common and not abnormal by itself, the combination of these characteristics with the person's age being between 74 and 108 years old is extremely rare. In a real-world context, it is unusual for someone who is that old to have such a short employment length (2 to 20 years) and a short credit history (less than 11 years). As a result, the risk of default is significantly high for these samples. Then the joint probability can be calculated.

If,

$$P(X1 = s2\_74\_108) = X, P(X4 = s1\_2\_20) = Y and P(X12 = s1\_below\_11) = Z \quad (3)$$

Then,

$$P(X,Y,Z) = P(X \mid Y,Z) \times P(Y) \times P(Z) = 0.9029 \times 0.9080 \times 0.00001246 = 0.0000125 \quad (4)$$

The same method is applied to the remaining 9 child nodes to calculate the joint probability. Then rank the logarithm of each joint probability in ascending value.

In the actual data set, there are 7108 default samples in 32,581 samples, which is 21.8% of the data set. Using the anomaly detection algorithm in this article, there are 5538 samples are default samples in the first 7100 samples of the ranking. This gives people an anomaly detection accuracy of 78%, which proves the feasibility and precision of the algorithm.

4. Conclusion

In summary, this paper uses the method based on the relationship between prior probability and conditional probability in a Bayesian network to analyze the sample, where the prior probability represents the support from the par-

ent node to a specific state of the child node and the conditional probability indicates the likelihood of that state occurring in the child node. Due to the causal relationship between the parent node's prior probability and the child node's conditional probability, there is also a causal relationship between support and likelihood. Through this reasoning, all relationships within the Bayesian network can be classified and analyzed, allowing the identification of two types of anomalous relationships. These anomalous relationships can then be used to filter all anomalous relationships within the Bayesian network, achieving the goal of anomaly detection. This unsupervised learning method does not rely on sample labels, and the Bayesian network can reasonably handle missing data and imbalanced samples, ensuring that learning performance is not affected. This avoids the distortion that can occur in credit loan default prediction when oversampling a small number of default samples.

References

[1] CFisher R A. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 1936, 7(2): 179-188.

[2] David V C, Loring M. Extragenital lesions of lymphogranuloma inguinale. Journal of the Americal Medical Association, 1936, 106(22): 1875-1879.

[3] Tsai K, Ramiah S, Singh S. Peer lending risk predictor. Palo Alto: Stanford University CS229 Project Report, 2014.

[4] Emekter R, Tu Y, Jirasakuldech B, et al. Evaluating credit risk and loan performance in online Peer-toPeer (P2P) lending. Applied Economics, 2015, 47(1): 54-70.

[5] Aksakalli, Vural, Malekipirbazari, et al. Risk Assessment in Social Lending via Random Forests. Expert Systems with Applications, 2015, 42(10): 4621-4631.

[6] Byanjankar A, M Heikkilä, Mezei J. Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach Computational Intelligence, IEEE Symposium. IEEE, 2015: 719-725.

[7] Leong C K. Credit Risk Scoring with Bayesian Network Models. Computational Economics, 2016, 47(3): 423-446.

[8] Ding Lan, Luo Pinliang. Research on Default Risk Early Warning in P2P Lending Based on Stacking Ensemble Strategy. Investment Research, 2017, 36(04): 41-54.

[9] Tan Zhongming, Xie Kun, Peng Yaopeng. Research on Credit Risk Assessment of P2P Borrowers Based on the Gradient Boosting Decision Tree Model. Soft Science, 2018, 32(12): 136-140.

[10] Chang Li. Research on Credit Risk Prediction in Internet Lending: Based on XGBoost Algorithm. Modern Business, 2019, 8: 86-87.

[11] Zhang Chen. Research on P2P Default Prediction Based on Ensemble Learning. China Prices, 2020, 5: 71-74.