science, engineering and health studies

Research Article

Science, Engineering and Health Studies
https://lio1.tci-thaijo.org/index.php/sehs
ISSN (Online): 2630-0087

Association rule mining framework for financial credit-risk analysis in peer-to-peer lending platforms

Tanatorn Tanantong[1,2] and Pakin Loetwiphut[1]*

[1] Thammasat Research Unit in Data Innovation and Artificial Intelligence, Thammasat University, Pathum Thani 12120, Thailand
[2] Department of Computer Science, Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand

ABSTRACT

This study demonstrates a comprehensive framework for financial credit-risk analysis in the context of peer-to-peer (P2P) lending, which is a rapidly expanding industry that enables people to lend and borrow money not using conventional financial institutions. However, the considerable default risk associated with P2P lending shows serious difficulties for investors. Difficulties can be overcome through a framework based on feature selection and data discretization approaches for mining association rules from P2P lending data. Providing useful information for credit-risk analysis in P2P lending, obtained association rules can be used to identify trends and connections in the data that indicate a borrower's creditworthiness and likelihood of payback.

Keywords: P2P lending; feature selection; data discretization; association rule mining; Apriori algorithm; financial risk-analysis

***Corresponding author**:
Tanatorn Tanantong
tanatorn@sci.tu.ac.th

***Received**: 24 March 2023*
***Revised**: 28 July 2023*
***Accepted**: 5 September 2023*
***Published**: 21 December 2023*

***Citation**:
Tanantong, T., and Loetwiphut, P. (2023). Association rule mining framework for financial credit-risk analysis in peer-to-peer lending platforms. Science, Engineering and Health Studies, 17, 23020006.

1. INTRODUCTION

Peer-to-peer lending (P2P lending) is a form of alternative lending in which individuals can borrow and lend money directly with each other, without the involvement of banks or other traditional financial institutions. P2P lending platforms connect borrowers with investors, who can then choose to fund the borrowers' loans. The borrowers and lenders interact directly through an online platform, which acts as a facilitator and intermediary, collecting the funds and distributing the payments. According to Acumen Research and Consulting (2023), P2P lending growth is driven by the increasing availability of digital technologies and the internet. This growth has increased rapidly in recent years, with the global market size reaching USD 82.3 billion in 2021. It is predicted to escalate significantly to USD 804.2 billion by 2030. This increasing availability of digital technologies and the internet has made it easier for borrowers and lenders to connect and transact with each other.

One of the most significant risks of P2P lending is credit risk, specifically regarding the risk that borrowers will default on their loans or fail to make their payments as agreed. This is because P2P lending platforms typically have less stringent lending criteria than traditional financial institutions. The default rate on P2P lending (Ko et al., 2022) is generally higher than that of traditional bank loans. In 2020, the Bank of Thailand (Bank of Thailand, 2022) mandated that businesses operating on P2P lending platforms must undergo testing services within a limited scope under the regulatory sandbox, through which providers can experiment with new technologies in a controlled environment. This fosters innovation in finance while ensuring appropriate risk management and user protection, which serves as a mechanism for testing and developing new technologies that support financial services.

Silva et al. (2019) demonstrated the data mining application in commercialization for extracting significant patterns from customer behaviors in small and medium-sized enterprises (SMEs), particularly in credit-risk analysis within the P2P lending context. Several studies (Vinod et al., 2016; Setiawan, et al., 2019; Riguzzi et al., 2020; Guo et al., 2020) have supported data mining utilization in credit-risk analysis of P2P lending. Vinod et al. (2016) presented a financial credit-risk analysis of the P2P lending system of a "Lending Club 2007–2020Q3" (n.d.) company. The study aimed to predict the likelihood of LC-associated default or risky credits publicly available from a 2013–2015 loan applications dataset. Setiawan et al. (2019) developed P2P lending platforms for loan prediction. The research proposed a tree-based classification method to predict loan defaults using data collected from LC platform (Yash, 2020). To enhance the classification performance, the study employed binary particle swarm optimization with a support vector machine for feature selection. Riguzzi et al. (2020) compared the effectiveness of different machine learning models in predicting default risks by investigating the determinants of credit risk in P2P lending platforms. They concluded that positive customer recommendations can be utilized to minimize default risks, while abnormal returns are a significant trigger for platform default risks.

Association rules mining (ARM) is another data mining technique widely applied to various domains, including finance that can potentially be used to identify patterns and relationships in large data sets (Tanantong and Ramjan, 2021). Desai and Kaiwade (2018) applied the Apriori algorithm to customer data for identifying behavior patterns for increasing deposits for banks. The results of the algorithm were used to construct a model, guiding managers in making decisions. Zhang (2021) conducted a study on the relationship between financial ratios and bankruptcy risk in SMEs listed on the New Third Board. By utilizing ARM with Apriori, the study discovered that key indicators, such as current ratio, quick ratio, return on assets, and cash ratio, exhibit a significant correlation with the risk of bankruptcy in these enterprises. Based on ARM, Hsueh and Kuo (2017) addressed several challenges, such as imbalanced data, high-dimensional feature space, and discretization of continuous variables, in order to identify patterns and relationships in the data through analysis of numerical data from zopa.com, the UK's largest P2P lending platform. One significant challenge was the presence of a large number of non-default loans and a small number of default loans.

To handle the imbalanced data issue, Chen (2021) applied three resampling methods, namely, synthetic minority oversampling technique (SMOTE), near-miss, and manual 1:1 random selection for developing credit-default-risk classification models in P2P lending. Unbalanced data is a considerable obstacle to association rule mining on big datasets. According to Mahdi et al. (2022), this issue makes the identification of unusual items or rare rules difficult, which do not appear regularly in data. It also emphasizes the importance of these rare items in providing essential and sensitive information in various fields, such as medicine. To address the imbalanced data issue in ARM, Chai et al. (2023) presented a credit-risk analysis method combining the Apriori algorithm and SMOTE-nominal continuous technique. This proposed method can help make mining the defaulting feature attributes of existing small business due to the data characteristics of small default costumers and large non-defaulting costumers less challenging.

Additionally, P2P lending data often contains a large number of features, making it difficult to identify the relevant features and patterns (Cheng et al., 2021). To address this, Suhada and Devasia (2022), in an approach that includes a feature selection module with recursive feature elimination with cross validation (RFECV) and a data synthesis module with borderline-SMOTE, proposed a credit-risk prediction system for the LC data. Discretization of continuous variables is another challenge faced by ARM in P2P lending data. Tan (2018) presented the two-step clustering (TSC) approach for data discretization in association rule mining. Unlike traditional methods, TSC automatically determines discretization intervals by considering the unique distribution of each attribute.

This study presents an approach for dealing with difficulties of P2P lending. It identifies important variables affecting loan default and borrower creditworthiness by using association rule mining with Apriori. Feature selection utilizing RFECV is used to omit weak or redundant features while at the same time keeping high-impact features. SMOTE is used to address imbalance of data, while data discretization is utilized to manage the dataset's continuous numerical properties. The obtained association rules from this study are valuable for credit-risk analysis purposes, offering insightful information that is beneficial for P2P platforms, investors, and regulators. The integration of these methodologies contributes significantly to the field of knowledge in risk analysis, making it a valuable tool for stakeholders in the P2P lending domain, where, by reducing default risks, optimizing investments, and improving risk management procedures, this research thus proves to be highly beneficial.

2. MATERIALS AND METHODS

This section explains the process of constructing risk analysis rules (Figure 1).

2.1 Data preprocessing

This study used a dataset obtained from the LC platform (Yash, 2020) to conduct research on P2P lending. The dataset includes 141 unique characteristics and a huge collection of 2,925,493 observations spanning the years 2007–2020. Specific examples are given in Table 1 to give an understanding of the type of data used in this study. The cloud platform Google Colab Pro, known for its powerful capabilities, was used in this investigation. While the study leverages a Tesla P100-PCIE-16GB GPU, contributing to the overall computational efficiency, the hardware environment utilized consists of an Intel(R) Xeon(R) CPU clocked at 2.20GHz, coupled with 25 GB of memory. The experiment presented in this study was developed within the scikit-learn framework, utilizing Python 3.10 programming language.

For the preprocessing of the dataset, attribute removal and missing value removal techniques were applied. Initially, in the attribute removal process, irrelevant columns were carefully eliminated, including identification numbers, postal codes, and cities of residence, to reduce noise in the dataset. A further 44 attributes that exhibited more than 50% missing values were subsequently excluded

to mitigate potential bias and inaccuracies in the analysis. Regarding the missing value removal process, all rows containing missing values were excluded to preserve the accuracy and reliability of the results. Consequently, the dataset was narrowed down to focus on two loan statuses: fully paid and charged off. As a result, the final dataset comprised 672,237 entries, featuring 538,917 fully paid loans and 133,320 charged off loans, with 97 attributes retained for analysis.

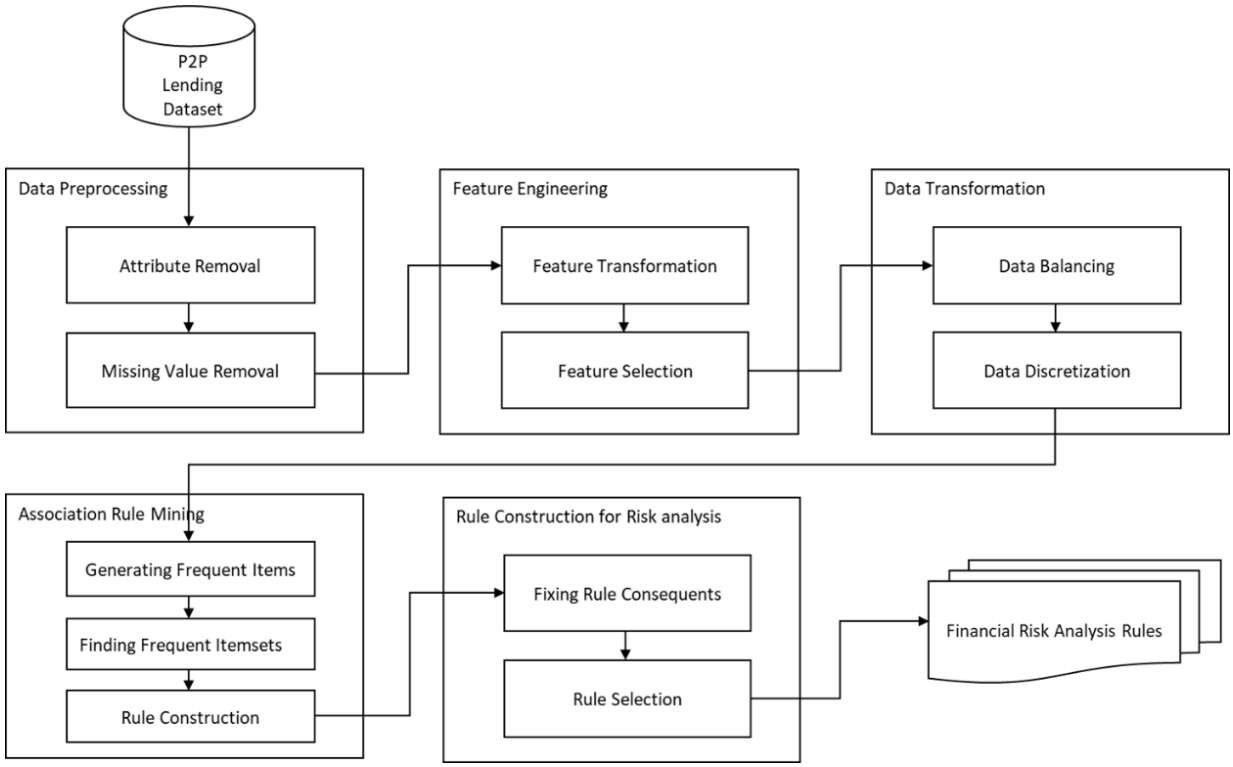

Figure 1. Components of the proposed framework for financial credit-risk analysis

Table 1. Examples of P2P lending data obtained from LC (Yash, 2020)

id	loan_amnt	deferral_term	revol_bal_joint	dti_joint	zip_code	addr_state
000203532	21,000	n/a	13,343	n/a	606xx	IL
000663451	6,400	n/a	n/a	n/a	365xx	AL
000443279	4,000	n/a	35,827	n/a	775xx	TX
000742186	11,500	n/a	30,324	n/a	238xx	VA
000597005	8,000	n/a	n/a	n/a	070xx	NJ
000997918	17,400	n/a	47,995	n/a	628xx	IL
000648213	20,000	n/a	0	17.67	021xx	MA
000792768	28,000	2	n/a	n/a	117xx	NY
000762087	11,200	n/a	10,052	n/a	145xx	NY
000847527	19,075	n/a	n/a	n/a	481xx	MI
000749019	3,000	n/a	n/a	17.92	617xx	IL
000569379	7,750	2	26,220	19.02	430xx	OH
000406211	9,000	n/a	n/a	n/a	774xx	TX
000494432	18,000	n/a	15,490	n/a	787xx	TX

Note: Id = identification; loan_amnt = loan amount; deferral_term = deferral term; revol_bal_joint = revolving balance for joint accounts; dti_joint = debt-to-income ratio for joint accounts; zip_code = zip code; addr_state = address sta

2.2 Feature engineering

The two main steps in feature engineering are feature transformation and feature selection. Feature transformation is a critical aspect in the methodology of conducting P2P lending risk analysis. It involves creating new variables or transforming the existing ones in a dataset to better capture relevant information and enhance the accuracy of predictions. Feature transformation can produce outcomes that are more insightful and understandable, increase the quality of the analysis, and minimize redundant information in the depiction of creditworthiness.

One significant variable in the LC data is the FICO score, a credit score created by the Fair Isaac Corporation (FICO) that assesses creditworthiness and assigns a score between 300 and 850. According to Kumar and Gunjan (2020), the Kumar score is based on credit use, credit history, and repayment of credit. In our analysis, with new variables capturing the borrower's creditworthiness without repeated information indicating the mean value of each range, redundancy in the representation of FICO scores was seen, as they were duplicated in two features, namely fico_range_low and fico_range_high, as well as last_fico_range_low and last_fico_range_high. To address this redundancy, the representation was streamlined by creating new variables, namely fico_score and last_fico_score.

During feature selection, 21 columns with >70% unique values were initially eliminated due to their inefficient value distribution, which hindered the identification of meaningful patterns and associations through Association Rule Mining (ARM). Notably, columns such as pub_rec_bankruptcies, recoveries and collection_recovery fee were among those removed due to their pronounced concentration of unique values. This refinement substantially enhanced its interpretability and resulted in more informative outcomes. In an approach in which RFECV was employed, which makes use of five-fold cross validation, high-impact elements were kept while redundant or weak features were removed after each iteration. By determining the ideal amount of features to employ during cross validation with RFECV, there was no concern regarding choosing the duplicate features, and the training time was significantly decreased (Lee et al., 2023), preventing overfitting during model training.

Figure 2 illustrates the experiments, indicating that prediction performance of LightGBM model reached its peak with 43 features when utilizing RFECV. Nevertheless, due to resource constraints on Google Colab Pro, a feature selection with 41 features (40 features for constructing models and 1 targeted feature) was used, which demonstrated accuracy close to maximum (Figure 3). Consequently, this set of 40 features was utilized in the final model for the financial risk analysis on P2P lending. Despite the slight reduction in the number of features, the selected set still delivered robust predictive capabilities, making it suitable for analysis given the computing limitations.

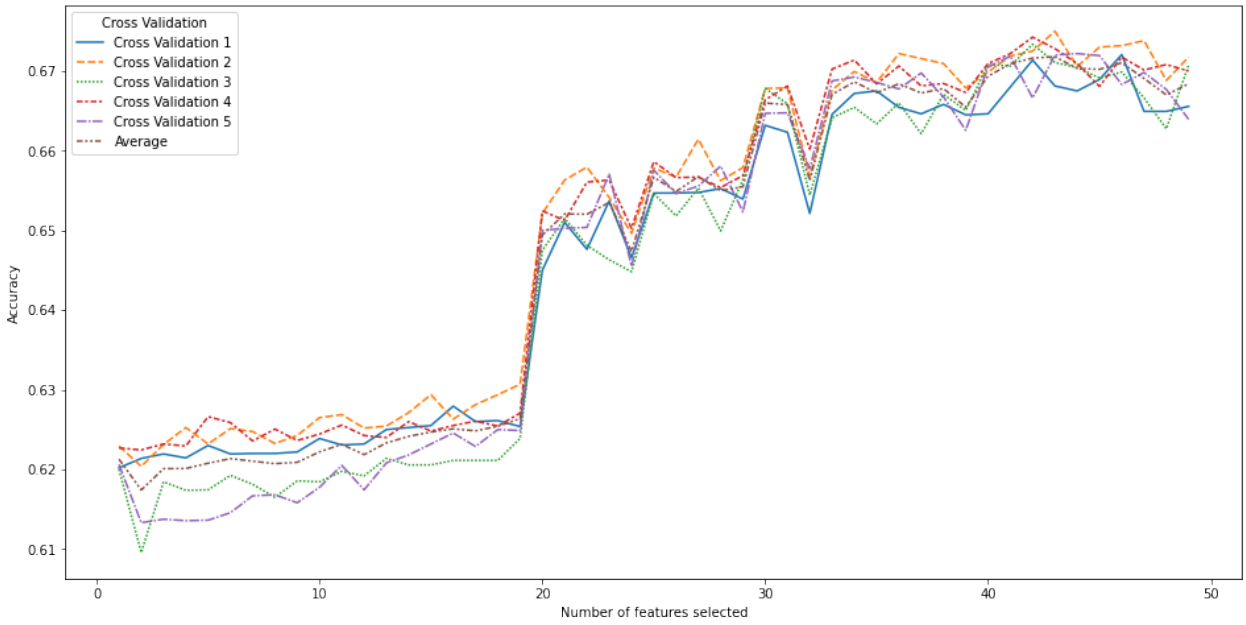

Figure 2. Number of features taken by RFECV based on LightGBM and five-fold cross validation

2.3 Data transformation

To address the issue of imbalanced data, the oversampling technique SMOTE was employed. To mitigate bias in the analysis ensuring that the predictive model is not skewed towards the majority class, SMOTE, initially proposed by Chawla et al. (2002), involves the synthetic generation of new samples for the minority class. This is a generation process that is based on the k-Nearest Neighbors (k-NN) algorithm, an algorithm that creates new samples in proximity to existing minority samples, as resulted in Table 2. In data discretization, and in a method that is particularly useful when the number of clusters is unknown and needs to be determined from the data (Tan, 2018), the 2-step clustering method in SPSS is an effective technique for grouping similar observations based on their characteristics.

In the first step, a hierarchical clustering analysis is performed to determine the maximum number of clusters, using a distance measure such as Log-likelihood and a cluster criterion such as BIC (Bayesian information criterion) to evaluate the clustering solutions. Once the maximum number of clusters is determined, the second step involves a k-means algorithm partitioning the data into k clusters, with each observation assigned to the cluster with the nearest mean, where k-means clustering analysis is using the determined number of clusters. The final result is a set of k clusters, each representing a group of similar observations.

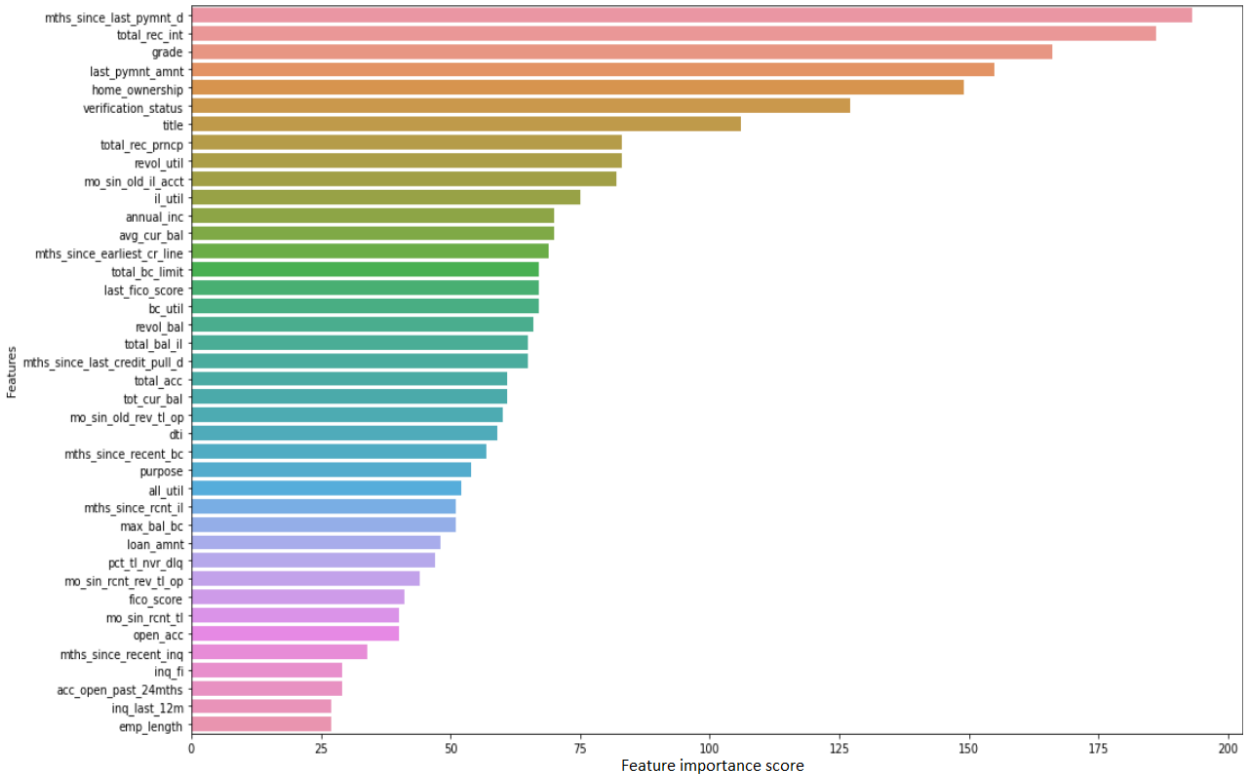

Figure 3. Feature importance ranking by RFECV (with 40 features)

Table 2. Percentage of default data before/after SMOTE

Before SMOTE			***After SMOTE***		
Loan status	**Number of data**	**Percentage**	**Loan status**	**Number of data**	**Percentage**
Fully Paid	538,917	80.17%	Fully Paid	538,917	50.00%
Default	133,320	19.83%	Default	538,917	50.00%
Total	672,237	100.00%	Total	1,077,834	100.00%

2.4 Association rules mining and rule construction

Association rule mining with the Apriori algorithm, proposed by Agrawal et al., (1993), discovers interesting relationships between variables in large databases. An association rule is represented as $X \rightarrow Y$, where X and Y are disjoint itemsets (with no items in common). The rule suggests that when X appears, Y also tends to appear. The three traditional measures for evaluating association rules are support, confidence, and lift. The support of the rule is the fraction of the database that contains X and Y. The confidence of an association rule $X \rightarrow Y$ is the proportion of the transactions containing X, which also contains Y. Lift measures the degree to which the X and Y occur together is greater than what would be expected if they were independent. If the lift $(X \rightarrow Y) < 1$, then X occurrence is negatively correlated with Y occurrence. Conversely, if lift $(X \rightarrow Y) > 1$, then X occurrence is positively correlated with Y occurrence. A lift $(X \rightarrow Y)$ close to 1 implies that X *and* Y are independent, thus not correlated. The formal definitions of these metrics are described in Equations (1), (2), and (3).

Support is defined as

$$\text{Support } (X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \quad (1)$$

Confidence is defined as

$$\text{Confidence } (X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (2)$$

Lift is defined as

$$\text{Lift } (X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X) \times \sigma(Y)} \quad (3)$$

3. RESULTS AND DISCUSSION

For association rule mining based on Apriori, the minimum support threshold was defined as 0.2, requiring an itemset to appear in at least 20% of the transactions to be considered frequent. By doing so, uninteresting rules that do not meet the minimum level of support or lift were filtered out. Additionally, the minimum lift threshold was set to 1, signifying a positive correlation between the antecedent and consequent of an association rule. The interesting rules that passed these thresholds underwent further evaluation to determine their relevance in financial risk analysis for P2P lending.

3.1 Association rule mining results

A total of 9,158,342 association rules were created using the Apriori algorithm. By using the developed association rules to recognize major financial risk elements and their combinations that affect the likelihood of loan status and credit score in P2P lending, consequents were established for financial risk analysis.

Table 3 presents the count of association rules for varying minimum support and confidence values. When the confidence range increases, the number of association rules also decreases. In the context above, support denotes the frequency of occurrence of itemsets in the dataset, while confidence represents the conditional probability of the consequent item occurring given the antecedent item. This occurs because the more confident and meaningful rules generated give higher confidence values indicating stronger relationships between the antecedent and consequent items.

Association rules analysis reveals that a significant portion of the rules falls within the 0.6–0.7 confidence range, indicating the generated rules may not be sufficiently strong. However, there are exceptions, as certain rules exhibit high confidence levels (>0.9) within the lowest support value range (0.2–0.3). To focus on the most interesting and meaningful rules, a filtering approach was applied by setting a minimum support threshold of 0.5. This required high confidence levels for each number of itemsets, with these high-confidence rules suggesting more robust relationships between the antecedent and consequent items. The above filtering process led to the identification of 19 particularly interesting rules (Tables 4–7). These rules are expected to provide valuable insights into the underlying patterns within the data, facilitating a deeper understanding of the relationships among variables in the context of financial risk analysis for P2P lending.

Table 3. Number of association rules when considering different support and confidence values

Support	Number of association rules					
	Conf. = 1	0.9 ≤ Conf. <1	0.8 ≤ Conf. <0.9	0.7 ≤ Conf. <0.8	0.6 ≤ Conf. <0.7	0.5 ≤ Conf. <0.6
0.2–0.3	570	539,161	919,967	1,435,224	3,404,638	2,307,622
0.3–0.4	7	46,781	76,372	108,231	131,286	131,466
0.4–0.5	0	5,696	8,949	11,549	12,814	10,893
≥0.5	0	1,146	1,716	2,083	1,667	504
Total	577	592,784	1,007,004	1,557,087	3,550,405	2,450,485

Table 4. Examples of two-itemset rules with a support value ≥0.5 and selected by highest confidence

Antecedents	Consequents	Support	Confidence	Lift
{'tot_cur_bal_low'}, {'total_bal_il_low'}	{'avg_cur_bal_low'}	0.50	0.999	1.33
{'inq_fi_low'}, {'total_bal_il_low'}	{'inq_last_12m_low'}	0.52	0.930	1.13
{'total_bc_limit_low'}, {'total_rec_prncp_low'}	{'last_pymnt_amnt_low'}	0.52	0.998	1.23
{'total_bc_limit_low'}, {'annual_inc_low'}	{'max_bal_bc_low'}	0.55	0.957	1.15
{'total_bc_limit_low'}, {'annual_inc_low'}	{'revol_bal_low'}	0.58	0.993	1.05
{'tot_cur_bal_low'}, {'avg_cur_bal_low'}	{'total_bal_il_low'}	0.50	0.964	1.15
{'max_bal_bc_low'}, {'fico_score_low'}	{'total_bc_limit_high'}	0.56	0.933	1.14

Table 5. Examples of three-itemset rules with a support value ≥0.5 and selected by highest confidence

Antecedents	Consequents	Support	Confidence	Lift
{'mths_since_recent_bc_low'},{'revol_bal_low'}, {'inq_fi_low'}	{'inq_last_12m_low'}	0.50	0.921	1.12
{'mths_since_recent_bc_low'},{'revol_bal_low'}, {'total_rec_prncp_low'}	{'last_pymnt_amnt_low'}	0.53	0.998	1.23
{'mths_since_recent_bc_low'},{'total_bc_limit_low'}, {'annual_inc_low'}	{'max_bal_bc_low'}	0.50	0.962	1.11
{'total_bc_limit_low'},{'annual_inc_low'}, {'max_bal_bc_low'}	{'revol_bal_low'}	0.55	0.994	1.05
{'revol_bal_low'}, {'max_bal_bc_low'}, {'total_acc_low'}	{'total_bal_il_low'}	0.52	0.934	1.11
{'revol_bal_low'}, {'max_bal_bc_low'}, ' {'fico_score_low'}	{'total_bc_limit_high'}	0.55	0.939	1.17

Table 6. Examples of four-itemset rules with a support value ≥0.5 and selected by highest confidence

Antecedents	Consequents	Support	Confidence	Lift
{'mths_since_recent_bc_low'}, {'revol_bal_low'}, {'total_rec_prncp_low'}, {'mo_sin_rcnt_rev_tl_op_low'}	{'last_pymnt_amnt_low'}	0.50	0.998	1.23
{'mths_since_recent_bc_low'}, {'avg_cur_bal_low'}, {'total_bc_limit_low'}, {'mo_sin_rcnt_rev_tl_op_low'}	{'max_bal_bc_low'}	0.51	0.961	1.11
{'avg_cur_bal_low'},{'total_bc_limit_low'}, {'max_bal_bc_low'}, {'last_pymnt_amnt_low'}	{'revol_bal_low'}	0.51	0.994	1.05
{'mo_sin_rcnt_rev_tl_op_low'}, {'revol_bal_low'}, {'max_bal_bc_low'}, {'fico_score_low'}	{'total_bc_limit_high'}	0.50	0.936	1.16

Table 7. Examples of five-itemset rules with a support value ≥0.5 and selected by highest confidence

Antecedents	Consequents	Support	Confidence	Lift
{'mths_since_recent_bc_low'},{'revol_bal_low'}, {'total_bc_limit_low'},{'mo_sin_rcnt_rev_tl_op_low'}, {'avg_cur_bal_low'}	{'max_bal_bc_low'}	0.50	0.962	1.11
{'mths_since_recent_bc_low'},{'total_bc_limit_low'}, {'max_bal_bc_low'},{'mo_sin_rcnt_rev_tl_op_low'}, {'avg_cur_bal_low'}	{'revol_bal_low'}	0.50	0.993	1.05

After P2P lending data ARM, several interesting patterns have been uncovered. One such intriguing rule is a finding that can be valuable for lenders in identifying potential borrowers who may pose a higher risk of defaulting on their loans (row 7 of Table 4). This indicates that borrowers with low maximum balances on their bankcards and low FICO scores are likely to have high total bankcard limits. The rule found in row 1 of Table 5 is noteworthy. It reveals if a borrower has opened a bankcard account recently, has a low revolving balance, and a low number of inquiries to financial institutions, they are likely to have made few inquiries in the past 12 months. The support value of this rule is 0.50, indicating that 50% of transactions in the dataset contain all of these antecedents and consequents. Moreover, the confidence of this rule is 0.921, implying that when these antecedents are present, there is a high probability (92.1%) that the consequents will also be present.

The rule found in row 3 of Table 6 is of particular significance. It is intriguing as it sheds light on the potential relationship between a borrower's credit utilization and their current account balances and payment behavior. It reveals that borrowers with lower average current balances, total bankcard limits, maximum bankcard balances, and last payment amounts are likely to have lower revolving balances. The support value of 0.51 indicates that this combination of antecedents and consequents occurs in 51% of the transactions in the dataset, which is relatively prevalent; the high confidence value of 0.994 indicates that when the antecedents are present, there is high likelihood (99.4%) of the presence of consequents. This rule suggests that borrowers who are capable of maintaining lower balances and demonstrating consistent payment behavior are likely to have lower revolving balances.

The rule described in row 1 of Table 7 holds significant implications. It indicates that if a borrower exhibits a low number of months since their most recent bankcard account opened, in addition to low values for their revolving balance, total bankcard credit limit, month since most recent revolving account opened, and average current balance, they are likely to have a low maximum current balance. Interestingly, this rule suggests that a borrower's credit utilization habits plays a vital role in determining their maximum current balance. In row 1 of Table 4, the rule suggests that borrowers with low total current balance and balance of installment loans are likely to have a low average current balance. This finding is sensible, as a borrower's current balance represents the sum of outstanding debts; low current balance implies that the borrower has fewer outstanding debts, logically corresponding to having low total balances across different loan types. Similarly, the rule in row 3 of Table 4 indicates that borrowers with a low total bankcard limit and a low total received principal are likely to have a low last payment amount. This finding is logical, as having a low bankcard limit signifies limited credit availability for the borrower. Consequently, the borrower's capacity to make larger payments may be constrained, leading to a lower last payment amount.

3.2 Results of financial risk analysis on P2P lending

Performing risk analysis on P2P lending requires a focused investigation of loan statuses, specifically "charge off" and "fully paid," by fixing the consequence to one of these values. This approach enables the identification of relevant association rules that can effectively predict loan default. It is essential for the success and sustainability of the lending platform to understand the contributing factors to financial credit risk. A key variable in creditworthiness assessment is the FICO score, widely used as a measure of borrowers' creditworthiness. Analyzing the relationship between FICO scores and other variables can yield valuable insights into borrower behavior and risk patterns, where, by fixing the consequents to FICO scores, associations between antecedents and credit scores can be unveiled, leading to informed lending decisions for various types of borrowers.

The rules presented in rows 1–8 of Table 8 provide association rules with loan status categories "fully paid" and "charge off." These rules highlight specific combinations of borrower characteristics and loan attributes associated with each loan status category. For instance, the first rule suggests that borrowers with a low number of inquiries about financial accounts (inq_fi_low) and a high last FICO score (last_fico_score_high) are likely to have "fully paid" loan status with a high-confidence level of 0.941. This finding indicates that borrowers with good credit scores and a lower number of inquiries are more likely to fully repay their loans. This finding suggests that borrowers who maintain low debt levels, high credit scores, and limited credit card balances are more likely to fully repay their loans. Another noteworthy rule states that borrowers with a high last FICO score (last_fico_score_high), a low revolving balance (revol_bal_low), a low total received interest (total_rec_int_low), and a low maximum balance on a bankcard (max_bal_bc_low) are likely to have a loan status of "fully paid" with a high-confidence level of 0.946.

Table 8. Risk analysis association rules with a support value ≥0.3

Antecedents	Consequents	Support	Confidence	Lift
{'inq_fi_low'}, {'last_fico_score_high'}	{'loan_status_fully paid'}	0.42	0.941	1.89
{'total_rec_int_low'},{'last_fico_score_high'}, {'inq_last_12m_low'}	{'loan_status_fully paid'}	0.31	0.948	1.89
{'last_fico_score_high'},{'revol_bal_low'}, {'total_rec_int_low'}, {'max_bal_bc_low'}	{'loan_status_fully paid'}	0.31	0.946	1.88
{'last_fico_score_medium'}, {'total_rec_prncp_low'}	{'loan_status_charge off'}	0.36	0.940	1.88
{'mths_since_recent_bc_low'}, {'last_fico_score_medium'}, {'total_rec_prncp_low'}	{'loan_status_charge off'}	0.33	0.942	1.88
{'mths_since_recent_bc_low'}, {'last_fico_score_medium'},{'total_rec_prncp_low'}, {'last_pymnt_amnt_low'}	{'loan_status_charge off'}	0.33	0.943	1.89
{'mths_since_recent_bc_low'},{'total_rec_prncp_low'}, {'last_fico_score_medium'}, {'mo_sin_rcnt_rev_tl_op_low'}, {'last_pymnt_amnt_low'}	{'loan_status_charge off'}	0.31	0.943	1.88
{'mths_since_recent_bc_low'}, {'revol_bal_low'}, {'max_bal_bc_low'}, {'last_fico_score_medium'}, {'mo_sin_rcnt_rev_tl_op_low'}, {'last_pymnt_amnt_low'}	{'loan_status_charge off'}	0.31	0.912	1.82
{'pct_tl_nvr_dlq_high'}, {'loan_status_fully paid'}	{'last_fico_score_high'}	0.36	0.904	1.91
{'pct_tl_nvr_dlq_high'},{'inq_last_12m_low'}, {'loan_status_fully '}	{'last_fico_score_high'}	0.31	0.909	1.92

The rule presented in row 7 of Table 8 sheds light on an important association between borrower characteristics and loan status. According to this rule, borrowers who have a low number of months since their most recent bankcard account opened (mths_since_recent_bc_low), low total received principal (total_rec_prncp_low), a medium last FICO score (last_fico_score_medium), fewer months since their most recent revolving account opened (mo_sin_rcnt_rev_tl_op_low), and a low last payment amount (last_pymnt_amnt_low) are likely to have a loan status of charge off, with a high-confidence level of 0.943. In this finding, borrowers with a shorter credit history, low loan amounts, and low payment amounts appear to be at higher risk of defaulting on their loans, where it is suggested that certain combinations of borrower attributes are associated with a higher likelihood of loan default.

The association rules in rows 9 and 10 of Table 8 highlight potential risk patterns associated with loan defaults. Borrowers with specific characteristics, such as a moderate last FICO score (last_fico_score_medium), low total received principal (total_rec_prncp_low), low last payment amount (last_pymnt_amnt_low), and a recent open credit line (mths_since_recent_bc_low), may have a higher likelihood of defaulting on their loans.

Indeed, the association rules presented in Table 8 offer valuable insights for lenders in financial credit-risk analysis and decision-making within the context of P2P lending. These rules provide a data-driven understanding of the relationships between borrower characteristics and loan status, enabling lenders to make informed and prudent lending decisions. This information can be leveraged by lenders to identify creditworthy borrowers and offer them more favorable loan terms and conditions. The obtained association rules reveal that borrowers with favorable credit profiles, such as good credit scores, lower debt levels, and fewer inquiries, are more likely to repay their loans in full. On the other hand, the rules also highlight potential risk factors associated with loan default, such as moderate credit scores, lower loan amounts, and shorter credit histories. Borrowers exhibiting these characteristics may have a higher likelihood of defaulting on their loans. By recognizing these risk patterns, lenders can implement risk mitigation strategies to minimize the occurrence of charge-offs.

4. CONCLUSION

The association rule mining framework for financial credit-risk analysis has been proposed. The study utilized the LC dataset, which is P2P lending data, and employed various data preparation techniques, i.e., feature transformation, feature selection, data balancing, and data discretization. For association rule mining, the selected 41 features were used to construct insight rules for P2P lending. The acquired results showed that the proposed framework was effective in identifying relevant features and association rules that can aid in the assessment of financial credit risk in P2P lending. To analyze a financial credit risk, the obtained rules have consequents, e.g., "loan_status_fully paid" and "last_fico_score_high," that were considered and then association rules for financial credit risk were extracted. Furthermore, by identifying patterns and relationships among borrower characteristics, lenders can make more informed lending decisions and reduce the risk of default. This study contributes to the field of credit-risk analysis by providing a comprehensive framework that combines various techniques and methodologies for analyzing P2P lending data. The findings can be useful for lenders, investors, and policymakers who are interested in assessing credit risk and making informed decisions in the P2P lending market.

ACKNOWLEDGMENT

This work was supported by Thammasat Research Unit in Data Innovation and Artificial Intelligence.

REFERENCES

Acumen Research and Consulting. (2023). *P2P lending market size - global industry, share, analysis, trends and forecast 2022–2030*. [Online URL: https://www.acumenresearchandconsulting.com/table-of-content/p2p-lending-market] accessed on February 24, 2023.

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2), 207–216.

Bank of Thailand. (2022). *List of peer-to-peer lending platform providers tested under regulatory sandbox*. [Online URL: https://bit.ly/45ITySX] accessed on February 26, 2023.

Chai, N., Shi, B., Meng, B., and Dong, Y. (2023). Default feature selection in credit risk modeling: evidence from Chinese small enterprises. SAGE Open, 13(2), 1–15.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Chen, T. (2021). Credit default risk prediction of lenders with resampling methods. In *Proceedings of the 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pp. 123–127. Taiyuan, China.

Cheng, Y.-C., Chang, H.-T., Lin, C.-Y., and Chang, H.-Y. (2021). Predicting credit risk in peer-to-peer lending: A machine learning approach with few features. In *Proceedings of the 2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pp. 295–300. Taichung, Taiwan.

Desai, D. B., and Kaiwade, A. (2018). Application of Apriori algorithm for analyzing customer behavior to improve deposits in banks. In *Proceedings of the International Conference on Advances in Computer Technology and Management (ICACTM) in Association with Novateur Publications*, pp. 188–191. Maharashtra, India.

Guo, H., Peng, K., Xu, X., Tao, S., and Wu, Z. (2020). The prediction analysis of peer-to-peer lending platforms default risk based on comparative models. *Scientific Programming*, 2020, 8816419.

Hsueh, S.-C., and Kuo, C.-H. (2017). Effective matching for P2P lending by mining strong association rules. In *Proceedings of the 3rd International Conference on Industrial and Business Engineering (ICIBE '17),* pp. 30–33. New York, USA.

Ko, P.-C., Lin, P.-C., Do, H.-T., and Huang, Y.-F. (2022). P2P lending default prediction based on ai and statistical models. *Entropy*, 24(6), 801.

Kumar, M. R., and Gunjan, V. K. (2020). Review of machine learning models for credit scoring analysis. *Revista Ingeniería Solidaria*, 16(1), 1–16.

Lending Club 2007–2020Q3. (n.d.). *Kaggle*. [Online URL: https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1] accessed on May 20, 2022

Lee, K. K. G., Kasim, H., Zhou, W. J., Sirigina, R. P., and Hung, G. G. T. (2023). Feature redundancy assessment framework for subject matter experts. *Engineering Applications of Artificial Intelligence*, 117(Part A), 105456.

Mahdi, M. A., Hosny, K. M., and Elhenawy, I. (2022). FR-Tree: A novel rare association rule for big data problem. *Expert Systems with Applications*, 187, 115898.

Setiawan, N., Suharjito, and Diana. (2019). A comparison of prediction methods for credit default on peer to peer lending using machine learning. *Procedia Computer Science*, 157, 38–45.

Silva, J., Varela, N., Borrero López, L. A., and Rojas Millán, R. H. (2019). Association rules extraction for customer segmentation in the SMEs sector using the Apriori algorithm. *Procedia Computer Science*, 151, 1207–1212.

Suhada, C., and Devasia, J. V. (2022). Peer to peer lending: Risk prediction using machine learning on an imbalanced dataset. In *Proceedings of the 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, pp. 511–519. Kannur, India.

Tan, S. C. (2018). Improving association rule mining using clustering-based discretization of numerical data. In *Proceedings of the 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pp. 1–5. Mon Tresor, Mauritius.

Tanantong, T., and Ramjan, S. (2021). An association rule mining approach to discover demand and supply patterns based on Thai social media data. *International Journal of Knowledge and Systems Science*, 12(2), 1–16.

Vinod. K. L., Natarajan., S., Keerthana., S., Chinmayi., K. M., and Lakshmi, N. (2016). Credit risk analysis in peer-to-peer lending system. In *Proceedings of the IEEE International Conference on Knowledge Engineering and Applications (ICKEA)*, pp. 193–196. Nanyang Ave, Singapore.

Zhang, Y. (2021). Application of data mining technology in financial risk management. In *Proceedings of the IEEE Conference on Telecommunications and Computer Science (TOCS),* pp. 319–323, Shenyang, China.

Appendix A

Table A. Descriptions of data

Variables	Description
mths_since_last_pymnt_d	Number of months since last payment was received
total_rec_int	Interest received to date
grade	LC assigned loan grade
last_pymnt_amnt	Last total payment amount received
home_ownership	Home ownership status provided by the borrower during registration or obtained from the credit report; our values: RENT, OWN, MORTGAGE, OTHER
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
title	Loan title provided by borrower
total_rec_prncp	Principal received to date
revol_util	Revolving line utilization rate, or credit amount the borrower is using relative to all available revolving credit
mo_sin_old_il_acct	Months since oldest bank installment account opened
il_util	Ratio of total current balance to high credit/credit limit on all install acct
annual_inc	Self-reported annual income provided by borrower during registration
avg_cur_bal	Average current balance of all accounts
mths_since_earliest_cr_line	Number of months since borrower's earliest reported credit line was opened
total_bc_limit	Total bankcard high credit/credit limit
.	Average of lower and upper range of the borrower's last FICO score
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts
revol_bal	Total credit revolving balance
total_bal_il	Total current balance of all installment accounts
mths_since_last_credit_pull_d	Number of months since LC pulled credit for this loan
total_acc	Total number of credit lines currently in the borrower's credit file
tot_cur_bal	Total current balance of all accounts
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
dti	Ratio calculated using borrower's total monthly debt payments on total debt obligations, excluding mortgage and requested LC loan, divided by the borrower's self-reported monthly income
mths_since_recent_bc	Months since most recent bankcard account opened
purpose	Category provided by borrower for loan request
all_util	Balance to credit limit on all trades
mths_since_rcnt_il	Months since most recent installment accounts opened
max_bal_bc	Maximum current balance owed on all revolving accounts
loan_amnt	Listed amount of the loan applied for by borrower; If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
pct_tl_nvr_dlq	Percent of trades never delinquent
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
fico_score	Average of lower and upper range of borrower's FICO at loan origination belongs to
mo_sin_rcnt_tl	Months since most recent account opened
open_acc	Number of open credit lines in borrower's credit file
mths_since_recent_inq	Months since most recent inquiry
inq_fi	Number of personal finance inquiries
acc_open_past_24mths	Number of trades opened in past 24 months
inq_last_12m	Number of credit inquiries in past 12 months
emp_length	Employment length in years; possible values are between 0 and 10 where 0 means <1 year and 10 means ≥10 years
loan_status	Current status of the loan

Appendix B

Table B. Feature values after discretization on numerical data

Features	Original values	Number of groups	Results from data discretization
mths_since_last_pymnt_d	[0, 56]	2	[0, 22.92] = mths_since_last_pymnt_d_low [22.93 ,56] = mths_since_last_pymnt_d_high
total_rec_int	[0, 31714.37]	2	[0, 3,122.73] = total_rec_int_low [3,122.74, 31,714.37] = total_rec_int_high
last_pymnt_amnt	[0, 42192.05]	2	[0, 7,707.31] = last_pymnt_amnt_low [7,707.32, 42,192.05] = last_pymnt_amnt_high
total_rec_prncp	[0, 40000]	3	[0, 8,112.99] = total_rec_prncp_low [8,113, 21,537.74] = total_rec_prncp_medium [21,537.75, 40,000] = total_rec_prncp_high
revol_util	[0, 165.8]	3	[0, 37.12] = revol_util_low [37.13, 62.38] = revol_util_medium [62.39, 165.8] = revol_util_high
mo_sin_old_il_acct	[0, 999]	3	[0, 103.57] = mo_sin_old_il_acct_low [103.58, 178.75] = mo_sin_old_il_acct_medium [178.76, 999] = mo_sin_old_il_acct_high
il_util	[0, 464]	4	[0, 49.036] = il_util_low [49.037, 73.177] = il_util_medium [73.178, 93.354] = il_util_high [93.355, 464] = il_util_very high
annual_inc	[20, 10999200]	4	[20, 83824] = annual_inc_low [83825, 160799] = annual_inc_medium [160800, 436761] = annual_inc_high [436762, 10999200] = annual_inc_very high
avg_cur_bal	[12, 425387]	2	[12, 18010] = avg_cur_bal_low [18011, 425387] = avg_cur_bal_high
mths_since_earliest_cr_line	[42, 1049]	3	[42, 211.15] = mths_since_earliest_cr_line_low [211.16, 326.81] = mths_since_earliest_cr_line_medium [326.815, 1049] = mths_since_earliest_cr_line_high
total_bc_limit	[100, 1105500]	2	[100, 34141] = total_bc_limit_low [34140, 1105500] = total_bc_limit_high
last_fico_score	[0, 639.18]	3	[0, 451.65] = last_fico_score_low [451.66, 639.18] = last_fico_score_medium [639.19, 847.5] = last_fico_score_high
bc_util	[0, 252.3]	3	[0, 40.15] = bc_util_low [40.16, 72.84] = bc_util_medium [72.85, 252.3] = bc_util_high
revol_bal	[0, 1696796]	2	[0, 39528.28] = revol_bal_low [39530, 1696796] = revol_bal_high
total_bal_il	[1, 1711009]	2	[1, 70414] = total_bal_il_low [70415, 1711009] = total_bal_il_high
mths_since_last_credit_pull_d	[0, 56]	4	[0, 7.932] = mths_since_last_credit_pull_d_low [7.933, 16.121] = mths_since_last_credit_pull_d_medium [16.122, 27.174] = mths_since_last_credit_pull_d_high [27.175, 56] = mths_since_last_credit_pull_d_very high
total_acc	[3, 176]	2	[3, 26.9] = total_acc_low [27, 176] = total_acc_high
tot_cur_bal	[52, 5445012]	4	[52, 94272] = tot_cur_bal_low [94273, 229132] = tot_cur_bal_medium [229131, 423093] = tot_cur_bal_high [423094, 5445012] = tot_cur_bal_very high
mo_sin_old_rev_tl_op	[1, 902]	3	[1, 131.86] = mo_sin_old_rev_tl_op_low [131.87, 230.70] = mo_sin_old_rev_tl_op_medium [230.71, 902] = mo_sin_old_rev_tl_op_high
dti	[0, 999]	5	[0, 14.54] = dti_very low [14.55, 21.60] = dti_low [21.61, 29.79] = dti_medium [29.80, 54.81] = dti_high [54.82, 999] = dti_very high

Table B. Feature values after discretization on numerical data (Continued)

Features	Original values	Number of groups	Results from data discretization
mths_since_recent_bc	[0, 569]	2	[0, 46.56] = mths_since_recent_bc_low [46.57, 569] = mths_since_recent_bc_high
all_util	[0, 204]	4	[0, 42.48] = all_util_low [42.49, 61.24] = all_util_medium [61.25, 77.37] = all_util_high [77.38, 204] = all_util_very high
mths_since_rcnt_il	[0,426]	7	[0, 8.13 = mths_since_rcnt_il_extreamly low [8.14, 14.62] = mths_since_rcnt_il_very low [14.63, 22.28] = mths_since_rcnt_il_low [22.29, 32.88] = mths_since_rcnt_il_medium [32.89, 51.89] = mths_since_rcnt_il_high [51.90, 87.25] = mths_since_rcnt_il_very high [87.26, 426] = mths_since_rcnt_il_extreamly high
max_bal_bc	[0, 776843]	2	[0, 9555] = max_bal_bc_low [9556, 776843] = max_bal_bc_high
loan_amnt	[1000, 40000]	3	[1000, 12651] = loan_amnt_low [12652, 23352] = loan_amnt_medium [23353, 40000] = loan_amnt_high
pct_tl_nvr_dlq	[5, 100]	2	[5, 89.69] = pct_tl_nvr_dlq_low [89.70, 100] = pct_tl_nvr_dlq_high
mo_sin_rcnt_rev_tl_op	[0, 404]	2	[0, 24.39] = mo_sin_rcnt_rev_tl_op_low [24.40, 404] = mo_sin_rcnt_rev_tl_op_high
fico_score	[662, 847.5]	2	[662, 705.95] = fico_score_low [705.96, 847.5] = fico_score_high
mo_sin_rcnt_tl	[0, 27.60]	7	[0, 3.44] = mo_sin_rcnt_tl_extreamly low [3.44, 6.34] = mo_sin_rcnt_tl_very low [6.35, 9.36] = mo_sin_rcnt_tl_low [9.37, 12.89] = mo_sin_rcnt_tl_medium 12.90, 18.139] = mo_sin_rcnt_tl_high [18.14, 27.60] = mo_sin_rcnt_tl_very high [27.61, 147] = mo_sin_rcnt_tl_extreamly high
open_acc	[2, 88]	3	[2, 10.8] = open_acc_low [10.9, 18.91] = open_acc_medium [18.92, 88] = open_acc_high
mths_since_recent_inq	[0, 24]	2	[0, 7.69] = mths_since_recent_inq_low [7.70, 24] = mths_since_recent_inq_high
inq_fi	[0, 27.174]	3	[0, 7.93] = inq_fi_low [7.934, 16.12] = inq_fi_medium [16.13, 27.174] = inq_fi_high [27.175, 56] = inq_fi_very high
acc_open_past_24mths	[0, 61]	2	[0, 5.79] = acc_open_past_24mths_low [5.78, 61] = acc_open_past_24mths_high
inq_last_12m	[0, 67]	2	[0, 4.22] = inq_last_12m_low [4.23, 67] = inq_last_12m_high