Electronic Commerce Research and Applications 62 (2023) 101332

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

journal homepage: www.elsevier.com/locate/elerap

Forecasting credit default risk with graph attention networks

Binbin Zhou [a,1], Jiayun Jin [a,1], Hang Zhou [a,b], Xuye Zhou [c], Longxiang Shi [a], Jianhua Ma [d], Zengwei Zheng [a,*]

[a] School of Computer and Computing Science, Hangzhou City University, No. 48 Huzhou Street, Hangzhou, 310015, Zhejiang, China
[b] College of Computer Science and Technology, Zhejiang University, 38 Zheda Road, Hangzhou, 310027, Zhejiang, China
[c] School of Digital Commerce and Trade, Zhejiang Institute of Mechanical & Electrical Engineering, 528 Binwen Road, Hangzhou, 310059, Zhejiang, China
[d] Faculty of Computer and Information Sciences, Hosei University, Chiyoda-ku, 1028160, Japan

ARTICLE INFO

Keywords:
Credit default risk
Graph neural network
Deep learning
Graph attention network
Multi-source data

ABSTRACT

The importance of credit default risk management has risen that companies can utilize it to identify and forecast future credit default risk. Several approaches have been proposed, however, they paid little attention on the various underlying relationships between users, which can provide significant improvement. In this paper, we propose a Graph Attention Network (GAT)-based model for predicting credit default risk, leveraging various types of data, including credit default history, credit status and personal profile. These data provide a comprehensive representation of users' overall status, including historical financial credit, recent financial credit and wealth status. Different graphs are constructed based on the similarities between users using these data, respectively. Then, for graphs, GAT modules are used to capture both the relationships with adjacent and high-order neighbors, as well as the linear and non-linear relationships. After fusing learned high-level features from GAT modules, final predictive results, whether users will default or not, are predicted. The effectiveness of our prediction model is validated using real-world datasets, and experimental results depict that our model can accurately predict credit default risks, outperforming several baseline methods. The codes and datasets are freely available at https://github.com/ZJUDataIntelligence/Foreknow.

1. Introduction

Electronic commerce has developed so rapidly in recent decades as a result of widespread Internet use that the volume of transaction conducted electronically has increased dramatically. There have been examples of how trade is performed in this fashion, such as online transaction payment system, supply chain management, and so forth. In electronic commerce, purchasers select things and pay for them using an online payment mechanism that is typically non-cash. The majority of buyers choose to pay using a credit card. Due to the convenience, transaction efficiency, mobility and little financial risk, it has become the most popular way of payment in recent years (Laudon and Traver, 2013). According to the data from People's Bank of China (People's Bank of China, 2021b), during the Double 11 period in 2021, which is an annual biggest shopping festival in China, similar to Black Friday in the U.S., financial services corporations (e.g. UnionPay) have completed 27.048 billion payment transactions totaling 22.32 trillion RMB, an increase of 17.96% and 14.9% year-on-year, respectively. Because of the convenience of electronic commerce trade, more consumers are opting to buy things online with their credit cards, promoting the widespread use of credit cards.

With the prevalence of credit card usage, credit default has been a rising concern that has gotten increasing attention in recent years. Consumers have a tendency to overbuy products and spend more money than they can afford. According to statistics, at least half of the total transaction amount of the Double 11 was completed via overdraft consumption (China Marketing, Research & Digital China Skinny, 2021). Due to their excessive spending, a large number of consumers are at risk of credit default, which implies they will not be able to repay the money on time (China Banking and Insurance Regulatory Commission, 2020). Along with the enormous increase in individuals using credit cards for overdraft purchases, credit card defaults are on the rise. At the end of 2020, the total number of credit cards that were six months past due was 83.86 billion RMB, and it reached 86 billion RMB by the end of 2021 (People's Bank of China, 2021a, 2022). In 2021, four large state-owned banks in China, i.e., ICBC, CCB, Agricultural Bank, and Postal Savings Bank of China, have credit card delinquency rates of

* Corresponding author.
E-mail addresses: bbzhou@hzcu.edu.cn (B. Zhou), jin.jyun@foxmail.com (J. Jin), hangz@zju.edu.cn (H. Zhou), zhouxuye@zime.zj.cn (X. Zhou), shilx@hzcu.edu.cn (L. Shi), jianhua@hosei.ac.jp (J. Ma), zhengzw@hzcu.edu.cn (Z. Zheng).
[1] These authors have contributed equally to this work.

https://doi.org/10.1016/j.elerap.2023.101332
Received 30 March 2022; Received in revised form 10 August 2023; Accepted 12 October 2023
Available online 13 November 2023

1.90%, 1.33%, 0.99%, and 1.66%, respectively (China Times, 2022). From a psychological perspective, some previous studies found that some characteristics, e.g., shopping contextual surroundings and high recommendations by online reviews, can help consumers feel pleasure, which would drive consumers in a positive state to more likely engage in impulse buying behavior (Sharma and Sivakumaran, 2015). Hence, it is necessary to conduct credit default risk-related studies in the field of e-commerce.

It is important to sense credit default risk in advance. Cardholders with credit default records may have a lower credit score, wiping out the primary benefit of a credit card and eventually jeopardizing their financial health. Financial institutions that provide credit card services may suffer serious economic losses (Ogiela, 2015; Hu et al., 2015; Luo et al., 2016; Wang et al., 2018a). Therefore, credit default risk management has grown in importance. Some companies can utilize it to identify and predict different-level potential credit default risk. For cardholders with high credit default risk, proper credit default risk management can save companies a lot of money when it comes to defaulters have overspent their credit cards. For cardholders with low credit default risk, financial institutions may not regard defaulters seriously and provide them the opportunity to keep their financial health if the due date is missed and they have the abilities to repay the money on time.

There have been numerous studies focused on credit default risk and various methods have been proposed to solve the problem. The existing methods can be grouped into two categories: statistic-based methods and machine learning-based methods. Statistic-based methods, such as Poisson regression model and forward intensity model, have been employed to predict future default risk by early studies (Chen and Yu, 2014; Guo et al., 2020; Agosto et al., 2019; Lee and Sohn, 2021; Duan et al., 2012; Sohn and Kim, 2012). In the past decades, machine learning-based methods have attracted lots of attentions in the study of credit default risk prediction, and have become a promising paradigm. Popular machine learning-based methods applied in this research fields include random forest, support vector machine (SVM), gradient boosting, long short-term memory (LSTM) and so forth (Zhu et al., 2019; Cowden et al., 2019; Shen et al., 2021; Liang and Cai, 2020; Guo et al., 2020). However, they paid little attention on the relationship construction between cardholders, which we believe can provide valuable improvement in the credit default risk prediction. The rational is two fold. First, predicting credit default risk depends on a comprehensive understanding of the intricate relationships between individuals. Conventional approaches often overlook the underlying relationships between users, focusing predominantly on their individual attributes and historical data. By investigating the relationship construction between cardholders, our study underscores the significance of capturing these relationships for more accurate credit default risk prediction. Second, the relationships between cardholders can provide valuable insights into their financial behaviors and potential risks. Individuals would influence others' credit default probability through financial transactions, shared characteristics, or similar financial habits. By considering the relationship construction between cardholders, this investigation aims to uncover these interdependencies and incorporate them into the credit default risk prediction model.

In this paper, we attempt to predict credit default risk taking the multiple relationship between cardholders into account. We introduce multiple kinds of data, including the recent loan records, loan history and their personal profile. These data can comprehensively represent the user's financial status, including the recent financial credit, the historical financial credit, and wealth status. In this study, we propose a Graph Attention Network (Veličković et al., 2017) (GAT)-based model for credit default risk prediction. Leveraging these types of data, different graphs are constructed based on the similarities between users. We apply GAT modules to capture relationship with other cardholders, both with adjacent and high-order neighbors, as well as the linear and non-linear relationships. Fully connected layers are adopted to fuse the learned high-level features from these GAT modules for final credit default risk prediction. The effectiveness of our proposed GAT-based prediction model are verified by using a real-world dataset, and the experimental results demonstrate that the proposed model can accurately predict the credit default risk, outperforming several baseline methods.

The remainder of this paper is organized as follows. Section 2 examines existing literature on relevant studies. Section 3 introduces the proposed method. The experiment results and analyses are depicted in Section 4. Finally, Section 5 gives some discussions and concludes this paper.

2. Literature review

In the past decades, there have been many studies on credit default risk prediction. We review the previous studies as follows.

Some well-acknowledged statistical models has been adopted for financial risk prediction tasks (Chen and Yu, 2014; Guo et al., 2020; Agosto et al., 2019; Lee and Sohn, 2021). A mixed Poisson model was used to develop the dependence structure of obligors based on the assumption that the probability of default is determined by common economic factors, with further empirical studies through an application to four Chinese industrial portfolios, which demonstrates the significance of mixed Poisson model in measuring credit portfolio risk (Chen and Yu, 2014). A binary spatial regression model was used to quantify the effects of business failure on contagion, and evidence of significant levels of contagion risk was presented, increasing the individual's credit risk (Agosto et al., 2019).

Recently, many studies have focused on improving the prediction performance by utilizing machine learning models, such as random forest (Zhu et al., 2019; Guo et al., 2020), fuzzy logic (Ashraf et al., 2020b), support vector machine (SVM) (Cowden et al., 2019), ensemble learning methods (Xia et al., 2017; Addo et al., 2018; Hamori et al., 2018; Ma et al., 2018; Shen et al., 2021) and deep learning-based methods (Kvamme et al., 2018; Liang and Cai, 2020). An event2vec method has been used to represent user online behaviors, such as click records and browsing records. After that, a bidirectional LSTM-based network with an attention mechanism has been designed with these event features to predict the default risk of borrowers (Wang et al., 2018b). A convolutional neural network-based model has been adopted for mortgage defaults forecasting of each consumer leveraging annual consumer transaction data. The experiments results demonstrates the effectiveness of the proposed methods (Kvamme et al., 2018). A novel deep learning-based model incorporating a LSTM-based network and an adaptive boosting method, has been developed to evaluate personal credit risk, while addressing imbalanced credit data using an enhanced synthetic minority oversampling method (Shen et al., 2021). Leveraging datasets from a peer-to-peer lending website Lending Club in the United States between 2008 and 2015, a novel LSTM-based model has been proposed to analyze the default risk of monthly fresh loans, with superior performance due to its unparalleled capacity to extract time-series information (Liang and Cai, 2020). These previous studies mainly focus on the personal credit risk domain, incorporating personal wealth profile-related data into designated models for credit risk prediction, e.g., income, historical loan records, wealth status, and so forth.

There have also been a number of studies utilizing machine learning methods to predict commercial credit risk (Luo et al., 2017; Lappas and Yannacopoulos, 2021; Wang and Song, 2022; Yao et al., 2022; Yang et al., 2022). Deep belief networks with restricted Boltzmann machines are applied in the corporate credit scoring domain to investigate the prediction capabilities of credit scoring (Luo et al., 2017). A decision tree based model combined with synthetic minority oversampling technologies has been developed to predict enterprise credit risk by fusing relevant supply chain data, and experimental results have confirmed the value of supply chain data (Yao et al., 2022). A deep neural network-based algorithm has been proposed for the high-dimensional

Table 1
Literature table concerning credit default risk prediction.

Research	Uniqueness	Major characteristics	Advantages	Disadvantages
Wang et al. (2018b)	Uses deep learning techniques to predict credit scoring	Builds an Event2vec model and adopts an attention mechanism to extract information	Flexible and scalable	Without considering the time interval between events
Kvamme et al. (2018)	Uses CNNs on financial transaction data	Takes in daily transaction data from multiple accounts and uses two CNNs to extract features and make predictions	Handles class imbalance and overfitting	Computationally expensive and not easy to interpret
Shen et al. (2021)	Combines multiple deep learning methods and SMOTE technique on imbalanced datasets	Uses multiple base learners to improve classification accuracy, and handles imbalanced datasets	Handles high-dimensional and complex datasets	Computationally expensive
Liang and Cai (2020)	Uses LSTM and GRU models	Captures long-term dependencies and uses memory cells	Strong learning and generalization abilities	Uses a single dataset and lack of consideration for external factors
Luo et al. (2017)	Study of DBN-based model in corporate credit rating	Applies DBN with Restricted Boltzmann Machines to handle high-dimensional data	Automatically learns features without manual feature engineering	Requires a large amount of data
Yang et al. (2022)	Uses neural networks with L1 and L2 regularization for feature selection	Uses deep neural networks to mine features, incorporates supply chain information	Effectively reduces data's dimensionality	Difficult to interpret
Yao et al. (2022)	Integrates supply chain information into credit risk prediction	Uses a decision tree ensemble model to deal with class imbalance	Handles class imbalance and interpretability	Not be suitable for all types of enterprises
Lee et al. (2021)	Uses a graph convolutional network to capture the relationship between borrowers	Captures high-order relationships between borrowers and incorporates soft information to improve credit default prediction	Provides augmented features	Without considering multiple relationships between borrowers

prediction of corporate credit risk with supply chain and network data (Yang et al., 2022). In contrast with studies on the personal credit risk domain, these studies paid lots of attention to the characteristics of companies, such as asset–liability ratio, profit ratio, and solvency, which are frequently issued quarterly or annual (Cisi et al., 2020).

Due to the powerful ability to represent complex relationships, graph structures have obtained increasing attention (Ashraf et al., 2020a). Graph neural networks that handle numerous tasks in graph-structured data have received a lot of attention (Liu et al., 2022). In graph structures, there are three levels of tasks: node-level, edge-level, and graph-level. By developing a word-document network utilizing word co-occurrence with word document frequency data, a text-GCN has been developed to predict the classification of a document (Yao et al., 2019). The edges, also known as graph links, have been extensively studied in order to determine if two nodes in a network would be connected (Zhang and Chen, 2018). Graph-level prediction gathers a complete graph characteristic with node and edge information, and then predicts the targeted values for the input graph, such as predicting the molecule and drug types utilizing chemical graph structures. For the credit default risk prediction task, a graph convolutional network has been employed that considers the nonlinear relationships between borrower attributes and high-order relationships between borrowers (Lee et al., 2021). However, previous studies do not consider the multiple relationships modeling which can be an effective augmentation for the prediction models. The comparison of previous studies are summarized in Table 1.

While the existing studies have contributed valuable insights, there remains an untapped potential to improve prediction performance by leveraging the full use of user data. Specifically, the multiple relationships of users are under exploration and remain a promising avenue for improvement, which can reflect the user's financial status from diverse yet complementary perspectives. In our study, we aim to comprehensively model the multiple relationships of users and construct graphs accordingly. Subsequently, we propose a novel Graph Attention Network-based model, proficient in the aggregation of multi-type information from neighboring and higher-order nodes, encompassing both linear and non-linear relationships, thereby improving the prediction performance holistically and automatically.

3. Methodology

In this paper, we propose a credit default risk prediction model based on Graph Attention Networks (GAT), taking relationship between users into consideration. The architecture of our proposed model is presented in Fig. 1. In the model, multiple types of data are leveraged for users relationship construction, including the latest detailed loan data, loan history data, and cardholders' personal data, which can represent the comprehensive credit status of cardholders, such as the recent credit, historical credit and trusted wealth status. Based on these data, we construct graphs representing their relationship, in which cardholders are defined as nodes in graphs, and similarities are extracted as the connection between them. Three graphs are constructed based on three types of similarities. Following that, we utilize GAT modules with the purpose of extracting both adjacent neighbors and high-order neighbors' information, as well as linear and non-linear relationships. A fully connected layer is then employed to fuse the extracted features from GAT modules, and output the final credit default risk prediction results to depict which nodes having credit default risk.

3.1. Graph construction

In this model, we aim to extract the various types of similarities of cardholders with the utilization of different types of data, including recent loan data, history loan data, and personal wealth-related data of cardholders, and then construct the corresponding graphs respectively.

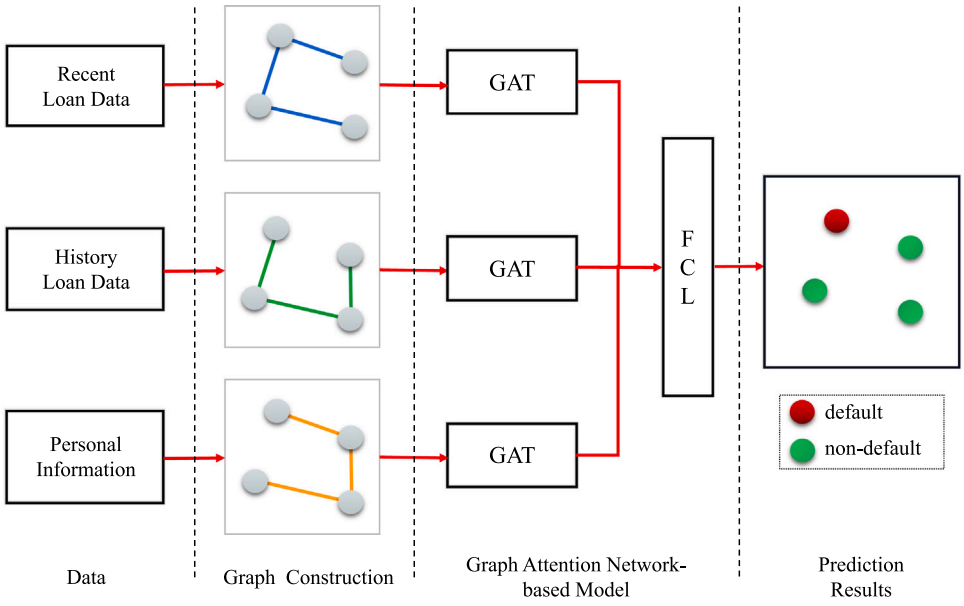


Fig. 1. Overview of our proposed model.

The recent loan data refers to latest detailed loan information of cardholder, such as the balance, contract status and loan purpose. Note that one cardholder may have multiple loan records. The history loan data means previous applications information, including status of reported credit, recorded currency, maximal amount overdue so far, current debt and so forth. The cardholders' personal data denotes some information relevant to represent the cardholders' trusted financial state, such as annual income, whether hold a house, whether hold a car, family status and so on. The detailed information of these data are presented in Section 4.1.

Leveraging these data, we compute the similarities between cardholders, to represent the edges connected nodes. Similarity between numerical data can be easily computed by the Euclidean distance. But the similarity between categorical data is not easy to compute, and even more difficult when dealing with mixed numerical and categorical data. Here, we adopt the Ahmad & Dey method (Ahmad and Dey, 2007) for similarity computation due to its capability in coping with mixed types of data.

The Ahmad & Dey method is a well-acknowledged method for mixed variable data to measure the difference between two data. The computed results would be a number between 0 and 1. A lower value depicts a more similarity between two data. In particular, the Ahmad & Dey distance with two data x_i and x_j is defined as $D_A(x_i, x_j)$. The computation of $D_A(x_i, x_j)$ is as follows:

$$D_A(x_i, x_j) = \sum_{k=1}^{l} s_k^A(x_i, x_j) \tag{1}$$

When they are categorical variables, the distance using the Ahmad & Dey method $s_k^A(x_i, x_j)$ between two variables x_i and x_j is calculated as follows.

$$s_k^A(x_i, x_j) = \frac{1}{l-1} \sum_{k=1, k\neq k'}^{l} \varphi^{kk'}(x_i, x_j) \tag{2}$$

$$\varphi^{kk'}(x_i, x_j) = p_k(\eta|x_i) + p_k(\overline{\eta}|x_j) - 1 \tag{3}$$

$$\eta = \arg\max_m \{p_k(m|x_i) + p_k(\overline{m}|x_j)\} \tag{4}$$

Here, $p_k(\eta/x_i)$ denotes the conditional probability that the jth attribute belongs to η when the ith attribute with data of x_i. $p_k(\overline{\eta}|x_j)$ denotes the conditional probability that the jth attribute belongs to $\overline{\eta}$ when the ith attribute with data of x_j. $s_k^A(x_i, x_j)$ satisfies the following requirements:

$$\begin{aligned} &0 \leq s_k^A(x_i, x_j) \leq 1 \\ &s_k^A(x_i, x_j) = s_k^A(x_j, x_i) \\ &s_k^A(x_i, x_i) = 0 \end{aligned} \tag{5}$$

When they are numerical variables, they need to be discretized into multiple intervals, as $u[1], u[2], \ldots, u[T]$. After the discretization, the distance of each pair of $u[i]$ and $u[j]$ can be calculated as shown in Eq. (2). For each numerical feature, the computation is as follows:

$$s_k^A(x_i) = \frac{2}{T(T-1)} \sum_{k=1}^{T} \sum_{j\geq k}^{T} s_k^A(u[k], u[j]) \tag{6}$$

Based on the similarity computation results, we then define a threshold value δ_A to determine whether to connect pairs of nodes with edges. In this way, the graphs can be constructed.

3.2. Graph attention network

Based on these constructed graphs, in this section, we present a Graph Attention Network (Veličković et al., 2017) based model. The basic structure has been shown in Fig. 2. The Graph Attention Network is able to extract similar users' features in different orders and fuse both linear and non-linear relationships, for the final credit default risk prediction.

Graph is a typical unstructured data and consists of a set of nodes, with a node feature matrix in most cases, and a set of edges which is usually represented by a adjacency matrix. Graph neural network (GNN) is a powerful machine learning method applied on graphs. Given a graph and the corresponding node feature matrix, GNN can learn a latent and expressive node feature representation which can be used to facilitate downstream tasks such as node classification and link prediction.

The graph attention network (GAT) is a spatial-based GNN method. Given the fixed adjacency matrix of the graph, GAT uses the node feature matrix as inputs, and outputs an updated node feature matrix. $(h_1, h_2, h_3, \ldots)$ in Fig. 2 represents the input node feature matrix, and h_i is one column of the node feature matrix which means the feature vector of node i.

A simple and effective message passing mechanism is used to aggregate the feature of neighbors to each node itself. With repeats of

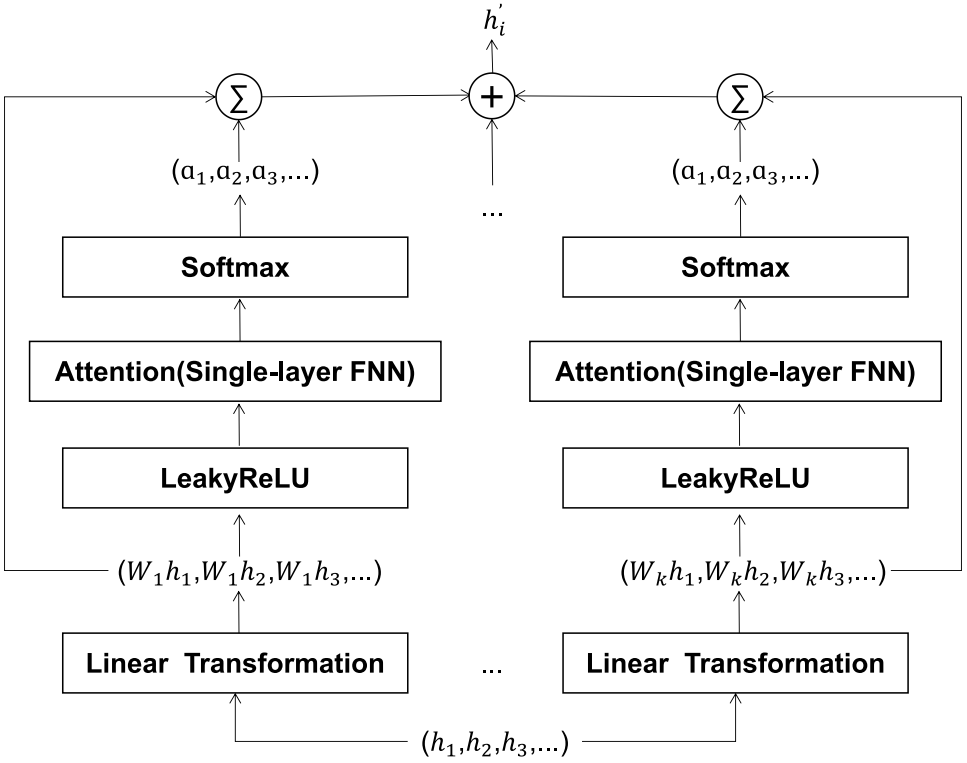


Fig. 2. A graph attention network.

this message passing process, both local and non-local features are contained in the final representations. Besides, GAT introduces an attention mechanism in the node feature aggregation process, which treats all neighbors unequally (i.e.,fully considers the difference between the contribution and importance of different neighbors). With the multi-head attention on the final layer of the network (i.e.,with multiple linear transformations), the final node representation is the sum or average of node representations learned by different attention heads, which improve the robustness and expressiveness of the model.

The detailed procedure and the corresponding mathematical formulas of GAT are described as following:

(1) First, a shared linear transformation with a learnable weight matrix W is applied to every node feature vector $\vec{h}_i$, then a shared attention mechanism a is used to compute the attention coefficient e_{ij}. The attention mechanism can be any function that can reflects the correlation of two objects, such as the cosine similarity function or MLP. Specifically, a single layer feed-forward neural network is chose as the attention mechanism in GAT.

$$e_{ij} = a\left(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j\right) \tag{7}$$

where $\vec{h}_i$ and $\vec{h}_j$ is the node feature vector, and $\mathbf{W}$ is a learnable parameter matrix. a denotes the attention mechanism, which is implemented as single layer feed-forward neural network in GAT.

(2) Then, a softmax function is applied on those attention coefficients to get the attention weights a_{ij}, which represent the importance of node j to node i.

$$\alpha_{ij} = \mathrm{softmax}_j\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(e_{ik}\right)} \tag{8}$$

where $\mathcal{N}_i$ is the set of neighbor nodes of node i and exp denotes the exponential function.

(3) Finally, for each node i, GAT update its feature vector by sum up its neighbors' feature vectors to itself with those attention weights a_{ij}. To aggregate the features of multi-head (i.e.,multiple linear transformations in step 1), multiple feature vectors of node i is averaged to generate the final representation of node i.

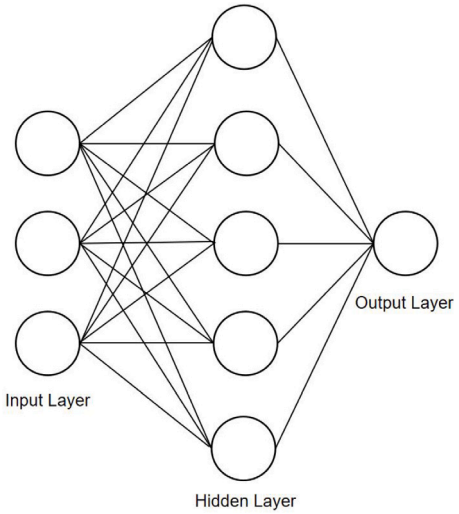


Fig. 3. Fully connect layers (FCL).

$$\vec{h}_i' = \|_{k=1}^{K}\ \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k\mathbf{W}^k\vec{h}_j\right) \tag{9}$$

where $\vec{h}_j$ the feature vector of node j. $\vec{h}_i'$ is the updated feature vector of node i. $\|$ represents the concatenation operation.

3.3. Credit default risk prediction

After getting the final feature representations of each node, we use fully connected layers to obtain the final prediction of each user. The

Table 2
Selected attributes list of recent loan data.

	Attributes	Definition
Key	SK_ID_CURR	ID of loan in our sample
	AMT_BALANCE	Balance during the month of previous credit
	AMT_CREDIT_LIMIT_ACTUAL	Credit card limit during the month of the previous credit
	AMT_DRAWINGS_ATM_CURRENT	Amount drawing at ATM during the month of the previous credit
	AMT_DRAWINGS_CURRENT	Amount drawing during the month of the previous credit
	AMT_DRAWINGS_OTHER_CURRENT	Amount of other drawings during the month of the previous credit
	AMT_DRAWINGS_POS_CURRENT	Amount drawing/buying goods during the month of the previous credit
	AMT_INST_MIN_REGULARITY	Minimal installment for this month of the previous credit
Expert	AMT_PAYMENT_CURRENT	Amount the client pay during the month on the previous credit
	AMT_PAYMENT_TOTAL_CURRENT	Amount the client pay during the month in total on the previous credit
knowledge	AMT_RECEIVABLE_PRINCIPAL	Amount receivable for principal on the previous credit
	AMT_RECIVABLE	Amount receivable on the previous credit
	AMT_TOTAL_RECEIVABLE	Total amount receivable on the previous credit
	CNT_DRAWINGS_ATM_CURRENT	Number of drawings at ATM during this month on the previous credit
	CNT_DRAWINGS_CURRENT	Number of drawings during this month on the previous credit
	CNT_DRAWINGS_OTHER_CURRENT	Number of other drawings during this month on the previous credit
	CNT_DRAWINGS_POS_CURRENT	Number of drawings for goods during this month on the previous credit
	CNT_INSTALMENT_MATURE_CUM	Number of paid installments on the previous credit
	NAME_CONTRACT_STATUS	Contract status on the previous credit
	SK_DPD	Days past due
	2 Attributes are excluded	

Table 3
Selected attributes list of history loan data.

	Attributes	Definition
Key	SK_ID_CURR	ID of loan in our sample
	CREDIT_CURRENCY	Recorded currency of the Credit Bureau credit
Liu et al. (2022)	DAYS_CREDIT	Days before current application did client apply for Credit Bureau credit
	AMT_CREDIT_SUM_LIMIT	Current credit limit of credit card reported in Credit Bureau
	AMT_CREDIT_SUM	Current credit amount for the Credit Bureau credit
Zhang et al. (2020)	AMT_CREDIT_SUM_DEBT	Current debt on Credit Bureau credit
	CREDIT_TYPE	Type of Credit Bureau credit (Car, cash, ...)
	CREDIT_ACTIVE	Status of the Credit Bureau (CB) reported credits
Expert	AMT_CREDIT_MAX_OVERDUE	Maximal amount overdue on the Credit Bureau credit so far
knowledge	AMT_CREDIT_SUM_OVERDUE	Current amount overdue on Credit Bureau credit
	AMT_ANNUITY	Annuity of the Credit Bureau credit
	6 Attributes are excluded	

structure of fully connect layers is shown in Fig. 3. Specifically, as we have constructed three graphs representing the three types of user relationships, we now get three feature vectors for each cardholder. To fuse these features and get the final prediction results, we first concatenate three feature vectors of each user and then feed it to the fully connect layer to get the value of risk probability prediction. The detailed operation is formulated as below:

$$\mathbf{H} = \|(\mathbf{H}_{recent}, \mathbf{H}_{hist}, \mathbf{H}_{info}) \tag{10}$$

$$\vec{p} = \mathbf{HW} \tag{11}$$

Where $H_{recent}, H_{hist}, H_{info}$ are the outputs of three GAT modules with the corresponding three graphs. Each of them has the shape of (N, D_O). D is the number of nodes (users) and D_O is the output feature dimension of the GAT module, which is a hyperparameter. After concatenation on the second dimension, we get H with the shape of $(N, 3D_O)$.

Then we use the efficient implementation of FCL with matrix multiplication, where W is a learnable parameter matrix with shape $(3D_O, 1)$. Finally, we obtain the final prediction vector $\vec{p}$ of shape $(N, 1)$, which represents the probability prediction of credit default risk for each cardholder. The ith element in $\vec{p}$ corresponds to the default probability of the ith user predicted by the model, which can be used to evaluate the credit level.

4. Experiments

4.1. Data acquisition and preprocessing

The datasets utilized in this study are from an international consumer finance provider, *Home Credit*, which was founded in 1997. This company focuses on offering various loan products to a large amount of users. In this paper, we use the credit default-related datasets from this company to evaluate the performance of our proposed method to predict the credit default risk. Specifically, we select recent loan data, history loan data and personal profile data of cardholders here. These datasets contain a total of 3,014,276 records, with 152,643 default records, accounting for almost 5% credit default ratio. This is a class-imbalance data. Hence, we connect information from multiple datasets according to users, and design features from multiple records belonging to each specific user. After the preprocessing, the datasets include 9841 cardholders, of which 2813 cardholders are labeled as with credit default risk. The class-imbalance has been relieved. We select some relevant features according to previous studies (Zhang et al., 2020; Xia et al., 2021; Liu et al., 2022) and expert knowledge, and present them in Table 2, 3 and 4. In the datasets with a total of 159 attributes, some attributes are irrelevant for predicting credit default risk. Thus, careful attribute selection is required. After a thorough literature review and analysis, we selected 44 relevant attributes and shown the rationale for why we select this attribute for further credit default risk prediction in these tables. We can also see that 2, 6 and 106 irrelevant attributes have been excluded, respectively.

Table 4
Selected attributes list of Cardholders' personal data.

	Attributes	Definition
Key	SK_ID_CURR	ID of loan in our sample
Zhang et al. (2020)	FLAG_OWN_CAR	Flag if the client owns a car
	AMT_ANNUITY	Loan annuity
	NAME_EDUCATION_TYPE	Level of highest education the client achieved
	NAME_FAMILY_STATUS	Family status of the client
	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
Xia et al. (2021)	CODE_GENDER	Gender of the client
Liu et al. (2022)	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
	FLAG_OWN_REALTY	Flag if client owns a house or flat
	AMT_INCOME_TOTAL	Income of the client
	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
	CNT_CHILDREN	Number of children the client has
Expert knowledge	AMT_CREDIT	Credit amount of the loan
	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave, ...)
	CNT_FAM_MEMBERS	How many family members does client have
	106 Attributes are excluded	

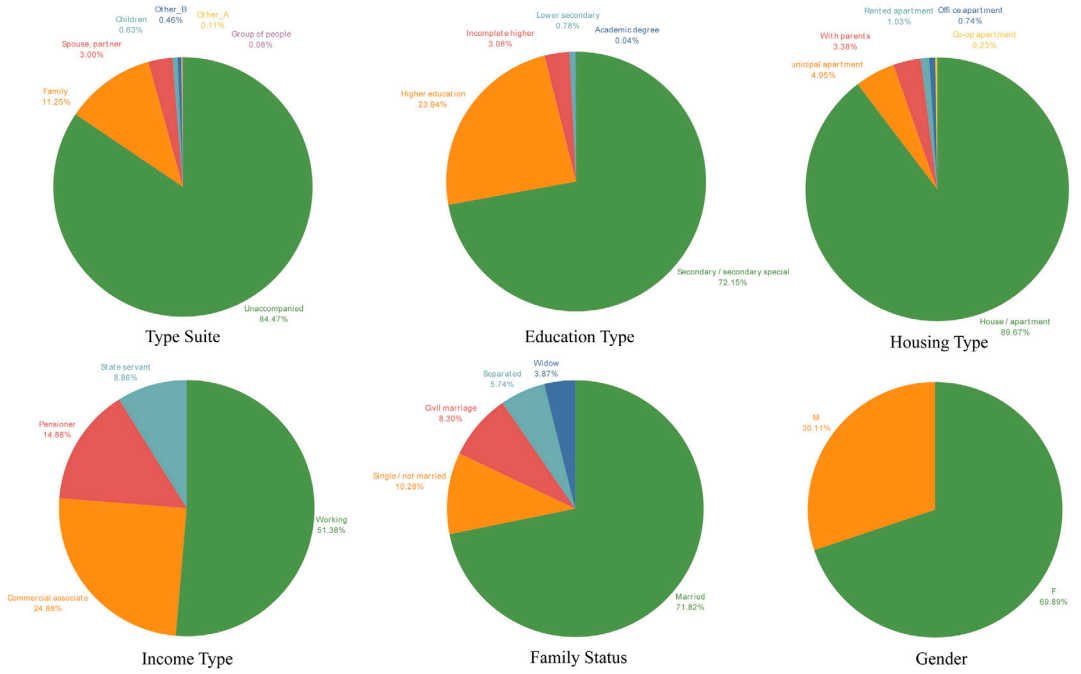


Fig. 4. Distribution of some attributes.

Notice that some cardholders have multiple records, such as loan records. Hence, we need to preprocess these data. Specifically, for several attributes as *CNT_CHILDREN*, *AMT_INCOME_TOTAL*, *AMT_CREDIT*, *AMT_ANNUITY*, *AMT_GOODS_PRICE*, *CNT_FAM_MEMBE RS* and *AMT_CREDIT_MAX_OVERDUE*, we select the maximum value of each attributes as the input data.

For attribute *CREDIT_ACTIVE* and *NAME_CONTR ACT_STATUS*, we design two extra numerical attributes to represent each categorical attribute, as the number of active credit and number of completed credit.

For attribute *DAYS_CREDIT*, we select the minimum value to represent it. For attribute *CREDIT_T YPE*, we select the most frequent type to represent it. For the other attributes with numerical values, we choose the mean value of each attributes as the input data.

For the attributes with categorical values, such as *CODE_GENDER*, *NAME_CONTRACT_TYPE*, *NAME_EDUCATION_TYPE*, *NAME_FAMILY_STATUS*, *NAME_HOUSING_TYPE* and *NAME_INCO ME_TYPE*, we present the pie charts to observe the distribution, as shown in Fig. 4. From these figures, we can observe that the preference of users with different education background that 72.15% users holding secondary degrees and 23.94% users holding higher education degrees. Moreover, from the perspective of family status, 71.82% cardholders are married and 10.28% are single. From the perspective of income type, 51.38% cardholders are having working income.

4.2. Implementation

We have presented a flowchart of our proposed method, as shown in Fig. 5. In particular, after we have collected all accessible data, we split them with a ratio 7:3 to construct training datasets and testing datasets. And both training and testing datasets consist of three types of data, i.e. recent loan data, history loan data, and personal information. For each type of data, the distance between users can be calculated according to the Euclidean distance or Ahmad & Dey distance, respectively. The top 10% of these distances are taken for the edge construction, leading to three graphs in the training set and three graphs in the testing set. The model is well trained by using the three graphs on the training dataset and fully connected layers for fusion.

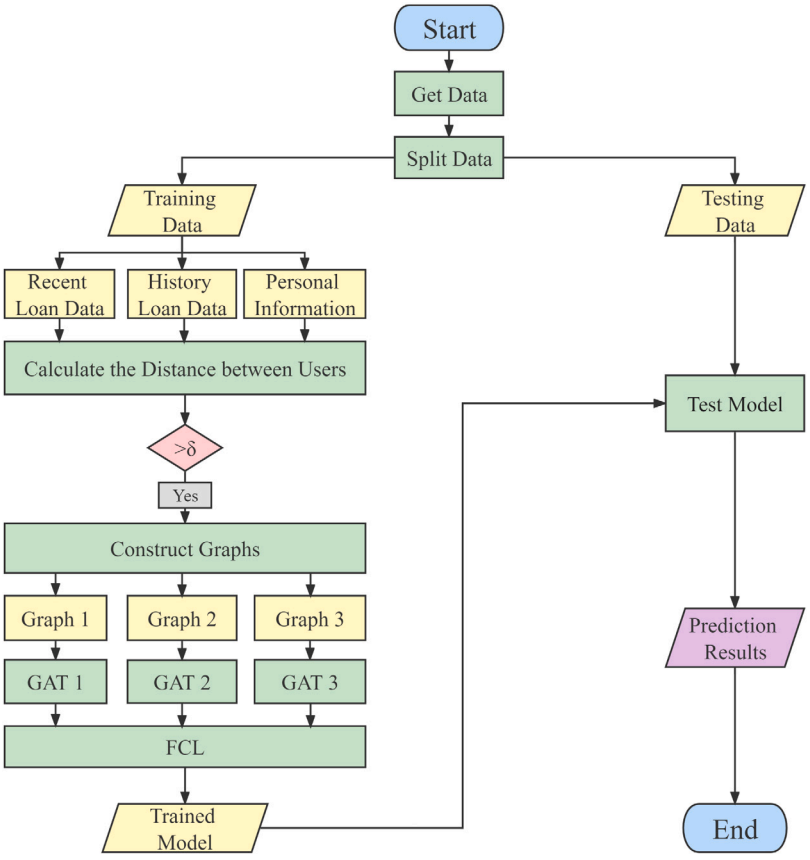


Fig. 5. A flowchart of our model.

And then we input the testing dataset into the trained model to generate the final credit default risk prediction results.

We implement our proposed method on a environment with one AMD EPYC 7502P CPU @ 3.35 GHZ and one NVIDIA RTX3090 24 GB card. The hyperparameters are determined by our proposed model's performance. The edges of three graphs are confirmed according to the top 10% similarities calculated, respectively. We select the Rectified Linear Unit (ReLU) as the activation function, and set 0.01 as the learning rate of our model.

4.3. Evaluation metric

In this study, we use the following metrics to evaluate the performance of our proposed method and other baseline methods, including Precision, Recall, F1-score, Accuracy and AUC-score. After the results prediction, we can obtain the confusion matrix, which present the number of true positives (TP), the number of false positives (FP), the number of true negatives (TN), and the number of false negatives (FN). Using the confusion matrix, we can compute the following metrics.

- Precision, this metric is used to depict the effectiveness of correctly predicted positives among all predicted positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall, this metric is used to present that the fraction of correctly predicted positives among all real positives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- F1-score, this metric is approximately an average of the Precision and Recall values.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Accuracy, this metric can show the correctly predicted values, including the positives and negatives.

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

- AUC-score, which refers to the area under the receiver operating characteristic curve, can be a useful metric to compare the prediction performance of different models. Specifically, we first compute the True Positive Rate (TPR) and False Positive Rate (FPR), then plot the receiver operating characteristic curve. After that, we normalize and divide the two values, followed by a Trapezoidal method for the AUC calculation.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$
$$\text{AUC} = Trapezoid([0; \text{FPR}; 1], [0; \text{TPR}; 1])$$

4.4. Baseline methods

To evaluate the prediction performance, we compare our proposed method with several popular baseline methods as follows.

- Support Vector Machine (SVM) (Cortes and Vapnik, 1995). SVM is a well-acknowledged and very robust prediction method for both classification and regression problems. For classification tasks, it maps each sample with class labels to points in the space and tries to maximize the gap width between different classes. When new samples are input, this method is able to map these samples into the same space and predict the exact side where the samples lie in. For non-linear classification tasks, this method provides different kernels to map the data samples into high-dimensional spaces.

Table 5
Comparison with baseline methods.

Methods	Accuracy	AUC-score	Precision	Recall	F1-score
SVM	0.8083 (2.98%)	0.7309 (5.69%)	0.7283 (1.98%)	0.5453 (16.03%)	0.6237 (9.56%)
DT	0.7494 (11.08%)	0.6910 (11.79%)	0.5725 (29.73%)	0.5512 (14.79%)	0.5616 (21.67%)
RF	0.8263 (0.74%)	0.7699 (0.34%)	0.7329 (1.34%)	0.6349 (−0.35%)	0.6804 (0.43%)
GNB	0.6082 (36.86%)	0.6822 (11.24%)	0.4163 (78.40%)	0.8593 (−26.37%)	0.5609 (21.82%)
Proposed method	0.8324	0.7725	0.7427	0.6327	0.6833

- Decision Tree (DT) (Safavian and Landgrebe, 1991). Decision tree is a tree-like structure that each node contains information on an attribute, in which leaf nodes refer to class labels and branches depict possible consequences. The paths from the root of the tree to the leaf nodes are rules extracted.
- Random Forest (RF) (Breiman, 2001). Random forest is an ensemble learning-type method for both classification and regression predictions, which consists of multiple independent decision trees. For classification tasks, for new data samples, each decision tree would output a classified result, and the final result that which class the data samples fall in would be selected from these decision trees.
- GaussianNB (GNB) (Hand and Yu, 2001; Perez et al., 2006). Naive Bayes method is a probabilistic-type method which applies Bayes' theorem on the data with a strong assumptions that features are independent. Gaussian Naive Bayes method assumes the data has a Gaussian distribution.

4.5. Experimental results

We conduct various experiments to verify the effectiveness of our proposed GAT-based model, including comparison with baseline methods, with extended baseline methods, component analysis, feature significance analysis and impact of parameters.

4.5.1. Baseline comparison

We compare the credit default risk prediction performance of our proposed method with several baseline methods. The comparison results are shown in Table 5. From this table, we observe that our model achieves the best prediction performance in terms of most metrics (Accuracy = 0.8324; AUC-score = 0.7725; Precision = 0.7427; F1-score = 0.6833), outperforming all the baseline methods. Compared with SVM, DT, RF and GNB, our GAT-based model achieves a substantial improvement, with 2.98%, 11.08%, 0.74%, and 36.86% in Accuracy, 5.69%, 11.79%, 0.34%, and 11.24% in AUC-score, 1.98%, 29.73%, 1.34%, and 78.40% in Precision, and 9.56%, 21.67%, 0.43%, and 21.82% in F1-score, respectively. These sets of experimental results demonstrate the efficacy and efficiency of our proposed GAT-based method. Using the same data, GNB obtains the worst prediction performance across four metrics and the best performance on the Recall metric. The explanation for this may be the data distribution assumption in GNB is not applicable in this problem. We also see that RF and SVM obtain comparable prediction performance with little difference, indicating the power and strong learning abilities of these two well-known approaches in typical classification problems.

Since our data is imbalance, we also adopt two methods to alleviate the imbalance issue between two classes, i.e. one under-sampling method and one over-sampling method. In particular, we conduct more experiments through applying classification methods after handling data imbalance issue by under-sampling and over-sampling methods, which we name them as extended-baseline methods.

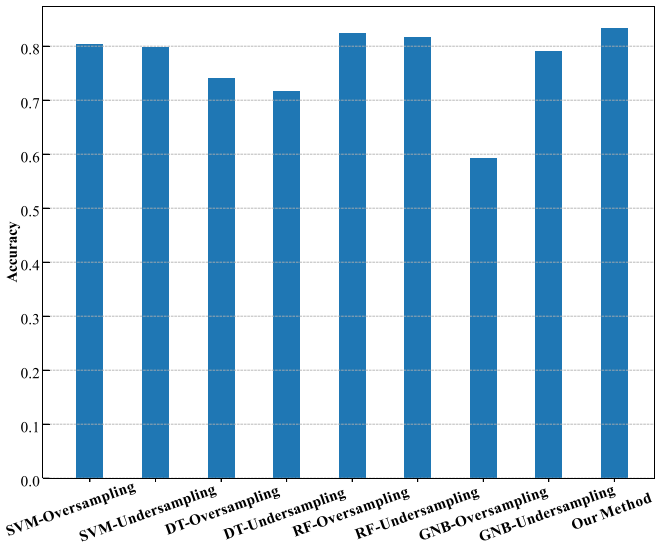


Fig. 6. Comparison with extended-baseline methods.

- RandomUnderSampler (Imbalanced-Learn, 2022). RandomUnderSampler is an easy-to-use undersampling method handling data imbalance issue in the way of selecting a subset of data points randomly for targeted classes.
- Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). SMOTE is an oversampling method in which the synthetic data points are generated for the minority category. It can alleviate the overfitting issue caused by random oversampling.

The comparison results between our GAT-based method and extended-baseline methods in terms of accuracy are demonstrated in Fig. 6. From the figure, we can observe that our method outperforms all these extended baseline methods, with both under-sampling operations and over-sampling operations, indicating the significance of relationship capturing in this problem. Extracting similarities between users and applying high-level features learned from similarities-constructed-based graphs, is an alternative way to alleviate the data imbalance issue. We can also observe that the RF-Oversampling and RF-Undersampling can achieve comparable performance with the RF baseline method. This represents the significance of random forest methods in dealing with data imbalance problems.

4.5.2. Component analysis

To further interpret the importance of components in this model, we conduct a component significance analysis study. We designate the following components configuration and compare their performance.

- P-GAT. This approach only takes personal profile of cardholders into account and then applies the Graph Attention Model after the graph construction for future credit default risk prediction.
- L-GAT. Similarly, this approach only considers the latest loan data, constructs the specific graph, and then uses the GAT model to predict credit default risk.
- H-GAT. Similarly, this approach only introduces the history loan records for graph construction and then feeds them into the GAT model for credit default risk prediction.
- PL-GAT. This approach utilizes both personal profile data and recent loan data for two graphs construction, and then applies GAT-based models and fully connected layers for further credit default risk prediction.
- PH-GAT. Similarly, this approach constructs different graphs employing both personal profile data and history loan records, then adopts the GAT-based model to predict credit default risk.

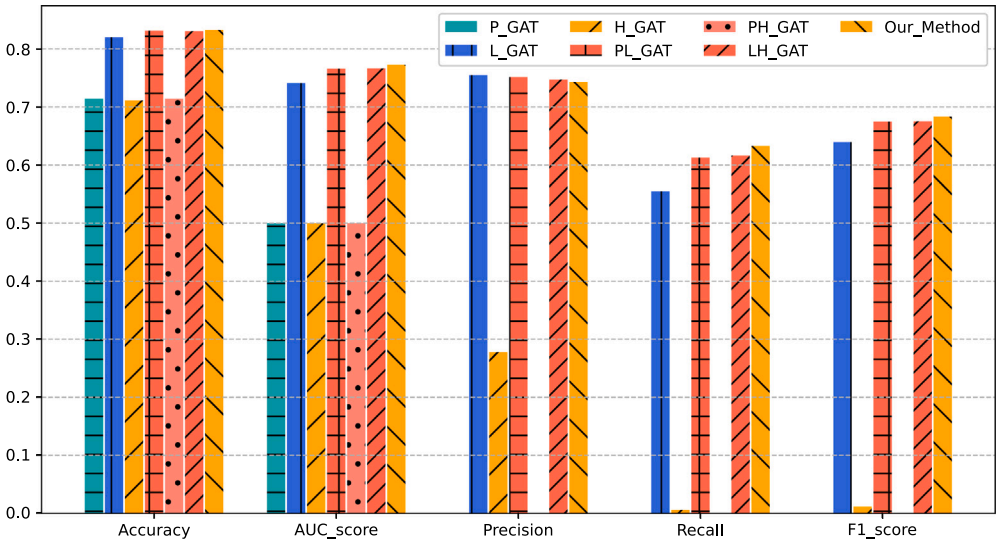


Fig. 7. Performance comparison between different components configuration.

Table 6
Top-10 attributes correlated with actual risk value.

	Method	Top-10 attributes	Coefficient
Actual risk value	Pearson correlation	CNT_INSTALMENT_MATURE_CUM	0.56
		Active_count_x	0.34
		Active_count_y	0.29
		Closed_count	0.25
		AMT_DRAWINGS_CURRENT	0.25
		AMT_CREDIT_LIMIT_ACTUAL	0.25
		CNT_DRAWINGS_CURRENT	0.22
		AMT_PAYMENT_CURRENT	0.20
		CNT_DRAWINGS_POS_CURRENT	0.20
		AMT_PAYMENT_TOTAL_CURRENT	0.19
	Random forest	CNT_INSTALMENT_MATURE_CUM	0.13
		Active_count_x	0.08
		Active_count_y	0.05
		AMT_DRAWINGS_CURRENT	0.05
		AMT_PAYMENT_CURRENT	0.04
		Closed_count	0.04
		CNT_DRAWINGS_CURRENT	0.04
		AMT_CREDIT_LIMIT_ACTUAL	0.03
		AMT_PAYMENT_TOTAL_CURRENT	0.03
		AMT_BALANCE	0.03

Table 7
Top-10 attributes correlated with predicted risk value.

	Method	Top-10 attributes	Coefficient
Predicted risk value	Pearson correlation	CNT_INSTALMENT_MATURE_CUM	0.79
		Active_count_x	0.48
		Active_count_y	0.42
		Closed_count	0.35
		AMT_CREDIT_LIMIT_ACTUAL	0.31
		AMT_DRAWINGS_CURRENT	0.29
		CNT_DRAWINGS_CURRENT	0.26
		AMT_PAYMENT_CURRENT	0.25
		AMT_PAYMENT_TOTAL_CURRENT	0.23
		CNT_DRAWINGS_POS_CURRENT	0.23
	Random forest	CNT_INSTALMENT_MATURE_CUM	0.29
		Active_count_x	0.11
		Active_count_y	0.07
		AMT_DRAWINGS_CURRENT	0.06
		Closed_count	0.05
		AMT_PAYMENT_CURRENT	0.05
		CNT_DRAWINGS_CURRENT	0.04
		AMT_PAYMENT_TOTAL_CURRENT	0.04
		AMT_CREDIT_LIMIT_ACTUAL	0.03
		AMT_BALANCE	0.03

- LH-GAT. Similarly, this approach select recent loan data and history loan records to compute the similarity and construct graphs. The GAT-based models and FCLs are utilized for credit default risk prediction.

The performance comparison between different components configuration are presented in Fig. 7. From the figure, we observe that our method, which consists of three types of graphs using the three types of data achieves the best performance, indicating the necessity and significance of the three graphs constructed based on the three datasets. From the perspective of single graphs, the L-GAT performs better than P-GAT and H-GAT, which means the importance of latest loan data. Compared the performance between L-GAT and bi-component configurations as PL-GAT and LH-GAT, we find that these bi-component configuration can obtain better performance than the single component L-GAT. The improvement indicates the importance of the P-GAT and H-GAT.

4.5.3. Feature significance analysis

We conduct comprehensive feature significance analysis study to investigate influential factors affecting final predictive outcomes. Two well-acknowledged methods, i.e., Pearson correlation and random forest, are employed for feature importance computation. We present the top-10 attributes correlated with actual credit default risk value in Table 6. From the table, we can observe that several attributes have obvious correlation with the actual credit default risk value, e.g. *CNT_INSTALMENT_MATURE_CUM*, our manually designated attributes *Active_count_x*, *Active_count_y*, *Closed_count* and so forth. Most of the top-10 attributes by Pearson Correlation method is the same as them calculated by Random Forest, which support the effect of these influential factors from different perspectives.

We also present the top-10 attributes correlated with predicted credit default risk value in Table 7. From this table, we can see that the top-10 attributes here is the same as them with the actual credit default risk value. And the corresponding correlation coefficient are higher than them with the actual risk value. For example, the Pearson correlation coefficient between the attribute *CNT_INSTALMENT_MATURE_CUM* and actual risk value is 0.56, while the coefficient is 0.79 between it and the predicted risk value. This indicates that the correlated attributes have more correlated relationship with the predictive values by our model, which further support the effectiveness of our proposed model for feature extraction and learning.

These correlated attributes can be used to measure the potential credit default risk. Take the attribute *CNT_INSTALMENT_MATURE_CUM* as an example. We analyze the data of the attribute, and discover that

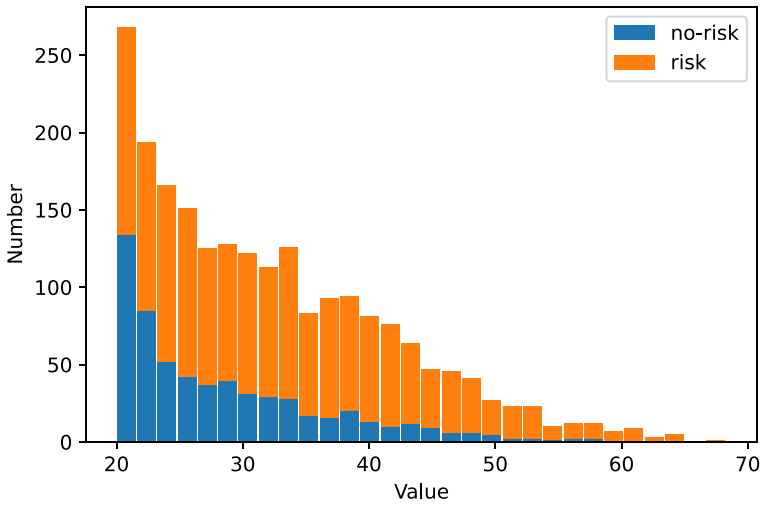
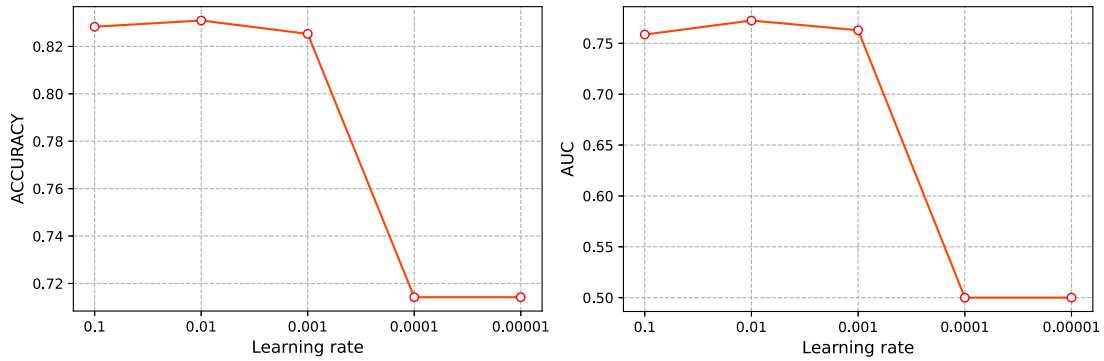


Fig. 8. Distribution of attribute *CNT_INSTALMENT_MATURE_CUM*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Accuracy (b) AUC score

Fig. 9. Impact of learning rate.

the probability of a credit default risk rises as the value of this attribute rises. In Fig. 8, we demonstrate the distribution of this attribute with values greater than 20. The orange bar represents the number of credit default risk, while the blue bar represents the number of credits without default risk. From the figure, when the attribute' value is 20, approximately 50% of the data corresponds to the default risk. As the value increases, so does the percentage of data corresponding to the default risk. When the value reaches 66, there is nearly a 100% probability of a credit default risk. Therefore, these correlated attributes can facilitate an explanation analysis and a rapid estimation of credit default risk.

4.5.4. Impact of learning rate

Experimental results of impact of learning rate are shown in Fig. 9. We observe that the value of learning rate has an obvious impact on the prediction performance, in terms of both accuracy and AUC score. This model would achieve a best performance when the learning rate set as 0.01. The prediction performance would drop drastically when the learning rate changes from 0.001 to 0.0001.

5. Discussion and conclusion

In this paper, we propose an effective model for predicting credit default risk from the perspective of various kinds of inherent relationships between users. The study contribution lies in the design of a GAT-based model, which aims to capture the latent relationships amongst users and harness them to enhance prediction precision. This study is well-targeted and relevant, which specifically targets the incorporation of GNN techniques to capture multiple relationships between users, addressing the issue of overlooking relationships of previous studies. To ensure producing meaningful results, our study focuses on the relationship construction between users, and finds that credit default risk is not solely determined by individual attributes but can be influenced by the dependencies between users. The main novelty of this study is two-fold. First, our GNN-based model facilitate the incorporation of complex relationships between users. Credit default risk is influenced by various factors, including the financial histories and behaviors of individuals. By incorporating multiple data types, our model is poised to capture the complex inter-relationships between users. This enhances our representation of the users' overall status, thereby refining the accuracy of our predictions. Second, our model can capture both linear and non-linear relationships between users in a flexible manner. A GAT module is leveraged that can effectively aggregate similar users' information by learning them from neighboring nodes, which captures both the relationships with adjacent and high-order neighbors, and finally improves the model's accuracy and contributes to the development of the financial domain. The final credit default risk predictive results, whether users will default or not, are predicted after fusing all these learned high-level features from GAT modules. Our study result can provide benefits for multiple corporations. For instance, risk management personnel of lending institutions, such as those at banks, can sense and identify potential economic losses at an early stage.

Extensive experiments on real-world datasets are conducted to verify the effectiveness of our GAT-based model. Experimental results demonstrate that our model is able to obtain accurate credit default

risk prediction and is superior to several baseline methods. Specifically, our model achieves the best prediction performance in terms of most metrics. Compared with SVM, DT, RF and GNB, our GAT-based model achieves a substantial improvement, with 2.98%, 11.08%, 0.74%, and 36.86% in Accuracy, 5.69%, 11.79%, 0.34%, and 11.24% in AUC-score, 1.98%, 29.73%, 1.34%, and 78.40% in Precision, and 9.56%, 21.67%, 0.43%, and 21.82% in F1-score, respectively. These sets of results depict the efficacy and efficiency of our GAT-based method. We also notice that our model obtains comparable performance to RF and worse performance than GNB in terms of the Recall metric, while GNB obtains much worse performance in terms of the other four metrics. The possible explanation may be that the data distribution assumption in GNB is not applicable in our case.

With the assumption that customers exhibiting similar financial behavior will exhibit similar risk profiles and financial outcomes, our study aims to predict the credit default risk of users by learning knowledge from similar users. These similar users can be identified and connected through leveraging datasets from multiple views, leading to various types of connections. For instance, users with similar loan histories would be regarded as similar users from the perspective of historical financial credit. When some financial institutions take irregular actions for some reason, data deviation or data noise will be caused. This may have some effect on the credit score. However, leveraging multi-source data could be a promising way to cope with this situation. The performance of methods using one single source of data would be highly dependent on the data quality. When multiple source datasets are collected, methods' performance are stable and robust in the face of data noise or deviation. This is because different source data can be used to improve the accuracy of the final prediction by enhancing the learned features, which would reduce the effect of data deviation or data noise to some extent.

We further analysis the theoretical and practical contribution of this paper. From a theoretical perspective, the most significant contribution is to provide a paradigm for financial index prediction in electronic commerce field using artificial intelligence algorithms. Multi-source data are leveraged to represent various underlying relationships in a comprehensive way. A GAT-based model is then proposed to augment the features, and then automatically learn significance between similar samples. From a practical perspective, our model offers significant potential in credit default risk management for companies and financial institutions. By accurately predicting credit default risks, companies can sense and identify potential risky customers and take appropriate actions to mitigate the risks at early stage. The model can help improve decision-making in the lending process and enable proactive risk management strategies.

While our study makes significant strides towards credit default risk prediction, there are some limitations that can serve as pathways for future research. From a data perspective, our study was constrained by the use of static data sourced from a single company, owing to challenges in accessing more comprehensive datasets. Consequently, a pertinent avenue for future exploration is to gather more dynamic data from multiple companies to learn a more general model. From a scenario perspective, our focus was primarily on individual credit default risk. In subsequent research, we aim to expand our scope to encompass corporate credit default risk prediction, or other domains.

CRediT authorship contribution statement

Binbin Zhou: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Jiayun Jin:** Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Hang Zhou:** Methodology, Writing – original draft. **Xuye Zhou:** Writing – review & editing. **Longxiang Shi:** Writing – review & editing. **Jianhua Ma:** Writing – review & editing. **Zengwei Zheng:** Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared all the code and data in the manuscript.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (62102349), and Scientific Research Project of Zhejiang Provincial Education Department, China (Y202248716) and Research Project of Science and Education Innovation Complex of Hangzhou City University, China (No. 22FHXM04). The authors would like to acknowledge the Supercomputing Center of Hangzhou City University for the support of the advanced computing resources.

References

Addo, P.M., Guegan, D., Hassani, B., 2018. Credit risk analysis using machine and deep learning models. Risks 6, 38.
Agosto, A., Giudici, P., Leach, T., 2019. Spatial regression models to improve P2P credit risk management. Front. Artif. Intell. 2, 6.
Ahmad, A., Dey, L., 2007. A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl. Eng. 63, 503–527.
Ashraf, S., Gao, M., Chen, Z., Naeem, H., Ahmad, A., Ahmed, T., 2020a. Underwater pragmatic routing approach through packet reverberation mechanism. IEEE Access 8, 163091–163114.
Ashraf, S., Muhammad, D., Shuaeeb, M., Aslam, Z., 2020b. Development of shrewd cosmetology model through fuzzy logic. J. Res. Eng. Appl. Sci. 5, 93–99.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Chen, R., Yu, H., 2014. Risk measurement for portfolio credit risk based on a mixed Poisson model. Discrete Dyn. Nat. Soc. 2014.
People's Bank of China, 2021a. Payment system report (2020). http://www.pbc.gov.cn/goutongjiaoliu/113456/113469/4213347/index.html.
People's Bank of China, 2021b. A steady growth in payment business and consumption during double eleven shopping festival. https://news.cctv.com/2021/11/12/ARTItjNTv2rYcXugoLGoa6Rk211112.shtml.
People's Bank of China, 2022. Payment system report (2021). http://www.pbc.gov.cn/zhifujiesuansi/128525/128545/128643/4523666/index.html.
China Banking and Insurance Regulatory Commission, 2020. http://www.gov.cn/fuwu/2020-07/06/content_5524417.htm.
China Marketing, Research & Digital China Skinny, 2021. Data analysis of 2021 double eleven shopping festival. www.chinaskinny.com/cn/blog-zh/2021-singles-day-data-2.
China Times, 2022. https://m.chinatimes.net.cn/article/116281.html.
Cisi, M., Devicienti, F., Manello, A., Vannoni, D., 2020. The advantages of formalizing networks: New evidence from Italian SMEs. Small Bus. Econ. 54, 1183–1200.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.
Cowden, C., Fabozzi, F.J., Nazemi, A., 2019. Default prediction of commercial real estate properties using machine learning techniques. J. Portfolio Manag. 45, 55–67.
Duan, J.-C., Sun, J., Wang, T., 2012. Multiperiod corporate default prediction—A forward intensity approach. J. Econometrics 170, 191–209.
Guo, H., Peng, K., Xu, X., Tao, S., Wu, Z., 2020. The prediction analysis of peer-to-peer lending platforms default risk based on comparative models. Sci. Program. 2020.
Hamori, S., Kawai, M., Kume, T., Murakami, Y., Watanabe, C., 2018. Ensemble learning or deep learning? Application to default risk analysis. J. Risk Financial Manag. 11, 12.
Hand, D.J., Yu, K., 2001. Idiot's Bayes—not so stupid after all? Int. Stat. Rev. 69, 385–398.
Hu, Y., Liu, K., Zhang, X., Xie, K., Chen, W., Zeng, Y., Liu, M., 2015. Concept drift mining of portfolio selection factors in stock market. Electron. Commer. Res. Appl. 14, 444–455.
Imbalanced-Learn, 2022. Under sampling. https://imbalanced-learn.org/stable/under_sampling.html.
Kvamme, H., Sellereite, N., Aas, K., Sjursen, S., 2018. Predicting mortgage default using convolutional neural networks. Expert Syst. Appl. 102, 207–217.
Lappas, P.Z., Yannacopoulos, A.N., 2021. A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. Appl. Soft Comput. 107, 107391.
Laudon, K.C., Traver, C.G., 2013. E-Commerce. Pearson Boston, MA.

Lee, J.W., Lee, W.K., Sohn, S.Y., 2021. Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. Expert Syst. Appl. 168, 114411.
Lee, J.W., Sohn, S.Y., 2021. Evaluating borrowers' default risk with a spatial probit model reflecting the distance in their relational network. PLoS One 16, e0261737.
Liang, L., Cai, X., 2020. Forecasting peer-to-peer platform default rate with LSTM neural network. Electron. Commer. Res. Appl. 43, 100997.
Liu, J., Zhang, S., Fan, H., 2022. A two-stage hybrid credit risk prediction model based on xgboost and graph-based deep neural network. Expert Syst. Appl. 195, 116624.
Luo, P., Chen, K., Wu, C., 2016. Measuring social influence for firm-level financial performance. Electron. Commer. Res. Appl. 20, 15–29.
Luo, C., Wu, D., Wu, D., 2017. A deep learning approach for credit scoring using credit default swaps. Eng. Appl. Artif. Intell. 65, 465–470.
Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., Niu, X., 2018. Study on a prediction of P2P network loan default based on the machine learning LightGBM and xgboost algorithms according to different high dimensional data cleaning. Electron. Commer. Res. Appl. 31, 24–39.
Ogiela, L., 2015. Intelligent techniques for secure financial management in cloud computing. Electron. Commer. Res. Appl. 14, 456–464.
Perez, A., Larranaga, P., Inza, I., 2006. Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes. Internat. J. Approx. Reason. 43, 1–25.
Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybern. 21, 660–674.
Sharma, P., Sivakumaran, B., 2015. Investigating impulse buying and variety seeking: Towards a general theory of hedonic purchase behaviors. In: Developments in Marketing Science: Proceedings of the Academy of Marketing Science. p. 61.
Shen, F., Zhao, X., Kou, G., Alsaadi, F.E., 2021. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. Appl. Soft Comput. 98, 106852.
Sohn, S.Y., Kim, J.W., 2012. Decision tree-based technology credit scoring for start-up firms: Korean case. Expert Syst. Appl. 39, 4007–4012.
Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. arXiv preprint arXiv:1710.10903.
Wang, G., Chen, G., Chu, Y., 2018a. A new random subspace method incorporating sentiment and textual information for financial distress prediction. Electron. Commer. Res. Appl. 29, 30–49.
Wang, C., Han, D., Liu, Q., Luo, S., 2018b. A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. IEEE Access 7, 2161–2168.
Wang, L., Song, H., 2022. E-commerce credit risk assessment based on fuzzy neural network. Comput. Intell. Neurosci. 2022.
Xia, Y., Li, Y., He, L., Xu, Y., Meng, Y., 2021. Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending. Electron. Commer. Res. Appl. 49, 101095.
Xia, Y., Liu, C., Liu, N., 2017. Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. Electron. Commer. Res. Appl. 24, 30–49.
Yang, M., Lim, M.K., Qu, Y., Li, X., Ni, D., 2022. Deep neural networks with L1 and L2 regularization for high dimensional corporate credit risk prediction. Expert Syst. Appl. 118873.
Yao, G., Hu, X., Zhou, T., Zhang, Y., 2022. Enterprise credit risk prediction using supply chain information: A decision tree ensemble model based on the differential sampling rate, synthetic minority oversampling technique and AdaBoost. Expert Syst. e12953.
Yao, L., Mao, C., Luo, Y., 2019. Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7370–7377.
Zhang, M., Chen, Y., 2018. Link prediction based on graph neural networks. Adv. Neural Inf. Process. Syst. 31.
Zhang, W., Wang, C., Zhang, Y., Wang, J., 2020. Credit risk evaluation model with textual features from loan descriptions for P2P lending. Electron. Commer. Res. Appl. 42, 100989.
Zhu, L., Qiu, D., Ergu, D., Ying, C., Liu, K., 2019. A study on predicting loan default based on the random forest algorithm. Procedia Comput. Sci. 162, 503–513.