Electronic Research Archive

ERA, 31(7): 4057–4076.
DOI: 10.3934/era.2023206
Received: 22 February 2023
Revised: 27 April 2023
Accepted: 03 May 2023
Published: 25 May 2023

http://www.aimspress.com/journal/era

Research article

Machine learning model of tax arrears prediction based on knowledge graph

Jie Zheng* and Yijun Li

School of Management, Harbin Institute of Technology, Harbin 150001, China

* **Correspondence:** Email: 416684500@qq.com.

Abstract: Most of the existing research on enterprise tax arrears prediction is based on the financial situation of enterprises. The influence of various relationships among enterprises on tax arrears is not considered. This paper integrates multivariate data to construct an enterprise knowledge graph. Then, the correlations between different enterprises and risk events are selected as the prediction variables from the knowledge graph. Finally, a tax arrears prediction machine learning model is constructed and implemented with better prediction power than earlier studies. The results show that the correlations between enterprises and tax arrears events through the same telephone number, the same E-mail address and the same legal person commonly exist. Based on these correlations, potential tax arrears can be effectively predicted by the machine learning model. A new method of tax arrears prediction is established, which provides new ideas and analysis frameworks for tax management practice.

Keywords: tax arrears prediction; knowledge graph; machine learning; enterprise association relationship; risk contagion

1. Introduction

As the paramount source of national public finance, taxation is a normative form of social resource allocation. Tax compliance refers to the compliance of taxpayers with national taxation regulations and obligations. There is no doubt regarding the enterprises' social responsibility in terms of compliance with tax payment [1]. A low level of tax compliance brings a serious risk of tax erosion to a country. Tax arrears of enterprise is a typical form of tax non-compliance. It is ubiquitous and hugely impactful worldwide. Accurate prediction of potential tax arrears is theoretically and practically

important in taxation management.

Current research on predicting tax arrears behavior mainly focuses on individual taxpayers and corporate financial indicators. Marghescu [2] predicts the tax arrears behaviors of 328 Finnish enterprises using logistic regressions. Scholars also deploy various machine learning algorithms to predict tax non-compliance [3–6]. The summary of the above research results is shown in Table 1.

Table 1. Tax arrears behavior and other tax non-compliance prediction results comparison.

Research literature	Data source/sample size	Prediction method	Independent variable	Prediction object	Predictive ability/accuracy
Marghescu. et al. (2010)	Finland/328	Logistic regression	Financial index	Tax arrears	61.6%
Su. et al. (2018)	China/120,000	Integration model	Financial indicators, basic attributes of enterprises	Tax arrears	90.6%
Ippolito. et al. (2021)	Brazil/604	XGBoost. et al.	Invoices, tax penalties	Tax crime	0.998 (AROC)
Vanhoeyveld. et al. (2020)	Belgium/950,000	FWAD exception detection	Tax regulatory ratio	VAT fraud	5 to 100 times higher than the fraud detected by random selection
Abedin. et al. (2021)	Finland/768	XGBoost. et al.	Financial index	Tax arrears	71.9%
Siimon and Lukason (2021)	Estonian/49,156	decision tree. et al.	Previous tax arrears	Tax arrears	95.28%

There are still several shortcomings in current research:

First, existing research relies heavily on enterprise financial data, which has high requirement for the integrity and authenticity of financial information. From a broader point of view, financial data is also one of the important factors to predict default risk. A typical one is Omega Score, which includes financial information such as enterprise financial ratios and transaction data [7]. However, a large number of small and medium-sized enterprises have poor information transparency. There are barriers to data source [8]. Filing false financial data to the government is one of the main methods of tax evasion. Therefore, the reliability of tax compliance risk analysis based on financial data is low. Moreover, the financial statements of non-listed enterprises are voluntarily disclosed. The financial statements of enterprises filed in government departments are difficult to obtain at will. Therefore, without a large amount of legally authorized financial data, it is difficult to carry out more extensive and in-depth research.

Second, the relationship between enterprises is seldom considered in tax arrears forecasting. According to the tax contract theory, tax is essentially a contract established between all taxpayers and the government. Tax non-compliance has obvious characteristics of contract breach [8]. Empirical research shows that default events may lead to a greater probability of similar risks in its affiliates. An enterprise that has incurred tax arrears or dishonest behavior will lead to a significant increase in the probability of tax arrears by its associated enterprises (with the same legal person, the same contact information and so on). There are significant correlations with conductivity and few risks exist in isolation. Risk contagion is reflected in various scenarios, for example, financial risk between countries

and tax compliance risk between taxpayers [9,10]. When considering the risk contagion, a network perspective is conducive to a deeper understanding of the complex connection between enterprises and their related events [11]. In fact, in addition to financial data, the relationship between enterprises is also an important factor to predict the risk of tax arrears. There have been studies focusing on the correlations between enterprises to predict default risk. However, they still rely on financial data. For example, some studies establish enterprise correlation network based on financial ratios and transaction data between enterprises, then use correlation network to forecast default risk [12,13]. The ideas and methods adopted in these default prediction studies are of great reference significance for tax arrears prediction.

Third, sample imbalance is a quite common nature in the field of risk prediction [14]. The measurement indexes of overall classification ability represented by literatures [3–6] include Accuracy, AROC and so on. It is suitable for the balanced distribution of positive and negative samples. However, the samples of tax arrears and financial fraud are not balanced. The above models neglect the ability to classify minority samples. For example, research [15] predicts the future risk of tax arrears based on an enterprise's own past tax arrears. Although the above dependence on financial data is avoided, the accuracy of the model evaluation method cannot fully reflect the prediction power for tax arrears. Research [15] also emphasizes that accuracy is not the best reference for model evaluation when samples are not balanced. Therefore, in the specific research process, other auxiliary indexes are also referred to evaluate the model.

The concept and method of knowledge graph can effectively analyze the correlations between enterprises and provide more variable choices for the prediction of tax arrears. It can further improve and optimize the tax arrears forecast research. In order to predict tax arrears of enterprise with more available data and more efficient model, in the remainder of this paper, a knowledge graph of enterprises is established. The correlations of enterprises with same telephone, E-mail and legal person are extracted to form new relations among tax arrears, trust-breaking and enterprises. Finally, a machine learning model for tax arrears prediction is established with the factors extracted from knowledge graph and the factors related with other aspects such as macroeconomic and business.

2. Construction of enterprise knowledge graph

Knowledge graph describes entities such as people, objects, concepts and their relationships in the form of graph. Comprising information on the property and structure of graph, the complex relationships of the real world can be better presented. The specific construction process involves designing pattern graph, data graph, knowledge extraction and knowledge graph storage [16,17].

The research on tax arrears behavior in this paper involves a typical industrial knowledge graph employing the bottom-up construction mode. It includes four types of entities: enterprises, contact information (telephone number, E-mail address), legal persons and events (tax arrears, trust-breaking). It also includes five types of initial relationships: "same corporate E-mail", "same corporate phone number", "same corporate legal person", "same tax arrears subject" and "same trust-breaking subject".

The set of all enterprise nodes is denoted as $N_d^{(c)}$. The set of E-mail node is denoted as $N_d^{(e)}$. For any enterprise instance $n_i^{(c)} \in N_d^{(c)}$ and E-mail instance $n_x^{(e)} \in N_d^{(e)}$, if $n_x^{(e)}$ is the E-mail address of $n_i^{(c)}$, the relation between enterprise and E-mail is established $r_{i,x}^{(c,e)}$: $\left(n_i^{(c)}\right) - \left[r_{i,x}^{(c,e)}: \mathrm{E-mail}\right] \rightarrow \left(n_x^{(e)}\right)$, and the set of $r_{i,x}^{(c,e)}$ is denoted as $R_d^{(c,e)}$. Similarly, the set of phone number nodes is denoted as $N_d^{(p)}$, and relation between enterprise and phone number is $r_{i,x}^{(c,p)}$, the relation set is denoted as $R_d^{(c,p)}$. The set of legal person nodes is denoted as $N_d^{(lp)}$, and relation between enterprise and legal person is $r_{i,x}^{(c,lp)}$, the

relation set is denoted as $R_d^{(c,lp)}$. The set of event nodes for tax arrears is denoted as $N_d^{(tr)}$, the relation between enterprise and event of tax arrears is $r_{i,x}^{(c,tr)}$, the relation set is denoted as $R_d^{(c,tr)}$. The set of event nodes for trust-breaking is denoted as $N_d^{(cr)}$, relation between enterprise and event of trust-breaking is $r_{i,x}^{(c,cr)}$, the relation set is denoted as $R_d^{(c,cr)}$.

The schema graph and data graph based on the above settings are shown in Figure 1.

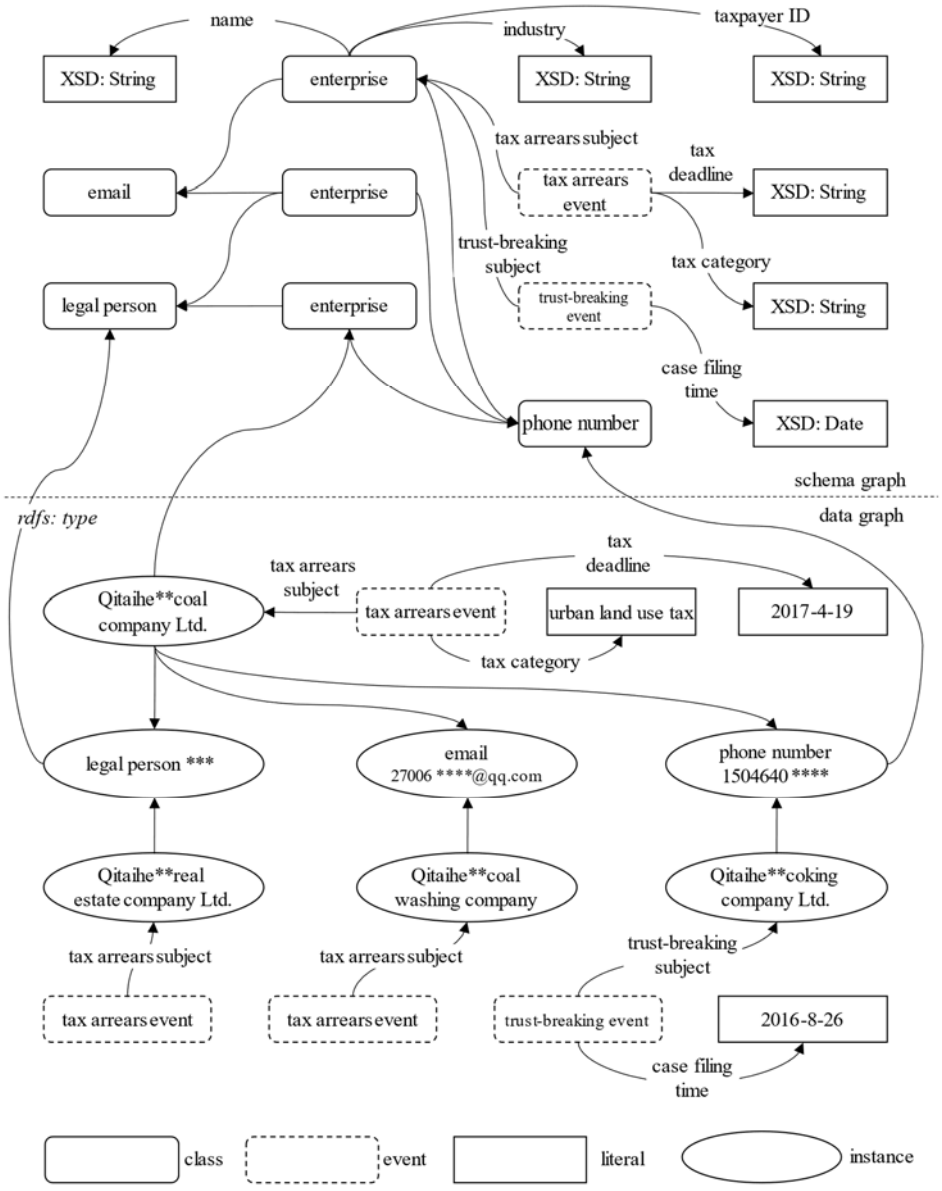

Figure 1. Schema graph and data graph of knowledge graph.

On one hand, Figure 1 can reflect the basic attribute information of the enterprise itself, including the enterprise name, industry, taxpayer identification number and so on. The information can be used as a variable to predict enterprise tax arrears and an important identifier for entity alignment. On the other hand, enterprise nodes can establish relationships through nodes, such as legal person, phone number and E-mail address. Furthermore, the correlations between different risk events can be

established. For example, if two different enterprises are at the two ends of a phone number node, they can establish an association. The tax arrears or trust-breaking events associated with the two enterprises can also form association relationships through telephone number. Learning from ideas of [18,19], trust-breaking events can be regarded as tax compliance risk related events. Thus, the relationships through telephone numbers can provide data support for the prediction of tax arrears in the later part of this paper.

In this paper, the data needed for constructing graphs mainly comprises information on enterprises, tax arrears events and trust-breaking events. The sources primarily include two types: One type comes from the official data publication channels, such as the Credit Information Disclosure System for Enterprises Nationwide and the tax arrears information announcements of taxation authorities, which can be extracted through web scraping. The other type originates from third-party data service providers such as TianYanCha.com and Qcc.com. As they have adopted more comprehensive anti-scraping measures, corporate credit reports are obtained in bulk through official channels.

This paper adopts the primary and secondary integrated data collection method. First, structured or semi-structured data is collected from secondary data sources such as TianYanCha.com and Qcc.com. Its main form is the third-party enterprise credit report provided by the data service company. Next, more direct original data is extracted from data sources as supplements: 1) national enterprise credit information publicity system (gsxt.gov.cn); 2) tax default notice of Heilongjiang Electronic Tax Bureau of State Administration of Taxation (etax. heilongjiang chinatax.gov.cn).

Furthermore, the unified social credit code or taxpayer identification number is taken as the sole identifier for an enterprise. Corporate entities of different data sources are merged and aligned. For the very few samples lacking a unified social credit code or taxpayer identification number, the entities are aligned on the basis of the textual similarity of corporate names using the method in literature [20].

Based on the above data collection methods, this paper obtains the basic information of all enterprises in Heilongjiang Province, China from January 1, 2014 to June 30, 2021. After cleaning and processing the original data, the knowledge graph is established. It contains nearly 2.85 million nodes and 3.59 million initial relationships in the knowledge graph, as shown in Table 2. The graph database Neo4j is used to realize the storage and query of knowledge graph.

Table 2. Numbers of nodes and initial relationships.

Node	Number of nodes	Relationship	Number of relationships
Enterprise	986,471	Enterprise - E-mail	1,266,347
E-mail	583,248	Enterprise - Phone number	1,223,126
Phone number	716,778	Enterprise - Legal person	986,471
Legal person	427,318	Enterprise - Tax arrears	96,764
Tax arrears	112,950	Enterprise - Trust-breaking	16,777
Trust-breaking	18,347		
Total	2,845,112	Total	3,589,485

Due to the large scale of the knowledge graph, a subgraph is selected for display. Figure 2 shows a subgraph containing 487 nodes and 437 relationships. Figure 3 shows a subgraph containing 11 nodes and 13 relationships.

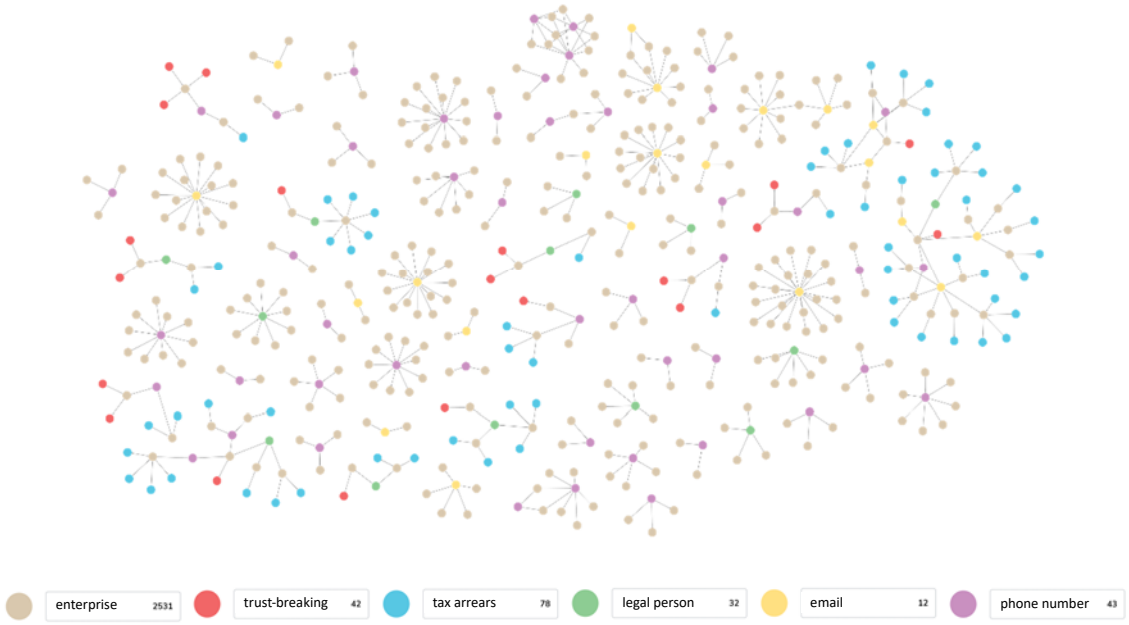

Figure 2. Example 1 of Knowledge Graph 1 (including 487 nodes and 437 relationships).

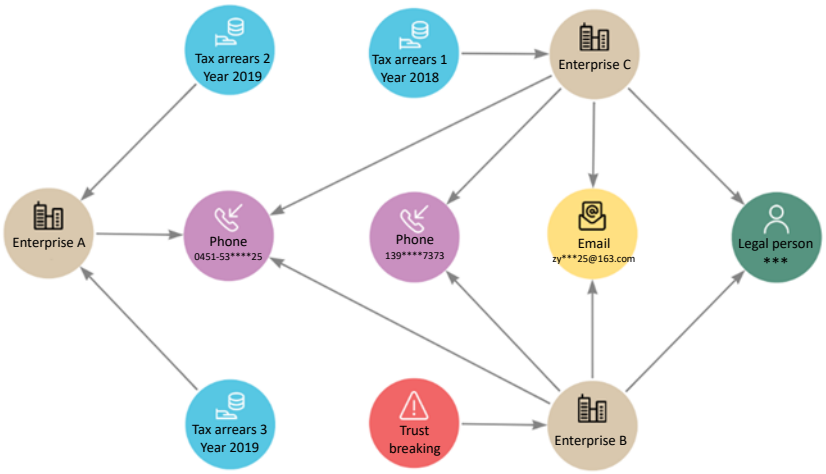

Figure 3. Example 2 of Knowledge Graph (including 11 nodes and 13 relationships).

3. Selection of tax arrears prediction factors

The influencing factors of default events include not only the own characteristics of an enterprise, but also social, industry and macroeconomic factors [21–24]. This paper adopts the analytical framework of BISEP model (considering five aspects of business, industry, sociological, economic and psychological to analyze enterprises' activities) for tax arrears prediction [25]. With this analytical framework, tax compliance including tax arrears is influenced by business, industry, sociological, economic and psychological factors respectively. These 5 dimensions of factors include enterprise's own attributes, general status of the industry, social relationship of enterprise, macroeconomic, risk

preference, previous interaction with Inland Revenue of enterprise and so on. However, the BISEP model mainly focuses on enterprise's isolated characteristics which influence tax arrears.

In BISEP, the correlations between enterprises are ignored. In fact, the tax compliance risk of an enterprise is not only closely related to its own credit level and business status, but also to other entities (individuals, enterprises and so on) in the social economic system. Related risks also influence and transform each other [26,27]. If various correlations between enterprises and the external environment are ignored, the tax compliance risk is only studied from the perspective of individual enterprises. It will make many enterprises with tax compliance risk slip through the net, resulting in serious tax loss.

Tax compliance risk is not a separate form of risk, but essentially a form of payment default risk. Tax compliance risk is closely related to bankruptcy risk. Lukason and Andresson compared the predictive ability of financial ratios and tax arrears to bankruptcy [8]. The result shows that tax arrears are more accurate than financial ratios in predicting bankruptcy risk. Different default behaviors can also be used for the evaluation and prediction of tax arrears. In this paper, we mainly consider tax arrears events and trust-breaking events. Tax Arrears mean the tax not paid in due time by the taxpayer in accordance with the procedure laid down in the tax law. Trust-breaking events are related to dishonest judgement debtor. The official explanation of dishonest judgement debtor is "a person or unit who has the ability to perform but refuses to perform an effective legal document-determined obligation". The trust breaking events can be understood as a private debt that is subjectively unwilling to be repaid, while the tax arrears can be understood as a public debt owed to the country that has not been repaid.

On one hand, the more similarities exist in telephone number, E-mail address and legal person between two different enterprises, the higher probability exists that the same person or highly correlated people involve in business of the two enterprises. Therefore, the behavior pattern and risk preference of tax compliance will be similar with higher probability, which lead to similar tax arrears between the two enterprises. On the other hand, there are correlations between different contract violations. Therefore, this paper considers not only enterprise's isolated attributes and contract violations, but also the related enterprise's events of tax arrears and trust-breaking. Considering the availability of data, in the framework of the extended BISEP model, this paper chooses the characteristics of the enterprise itself, the industry situation of the enterprise, the macro situation, the social relationships of the enterprise, the tax arrears behavior and the trust-breaking behavior as the factors for tax arrears prediction.

In order to reduce the collinearity between industry and economic factors, and the collinearity between sociological and psychological factors, these factors are merged to form three categories, which are shown in Table 3 together with dependent variable. The first category is the macroeconomic indicators of the region where the enterprise is located (F1), which has a fundamental impact on the operation of the enterprise, and then affects the ability to pay taxes. The second category is the enterprise's own attributes (F2), which mainly includes registered capital, industry, city, district, county, enterprise type and so on. The third category is enterprise-related tax arrears and trust-breaking events (F3). If one enterprise or its associated enterprises (with same contact phone number, same E-mail address, same legal person and so on) have ever been involved in tax arrears or trust-breaking events, the probability of tax arrears will be significantly increased.

Table 3. Prediction variables with merged classification.

variable classification	variable name
Macroeconomic in year *t*(F1)	Index of tax revenue of the city where the enterprise is located
	GDP growth index of the city where the enterprise is located
	GDP proportion of the industry
	GDP growth index of the industry
Business(F2)	Registered capital
	Industry
	Location city
	Location county/district
	Type of enterprise
Enterprise-related tax arrears/trust-breaking events in year *t*(F3)	The average number of tax arrears of E-mail related enterprises
	The average number of trust-breaking of E-mail related enterprises
	The average number of tax arrears of phone number related enterprises
	The average number of trust-breaking of phone number related enterprises
	The average number of tax arrears of legal person related enterprises
	The average number of trust-breaking of legal person related enterprises
	The number of tax arrears of the enterprise
	The number of trust-breaking events of the enterprise
dependent variable	tax arrears events of the enterprise in year $t + 1$ (dummy variable, 1 for observations with tax arrears, 0 for observations without tax arrears)

4. Construction of machine learning model for tax arrears prediction

4.1. Choice of machine learning model algorithm

Boosting is a common machine learning method that builds an integrated learning model capable of solving complex tasks by combining several simple learning models. As a boosting-based machine learning algorithm, gradient boosting decision tree (GBDT) boasts advantages such as fast computation, excellent generalization expression and strong robustness. Research has shown that the GBDT methods are most suitable for default risk prediction, as they perform better than neural networks and logistic regression [28]. XGBoost is an engineering implementation and functional advancement of the GBDT algorithm. LightGBM, in turn, further improves and refines XGBoost by 1) adopting a histogram-based decision tree algorithm to reduce computational complexity; 2) using a decision tree leaf node splitting strategy with a depth limit to decrease errors and enhance accuracy; 3) increasing the attention paid to large-gradient samples and computational efficiency without changing the original distribution characteristics of the data through gradient-based one-side sampling (GOSS); 4) bringing down the feature dimension and further improving computational speed through a mutually exclusive feature bundling. LightGBM can perform better than logistic regression, neural networks and other algorithms like SVMs for default risk prediction [29]. It has been widely used in the research of default risk prediction due to its advantages [30–32]. Considering the research requirements and the actual selection of related research, this paper selects the LightGBM algorithm framework to construct the machine learning model.

4.2. Choice of model evaluation method

For both in-sample and out-of-sample forecasting, choice of model evaluation method is always an important aspect to consider [33,34]. For the binary classification problem, a positive example is true positive (TP) or false negative (FN) if it is judged as positive or negative by the model, respectively; a negative example is false positive (FP) or true negative (TN) if it is judged as positive or negative, respectively. The confusion matrix [35] built on the above situations is shown in Table 4:

Table 4. Confusion matrix.

		Actual values	
		Positive	Negative
Predicted values	Positive	TP	FP
	Negative	FN	TN

According to the confusion matrix, the corresponding machine learning model evaluation indicators can be worked out, specifically including $Accuracy = (TP + TN)/(TP + TN + FP + FN)$, $Recall(hit\ rate, true\ positive\ rate, sensitivity) = TP/(TP + FN)$, $FPR = FP/(TN + FP)$ and $Precision(hit\ rate) = TP/(TP + FP)$.

Based on the above indicators, an ROC curve can be generated with FPR as the horizontal axis and TPR as the vertical axis. A PR curve can be created with Recall as the horizontal axis and Precision as the vertical axis. The areas below these two curves are labelled ROC-AUC and PR-AUC, respectively, which can serve as important indicators to evaluate the performance of machine learning models. For ROC-AUC, the greater the curve convexity is up to the left, the better. For PR-AUC, the greater the curve convexity is up to the right, the better. For ROC-AUC and PR-AUC, the closer the value is to 1, the better the classification performance of the model [36].

Because of the better robustness, the ROC curve enables ROC-AUC to not change drastically in the context of unbalanced samples, making it an outstanding evaluation indicator, as it considers positive and negative samples in measuring the overall performance of model classification [37]. However, in scenarios such as tax arrears prediction and financial fraud, where positive and negative samples are maldistributed, more attention should be paid to positive samples (i.e., the presence of tax arrears and financial fraud). ROC-AUC poorly reflects the impact of changes in sample equilibrium on the model classification and cannot explicitly compare the advantages and disadvantages of different models. In contrast, the PR curve produces greater differences when samples are not balanced, thus proving to be more suitable for tax arrears prediction.

In model evaluation, classification ability is generally considered acceptable for models with an AUC of more than 0.7 and regarded as outstanding if AUC surpasses 0.9. Evaluation results corresponding to different AUC values [38] are shown in Table 5:

Table 5. Evaluation level of AUC for model classification ability.

AUC interval	Model classification ability
$(0.5, 0.7)$	poor
$[0.7, 0.8)$	acceptable
$[0.8, 0.9)$	excellent
$[0.9, 1.0]$	outstanding

4.3. Model parameter determination based on knowledge graph

For the three types of predictor variables in Table 3, F1 can be obtained from public government information resources such as statistical yearbooks and statistical bulletins. F2 can be extracted from the base data obtained from the aforementioned data service sources. While F3 needs to be extracted from the generated knowledge graph. The following focuses on the extraction method of F3.

First, based on the initial relations in the knowledge graph, the following new relations are constructed:

Generate E-mail relation between enterprises as $r_{i,j}^{(c,e,c)}$: $\left(n_i^{(c)}\right) - \left[r_{i,j}^{(c,e,c)}: \mathrm{E-mail\ related\ enterprises}\right] \rightarrow \left(n_j^{(c)}\right)$. The relation set is denoted by $R_d^{(c,e,c)}$. The generation method is as follows:

$$\left(n_i^{(c)}\right) - \left[r_{i,x}^{(c,e)}: \mathrm{E-mail}\right] \rightarrow \left(n_x^{(e)}\right) \leftarrow \left[r_{j,x}^{(c,e)}: \mathrm{E-mail}\right] - \left(n_j^{(c)}\right) \tag{1}$$

Generate the relation of tax arrears events related with the E-mail of the enterprise as $r_{i,y}^{(c,e,tr)}$: $\left(n_i^{(c)}\right) - \left[r_{i,y}^{(c,e,tr)}: \mathrm{E-mail\ related\ with\ tax\ arrears\ events}\right] \rightarrow \left(n_y^{(tr)}\right)$. The relation set is denoted by $R_d^{(c,e,tr)}$. The generation method is as follows:

$$\left(n_i^{(c)}\right) - \left[r_{i,x}^{(c,e)}\right] \rightarrow \left(n_x^{(e)}\right) \leftarrow \left[r_{j,x}^{(c,e)}\right] - \left(n_j^{(c)}\right) \leftarrow \left[r_{j,y}^{(c,tr)}: \mathrm{Tax\ arrears\ subject}\right] - \left(n_y^{(tr)}\right) \tag{2}$$

Generate the relation of trust-breaking events related with the E-mail of the enterprise as $r_{i,y}^{(c,e,cr)}$: $\left(n_i^{(c)}\right) - \left[r_{i,y}^{(c,e,cr)}: \mathrm{E-mail\ related\ with\ trust-breaking\ events}\right] \rightarrow \left(n_y^{(cr)}\right)$. The relation set is denoted by $R_d^{(c,e,cr)}$. The generation method is as follows:

$$\left(n_i^{(c)}\right) - \left[r_{i,x}^{(c,e)}\right] \rightarrow \left(n_x^{(e)}\right) \leftarrow \left[r_{j,x}^{(c,e)}\right] - \left(n_j^{(c)}\right) \leftarrow \left[r_{j,y}^{(c,cr)}: \mathrm{Trust-breaking\ subject}\right] - \left(n_y^{(cr)}\right) \tag{3}$$

Similarly, the telephone number relation between enterprises is generated as $r_{i,j}^{(c,p,c)}$; the tax arrears events relation of the enterprises linked by telephone number is generated as $r_{i,y}^{(c,p,tr)}$; the trust-breaking events relation of the enterprises linked by telephone number is generated as $r_{i,y}^{(c,p,cr)}$; the relation sets are denoted as $R_d^{(c,p,c)}$、$R_d^{(c,p,tr)}$、$R_d^{(c,p,cr)}$. The legal person relation between enterprises is generated as $r_{i,j}^{(c,lp,c)}$; the tax arrears events relation of the enterprises linked by the legal person is generated as $r_{i,y}^{(c,lp,tr)}$; the trust-breaking events relation of the enterprises linked by the legal person is generated as $r_{i,y}^{(c,lp,cr)}$; the relation sets are denoted as $R_d^{(c,lp,c)}$, $R_d^{(c,lp,tr)}$, $R_d^{(c,lp,cr)}$.

So far, nine types of new relations are generated in the knowledge graph, with a total of nearly 140 million relations. The quantity statistics are shown in Table 6:

Table 6. Statistics of the number of newly generated relationships.

Relation	Relation description	Number of relations
$R_d^{(c,e,c)}$	Enterprise - E-mail - Enterprise	57,684,392
$R_d^{(c,e,tr)}$	Enterprise - E-mail - Enterprise - Tax arrears events	53,44,188
$R_d^{(c,e,cr)}$	Enterprise - E-mail - Enterprise - Trust-breaking events	963,331
$R_d^{(c,p,c)}$	Enterprise - Phone - Enterprise	44,216,538
$R_d^{(c,p,tr)}$	Enterprise - Phone - Enterprise - Tax arrears events	4,441,975
$R_d^{(c,p,cr)}$	Enterprise - Phone – Enterprise- Trust-breaking events	768,567
$R_d^{(c,lp,c)}$	Enterprise - Legal person - Enterprise	23,840,712
$R_d^{(c,lp,tr)}$	Enterprise - Legal person - Enterprise - Tax arrears events	1,764,792
$R_d^{(c,lp,cr)}$	Enterprise - Legal person - Enterprise- Trust-breaking events	331,365
	Total	139,355,860

Based on the above relations, the total number of related enterprises, the total number of tax arrears events, the total number of trust-breaking events, the average number of tax arrears events and the average number of trust-breaking events can be calculated respectively. The actual data samples of class F3 variables are extracted from the calculation. The specific composition of each sample data is further determined. First, enterprises with a duration of at least 2 years are selected. For any qualified enterprise $n_i^{(c)}$, if number of p tax arrears occur, the values of F1, F2, F3 variables of the previous year corresponding to each event are taken as the features of a positive sample. Therefore, p times of tax arrears will form p positive examples of samples. For an enterprise that does not have tax arrears in a certain year, the values of F1, F2 and F3 variables corresponding to the enterprise in the previous year are taken as the characteristics of a negative example of a sample. That is, only one negative example of a sample is generated in the year without tax arrears event. Finally, 62,783 positive samples and 2,036,352 negative samples are generated, and the total number of samples is 2,099,135.

5. Model implementation and optimization

5.1. Model implementation

There are various sampling methods for machine learning in the area of fintech and taxtech [39]. In this paper, data samples are maldistributed, as the ratio of positive samples to negative samples hits 1:32. To improve the classification accuracy of an unbalanced dataset, improvement needs to be made on the data and algorithm fronts.

On the data front, resampling, which includes under-sampling and oversampling, is deployed. ENN [40] and Tomelink [41] typify under-sampling models, while SMOTE [42] and Borderling-SMOTE [43] are representative of oversampling models. The shortcomings of resampling are that 1) the distance between samples must be calculated to describe the distribution characteristics of the samples. As the number of samples increases, the corresponding data computation costs skyrocket. 2) Oversampling leads to many new samples, which can enhance the importance of minority-class

samples but at the same time increase computational costs.

On the algorithmic front, modifications are made to balance the unequal preferences for majority and minority classes, with the core idea being cost-sensitive learning. Fundamentally, sample imbalance can be summarized as sample gradient imbalance [44,45]. As the idea akin to cost-sensitive learning, LightGBM employs GOSS in sample processing and thus solves the gradient imbalance of samples, which is basically done by emphasizing samples with large gradients and duly abandoning samples with small gradients when calculating loss function gradients.

The Boolean hyperparameter $is\_unbalance$ is included in the LightGBM algorithm and is set as $is\_unbalance = true$ under sample imbalance. At this time, the model adds an extra $weight = (Number\ of\ negativecases)/(Number\ of\ positive\ cases)$ when calculating the gradient of minority-class samples to increase their importance among all samples.

In the process of implementing the machine learning model, the hyperparameters is set: $is\_unbalance = true$. The corresponding weights are added to the positive samples of the minority class. Different combinations such as F1 + F2, F3, F1 + F3, F2 + F3 and F1 + F2 + F3 are used as the characteristics of the samples to generate different models. The ROC curve and PR curve obtained are shown in Figures 4–8.

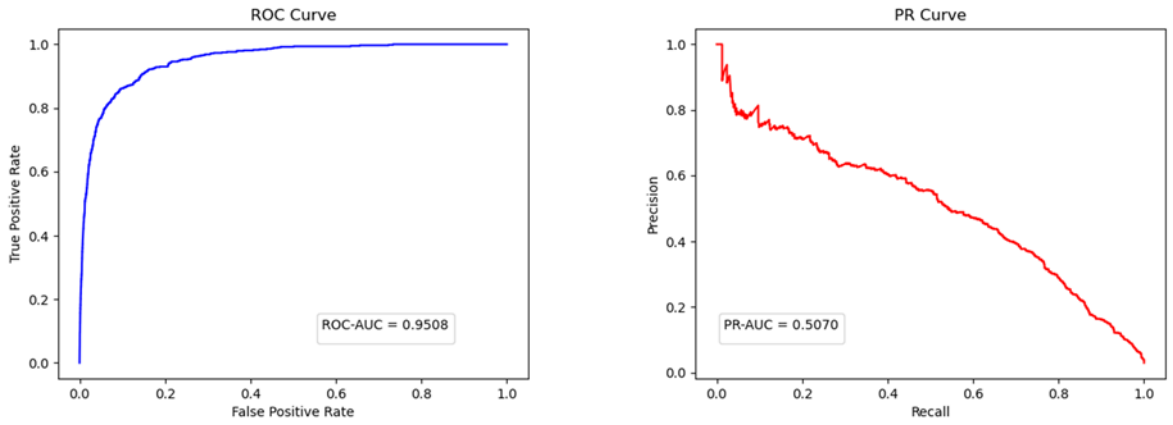

Figure 4. Model learning results with F1 + F2 as characteristics.

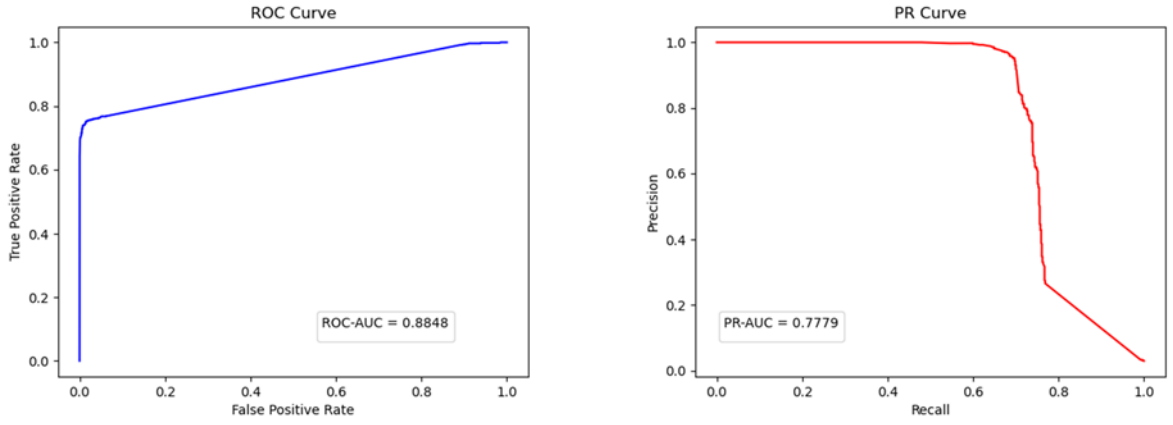

Figure 5. Model learning results with F3 as characteristic.

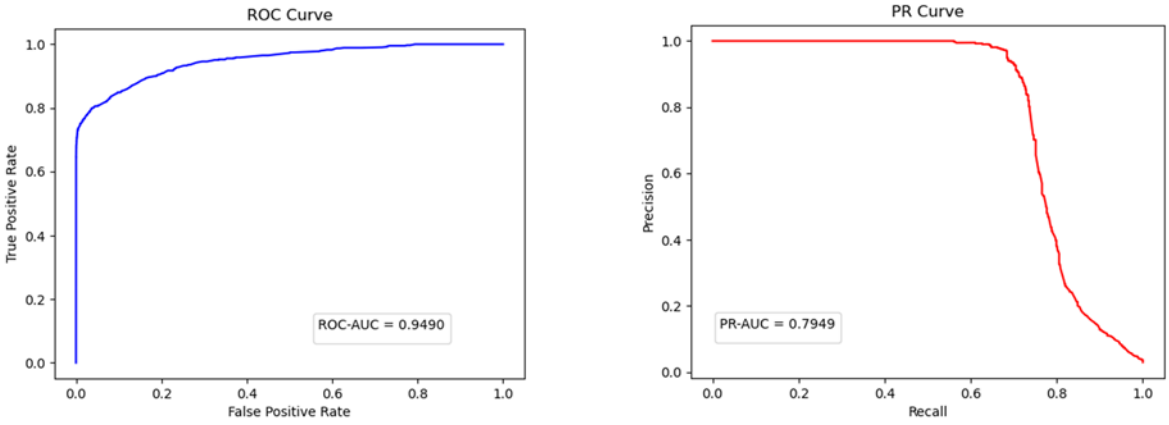

Figure 6. Model learning results with F1 + F3 as characteristics.

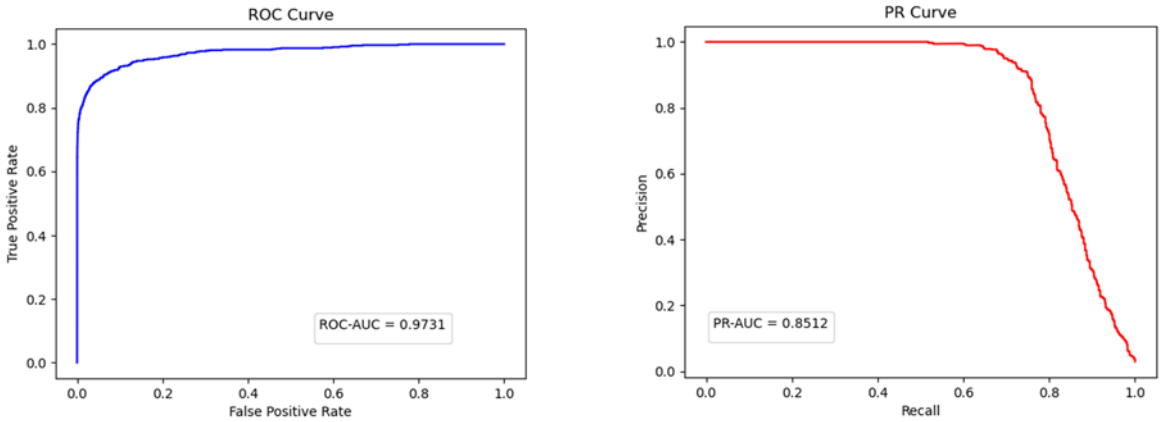

Figure 7. Model learning results with F2 + F3 as characteristics.

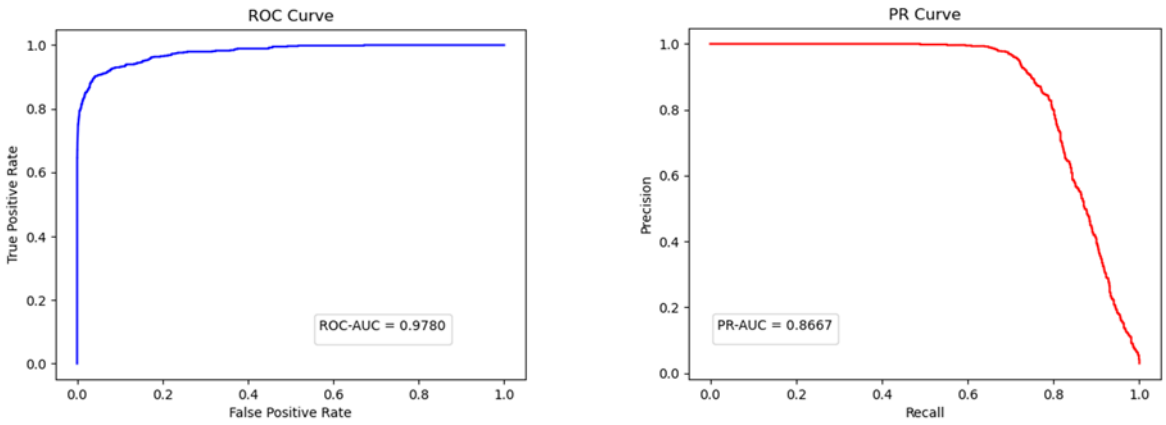

Figure 8. Model learning results with F1 + F2 + F3 as characteristics.

ROC-AUC and PR-AUC values corresponding to model learning results are shown in Table 7:

Table 7. Model learning results of different characteristics.

characteristics	ROC-AUC	PR-AUC
F1 + F2	0.9508	0.5070
F3	0.8848	0.7779
F1 + F3	0.9490	0.7949
F2 + F3	0.9731	0.8512
F1 + F2 + F3	0.9780	0.8667

The analysis of the model learning results is as follows:

1) When F1 and F2 are taken as sample characteristics, ROC-AUC reaches 0.9508. According to the aforementioned criteria, the overall classification ability is "outstanding". However, the corresponding PR-AUC is only 0.5070, which means that the prediction accuracy of tax arrears event is "poor".

2) When F3 is used as the feature alone, ROC-AUC is 0.8848. Overall classification ability evaluation is "excellent". At the same time, PR-AUC has been greatly improved, reaching 0.7779, and its forecasting ability for tax arrears is "acceptable".

3) When F1 and F3 are taken as sample characteristics, ROC-AUC is 0.9490. The overall classification performance is slightly lower than in case 1, but still "outstanding". PR-AUC is further improved to 0.7949, with the forecast for tax delinquencies approaching "excellent".

4) When F2 and F3 are taken as sample characteristics, ROC-AUC is 0.9731 and the overall classification ability is "outstanding". PR-AUC reaches 0.8512, and the prediction accuracy of tax arrears is "excellent".

5) When F1, F2 and F3 are used as sample features together, the ROC-AUC of the model is 0.9780 and the PR-AUC is 0.8667. Both the classification ability of the sample population and the forecasting ability of tax arrears are improved.

Since only 2.99% of the positive cases (tax arrears events) in the sample, even if all the samples predicted the negative cases, the prediction accuracy would be about 97%. As a result, no matter what combination of F1, F2 and F3 is adopted as sample characteristics, the obtained model has a good overall sample classification ability.

However, if we focus on the classification of positive examples, models based on F1 and F2 characteristics are not feasible. The model based on F3 features performs better. Even if F3 is taken as a feature alone, the model can obtain "acceptable" positive example classification capability. More recently, when F3 is combined with F1 or F2, the classification ability is gradually improved. When F1, F2 and F3 are used together as features, the optimum is achieved.

Based on the above analysis, it can be seen that the associations between enterprises based on the same telephone number, the same E-mail address and the same legal person are very important in predicting the tax arrears of enterprises. Trust-breaking events and tax arrears events, different tax arrears events are also related to each other.

5.2. Model optimization

In addition to the abovementioned hyperparameter $is\_unbalance$, in the event of sample imbalance, another hyperparameter of LightGBM, $scale\_pos\_weight$, can be used to adjust the

weight of minority-class samples.

The choice of $scale\_pos\_weight$ should maximize the model's attention to minority-class samples while avoiding the introduction of excessive small-gradient samples and noisy samples.

For the sample in this paper, when $is\_unbalance = true$ is set, the corresponding weight for the positive example is $weight = 32.43$. Therefore, in the specific tuning process, first set $is\_unbalance = false$, and then search for the $scale\_pos\_weight$ value that makes the positive example classification ability of the model better. Steps are as follows:

1) Other Settings remain unchanged. With 1 as the stepping, 130 PR-AUC values of the model are calculated when $scale\_pos\_weight$ from 1 to 130.00 (namely, four times the above weight value). The results are shown in Figure 9. Among all the PR-AUC results, the largest result is 0.8783 with $scale\_pos\_weight = 2$.

2) With $scale\_pos\_weight = 2$ as the interval center and 0.1 as the stepping, 20 PR-AUC values corresponding to $scale\_pos\_weight$ from 1.1 to 3.0 are calculated respectively. The maximum result is 0.8804 with $scale\_pos\_weight = 2.7$.

3) Further, with $scale\_pos\_weight = 2.7$ as the center of the interval and 0.01 as the stepping, 20 PR-AUC values corresponding to $scale\_pos\_weight$ from 2.61 to 2.80 are calculated respectively. The maximum result is 0.8860, where $scale\_pos\_weight = 2.77$.

After the completion of the above tuning process, the final tax arrears prediction machine learning model is obtained. The corresponding PR-AUC value of this model has reached 0.8860, which means "excellent" prediction ability.

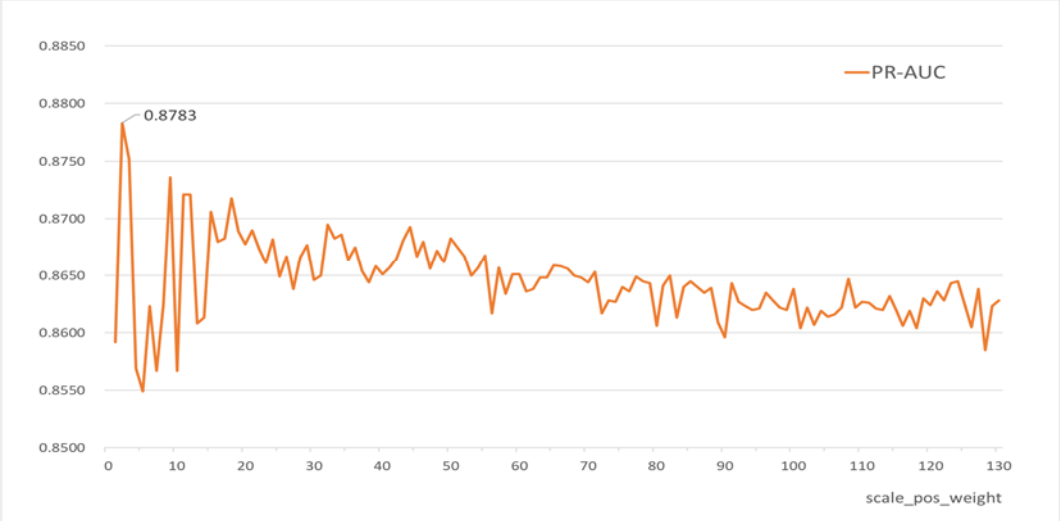

Figure 9. PR-AUC for different $scale\_pos\_weight$ values (1–130).

6. Conclusions

By constructing enterprise knowledge graph, this paper extracts the correlations between tax arrears events, trust-breaking events and enterprises. Taking them as important factors to predict tax arrears, a machine learning model for tax arrears prediction is established. The results and innovations are as follows:

1) It is common for enterprises to establish connections through the same telephone number, the same E-mail address and the same legal person. In this way, a large number of correlation relations can also be established between enterprises, tax arrears events and trust-breaking events. The results

show that there are 140 million links between nearly 1 million enterprises and 130,000 cases of tax arrears events and trust-breaking events. This paper does not use corporate financial data and transaction relations to establish corporate association network. The information, such as phone numbers, corporate legal person and E-mail addresses is easily obtained in real life. Based on the information, an enterprise association network with sufficient coverage can be established. The information extracted from it can fully support the prediction of tax arrears events. Compared with the existing tax arrears prediction research, the prediction model in this paper does not rely on the specific financial data of enterprises, avoiding the source obstacles of financial data. This model only needs to get data from the public to obtain good model predictive power. It has lower data acquisition cost and wider universality, especially for those small and medium-sized enterprises whose financial information is difficult to obtain. Through the analysis method of multi-source data fusion, it provides a new research framework for enterprise tax arrears forecast. It also has important practical significance for the tax regulation of enterprises.

2) The correlations between enterprises, tax arrears events and trust-breaking events extracted from the knowledge graph can be used as the prediction variable of the machine learning model for tax arrears prediction. The model has "acceptable" predictive power with the correlation relationships as the variable alone. After the further introduction of macroeconomic variables and the enterprise's own attributes, the forecasting ability reaches "excellent". Among the existing studies on tax arrears prediction, the prediction model of Siimon and Lukason [15] has the highest accuracy, with the evaluation index accuracy reaching 95.28%. However, Siimon and Lukason [15] also point out that accuracy is not the best model evaluation index in the case of imbalance sample prediction of tax arrears. This point also applies to this study. As previously discussed in this paper, since the proportion of positive cases (events of tax arrears) in samples is only 2.99%, the model will have a prediction accuracy of about 97% even if all samples are predicted as negative cases (non-events of tax arrears). This result is already higher than the model accuracy of Siimon and Lukason [15]. The accuracy index of the final model in this paper has reached 99.17%, but this situation is for the common prediction results of all positive and negative samples. In practice, more attention should be paid to the accuracy of the positive example prediction. Therefore, this paper does not use accuracy but uses PR-AUC as the evaluation index of the model. Based on the idea of cost-sensitive learning, the gradient weights of minority samples in the process of model learning are adjusted by selecting appropriate model hyperparameters. It can alleviate the influence of sample imbalance on the prediction ability of the model and improve the classification prediction ability of the model for positive cases. The resulting final model not only has a good prediction ability for the sample population, but also has a stronger pertinence for the tax arrears.

There are still some further improvements to be made in the follow-up research. In this paper, a common telephone number, E-mail address and legal person are used to establish the association network between enterprises. Its shortcomings are as follows: on one hand, the specific connotation of the correlations between enterprises is not clear enough. For example, if there are two enterprises with the same contact number, it may mean that they have the same core employees, or it may mean that both enterprises purchase the same outsourcing services, such as bookkeeping services. In this case, there is likely to be a mediating effect discussed by Li [46] in our prediction model. It is difficult to accurately describe the correlation transmission mechanism of risk events through the correlation network. This is one of the key points to focus on in the follow-up research. On the other hand, the relationships between enterprises are not limited to phone numbers, E-mail addresses and legal persons. There are other linkages such as enterprise locations, name similarity and so on. These links may also affect the transmission of tax arrears between enterprises. Therefore, on the basis of existing studies,

how to add other forms of association relations to form a more comprehensive enterprise association network is also an important direction to be improved in the future.

Acknowledgments

The authors are very grateful to the editors and reviewers for all the work they have done on this paper. They have spent a lot of time and effort in improving this paper, which has greatly enhanced the comprehensiveness of this paper.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. Krut, X. Peng, Does corporate social performance lead to better financial performance? Evidence from Turkey, *Green Finance*, **3** (2021), 464–482. https://doi.org/10.3934/gf.2021021
2. D. Marghescu, M. Kallio, B. Back, Using financial ratios to select companies for tax auditing: a preliminary study, in *Communications in Computer and Information Science*. Springer, Berlin, 2010. https://doi.org/10.1007/978-3-642-16324-1_45
3. A. Su, Z. He, J. Su, Y. Zhou, Y. Fan, Y. Kong, Detection of tax arrears based on ensemble learning model, in *Proceedings of the 2018 International Conference on Wavelet Analysis and Pattern Recognition*, Piscataway, NJ, (2018), 270–274. https://doi.org/10.1109/icwapr.2018.8521362
4. A. Ippolito, A. C. G. Lozano, Sammon mapping-based gradient boosted trees for tax crime prediction in the city of São Paulo, in *Enterprise Information Systems, ICEIS 2020*, (2020), 293–316. https://doi.org/10.1007/978-3-030-75418-1_14
5. J. Vanhoeyveld, D. Martens, B. Peeters, Value-added tax fraud detection with scalable anomaly detection techniques, *Appl. Soft. Comput.*, **86** (2020), 1–38. https://doi.org/10.1016/j.asoc.2019.105895
6. M. Z. Abedin, G. Chi, M. M. Uddin, M. S. Satu, M. I. Khan, P. Hajek, Tax default prediction using feature transformation-based machine learning, *IEEE Access*, **9** (2021), 19864–19881. https://doi.org/10.1109/access.2020.3048018
7. E. I. Altman, M. Balzano, A. Giannozzi, S. Srhoj, Revisiting SME default predictors: The Omega Score, *J. Small Bus. Manage.*, **2022** (2022), 1–35. https://doi.org/10.1080/00472778.2022.2135718
8. O. Lukason, A. Andresson, Tax arrears versus financial ratios in bankruptcy prediction, *J. Risk Financ. Manag.*, **12** (2019), 187–200. https://doi.org/10.3390/jrfm12040187
9. S. Chen, J. Zhong, P. Failler, Does China transmit financial cycle spillover effects to the G7 countries, *Econ. Res. -Ekon. Istraz.*, **35** (2022), 5184-5201. https://doi.org/10.1080/1331677X.2021.2025123
10. F. Misra, R. Kurniawan, The role of audit information dissemination in curbing the contagion of tax noncompliance, *J. Innov. Bus. Econ.*, **4** (2020). 1–11. https://doi.org/10.22219/jibe.v4i01.10223

11. Z. Li, J. Zhu, J. He, The effects of digital financial inclusion on innovation and entrepreneurship: A network perspective, *Electron. Res. Arch.*, **30** (2022), 4697–4715. https://doi.org/10.3934/era.2022238
12. G. Kou, Y. Xu, Y. Peng, F. Shen, Y. Chen, K. Chang, et al., Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection, *Decis. Support Syst*., **140** (2021), 113429. https://doi.org/10.1016/j.dss.2020.113429
13. P. Giudici, B. H. Misheva, A. Spelta, Network based credit risk models, *Qual. Eng*., **32** (2020), 199–211. https://doi.org/10.1080/08982112.2019.1655159
14. K. Peng, G. Yan, A survey on deep learning for financial risk prediction, *Quant. Finance. Econ.*, **5** (2021), 716–737. https://doi.org/10.3934/qfe.2021032
15. Õ. R. Siimon, O. Lukason, A decision support system for corporate tax arrears prediction, *Sustainability*, **13** (2021), 8363. https://doi.org/10.3390/su13158363
16. V. Chaudhri, C. Baru, N. Chittar, X. Dong, M. Genesereth, J. Hendler, Knowledge graphs: introduction, history and, perspectives, *AI Mag.*, **43** (2022), 17–29. https://doi.org/10.1609/aimag.v43i1.19119
17. R. Angles, C. Gutierrez, Survey of graph database models, *ACM Comput. Surv.*, **40** (2008), 1–39. https://doi.org/10.1145/1322432.1322433
18. N. Ahbali, X. Liu, A. Nanda, J. Stark, A. Talukder, R. P. Khandpur, Identifying corporate credit risk sentiments from financial news, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, (2022), 362–370. http://dx.doi.org/10.18653/v1/2022.naacl-industry.40
19. Z. Li, L. Chen, H. Dong, What are bitcoin market reactions to its-related events, *Int. Rev. Econ. Finance*, **73** (2021), 1–10. https://doi.org/10.1016/j.iref.2020.12.020
20. T. Ruan, L. Xue, H. Wang, F. Hu, L. Zhao, J. Ding, Building and exploring an enterprise knowledge graph for investment analysis, in *International Semantic Web Conference 2016*, (2016), 418–436. https://doi.org/10.1007/978-3-319-46547-0_35
21. X. Chang, The impact of corporate tax outcomes on forced CEO turnover, *Natl. Account. Rev.*, **4** (2022), 218–236. https://doi.org/10.3934/nar.2022013
22. A. Sousa, A. Braga, J. Cunha, Impact of macroeconomic indicators on bankruptcy prediction models: Case of the Portuguese construction sector, *Quant. Finance. Econ.*, **6** (2022), 405–432. https://doi.org/10.3934/qfe.2022018
23. Z. Li, Z. Huang, Y. Su, New media environment, environmental regulation and corporate green technology innovation: Evidence from China, *Energy Econ.*, **119** (2023), 106545. https://doi.org/10.1016/j.eneco.2023.106545
24. Y. Liu, Z. Li, M. Xu, The influential factors of financial cycle spillover: evidence from China, *Emerg. Mark. Finance Trade*, **56** (2020), 1336–1350. https://doi.org/10.1080/1540496x.2019.1658076
25. G. Aytkhozhina, A. Miller, State tax control strategies: Theoretical aspects, *Contaduría y Administración*, **63** (2018), 25. https://doi.org/10.22201/fca.24488410e.2018.1672
26. Z. Li, B. Mo, H. Nie, Time and frequency dynamic connectedness between cryptocurrencies and financial assets in China, *Int. Rev. Econ. Finance*, **86** (2023), 46–57. https://doi.org/10.1016/j.iref.2023.01.015

27. Z. Li, H. Dong, C. Floros, A. Charemis, P. Failler, Re-examining bitcoin volatility: a CAViaR-based approach, *Emerg. Mark. Finance Trade*, **58** (2022), 1320–1338. https://doi.org/10.1080/1540496x.2021.1873127
28. A. Chang, L. Yang, R. Tsaih, S. Lin, Machine learning and artificial neural networks to construct P2P lending credit-scoring model: A case using Lending Club data, *Quant. Finance Econ*., **6** (2022), 303–325. https://doi.org/10.3934/qfe.2022013
29. D. Wang, L. Li, D. Zhao, Corporate finance risk prediction based on LightGBM, *Inf. Sci.*, **602** (2022), 259–268. https://doi.org/10.1016/j.ins.2022.04.058
30. B. Gao, V. Balyan, Construction of a financial default risk prediction model based on the LightGBM algorithm, *J. Intell. Syst.*, **31** (2022), 767–779. https://doi.org/10.1515/jisys-2022-0036
31. L. Zhang, Q. Song, Multimodel integrated enterprise credit evaluation method based on attention mechanism, *Comput. Intell. Neurosci.*, **2022** (2022), 1–12. https://doi.org/10.1155/2022/8612759
32. J. G. Ponsam, S.V. J. B. Gracia, G. Geetha, S. Karpaselvi, K. Nimala, Credit risk analysis using LightGBM and a comparative study of popular algorithms, in *International Conference on Computing and Communications Technologies (ICCCT)*, 2021. https://doi.org/10.1109/iccct53315.2021.9711896
33. D. G. Kirikos, An evaluation of quantitative easing effectiveness based on out-of-sample forecasts, *Natl. Account. Rev.*, **4** (2022), 378–389. https://doi.org/10.3934/nar.2022021
34. F. Corradin, M. Billio, R. Casarin, Forecasting economic indicators with robust factor models, *Natl. Account. Rev.*, **4** (2022), 167–190. https://doi.org/10.3934/nar.2022010
35. P. Harrington, *Machine Learning in Action*, Manning Publications, (2012), 143–149.
36. J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, in *William C. ICML '06: Proceedings of the 23rd international conference on Machine learning*, (2006), 233–240. https://doi.org/10.1145/1143844.1143874
37. T. Fawcett, An introduction to ROC analysis, *Pattern Recognit. Lett.*, **27** (2006), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
38. W. H. J. David, S. Lemeshow, R. X. Sturdivant, *Applied Logistic Regression*, 3 edition, John Wiley & Sons, (2013), 177–178. https://doi.org/10.1002/9781118548387
39. Z. Li, C. Yang, Z. Huang, How does the fintech sector react to signals from central bank digital currencies, *Finance Res. Lett*., **50** (2022), 103308. https://doi.org/10.1016/j.frl.2022.103308
40. D. L. Wilsin, Asymptotic properties of nearest neighbor rules using edited data, *IEEE Trans. Syst. Man Cybern.*, **3** (1972), 408–421. https://doi.org/10.1109/tsmc.1972.4309137
41. I. Tomek, Two modifications of CNN, *IEEE Trans. Syst. Man Cybern.*, **6** (1976), 769–772. https://doi.org/10.1109/tsmc.1976.4309452
42. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, **16** (2002), 321–357. https://doi.org/10.1613/jair.953
43. H. Han, W. Y. Wang, B. H. Mao, Borderline-smote: a new over-sampling method in imbalanced data sets learning, in *International Conference on Intelligent Computing*, (2005), 878–887. https://doi.org/10.1007/11538059_91
44. B. Y. Li, Y. Liu, X. G. Wang, Gradient harmonized single-stage detector, in *The 33rd AAAI Conference on Artificial Intelligence*, (2019), 8577–8584. https://doi.org/10.1609/aaai.v33i01.33018577

45. T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. https://doi.org/10.1109/iccv.2017.324
46. T. Li, J. Wen, D. Zeng, K. Liu, Has enterprise digital transformation improved the efficiency of enterprise technological innovation? A case study on Chinese listed companies, *Math. Biosci. Eng.*, **19** (2022), 12632–12654. https://doi.org/10.3934/mbe.2022590

AIMS Press