Financial Engineering and Risk Management (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2024.070524
ISSN 2523-2576 Vol. 7 Num. 5

Research on Identification of Taxpayers' Fraudulent Invoicing Behavior Based on Feature Engineering

Wei Liu[1,a], Jiyuan Chen[1,b], Lingjun Xiao[1,c], Yin Si[1,d], Jun Tang[1,e,*]

[1]School of Information and Intelligent Engineering, Guangzhou Xinhua University, Guangzhou, China
[a]6444617704@qq.com, [b]2436898250@qq.com, [c]xlj1031400301@163.com, [d]sy987110@163.com, [e]9399576@qq.com
[*]Corresponding author

Keywords: False invoicing; Feature engineering; Integrated learning; Stacking

Abstract: Fraudulent invoicing is a key part of tax risk work, and how to accurately identify whether taxpayers have fraudulent invoicing behaviors from massive tax data to reduce the loss of tax is the focus of tax risk work. The existing tax data is large in volume, with fuzzy data features, and traditional machine learning models have limited generalization ability, which leads to poor performance in identifying false invoicing behaviors. To address the above problems, this paper establishes a high-quality sample dataset by establishing a tax feature project and proposes a learning model based on Stacking integrated ideas to identify taxpayers' false invoicing behavior. Taking the Stacking-based false invoicing behavior recognition model proposed in this paper as the core, on the tax sample dataset, the classification effect of the proposed model is compared with that of single models, and the results show that the Stacking-based recognition model is superior to others in terms of AUC value, accuracy, and F1 score. The experimental results validate the superiority of the model.

1. Introduction

Since China's tax reform in 1994, the identification of taxpayers' fraudulent VAT invoicing behavior has been a major problem plaguing China's tax risk management department.[1]By using false VAT invoices, taxpayers can obtain illegal deductions, creating significant challenges for tax enforcement officials. Therefore, it is necessary to accurately and comprehensively identify the behavior of taxpayers with false VAT invoices through the existing technical methods and their expertise.

Machine learning provides new technical means for the identification of false invoicing behavior. In this paper, we analyze tax data through feature engineering, algorithmic modeling and other machine learning methods to study the features and models that can accurately identify the behavior of false invoicing.We establish tax feature engineering to screen out features highly relevant to fraudulent invoicing behavior. A stacking integration-based identification model is then proposed[2], which can accurately and reliably detect fraudulent invoicing behavior. This model provides strong technical support for tax departments to identify taxpayers involved in such behavior and has significant practical application value.

2. Related work

The research related to the recognition of false invoicing invoice behavior is as follows: JI Y L, Wang W Q proposed establishing a false VAT invoice recognition model through random forest algorithm, combining artificial modeling with data analysis, and analyzed the model recognition accuracy, robustness and reliability through the experimental validation of the data.[3]The model identification is accurate, robust and reliable. Zhu J T proposes that the loopholes in the tax department's supervision of taxpayers' false invoicing come from the lack of information sharing among relevant law enforcement departments, and that by establishing a comprehensive system for taxpayers' business chains and a cross-departmental information-sharing platform, breaking down barriers between law enforcement agencies. we can identify whether taxpayers engage in false invoicing through the monitoring of taxpayers' invoicing capital chain.[4]The information sharing platform has been set up to realize the information sharing. Based on the characteristics of tax business, Chen Z Sand Zhang J P proposed the indicator of "accumulated unaccounted tax source" to measure the risk of false invoicing.[5]

On the basis of the above research, this paper proposes a virtual invoice behavior recognition model based on Stacking integration to address the problem of poor recognition of virtual invoice behavior by a single model. This paper uses the Stacking integration learning approach to construct a two-layer fusion model for identifying fraudulent invoicing behavior, after analyzing the classification effects and principles of different types of base models, LightGBM and other models are selected to form the first layer of the model.The second layer, which is a logistic regression classifier, corrects the biases in the predictions from the first layer to obtain the final classification results after fully learning from the predicted sample data.

3. Dataset Construction and Experimentation

The daily tax business data in the tax system are large in volume, and the business relationship between the data is complex and characterized by vague features. To address these problems, this chapter extracts features through manual empirical analysis of the characteristics of false invoicing behaviors manifested in the tax business after combing the tax business data. By applying the feature engineering method to screen and preprocess the features, a high-quality tax sample dataset is finally obtained, which lays the data foundation for the subsequent experimental research. The data extraction preprocessing process in this paper is shown in Figure 1.

3.1. Data extraction

The samples of taxpayers with good credit (white sample) and taxpayers with false invoices (black sample) are screened according to the tax business rules in the tax database of Province A. Finally, a total of 40,053 taxpayers from the black and white lists are selected from the database, among which 4,403 are taxpayers with false invoices (black sample), and 35,650 are taxpayers with excellent tax credit (white sample). This dataset is unbalanced, affected by practical factors, and the ratio of black to white samples is about 9:1.

The process of extracting tax features is based on the idea of constructing the business characteristics of fraudulent invoicing behaviors manifested under the business domains of tax registration, invoicing, collection, approval, declaration, and preference, and setting the relevant thresholds according to the relevant experience of tax personnel in the tax field, and extracting the tax data from the original database. In this paper, a total of 94 features of nine major types related to fraudulent invoice taxpayers are extracted, which are labeled as TZ01 to TZ94.

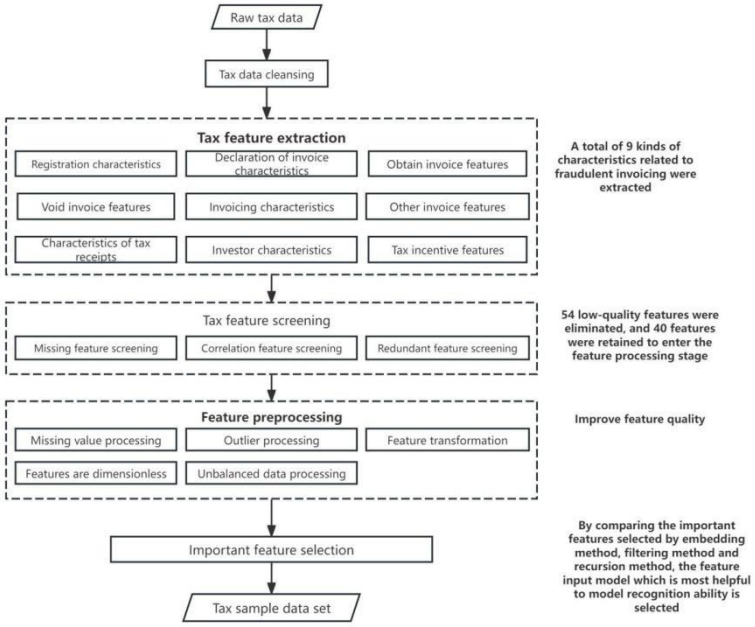


Figure 1: Data extraction preprocessing flow.

3.2. Feature engineering

3.2.1. Feature Screening

The more features into the model is not the better, too high dimensional features are easy to cause model overfitting, affecting the generalization ability of the model, increasing the cost of model computing, and the prediction results are often greatly disappointed. In this paper, after exploring and analyzing the missing features, related features and redundant features, we screened the features with a missing proportion higher than 10% of the features, excluded the features whose absolute value of the Pearson's correlation coefficient between them and the labels was lower than 0.05, and finally excluded a small number of highly repetitive covariate features by using the manual selection method. After feature screening, a total of 54 low-quality, low-correlation and redundant features were removed.

3.2.2. Feature preprocessing

After feature screening, the missing values, outlier points in the features and the amount between different features just will have an impact on the fitting ability of the model. In this section, this paper analyzes and processes the missing values and outliers in the tax dataset.The methods for filling the missing values of the feature items are mainly:1)filling the missing values according to the statistical laws of the features, such as the median, the multitude, and the mean of the features to fill the missing values. 2) custom value filling.

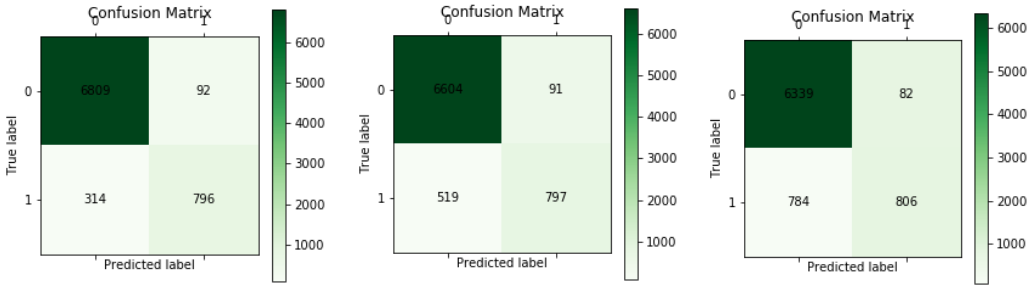
Then the correlation function is applied to transform and dimensionless processing of features to improve the quality of data. The commonly used methods for dimensionless processing of feature data are normalization and standardization. This paper aims to retain the meaning and distribution information of the data while processing the features in a dimensionless way, therefore, normalization is mainly used for this purpose.

Finally, the sampling strategy of the unbalanced dataset is discussed, and the sample dataset with higher data quality is generated. Due to the influence of objective factors, the sample data of taxpayers

with false invoices and taxpayers with good credit in the tax dataset are unbalanced.This sample imbalance is likely to cause the model to be more inclined to label the samples in the test set as the majority of the classes, resulting in an inflated accuracy rate. Taken together, in order to increase the sample information of taxpayers with false invoices, this paper processes the dataset by using oversampling in the data layer and applies integrated learning algorithms in the algorithm layer to correct the classification surface.The cross-validation method is then used to verify the effectiveness of the model during the modeling process.

3.2.3. Selection of important features

The result of feature selection directly affects the model classification effect[6].In this paper, after the extraction of tax features, the tax features with low quality, weak relevance and redundancy are filtered out, and finally the tax features are preprocessed, so that a total of forty features with higher quality are obtained. In order to pursue the modeling efficiency and better modeling effect, this paper filters the more valuable features into the model according to the importance of the features, which improves the model training efficiency while ensuring the model classification quality.



(a) Characteristics of filtering screening (b) Characteristics of Embedding Method Screening
(c) Characterization of screening by wrapping method

Figure 2: The confusion matrix for random forest prediction results.

In this section, three feature selection methods, namely, filtering (chi-square test), embedding (XGBoost), and wrapping (RFE), are used to screen and analyze the important features of tax data. The top ten filtered important features are input into the model, and the confusion matrix obtained by taking the random forest model as an example is as Figure 2.

From the experimental confusion matrix, it can be seen that the model trained by the features screened through the filtering method has the strongest ability to identify the behavior of false invoicing of taxpayers, and the number of correctly classified samples is the largest. From the experimental results, it can be seen that the features screened by the filtering method have the greatest effect on model identification of false invoicing behavior.Therefore, the features screened by the filtering method are selected to train the model in the subsequent experiments.

3.3. Model building

3.3.1. Model structure

In recent years, integrated learning has received high attention in the field of machine learning due to its outstanding performance in various prediction problems[7]. By combining different learners, the integration learning algorithm makes the integrated model have higher generalization ability and better classification ability, which can avoid the model from falling into the local minimum dilemma.Additionally, the coordination between models enriches the sample hypothesis space of the

fusion model.

In this paper, we apply the idea of integrated learning to propose a virtual invoice recognition model based on Stacking integration, and the model structure is shown in Figure 3.

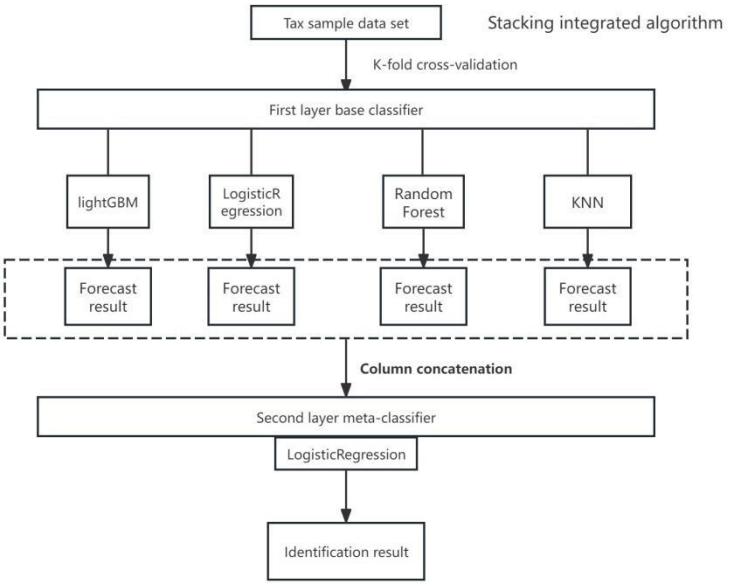


Figure 3: The stacking-based false invoice recognition model.

The first layer of the integrated model consists of multiple heterogeneous classifier models[8-11],and the tax sample dataset is input into the first layer model, which is trained using five-fold cross-validation to improve the model prediction accuracy. The output results of the lower layer model are used as the input of the second layer meta-model, and the logistic regression meta-model is trained to output the identification results of false invoicing behavior.

By applying the integration idea of "better but different" in integrated learning, the heterogeneous model with high prediction accuracy is selected as the base model for fusion, so that the fusion model is characterized by diversity in structure, which reduces the risk of insufficient model assumption space and falling into local minima. Through the coordination of the two-layer model, the meta-model can correct the deviation of the base model predictions, and ultimately achieve a better classification effect.

3.3.2. Integrating the model training process

In the process of base learner training, if the training set is used directly for training, it is easy to cause overlearning of the model, so that the model prediction does not have the expected effect. Therefore, in this paper, we use five-fold cross-validation to train the base learner to improve the prediction accuracy. In this paper, the five-fold cross-validation process involves first dividing the tax sample data training set into five sub-datasets during the training process. Let the number of samples in the training set be 5N;then the number of samples in each sub-data is N. In each cross-training, four copies of the five sub-sets are selected as the model training set, and one copy is selected as the model test set, so that a total of five combinations of training and test sets can be obtained. The five combinations of training and test sets are traversed for iterative training by the base learner. Use the training set trained model on the test set (with N samples) to predict, get the prediction data (the number of N), five iterations of the model prediction data by row splicing, to get the base model of the five-fold cross-validation of the output of the output (5N rows and 1 column), each base learning output results by column splicing, as a new training samples of the data. In the process of iterative

training of the model, at the same time, the tax sample data test set input model for prediction, the prediction results by rows summed to take the average, and then with other model prediction results by column splicing, to get with the new test sample data. Take the base learner 1 as an example, the 50% discount cross-validation training process for base learner 1 is shown in Figure 4.

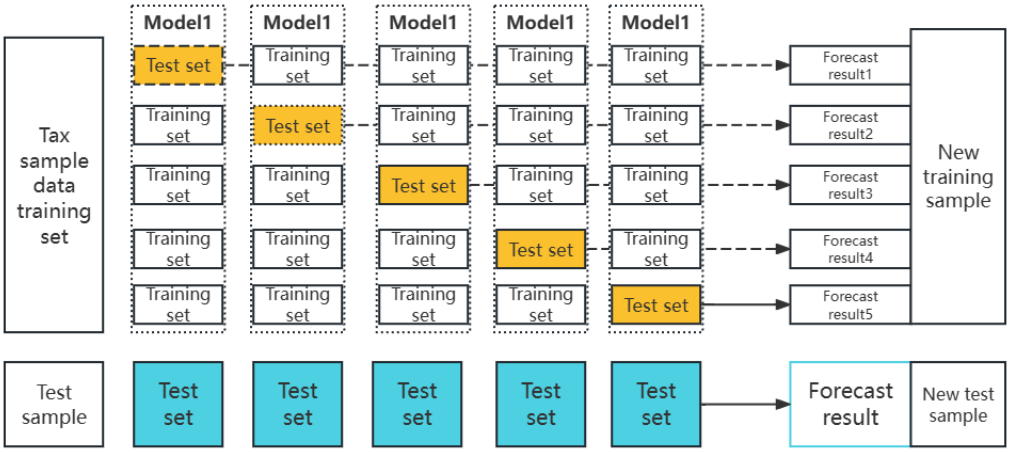


Figure 4: 5-Fold CV for Base Learners.

Taking the KNN model in the proposed model in this paper as an example, the KNN model five-fold cross-validation training process is as follows:

(1) Split the tax sample dataset D into a training set D_{train} and a test set D_{test} ;

(2) Divide the training set D_{train} into five equal sub-datasets, $D_{train} = \{D_1, D_2, D_3, D_4, D_5\}$;

(3) The KNN model is cross-trained five times, let the number of iterations is k, then each time, respectively, using D_k as the test dataset of the KNN model, the rest of the four copies of the sub-data as the training dataset of the KNN model, the kth iteration of the training to hit the prediction results for p_k , repeated iterations of the five times of the training to get the prediction results of $P = \{p_1, p_2, p_3, p_4, p_5\}$ for the new training set segments, the number of samples in P and the number of samples in D_{train} remain the same;

(4) Meanwhile in each cross-training, the trained KNN model predicts the test set D_{test} to get the prediction results as t_k , and the new test set fragment T is averaged as the sum of each 5 prediction results;

(5) Repeat steps (2)(3) for the remaining three models to obtain new predicted training set fragments and new test set fragments, and splice the training set fragments and test set fragments obtained from the prediction of each model by columns to obtain new training set features and test set features.

At the end of the base model training to get a new training set and test set, the new training set features and test set will be input into the second layer logistic regression model for prediction, and finally the meta-learner outputs the result of false invoicing behavior recognition.

3.4. Experiments and Results

The features screened out by feature engineering are inputted into the traditional model for training respectively and are compared with those in the false invoicing behavior recognition model based on

Stacking integration proposed in this paper, and the scores of the model in terms of AUC value, accuracy, precision, recall and F1 value are shown in Table 1.

As can be seen from the table, by comparing the classification effect of identifying taxpayers with fraudulent invoices on the tax sample dataset between the base model and the Stacking-based integrated fraudulent invoice behavior recognition model proposed in this paper, the integrated model proposed in this paper performs optimally in terms of the AUC score, the accuracy rate, and the F1 value, which reach 93.58%, 97.17%, and 87.43%, respectively.Furthermore,the AUC value of the integrated model proposed in this paper increases by 0.9 percentage points compared with the highest value of the single model. The AUC value of the model is 0.9 percentage points higher than the highest value of the single model, indicating that the model has a good ability to identify taxpayers with false invoices and that the model effect is stable. However, the performance of the integrated model is slightly inferior in terms of precision rate and recall rate.

Table 1: The stacking integration model vs. single model comparison experiments.

Model	AUC	Accuracy	Precision	Recall rate	F1 Value
Stacking Integration Model	**0.9358**	**0.9717**	0.8596	0.8896	**0.8743**
LightGBM	0.9243	0.9390	0.6650	**0.9054**	0.7668
RandomForest	0.9262	0.9493	0.7171	0.8964	0.7968
KNN	0.9159	0.9679	**0.8597**	0.8491	0.8544
LogisticRegression	0.8972	0.8916	0.5063	0.9043	0.6492

4. Conclusion

This paper mainly focuses on the research on the identification of false invoicing behavior of taxpayers. Firstly, we extract features by analyzing the characteristics of fraudulent invoicing behavior, and use feature engineering methods to process tax data to obtain high-quality tax datasets. Then, by combining with the idea of Stacking integration algorithm, a two-layer false invoice recognition model is constructed. Finally, we compare the prediction results of different models on the tax sample dataset to verify the superiority of the model.

Acknowledgements

This work was supported by the 2023 Guangdong Undergraduate University Teaching Quality and Teaching Reform Project with No. 2023CYXY003-2; Guangzhou Xinhua University YueQian Information Technology Industry College Project with No. 2023CYXY003.

References

[1] Li Xiangrong, Zhu Keshi. Analysis of Tax Risk Prevention Countermeasures of VAT Invoice Management[J]. China Accountant General, 2020(09):52-54.

[2] Wolpert D H. Stacked generalization [J]. Neural Networks, 2017, 5(2):241-259.

[3] Ji Yanli, Wang Wenqing. Research on the stock of accuracy of tax risk identification in the context of big data - based on the perspective of machine learning[J]. Fiscal Research, 2020(09):119-129.DOI:10.19477/j.cnki.11-1077/f.2020. 09. 010

[4] Zhu Jiangtao. Utilizing "Internet+" thinking to crack the problem of export tax fraud[J]. Tax Research, 2016(05):22-27.DOI:10.19376/j.cnki.cn11-1011/f.2016.05.003.

[5] Chen Zaosheng, Zhang Junping. VAT tax source risk control model and empirical analysis[J]. Taxation Economics Research, 2015, 20(02):66-71.DOI:10.16340/j.cnki.ssjjyj.2015.02.011.

[6] Yao X, Wang Xiaodan, Zhang Yuxi, Quan Wen. A review of feature selection methods[J]. Control and Decision

Making, 2012, 27(02):161-166+192.DOI:10.13195/j.cd.2012.02.4.yaox.013.
[7] J. W. Xu, Y.Y. Yang. Integrated learning methods:A research review[J]. Journal of Yunnan University(Natural Science Edition), 2018, 40(06):1082-1092.
[8] Hart P E. The Condensed Nearest Neighbor Rule[J]. IEEE Transactions on Information Theory, 1968, 14(3):515-516.
[9] Pregibon D. Logistic Regression Diagnostics[J]. Annals of Statistics, 1981, 9(4):705-724.
[10] Breiman L. Random Forests [J]. Machine Learning, 2001.
[11] Ke G L, Meng Q, Finley T, et al. Light GBM: A Highly Efficient Gradient Boosting Decision Tree[C]// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA.