Fuzzy Systems and Data Mining VIII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220371

Forecasting Tax Risk by Machine Learning: Case of Firms in Ho Chi Minh City

Nguyen Anh Phong[a, b,1], Phan Huy Tam[a, b] and Le Quoc Cuong[a,b]
[a] *University of Economics and Law, Vietnam*
[b] *Vietnam National University, Ho Chi Minh City, Vietnam*

Abstract. Tax is the main source of income for the State. However, managing tax collection effectively and limiting the tax risks is a challenge for state tax authorities. This study applies machine learning to assess and predict firms with tax risks using logistic regression algorithm. The data set includes 872 observations of firms in Vietnam market. The machine learning approach is used to classifies the firms into 2 categories which has tax risk or not based on 6 main factors: (i) revenue and other income; (ii) expenses; (iii) liquidity; (iv) asset; (v) liabilities; and (vi) equity. The results show that the machine learning method is effective and accurate in identifying and predicting risks in tax declaration. The authors recommend that the tax agencies could apply machine learning methods and go further with big data and artificial intelligence approach to identify and classify enterprises.

Keywords. tax risk, machine learning, enterprises

1. Introduction

Tax is the main source of state revenue and a major contributor to the national budget to ensure the expenditure of the State, an important tool for redistributing gross social product and national income [1]. Thus, it could be seen that tax is an economic measure of every state. Because of that importance, the tax system reform strategy for the period 2011-2020 was approved by the Government, along with the completion and development of new laws on tax policy, the Law on Tax Administration was promulgated, amending and supplementing to meet the requirements of socio-economic development and international economic integration, establishing a common legal framework and uniformly applying it in the process of implementing all tax policies, overcome the situation of separation in management methods among taxes, creating a foundation for the self-declaration and self-payment mechanism.

However, due to the rapid increase in workload and the number of taxpayers, the self-declaration management mechanism is facing many risks such as the risk of

[1] Corresponding Author, Nguyen Anh Phong, Faculty of Finance and Banking, University of Economics and Law, VietNam National University, Ho Chi Minh City, Vietnam. Email: phongna@uel.edu.vn
JEL Classification Code: C01, C81, G38, H11
This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2022-34-03

intentionally or unintentionally declaring dishonestly the amount of tax payable, actual tax paid less than the declared amount. Besides, many business owners are lack of tax law knowledge, not voluntarily fulfilling their tax obligations, taking advantage of the above mechanism for tax evasion, tax fraud, increase tax debt or deliberately finding all tricks and forms of cheating on payable tax amounts such as making false declarations to appropriate VAT refunds from the state budget, increasingly sophisticated.

This is a challenge for state tax authorities in managing tax collection effectively from these enterprises. In which, limiting the risk of tax declaration is always an important factor that attract great attention by tax authorities. Assessing tax declaration risks of businesses and develop an appropriate and effective strategy in the practice of tax administration and management is the core task for state tax agencies. With a large number of firms, as well as the increasing complexity in tax behaviors of these firms, applying machine learning technology to detect tax risks timely and accurately for management agencies to make decisions is absolutely necessary.

2. Empirical Studies

Andreoni et al [2] argue that tax compliance should be defined as the willingness of taxpayers, the observance of tax laws to achieve equilibrium in the economy of a country. Song and Yarbrough [3] argue that the operation of the US tax system is mainly based on voluntary self-assessment and compliance, possibly due to the view that tax compliance should be determined as the capability of the taxpayer and willingness to comply with tax law is determined by ethics, regulatory environment and other situational factors at a particular time and place.

Tax compliance is also determined by some tax authorities such as the ability and willingness of taxpayers to comply with tax laws, report income accurately each year, and pay taxes on time (IRS, 2009; ATO, 2009 and IRB, 2006). The broader view of tax compliance requires a degree of honesty, adequate tax knowledge and the ability to use this knowledge in a timely and accurate manner, to be able to prepare tax refund dossiers and related documents [4]. McBarnet [5] suggests that tax compliance should be viewed in three ways: (i) “absolute compliance”: taxpayers voluntarily pay taxes without complaint; (ii) “conditional compliance”: reluctance to pay taxes; (iii) “creative compliance”: demonstrate by the ability to apply provisions of tax law to redefine taxable income and expenses.

Studies using probabilistic models to assess compliance or risk of tax declaration for the purpose of classifying enterprises have widely applied. However, with the trend of applying machine learning technology, and furthermore can use machine learning and artificial intelligence to predict information quickly and accurately, there are currently few studies, both domestically and internationally. Research on machine learning application to evaluate the probability of fraud (tax avoidance) of typical enterprises such as Rahayu Abdul Rahman et al [6], designed and deployed a machine learning model based on selected algorithms with the data set of 3,365 listed Malaysian companies from 2005 to 2015. The performance of each machine learning algorithm on the test dataset was observed based on two training approaches.

In addition, Yin & Luo [7] argues that with the complexity of massive data and the secrecy of modern transactions, traditional tax risk identification can no longer adapt to the development of the times. The application of complex machine learning algorithms shows the suitability to process tax data and also the ability to produce higher accuracy

in tax risk identification. Other research also provides certain evidence for better result of machine learning application in tax risk management, reduce tax risk and tax loss [8], [9], [10], [11].

Besides, the combination of machine learning approaches and data processing methods shows potential in enhancing the prediction power for tax data. For example, Oliveira et al [12] confirm machine learning techniques in the search of tax defaults evidence. Baghdasaryan et al [13] combine information contained in the supplier and buyer network of the taxpayer with machine learning models to identify tax fraud probability. Savić et al [14] analyze the tax risk management by using a hybrid method which mainly focus on outlier detection approach and unsupervised machine learning algorithms. Also, different studies have applied various machine learning models to process data related to tax risk problems [15], [16], [17]

Furthermore, with cross-validation training approach, the performance of each machine learning algorithm was tested on selection groups with different attributes, namely industry, governance, year and characteristics of the company. The findings indicate that the machine learning model exhibits better reliability with industry, governance, and company-specific features. In that, the Random Forest and Logistic Regression algorithms are more efficient. Andr´e Ippolito et al [18], applied machine learning to predict tax fraud crimes. In their methods, the authors have structured a process that includes the following steps: feature selection; data partitioning; model training and testing; model evaluation. The results show that the Random Forest and Logistic Regression methods are more effective. With better predictions, the audit plan will become more assertive. Thus, increasing the taxpayer's compliance with tax laws and increasing tax revenue.

3. Research methodology & data

The tax risk dataset consists of 872 observations. In which, the target variable "tax_risk" is represented as a dummy with 2 values, 0 representing the absence of tax risk and 1 representing the observation with tax risk respectively. Our goal here is to build a machine learning model with Logistic regression to classify the target variable “tax_risk” based on the remaining variables. The classification of enterprises with risk (1) or no risk (0) is based on the tax industry's set of criteria and adjusted according to the actual tax inspection.

In this study, the authors apply the Logistic model with machine learning approach to evaluate the probability of risks in tax declaration of enterprises. Accordingly, the tax declaration and tax compliance of an enterprise depends on the following groups of factors:

- The group of factors reflecting the financial situation of the enterprise includes evaluation criteria such as: (i) Revenue and other income; (ii) Expenses; (iii) Liquidity; (iv) Asset; (v) Liabilities; (vi) Equity.
- The group of factors related to business performance includes: (i) Return on Assets (ROA); (ii) Return on Sales (ROS); (iii) Return on Equity (ROE).
- The group of factors that reflect the characteristics of the enterprise's operations, such as: (i) Business lines of enterprises; (ii) Enterprise characteristics.

- Group of factors to assess the performance of tax declaration and payment obligations of enterprises: (i) Related-party transactions factors; (ii) Tax payable; (iii) Management mechanism of the tax industry.

4. Results

This study uses the Standard Scaler method for the "tax_risk" dataset. Note that the scaler function is only fit to the training set and then transforms for both the training and test sets, to avoid the scaler function remembering the data in the test set. The purpose of dividing data into 2 separate sets (training set and test set) is to ensure that the performance evaluation of the model is objective when making predictions on the data set that the model has not encountered in the training process. So the scale synchronization for the variables is only fit on the training set. In addition, for this classification problem, the target variable should be kept as categorical value in order to properly represent the categorical nature that this variable represents. So, data scaling is only done with the independent variable. The following steps of processing and conducting the model includes: (i) build Logistic regression model and train the model with the data in the training set which has been divided above; (ii) perform the target predictor with the data in the test set ; (iii) performance evaluation of Logistic regression model.

According the test result, the accuracy scores show that the model has more than 80% accurate prediction results. This indicates that the Logistic regression model is relatively suitable for the data set "tax_risk". However, in order to evaluate the performance of the above model more comprehensively and accurately, this study take a closer look at the performance of the classification model on the basis of this Logistic regression with confusion matrix, classification report and ROC curve. The graph showing the results of the confusion matrix as follows:

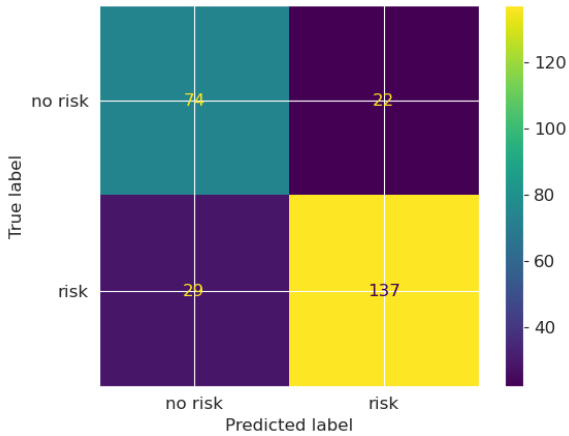

Figure 1: Confusion matrix for logistic model

Source: Author's calculation

The Confusion matrix shows that there are 74 correctly predictive observations for the "no-risk" class, 22 observations which are "no-risk" but are mistakenly predicted for the "risk" class, and 29 observations which are "risk" but mistakenly assigned to the "no-

risk" class, 137 observations were correctly predicted for the "risk" class. The results show that the model has a good predictive performance when the number of correct predictions is much higher than the number of false predictions in both classes of the target variable.

Table 1: Classification report

	precision	recall	f1-score	support
no risk	0.72	0.77	0.74	96
risk	0.86	0.83	0.84	166
accuracy			0.81	262
macro avg	0.79	0.80	0.79	262
weighted avg	0.81	0.81	0.81	262

Source: Author's calculation

Classification report shows that the "risk" class has higher precision and recall scores, so the f1-score score of this class is also higher. In addition, the number of observations belonging to the "risk" class is much higher than that of the "no-risk" class, so the weighted avg is higher than the mean (macro avg), but the difference is not significant.

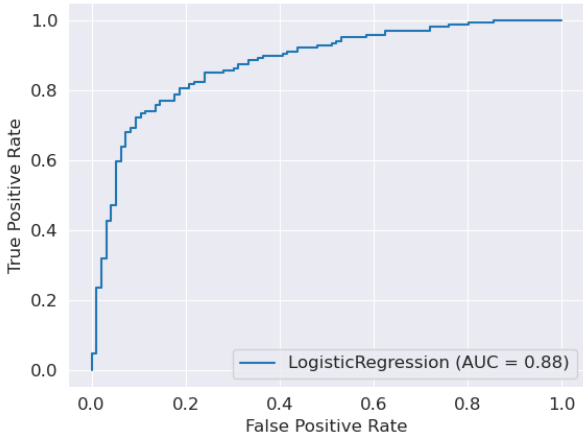

Figure 2: ROC curve

Source: Author's calculation

The ROC chart of the logistic regression model shows quite good results, the AUC score is quite high at 0.88. In general, almost all the performance evaluation factors of the classification model are at good level, which shows that the application of the logistic regression model with the tax risk data set "tax_risk" is reasonable.

5. Conclusion & recommendations

With the classification of firms into 2 groups (tax risk and no tax risk) with a sample of nearly 900 companies, applying logistic regression model by machine learning method shows the feasibility and accuracy. This study recommend that the management agencies can scale across the entire enterprise database, with big dataset, we can conduct identification and ranking using artificial intelligence. This will create efficiency in tax control and management, reduce tax fraud, and increase budget revenue.

References

[1] Nguyen Thanh Son (2010). Tax Textbook, Social Labor Publishing House, pages 3-7
[2] Andreoni, J., Erard, B. and Feinstein, J. (1998), Tax compliance. Journal of Economic Literature, Vol. 36, No. 2, pp. 818-860
[3] Song, Y., D. and Yarbrough, T., E. (1978), Tax ethics and taxpayer at 3 itudes: A survey, Public Administration Review, Vol. 38, No. 5, pp. 442-452.
[4] Singh, V. and Bhupalan, R. (2001), The Malaysian self assessment system of taxation: Issuse and challenges, Tax Nasional, 3rd quarter, pp. 12 – 17.
[5] McBarnet, D. (2001), When compliance is not the sulotion but the broblem: From changes in law to changes to attitude, Canberra: Australian National University, Centre for Tax System Integrity.
[6] Rahayu Abdul Rahman et al (2020), An application of machine learning on corporate tax avoidance detection model, Vol. 9, No. 4, December 2020, pp. 721~725, IAES International Journal of Artificial Intelligence (IJ-AI).
[7] Yin, M., & Luo, N. (2021). Tax Risk Prediction of Real Estate Based on Convolutional Neural Network. In Modern Management based on Big Data II and Machine Learning and Intelligent Systems III (pp. 49-56). IOS Press.
[8] Neñer, J., Cardoso, B. H. F., Laguna, M. F., Gonçalves, S., & Iglesias, J. R. (2022). Study of taxes, regulations and inequality using machine learning algorithms. Philosophical Transactions of the Royal Society A, 380(2224), 20210165
[9] Rathi, A., Sharma, S., Lodha, G., & Srivastava, M. (2021). A Study on Application of Artificial Intelligence and Machine Learning in Indian Taxation System. PSYCHOLOGY AND EDUCATION, 58(2), 1226-1233
[10] Zumaya, M., Guerrero, R., Islas, E., Pineda, O., Gershenson, C., Iñiguez, G., & Pineda, C. (2021). Identifying tax evasion in Mexico with tools from network science and machine learning. In Corruption Networks (pp. 89-113). Springer, Cham.
[11] Wang, Y., & Tang, G. (2019). The New Model of the Tax Vocational Education and the Application of Knowledge Management in China. Education Journal, 8(5), 196-201.
[12] Oliveira, V., Chaim, R. M., Weigang, L., Neto, S. A. P. B., & Filho, G. P. R. (2021). Towards a Smart Identification of Tax Default Risk with Machine Learning.
[13] Baghdasaryan, V., Davtyan, H., Sarikyan, A., & Navasardyan, Z. (2022). Improving Tax Audit Efficiency Using Machine Learning: The Role of Taxpayer's Network Data in Fraud Detection. Applied Artificial Intelligence, 1-23
[14] Savić, M., Atanasijević, J., Jakovetić, D., & Krejić, N. (2022). Tax evasion risk management using a Hybrid Unsupervised Outlier Detection method. Expert Systems with Applications, 116409.
[15] Agustiani, D., Wardani, S., & Riyadi, A. (2021, March). OpenCV and Machine Learning Implementation for the Vehicles Classification and Calculation in the Parking Tax Monitoring System at the Bantul Regency Regional Financial and Asset Agency (BKAD). In Journal of Physics: Conference Series (Vol. 1823, No. 1, p. 012062). IOP Publishing.
[16] Pavlova, K. S., & Knyazeva, N. V. (2021, April). Artificial Intelligence Technologies in Tax Consulting and Forensic Tax Expertise. In International Scientific Conference "Digital Transformation of the Economy: Challenges, Trends, New Opportunities" (pp. 291-300). Springer, Cham.
[17] Jamshidi, R., Barzegar, B., & Mohseni, A. (2022). Developing A Model To Improve The Quality Of Tax Audits. Iranian Journal of Accounting, Auditing and Finance.
[18] Andr´e Ippolito and Augusto Cezar Garcia Lozano (2019), Tax Crime Prediction with Machine Learning: A Case Study in the , Municipality of S˜ao Paulo, Tulane Economics Working Paper Series.