IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 9, No. 4, December 2020, pp. 721~725
ISSN: 2252-8938, DOI: 10.11591/ijai.v9.i4.pp721-725

An application of machine learning on corporate tax avoidance detection model

Rahayu Abdul Rahman[1], Suraya Masrom[2], Normah Omar[3], Maheran Zakaria[4]
[1]Faculty of Accountancy, Universiti Teknologi Mara, Perak Branch, Tapah Campus, Malaysia
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara, Perak Branch, Tapah Campus, Malaysia
[3]Accounting Research Institute, Universiti Teknologi Mara, Selangor, Malaysia
[4]Faculty of Accountancy, Universiti Teknologi Mara, Kelantan Branch, Machang Campus, Malaysia

Article Info

Article history:

Received Jul 28, 2020
Revised Oct 10, 2020
Accepted Nov 8, 2020

Keywords:

Machine learning
Malaysia
Prediction
Tax avoidance

ABSTRACT

Corporate tax avoidance reduces government revenues which could limit country development plans. Thus, the main objectives of this study is to establish a rigorous and effective model to detect corporate tax avoidance to assist government to prevent such practice. This paper presents the fundamental knowledge on the design and implementation of machine learning model based on five selected algorithms tested on the real dataset of 3,365 Malaysian companies listed on bursa Malaysia from 2005 to 2015. The performance of each machine learning algorithms on the tested dataset has been observed based on two approaches of training. The accuracy score for each algorithm is better with the cross-validation training approach. Additionationally, with the cross-validation training approach, the performances of each machine learning algorithm were tested on different group of features selection namely industry, governance, year and firm characteristics. The findings indicated that the machine learning models present better reliability with industry, governance and firm characteristics features rather than single year determinant mainly with the Random Forest and Logistic Regression algorithms.

Corresponding Author:

Rahayu Abdul Rahman
Faculty of Accountancy
Universiti Teknologi Mara
Perak Branch, Tapah Campus, 34500 Tapah Road, Tapah, Perak, Malaysia
Email:rahay916@uitm.edu.my

1. INTRODUCTION

The development of the corporate tax evasion prediction model has long been regarded as an important issue in the academic, tax authorities and business community. It is because corporate tax evasion might has significant impact on countries revenues as well as public budget which limit country development and continuity of the people well-being. Although corporate tax is the highest contributors to the government revenues, such taxes represent the most considerable cost incurred by the firms [1]. Thus, managers attempt to minimize the tax burden using various legal and illegal plans knowns as tax avoidance and tax evasion strategies. Tax avoidance is one of the various legal plans firms may use to minimize corporate tax liability [2]. Meanwhile, tax evasion, is illegal, deceptive and fraudulent practice engages by firms to avoid paying actual tax liability [2].

The prevalence and negative impact of corporate tax evasion has sparked the interest to study on corporate tax evasion detection models [3-5]. Despite the growing body of literature on tax evasion prediction, very little attention has been devoted to the other tax plan strategies including corporate tax

avoidance, making it as interesting topic to be studied. Motivated by the limitation, this paper attempts to fill the gap by developing model to predict tax avoidance strategies. In the recent Industrial 4.0 era, many recent studies have demonstrated that machine learning and big data mining approaches are effective tools for many problems [6-10] and also for the detection of financial fraud including tax fraud [11]. Despite the wiser used machine learning in many applications, there is limited literature on the development of related to tax avoidance.Therefore, this study has been initiated to fill the gap by looking at the experimental methods of developing tax avoidance classification model based on machine learning.

The contribution of this paper is two-fold. Firstly, it presents the study that deepens current understanding on the effectiveness of machine learning approach in predicting corporate tax avoidance prediction. Secondly, it provides another design and implementation approaches that extends the method used in [11] that exclude the elements of governance and firms sectors.

2. RELATED WORK

2.1. Tax avoidance prediction

The first study on corporate tax avoidance prediction model conducted by [11]. This study used two main input factors; network characteristics and firm specific characteristics to predict tax avoidance. Using three machine learning techniques which are logistic regression, decision trees and random forests, the findings revealed that network characteristics have a significant contribution to the improvement of predictive ability for tax avoidance model.

2.2. Tax evasion prediction

Study by [6] aims to detect tax evasion on Taiwan value-added tax (VAT) data by using data mining technique. Data mining technique applies in this study as it ables to filter non- compliant VAT report. The findings show that data mining technique able to enhance the tax evasion detection which in turn mitigate VAT evasion practice and lossess. Meanwhile, [4] uses Gaussian process prediction technique to propose income tax fraud prediction model. The performance of the prediction model of this study has been measured by using normalized root mean square error (NRMSE) and coefficient of determination (COD) with varying hyper parameters. Recently, [5] attempts to predict tax evasion using hybrid intelligent system for the Iranian textile and food sectors firms. Using combination of multilayer perceptron (MLP) neural network, support vector machine and logistic regression classification model with harmony search optimization, the results show that MLP neural network outperforms other combinations for both sectors.

3. RESEARCH METHOD

3.1. Dataset and features selection

The sample of this study consists of 3,365 Malaysian listed firms from 2005 to 2015. Similar to prior research [12-20], the effective tax rates (ETR) is used to measure the tax avoidance strategies. This study uses four main features to predict tax avoidance. The first category of features is firm specific characteristics. It consists of four features namely firm size (SIZE), firm leverage (LEVERAGE), firm growth (GROWTH) and firm profit (PROFIT). Following [20-21], this study use SIZE as feature as large firms often receive more media attention, have a higher analysts following and face a greater level of public scrutiny that results in less tax avoidance. Second, the study use LEVERAGE as feature as firms with higher levels of debt have lower ETR because of the deductibility of interest payments for tax purpose [21]. Third, the study use GROWTH as feature as it represents the firms' investment opportunities. In [16] argues that firms with greater investment opportunities have higher ETRs. Finally this study uses PROFIT as feature as [15] argue that firms with good performance are aggressive tax planner.

Further, this study select corporate governance regime periods as features. It has been widely accepted that corporate governance mechanism enhance best practice in the form of corporate performance [22] and transparency [8]. The effective governance system can reduce tax avoidance as the system has the ability to govern and monitor corporate tax decisions [13]. As many other Asian Pacific countries, the importance of corporate governance in Malaysia rose after the Asian Financial Crisis in 1997. Following the crisis, Malaysian government established a high level finance committee on corporate governance (FCCG) who rules is to review governance practice in corporate sector and recommend legal reform to strengthen their effectiveness [21-24]. In 2000, Malaysia code of corporate governance was issued by FCCG. The code essentially aim to set out principle and best practices on structures and process that companies may use in their operation toward achieving the optimal governance framework. All listed companies are required to disclose their level of compliance with its recommendations in view to provide a strong facilitative regulatory regime including corporate accountability and high quality corporate governance mechanism that would

strengthen investor confidence [23]. The MCCG was reviewed in 2007 with objectives to strengthening the board directors (BOD), audit committees and the internal audit function. This is to ensure that the BOD and the audit committee discharge the roles and responsibility efficiently [4]. The MCCG was again being revised in 2012. Areas that have been strengthened in the revision include the roles, responsibilities, composition of the board, directors (commitment, independence, and remuneration), risk-management framework, internal controls system, the integrity of reporting for the financial and lastly is the relationship between the company and the shareholders. Since December 31, 2012 when MCCG (2012) was established, all the listed companies were required to provide their annual report that compliance to the principles of MCCG (2012) [22].

The third category of features is firms' industry/sector (Properties, Reits, Technology, Finance, IndustrilProd, Cons, Const, Plant, IPC, Trad/ser). Firm from different industries has different tax implication which in turn has different opportunities to reduce its tax burden. Some industries are highly competitive and very reactive to economic condition and political event, some industries are protected by the government and the rest rather be in safe environment. In [13] suggests that industrial effects might be very important factors that will explain the differences in ETR for non-western firms due to the long standing industry policy in these countries to protect certain sectors. Consistent with the argument, [13] in their study find that manufacturing firm and hotels pay significant lower effectives tax compared to another firms in other industries in Malaysia. Meanwhile, [25] mentioned that corporate effective tax rates during the year between 2000 to 2004 in Malaysia differ considerably between companies from the same sector and between sectors. The findings reveal that firm from trading and services, properties and construction sector paid higher effective taxes. The final input feature is year. The year period is from 2005 to 2015 to capture the changes of country corporate tax rate.

3.2. Machine learning algorithms

The five algorithms used in this study were logistic regression, K-Nearest Neighbour, gaussian nave bayes (NB), decision tree and random forest. The configurations of hyper-parameters for each algorithm was identified based on a series of preliminaries experiments and literatures.

3.3. Training and validation approach

Simple split and cross validation (CV) approaches have been employed to each algorithm. The configuration has been set to a ratio of 80:20 between the training and validation dataset. The advantage of CV approach is wiser used of dataset when it randomly divides the training and validation datasets with multiple times (depends on the number of CV). The CV used *ShuffleSplit* cross validation technique (split numbers=5, random state=50).

3.4. Software and hardware platforms

The experiments for were implemented with *Python* programming in Jupyter notebook platform and run in a notebook Intel i7 processor with 16GB RAM.

4. RESULTS AND ANALYSIS

The five algorithms used in this study were logistic regression, K-Nearest Neighbour, gaussian nave bayes (NB), decision tree and random forest. Experiments for each algorithm on the four types of feature were run for five times and the mean of accuracy were recorded. At first, the split training approach was employed, and the result is listed in Table 1.

Table 1. The accuracy score of each algorithm with split training approach

Algorithm	Industry	Governance	Year	Firm Characteristics
Logistic regression	65.82	66.87	68.82	68.52
KNN	61.02	66.87	69.42	57.42
Gaussian NB	58.92	66.87	69.42	52.62
Decision Tree	66.27	66.87	69.87	71.81
Random forest	65.82	66.87	69.87	70.01

The industry types feature set contributed high accuracy score (> 60%) except with gaussian NB with slightly lower (58.92). All the algorithms also worked better on the governance and year features. On the firm characteristics feature set, decision tree and random forest has shown better performances with accuracy scores higher than 70%. Kruskal Wallis test was applied to check if there were statistical differences among the mean validation scores of each algorithm in the four types of determinant.

The p-values obtained from industry types, governance, year and characteristics was 0.011,0. 010, 0.001 and 0.011 respectively at a significance level of 95%, which shows that all the samples were not generated in the same distribution. Furthermore, Table 2 shows the results of accuracy scores with the CV training approach.

Table 2. The accuracy score of each algorithm with cross-validation training approach

Algorithm	Industry	Governance	Year	Firm Characteristics
Logistic regression	0.78	0.81	0.81	0.80
KNN	0.71	0.74	0.66	0.72
Gaussian NB	0.68	0.81	0.79	0.64
Decision Tree	0.80	0.81	0.80	0.80
Random forest	0.80	0.81	0.80	0.80

Improvements of accuracy scores have been achieved by all algorithms when implemented with CV training approach. Majority of algorithms performed very well with all feature sets (higher than 80% of accuracy score). The Gaussian NB produced slightly lower accuracy score with the industry types and characteristics feature sets, while KNN with the year feature. Kruskal Wallis test on all results of experiments with the CV training approach was also showed that all the determinant types have p-values were less than 0.05 (0.0-0.01), hence rejecting null hypothesis that all results have been produced with the same distribution.

Furthermore, the area under curve (AUC) results from CV training approach presented in Table 3. AUC is used to measure the model performance by mean of reliability of the model. AUC calculates the entire two-dimensional area underneath the entire receiver operating characteristic (ROC) curve graph from (0,0) to (1,1). The ROC graph plots two parameters; namely true positive rate (TPR) and false positive rate (FPR). TPR representing how often is the tax avoidance occurred or detected from the sample dataset. In order words is the ability of the model to recall the existence of tax avoidance. On the other hand, FPR in this case is the number of 0 tax avoidance that detected as 1. It is calculated as the ratio between the negative condition wrongly classified as positive and the total number of actual negative condition. Compared to accuracy score, the AUC measures the performance of a binary classifier averaged across all possible decision thresholds. Therefore, the AUC results are smaller than accuracy score and would be more reliable to measure the model performance. The AUC of year feature of all algorithms are lower compared to the other feature set. Random forest and logistic regression classifiers outperformed another three algorithms when tested on industry types, governance and characteristics features set.

Table 3. The AUC of each algorithm with different types of feature sets

Algorithm	Industry	Governance	Year	Firm Characteristics
Logistic regression	0.634	0.612	0.430	0.634
KNN	0.485	0.544	0.512	0.485
Gaussian NB	0.611	0.573	0.570	0.611
Decision Tree	0.571	0.557	0.565	0.571
Random forest	0.635	0.605	0.574	0.635

5. CONCLUSION

This paper presents the review and empirical research works for the design and implementation of machine learning classification model on corporate tax avoidance among Malaysian listed companies. Based on real dataset, the performances evaluation of different machine learning models that employed different training approaches and features selection are extensively presented. Generally, all algorithms produced good accuracy results with cross-validation training approach compared to simple split aproach. From the reliability perspective, year feature has the lowest contribution to majority of algorithms performance compared to industry types, governance and firms specific characteristics. This work can be further enhanced in the future by considering different aspects of tax avoidance and implementation approaches.

ACKNOWLEDGEMENTS

The authors would like to thank the financial support granted by the Universiti Teknologi MARA and Ministry of Education, Malaysia for this project under FRGS grant No 600-IRMI/FRGS5/3 (140/2019).

REFERENCES

[1] S. Chen, X. Chen, Q. Cheng, and T. Shevlin, "Are family firms more tax aggressive than non-family firms?," *Journal of Financial Economics*, vol. 1, no. 95, pp. 41-61, 2010.

[2] M. Hanlon, and S. Heitzman, "A review of tax research," *Journal of accounting and Economics*, vol. 50, no. 2, pp. 127-78, 2010.

[3] S. K. Babu, and S. Vasavi, "Predictive Analytic as a Service on Tax Evasion Using Feature Engineering Strategies," *InSmart Intelligent Computing and Applications, Springer*, pp. 393-402, 2019.

[4] S. K. Babu, and S. Vasavi, "Predictive Analytics as a Service on Tax Evasion using Gaussian Regression Process," *HELIX*, vol. 7, no. 5, pp. 1988-93, 2017.

[5] E. Rahimikia, S. Mohammadi, T. Rahmani, and M. Ghazanfari, "Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran," *International Journal of Accounting Information Systems*, vol. 1, no.25, pp.1-7, 2017.

[6] R. S.Wu, C. S. Ou, H. Y. Lin, S. I. Chang, and D. C. Yen, "Using data mining technique to enhance tax evasion detection performance," *Expert Systems with Applications*, vol.39, no. 10, pp. 8769-77, 2012.

[7] M. Nawir, A. Amir, N. Yaakob, and O. B. Lynn, "Effective and efficient network anomaly detection system using machine learning algorithm," *Bull. Electr. Eng. Informatics*, vol. 8, no. 1, pp. 46–51, 2019.

[8] F. R. Kamala, P. R. J. Thangaiah, and A. Info, "An improved hybrid feature selection method for huge dimensional datasets," *IAES Int. J. Artif. Intell.*, vol. 8, no. 1, pp. 77–86, 2019.

[9] V. S. Padala, K. Gandhi, and D. V Pushpalatha, "Machine learning : the new language for applications," *IAES Int. J. Artif. Intell.*, vol. 8, no. 4, pp. 411–421, 2019.

[10] A. Zakrani, M. Hain, and A. Idri, "Improving software development effort estimation using support vector regression and feature selection," *IAES Int. J. Artif. Intell.*, vol. 8, no. 4, pp. 399–410, 2019.

[11] J. Lismont, E. Cardinaels, L. Bruynseels, S. De Groote, B. Baesens, W. Lemahieu, and J. Vanthienen, "Predicting tax avoidance by means of social network analytics," *Decision Support Systems*, vol. 1, no. 108, pp. 13-24, 2018.

[12] H. A. Annuar, I. A. Salihu, and S. N. Sheikh Obid, "Corporate ownership, governance and tax avoidance: An interactive effects," *Procedia-Social and Behavioral Sciences*, no. 164, pp. 150-60, 2014.

[13] C. Derashid, and H. Zhang, "Effective tax rates and the industrial policy hypothesis: evidence from Malaysia," *Journal of international accounting, auditing and taxation*, vol. 12, no. 1, pp. 45-62, 2003.

[14] J. Kasipillai, M. Y. Lee, and S. Mahenthiran, "Political connections, corporate governance and effective tax rates in Malaysia," *Austl. Tax F.*, vol. 32, pp. 493, 2017.

[15] S. Mahenthiran, and J. Kasipillai, "Influence of ownership structure and corporate governance on effective tax rates and tax planning: Malaysian evidence," *Austl. Tax F.*. vol. 27, pp. 941, 2012.

[16] I. A. Salihu, H. A. Annuar, and S.N. Obid, "Foreign investors' interests and corporate tax avoidance: Evidence from an emerging economy," *Journal of Contemporary Accounting & Economics*, vol. 11, no. 2, pp.138-47, 2015.

[17] G. M. Spooner, "Effective tax rates from financial statements," *National Tax Journal*, 1986, vol. 39, no. 3, pp. 293-306. 1986.

[18] E. A. Wahab, A. M. Ariff, M. M. Marzuki, and Z. M. Sanusi, "Political connections, corporate governance, and tax aggressiveness in Malaysia," *Asian Review of Accounting*, 2017.

[19] I. A Salihu, S. N. Sheikh Obid, and H. A. Annuar, "Measures of corporate tax avoidance: empirical evidence from an emerging economy," *International Journal of Business & Society*, vol. 14. No. 3, 2013.

[20] K. A. Kim, and P. Limpaphayom, "Taxes and firm size in Pacific-Basin emerging economies," *Journal of international accounting, auditing and taxation*, vol. 7, no. 1, pp. 47-68, 1998.

[21] K. H. Chan, P. L. Mo, and A. Y. Zhou, "Government ownership, corporate governance and tax aggressiveness: evidence from China," *Accounting & Finance*, vol. 53. No. 4, pp. 1029-51, 2013.

[22] P. R. Bhatt, "Corporate governance in Malaysia: has MCCG made a difference," *International Journal of Law and Management*, 2016.

[23] A. A.Haji, "The relationship between corporate governance attributes and firm performance before and after the revised code," *International Journal of Commerce and Management*, 2014.

[24] M. F. Rahim, R. J. Johari, N. F. Takril, "Revisited note on corporate governance and quality of audit committee: Malaysian perspective," *Procedia Economics and Finance*, pp. 213-21, 2015.

[25] R. M. Noor, N. A. Matsuki, and B. Bardai, "Corporate effective tax rates: a study on Malaysian public listed companies," *Management and Accounting Review*, vol. 7, no. 1, pp. 1-20. 2008.