An Ensemble Machine Learning Model to Detect Tax Fraud: Conceptual Framework

Kudzanai Charity Muchuchuti

Lecturer, BA ISAGO University, Department of Accounting and Finance, Botswana

DOI : https://dx.doi.org/10.47772/IJRISS.2024.806171

Received: 16 May 2024; Revised: 06 June 2024; Accepted: 11 June 2024; Published: 16 June 2024

ABSTRACT

Most governments throughout the world, especially in developing and underdeveloped countries, depend more on tax revenue to fund public expenditure and investments. In the wake of Covid19, even governments that did not depend largely on tax revenue are forced to do that since their other sources of income were affected by the pandemic as borders were closed and nations were on lockdowns. Research, however, has shown that tax fraud is rampant especially in less developed countries. Traditional methods of detecting tax fraud are costly and they largely depend on the experts' past experience. This renders them less effective where new mechanisms of tax fraud are involved. In this work I provide a conceptual framework on the use of ensemble machine learning models to detect tax fraud. I use decision trees, support vector machines and logistic regression as the base models. I hypothesize that ensemble methods outperform unsupervised machine learning models and the use of a single algorithm under supervised machine learning models. The outcomes of this research will serve to provide a framework that will help tax authorities to detect tax fraud thereby increasing the revenue collected.

Keywords: tax fraud, machine learning, supervised machine learning methods, ensemble methods

INTRODUCTION

Tax fraud poses a great challenge to the fiscal stability of economies worldwide. It drains the economies of essential resources needed to finance public expenditure. This also exacerbates income inequality. Economies are losing billions of dollars annually due to various forms of tax evasion. Tax is the largest source of revenue to some governments globally in particular countries in sub-Saharan Africa, the Caribbean, South Asia and Latin America. However, tax fraud is rampant in these countries leading to budget deficits and limited public investment. (de Roux et.al., 2018). Tax fraud is defined as deliberately evading tax through deliberately submitting false statements, producing fake documents and this is a criminal act which is punishable (Nosiri et al, 2021; Saxunova and Szarkova, 2018). Tax fraud is difficult to measure as people try to conceal it since it is a criminal activity. Countries, however, try to measure it by using the tax gap (Alm, 2012). The IRS defines tax gap as the difference between the expected tax revenues and the actual tax revenues collected timely. According to Kurauone et. al. (2021), statistics indicate that countries lose billions of dollars through tax evasion. The HM Revenue and Customs (HRMC) estimates that the UK government loses billions of pounds every year. The estimated tax gap for the 2019 / 2020 fiscal year was £35 billion pounds (HRMC; 2019). According to the IRS report, America estimated an annual average gross tax gap of US$441 billion and an annual average net tax gap of US$381 billion representing a net compliance rate of 85.8% in the years 2011 – 2013. (IRS; 2019). This indicates a non-compliance rate of almost 15% which is worrisome. In the 2018 – 2019 fiscal year, the Australian Taxation Office (ATO) estimated a net tax gap of $33.5 billion. (ATO, 2019). While developed countries estimate the tax gaps, very little information is available for developing countries. Dare et al (2019) asserts that in South Africa, the only available tax gap is that of value added tax (VAT). However, Kyle Mandy, a partner, and head of national tax technical at PwC SA, at a conference hosted by the South African Institute of Taxation (Sait) in 2021, reported that some studies believe that the overall tax gap in South Africa is R200 billion. These statistics show that tax fraud is an issue that needs to be dealt with. After the Covid-19 pandemic, most economies are relying more on tax revenue again since their other main sources of revenue were disrupted by the pandemic.

The two traditional methods that are commonly used to detect tax fraud are the auditors experience and rule-

based systems. (de Roux et al, 2018). Auditors' experience involves the random selection of tax declarations and then auditing them. The auditor uses his or her experience, domain knowledge and intuition. With a rule-based system, some sets of if-then rules are developed by the auditors after reviewing the characteristics of a fraud case. They use those characteristics to set a rule that then triggers a signal whenever fraud is detected. The authors argue that these methods may not detect new fraud mechanisms on their own because they are based on past experience. The other challenge is that rule-based systems are costly in terms of building, maintaining, and updating them. This is because of their reliance on the subjective judgement of the tax auditors. Research has been done on the use of machine learning to detect tax fraud using supervised machine learning techniques (Mittal et al, 2018; da Silva et al, 2016), semi-supervised techniques (Mi et al, 2020, Wu et al, 2020, Kleanthous & Chatzis, 2020) and unsupervised machine learning techniques (de Roux et al, 2018; Mehta et al, 2020, Savic et al, 2021). While standalone machine learning algorithms have been used to detect tax fraud with promising results, ensemble methods have outperformed individual machine learning algorithms in other domains as in Khedr et al (2021), Bagga et al (2020), Twala (2010) and Olowookere & Adewale (2020). Standalone methods may struggle to cope with the evolving tasks of fraudsters who are always working hard to outsmart the traditional fraud detection systems. This study hypothesises that applying ensemble machine learning methods will enhance accuracy and adaptability. The purpose of this study is therefore to provide a conceptual framework on the detection of corporate tax fraud using ensemble methods.

The following sections are organised as follows; Section 2 focuses on related works; Section 3 shows the proposed ensemble machine learning model; Section 4 details the model limitations and Section 5 presents the conclusion.

LITERATURE REVIEW

2.1 Importance of Tax Fraud Detection

Tax fraud is a social ill that has troubled many countries especially the countries that depend on it for their fiscal planning. (de Roux, 2018). Cobham (undated) identified the purpose of taxation using 4Rs which are revenue, redistribution, repricing and representation. Governments use tax as revenue to finance public expenditure. They also use it to redistribute wealth to raise the poor out of poverty. Representation looks at the link between tax and political representation and higher quality governance. Repricing refers to how the government uses taxation to encourage certain economic activities and discouraging certain economic activities. These 4Rs show the importance of taxation and governments will not be able to achieve this in cases where tax fraud is high. Tax evasion hinders a country's economic growth, and it leads to losses that reduce the revenue needed to finance public expenditure (Mehta et al (2020); Junqué de Fortuny et al (2014)) and this also leads to unfair redistribution of wealth and tax increases. (Junqué de Fortuny et al, 2014).

Traditionally, tax authorities have used auditors' experience and rule-based systems (de Roux, 2018), analytical procedures, review of documents, observation, and informants. (Malaszczyk & Purcell, 2017). These traditional methods have been found to be costly and time consuming. They are also subjective since they are based on the auditor's experience. They might not be able to pick the new trends in tax fraud. Therefore, the related works that will be looked at in this study are related to detecting tax fraud using machine learning methods.

2.2 Machine Learning for Tax Fraud Detection

Recent studies have been done to detect tax fraud using machine learning techniques. This section therefore focuses on the previous works done using different machine learning techniques. In 2015, Matos et al, in collaboration with the Brazilian Tax Agency, used a supervised ML technique where they created a fraud scale that ranks taxpayers based on their potential to commit a fraud. Preliminary results showed that the method may have an 80% accuracy level in detecting fraudsters. Mittal et al (2018) developed a system that used one sided labels to enable the tax authority to identify bogus firms to be targeted for physical inspections. They applied a classifier to Vat returns to enable this. The aim was to increase tax compliance. da Silva et al (2016) used Bayesian networks to develop a predictive model that selects taxpayers for audit purposes.

Kleanthous & Chatzis (2020), in collaboration with the Cyprus Tax Agency, developed a semi-supervised ML technique that helps the tax authority to select companies for VAT audits. They developed a gated mixture variational autoencoder deep network. Gao et al (2021) also used semi-supervised ML techniques to develop a multi-stage method that detects tax evasion using a novel ML algorithm known as PnCGCN. Mi et al (2020) argued that detecting tax evasion in real life must be formalised as a positive learning problem. They therefore proposed a new method to detect tax evasion based on Positive Unlabooooooelled Learning with Network Embedding features.

Wei et al (2019) proposed unsupervised conditional adversarial networks to detect tax evasion. They combined the label predictor and the distribution adapter to get the end-to-end learning of the unsupervised feature transfer. de Roux et al (2018) utilised unsupervised machine learning models for the identification of taxpayers who underreport their tax liabilities. Their model increases operational efficiency in supervising the tax process by marking some declarations as suspicious and labelling as suspicious previously undetected tax declarations. Mehta et al (2020) designed a method that tracks those taxpayers who do not submit their tax returns to evade indirect taxes. They derived ten parameters using the TrustRank algorithm.

A significant gap still exists in the ability of the current models to synthesize insights derived from different machine learning approaches to handle tax fraud effectively. Fraudsters are rapidly coming up with new sophisticated tactics therefore standalone methods may fail to capture all the aspects of fraudulent behavior. Ensemble methods have been used in other domains for example in detecting financial statement fraud and the studies have shown that ensemble methods outperform single algorithms. This work therefore seeks to provide a conceptual framework in using these ensemble methods to detect corporate tax fraud.

DATA AND PROPOSED METHODOLOGY

This section outlines the proposed methodology of the conceptual model.

The schematic of the methodology is as depicted in Figure 1.

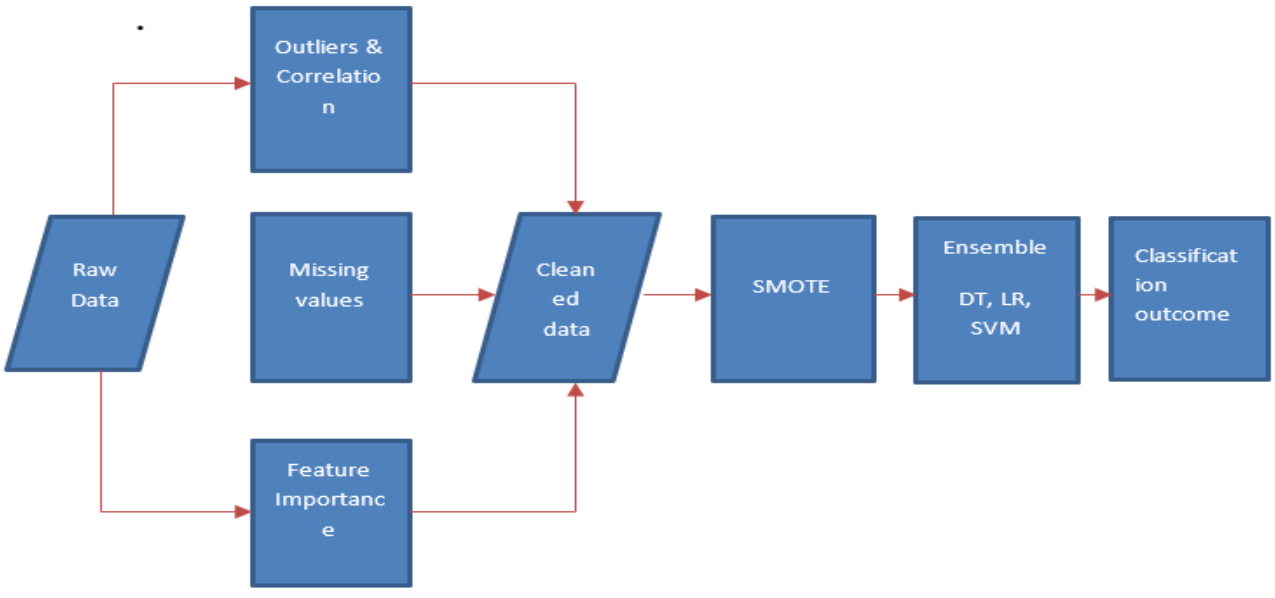

Fig 1 – Proposed model

Raw corporate tax data is passed through a pre-processing module which includes dealing with outliers, missing values and feature selection. The Synthetic Minority Oversampling Technique (SMOTE) module is aimed at correcting class imbalance. Balanced cleaned data becomes input to the ensemble of three base algorithms – Decision tree (DT), Sector Vector Machine (SVM) and Logistic regression (LR) where voting is used to settle at an agreed class.

ENSEMBLE MODEL

4.1 Data Pre-processing

Data pre-processing entails preparing data for modelling. Pre-processing sub-processes include outlier detection, dealing with missing values and feature selection.

4.1.1 Outlier detection

It is assumed D is the input tax dataset with attribute features f_1 to f_d, where d is the total number of features. The k- means clustering algorithm which is used to detect outliers in D (as used in Savic et al, 2021). Using domain knowledge, a tax expert is used to assign weights to each of the features. This has the effect of modifying the equation of the eucledian distance between instances p and q as follows:

Distance (p, q)= $\sqrt{\sum_{i=1}^{d} w_i^2 (f_i^1 (p)) - f_i^1 (q)^2}$ (1)

Where w_1 is f's weight and $f^1_i(r)$ is …….f_i's value on a unit interval scale.

The k- means identified clusters are used to identify outliers by computing the outlierness for each of the small clusters thus:

Outlierness (s) = minimum $c \epsilon C_1$ {distance (centroid(s), centroid(c))} (2)

where C_1 is the set of large clusters.

4.1.2 Missing values

Strategies for dealing with missing values broadly fall into three categories namely exclusions, simple mean imputations, and interpolation (Feng et al, 2021). A multiple imputation method is used to deal with missing data as it minimises imputation standard errors. (Feng et al, 2021).

4.1.3 Feature importance

We compute the Gini importance value of each feature selecting features with the highest Gini importance. Features with smaller Gini importance are deemed to be least linked to the target variable and are therefore dropped. (Khedr, 2021).

4.2 SMOTE

Typically, tax fraud data will be imbalanced, favouring the non-fraudulent class as opposed to the fraudulent class (the minority class). Modelling with imbalanced data leads to biased classification performance. One way to deal with this problem is to under sample the majority class but owing to the small size of tax fraud data sets, we resort to oversample the minority class using SMOTE. (Chawla et al, 2002). This technique was chosen for its effectiveness and also because it is the most commonly used in literature. (Khedr et al. 2021). SMOTE addressing the over-fitting challenge caused by random over-sampling techniques.

4.3 Base classifiers

Studies have demonstrated that SVM, LR and DT have performed well for binary fraud detection problems (Khedr et al. 2021), as individual classifiers.

4.3.1 Decision Tree

According to Osisanwo et al (2017) decision trees classify instances through sorting them based on feature values. The nodes in a DT have multiple levels with the root node being the top-most node. The child (internal) nodes epitomise the tests on input variables. The classification algorithm moves towards the appropriate internal

node based on the test outcome. Uddin et al. (2019) explained that the testing and branching process continues until it reaches the leaf node which corresponds to the decision outcomes.

4.3.2 Support Vector Machines

SVM begins by mapping all data items into a feature space and then it identifies the hyperplane that divides the data items into 2 classes at the same time maximising the marginal distance between both classes while minimising the classification errors. (Uddin et al., 2019). They further define the marginal distance as the distance between the hyperplane and its closest instance.

4.3.3 Logistic Regression

The LR is only a binary classifier that represents the non-occurrence or occurrence of an event. It assists in getting the probability that a new instance belongs to a certain class. Since it is a probability, the outcome ranges from 0 and 1. A threshold must be set to distinguish the two classes. (Uddin et al., 2019).

4.4 Ensemble Methods

Ensemble methods combine outcomes from base ML models to produce an optimal predictive model. Each of the three models makes a prediction (votes) for each test instance and the prediction that receives the highest votes becomes the ultimate class. (Khedr et al., 2021). To maintain simplicity and avoid lengthy training times, the model is going to employ simple majority voting mechanism.

4.5 Performance Evaluation

The performance of the model will be assessed in terms of accuracy, precision, recall and F1 score with the help of the confusion matrix. A depiction of the confusion matrix is as shown in Table 1.

Table 1 – Confusion Matrix

Real Label		
	Fraud	Normal
Predicted Label Fraud	T_P	F_P
Non-fraudulent	F_N	T_N

In Table 1,

True Positive (T_P) is the number of fraudulent transactions predicted as fraudulent.

False Positive (F_P) is the number of non-fraudulent transactions predicted as fraudulent.

False Negative (F_N) is the number of fraudulent transactions predicted as non-fraudulent.

True Negative (T_N) is the number of normal transactions predicted as normal.

$$Accuracy = \frac{T_p+T_N}{T} \quad (3)$$

$$Recall = \frac{T_p}{F_N+T_p} \quad (4)$$

$$Precision = \frac{T_p}{F_p+T_p} \quad (5)$$

$$F1 - Score = 2 * \frac{precision * recall}{precision+recall} \quad (6)$$

$$T = T_p + T_N + F_N + F_p \quad (7)$$

MODEL LIMITATIONS AND RECOMMENDATIONS FOR FURTHER STUDY

5.1 More time and resources are spent on training the model since multiple algorithms are involved.

5.2 Lack of explainability – Ensembling makes it harder to investigate the decisions made by the AI algorithm.

5.3 While the SMOTE technique addresses the class imbalance problem, it does not consider that neighbouring examples can be from other classes. This can introduce additional noise.

5.4 The efficacy of the ensemble model is yet to be evaluated with real tax fraud data.

5.5 It is recommended that the number of base models be increased beyond the three used in this work and that alternative ensemble techniques like AdaBoost be explored. These could further improve the performance of the model.

CONCLUSION

Tax revenue remains one of the highest contributors to the Gross Domestic Product of many economies especially for developing countries. Tax fraud erodes these earnings delaying the advancement of these economies to high income status. This research presented a concept of ensemble machine learning model for tax fraud prediction. The model will help plug potentially fraudulent corporate tax filings and help focus tax fraud investigation. This will potentially help recover lost revenue and even deter would be fraudsters.

REFERENCES

1. Alm, J. (2012). Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies. International tax and public finance, 19, 54-77.
2. ATO (2019)., https://www.ato.gov.au/about-ato/research-and-statistics/in-detail/tax-gap/australian-tax-gaps-overview/?anchor=Summaryfindings (Accessed on 18.11.2021)
3. Bagga, S., Goyal, A., Gupta, N., & Goyal, A. (2020). Credit card fraud detection using pipeling and ensemble learning. Procedia Computer Science, 173, 104-112.
4. Cobham, A., (undated), https://www.taxjustice.net/cms/upload/pdf/OCGG_-_Alex_Cobham_-_Taxation_Policy_and_Development.pdf
5. da Silva, L. S., Rigitano, H., Carvalho, R. N., & Souza, J. C. F. (2016). Bayesian networks on income tax audit selection - A case study of Brazilian tax administration. In R. N. Carvalho & K. B. Laskey (Eds.), Proceedings of the 13th UAI Bayesian.
6. Dare, C., du Plessis, S., & Jansen, A. (2019). Tax revenue mobilisation: Estimates of South Africa's personal income tax gap. South African Journal of Economic and Management Sciences, 22(1), 1-8.
7. de Roux, D., Perez, B., Moreno, A., Villamil, M. D. P., & Figueroa, C. (2018, July). Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 215-222).
8. Feng, S., Hategeka, C., & Grépin, K. A. (2021). Addressing missing values in routine health information system data: an evaluation of imputation methods using data from the Democratic Republic of the Congo during the COVID-19 pandemic.
9. Gao, Y., Shi, B., Dong, B., Wang, Y., Mi, L., & Zheng, Q. (2021). Tax evasion detection with FBNE-PU algorithm based on PNCGCN and PU learning. IEEE Transactions on Knowledge and Data Engineering. Advance online publication. https://doi.org/10.1109/TKDE.2021.3090075
10. Khedr, A. M., El Bannany, M., & Kanakkayil, S. (2021). An Ensemble Model for Financial Statement Fraud Detection. ARPHA Preprints, 1, e69590.
11. Kleanthous, C., & Chatzis, S. (2020). Gated mixture variational autoencoders for value added tax audit case selection. Knowledge-Based Systems, 188, 105048. https://doi.org/10.1016/j.knosys.2019.105048
12. Kurauone, O., Kong, Y., Sun, H., Famba, T., & Muzamhindo, S. (2021). Tax evasion; public and political corruption and international trade: a global perspective. Journal of Financial Economic Policy, 13(6), 698-729.

13. Malaszczyk, K., & Purcell, B. M. (2017). Big data analytics in tax fraud detection. Northeastern Association of Business, Economics and Technology, 233.
14. Matos, T., de Macedo, J. A. F., & Monteiro, J. M. (2015, July). An empirical method for discovering tax fraudsters: A real case study of brazilian fiscal evasion. In Proceedings of the 19th International Database Engineering & Applications Symposium (pp. 41-48).
15. Mehta, P., Mathews, J., Bisht, D., Suryamukhi, K., Kumar, S., & Babu, C. S. (2020). Detecting tax evaders using Trust Rank and spectral clustering. In Business Information Systems: 23rd International Conference, BIS 2020, Colorado Springs, CO, USA, June 8–10, 2020, Proceedings 23 (pp. 169-183). Springer International Publishing.
16. Mi, L., Dong, B., Shi, B., & Zheng, Q. (2020). A tax evasion detection method based on positive and unlabeled learning with network embedding features. In Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23–27, 2020, Proceedings, Part II 27 (pp. 140-151). Springer International Publishing.
17. Mittal, S., Reich, O., & Mahajan, A. (2018). Who is bogus? Using one-sided labels to identify fraudulent firms from tax returns. In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '18). Association for Computing Machinery. https://doi.org/10.1145/3209811.3209819
18. Modeling Applications Workshop (BMAW 2016) co-located with the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016). (2016, June 25). CEUR Workshop Proceedings, 1663, 14-20. CEUR-WS.org.
19. Nosiri, H., Abiahu, M. F. C., & Uwakwe, T. N. (2021). Forensic Accounting and Tax Fraud Detection in the Nigerian Public Sector. In Forensic accounting and tax fraud detection in the Nigerian public sector. In Taxation, Social Contract and Economic Development. Proceedings of 4th Annual International Academic Conference of the Chartered Institute of Taxation of Nigeria.
20. Olowookere, T. A., & Adewale, O. S. (2020). A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach. Scientific African, 8, e00464.
21. Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), 128-138.
22. Pérez López, C., Delgado Rodríguez, M. J., & de Lucas Santos, S. (2019). Tax fraud detection through neural networks: an application using a sample of personal income taxpayers. Future Internet, 11(4), 86.
23. Savic, M., Lukic, M., Danilovic, D., Bodroski, Z., Bajović, D., Mezei, I., ... & Jakovetić, D. (2021). Deep learning anomaly detection for cellular IoT with applications in smart logistics. IEEE Access, 9, 59406-59419.
24. Saxunova, D., & Szarkova, R. (2018). Global efforts of tax authorities and tax evasion challenge. Journal of Eastern Europe Research in Business and Economics, 2018, 1-14.
25. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1), 1-16.
26. Warren, N. (2018). Estimating tax gap is everything to an informed response to the digital era. eJTR, 16, 536.
27. Wei, R., Dong, B., Zheng, Q., Zhu, X., Ruan, J., & He, H. (2019). Unsupervised conditional adversarial networks for tax evasion detection. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 1675-1680). https://doi.org/10.1109/BigData47090.2019.9005656
28. Wu, Y., Dobriban, E., & Davidson, S. (2020, November). Deltagrad: Rapid retraining of machine learning models. In International Conference on Machine Learning (pp. 10355-10366). PMLR.
29. Zhang, C., & Ma, Y. (Eds.). (2012). Ensemble machine learning: methods and applications. Springer Science & Business Media.