Supervised Machine Learning for Tax Evasion Detection: A Case Study with the Brazilian Tax Administration

Cleyton Andre Pires[1]

[1]Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brazil

cleyton07@gmail.com

Abstract. *In this study, we present an innovative approach to enhance the audit case selection process within the Brazilian Tax Authority (RFB) by integrating Artificial Intelligence techniques. We employ supervised learning algorithms to predict taxpayers' annual income coupled with outlier detection techniques to strategically prioritize cases of heightened fiscal interest. This involves leveraging a comprehensive dataset of socioeconomic variables available to the Tax Administration. A pivotal facet of our methodology is its commitment to model explainability for ensuring fairness and compliance with legal and ethical considerations. Preliminary findings demonstrate promising results, positioning our model as a valuable complement to the existing rule-based system.*

1. Introduction

Tax evasion undermines the tax base and compromises fiscal equity. Income tax requires taxpayers to file annual returns detailing income, deductions, credits, and financial information. Audits by state tax agencies ensure compliance but are resource-intensive.

The Brazilian Tax Authority (RFB) uses a rule-based audit case selection system, comprising 2 stages: (1) cross-referencing data to identify inconsistencies and (2) individual analysis to validate the preliminary indicators. However, this system struggles to adapt to new fraud patterns and needs constant updates [OECD 2017].

This paper proposes a data mining-based audit selection method to complement the traditional system. Using machine learning and outlier detection, the goal is to uncover patterns and anomalies in large datasets, enhancing scalability and adaptability in detecting evolving tax evasion tactics. The study employs the CRISP-DM framework [1] to guide the data mining process.

The paper is organized as follows: Section II reviews relevant literature on AI in tax evasion. Section III outlines the proposed solution. Section IV presents a case study using RFB data. Section V discusses the results, and Section VI concludes the analysis.

2. Related Work

Due to the scarcity of publicly available tax data, published literature on combating tax fraud and evasion is limited. Tax administrations often refrain from disclosing internal projects to protect taxpayer information.

[1] [Wirth and Hipp 2000]

[de Roux et al. 2018] presents an approach to detect property tax evasion in Bogotá, Colombia, using unsupervised learning techniques. In [Zumaya et al. 2021], the authors utilize electronic invoices from the Mexican federal government to analyze taxpayer interaction patterns and identify evaders through temporal networks. [Savić et al. 2021] proposes the HUNOD (Hybrid UNsupervised Outlier Detection) method, achieving a 90%-98% validation rate, to identify outliers in Serbian tax data. [Lin et al. 2021] introduces TaxThemis, an interactive visual analysis system that helps tax auditors identify suspicious tax evasion groups through data analysis and visualizations.

In Brazil, the significant work of [da Silva L. S. et al. 2016] explores Bayesian networks to enhance the efficiency of the tax audit process for Income Tax declarations. Additionally, [Xavier et al. 2022] introduces an innovative methodology using Random Forest, Neural Networks, and Graphs to identify potential tax evaders in Goiás, Brazil, by distinguishing between ”default”and ”reputable”company profiles using open and public data.

3. Proposed Solution

This study proposes using supervised learning algorithms to predict individual taxpayers' annual income (dependent variable - Y) based on internal data provided to the Brazilian Tax Authority (RFB) (independent variables - X). The positive difference between the predicted annual income and the declared value in the Individual Income Tax Return (DIRPF) is termed **Estimated Omission Income (EOI)**.

Taxpayers with an EOI exceeding the **Mean Absolute Error (MAE)** of the predictive model are the focus, as minor deviations are deemed inconsequential. Taxpayers are stratified into three groups based on EOI: **Group 1** includes those with predicted values greater than declared values, where discrepancies exceed the MAE threshold. **Group 2** comprises taxpayers whose predicted values approximately match their declared values, with variances within the range of -MAE to +MAE. **Group 3** consists of taxpayers whose declared income exceeds the predicted value.

To refine the selection process and reduce false positives, statistical outlier detection techniques are applied within Group 1. In essence, the method comprises the following steps:

Step 1 Predicting the annual income using supervised learning.

Step 2 Computing the **EOI** by comparing predicted and actual values.

Step 3 Selecting taxpayers with EOI > MAE.

Step 4 Identifying outliers within this group and ranking them by EOI.

The model serves as a ”selection rule”to generate a preliminary list of taxpayers ordered by estimated omission income (EOI). This list prioritizes taxpayers for individual analysis in the next stage of the audit case selection process.

4. Experiment

4.1. Business Understanding

The RFB's Individual Taxpayer Registry holds data on over 200 million Individual Taxpayer Identification Numbers (Cadastro da Pessoa Física – CPF). Despite widespread compliance, the vast dataset poses a challenge in prioritizing economically significant cases.

For proof of concept, this study focused on **individual** taxpayers for the calendar year **2019**, specifically those whose occupation is **”member or public servant of the federal direct administration”**.

The scoping limitation was deemed necessary to reduce the size of the dataset, originally comprising tens of millions, to approximately 400,000 rows, facilitating in-depth analysis of a more homogeneous group. The focus was on taxpayers classified as ”*member or public servant of the federal direct administration*”, as their income data (target variable), reported by the Federal Government, is considered reliable and suitable for the machine learning model.

4.2. Data Understanding

The RFB's taxpayer data repository is vast, comprising thousands of terabytes across multiple databases within a comprehensive datalake. To predict taxpayer annual income, this study focused on variables indicating socio-economic status.[2]

Queries across databases were consolidated using CPF and calendar year (CY), resulting in a unified file with **395,560 rows** and **59 columns**. Each row corresponds to a unique CPF, with key features detailed in Table 1.[3]

Tabela 1. Description of the main features

Feature	**Description**
cd_ocu	Primary occupation as reported in DIRPF.
gender	Gender of the taxpayer
mar_stat	Marital status
age	Age of the taxpayer
qt_dep	Number of dependents reported in DIRPF
qt_comp	Number of companies in which the taxpayer is registered as a shareholder in the QSA (Shareholders and Administrators Registry).
qt_resp	Number of companies where the taxpayer is listed as a responsible party in the QSA.
vl_income	Annual income amount reported in DIRPF (target variable).
vl_asset	Value of assets and rights reported in DIRPF
vl_dirf	Annual income amount reported in DIRF by the withholding entity
vl_fin	Total value of credits in the account as reported by the financial institution(s).
vl_buy	Total value of electronic purchase invoices reported in SPED NF-e.
vl_cred	Value reported in DECRED by the credit card operator, referring to the total credit card transactions carried out by the taxpayer in a specific year.

The first four variables are text type, while the remaining ones are numeric. These

[2] To comply with legal confidentiality, this study used aggregated and/or anonymized data to prevent identifying individual taxpayers. This deidentification does not hinder the study's comprehension, development, or reproducibility.

[3] DIRF – Income Tax Withholding Statement reported by the withholding agent. SPED-NFe – Digital Accounting System for Electronic Invoices. DECRED – Statement of credit card transactions reported by credit card issuers.

variables were selected because they indicate income and/or wealth, and are expected to correlate with the target variable (*vl_income*).

4.3. Data Preparation

The data preparation phase involved several steps to convert the data into a suitable format for modeling. Missing values in key columns were replaced with 0 to indicate the absence of specific events, while rows with missing data in less critical columns were removed. Data transformation included encoding categorical variables using one-hot encoding, normalizing numerical variables to a range between 0 and 1, and discretizing continuous variables into categorical intervals.

Feature engineering created new features such as *vl_var* (*vl_asset* (2019) - *vl_asset* (2018)), *qt_qsa* (qt_comp + qt_resp), *rat_1* (ratio between *vl_fin* and *vl_dirf*), and *rat_2* (ratio between *vl_var* and *vl_dirf*) to enhance predictive accuracy. Outliers were managed using a clipping technique, capping values beyond the 99th percentile to maintain data structure and improve model performance. The final dataset contained **390,165** records.

4.4. Modeling

To tackle the presented challenge, we employed regression methods to predict the dependent variable (*vl_income*) based on independent variables. After testing several algorithms, we selected Gradient Boosting, implemented by the *xgboost* [4] Python package, as it achieved the highest R² score with default hyperparameters (Linear regression: 0.68, Decision tree: 0.77, Random Forest: 0.80, xgboost: 0.82).

4.4.1. Metrics

For a comprehensive assessment of the model's performance, we employed two metrics: R² Score and MAE.

The **R² Score** measures the proportion of variability in the dependent variable explained by the model, ranging from 0 (no explanation) to 1 (perfect fit). The **MAE** (Mean Absolute Error) calculates the average absolute differences between predictions and actual values, offering an easily interpretable measure in the same units as the target variable (*vl_income*).

4.4.2. Baseline

To evaluate the model effectively, establishing a baseline parameter for comparison was crucial. This baseline acts as a naive guess against which the model's results can be measured. Given the lack of similar works predicting income, we considered the median value of the target variable as a reasonable baseline. Using the median value of 179,604.32 for all instances in the dataset, the resulting **MAE** is **122,220.36**, and the **R² Score** is **-0.05**. The supervised model is expected to significantly improve upon these metrics.

[4] [Chen and Guestrin 2016]

5. Results

The dataset was divided into 90% for the training set and 10% for the test set. Cross-validation was performed on the training set using the k-fold method with k = 5 to ensure robust performance evaluation. Table 2 presents the results achieved by the proposed model across the selected metrics

Tabela 2. Performance metrics results

Metric	Train	Validation	Test
MAE	23,432.17	26,832.37	26,839.25
R² Score	0.91	0.84	0.84

The model demonstrated excellent forecasting ability with an R² score of 0.84 on the test set, indicating strong generalization performance. The model's MAE was **R$ 26,839.25**, significantly lower than the baseline MAE of **R$ 122,297.75**, showing an approximately fivefold improvement in precision. This substantial difference highlights the model's effectiveness in producing more accurate and reliable estimates.

The graph in Figure 1 illustrates how taxpayers were classified by EOI according to the criteria described in Section 3.

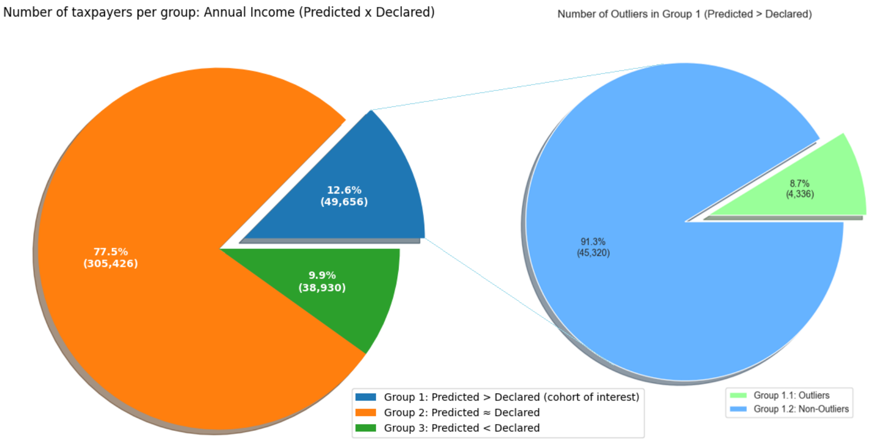

Figura 1. Number of taxpayers per group and number of outliers in Group 1

The identified Group 1, indicating signs of income omission, constitutes only **12.6%** of the initial taxpayer population, reducing the dataset to **49,656** taxpayers for further analysis. The next step is to identify outliers within Group 1 using the **boxplot** method, which reveals outliers as points above the upper whisker, calculated as $Q3 + 1.5 * IQR$. This analysis shows several outliers with a threshold of **R$ 117,200.31**.

Group 1 was further divided into two subgroups: **Group 1.1 (outliers)** with EOI greater than 117,200.31, and **Group 1.2 (non-outliers)** with EOI less than or equal to 117,200.31. This division refined the initial list, resulting in the pre-selection of **4,336** taxpayers in Group 1.1, which constitutes about **1%** of the initial dataset.

The key summary statistics for the Estimated Omission Income (EOI) of taxpayers in Group 1.1 show a maximum EOI of R$ 1,248,688.00, a mean of R$ 190,160.54, a total sum of R$ 872,236,297.69, indicating a substantial potential financial impact.

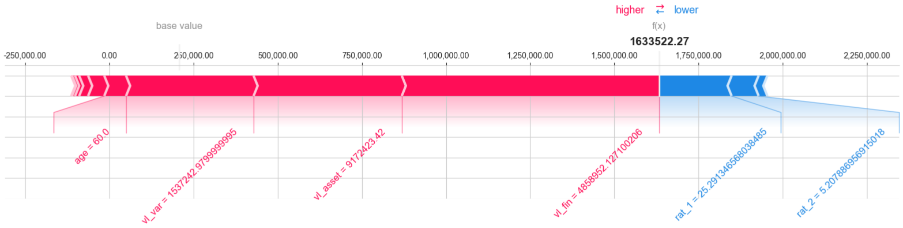

Figura 2. Local Explainability: Red color indicates features that increased the prediction, while blue color suggest features that decreased the prediction

The ultimate result of the proposed audit case selection process is a descending ordered list based on EOI values from taxpayers in Group 1.1, which represents the cohort with the highest risk of tax evasion and, hence, should be prioritized for further scrutiny.

5.1. Explainability

Explainable Artificial Intelligence (**XAI**) enhances the transparency and interpretability of AI models, which is crucial for fairness, compliance, and trust in taxpayer audit selections. XAI helps to understand and justify ML decisions, fostering responsible AI use in tax administration.

In this study, we used *SHAP (SHapley Additive exPlanations)* [5], based on cooperative game theory, to measure feature importance for each prediction. The SHAP summary plot illustrates how each feature influences the model's output, aiding in understanding the model's behavior and providing insights into specific factors influencing individual predictions, highlighting atypical behavior and potential risk factors.

For example, Figure 2 shows the SHAP values for the highest Estimated Omission Income (EOI) instance. The features *vl_fin* (R$ 4,858,952.13), *vl_asset* (R$ 9,172,423.42), and *vl_var* (R$ 1,537,242.98) significantly impacted the prediction of R$ 1,633,522.27, compared to the reported value of R$ 384,834.27, resulting in an EOI of R$ 1,248,688.00.

6. Conclusion

This study proposes a novel approach to enhance audit case selection in the Brazilian Tax Administration by applying supervised learning and data mining techniques to a real-world dataset of taxpayers' socioeconomic situations. The use of advanced Machine Learning and *eXplainable AI* not only improves data analysis but also addresses legal and ethical considerations, paving the way for future research and audit improvements.

Our approach has the potential to revolutionize the RFB's taxpayer selection process by integrating AI into the traditional rule-based system, leveraging big data. This research also enriches academic literature on AI's role in combating tax fraud and evasion, filling a crucial gap.

Future research could include: individual analyses of Group 1.1 to validate income omission indications; testing the model on taxpayers with different occupations; incorporating new independent variables from RFB's databases to improve predictive capacity; creating derived features; and employing unsupervised outlier detection algorithms to compare results.

[5] [Lundberg and Lee 2017]

Referências

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

da Silva L. S., de C., R. H., N., C. R., and F, S. J. C. (2016). Bayesian networks on income tax audit selection —a case study of brazilian tax administration. In *Bayesian Modeling Application Workshop (BMAW).*

de Roux, D., Perez, B., Moreno, A., Villamil, M. D. P., and Figueroa, F. (2018). 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 (KDD '18)*, pages 215–222.

Lin, Y. et al. (2021). Taxthemis: Interactive mining and exploration of suspicious tax evasion groups. *IEEE Transactions on Visualization & Computer Graphics*, 27(02):849–859.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

OECD (2017). Tax audits in a changing environment. In *The Changing Tax Compliance Environment and the Role of Audit*, pages 72–77. OECD Publishing, Paris.

Savić, M. et al. (2021). Tax evasion risk management using a hybrid unsupervised outlier detection method. https://arxiv.org/pdf/2103.01033.pdf.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. pages 29–39.

Xavier, O. et al. (2022). Tax evasion identification using open data and artificial intelligence. *Revista de Administração Pública*, 56:426–440.

Zumaya, M. et al. (2021). *Identifying Tax Evasion in Mexico with Tools from Network Science and Machine Learning*. Springer.