Deep Science Publishing
https://doi.org/10.70593/978-93-49910-54-6_2

DeepScience
Open Access Books

Chapter 2: Leveraging predictive analytics and machine learning for advanced fraud detection in tax systems

2.1 Introduction

Tax fraud is a global problem of serious dimensions, negatively affecting the economic development of different countries, both developing and developed. Its implications go beyond the public finance sector and also impact social behavior. Tax fraud encompasses higher tax rates to compensate for the loss of revenue and may induce psychological distortions in payment decisions, as citizens feel cheated when they pay their taxes and see that somebody else did not pay theirs. There is an exhaustive literature dealing with tax fraud, covering both theoretical and empirical approaches. Theoretical models focus on the incentives to evade taxes and build diverse motives favoring or objecting to tax fraud. Empirical approaches intend to validate or invalidate theoretical models, assessing the relative weight of the different factors invoked to explain tax fraud decisions or the correlations amongst those variables. Those analyses focus on different countries, with different degrees of development and with different characteristics. In the last few years, there have been several changes in population behavior regarding tax fraud. In some countries, tax fraud has not gained relevance, whereas, in other places, it has broadened considerably or contains alarming projections to broaden. Money laundering and digital currencies have undoubtedly favored, to some extent, tax fraud, and it cannot be neglected that it may have gained relevance during the pandemic. In this sense, income taxes and tax on the transfer of assets are the ones that most suffer tax fraud. Security Economics states that there exist two types of fraud incentives: those concerning the likelihood of detection versus the fines to be paid for fraudulence; and those regarding the probability of audit after filing the declaration. Moreover, decisions such as when to file the declaration and register fraudulent deductions also favor fraudulent activities. The growing technological and digital transformation that characterizes current society allows for the exploration of variables that favor tax fraud detection. In this sense, different tools associated with machine learning and predictive analytics may contribute to exploiting and analyzing those variables so that nowcasting and forecasting can be performed with higher predictive power (Chen et al., 2004; Nuseir, 2020; Hamid et al., 2021).

2.1.1. Overview of Tax Fraud Implications

To first assess the feelings of taxpayers and markets towards such an appropriate use of predictive systems, we first introduce and remind some of the most significant and basic concepts regarding fiscal policy, tax systems, and tax fraud. Each country develops and implements tax systems that directly affect how citizens behave and interact with markets. Once implemented, planned, and carried on with carefulness, these systems guarantee resources to governments, and so taxes have the function of financing public expenditure, essential to maintaining adequate levels of education, health, and infrastructure. The spending policies that governments implement should promote factors that lead to sustainable and lasting economic growth. However, different tax systems imply different impacts on market behavior and ultimately on the economy. As known, among voluntary payment and evasion, a taxpayer can choose to deflect from the prescribed conduct. In other words, tax systems should stimulate voluntary compliance. In doing so, they should guarantee low rates, not too-complicated procedures, and affordable costs in carrying out compliance. Enforcement must be fair to intimidate dishonest taxpayers. These simple rules are needed to avoid hampering economic growth, as they foster trust in the government, affecting the underground economy.

Normally, not all tax obligations are satisfied, just as it is known that tax evasion is a problem common to all nations and growing with the economy's overall involvement with markets. As known, tax fraud lowers potential economic growth, but also creates inequities, distortions, and economic problems. Most developed countries know that their advanced economies and growing dependence upon global markets impose new burdens on taxpayer compliance. Because tax and revenue system functions and operations have become more complex and the costs of non-compliance are higher than ever, an increasing number of tax and revenue authorities now consider non-compliance management as a priority.

2.2. Understanding Tax Fraud

Taxation is a source of Government revenue through compulsory contributions levied on individuals and organizations. There are different types of taxes such as income taxes, excise taxes, property taxes, sales taxes, and others. The government uses tax revenue for various purposes such as salary payment, infrastructure development, transfer payments, and interest on public debt. Taxes can be classified as direct and indirect; with direct taxes being personal income tax and corporate income tax, while all other taxes are classified as indirect taxes. Tax roles are the basis for collecting an individual's or organization's taxes. This information is presented to tax agencies in the form of tax returns (Wang et al., 2020; Purohit & Singhal, 2022).

Tax fraud is considered fraud that is committed to avoid taxation. It describes the illegal acts of omitting income, inflating expenses, inappropriately claiming deductions, and failing to file. This act of deception may be conducted by either an organization or an individual. It refers to situations in which taxpayers report false information on their income tax return to deceive tax authorities. Tax laws impose heavy penalties on people who evade taxes. Tax fraud typically involves the underpayment and underreporting of income sources, such as interest income, dividends, rental, and sales of property.

Tax evasion refers to the voluntary, unlawful act of failing to pay taxes that are due to reduce tax liability. It is achieved by concealing relevant financial information from tax authorities responsible for collecting taxes. The critical elements of tax evasion include the willful attempt to defeat the tax law to pay less tax than is required. In contrast, tax avoidance refers to the acceptable use of tax rules to reduce a person's tax liabilities. While tax evasion is illegal, tax avoidance is not.

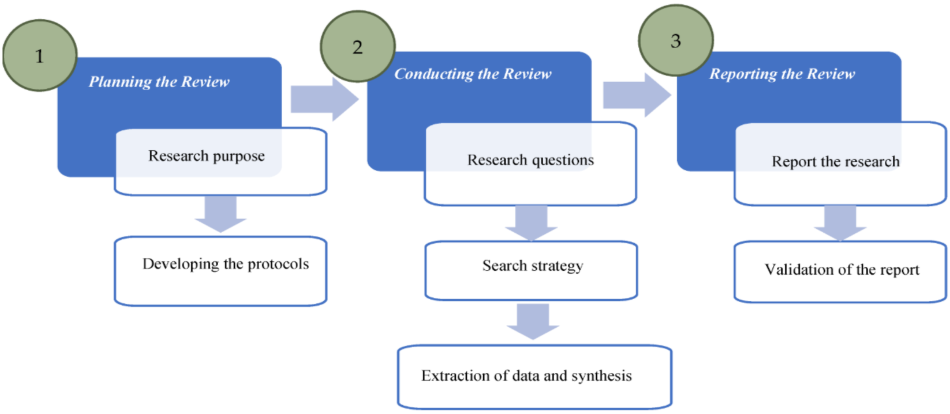

Fig 2 . 1 : Financial Fraud Detection Based on Machine Learning

2.2.1. Types of Tax Fraud

Taxation has played an important role in the operation of public authorities by supplying the resources necessary for public expenditure. There are three essential functions, which taxes fulfill: the supply of revenues, the distribution of wealth, and the regulation of the economy. However, all societies also face the fact of tax fraud, which undermines the role of taxation in economic and social life: at its most general level, tax fraud may be defined as the failure to comply, completely or partially, with the promises made to the state concerning the payment of taxes. The consequences of tax fraud are two: the diminishing of resources necessary for public expenditure and the distortion of conditions of competition in favor of those who evade or avoid taxes.

Different types of tax fraud can be found. The first distinction can be made between tax evasion and tax avoidance. The most common definition of tax evasion refers to all illegal activities using which a taxpayer obtains a tax saving; it is the cheating that costs other taxpayers money. Tax evaders misrepresent their financial situation to the tax authority. They fail to declare income, usually from offshore accounts, and they underestimate expenses, claiming tax deductions to which they are not entitled. Tax avoidance, on the other hand, is defined as the reduction of tax liabilities using legal arrangements. It is a defining principle of every tax system that an individual is entitled to minimize his tax liability, so long as he acts within the law. Tax avoidance is primarily detrimental to governments because it reduces their revenues. It affects tax policy decisions by governments because the potential presence of tax avoidance will usually be a further consideration in deciding how high to set tax rates, or how generous tax concessions should be.

2.2.2. Impact of Tax Fraud on Revenue

Fraud or violation of tax laws in any country adversely impacts the government's capability to provide essential services to citizens. The government may not be able to invest and build infrastructure or execute social welfare programs for the upliftment of the poor and downtrodden if the collection of tax revenue is not on a proper footing. Tax fraud has been a serious research problem and is also an ongoing issue for tax authorities across the globe. The revenue losses due to fraudulent tax returns not only tax authorities but also common taxpayers. Countering tax fraud is something that all responsible citizens of the nation would like to support. We have also expressed strong support for countering tax fraud in several public speeches. Tax fraud is widely understood but often overlooked. Tax fraud is perceived to be an event happening in small-scale businesses. The truth is that tax fraud is happening in the biggest conglomerates in the world. Large tax frauds cheat tax authorities in billions of dollars. Then come professionals both tax preparers and return filers. These tax preparers prepare false returns to get refunds for their clients. As a result, limits have been put on claiming and also impose penalties on tax preparers with the intent of reducing violations. Taxpayers who intentionally fail to report all of their profit are considered to be engaging in tax fraud and are subject to penalties. In addition, individuals, corporations and partnerships also occasionally commit tax fraud. Some corporations provide false or misleading financial statements in their attempts to reduce their tax liabilities or to hide profits. Partnerships may also engage in tax fraud by filing misleading or false returns or by failing to file a tax return altogether.

2.3. Predictive Analytics in Tax Systems

Tax administrations are increasingly aware of the fact that the development of digital innovation offers them great opportunities to modify their functioning and improve their performance and impact. Predictive analytics, which includes predictive modeling, machine learning, and pattern recognition, is seen as one of the most important areas of development for tax administrations in the future. However, it is also recognized as being complicated, highly technical, and demanding many advanced skills and expertise that tax administrations may lack. Moreover, the tools of predictive analytics also need to be integrated into existing processes and practices, which could take time and effort. In addition to these technical challenges, questions can be raised about if and how the use of predictive analytics can modify or transform the existing tax culture. More broadly, the impact of predictive analytics on the relationship between taxpayers and tax administrations also needs to be discussed, since predictive analytics may increasingly help to detect fraud, corruption, and non-compliance with the law, resulting in sanctions against some taxpayers based on their behavior and profile.

Predictive modeling is one of the core areas of predictive analytics, which attempts to identify the real intention behind a person’s actions by creating a model that describes the relationship between certain risk indicators and the decision a taxpayer took, whether legal or illegal. The aim is thus to provide a tool to enhance the tax administration’s capability to arbitrarily discriminate between legal and illegal behavior by taxpayers. Identifying a fraudulent taxpayer’s profile is complicated, as it may be composed of disparate and sometimes contradictory elements. We define predictive tax fraud modeling as the set of processes that allows tax administrations to study past tax fraud cases to create an empirically grounded model that can be used to assess the probability of the recurrence of tax fraud in a relevant pool of taxpayers in the future.

2.3.1. Definition and Importance

In today's tax systems, tax agencies encounter sizable challenges when estimating the demands and needs of various taxpayers in terms of risk in their compliance and revenue collection. Predictive analytics is an emerging approach for data analytics that is powerful, but easy to understand. Predictive analytics transform access to and the perception of risk, by focusing on the analysis of observable aspects of tax compliance behavior, and optimizing the allocation of limited resources for detection and intervention against that predictive risk. Predictive analytics help in modeling a clear picture of who these tax-payers likely are, which data is useful and when it should be collected; patterns, trends, relationships, and feedback that reveal taxpayer intentions and behavior; and the impact that the application of limited resources will have on compliance.

Tax agencies need to assess the demand and need for predictive analytics-based solutions in their environment, before embarking on a journey that may lead to misaligned investments, resulting in inadequate returns and unnecessary waste of taxpayer resources. This involves exploring what predictive analytics can do to improve tax compliance policies and revenue, as well as gaining insights into their demographics, operational environment, tax structure, compliance gaps, stakeholder needs, resource constraints, and likely risk profiles going forward. Presenting understandings from both of these domains, in a clear, audience-understandable way, is likely to lead to synchronizing perceived needs, and educating stakeholders so that they have an aligned interest in supporting investments in predictive modeling.

2.3.2. Historical Context of Predictive Analytics

The use of "predictive analytics" may appear to be an innovation, a hallmark of contemporary high-tech environments, but in fact, the combination of predictive and prescriptive analytics has a long history, especially in government. Kept at arm's length until the last decade by risk-averse organizations, predictive analytics solutions are used throughout government, in areas such as public health, weather forecasting, terrorist and criminal threat detection, and pension, unemployment, and disability fraud detection. Indeed, analytics covering the full spectrum have been in use for more than five decades to identify specific long-standing criteria for the detection of tax and customs cases. This approach—"the traditional method"—has relied on years, if not decades, of institutional knowledge about the "characteristics" of individuals and organizations who might evade taxes or commit customs fraud in the overall taxpayer/unemployed/pensioner/citizen pool. Predictive analytics uses data and statistical algorithms to forecast the future.

For at least the last two decades, statistical models have been used to score returns for audit selection that contain an undue understatement of income, which models have been updated approximately every three years for use by outside contractors. The traditional "audit selection model" identifies the returns most likely to be noncompliant and selects them for review. Other tax agencies also apply similar internal algorithms to select tax returns for examination. These audit selection models have influenced the timing of tax return audits. The increased availability of taxpayer data complicates relationships; tax officials are aware that what taxpayers share with agencies plays a key role in compliance levels.

2.4. Machine Learning Fundamentals

1. Machine Learning Machine Learning (ML) is a sub-area of Artificial Intelligence that allows the development of statistical models capable of performing some cognitive tasks on behalf of humans. These tasks are designed based on the analysis of examples and the learned model's predictions, that is, without being explicitly programmed to do so, which gives ML the ability to adapt to new tasks and new environments. The ML internal mechanism is based on the use of algorithms that still require to be configured and trained to achieve a desired level of performance. ML has achieved noteworthy success across multiple domains leading it to be used for real-world applications such as recommendation systems, video surveillance, speech recognition, medical diagnosis, security, gaming, advertising, robotics, and so forth. Notably, even though ML techniques have been available for decades, they have gained renewed interest recently, mainly as a result of the availability of big amounts of data, the development of new algorithms, and the emergence of the capacity for processing big amounts of data.

2. Types of Machine Learning Techniques ML techniques can be classified into two classes, without implying that the division is always clear. The first class of techniques is called supervised if they make use of a training dataset in which the examples are composed of the predictor variables together with a target variable that represents the desired value that the ML algorithm has to predict for unseen examples. The second class of techniques is called unsupervised if they make use of a training dataset that only includes predictor variables to understand the structure of the dataset. Both classes of techniques can be applied to different kinds of specific tasks: classification (for supervised techniques), regression (for supervised techniques), clustering (for unsupervised techniques), and anomaly detection (for supervised and unsupervised techniques).

2.4.1. Overview of Machine Learning

Machine learning can be described as the study that aims to provide computers with the ability to learn and adapt, through experience, to perform a specific task without being repeatedly programmed for it. Within this context, learning is a process that modifies data structures of interest and is based on the identification of patterns in data. Therefore, it is acknowledged that a key part of the learning process is the experience, understood as the collection of examples that enables the identification of data patterns. The task that computers perform by learning is often defined in terms of an objective function.

An objective function is a measure of the performance of a particular task of interest that reflects its specific characteristics.

More formally, machine learning specifically refers to a subset of artificial intelligence (AI) focusing on the development of methods that enable computers to learn while acting in the world and to be able to incorporate the knowledge acquired by experience to solve new but related challenges in the future. AI focuses on endowing a machine with intelligent features. This is done by creating systems that can emulate certain functions of the human brain related to processing sensory data and making decisions in real-time. Additionally, to be considered intelligent systems, AI systems have to display “generalization capabilities”. This means that the response back to certain sensory input, which could constitute the learning of that specific input pattern, must be applied to new yet similar situations, allowing the provision of a timely, meaningful, and automatically determined response.

2.4.2. Types of Machine Learning Techniques

Machine learning algorithms can be divided into two different categories: supervised learning and unsupervised learning. Supervised learning techniques infer a function from a set of labeled training data, mapping input variables to an output variable, which can be used for prediction. Those functions can be simple linear functions, or more complicated functions, where the map is non-linear. Traditional supervised learning techniques utilize deep or supervised networks or Bayesian learning algorithms, such as Logistic Regression, Bayesian Neural Networks, and Bayesian Support Vector Classifiers. These algorithms are statistical approaches that learn a linear or non-linear decision surface. Supervised learning algorithms are often used for classification tasks, as neural networks are commonly utilized in situations where the relationship between the input data and classification is highly nonlinear. Unsupervised learning techniques, on the other hand, infer functions from inputs alone and are employed if no training labels are given. The approach is to formulate the problem in such a way that the unsupervised machine-learning algorithm can generate its training predictions. For instance, a clustering task has as its output a function mapping every input into a unique label corresponding to a cluster. Clustering is the most common unsupervised learning task. Statue clustering can also build on supervised clustering, iteratively refining the labels to optimize the supervised performance metric.

When a problem is in the supervised-to-supervised or supervised-to-void labeling paradigm, the appropriate algorithm is traditionally supervised. The latter is a more natural approach to predictive classification problems and is also generally more efficient than unsupervised learning. However, traditional supervised learning suffers from two main drawbacks: it requires a large amount of labeled data, which may be very expensive to obtain, and it does not allow transfer learning. Supervised machine learning aims to expedite data labeling and provides more information relating to different areas.

2.5. Data Sources for Fraud Detection

The challenge of tax fraud detection is compounded by the diverse, decentralized, and often fragmented nature of tax entities and their underlying systems. Consequently, meaningful and effective prediction of fraud in tax systems necessitates access to a rich array of data sources, encompassing both internal data and external data related customarily to the entities being taxed, but sourced externally from multiple third party public and confidential data sources. These are discussed below.

Internal data sources are historical tax data relevant to a predicting entity, used for model training for tax prediction. Internal data pertinent to tax fraud detection can be divided broadly into two classes as follows: a) Time-dependent Internal Data Sources – For tax fraud detection, these data elements relate typically to the transactions involving the taxpayer during a tax compliance cycle, as well as other qualitative historical data about the taxpayer. Concerning business taxpayers, these can include: Trademark registration information, Company information at Taxpayer Identification Number (TIN) (Company name, tax type, status, opening date, expiration date, cancellation date, date of Tax Registration Certificate, address, phone number, fax, issued TRC number for Company TIN, number of employees during operation, number of employees declared to Social Insurance Agency), business plan, borrowing credit transactions, Company financial statement, Tax Administration Policy (Number of years of registered tax debts), Labor Management Policy, records of suspension/reinstatement of business activities, complaints related to Company business; b) Time-Independent Internal Data Sources – These data elements can be acquired from tax authorities and allocated for each predicting entity. They are indicative of the entity’s public fraud propensity and highlight the common traits of fraudulent entities (for example, age of tax registered business, time in tax debts, time of tax outstanding declaration).

2.5.1. Internal Data Sources

Next to the algorithms chosen, an important aspect of successful fraud detection is the data used for model development. Since large parts of the data generation process are not available for observation or are not fully reliable, and the vast number of different forms and solutions are possible, developing appropriate proxies for missing alternative data is not trivial. Therefore, if you have large amounts of internal agency data available, the preferred approach is to start model development using this data. Since the fraud methods in a country do not suddenly change from one day to the next, the best prediction models are those based on this country's data. As a country’s tax reporting and compliance system matures, internal data should become the main source of available data, and external data can function to augment it.

The kinds of internal data that can be used to build anti-fraud models in the area of direct taxes include tax returns, audits, and third-party reporting systems. In tax returns, together with the financial data for the tax year under review, auxiliary data is filled out, such as information on sources of income, business structure, assets, liabilities, etc. For a unique taxpayer number, tax returns for the last years are on file. In characteristics for other fiscal years, for which tax returns are available too, one can find both qualitative and quantitative data. Commercial databases have financial data on numerous businesses, but these databases are costly and often do not contain the whole population. The information contained in tax returns is of utmost relevance for the detection of tax fraud, especially when done recurrently.

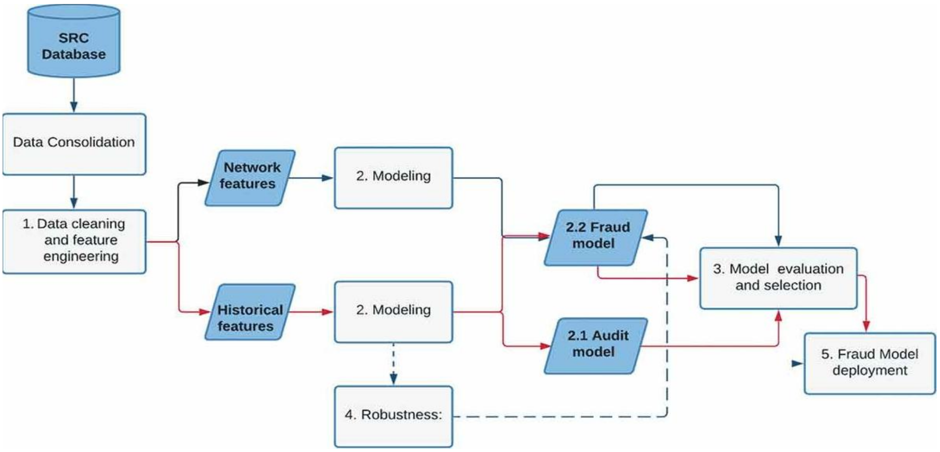

Fig 2 . 2 : Improving Tax Audit Efficiency Using Machine Learning

2.5.2. External Data Sources

Tax administrators have long recognized that many data sources outside the tax systems can point to potential tax fraud and non-compliance. Researchers have used all sorts of external data, such as rainfall patterns, palm oil prices, bank credit reports, and income estimates from satellite images of house conditions and the quality of cars parked in them, as well as estimates of energy costs and purchases, to detect potential fraud and non-compliance in tax systems. All these data are often not collected or available to tax administrations. These data are instead collected privately by firms, banks, satellite companies, and energy consumption monitoring firms for their business purposes.

We agree that so-called "big data" is uniquely available outside tax administrations with firm and commercial databases than those sources traditionally circulated among tax administrations, such as management booklets, accounting procedures, and business guidelines for compliance materials. Fortunately, we are at the threshold of making more and better use of the big data revolution in our fight against tax fraud and non-compliance. And while using that data requires the agreement of the private owners, which is not always easy to find and coordinate, it likewise does not cost tax administrations anything. It need not even extract the data from the outsourcer. When well-designed, the contractor usually can automatically extract the data every day, week, or month.

Using big data raises special challenges that will complicate matters for tax administrations. Accessing, cleaning, and validating data from a multitude of external sources can be time-consuming, costly, and fraught with errors, especially if the data stream is noisy, inconsistent, and disorganized. Such econometric considerations remain paramount for tax administrations in deriving benefits from their investments in external databases and have often put them off.

2.6. Building Predictive Models

The last part of the predictive modeling phase relies on feature selection and model training and validation. The latter comprises the process in which we define, train, and validate predictive models to achieve the results contained in the section.

As detailed in the section, different statistical, econometric, and machine learning techniques may be applied to create specific tax analytical models to reach specific goals. The decision tree is used as a starting point for model training and validation. It provides a set of if/then rules, which, compared with surrogate models, are easy to interpret and visualize, allowing a logical analysis of the model and scrutiny of its predictions. A few techniques applied to various datasets are detailed in the Results section.

In the choice of the set of features for training, it is important to avoid over-fitting: the model fits well to the training data, but it loses predictive power for unseen data. To reduce overfitting, we apply the following practices. First, in the training and testing splits of the data, we do not perform stratified sampling, allowing flexible partitions of negative and positive datasets, in the training and validation datasets. By balancing false positives and false negatives during validation, we can avoid excessive compartmentalization of the model performances on one side or the other. The objective is to test the model performance over a wide range of different balancing thresholds. Second, instead of the straightforward simple split of the data in training and test subsets based on a fixed percentage of the complete dataset, we assess the models using crossvalidation. The data is split up into training and testing sub-groups multiple times and the overall average validation is retained.

2.6.1. Feature Selection

The predictive power of a model is inextricably linked to its underlying features. In supervised learning problems, when the output value is stipulated, there is a probability of an inverse relationship being formed between the predictor and the response. Models fueled with relevant features built with the right level of abstraction can better explain, and thus better predict, the behavior of the response variable with less mean squared error. Conversely, models bolstered by irrelevant features or noise will suffer unforeseen problems during operationalization, a fate such models are often condemned to. Due to the large volume and wide availability of structured, semi-structured, and unstructured data in today's world, as well as the vast number of predictors typically involved in modern modeling problems, careful feature selection before model building is paramount for the performance of the end model. Feature selection, or dimensionality reduction, is the process of selecting a subset of relevant features from a pool of potentially redundant, irrelevant, or noisy predictors. Feature selection is driven by practical considerations, as the resulting models are typically easier to interpret, computationally inexpensive, less prone to overfitting, and more efficient for both prediction and theory building. The multicollinearity problem of high-dimensional data might lead to spurious regression results when the response variable is continuous. This is especially true in cases of small sample sizes, where shrinkage methods are often employed instead of dimension reduction. In tax fraud studies, observer subjectivity often comes into play when selecting predictors.

2.6.2. Model Training and Validation

For building the predictive model, data is split into the training and test sets, where the model is fitted using the training data, and the predictions are computed using the test data. Further, the performance of the model is evaluated using the prediction error metrics. Majority of the predictive models rely on historical data to learn the correlation between predictor variables and target variables, through a process referred to as model training. Training a predictive modeling involves implementing and optimizing a prediction algorithm based on training data.

The quality of the fitted model is crucial for successful predictive analytics and machine learning implementations. Essentially, the better trained a model is, the better its predictions will be on the data it has not seen before. If a model cannot make reliable predictions, machine learning is not helping achieve the end goal. Most statistical learning algorithms are designed to minimize prediction errors on the training data through some technique. Choosing a prediction algorithm based only on the prediction accuracy in the training data will yield models that generalize poorly on data that was not used for training. This is because the model gets tailored too closely to the idiosyncrasies of the training data rather than the underlying association between predictors and the response. As a result, these models will make grossly inaccurate predictions on the new test data or any new data. To reduce the overfitting risk, it is recommended to implement methods such as justifying the model, using a large training data, mixing up training strategy, and others, which are discussed in the following section. In the next step, the predictive power of the models is measured using the test data, prediction error metrics, and receiver operating characteristic curves, without any tuning.

2.7. Algorithms for Fraud Detection

Fraud detection is an essential component of many areas in today's economy. In general terms, fraud is an activity that misuses some type of asset. Fraud in the area of taxation occurs when taxpayers are dishonest in their reporting to tax authorities, resulting in a loss of revenue that those authorities depend upon for ongoing government operations. The cost of dealing with fraud in tax systems comes from two areas: first is the cost of implementing policy designed to prevent fraud from happening, which may be as simple as systems that look for unusual financial events on tax filings, to procedures of tax authorities regarding the processing of fraud cases once they are discovered. The second cost area flows from the loss of revenue from the fraudulent activity, and this loss may be substantial. This section presents several approaches to the use of fraud prediction models based on historical patterns of accepted and rejected tax filings.

A machine learning framework composes predictive modeling as the basic technique for the further probabilistic prediction of tax fraud given a set of historically resolved records. A second component of the framework involves using unsupervised detection models that are capable of predicting unknown fraud cases so that suspicious newly detected records can be sent for further investigative action together with the reported predicted probabilities of being involved in tax fraud. The particular algorithms for predictive fraud modeling should be selected from the supervised and unsupervised machine learning algorithms based on their suitability for the tax fraud detection context. A specific supervised algorithm that caters to the special attributes of tax fraud detection is decision tree induction using decision trees.

2.7.1. Supervised Learning Algorithms

Many types of supervised learning algorithms exist. Generally, they apply to classification or regression problems. In classification problems, the goal is to predict a discrete label for each sample. In regression problems, the goal is to predict a numeric value for the samples. The two problems can be seen as different, motivating researchers to customize techniques. On the other hand, we can also see them as two sides of the same coin, using the same inputs. Many algorithms can handle both types of problems. So to give a more general overview, we split algorithms into probabilistic and discriminative. In the case of a small number of examples, the most reliable learning strategy consists of using the model that is most specific to the family of distributions.

Many models are generative models, which means they represent the distribution of examples conditional to their class. These models have the advantage that learning is simple when estimating model parameters. However, to get percentages or probabilities of class membership in classification tasks, it is necessary to make a validation set containing a sufficient number of examples for every possible output class. This is an issue, in case some classes have few examples. For instance, in tax fraud detection, taxpayers' accounts are the majority class, which includes many characteristic examples of benign behavior. Thus, the prediction is based on the estimated distribution of the conditional probability of the class given an input example. The previous solution requires a validation set large enough. It seems unlikely that a validation set dedicated to estimating the conditional distribution of the class would contain enough examples.

2.7.2. Unsupervised Learning Algorithms

In contrast to supervised learning methods, unsupervised learning models do not determine predictions via functions that best relate training inputs to the desired outputs. Given that most fraud measurement and evaluation problems could lack information in terms of either labels indicating fraudulent versus non-fraudulent examples or costs in the sense of fraud losses, VAT and corporate tax systems are good examples for such instances, unsupervised learning techniques are promising – and indeed necessary – for fraud detection applications. Provided that the tax system data for corporate tax and VAT represent cumulative transactions and account entries over the complete financial year, in addition to tax monitoring over the tax year, for both tax types, the risk of false positives is decisively low for fraud detection. That is, the model outcome provides yes-no answers for fraud payments.

In terms of recent research that relies beyond anomaly detection methods, for both VAT and corporate tax fraud monitoring purposes, the relative literature for unsupervised learning algorithms is limited. For the sole VAT fraud type, e.g., destination-based fraud, a clustering technique for credit card transaction data is proposed. Employing a heuristic evaluation, the approach relies on clustering and the elbow method. In the domain of corporate tax fraud, the problem of low fraudulent rates is addressed. The major drawback, however, is the insensitivity of cluster methods in a similarity index definition and choice of variables, and that fraud-dependent clusters may not exist. Hence, closeness to fraud activities can also differ in companies other than fraud ones.

2.7.3. Anomaly Detection Techniques

Different tax fraud detection methods can be classified into three main techniques: anomaly detection, supervised classification, and social network analysis. These three techniques provide a different angle to the fraud detection problem. Each offers both advantages and disadvantages, but real-world applications usually combine at least two of them.

Anomaly detection techniques use a set of normal samples to characterize the allowed behavior – for example, in the form of a distribution – and identify anomalous samples differing by a sufficient margin. The margins can be chosen according to different types of anomaly scores, for example, density differences or Mahalanobis distances. Such techniques regard fraud as rare behavior and do not need examples of fraud.

Anomaly detection offers the advantage of being unsupervised, therefore usually requires less data and generally relies on less biased assumptions than supervised learning. Multi-dimensional data are common in tax-deductible expenses or balanced sheets of law companies, and fraud can also happen in different spaces and times, especially for big companies that operate in multiple countries. Hence, detecting and characterizing anomalies in multi-dimensional, temporal tax data is a challenge that must often be faced by tax administrations. Some anomaly detection methods, invented in statistics, even go back to the nineteenth century. There are many methods for anomaly detection, and they can all be adapted to the task of fraud detection. Also, these methods can be classified according to different criteria: supervised/unsupervised, parametric/non-parametric, and distance-based/statistical/decisional kernel. Each of these methods offers advantages and disadvantages, usually intrinsic in the used assumptions and the specific use case.

2.8. Conclusion

Fraudulent practices are a serious concern due to their negative impact on tax systems, but technology is advancing rapidly. This may open the door for tax authorities, and although they would still need the collaboration of other institutions to overcome legal and confidentiality issues, new fraud detection strategies based on these emerging technologies may be implemented. Advanced techniques not only can expedite any decision related to who must undergo additional validation procedures; they also can generate a priority list to have the best ratio of successful detections to people checked, taking into consideration the resources available, the monitoring frequency, and the possible consequences for the taxpayer as well. The constant monitoring provided by these systems may even allow the monitoring model to be adapted, detecting deviations in taxpayers' behavior in the case of any change.

Tax authorities have implemented e-government initiatives that allow interaction with citizens, making filing income tax returns electronically easier, which facilitates the automatic detection of fraud; however, collaboration with other entities may help improve the detection of fraudulent behavior, taking advantage of the knowledge of these other institutions. For instance, universities could efficiently apply data mining algorithms to detect taxpayers with unusual behavior and thus collaborate with tax authorities by notifying them of any wrongdoing and performing external audits. The model developed may become obsolete as time passes, mainly due to changes in the surrounding social and economic environment, which will affect taxpayers, making it necessary to refresh the model periodically. Furthermore, algorithms that in the past had a good performance may be surpassed as they have been improved with time. These aspects have been faced to a certain extent thanks to ensemble learning; however, more efficient solutions would be welcome.

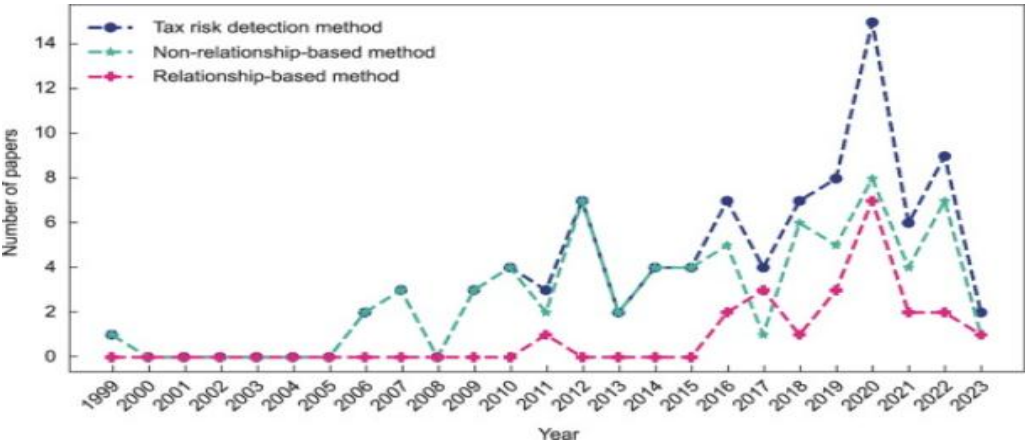

Fig 2 . 3 : Tax Risk Detection Using Data Mining Techniques

2.8.1. Final Thoughts and Future Directions

Detecting and preventing tax fraud is a challenging activity that requires complex models, developed utilizing capabilities of other domains, like technology advancement that employs big data, predictive analytics, machine learning, and artificial intelligence. Data is becoming a significant asset for countries, and taking advantage of data science and data analytics techniques, applied by skilled professionals to process, analyze, and visualize the data to discover insights and extract value, is the way forward to combat tax fraud. When targeting tax fraud detection solutions targeting return prediction, on which most previous research has concentrated, it is imperative that the machine learning solution developed focuses on taxpayer returns associated with a certain risk parameter, for example, fraudulent activity in prior years. This is crucial since most of the returns filed would not have this label; accuracy in the results would be significantly high if all the taxpayer returns were used to train the model target applying strict thresholds for predictive power and many of the taxpayer returns for the risk parameter would belong to other classes. This work describes a generalized data science procedure that is not domain-specific. We present specifically how the different phases are employed in tax systems and preliminary findings from the model and combine the model's shortcomings and the domain challenges to propose future research. Currently only a handful of countries in the world have integrated fraud detection models to inform risk-based decision-making fully, and several other countries are using machine learning spam detection models targeting countries that already employ analytics within their tax structure. We end this work by suggesting guidance to start-up tax systems for detecting fraud to take advantage of machine learning in the tax fraud domain.

References

Hamid, M. A., Ahmad, A., & Sulaiman, S. (2021). A Predictive Analytics Model Using Machine Learning for Tax Fraud Detection. International Journal of Advanced Computer Science and Applications, 12(5), 282–289. https://doi.org/10.14569/IJACSA.2021.0120536

Nuseir, M. T. (2020). The Impact of Artificial Intelligence on Tax Evasion Detection in the Era of Big Data. Journal of Theoretical and Applied Information Technology, 98(20), 4036–4046. https://doi.org/10.5281/zenodo.4280147

Purohit, H., & Singhal, A. (2022). Machine Learning Approaches for Tax Fraud Detection: A Comparative Analysis. Procedia Computer Science, 199, 1020–1027. https://doi.org/10.1016/j.procs.2022.01.131

Chen, H., Chung, W., Xu, J. J., Wang, G., & Qin, Y. (2004). Crime Data Mining: A General Framework and Some Examples. Computer, 37(4), 50–56. https://doi.org/10.1109/MC.2004.1297361

Wang, Y., Wang, J., Ren, J., & Liu, S. (2020). Applying Machine Learning to Detect Fraudulent Transactions in Tax Systems. Journal of Intelligent & Fuzzy Systems, 38(5), 5735–5743. https://doi.org/10.3233/JIFS-179502