Big data analytics in tax fraud detection

Karl Malaszczyk
Holy Family University

Bernice M. Purcell
Holy Family University

ABSTRACT

Big data is the term applied to datasets exceeding the normal confines of traditional database technology. Datasets collected range from all professional fields, including taxation. One use of big data analysis – or analytics – in the taxation field is discovery of tax fraud.

Big data is characterized by the terms volume, velocity, variety, and veracity. The characteristics mean that big data employs large amounts of storage space, gathered from diverse sources, stored in diverse formats, and updated at different intervals. The specific processing used in tax fraud analysis of big data is data mining; the process is now often referred to as analytics. Datamining itself is one step in a larger process referred to by practitioners as knowledge discovery in databases (KDD). Two key groups of datamining tasks employed in fraud discovery are predictive tasks and descriptive tasks. Predictive tasks make a prediction for each observation, whereas descriptive tasks essentially describe the data examined.

Various agencies impose numerous taxes on society, all of which are subject to fraud. Fraud exists in many forms and the Internal Revenue Code defines fraud in several places. Investigators use various methods to detect fraud including direct and indirect procedures. Direct methods include the matching of reported data to information returns received by the Internal Revenue Service. Indirect methods include analytical procedures, review of documents, observation and informants. These traditional methods of finding fraud can be greatly enhanced using analytics.

Key words: Big data, analytics, taxation, tax fraud, fraud analytics

Copyright statement: Authors retain the copyright to the manuscripts published in AABRI journals. Please see the AABRI Copyright Policy at http://www.aabri.com/copyright.html

INTRODUCTION

Big data is a buzzword that has infiltrated every sector of business, including taxation. Emerging from the computing field, big data is the term applied to datasets that exceed the normal confines of traditional database technology. The datasets collected range from all fields – healthcare, manufacturing, retail, and the public sector (Manyika, 2011). The area of finance and accounting is also affected by the big data revolution, including taxation. Fraud is a significant problem in today's world especially in the area of taxation. The Internal Revenue Code defines various types of fraud but does not address the issue of finding it. There are several means of discovering fraud including both direct and indirect means of detection. It is essential to gather evidence to prove the fraud perpetrated. The quality and quantity of this evidence should be sufficient to meet the investigator's goals. Analytics can be used to aid the tax practitioner, accounting firm and governmental unit trying to discover the fraud.

BIG DATA AND ANALYTICS

The characteristics of big data lead to the need for new storage and analysis techniques. The three most commonly described characteristics of big data are volume, velocity, and variety, with some researchers adding a fourth characteristic, veracity (Mills, 2012). Volume refers to the amount of data collected from myriad devices. Velocity is the speed with which the data can change. Variety references to the sources of data (mobile, social media, videos, chart, genomics, and sensors) as well as format (structured data, unstructured data, semi-structured data, and complex data). Veracity indicates the quality of the data (Mills, 2012). Essentially, big data employs large amounts of storage space, gathered from diverse sources, stored in diverse formats, and updated at different intervals. The characteristics of big data require use of new processing methods. The processing methods used for big data are commonly referred to alternately as analytics or data mining.

Data mining is the process of analyzing volumes of data to discover previously unrecognized patterns in the data set. The patterns discovered are generally "unsuspected associations" (Thillainayagam, 2012). Considered a step in the larger overall process of knowledge discovery in databases (KDD), data mining leverages statistical analysis and database technology to find these patterns (Chen, 1996; Fayyad, 1996). The steps in the KDD process are: 1.) data cleansing or data cleaning – removal of erroneous or irrelevant data, 2.) data assimilation – consolidation of data from diverse data sources, 3.) data assortment – retrieval of data selected for analysis from the overall dataset, 4.) data transformation – conversion of data into form necessary for chosen analyses, 5.) data mining – extraction of useful discovered associations, 6.) pattern evaluation – identification of patterns and definition of measurements for the patterns, and 7.) knowledge representation – data visualization to enable better understanding of patterns (Thillainayagam, 2012). The data mining process alone requires several stages: 1.) problem definition – identifying the goals of the specific analysis, 2.) data exploration – ensuring data collected is suitable for analysis selected; if not, recommendations for changing collection process are developed, 3.) data preparation – ensuring rules exist for missing and invalid data as well as consistency of valid data, 4.) modeling – selecting data mining algorithm (analysis techniques), and 5.) evaluation and deployment -- conducting and interpreting the analysis (Thillainayagam, 2012).

TAXATION AND TAX FRAUD

Several types of taxes exist in society today. Taxation is a necessary means of raising revenue so that government can provide goods and services (Hopwood, 2012). The taxes include individual income taxes imposed by both the federal and state governments. Corporate income taxes are also levied by both levels of government. Income received by the taxpayer, as adjusted by certain deductions and credits allowed by the law, is used to calculate income taxes, including individual and corporate taxes. While tax laws change over time (and political administrations), the laws are a constant in society. Other types of taxes exist, including sales and use taxes, employment or payroll taxes, estate, gift and excise taxes but income taxes have consistently been the largest part of our government's tax collections("Policy basics: Where do Federal tax revenues come from?," 2016).

Fraud is a serious problem currently causing financial hardship. The fallout of the Enron Corporation financial accounting scandal was estimated to be in excess of $63 billion dollars (USA Today, 2012). The Enron disaster rocked the United States economy for years. While the financial fraud has been well publicized, it pales in comparison to the hidden, almost socially acceptable cost of tax fraud. The Internal Revenue Service (IRS) estimates that tax fraud costs our federal government an average of $458 billion per year between 2008 and 2010 (Matthews, Fortune 2016). The University of Wisconsin-Madison estimates that the cost could be as high as $600 million per year (Clark, 2014). Tax fraud is perpetrated by professional criminals and by ordinary citizens as well. It is a seriously overlooked problem facing the government, all citizens and the accounting industry alike.

Black's law dictionary defines fraud as "all multifarious means, which human ingenuity can devise, and which are resorted to by one individual to get an advantage over another by false suggestions or suppression of the truth. It includes all surprises, tricks, cunning or dissembling, and any unfair way which another is cheated" (Garner, 2016). The essential elements in this definition are deception and the intent to deceive (Liuzzo, 2013). It is not enough for the target to be taken advantage of; it is essential that the perpetrators intent was to deceive the victim.

Fraud takes place in various types of financial transactions, including taxation. The Internal Revenue code defines tax fraud under IRC section 7201. Under this provision, "any person who willfully attempts to evade or defeat any tax imposed by [the IRS] or the payment thereof shall, in addition to other penalties provided by law, be guilty of a felony" (26 USC §7201). The IRS criminal investigation division generally mentions several types of fraud that are currently being focused upon. These include deliberately underreporting or omitting income, overstating the amount of deductions, keeping two sets of books, making false entries in books and records, claiming personal expenses as business expenses, claiming false deductions and hiding or transferring assets or income ("Types of fraudulent activities: General fraud," 2016).

More specifically, the statues define different types of tax fraud and other offenses including:

1) willfully failing to pay tax - 26 USC §7202
2) willfully failing to file a tax return – 26 USC §7203
3) furnishing fraudulent statements and/or not filing essential information–26 USC §7204
4) furnishing false identification numbers – 26 USC §7205(b)
5) false statements on returns requiring declarations of information – 26 USC §7206(1)
6) aiding or advising in the preparation of false tax returns – 26 USC §7206(2)
7) attempting to evade tax collection by hiding assets – 26 USC §7206(4)
8) falsifying or destroying essential taxpayer records – 26 USC §7206(5)
9) willfully delivering returns, statements or other documents – 26 USC §7207
10) interfering or threatening officers or the IRS – 26 USC §7212(a)
11) failure to collect or pay trust funds – 26 USC §7215
12) conspiring to commit a fraudulent offence against the United States – 18 USC §371

(all refrences are to the Internal Revenue Codes of 1939 and 1954 as amended).

The Internal Revenue service has several divisions within its framework. These include the wage and investment division, the small business/self-employed division, the large and mid-size business division and the tax-exempt and government entities division (Hopwood, 2012). More importantly, the IRS Criminal investigation Division was created to investigate potential criminal offenses and assure the collection of tax revenue. The above offenses exemplify the types of issues the IRS criminal investigation division must address within the realm of tax fraud. Finding this fraud is a difficult charge for often-overburdened governmental resources.

Fraud has been discovered by utilizing several different methods including direct and indirect procedures (Hopwood, 2012). Direct methods include specific identification of items listed on tax returns. These are the most infallible methods utilizing the matching of information IRS already has to information reported by the taxpayer. An example would be matching of information returns such as W-2's, 1099's and other statements provided by reporting agencies to the amounts listed on each specific tax return. These types of checks and searches by IRS are relatively easy and the IRS is very good at identifying incorrect data reporting. However, matching data within the tax system does not address the problem of falsely reported data to begin with. That requires the use of more indirect methods of investigation.

Indirect methods of detecting fraud require investigation and analysis. Sometimes these are referred to as analytical procedures. Three main methods have been used to provide what can be best described as circumstantial evidence of wrongdoing. The first method is the "net worth" type of analysis. Here the examiner looks at the change in a taxpayer's net worth over a year period to see if it is reasonable based on the amount of income reported. The second method is the "cash expenditures" analysis, which compares the taxpayer's known sources of income and comparing that to the taxpayers spending habits. The third indirect investigation method is called the "bank deposit" analysis. Here the investigator looks at all bank deposits and eliminates transfers between accounts to see how much actual money is flowing in to the taxpayers accounts.

Each of these methods are effective but much harder to accomplish than directly comparing numbers reported on returns.

All of these indirect methods of finding fraud have a common theme – a search for reasonableness. (*IRS Internal Revenue Manual, Part 4, Chapter 10, Section 4 - The examination of income*, 2016). Tax practitioners and Internal Revenue agents come across fraud often in the regular course of tax compliance duties The trouble is not finding the fraud it is in the enforcement. By its very nature, reasonableness is a fluid judgment, and judgment needs proof. Proof of fraud requires evidence which can be a difficult task for the IRS and its criminal investigation division. Evidence comes in many forms, but all the different types of evidence must be collected to build a strong case.

The types of evidence needed to prove fraud include physical evidence such documents and records. (*IRS Internal Revenue Manual, Part 4, Chapter 10, Section 4 - The examination of income*, 2016). This includes original documents such as accounting records either in paper form or on computer hard drives, flash drives and servers. All documents can prove incidents of fraud whether they exist in filing cabinets or on the internet. Helpful information not traditionally associated with accounting records can include personnel files, resumes, court and real estate records. Larger companies have numerous filings with securities authorities. Commercial databases also exist that provide a great deal of financial information that can be compared to accounting and tax records. Banks are required to file certain financial disclosure reports based on the size and nature of deposits. Documents are not however the only means of proving tax fraud.

Observational evidence is also key to proving fraud. (Hopwood, 2012). Counts of inventory are not the only way of finding observational evidence. Auditors base heavy emphasis on testing procedures within a company. Investigators can use these techniques as well in finding fraud. Merely observing how many people exist within a business can be determinative of so many items included on a tax return. Procedures used within the business are also helpful in testing the accuracy of the information given by the taxpayer. As easy as it seems to say that just looking at the business itself can tell so much about the value of the disclosures reported on the tax returns, the most effective method of indirect fraud detection may still be as simple as talking to people. The IRS has provided several publications that be used by taxpayers, tax preparers and businesses to aid in the prevention and detection of fraud. Primary among this information is IRS publication 4524 – Security awareness for taxpayers and publication 4557 Safeguarding taxpayer data.

As hard as the IRS Criminal Investigation Division works to find fraud, still one of its greatest resources are informants. The so-called "Whistleblower" statutes of 7263 were changed in 2008 to give more incentives to citizens to provide information on tax fraud. Information provided to the IRS must lead to a judicial or administrative action (an audit or an investigation) resulting in the collection delinquent tax. 26 IRC 7623. The statue provides incentives for informants of up to 15% of the underpayment, up to $10 million dollars. 26 IRC 7623(a). Many rules relate to the payout of the awards, such as the informant not participating in the creation of the underpayment. However, there exists a very large incentive to provide information on tax fraud. The IRS provides a specific form allowing informants to report and follow the correct procedures for attesting to information. (IRS Form 3949-A). The incentives are becoming more popular in bring fraud to the criminal investigation division's attention (Novack, 2009).

All of these methods of finding and proving tax fraud are labor intensive and need to be undertaken with the caution required to preserve proper trails and backup to demonstrate the underlying felony of fraud in a court. If items of evidence are to be taken to court, a proper litigation trail must be established to prove the veracity of the information provided. It can be an overwhelming task for our government and the accounting industry as a whole. The introduction of analytics into an already overburdened system would aid in the pursuit of tax fraud. The use of big data has slowly been introduced to some state revenue agencies, such as Maryland, to great success (Shacklett, 2016). Politicians are fast to cut budgets on the Internal Revenue Service. The agency is an easy target for blame when there are not enough resources to go around. Aiding in the enforcement of our current tax system can, however, generate more revenue while using less resources. Making our tax and accounting systems more efficient through the use of big data and analytics is the correct next best step to making sure that everyone pays their fair share.

BIG DATA ANALYTICS AND TAX FRAUD

Data mining tasks and techniques are used to find patterns indicative of financial fraud. The patterns discovered could be used either in detection or prevention of financial fraud. Two broad subgroups of data mining tasks are predictive tasks and descriptive tasks. Predictive tasks are so named because along with machine learning and related technologies these tasks make a prediction for each observation. Descriptive tasks, which include association rules and cluster analysis, describe the data being examined (Gupta, 2012).

Prediction utilizes regression analysis to examine relationships between one or more independent variables and dependent variables (Thillainayagam, 2012). The complexity of financial situations requires the volume of variables provided by big data to make more accurate predictions. The statistical techniques for these include linear regression, multivariate linear regression, nonlinear regression, and multivariate nonlinear regression, as well as the more complex logistic regression, decision trees, and neural networks (Thillainayagam, 2012). Neural networks are comprised of sets of connected input/output units, each of which has a connected weight. The weight is adjusted in the network analysis during the “learning phase” to allow the network to extract patterns which can be used in prediction (Thillainayagam, 2012). Other, more complex predictive techniques of data mining appropriate to fraud detection or prevention include rule-based fuzzy reasoning, genetic algorithms, Bayesian belief networks and fuzzy neural networks (Jans, 2009).

Descriptive tasks can be used to create models of behaviors or transactions that could be suspicious. The descriptive tasks might be types of association rule analysis including multilevel association rules, multidimensional association rules, and quantitative association rules (Thillainayagam, 2012). Association rule algorithms generate rules that describe potentially fraudulent situations. Cluster analysis collects data into related subsets patterns, or “high quality clusters with high intra-class similarity and low interclass similarity” (Gupta, 2012). As with association rules, cluster analysis or clustering involves discovery of patterns; the patterns discovered can be used to discover or prevention financial fraud.

Analytics requires big data, meaning use of multiple data sources. In a company audit, it would previously be typical to examine only the data of the company being audited with some limited exploration of "industry norms" (*IRS Internal Revenue Manual, Part 4, Chapter 10, Section 4 - The examination of income*, 2016). An audit undertaken to discover fraud would integrate large internal and external datasets, including such data as demographics, taxpayer or corporate profiles, previous filings, call center data, and audit histories (Opentext, 2015). The data analyzed could therefore include many years of historical data as well as external data. The volume and variety of data would be difficult to analyze without the analytics toolset.

All the tasks and techniques used in data mining emerge from the fields of statistics and computer science. Application of the techniques, therefore, will require analytical skills based in these fields. Ability to use statistical software packages and advanced spreadsheet applications should be sufficient for most processing. For some applications cited (e.g. neural networks, rule-based fuzzy reasoning, genetic algorithms), computer programming skills would be needed. New software packages are also now emerging that are specialized for the task of fraud analytics; many of these will likely still rely on ability of the user to interpret statistical analyses.

Infrastructure considerations must be made when considering the adoption of analytics to fraud detection and prevention. The big datasets typical in big data analytics requires larger capacity than regular analysis. Larger companies may have the resources and staff to manage larger systems. Small to mid-sized firms can employ cloud computing to access the infrastructure (Purcell, 2013) needed for analytics (Purcell, 2013). Many vendors provide cloud services for a reasonable fee, usually based on level of usage (Purcell, 2013). Types of cloud services provided are public, private, and hybrid. The public cloud is cloud services used at a vendor's site; the private cloud is an internal company data center based on cloud technology inaccessible to the public; the hybrid cloud is use of an outside vendor for cloud technology that is secured, with access to data protected from unauthorized users (Purcell, 2013). Small to medium sized businesses could leverage cloud computing for big data technology implementation to reduce hardware and processing costs inherent with big data processing.

RECOMMENDATIONS FOR USING ANALYTICS IN TAX FRAUD DISCOVERY

Accounting firms and government agencies will increasingly be called upon to use analytics in audit work. Considerations for a firm moving to analytics use involve employee skillsets and the firm's infrastructure. Some of the skills and infrastructure will remain the same. The accounting skills and infrastructure are essentially unchanged – the ability to plan an audit, review accounting and tax data, communicate the information obtained and report the findings. The new skillset will be the need for statistical analysis ability and computer skills. Several means of adding these skills exists. A firm could train existing staff members showing an aptitude in working with numeric analysis. Statistics and computer courses are available at universities, online, and through professional associations. The educational opportunities range from university credit courses in analytics to certificates for continuing professional education (CPE). When considering courses for employees, look specifically at the course content to ensure the course teaches the skills required. Another option is to hire new employees with the statistics and computer backgrounds. Many accounting students are taking coursework that adds the statistical and computer skills needed for analytics. A larger firm could consider hiring someone with the needed skills only (little or no accounting background) to add to an audit team.

Infrastructure considerations depend on the level of analytics the firm plans to employ. Cloud computing is a viable option for most small to mid-sized firms. An increasing number of vendors provide software and cloud services specific to fraud detection (Opentext, 2015; SAS, 2016; Teradata, 2016). Examination of infrastructure needs and service options provided by the software vendor (including cloud computing options) will provide a basis for determining infrastructure needs.

Before considering infrastructure and skills acquisition, decision makers at the firm should understand the options. Professional societies are beginning to address the need for big data analytics in accounting. Attending society meetings and reading related publications will aid professionals in learning how to examine their needs and develop a solution that is a best fit for the particular firm.

CONCLUSION

Analytics uses statistical and computing skills to study financial big data to determine previously undiscovered patterns of activity – some of which could be used in detection and prevention of tax fraud. The tasks used in tax fraud detection and prevention are predictive or descriptive in nature. Skills needed for performing these tasks go beyond the typical accounting and taxation skillset present in firms to include statistical and computer analysis. Infrastructure considerations can be grounded in the choice of analytics platform adopted; choices often involve cloud computing-based platforms provided by the vendor. Decision makers need to understand the terminology and options available to them. Actively engaging with professional societies with an understanding of the option addressed will allow decision makers to determine the best tax fraud analytics option for the firm and governmental agency.

References

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866 - 893.

Clark, C. (2014). Fighting tax fraud with big data. . Retrieved from http://www.cnbc.com/2014/04/11/fighting-tax-fraud-with-big-dataibmcommentary.html

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *American Assoication for Artificial Intelligence*, 37 - 53.

Garner, B. A. a. B., H. C. . (Ed.) (2016). Thomson Reuters West.

Gupta, R. G. N. S. (2012). Prevention of financial statement fraud using data mining. *International Journal of Computer Science and Information Security, 10*(4), 55 - 59.

Hopwood, W. S., Leiner, J. J., and Youg, G., R. (2012). *Forensic accounting and fraud examination*: McGraw-Hill.

IRS Internal Revenue Manual, Part 4, Chapter 10, Section 4 - The examination of income. (2016). Retrieved from https://www.irs.gov/irm/part4/irm_04-010-004.html.

Jans, M., Lybraert, N., & Vanhoof K. (2009). A framework for internal fraud risk reduction at IT integrating business processes: The IFR2 framework. *International Journal of Digital Accounting Research, 9*, 1 - 29. doi:10.4192/1566-8517-v9_1

Liuzzo, A. (2013). *Essentials of business law*. New York: McGraw-Hill.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A.H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Retrieved from http://www.mckinsey.com/Insights /MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_ innovation

Mills, S., et. al. (2012). *Demystifying Big Data: A practical guide to transforming the business of government*. Retrieved from http://breakinggov.com/documents/demystifying-big-data-a-practical-guide-to-transforming-the-bus/

Opentext. (2015). Advanced and Predictive Analytics for Tax Authorities.   Retrieved from http://birtanalytics.actuate.com/download/birt-analytics-advanced-and-predictive-analytics-for-tax-authorities.pdf

Policy basics: Where do Federal tax revenues come from? (2016).   Retrieved from http://www.cbpp.org/research/policy-basics-where-do-federal-tax-revenues-come-from

Purcell, B. (2013). Big Data using cloud computing. *Journal of Technology Research, 4*.

SAS. (2016). The impact of the underground economy, and how analytics can fight it. Retrieved from http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/impact-of-underground-economy-how-analytics-can-fight-it-108445.pdf

Shacklett, M. (2016). Fighting tax return fraud with analytics. Retrieved from http://www.techrepublic.com/article/fighting-tax-return-fraud-with-analytics/

Teradata. (2016). Targeting tax fraud with advanced analytics. Retrieved from http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB7183_GT16_CASE_STUDY_Teradata_V.PDF?processed=1

Thillainayagam, V. (2012). Data mining techniques and applications - A review. *I-manager's Journal on Software Engineering, 6*(3), 44 - 48.

Types of fraudulent activities: General fraud. (2016). Retrieved from https://www.irs.gov/uac/types-of-fraudulent-activities-general-fraud