Machine Learning and Advanced Analytics in Tax Fraud Detection

Conference Paper · October 2019

DOI: 10.1109/AICT47866.2019.8981758

Abzetdin Adamov

ADA University

Machine Learning and Advanced Analytics in Tax Fraud Detection

Abzetdin Z. Adamov
Center for Data Analytics Research (CeDAR)
ADA University
Baku, Azerbaijan
aadamov@ada.edu.az

Abstract – We all are passing through Big Data Renaissance right now. Academic literature shows increasing interest to subject of application of Big Data technology to cross-sectoral business and relevant scientific research. But portion of research papers devoted to application of Big Data to Taxation is relatively small, despite of the fact that tax generates huge amounts of Data and has enormous potential to benefit from Big Data Analytics. This paper is a conceptual approach to build theoretical and methodological foundation for Data Analytics application in taxation. It is described how tax authorities can benefit from their operational data. Primary prerequisites for efficient application of Advanced Data Analytics are revealed.

Key words – Data Analytics, Machine Learning, Big Data, Hadoop, Taxation.

I. Introduction

Ubiquitous application of information technology and automation of business processes in enterprises eliminates human involvement widening space for fraudulent activities. In order to combat fraud effectively, the same environment that enables variety of fraud operation should be used against it. Digital environment and online services enable collection of massive amounts of data. New approaches, platforms, tools and technologies now enable cost-effective processing and analysis of almost any size data that was not possible in near past. All these technological advancements increase not just quality of services, reduce cost and increase customer satisfaction, but also build powerful foundation for anti-fraud monitoring and prevention. Business process automation and online services bring to generation of huge amounts of digital data. This stored data can be employed as millions or billions use-cases to analyze and learn from finding specific attributes those indicate possible fraud. Even it may seem appealing to avoid tax payments totally or partially, but it brings many risks of imposing higher fines by tax authorities and loose credit and reputation. The most recent advancements in fraud detection even increase the likelihood of exposure and detection.

Despite of the fact that Data Analytics is accepted as most promising technology in fraud detection, according to [1] only 3% of fraud cases were detected using fraud-focused analytics, while 44% were found based on intuition or whistle-blower mechanisms.

Since fraudsters are evolving their knowledge using more sophisticated techniques and tools, it is not enough productive to try to identify fraud following to well known patterns. Thanks to technological advancements traditional fraud types evolve to cyber-enabled like ATM fraud, application fraud, card not present fraud, internet banking fraud, mobile banking fraud, contactless card fraud, financial reporting fraud, etc.

The problem of Data Analytics being not employed adequately is in the lack of understanding by authorities what Advanced Data Analytics can do for combatting fraudulent actions and how effective it can be on this particular task.

In such an evolving environment where fraudsters are in the front of the progress, it is impractical to tackle this issue based on traditional approaches. The only tool that can help is Machine Learning and Advanced Data Analytics that can help to find never before seeing patterns and build adaptive models that can be employed for particular fraud case or continuously evolving fraudulent behavior.

II. Understanding Big Data Problem in Tax

Data Analytics helps to identify Fraud in the way of applying particular tools, algorithms and approaches to find out trends, regularity, patterns and anomalies within huge amounts of complex data that can't be found manually. Advanced (Deep) Data Analytics can significantly increase the coverage of audit not limiting with LTO (Large Taxpayer Offices) only, but extending it to midsize and small taxpayers. To apply Data Analytics to Fraud Detection effectively, it is important to:

1. Build Analytics Model based on industry-oriented indicators and utilize knowledge base of organization;
2. Employ effectively technology including platforms, algorithms and proven concepts;
3. Pre-process Data turning it to Quality Data - Quality results require high quality data;
4. Utilize traditional fraud detection skills in Data Analytics changing scale, dimension and depth.
5. Continuously improve Fraud Detection Models and Indicators.

Some of data sources in taxation that can be Analyzed:

- Declarations;
- Individual Taxpayer;
- Legal Entities;
- Electronic Invoices;
- Quarterly and Annual Tax-reports;
- Bank Transactions;
- Audits.

Unstructured data sources for sophisticated Data Analytics:

- Emails and Messaging Data;
- Social Media activity;
- Description of Payments;
- Documents, contracts, etc.;
- News Media.

Some solutions towards building more transparent business environment:

1. Segmentation of Taxpayers depending on dept size, likelihood of delinquency and fraud for deeper analysis;
2. Work more intensively with third party data owners - Real Estate, Utilities, Banks, Customs, etc.
3. Define industry-specific criteria for auditing;
4. Work on incentives towards reducing cash flow decreasing shadow economy:
    a. apply limits to maximum amount of cash payments;
    b. tax discounts for non-cash payments for consumers and merchants;
    c. demonetization of big banknotes;
    d. obligation to use POS-terminals by merchants;
    e. obligation to pay salaries by companies and state organizations in non-cash;

II. Literature Review

According to literature review, there are many fraud detection methods those perform differently depending on specific type of fraud.

Authors in following paper [2] propose a way to find inconsistences and fraud not just using numerical data, but also textual. When it comes to narratives in annual report by companies, these reports mainly consist of free style text. Authors assume that narratives in financial reports may include some contradiction and inconsistency with numerical data. In order to compare text with numbers, both should be brought to the same format: both numerical, or both to categorical. NLP and text analytics are used to discover the content of annual narratives.

In [3] authors investigated the role of data mining techniques in fraud detection in several sub-areas of the financial domain. Authors applied multiple data mining algorithms including logistic regression, decision tree, SVM, NN and Bayesian networks.

The following research [4] is focused on applying several machine learning techniques to detect financial fraud including classification, clustering and regression. Authors also classified most of well known types of financial fraud. The main objective of the study is to identify the techniques and methods that give the best results.

Many authors devoted their research not just to solutions built on the top of technical solutions, but some of them also try to find socio-behavioral aspects of committing tax evasion. The authors of [5] paper try to analyze the relation between the taxpayer behavior and their social status. At the end they offer a platform that comes as combination of tax evasion detection models and big data processing capabilities.

The authors in [6] employed a new methodology of fraud detection in financial statement. According to authors, the combination of specific features derived from financial information and managerial comments in corporate annual reports allow to detect non-fraudulent companies. Bayesian belief networks (BBN) outperforms other methods in terms of True-Negative rate. To identify fraudulent companies with high True-Positive rate, many other variables are required.

This paper [7] addresses another widely used method of identifying fraud by Analysis of Financial Ratios. This method is based on idea of finding certain correlation between variables that can indicate financial fraud.

The following study [8] makes emphasize on commonly accepted approach of using special information system that automates the process of data analytics. Authors outline the list of most popular tools and define greatest barriers in leveraging data analytics.

III. Digital Transformation through Automation of Core-Functions in Taxation

"Data is Oil of the 21st Century". Data becomes primary asset of industries and organization operating in the modern digital age. Most effective governments are trying to keep up with businesses and get benefit of data they generate and have administrative power to manage. For many countries one of the largest sources of government revenue is taxation [9]. This is why tax authorities around the world have strong intention to increase tax revenue through closing tax under-reporting, omitting income, fraud in all types of taxation including but not limited to: PIT (Personal Income Tax), CIT (Corporate Income Tax), VAT (Value Added Tax) etc.

As a result, the search for result of Data Analytics for government and taxation returns a lot of commercial tools and in-box solutions, but at the same time there is a deficit of academic and scientific research devoted to this area.

Automation of core tax functions, allow streamlining of main business processes and collect digital data that enable validation, testing and analysis. Digital Transformation announced by industries and government institutions as a way of becoming more efficient, transparent and productive, should include certain KPIs to measure success of initiatives. Talking about tax administration, there should be at list several primary indicators of success including following three functions:

- e-Invoicing;
- e-Filing;
- e-Accounting;

Efficient implementation of these projects can enable another promising initiative e-Auditing which provides better results while requires much less resources in terms of time and workload. e-Audit means application of Data Analytics to the data submitted by taxpayer and third-party organizations including publicly available data.

A. e-Invoicing System

e-Invoicing is an Information Technology enabled reporting system that suppose post-reporting of all business transaction. e-Invoicing systems have been already implemented by some of leading tax authorities and is expected to be widespread within the next couple of years. Generally, an e-Invoicing system is offered as REST-application that consists of pre-defined fields indicating amount of transaction, Taxpayer Identification Numbers (TIN) of sides, quantities, price, etc. The analysis of transactions provided by e-Invoicing system allow tax-authorities to cross-check and verify amounts of indirect taxes (VAT) paid by businesses.

B. e-Filing System

e-Filing is another system that allows to individuals and legal entities to report Income Tax Returns (PIT and CIT) online. The system may provide API that allows to integrate e-Filing system into existing software (ERP, CRM, etc.) solutions within businesses avoiding redundant operations. Utilization of the e-Filing together with e-Invoicing allows tax authorities to better monitor tax-compliance and identify any potential under-reporting and fraud.

C. e-Accounting System

e-Accounting is another rising trend encouraged by tax authorities. Generally, e-Accounting is not certain developed information system, rather it is a set of standards or format of financial data. In some countries (including EU) taxpayers are required to provide financial data in accordance to this format. If e-Accounting is required, legal entities prefer to use financial software that generates reports in accordance with requirements or they may develop software appliance that adds the certain functionality to existing software.

All three above mentioned system can help tax authorities to collect relevant data from taxpayers that allows not just to verify tax returns of specific taxpayer, but also to apply Advanced/Predictive Analysis and Machine Learning (Artificial Intelligence) enabling Data Science capabilities in taxation:

- To build full tax picture in business-sector or geographically;
- To Classify (Clusterize) taxpayers by behavior and risk factors;
- To identify key predictors for the future and highlight high-risk areas;
- To determine key indicators to classify fraudulent taxpayers.

At the same time, having data in premises and having capabilities to effectively employ that data are totally different things. Many of tax authorities successfully apply basic data analytics using affordable (in terms of cost and demanded proficiency) tools to their data stored on traditional platforms (RDBMS/EDW). But when it comes to huge amounts of data that continues to increase at a fast pace, traditional analytical tools and platforms are no longer enough. Instead, Big Data technology is required known as special platform that supports Cluster/Distributed Computing and MPP (Massive Parallel Processing) built on top of distributed and highly scalable storage.

IV. Importance of Data Collection for Fraud Detection in Tax

Quality Data Analytics requires quality data. But it may be not enough to analyze just data, the analysis of business processes those generates specific data is also critical. What is role of customer, client, employee in the lifecycle of process? How much each type of client can influence on transaction amount? Are there any validation procedures that ensure the validity of attributes?

Area of taxation is one of the most suitable in terms of implementation of Supervised Machine Learning (SML). Tax administrations aggregate huge amounts of data as a result of tax submissions by taxpayers. Even huge amount of data with many attributes is available, the production data is not directly applicable for SML. Data is not labeled what is essential requirement for SML. One of the best candidates that can utilize and largely benefit from ML is auditing departments. SML can help them to identify fraudulent submissions and focus on them, instead of relying on just own experience. As a result of each day activities of auditors and collectors during remote auditing they collect unstructured data in form of comments and reports. This kind of data can't be directly used for ML model training. If tax authorities plan to use ML, they (IT departments or contractors) should develop clear instructions for auditors to collect data in the form that can be easily processed and loaded to the system. And they should start this process well in advance before starting Data Analytics project. Another, more efficient way of data collection in ML-friendly way is to provide auditors with mobile (remote access) devices to enter results of investigation directly to the system in classified way.

Big Data is a term applied to any dataset with the size that can't be effectively stored and processed using traditional systems and platforms.
Financial sector, in general, and its particular area Taxation can't stay aside from data-related trends. Since taxation mainly relies on self-declaration of income, it largely open for fraud. There are many fraud methods in taxation well known by tax authorities, but there are no commonly admitted standard identification approaches those can help to tackle this global issue.

Taxation, generally, generates and operates with number of large datasets. Traditional tools like spreadsheets are not enough good to process and analyze large datasets, respectively special infrastructure and tools with parallel processing capabilities are required. It obvious that small and medium companies can't invest enough into building such a complex and expensive infrastructure. They can employ cloud-based solutions from vendors or develop own systems on the top of open-source products.

V. Data Storage Platform Solution

Hadoop is open source common platform that combines two main tasks of any operating system: storing and processing data. Unlike to traditional systems, Hadoop accomplishes that tasks towards Big Data. The popularity of Hadoop increases day by day, because of simplicity, scalability and affordability that it enables thanks to its distributed architecture. Although, the Hadoop Core consists of two main components (HDFS and MapReduce) and has limited functionality, but thanks to many other components available in the Hadoop Ecosystem under Apache License, this platform can cover any requirements to manage and process data regardless to its size and format [10].

Data Analytics in scale that is enabled by Hadoop Ecosystem opens new horizons in turning operational data of businesses into actionable knowledge and consequently into value. In most cases, operational data is generated in structured format and stored in RDBMS. These databases and warehouses are still important, but now in era of Big Data businesses deal with amounts of data that can't feet into traditional RDBMS. Another challenge here is that in order to keep competitive advantage, businesses want to use for analytics all available data including website clickstream data, text from call centers, emails, instant messaging repositories, open data initiatives from public and private entities. It is clear that this goal can't be achieved based on traditional RDBMS systems. In contrast, Hadoop is a platform which consists of many components designed to accomplish specific tasks using particular data format. Hadoop Ecosystem components are classified into several categories to make it easier for user to choose appropriate components in accordance to the functions they designed for. There are following categories of the Hadoop components:

- Core Hadoop
- Governance, Integration
- Data Access and Storage
- Operations, Monitoring, Orchestration
- Security
- Data Intelligence

Conclusion and Future Work

Taxation administrations are responsible for most data-intensive functions in government operations. Tax authorities understand that organization can be data-intensive and at the same be far away from being Data-Driven. At the age of Big Data and Analytics, tax authorities and state agencies will increasingly be called towards implementation of Advanced Data Analytics and Machine Learning for specific functions including auditing and fraud detection. These technologies help to discover patterns in taxpayer's behavior that can enable detection and prevention of fraud.

Not all tax authorities, same as with other government organizations, have same capacity and capabilities to employ Machine Learning and Data Analytics. Development of such a capacity is a long process that includes, but not limited with following components:

- Understanding how taxation can benefit from operational data;
- Understand own data and have capability to transform data into quality data;
- Deployment of right applications and provision of online services to harvest right data;
- Making all data available (using special distributed platforms);
- Capability to apply Machine Learning and Data Analytics at the scale;
- Development of Data Science (scientific) capabilities.

As a future work it is planned to apply various Machine Learning approaches to large-scale real (anonymized) data provided by tax authorities. The purpose of this study will be identification of certain ML methods which perform best with identification of specific type of fraud.

Acknowledgments

This research was supported by Center for Data Analytics Research (CeDAR) at the ADA University.

References

[1] Phill Ostwalt, Global Data & Analytics, Trusted Analytics article series, KPMG International, July 2016

[2] Pascal Balata, Gaétan Breton, (2005),"Narratives vs Numbers in the Annual Report: Are They Giving the Same Message to the Investors?", Review of Accounting and Finance, Vol. 4 Iss 2 pp. 5 - 14 Permanent link to this document: http://dx.doi.org/10.1108/eb043421

[3] Mousa Albashrawi, Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015, Journal of Data Science 14(2016), 553-570

[4] Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. Procedia Computer Science, 148, 45–54. doi:10.1016/j.procs.2019.01.007

[5] Stankevicius, E., & Leonas, L. (2015). Hybrid Approach Model for Prevention of Tax Evasion and Fraud. Procedia - Social and Behavioral Sciences, 213, 383–389. doi:10.1016/j.sbspro.2015.11.555

[6] Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. Knowledge-Based Systems, 128, 139–152. doi:10.1016/j.knosys.2017.05.001

[7] Kanapickienė, R., & Grundienė, Ž. (2015). The Model of Fraud Detection in Financial Statements by Means of Financial Ratios. Procedia - Social and Behavioral Sciences, 213, 321–327. doi:10.1016/j.sbspro.2015.11.545

[8] Bănărescu, A. (2015). Detecting and Preventing Fraud with Data Analytics. Procedia Economics and Finance, 32, 1827–1836. doi:10.1016/s2212-5671(15)01485-9

[9] International Monetary Fund, World Bank and OECD GDP estimates., 2017 https://data.worldbank.org/indicator/GC.TAX.TOTL.GD.ZS

[10] Large-scale Data Modelling in Hive and Distributed Query Processing using MapReduce and Tez, DiVAI 2018 - Distance Learning in Applied Informatics, 02 - 04 October, 2018, Štúrovo, Slovakia

[11] Adamov, A. Z. (2018). Mining Term Association Rules from Unstructured Text in Azerbaijani Language. 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT). doi:10.1109/icaict.2018.8747143