BIG DATA ANALYTICS & ITS APPLICATIONS IN THE TAX DOMAIN

Mr. N. Koteswara Rao, Asso.Professor, Department of CSE Narayana Engineering College (Autonomous), Gudur, SPSR Nellore, AP, India
P.V.Bhavankumar , V .Rajagopal , P.Saiteja UG Student, Department of CSE Narayana Engineering College (Autonomous), Gudur, SPSR Nellore, AP, India

Abstract

The aim of this paper is to demonstrate how integrating Big Data Analytics into the Fraud Framework aids in the creation of an integrated predictive model that can integrate risk scoring in terms of predicting the likelihood of tax evasion, non-compliance, and fraud in the future for a current or new taxpayer. This would aid in the formulation of the appropriate solution or interventions prior to the occurrence of the case, and this lateral transition to predictive mode from the current reactive mode would result in revenue optimization. Working on a sample of 20,000 dealers from a database of 7.5 lakh specific dealers of a leading commercial tax department in India is part of the technique. The approach included testing multiple competitive statistical models and zeroing on the best model to score on the entire dealer base.

Keywords and phrases: Tax evasion, risk scoring, predictive modelling, statistical modelling, and big data analytics.

Introduction

Every year, billions of dollars are lost due to non-compliance, evasions, frauds, and non-collection by tax administrations all over the world. Tax authorities have access to massive amounts of data from a variety of sources (e.g., financial institutions, services, bank transfers, social media data, etc.), both structured and unstructured (text, video, pdfs, etc.). Using emerging techniques such as rule-based tracking, predictive modelling, and outlier identification, tax authorities will increasingly use big data and advanced analytics techniques to perform audits and discover patterns and discrepances[1][2].

Percentage of GDP

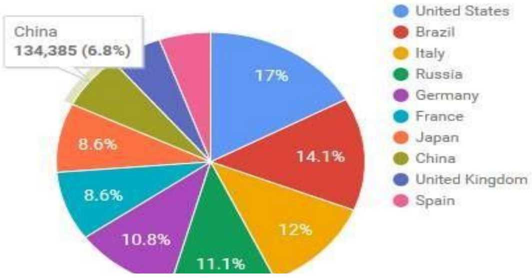


Figure 1

The figure above showcases the percentage of GDP lost as revenue because of non-compliances, evasion, and fraud in taxes across different continents. This paper will highlight examples of forward-thinking tax administrations in both developing and developed countries. Using big data analytics on a regular basis, you can improve service and enforcement efficiency.

PURPOSE OF THE STUDY

The aim of this piece during the paper implementation is to assist in the following:

I. Identifying the significant factors leading to tax evasion/non-compliance across various countries, types of businesses, types of tax payers, and so on, and thus assisting in determining what are the impact factors and non-impact factors among the massive data that the tax authorities have after the investigation.

II. Statistically assisting in the development of optimised and most impactful business rules by prioritising significant factors in identifying potential or actual evaders/non-compliances/fraud[3].

III. Risk Scoring, which involves rating and quantifying risky individuals and businesses using advanced analytics and predictive modelling on the entire population of tax payers.

Burning questions for the different tax authorities

I. Based on the massive data that is currently deposited with various tax authorities, what are the causes of tax evasion, abuse, and non-compliances[7][8].

II. Can significant factors (both direct and indirect) integrated with multiple sources from multiple systems provide a quantifiable risk score to existing dealers/organizations/individuals for the right approach through a scientific automated process? Is there a rise in taxes and the tax base as a result of this?

III. Can a scientific and methodical statistical process be developed to aid in the development of optimised and most impactful business rules by prioritising through significant factors in identifying potential or actual evaders/non-compliances/frauds? Can a current decision system assist subject matter experts in validating possible market laws of alleged evaders/frauds?

IV. Currently, there are silos among the various stakeholders in terms of using the entire information, and as information is used in bits and pieces or on an ad hoc basis. Is it possible to set up a process/platform that could use all of the data to construct an automated process for identifying.

V. Can a statistical or tailored predictive model be implemented for risk scoring in terms of forecasting the likelihood of potential tax avoidance, non-compliance, and frauds for an existing or new taxpayer[4][5].

How the project answers the burning questions and mitigates the pain areas of the different tax authorities:

A. The project entails first combining different sources from various systems and then creating a single, unified table.

B. The second stage entails using data quality measures such as data standardisation, deduplication, imputations, transformations, and so on, in order to improve or preserve a standard data quality.

C. The third stage entails developing statistical modelling to distinguish important from insignificant factors in a complex situation, thus addressing one of the most pressing market questions.

Methodology/ approach

Working on a sample of 20,000 dealers from a database of 7.5 lakh specific dealers of a leading commercial tax department in India is part of the technique. Testing several competitive statistical models and settling on the best model to score on the entire dealer base was part of the strategy. The entire process was automated, and the Fraud Framework used Big Data Analytic in India.We should ideally have a list of verified risky/non-risky dealers so that we can train an equation-based model on historical data and then check it on validation data to see if the predictions are accurate. The best candidate model[6].

Our Approach which is very effective considering the limitation of the confirmed cases

In the example above, a sample size of at least 5% of the population is needed. We used the business rule because the number of reported cases was so small, and we tried to statistically test it, assuming the business rule was right.

Data Partition Mechanism incorporate

The Data Partition Node divides the data into two sections: training and validation. In our case, 50 percent of the data was used for training and 50 percent was used for validation. In both preparation and testing, the ratios of 1:0 (probable risky vs genuine dealers) are retained.

For example, if we have 20,000 samples, ten thousand of which are risky and ten thousand of which are not, and we divide them into 50:50 training and validation, the number of samples in training would be 20,000. = 10000 (5000 risky dealers, 5000 genuine dealers) & number of validation samples= 10000 (5000 risky & 5000 genuine dealers). This is stratified sampling, which means that each strata/category has an equal percentage of participants (1 & 0)[9].

Layers of the Methodology Adopted:

The procedure consolidates five layers including

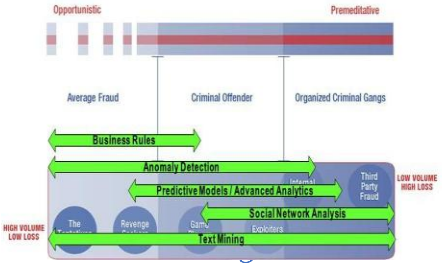
i) Business Rules: A rundown of business rationale and rules tried, appreciated and exceptionally powerful across the diverse business charge divisions can be fused with required customization. The principles at that point can be approved both measurably also by the business clients as business rules.

ii) Anomaly Detection Techniques: The peculiarity discovery strategies incorporates both factual exception identification just as business rules to comprehensively see any abnormality/possible non-compliances and fakes across a rundown of key execution pointers.

iii) Predictive Models/Advance Analytics/Data Mining Techniques: The prescient models, advance examination are utilized for hazard scoring for existing citizens just as for another citizen/vendors in future.

iv) Social Network Analysis: The Social Network Analysis or Link Analysis would assist with connecting distinctive key execution pointers, data as far as ledgers, service charges, paper data, web data and so on and make affiliation rules for plausible dodgers. This method is extremely powerful for party exchanging or expected round exchanging and has had the option to zero in on some large organizations including in enormous exchanges.

v) Text Mining and Sentiment Analysis: The content mining and assessment investigation helps in ordering the various writings, unstructured information, paper data, recordings, pdfs, web data accessible in the organize and subsequently can be utilized in a state of harmony with the prescient models to refresh the danger scores.



Business Objective:

i)To appropriately order a likely dodger (hazardous vendor) as a plausible dodger and a decent seller (non-dangerous seller) as a decent vendor

ii)What are the huge factors and the immaterial factors from the informational index gave as far as separating between a hazardous and non-dangerous vendor.

Information quality: The exercises fused for making a disinfected CAD (Core Analytics Data) for rule building region) Data Profiling b) Deduplication c) Standardization d) Imputation

Business Rule:

Limit of the sellers' data to be acknowledged for measurable demonstrating:

i) At least 3 years of enlistment

ii) Input Tax > 10, 000

Business Rule for First Bucket Cases of Tax Evasions/Risky Dealers

Information Tax> 10000 and Output Tax = 0

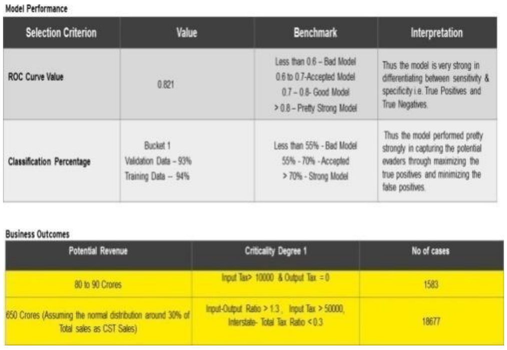
Business Rule for Second Bucket of basic Cases of Tax Evasions/Risky Dealers

Information Output Ratio > 1 , Input Tax > 10000 , Gross Turnover > 1 crore , Interstate-Total Tax Ratio < 0.5, Local Sales/Total Sales > 0.5 (More Local Sales than Interstate deals) , Avg. Absolved products/Gross Turnover < 0.5(Not enough excluded merchandise) ,Avg. Export= 0 (No fare) and Should record returns in the last two monetary year.

Exception identification:

Stocks to Gross Turnover & ITC Mismatches were used to look for these cases for justification. Also logistic regression, decision tree & neural network are used as automated processes to capture the exceptional cases.

Significant variables – Selection & Ranking Process for one of the commercial tax agencies Statistical Model Performance &the potential

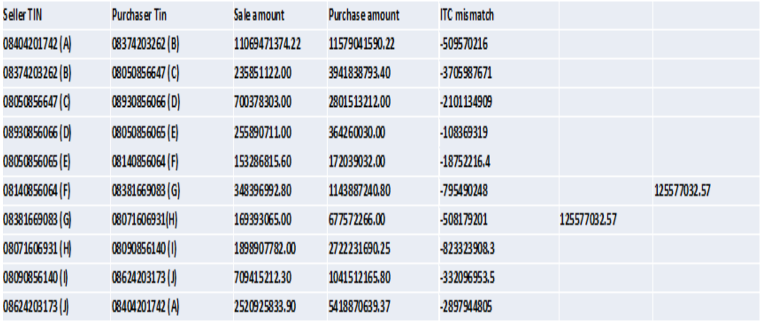


Model Performance

Selection Criterion	Value	Benchmark	Interpretation
ROC Curve Value	0.821	Less than 0.6 – Bad Model 0.6 to 0.7-Accepted Model 0.7 – 0.8- Good Model > 0.8 – Pretty Strong Model	Thus the model is very strong in differentiating between sensitivity & specificity i.e. True Positives and True Negatives.
Classification Percentage	Bucket 1 Validation Data – 93% Training Data -- 94%	Less than 55% - Bad Model 55% - 70% - Accepted > 70% - Strong Model	Thus the model performed pretty strongly in capturing the potential evaders through maximizing the true positives and minimizing the false positives.

Business Outcomes

Potential Revenue	Criticality Degree 1	No of cases
80 to 90 Crores	Input Tax> 10000 & Output Tax = 0	1583
650 Crores (Assuming the normal distribution around 30% of Total sales as CST Sales)	Input-Output Ratio > 1.3 , Input Tax > 50000, Interstate- Total Tax Ratio < 0.3	18677



Seller TIN	Purchaser Tin	Sale amount	Purchase amount	ITC mismatch		
08404201742 (A)	08374203262 (B)	11069471374.22	11579041590.22	-509570216		
08374203262 (B)	08050856647 (C)	235851122.00	3941838793.40	-3705987671		
08050856647 (C)	08930856066 (D)	700378303.00	2801513212.00	-2101134909		
08930856066 (D)	08050856065 (E)	255890711.00	364260030.00	-108369319		
08050856065 (E)	08140856064 (F)	153286815.60	172039032.00	-18752216.4		
08140856064 (F)	08381669083 (G)	348396992.80	1143887240.80	-795490248		125577032.57
08381669083 (G)	08071606931(H)	169393065.00	677572266.00	-508179201	125577032.57	
08071606931 (H)	08090856140 (I)	1898907782.00	2722231690.25	-823323908.3		
08090856140 (I)	08624203173 (J)	709415212.30	1041512165.80	-332096953.5		
08624203173 (J)	08404201742 (A)	2520925833.90	5418870639.37	-2897944805		

Table 1 Table 2

Business Outcomes for one of the Commercial Tax Agencies

Round exchange model for one of the significant business charge assortment authority of India through Network Analysis (Identification of Linkages—requires digitized information) Sector: Iron Ore discount exchange

CONCLUSION:

In summary, in the era of big data, if no reform is carried out to the existing tax collection and administration system, it will seriously affect China's taxation work and bring severe economic losses to our country. Facing the problems of imperfect tax legal system, not high professional quality of collection and administration staff, and inadequate enforcement of tax collection and administration, tax authorities should cooperate and negotiate with local governments and other departments actively to promote informationization of tax collection and administration jointly.

REFERENCES:

1. Zhu Yuanguang. Speed up the change of expense assortment and organization upheld 2016, (12):30-31.
2. Wang Shuai. Driving assessment organization development with enormous information innovation [J]. Fujian quality administration, 2017(4): 83-83.
3.Wei Junbo, Wang Chao, Xue Jianqiu, Zhu Lan, Su Ping, Yan Liping. The impact of large information on current assessment organization under the foundation of Internet+ [J]. Tax collection Research, 2016, (10): 108-112.
4. Tian Jinying, Sun Meiqi. Investigation of Effective Strategies for Building a Modern Tax Collection and Administration System in the Big Data Era [J]. China Journal of business, 2017, (19): 177-178.
5. S. Banerjee and H. Al-Qaheri, "An insightful half breed conspire for streamlining parking spot: An unthinkable analogy and harsh set based approach,"Egyptian Informat. J., vol. 12, no. 1, pp. 917, Mar. 2011.
6.B. Ramachandran, S. K. Srivastava, C. S. Edrington, and D. A. Cartes, "A clever sale conspire for brilliant lattice market utilizing a cross breed safe calculation," IEEE Trans. Ind. Electron., vol. 58, no. 10, pp. 46034612, Oct. 2011.
7. A study on potential of big visual data analytics in construction Arena Bhargava, M.G., Vidyullatha, P., Venkateswara Rao, P., Sucharita, V.International Journal of Engineering andTechnology(UAE), 2018, 7(2.7 Special Issue 7), pp. 652–656
8. Mandava Geetha Bhargava, Modugula TS Srinivasa Reddy, Shaik Shahbaz, P Venkateswara Rao, V Sucharita Potential of big data analytics in bio-medical and health care arena: An exploratory study, Global Journal of Computer Science and Technology 2017/8/5
. 9.D. Mackowski, Y. Bai, and Y. Ouyang, "Parking space the executives through unique execution based valuing," Transp. Res. Procedia, vol. 7,pp. 170191
,