Audit Selection Strategy for Improving Tax Compliance – Application of Data Mining Techniques

Manish Gupta[1] and Vishnuprasad Nagadevara[2]*

ABSTRACT

The tax administration is required to audit some or all its taxpayers to check the evasion of tax and ensure compliance. Conducting of audits involves costs to the tax department as well as to the taxpayer. Thus, audit is not a very welcome procedure both for the taxpayers as well as the economy. Tax administration agencies must therefore use their limited resources very judiciously to achieve maximal taxpayer compliance, minimum intrusion and minimum costs. This paper analyzes the use of data mining algorithms as the best cost effective option to make audit selection more efficient and effective.

Keywords: Tax audit; Tax Compliance; Data Mining Techniques; Audit Selection Strategy

1. Introduction

Tax on sale of goods is a state subject in India. It is an indirect tax on the consumption of goods and is borne by the consumer. Previously it used to be administered under the names of 'Sales Tax', 'Commercial Tax' or 'Trade Tax'. However, since 2005, almost all the states have switched over to a new system known as 'Value Added Tax' or VAT. The tax administration is required to audit some or all its taxpayers to check the evasion of tax and ensure compliance. A (tax) audit is a detailed exploration into the activities of a taxpayer to determine whether he/she has been correctly declaring the tax liabilities. Audits indirectly drive voluntary compliance and directly generate additional tax collections, both of which help tax agencies to reduce the 'tax gap' between the tax due and tax collected. Audit plays the pivotal role in the administration of tax and achieving the revenue objectives, ensuring the fiscal health of country and ensures a level playing field for an honest taxpayer.

On the other hand, being an intrusive method, audit is not a very welcome procedure both for the taxpayers as well as the economy. Conducting of audits involves costs to the tax department as well as to the taxpayer. Tax administration agencies must therefore use their limited resources very judiciously to achieve maximal taxpayer compliance, minimum intrusion and minimum costs. Planning an adequate audit strategy therefore, is a key success factor in *a posteriori* fraud detection (where audits are intended to detect tax evasion and fraudulent claims) as well as proactively preventing frauds and tax evasions. Most of the state sales tax departments in India had been auditing all taxpayers until introduction of VAT. Rapid development of the economy and introduction of VAT resulted into sharp increase in the ratio of taxpayers per taxman and thus spreading the available manpower too thin. The methodology followed so far (under VAT) has been to allow all taxpayers to opt for self-assessment. A fixed percentage of returns of taxpayers

[1] Ministry of Textiles, Government of India, New Delhi, India
[2] Indian Institute of Management, Bangalore, India
* *Corresponding Author:* (Email: nagadev@iimb.ernet.in, Telephone: +91-8026993144)

opting for this system are picked up for audit, mostly on random basis. Some states have introduced audit selection based on information, procedural non-compliance, (lack of) growth etc. The major drawback of random selection is its equal treatment meted to honest and dishonest taxpayers, as probability of selection is same for both. The criteria/information based selection system too has a drawback that it pre-supposes certain symptoms of non-compliance, which may actually be symptoms of other things such as change in economic situation in that particular trade.

Selecting returns for Audits is like looking for a needle in a haystack. Every year, a large number of taxpayers fail to declare their tax liabilities correctly, and the Tax Administration is forced to tackle a tough task – to detect them (and enforce compliance from them) without increasing the compliance costs of the tax compliant taxpayers. It is not possible to identify the likely tax-evaders by simple query and reporting tools. Tax departments have access to enormous amounts of taxpayer data. However, it is impossible to ascertain the legitimacy of and intention behind a claim or a declaration in the tax return by simply looking at the return or a profile of a taxpayer. Given this reality, the best cost effective option is to tease-out possible indications of fraudulent claims/declarations from the available data using data mining algorithms.

VAT being new in India, most of the states are experimenting with various methods of audit selection to come up with the best and most effective strategy. On the other hand, it is not possible for the State governments to obtain and utilize the strategy from other countries for two reasons. First, getting such a system from any other country is not feasible due to high level of secrecy attached to such systems. High level of secrecy is to ensure that the method of selection is not known to outsiders else unscrupulous taxpayers may manipulate their returns and other documents to avoid selection while making wrong reporting in returns and other documents. Secondly, each country has its own unique systems. Thus audit strategy of other countries is unlikely to be effective in India and the States have to come up with their own strategies in this regard. Data mining techniques have been widely used in learning the patterns in the past data and these learnt patterns have been used to predict the behavior of the current data. Knowledge Discovery in Databases (KDD) is a hot and emerging area of research as well as applications. An Audit Strategy based on KDD can meet the conflicting issues of audit planning, viz., trade off between maximizing audit benefits and minimizing audit costs.

2. Review of literature

Alm, Blackwell and McKee (2004) have studied the selection rule for Sales Tax Audits in the US and its impact on tax compliance. This paper estimates the process by which firms are selected for a sales tax audit and the determinants of subsequent firm compliance behaviour, focusing upon the Gross Receipts Tax in New Mexico. Results indicate that auditors select returns based upon a systematic, even if informal, audit rule and that firms that exhibit greater variation in deductions, provide services, miss filing deadlines, and have an out–of –state mailing address have a lower compliance rate.

Murray (1995) explores the subject of sales tax audit selection and firm underreporting of statutory sales tax liabilities. The analysis relies on sample selection estimation techniques in identifying systematic audit selection rules and the determinants of sales tax underreporting. The results support the view that sales tax accounts are chosen for audit non-randomly. The analysis also provides strong evidence that taxpayer opportunities for underreporting are correlated with the observed behaviour of firms. Fisher (1985) gives an overview of the methodology that is adopted by the State of Texas and Tennessee in designing the Audit Selection formula. Both the states are using regression and other statistical techniques to help perform sales tax audit selection. Their approaches are somewhat different.

Wedick (1983) describes the Discriminant function analysis, also known as Discriminant Index Function (DIF) as the most important method used by the IRS to select the 1%-2% of returns audited. DIF uses

multiple variables or criteria to differentiate between 2 populations. DIF formulas are developed from data collected as part of the IRS's Taxpayer Compliance Measurement Program (TCMP). When a return is processed at an IRS service centre, selected line items are transcribed onto magnetic tape. To achieve a DIF score, the items are multiplied by a DIF coefficient, and then added to obtain a total DIF score. After the returns have been scored, IRS managers can determine the number of returns that will be audited based on their available staff power, using the DIF score.

Manasan (2003) has developed industry benchmarks for (i) the Input-output ratio and (ii) effective VAT rate (i.e., the ratio of VAT liability to VAT output) for various industry groupings. Such benchmarks are envisioned to form part of toolkit in selection of VAT returns for audit. The paper has concluded that excessive claims for VAT credit are a major source for revenue leakage in the Philippines. The tax administrators can predict such excessive claims by comparing the Input-Output ratio in returns with the business benchmarks. Wiersema (1997) provides an overview of the parameters that IRS uses to screen the returns and select them for audits. According to this paper, The IRS tests the returns on items like (i) mathematical accuracy, (ii) matching information returns, (iii) unusual or large deductions or losses.

Cecil (1998) classifies the non-random methods used by IRS in selection of returns for audits. In the year 1996 in the US, unallowable items accounted for 42% of audited returns. The second most important source (18%) of audited returns was Discriminant Function Analysis (DIF). The third most important audit source (11%) was the nonfiler programme, while the fourth (4%) was the state information programme. Ho and Lau (1999) give an overview of procedure of selection of returns for audits in Hong Kong. It states that returns are selected on the basis of criteria (e.g. turnover volume or gross profit ratio) however there is no exhaustive list of selection criteria.

Kumar P (1999) has developed a criteria based selection system for sales tax in India, examining various parameters in the Balance Sheet of the taxpayer such as Turnover growth, Current ratio, Asset to turnover ratio, Coverage etc. with tax compliance. He has developed a model using Discriminant Function and benchmarking of these parameters with norms. European Commission Report (2006) provides background information / guide for tax officials on Risk Management for Tax Administration. It states that Risk analysis normally requires the use of computer systems because of the sheer volume of data. However, manual analysis and weighting is possible and sometimes even preferable and in some cases human intelligence and professionalism is indispensable. A combination is possible with a rough mass weighting by machine followed by a specific weighting by people. Risk analysis can be done within the administration, centrally or locally, or in combination.

Executive report by Micci-Barreca and Satheesh describe a case study of an audit selection strategy for the State of Texas, where the state has effectively used data-mining tool (Clementine) for design and implementation of the strategy. The Audit Division's Advanced Database System (ADS) uses predictive models to identify sales tax audit leads. It uses five sources of data to create taxpayer profile, which is used to train the model. The model predicts an 'Audit Score' that signifies the likelihood of the taxpayer under-reporting tax. Results indicate that the tax adjustments (i.e., outcomes of audits) increased significantly in proportion to the Audit Score. The paper also describes the phases of application of data mining techniques, viz., Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment.

In their technical report, Micci-Barreca and Satheesh present a case study to show how a tax and finance department built models based on previously audited returns to identify potential non-compliant taxpayers. It gives the steps that are required to be undertaken to implement a data mining technique. In the instant case, linear regression and Neural Network models were used. Bonchi, Giannoti, Mainetto and Pedreschi (1999) illustrate how techniques based on classification can be used to support the task of planning audit

strategies. It presents a methodology for constructing profiles of fraudulent behaviour, aimed at supporting audit planning. Emphasis is placed on the methodological issues of design and control of the process. The paper describes the actual working with the decision tree classification in detail.

Phua, Alahakoon and Lee (2005), in their paper, categorize, compare and summarize fraud detection methods and techniques published in academic and industrial research during the last 10 years. Within the business context of mining the data to achieve higher cost savings, it presents methods and techniques for fraud detection together with problems. It describes four major methods commonly used for applying Data-mining algorithm, viz., (a) Supervised approaches on labelled data; (b) Hybrid approaches with labelled data; (c) Semi-supervised approach with non-fraud data; and (d) Unsupervised approaches with unlabelled data.

Shao, Zhao and Chang (2002) describes the building of fraud detection model for Qingdao customs port at China to give decision rules to the custom officials for inspection of goods on the basis of past transaction data, with the objective of improved hit rate. The model so developed (named Intelligent Eyes) is successfully implemented with high predictive accuracy. Kumar Anuj (2005) has developed risk assessment (predictive) model for selective customs examinations in Indian customs. The model has been developed using classification tree algorithm and is expected to detect over 90% of the total duty short declarations with mere 30% of the original examination efforts.

3. Objectives

The objective of this paper is to develop a model that can identify dealers/ returns that have maximum likelihood of tax under-reporting in the large volumes of tax returns filed by the dealers in VAT system as well as to minimize the examination effort and costs, so that the scarce audit resources can be effectively deployed by the tax department. These two objectives are contrary to each other, as enhancing the detection would certainly mean the increase in examination effort. Therefore an optimal trade-off has to be achieved by the model to suit the end objective. It will be a predictive model, as it would predict the likelihood of a dealer under-declaring / evading tax in the return. If utilized, the Department would be able to allocate its limited resources for more productive and specific purposes.

4. Methodology and Description of data.

For the purpose of Data mining, the dealers are divided into two groups on the basis of prior information available. After categorizing the taxpayers into groups, suitable data-mining techniques are selected and applied to identify significant parameters that effect evasion/ under-reporting of tax. The first group in the data mining exercise is the sample of dealers found to be evading/ under-declaring tax. Delhi VAT had approximately 1,80,000 dealers registered with the department. It is impossible to know about the complete set of dealers who had evaded or underreported the tax. Therefore, for the preparing the first set, a study of business practices in the department was conducted. It was decided to select the dealers who had been assessed additional tax and where such assessed tax was sustained in the first appeal. The year for which the selection would be done was chosen as 2003-04 as this was the latest year for which the appeals had been decided. Manual collection of data (from the files / registers of the Appellate Authorities) resulted in a sample containing 402 dealers, from whom, the recoveries in respect of year 2003-04 were of Rs 50,000/- or more.

The second sample required is of ‘good’ dealers, which, with reasonable level of confidence, can be said to have been correctly reporting the tax. For the second group, businesses that pay high tax were assumed to be ones that correctly declare their taxes. However, in identifying the sample, sufficient precautions were required to ensure it is sufficiently representative. For this purpose, (major) commodity and the nature of business (retailer / whole seller/ exporter etc.) of the 402 dealers were identified. Thereafter, for each of the

dealer in the 'evading' category, three highest tax paying dealers dealing in same commodity as well as having same nature of business were identified. The tentative list of 'good' dealers so generated was manually updated from the respective wards to examine if there is any adverse material, based on surveys, raids etc. from official records. Such dealers, if found, were deleted from the list. The said procedure generated approx 1200 dealers. Three times the number of evading dealers has been taken so as to strike a balance between representative-ness of the data-set and skew-ness of the data set.

Based upon the literature survey, parameters used in other VAT jurisdictions, domain expertise of Tax department officers and experts, as well as usability (on the basis whether the parameter can be extracted from the returns/computer database), the following input variables were identified initially for the data mining exercise.

A. Dealer Profile
- A1. New Registrant (Y/N)
- A2. Deals in high Tax rate items (Y/N)
- A3. Any other business operating from same Address? (Y/N)
- A4. Any other business having same Tel No? (Y/N)

B. Return Compliance
- B1. Any Return default (Non filing)? (Y/N)
- B2. Delay in filing returns? (No of days)
- B3. No of returns that are NIL return ?

C. Returned Values & Ratios
- C1. Tax : Turnover
- C2. Gross profit %
- C3. Exempt Sales : Turnover
- C4. Inventory : Turnover
- C5. Purchases : Sales
- C6. Refund Claimed > Rs. 1000? (Y/N)
- C7. Tax credit carryforward > Rs 1000? (Y/N)
- C8. Output Tax or Input Tax Credit Adjustments > 1000? (Y/N)

D. Variations in the returns across tax-periods.
- D1. Turnover growth (compared to last year)
- D2. Tax Growth (compared to last year)
- D3. Variance of Turnover across tax-periods
- D4. Changes/variation in Sales mix (local/ interstate/ export)
- D5. Changes/variation in Purchase mix (local/ interstate/ import)
- D6. Changes/variation in product mix for sales (exempt, taxable at various rates)
- D7. Changes/variation in product mix for purchases (exempt, taxable at various rates)

E. Benchmarking vis-à-vis dealers of similar trade/industry, in respect of following parameters-
- E1. Tax: Turnover
- E2. Gross profit %
- E3. Taxable sales: Turnover
- E4. Inventory: Turnover
- E5. Turnover growth
- E6. Tax growth

The list of variables set out above is a wish-list of input variables, and many of them, although technically possible, could not practically be calculated from the database because of non-availability of data, data inconsistency, high programming & computer resource requirement etc. At the time of data-preparation/ extraction, many of them were required to be dropped because of practical considerations. The target variable in this exercise is the fact whether the dealer has under-reported tax or not. There are two options

for defining the target variable. It can be (a) the amount of tax under-reporting detected, or alternately, (b) the fact whether under-reporting has been detected (yes/no). Under the Sales Tax/VAT system, 96-97% of revenues come with the returns, and the Audits contribute only 3-4% of revenues directly. However, it should not undermine the importance of Audits. It is the fear of penalties in case of Audits that fetches the 96% of the revenues with the returns. The objective of the Audit strategy is not to maximize the revenues out of the audits, but to maximize the strike rate in Audits, so that fear of god is put in the tax-evaders. The Literature review suggests that in the IRS system, the latter, i.e. (b) as the target variable has yielded better results in the US. Therefore, target variable, TARGET was set at '1' for dealers detected to have been undeclared tax. It is set to '0' for dealers that are high taxpayers of the category, and nothing adverse has been detected against them.

4. Analysis and Results

Evaluation methodology

Any Audit Selection Strategy cannot hope to, and cannot practically catch all the tax-evaders. However, it is immaterial whether it is able to catch all the under-declarers/ evaders. The tax administration would like to deploy its resources where it is able to get high strike-rate, so that the fear of audits may automatically increase the tax revenues with the returns. In the process no harm would be done even if some guanine returns are selected for audit. Therefore, the objective of the strategy would be to catch the under-declarers and make an example of such catches. By this method, the fear of Audits will automatically weigh upon the tax-payers and the tax administration will realize higher revenues with the returns. The models therefore need to be evaluated on this basis. Based upon the actual fact whether the dealer is evading or not and prediction, the four possibilities (known as the Confusion Matrix) that emerge are shown in Table 1.

Table 1: Confusion Matrix

Predicted→ **Actual ↓**	**0** **(Tax Complying)**	**1** **(Tax-evading)**
0 (Tax Complying)	TN (True Negative)	FP (False Positive)
1 (Tax Evading)	FN (False Negative)	TP (True Positive)

Accordingly following indices are defined in line with the desired objectives of the Audit Strategy.
Prediction efficiency (PE) = percentage of tax-evasions cases correctly predicted by the model.

$$PE = TP / (TP + FN)$$

Reduction in examination effort (EF) = Reduction in the number of cases to be examined vis-à-vis the traditional method where all cases were being examined.

$$EF = 1 - (FP + TP)/ \text{Total cases}.$$

Strike Rate (SR) = percentage of cases where evasion is likely to be detected if predicted cases are taken for audit.

$$SR = TP / (TP + FP)$$

Since the prime objective is to most effectively deploy our resources, the most important parameter for model evaluation is Strike Rate (SR). However the prediction efficiency (PE) cannot be ignored either, because if the prediction efficiency of the model is low, the model itself is useless, and one can rather resort to random selection.

Construction of Models

Various data mining models were constructed and were evaluated based on the performance indices described above. These models are discussed in this section.

Model 1: Initially, Classification Tree Model was attempted by dividing the sample in training set and testing set. The data binned into binary data-set was utilised for the purpose. Initially, the model was run using the un-massaged (but binned) data, i.e., the data with the holes etc remaining as it is. The performance of the model is shown in Table 2.

Table 2: Evaluation Parameters for Models 1 to 6

Parameter	PE	EF	SR
Training Data	37%	89%	83%
Test Data	31%	89%	75%
Model 2			
Training Data	41%	88%	84%
Test Data	42%	85%	69%
Model 3			
Results same as Model 2			
Model 4			
Complete data	32%	90%	79%
Model 5			
Original Data	59%	72%	51%
Cross validated data	56%	71%	48%
Model 6			
Selecting 2 only	30%	82%	40%
Selecting 1 & 2	68%	63%	45%

The Model was found to have poor prediction efficiency. It could predict only 31% (in the test data) as positive (out of total positives). Also, some of the Classification Rules resulting out of model, like "Growth_of_Turnover is not available" does not make any sense for practical usability. Therefore this model was discarded.

Model 2: In order to improve upon the prediction efficiency, Another Classification Tree Model was attempted. This time the massaged data, i.e., the data with kinks and holes etc removed was used. The input variables were binned into 4 equi-depth groups. The resultant performance is shown in Table 2.

The prediction efficiency in this model improved to 41-42%. Also, the Strike rate of about 70% was fairly good for an effective audit strategy.

Model 3: The Classification tree models developed so far have prediction efficiency of about 40% only. Another Classification Tree Model was therefore attempted by reclassifying the target variable. Earlier it was 0 (Tax Complying) and 1 (Non-complying). The data was further classified to select dealers from whom recoveries of Rs 5.00 lakhs or more have been made. In respect of these dealers, the target variable was assigned a value of '2' and the classification tree model was constructed again. However, this model could not do any better then model 2 as it could not predict any of the 'high recoveries' dealers. The Model was therefore discarded.

Model 4& 5: As no further improvement in the prediction efficiency was obtained in the Classification tree models, and also considering the fact that many of the input variables were ratio variables, Logistic Regression and Discriminant Function were attempted. The performance is presented in Table 2.

The logistic regression model could not do any better then model 2 and both prediction efficiency and strike rate were lower. The Discriminant model on the other hand improved the prediction efficiency to about 57%, however, it was at the cost of Strike rate, which came down to around 50%. The model was

useful, but had a disadvantage of low strike-rate.

Model 6: The Discriminant model had achieved an improvement in prediction efficiency, but it was through increase in examination effort. The next model tried was Discriminant function only, but this time the target variable was coded at three levels as in model 3. The performance was presented in Table 2.
This Discriminant model has further improved the prediction efficiency to about 68%; however, it has further brought down the strike rate as well, and increased the examination effort. The model too was useful, but had a disadvantage of low strike-rate.

Model 7: So far the Discriminant model was better in prediction efficiency and the classification tree better in the Strike-rate parameter. In order to take advantage of both, a hybrid model was developed. The data was first processed through Discriminant Function Model (Model 6) and the predicted output of Discriminant function was put in classification tree model. This could be done in two ways. First, an independent classification tree could be trained and tested on this data, and second, the Classification tree prepared in Model 2 could be used to process the Discriminant function Output.

Initially, an independent Classification tree model was prepared from the Cases selected as '1' or '2' in the Discriminant Model (Model 6). There are 610 such cases predicted as '1' or '2'; these were divided in Training set and test set and the model was run.

Table 3: Evaluation parameters for Model 7 to 8

Parameter	PE	EF	SR
Model 7			
Out of Selected data	82%	55%	54%
Compared to total Data	36%	83%	54%
Model 8			
Out of Selected data	58%	70%	86%
Compared to total Data	39%	89%	86%

As may be seen, within the dataset of the Discriminant output, this model correctly predicted 82% of the cases, but the strike rate was low because of the high number of false positive cases that have passed through the model.

Model 8: The Classification tree in the Hybrid Model above was letting large number of FP cases pass. However the Classification tree in Model 2 had a very good strike rate. Therefore, the Classification tree model developed in Model 2 was applied to the selected cases of Discriminant output to see if it can achieve any improvement in the performance parameters. This hybrid model was found to have extremely good strike-rate and very low examination effort is required. It has an added advantage that in case resources permit, balance cases out of the Discriminant output can be taken up for Audits.

As mentioned earlier, for the purpose of Audit Strategy, the model should have a reasonable Prediction efficiency but a good strike rate. Figure 1 below presents a comparative analysis of the performance indices of the Models. As may be seen, there is a trade off between the two parameters.

An interesting comparison of the efficiency of the model is with random selection. If a random sample is chosen from our dataset, since there are 3 tax complying dealer for every tax-evading dealer in the sample, it would achieve a strike rate of 25%. It is observed that all the models are giving much better strike rate than random selection. Chart below compares the performance efficiency of the model with hypothetical performance efficiency if same size sample as that in the model is chosen randomly instead. The

performance of the models are always much better than random selection. In Model 8, the performance efficiency has improved almost 3 ½ times.

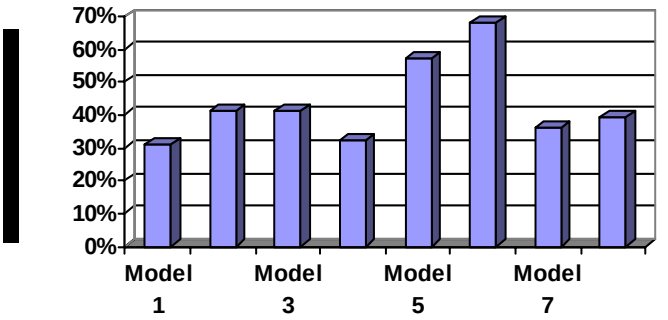

Figure 1a: Performance Efficiency for Different models

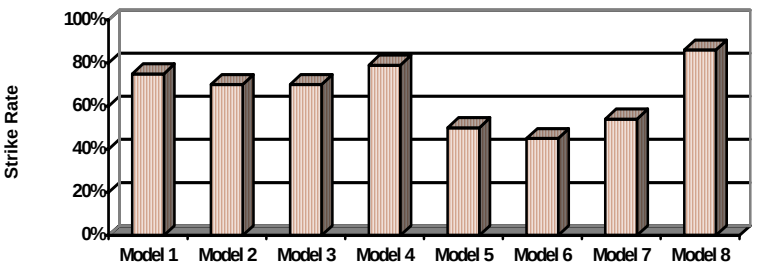

Figure 1b: Strike Rate for Different models

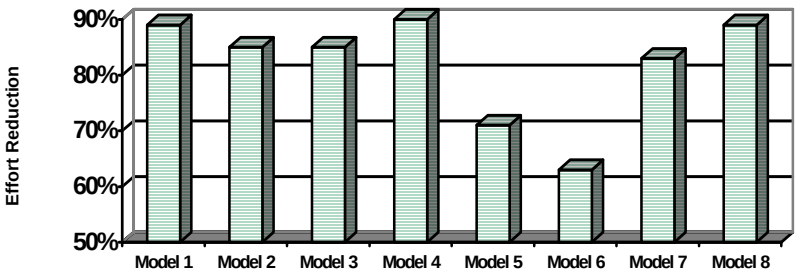

Figure 1c: Effort Reduction for Different models

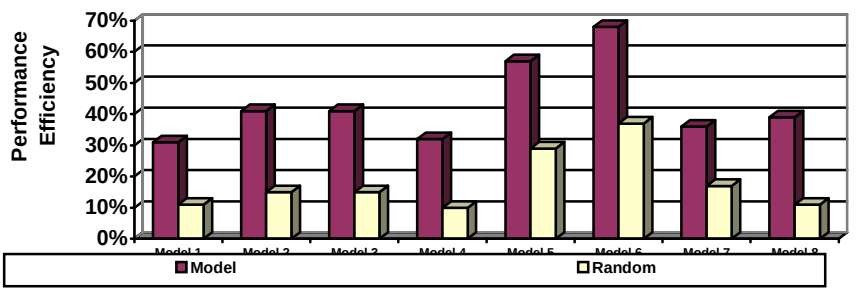

Figure 2: Comparison of Efficiency of Different Models With Random Selection

5. Concluding Remarks

There is a trade-off between strike rate and performance efficiency and the models studied could-not achieve both simultaneously. However, as has been mentioned earlier, the objective of Audit Strategy is not to catch all the evaders but to effectively deploy our resources so that the 'fear of audits' gets the department better taxes with the return. Therefore, in selecting a model, one can compromise on the Performance efficiency in quest for a better strike rate. Model 2 (Classification tree) and Model 8 (Hybrid) both have a good Strike rate (70% and 86%) and a reasonable performance efficiency (41% and 39%). Further performance of the models can be improved by including additional input variables, which had been dropped due to various reasons in this study. All the models developed through Data-mining technique were better than random selection. The improvement in the strike rate (compared to random) being 2 times in case of Discriminant function; 2.5 times in case of classification tree; 3 times in case of logistic regression and 3.5 times in case of hybrid model.

References

1. Alm, James, Blackwell C. & McKee Michael. (2004) Audit Selection and Firm Compliance with a Broad-based Sales Tax. *National Tax Journal*, 57, (2), 1, 209-27.
2. Bonchi F, Giannoti F, Mainetto G, Pedreschi D. (1999) A Classification – Based Methodology for Planning Audit Strategies in Fraud Detection, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM 1999, 175-184.
3. Cecil, Wayne H. (1998) Assuring Individual Taxpayer Compliance: Audit rates, Selection Methods, and Electronic Auditing. *The CPA Journal*, 68, (12), available at http://www.nysscpa.org/cpajournal/1998/ 1198/ Departments/D661198.html, last accessed 27 Sep 2007.
4. Ho, Daniel, Peter Lau. (1999) Tax Audits in Hong Kong. *The International Tax Journal New York*, 25, (3), 61-71.
5. Fisher, Vickie L. (1985) Recent Innovations in State Tax Compliance Programs. *National Tax Journal*. 38, (3), 365-71.
6. Kumar, Anuj and V. Nagadevara. (2006) Development of Hybrid Classification Methodology for Mining Skewed Data Sets – A Case Study of Indian Customs Data, *Proceedings of the 4th ACS/IEEE International Conference on Computer Systems and Applications*, Mar 8-10, 2006, Sharjah, UAE.
7. Manasan Rosario G. (2003) Estimating Industry Benchmarks for the Value Added Tax. *Philippine Journal of Development*, 55, (30), 1, 71-90.
8. Micci-Barreca Daniele, Ramachandran Satheesh. (2006) Analytics Elite. Improving Tax Administration with Data Mining., http://www.spss.com Last accessed April 26, 2006
9. Micci-Barreca Daniele, Ramachandran Satheesh. (2006) Analytics Elite. Predictive Tax Compliance Management. http://www.spss.com Last accessed April 26, 2006,
10. Murray, Mathew N. (1995) Sales Tax Compliance and Audit Selection. *National Tax Journal*. 48, (4), 515-30.
11. Wedick, John L. (1983) Looking for a Needle in a Haystack – How the IRS Selects returns for Audit. *The Tax Advisor*, New York, 14, (11), 675-675.
12. Phua C, Alahakoon D, Lee V. Minority Report in Fraud Detection: Classification of Skewed Data, SIGKDD Explorations, Vol 6, (1), 50-59.
13. Shao H, Zhao H, Chang G. (2002) Applying Data Mining to Detect Fraud Behavior in Customs Declaration, *Proceedings of the International Conference on Machine Learning and Cybernetics*, Vol 3, IEEE 2002, 1241-1244
14. Wiersema, William H. (1997) Will the IRS Audit you? *Electrical Apparatus*. Chicago: Vol 50, (3), 38-40.

About Authors

Manish Gupta, is an officer of the Indian Administrative Service currently working as Director in Government of India. He is a B.Tech from IIT Kanpur, M.Tech from IIT Delhi and has done PGD in Public Policy & Management from IIM Bangalore. In the past, he has worked as Additional Commissioner and Commissioner of Sales Tax in Delhi and Arunachal Pradesh. He has also worked extensively with the Empowered Committee of Finance Ministers during 1998 – 2002 in formulating the VAT policy recommendations.

Vishnuprasad Nagadevara, obtained his Ph D from Iowa State University, Ames Iowa. He is currently Professor in the Quantitative Methods and Information Systems Area at the Indian Institute of Management Bangalore. His current research interests are Data Mining, Application of Management Techniques to Public Policy and Entrepreneurship.