APPLIED ARTIFICIAL INTELLIGENCE
2018, VOL. 32, NO. 1, 48–64
https://doi.org/10.1080/08839514.2018.1451032

Fraudulent Firm Classification: A Case Study of an External Audit

Nishtha Hooda, Seema Bawa, and Prashant Singh Rana

Computer Science and Engineering Department, Thapar University, Patiala, India

ABSTRACT

This paper is a case study of visiting an external audit company to explore the usefulness of machine learning algorithms for improving the quality of an audit work. Annual data of 777 firms from 14 different sectors are collected. The Particle Swarm Optimization (PSO) algorithm is used as a feature selection method. Ten different state-of-the-art classification models are compared in terms of their accuracy, error rate, sensitivity, specificity, F measures, Mathew's Correlation Coefficient (MCC), Type-I error, Type-II error, and Area Under the Curve (AUC) using Multi-Criteria Decision-Making methods like Simple Additive Weighting (SAW) and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). The results of Bayes Net and J48 demonstrate an accuracy of 93% for suspicious firm classification. With the appearance of tremendous growth of financial fraud cases, machine learning will play a big part in improving the quality of an audit field work in the future.

Introduction

Fraud is a critical issue worldwide. Firms that resort to the unfair practices without the fear of legal repercussion have a grievous consequence on the economy and individuals in the society. Auditing practices are responsible for fraud detection. Audit is defined as the process of examining the financial records of any business to corroborate that their financial statements are in compliance with the standard accounting laws and principles (Cosserat 2009). It is a very exacting task to detect firms in spotting frauds, detecting errors, and disclosing employees guilty of abetting illegal transactions. Data analytics tools for an effective fraud management have become the need of the hour for an audit. The possibilities that how data analytics can improve the quality of the process is published in Emerging Assurance Technologies Task Force of the AICPA Assurance Services Executive Committee (ASEC) (Staff 2014). Generally, audits are classified into two categories as internal and external auditing (Cosserat 2009). Internal-audit, although is an

CONTACT Nishtha Hooda 27nishtha@gmail.com 1119-B, Sector 41-B, Chandigarh-160036
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uaai.

independent department of an organization, but resides within the organization. These are company-employees who are accountable for performing audits of financial and nonfinancial statements as per their annual audit plan. External audit is a fair and independent regular audit authority, which is responsible for an annual statutory audit of financial records. The external audit company has a fiduciary duty and is critical to the proper conduct of business. For instance, their work is to audit the receipts, expenditures, accounts related to the trading, profit, contingency funds, balance sheets, public accounts, etc. kept in any government office. It is their duty to ensure that the funds allocated to any government department have been put to use as per law. On successful completion of an audit process, auditors deliver an audit and inspection summary report called *audit paras* to the company comprising of the details of all the findings from the audit. This may include discrepancies, noncompliance of accounting rules, leakage of revenue, inaccurate calculations, etc. The whole audit process flow is summarized in Figure 1. In order to improve the advancement of data analytics tools for auditing, AICPA and Rutgers Business School in 2015 announced a research initiative on how data analytics can improve the quality of an audit

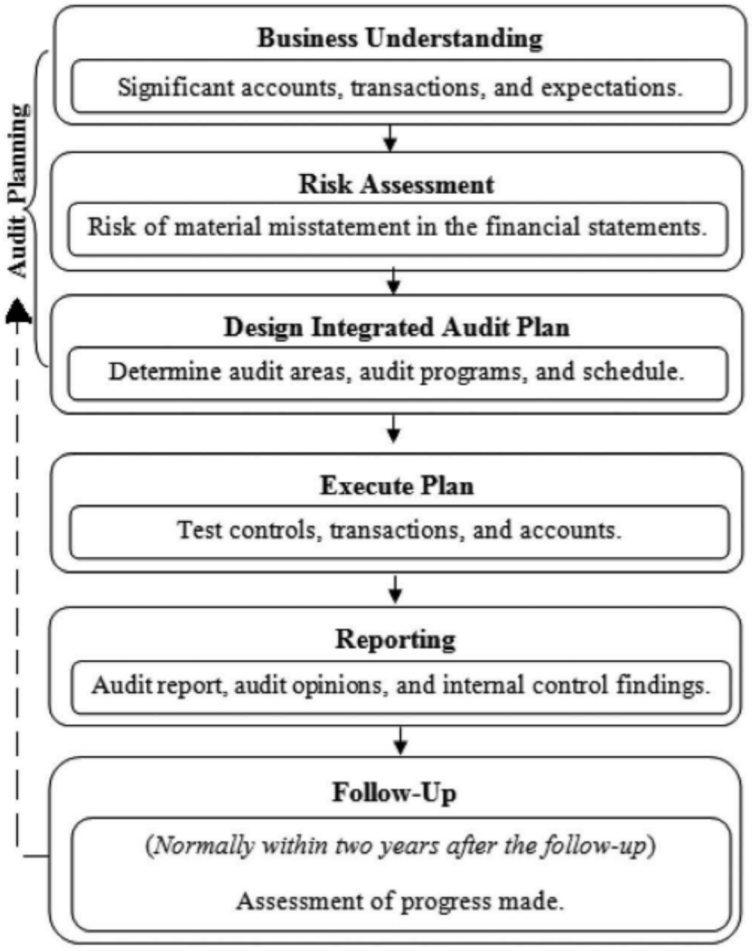


Figure 1. Audit work-flow.

(Tschakert 2016). Auditing Standards Board Task Force (ASBTF) is also working on developing an innovative *Audit Data Analytics Guide* in order to integrate the data analytics tools for the auditing tasks (Maria, Murphy and Tysiac 2015).

Machine learning has got much attentiveness in the data analytics as it offers new computational as well as epistemological techniques to produce better results. Machine learning proposes several algorithms that are derived from the area of statistics and artificial intelligence. Many researchers have employed algorithms like artificial neural network, logistic regression, decision trees, and Bayesian belief networks for detecting management fraud in the financial statements (Fanning and Cogger 1998; Green and Choi 1997; Spathis 2002). The ensemble machine learning method is also applied successfully for improving the classification accuracies of the auditing task (Kotsiantis 2006). Machine learning algorithms like support vector machine, logistic regression, probabilistic neural network, genetic algorithm, etc. are also combined with feature selection methods in order to prove their usability in detecting fraud in the Chinese firms (Ravisankar 2011). In a review of data analytics tools like for fraud prediction, clustering, and outlier detection that are used for fraud management task, researchers listed algorithms like neural network, decision tree, Bayesian network, etc. as most commonly used methods (Sharma 2013).

The prime goal of an auditor during an audit-planning phase is to follow a proper analytical procedure to impartially and appropriately identify the firms that resort to high risk of unfair practices. Predictive analytics is also implemented using machine learning methods because it provides actionable insights for the audit companies. One of the most common applications of predictive analytics in audit is the classification of suspicious firm. Identifying fraudulent firms can be studied as a classification problem. The purpose of classifying the firms during the preliminary stage of an audit is to maximize the field-testing work of high-risk firms that warrant significant investigation. According to a research, data analytics has benefited internal auditing more as compared to advancements it has contributed for the external audits (Tysiac 2015). This research work is a case study of an external government audit company which is also an external auditor of government firms of India. During audit-planning, auditors examine business of different government offices but target to visit the offices with very-high likelihood and significance of misstatements. This is calculated by assessing the risk relevant to the financial reporting goals (Houston, Peters, and Pratt 1999). The three main objectives of the study are as follows:

- To understand the audit risk analysis work-flow of the company by in-depth interview with the audit employees, and to propose a decision-making framework for risk assessment of firms during audit planning.
- To examine the present and historical risk factors for determining the Risk Audit Score for 777 target firms, to implement the Particle Swarm Optimization (PSO) algorithm to rank examined risk factors, and evaluating the Risk Audit Class (Fraud and No-Fraud) of nominated firms.
- To explore and test the applicability of 10 classification models in the prediction of a Risk class, and to check the collective performance of the models for the fraud prediction using Multi-Criteria Decision-Making (MCDM) methods like SAW, TOPSIS, etc.

Ten machine learning classifiers are explored, implemented, and tested by K-fold cross validation to check their applicability in the prediction of the high risk firms. The rationale is to build a classification model that can predict the fraudulent firm on the basis of the present and historical risk factors. Through the rigorous experiments, it has been found that Bayes Net and J48 classifiers outperform over the other machine learning methods in terms of their performance in predicting the risk-class (Fraud or No-Fraud) for an Audit Field Work Decision Support System. The rest of the paper is organized as follows: Section 2 presents the audit dataset, considered features, risk assessment procedure, and the complete methodology. Experiments, performance evaluation, and result-analysis are discussed in Section 3. Finally, the conclusion and future scope are discussed in Section 4.

Materials and methods

Data collection

Comptroller and Auditor General (CAG) of India is an independent constitutional body of India. It is an authority that audits receipts and expenditure of all the firms that are financed by the government of India. While maintaining the secrecy of the data, exhaustive one year nonconfidential data (2015 – 2016) of firms is collected from the Auditor General Office (AGO) of CAG. There are total 777 firms from 46 different cities of a state that are listed by the auditors for targeting the next field-audit work. The target-offices are listed from 14 different sectors. The information about the sectors and their counts are summarized in Table 1.

Features

Many risk factors are examined from various areas like past records of audit office, audit-paras, environmental conditions reports, firm

Table 1. Target sectors.

Sector ID	Target sector	Information	Number of target firms
1	IR	Irrigation	114
2	P	Public Health	77
3	BR	Buildings and Roads	82
4	FO	Forest	70
5	CO	Corporate	47
6	AH	Animal Husbandry	95
7	C	Communication	1
8	E	Electrical	4
9	L	Land	5
10	S	Science and Technology	3
11	T	Tourism	1
12	F	Fisheries	41
13	I	Industries	37
14	A	Agriculture	200

Table 2. Risk factors classification and other features in model.

Inherent risk factors		Control risk factors	
Feature	Information	Feature	Information
Para A value	Discrepancy found in the planned-expenditure of inspection and summary report A in Rs (in crore).	Sector score	Historical risk score value of the target-unit in the Table 1 using analytical procedure.
Para B value	Discrepancy found in the unplanned-expenditure of inspection and summary report B in Rs (in crore).	Loss	Amount of loss suffered by the firm last year.
Total	Total amount of discrepancy found in other reports Rs (in crore).	History	Average historical loss suffered by firm in the last 10 years.
Number	Historical discrepancy score.	District score	Historical risk score of a district in the last 10 years.
Money value	Amount of money involved in misstatements in the past audits.		
Other features			
Feature	Information	Feature	Information
Sector ID	Unique ID of the target sector.	Location ID	Unique ID of the city/province.
ARS	Total risk score using analytical procedure.	Audit ID	Unique Id assigned to an audit case.
Risk class	Risk Class assigned to an audit-case. (*Target Feature*)		

reputation summary, on-going issues report, profit-value records, loss-value records, follow-up reports etc. After an in-depth interview with the auditors, important risk factors are evaluated and their probability of existence is calculated from the present and past records. Table 2 describes the various examined risk-factors that are involved in the case study. Various risk factors are categorized, but combined audit risk is expressed as one function called an Audit Risk Score (ARS) using an audit analytical procedure. At the end of risk assessment, the firms with high ARS scores are classified as "*Fraud*" firms, and low ARS score companies are classified as "*No-Fraud*" firms. Sample audit data of the corporate sector are shown in Table 3.

Table 3. Sample data of the corporate sector unit.

Audit ID	Loc ID	Para A	Para B	Total	Sector score	Loss Score	Num.	Money value	History score	District score	ARS	*Risk class*
26	4	5.78	57.92	73.70	3.89	0	5	11.16	1	2	4	F
27	4	7.42	2.24	19.66	3.89	1	1	1.25	2	2.5	2.4	F
28	4	0	1.10	4.11	3.89	0	3	0.007	2	2	2	N
29	14	6.85	31.76	58.61	3.89	2	5	1.46	1	4	3.6	F
30	14	0	1.03	5.03	3.89	0	5	0	2	2	1	N
31	37	0	0.75	3.75	3.89	0	5	6.78	2	2	2.2	F
32	37	2.4	16.63	29.73	3.89	0	3	1.16	0	4	3.6	F
33	5	0	0.05	1.23	3.89	1.3	5	152.41	2	2	2.4	F
34	5	0	1.76	4.76	3.89	0	2	1.08	2	2	2	N
35	5	0	2.97	6.97	3.89	0	5	2.84	1	2	2	N

Audit Risk Assessment (ARA)

Audit Risk Assessment (ARA) is a deliberate process of evaluating the likelihood of discrepancies or misstatements (event E) (Nikolovski et al. 2016). Risk is often measured as the expected value of any an unenviable outcome. During audit-planning, external auditors first quantitatively evaluate the risk of fraud in an organization in order to estimate the need for audit-field work. As Event E is a negative event, so historical data are also analyzed. For calculating the probability of discrepancies or misstatements (event E), formal methods are preferred. The associated formula for calculating the risk R is expressed as

$$R = (P_\xi(L_\xi)) \tag{1}$$

where P_ξ is the probability of discrepancy and (L_ξ) is the loss involved in the discrepancy. Due to different categories of risk, situations in an audit are sometimes more complex than the simple possibility case of one risk. In a situation with several possible risk types, the total risk is the sum of the different risk type and can be expressed as

$$R = \sum_i (P_\xi(L_\xi)) \tag{2}$$

where i is the total number of considered risk types.

When the audits are performed by any external audit company, the risk assessment plays a vital role in deciding the amount of field work that would be required before actually visiting the official firms. According to ISA315, an auditor should always obtain a clear understanding of the firm including all its internal environments, controls, entities, etc. for a complete risk assessment before actually visiting the firm (of Certified Public Accountants (AICPA) 2006). This process acts as initial evidence for performing an effective audit at client's firm. As a formula, audit risk is the product of inherent risk (IR), control risk (CR), and detection risk (DR) (Srivastava and Shafer 1992). It can be calculated as

$$\begin{aligned} AR &= IR \times CR \times DR \\ &= (Combined\,Risk) \times DR \\ &= (P_{IN}(L_{IN}) \times P_{CO}(L_{CO}) \times DR \end{aligned} \tag{3}$$

Inherent Risk (IR) is the risk present due to the discrepancies present in the transactions. For instance, transaction which involves settlement by checks has lower IR as compared to the transaction which involves exchange of cash. CR is the risk due to the discrepancies which are left undetected by an internal control system. For instance, CR risk is high when the separation of duties is not properly defined. DR is the risk of discrepancies present in the firm which are not even detected by the audit procedures. Human or sampling error, for instance. Considering all risk factors, a complete equation for evaluating an audit risk using Equation (2) and Equation (3) can be expressed as

$$AR = \sum_{i=1}^{\alpha}((P_{IN}(L_{IN}) \times \sum_{i=1}^{\beta}((P_{CO}(L_{CO}) \times DR\{\eta = \alpha + \beta\} \tag{4}$$

where α and β are the number of risk factors causing inherent risk and control risk, respectively. For this case study, the complete equation for the risk factors (risk factors categorized in Table 2) can be expressed as

$$\begin{aligned} AR = (((&P_{PARA\ A}(L_{PARA\ A})) + ((P_{PARA\ B}(L_{PARA\ A})) + ((P_{Total}(L_{Total})) \\ &+ ((P_{Number}(L_{Number})) + ((P_{Money\ Value} + (L_{Money\ Value})) \\ &\times ((P_{Sector\ Score}(L_{Sector\ Score})) + ((P_{District}(L_{District})) + ((P_{History}(L_{History})) \\ &+ (P_{Loss}(L_{Loss}))) \times DR \end{aligned} \tag{5}$$

For calculating the audit risk of a firm, the probability of each risk factor is calculated using an analytical procedure and an audit risk score is calculated for each firm. In order to understand the complete step-by-step process, it is presented as a *Risk Assessment Algorithm* 1.

Proposed framework

The goal of the research is to design and develop a prediction model for the proposed audit field work decision support framework. The proposed framework which can also work as a *Decision-Making System* is presented in an abstract view in Figure 2.

The selected features (as described in Table 2) are used as candidates for the input vector of the model. The outcome of the proposed framework will be available in the form of a web-based application that helps an auditor to predict an audit risk class (Fraud or No Fraud). The complete flow of the prediction model for the proposed audit field work decision support framework is described in Figure 3 and discussed in this section.

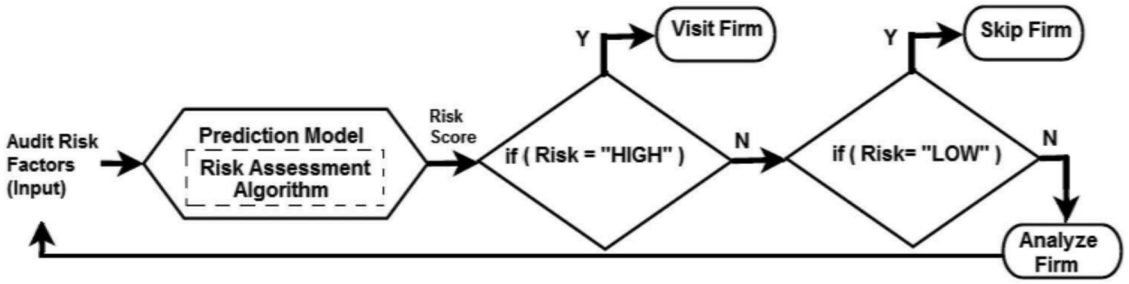


Figure 2. Proposed framework for an audit field work decision-making.

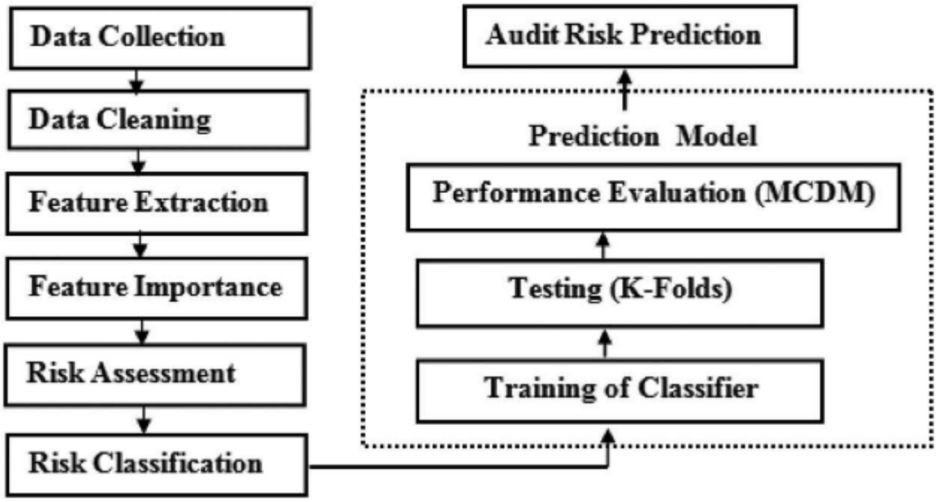


Figure 3. Prediction model.

Algorithm 1 Risk Assessment Algorithm

Input: Agenda list A.
Output: Risk class of each sample in the agenda list A.
Comment: {Let X denote the number of offices to be examined. Assume X is finite. Let A is a list of nominated set of offices called Agenda A, extracted from set X (Here Agenda is a set of all offices that are not being visited since last three years).}

1. **Data collection**: Collect the unstructured data of all the offices under Agenda A.
2. **Feature extraction**: Examine k features (risk factors) ξ_k that may be needed for the inherent and control risk assessment.
3. **Risk calculation**:
   - Calculate the loss L_ξ that may be involved for each risk factor.
   - Calculate the probability of loss (p_ξ) for each risk factor ξ.
   - Calculate the risk (R_ξ) for each risk factor ξ as $R = (P_\xi(L_\xi))$.
4. **Risk classification**: Classify the risk factors into inherent-risk class(IN) and control risk class (CO) and calculate the sum of risk for each class as Sum_{IN} and SUM_{CO}, respectively.
5. **Combined risk**: Calculate the combined risk of class CO and IN as $Combined\ Risk = (SUM_{IN}) \times (SUM_{CO})$
6. **Detection risk**: Define the detection risk value DR.
7. **Audit risk score**: Calculate the audit risk as the $Audit\ Risk = ((Combined\ Risk) \times DR)$
8. **Risk assessment**: Calculate the average of audit risk as $Audit_{avg}$. Classify the audit risk a_i for each audit case as high (fraud class) and low (no fraud class) by the following rules:
   - *if* the audit risk $(a_i <= 1)$, label it as **No Fraud**.
   - *else* label it as **Fraud**.

Data cleaning and feature extraction

In real applications, the learning algorithm is rarely such powerful and perfect. Data suffer from noise, missing values, errors, inconsistencies, class-imbalance problem etc. After cleaning and preparing the collected unstructured data from numerous files, different types of risk factors are explored. The data are organized in 777 rows and 9 important risk factors (columns).

Feature importance and selection

Proposed by Kennedy and Eberhart, Particle swarm optimization (PSO) is one of the simple and widely preferred optimization techniques (Kennedy 2011, Couceiro 2016). PSO is a heuristic global optimization technique and is successfully used as a feature importance and selection method in many physical problems (Kothari 2012, Couceiro 2016). Table 4 gives the optimal weights to each risk factor according to Equation (6) using Particle Swarm Optimization (PSO) as described below:

$$ObjectiveFunction = min\left(\sum_{i=1}^{T}\sqrt{\left(\alpha_i - \sum_{j=1}^{n} w_j.F_{i,j}\right)^2}\right) \tag{6}$$

where T is the total number of instances in training dataset, α is the Audit Risk Score (ARS) value of training dataset, n is the number of features, F is the feature, and w is the weight given to each feature defined in [0,1]. The weight of F3 is the lowest, so it can be eliminated.

The final formula used for all the machine learning models is given by

$$Fraud \sim f(Para\,A, Para\,B, Number, Money\,Value, Sector\,Score, District, History, Loss) \tag{7}$$

Table 4. Risk factors weight assignment using Particle Swarm Optimization (PSO).

Feature ID	Feature name	PSO weight	Feature rank
F1	Para A value	0.18	5
F2	Para B value	0.16	6
F3	Total	0	9
F4	Money value	0.01	8
F5	Number	0.07	7
F6	Sector score	0.21	4
F7	History score	0.69	2
F8	District score	0.50	3
F9	Loss score	0.93	1

Classification models

Identifying a firm is fraudulent or not using an input vector (risk factors) can be considered as a binary classification problem. Ten state-of-the-art classification methods are employed in the case study are discussed in this section.

- **Decision Trees (DT)**: This model is an extension of the C4.5 classification algorithm. It is described by Quinlan and works by classifying samples by sorting them down the tree (Quinlan 1986).
- **AdaBoost (AB)**: It is a successful ensemble classifier by Schapire and Freund. It employs multiple learners to finally make a more powerful learning algorithm (Schapire 1999).
- **Random Forest (RF)**: It is an ensemble learning algorithm which builds a forest of decision trees using random inputs to improve the classification rate (Liaw and Wiener 2002).
- **Support Vector Machine (SVM)**: SVMs search for data points that are present at the edge of an area in a space (boundary between two classes) and refer them as support vectors. It is a preferred technique for classification (Keerthi and Gilbert 2002).
- **Probit Linear Models (PLM)**: The linear model is a traditional regression method for fitting the data. For binary classification, it is transformed using a logistic or probit function and offers similar results to the logistic regression (Chambers 1977; Finney 1992).
- **Neural Network (NN)**: It is inspired from biological neural networks and used to model complex relationships, and useful patterns in statistical data (Russell et al. 2003).
- **Decision Stump Model (DSM)**: It is a one-level decision tree. It is also used as a base learner in ensemble models (Iba and Langley 1992).
- **J48**: It builds decision tree based on the theory of information entropy. J48 is an open source java implementation of the C4.5 algorithm (Quinlan 1996).
- **Naive Bayesian (NB)**: It computes the conditional a posterior probabilities of a categorical class variable of a given independent predictor variable using the Bayes Rule (Rish 2001).
- **Bayesian Network (BN)**: This model is based on probabilistic and directed acyclic graph theory. It builds a graphical model that represents a set of features and their conditional dependencies via a directed acyclic graph (Buntine 2016; Neapolitan et al. 2004).

Performance evaluation

To check the performance of 10 classifiers, K fold (K = 10) validation is implemented and 10 performance metrics namely Type-1 error, Type-2 error, accuracy, error rate, sensitivity, specificity, AUC, F1 score, MCC, and

F2 score are evaluated using the results of the confusion matrix as described in Table 6. The confusion matrix commonly known as the *Error Matrix* in machine learning is a popular method to measure the performance of a binary classification model on a test dataset. The classifiers have been trained to distinguish between "*Fraud*" and "*No-Fraud*" firms, and the confusion matrix summarizes the results as described in Table 5 for class Fraud. Here, X (true positive) is equivalent to the hits. X gives the number of actual fraudulent firms predicted correctly as fraudulent. Z (false negative) gives the number of fraudulent firms that are incorrectly marked as nonfraudulent firms. It is equivalent to the miss and commonly known as the *Type-II error*. Q (false positive) gives the count of nonfraudulent firms that are incorrectly labeled as fraudulent. It is known as the Type-I error. Y (true negative) indicates the number of nonfraudulent firms that are correctly classified as nonfraudulent. The accuracy and error rate are commonly used metric to test the performance of a classifier. The accuracy calculates the number of correct predictions from all predictions. Sensitivity is the true positive rate (TPR), measuring the hit rate of a classifier in predicting fraudulent firms. Specificity measures the true negative rate (TNR) of a model. The area under the curve (AUC) is equal to the probability that a classification model will rank a randomly chosen positive sample (fraud class sample) higher than a randomly chosen negative sample (nonfraud class sample). A graph is generated by plotting TPR against the FPR, depicting the relative trade-offs between true positive and false positives (Fawcett 2006). The accuracy is sometimes misleading and AUC (the area under the curve) is a more preferred approach as compared to accuracy (Bradley 1997). F measure (F1 score) is a balanced score of sensitivity and specificity. F2 score weights sensitivity value higher

Table 5. Confusion matrix.

	True reference	
Predicted condition	Fraud	No fraud
Fraud	True positive X	False negative Z
No fraud	False positive Q	True negative Y

Table 6. Performance evaluation metrics.

Performance metric	Formula
Type-I error	Q
Type-II error	Z
Sensitivity	$X/(X+Z)$
Specificity	$Y/(Q+Y)$
Accuracy	$(X+Y)/(X+Z+Q+Y)$
Error rate	$1-Accuracy$
F1 score	$(2*X)/(2*X)+(Q+z)$
F2 score	$(5*X)/((5*X)+(4*Z)+Q)$
MCC	$(X*Y)-(Q*Z)/SQRT((X+Q)+(X+Z)+(Y+Q)+(Y+Z))$

than specificity. Matthew's correlation coefficient (MCC) is another balanced measure that focuses on true and false positives and negatives.

Experiments and result analysis

Ten machine learning models are implemented for prediction of an audit risk class (fraud or no-fraud). To test the robustness of designed the framework, K fold cross validation method ($K = 10$) is implemented and performance of the models are compared using 10 different performance metrics. The results obtained from the 10 different metrics (discussed in the Section 2.4.4) are analyzed and collective performance scores of all the classifiers are calculated using two different multi-criteria decision-making methods TOPSIS and SAW are discussed here.

Experiment setting

The R "caret" package is used to implement the various classification models. The models are available in R open source software. R is licensed under GNU GPL. The brief implementation details of the models discussed in Section 2.4.3 are summarized in Table 7. One shot training and testing technique is adopted here. Training data i.e. 70 percent of the sample data are fed to enable the classifier. The model is trained on these training data and then tested on independent new samples (30 percent of the samples) of testing data. The purpose is to measure the prediction performance of the model when it is up and running and then predicting the risk-class of the new samples without the benefit of knowing the true risk-class of the samples.

Table 7. Machine learning classification methods.

Model	Method	Package	Tuning parameter	Reference
DT	rpart	rpart	MinSplit = 20, MaxDepth = 30, MinBucket = 7	(Quinlan 1986)
AB	adaboost	fastAdaboost	Default	(Schapire 1999)
RF	rf	randomForest	mtry = 500, sampling = bagging	(Liaw and Wiener 2002)
SVM	ksvm	e1071	nu = 10, epsilon = 0.5	(Keerthi and Gilbert 2002)
PLM	bayesglm	arm	Default	(Chambers 1977, Finney 1992)
NN	neuralnet	nnet	size = 10, linout = TRUE, skip = TRUE, MaxNWts = 10000,trace = FALSE, maxit = 100	(Russell et al. 2003)
DSM	decision stump	RWeka	rules = 6,pruned = 25,smoothed = 0.9	(Iba and Langley 1992)
J48	J48	RWeka	Default	(Quinlan 1996)
NB	NaiveBayes	RWeka	Default	(Rish 2001)
BN	BayesNet	RWeka	Default	(Pearl 1985, Neapolitan et al. 2004)

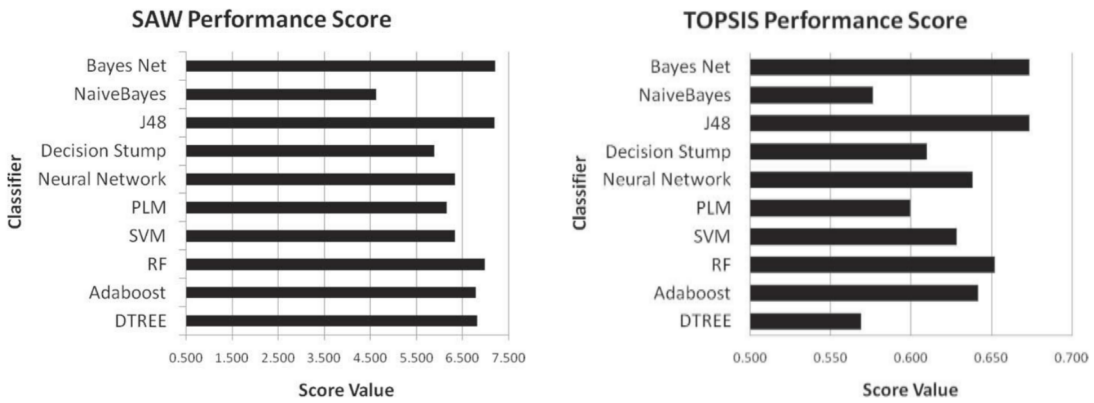
Table 8. Average performance comparison of machine learning methods for the prediction of an audit risk on testing dataset.

Classifier	Type-I error	Type-II error	Sens.	Spec.	Acc.	Error rate	F1 score	F2 score	MCC	AUC
DT	0.09	0.04	0.88	0.91	0.89	0.11	0.91	0.89	0.78	0.95
AB	0.05	0.11	0.87	0.95	0.91	0.09	0.92	0.89	0.81	0.94
RF	**0.01**	0.12	0.88	**0.99**	0.92	0.08	0.93	0.90	0.85	0.96
SVM	0.10	0.14	0.85	0.90	0.87	0.13	0.89	0.86	0.73	0.92
PLM	0.14	0.11	0.84	0.86	0.85	0.15	0.87	0.84	0.68	0.91
NN	0.07	0.12	0.84	0.93	0.88	0.12	0.89	0.86	0.76	0.78
DSM	0.08	0.22	0.78	0.92	0.84	0.16	0.85	0.81	0.69	0.89
J48	**0.01**	**0.09**	**0.91**	**0.99**	**0.94**	**0.06**	**0.95**	**0.92**	**0.88**	0.96
NB	0.05	0.40	0.60	0.95	0.74	0.26	0.73	0.65	0.55	0.93
BN	**0.01**	**0.09**	**0.91**	**0.99**	**0.94**	**0.06**	**0.95**	**0.92**	**0.88**	**0.97**

Performance score and result analysis

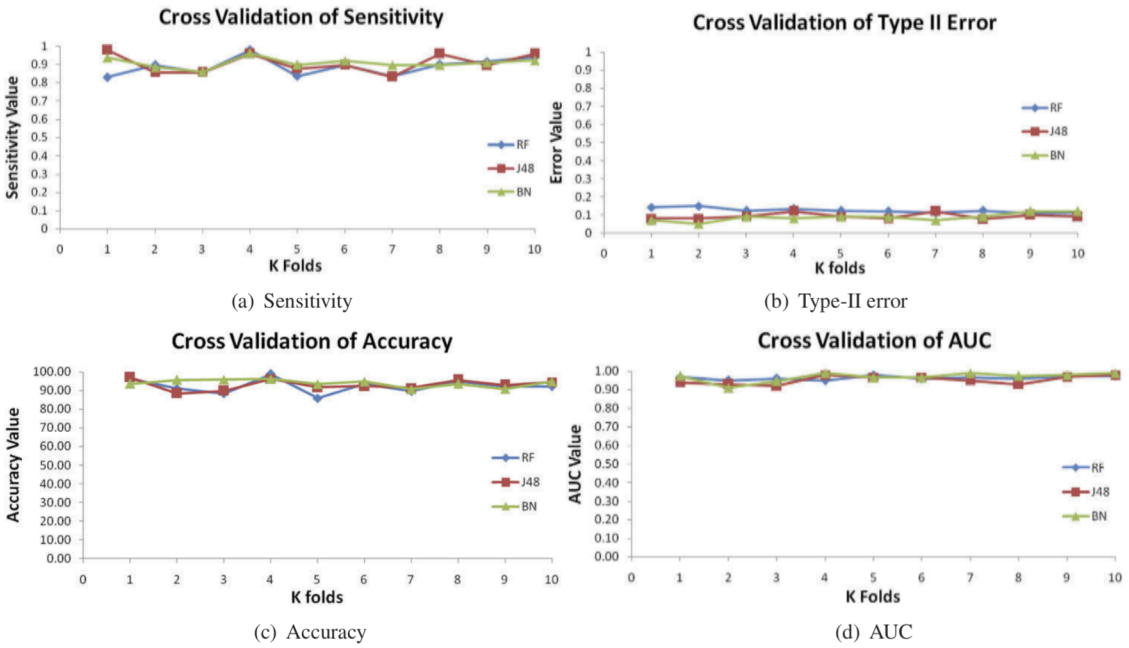
The performance of the proposed framework is evaluated with 10 different performance parameters, and the results are summarized in Table 8. Researchers have proved that there is no classifier present that works perfectly for all the data problems (Ali and Smith 2006). Similarly, there are numerous measures to evaluate the performance of classifiers and there is no best metric for all the classification problems (Ali and Smith 2006; Smith-Miles 2009). Based on these two important considerations, this case study uses 10 performance metrics to compare 10 state-of-the-art classifiers. In an attempt to perform comprehensive performance evaluation, classifiers are evaluated against multiple criteria or metrics. For selecting the best classification algorithms against multiple criteria, selection problems can be modeled as the Multi-Criteria Decision-Making (MCDM) problem (Triantaphyllou 2013) . For unbiased ranking of 10 classifiers using MCDM methods, Simple Additive Weighting (SAW) and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) methods are implemented. The reason for choosing these methods is their simplicity and involvement of subjective ranking of different performance criteria by experts.

Different metrics are analyzed and their importance (weight of evaluation criteria) is judged for the prediction of an audit risk. It is observed that the proposed framework utilizes false positives routinely in order to predict firms that are under the risk of frauds. By examining historical and present information (summarized in Table 2), the framework should have high sensitivity (the hit rate of predicting fraudulent firms). In the present scenario, the situation of not detecting the firm as fraud (low sensitivity) could be menacing for auditors whilst the low specificity (predicting honest firm as fraud) may only cause a further inspection. So, higher relative importance is given to sensitivity than the specificity. Similarly, the F score measure (balanced measure of sensitivity and specificity) is given lesser preference than the F2 score. Accuracy, error rate, and AUC are equally important but while analyzing the Type-I error (incorrect prediction of fraud firm) and



(a) TOPSIS (b) SAW

Figure 4. Performance score of SAW and TOPSIS MCDM methods.



(a) Sensitivity (b) Type-II error

(c) Accuracy (d) AUC

Figure 5. Ten-fold cross validation of Type-II error, sensitivity, accuracy, and AUC on the testing dataset in the audit risk prediction using Bayes Net, J48, and Random Forest.

Type-II error (failing to detect the fraud firm), the Type-II error is given more preference than the Type-I error. The collective score of SAW and TOPSIS are graphically represented in Figure 4. It is reflected in the figure that J48 and Bayes Net outperform the other nine models in terms of the overall performance. Besides J48 and Bayes Net, Random Forest has also achieved a satisfactory performance on the audit dataset. Other models show considerably low performance on the data. To check the robustness of J48, Bayes Net and Random Forest further, the results of K-cross validation are graphically analyzed in Figure 5. It is observed that the J48 and Bayes Net classifiers are quite robust in their performance, hence can be recommended as the prediction models for the Audit Field Work Decision Support System.

Conclusion

This paper presents a case study of Comptroller and Auditor General (CAG) of India to check the applicability of machine learning methods to predict the fraudulent firms during audit planning. A complete Audit Field Work Decision Support framework is proposed to help an auditor to decide the amount of field work required for a particular firm and to skip visiting low risk firms. Fraudulent firm prediction is an important step at the preliminary stage of an audit planning as high-risk firms are targeted for the maximum audit investigation during field engagement.

After collecting the data of 777 firms from 14 different sectors, it is cleaned, transformed, and useful risk factors are examined with the help of an in-depth interview with the auditors. Different types of risks are explored and then calculated mathematically for the audit dataset using the audit risk formula. The Particle Swarm Optimization (PSO) algorithm is implemented for feature selection and feature importance. The Risk Assessment Algorithm is presented in the paper to clearly understand the complete risk assessment process. Ten state-of-the-art classifiers like SVM, NN, BN, RF, PLM, AB, DS, J48, etc. are implemented. For comprehensive assessment of all the classifiers, performance scores of 10 different evaluation criteria using subjective ranking of criteria by audit experts are considered. The results of two multi-criteria methods, TOPSIS and SAW, indicated that Bayes Net and J48 perform the best for this particular audit dataset. BayesNet and J48 also give stable results on K-fold validation testing, serving as a proof of eligibility of classifiers to perform an efficient risk assessment of the suspicious firms in the audit field work decision-making process.

For future works, we are targeting to improve the performance of the classifiers by the ensemble machine learning approach (using a hybrid of the best performing classifiers). In the next step, we offer the auditors to handle the last 10 years data of firms on the top of advance big data techniques like Hadoop, Spark, etc.

Acknowledgments

The authors wish to thank the auditors for their assistance, time, and continued support. The authors are also grateful for their helpful feedback and comments on the earlier version of this work.

Funding

This work was supported by the Ministry of Electronics and Information technology (MEITY), Govt of India (http://meity.gov.in/) with Grant No. 1633.

References

Ali, S., and K. A. Smith. 2006. On learning algorithm selection for classification. *Applied Soft Computing* 6 (2):119–38. doi:10.1016/j.asoc.2004.12.002.

Bose, I. et al. 2011. Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems* 50 (2):491–500. doi:10.1016/j.dss.2010.11.006.

Bradley, A. P. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30 (7):1145–59. doi:10.1016/S0031-3203(96)00142-2.

Buntine, W. 2016. Learning classification rules using bayes. *Proceedings of the sixth international workshop on Machine learning*, Sydney, Australia, ACM,94–98.

Ramos, M. J. 2006. Wiley Practitioner's Guide to GAAS 2006: Covering All SASs, SSAEs, SSARSs, and Interpretations. In *Understanding the entity and its environment and assessing the risks of material misstatement* 52–53. John Wiley & Sons.

Chambers, J. M. 1977. Computational methods for data analysis. Technical report, New York.

Cosserat, G. 2009. Accepting the engagement and planning the audit. In *Modern auditing*, ed. G. Cosserat and N. Rodda, 3rd ed., 734–36. John Wiley & Sons.

Couceiro, M. 2016. Particle swarm optimization. In *Fractional order darwinian particle swarm optimization: Applications and evaluation of an evolutionary algorithm*, 1–10. Boston, MA: Springer.

Fanning, K. M., and K. O. Cogger. 1998. Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management* 7 (1):21–41. doi:10.1002/(SICI)1099-1174(199803)7:1<21::AID-ISAF138>3.0.CO;2-K.

Fawcett, T. 2006. An introduction to roc analysis. *Pattern Recognition Letters* 27 (8):861–74. doi:10.1016/j.patrec.2005.10.010.

Finney, D. J. 1992. Miscellaneous problems. *Probit Analysi* 4:140–45. JSTOR.

Freund, Y., R. Schapire,and N. Abe, 1999. A short introduction to boosting. *Journal of Japanese Society For Artificial Intelligence* 14:771–80.

Green, B. P., and J. H. Choi. 1997. Assessing the risk of management fraud through neural network technology. *Auditing* 16 (1):14–16.

Houston, R. W., M. F. Peters, and J. H. Pratt. 1999. The audit risk model, business risk and audit-planning decisions. *The Accounting Review* 74 (3):281–98. doi:10.2308/accr.1999.74.3.281.

Iba, W., and P. Langley 1992. Induction of one-level decision trees. In Proceedings of the ninth international conference on machine learning, Moffett Field, California. 233–40.

Keerthi, S. S., and E. G. Gilbert. 2002. Convergence of a generalized smo algorithm for svm classifier design. *Machine Learning* 46 (1–3):351–60. doi:10.1023/A:1012431217818.

Kennedy, J. 2011. Particle Swarm Optimization. In Encyclopedia of Machine Learning, Boston, MA: Springer.

Kothari, V. 2012. *A survey on particle swarm optimization in feature selection*, 192–201. Berlin, Heidelberg: Springer.

Kotsiantis, S. 2006. Forecasting fraudulent financial statements using data mining. *International Journal of Computational Intelligence* 3 (2):104–10.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomforest. *R News* 2 (3):18–22.

Maria, L., C. Murphy, and K. Tysiac. 2015. *Data analytics helps auditors gain deep insight*, 52–54, New York: Journal of Accountancy.

Neapolitan, R. E., et al. 2004. Introduction to Bayesian Networks. In *Learning Bayesian Networks*, 40. Chicago: Pearson.

Nikolovski, P., I. Zdravkoski, G. Menkinoski, S. Dicevska, and V. Karadjova. 2016. The concept of audit risk. *International Journal of Sciences Basic and Applied Research (IJSBAR)* 27 (3):22–31.

Pearl, J. 1985. Bayesian networks: A model CF self-activated memory for evidential reasoning. In Proceedings of the 7th Conference of the Cognitive Science Society, Irvine, California, Pp. 329–334.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning* 1 (1):81–106. doi:10.1007/BF00116251.

Quinlan, J. R. 1996. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research* 4:77–90.

Rish, I. 2001. An empirical study of the naive bayes classifier. In IJCAI workshop on empirical methods in artificial intelligence 3 (22):41–46. Seattle, Washington: IBM.

Russell, S. J., P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards. 2003. *Artificial intelligence: A modern approach*, Vol. 2, 65–71. Malaysia: Pearson Education Limited

Sharma, A. 2013. A review of financial accounting fraud detection based on data mining techniques. *International Journal of Computer Applications*.

Smith-Miles, K. A. 2009. Cross-disciplinary perspectives on meta-learning for algorithm selection. *ACM Computing Surveys (CSUR)* 41 (1):16–19.

Spathis, C. T. 2002. Detecting false financial statements using published data: Some evidence from greece. *Managerial Auditing Journal* 17 (4):179–91. doi:10.1108/02686900210424321.

Srivastava, R. P., and G. R. Shafer. 1992. Belief-function formulas for audit risk. *Accounting Review* 67 (2):249–83.

Staff, A. 2014. Reimagining auditing in a wired world1. Technical report, University of Zurich, Department of Informatics. Zurich: Citeseer.

Triantaphyllou, E. 2013. *Multi-criteria decision making methods: A comparative study*, Vol. 44. Boston, MA: Springer.

Tschakert, N. 2016. The next frontier in data analytics. *Journal of Accountancy*. Accessed September 12, 2016. http://www.journalofaccountancy.com/issues/2016/aug/data-analytics-skills.html.

Tysiac, K. 2015. Data analytics helps auditors gain deep insight. *Journal of Accountancy*. Accessed September 12, 2016. http://www.journalofaccountancy.com/issues/2015/apr/data-analytics-for-auditors.html.