Machine Learning Framework for Audit Fraud Data Prediction

Article · December 2018

CITATIONS
0

READS
747

2 authors, including:

Nishtha Hooda
Thapar University
12 PUBLICATIONS **14** CITATIONS

SEE PROFILE

Machine Learning Framework for Audit Fraud Data Prediction

Arpit Tiwari
Computer Science Engineering Department
Chandigarh University
Punjab, India
arpittiwari9898@gmail.com

Nishtha Hooda
Computer Science and Engineering Department
Chandigarh University
Chandigarh
27nishtha@gmail.com

***Abstract*—This research is about visiting an audit company to explore the practicality of machine learning for an audit data work. Data of 777 different firms are collected from six distinct sectors. In recent years, machine learning has developed and received major attention in the predictive analytics in audit research. The main objective is to produce an efficient and effective prediction model, that will be hybrid of various machine learning algorithmic characteristics, which will be capable of predicting whether any fraud has been committed in any firm or not. The experiments are carried out on high dimensional audit data and achieved an accuracy of 94.52% and AUC of 0.96. Ten different classification models like SVM, Random Forest, J48, Bayes net, etc. are compared in terms of their accuracy measure, FP rate, TP rate, F-Measure, error rate, Mathew's Correlation Coefficient (MCC), and Area Under the Curve (AUC). Machine Learning is bound to be an absolute necessity in the future because of emergence of unprecedented growth of toxic financial fraud**.

Keywords—machine learning; prediction; audit; fraud analysis ; data analytics.

I. INTRODUCTION

Fraud is intentional or purposely false commited crimes for the benefit of an individual or group of person. Individuals are able to commit fraud because they see an opportunity or due to pressure or greed or due to rationalization. Auditing practices are conducted or accountable for detection of an committed fraud. It is the process of verification of on-site financial detail of any firms to check whether their financial details are verified with different principle and standard accounting laws [1]. It is challenging task to spot firms in detecting frauds, spotting errors and exposing employee of being guilty of aiding an illegal transaction.

Audits performed by external groups over private companies can be exceedingly supportive in eliminating unfairness when it narrows down to company's finances. Audits actually seek for what is called "material error" in any statement on any distinct object.

Machine Learning is an active area of research and researchers are working on solving fraud prediction issues using machine learning techniques [8]. So, designing a machine learning problems using different machine learning classifiers helps in building a prediction model that would predict whether the firm has done any fraud or not.

In this research work, after pre-processing of data, various machine learning algorithms are trained to check the performance of classification. The performance and quality measure of the best classifier is being compared with the state-of-the-art classifiers models like support vector machine, random forest , Adabag etc. Promising results are achieved, when the performance of the proposed framework is compared with the standard classifiers like SVM, random forest, adabag etc. using various performance metrics like accuracy, FP rate, MCC, area under the curve, sensitivity, etc.

This paper is organized into various section as follows: Section 2 briefly describes the classification technique that are used in the proposed framework. Section 3 discusses the data, its features and experimental setup. Section 4 summarizes the result of the experiments performed and performance comparison. And at last in Section 5 conclusion and future scope has been discussed.

II. MATERIAL AND METHODS

The main objective of research in machine learning is to develop the programmed computing machine so that based on available data classification of pre-diction of objects can be constructed. The outcome of proposed framework helps to predict whether any fraud had been committed in any firm or not.

A. Proposed framework

The complete overflow of audit is presented in Figure 1. The outcome of proposed framework helps to predict the fraudulent of risk audit firm of any nominated firm and to understand and analyze the audit risk analysis and the work-flow of the company by in depth interview with the audit expertise, and to proffer a decision making foundation for risk assessment of firms in the time of audit planning.

B. Machine Learning Classifiers

i. Random Forest: This technique is based on ensemble learning. It is used for classification regression by constructing decision trees.[2].

ii. Neural Network: Basically, this model works on the principle of biological neurons and are used train and frame different complex and complicated relationships, and constructive patterns in statistical data [3].

iii. Support Vector Machine: SVMs are the most widely used technique for classification of different type of type datasets. It searches for data in a space i.e. boundary between two classes that are present at the edge of an area and pass them as support vectors. It is a preferred technique for classification [4].

iv. Adaboost: It is one of the most successful ensemble classifier developed by Schapiro and Freund. It assigns distinct learners to finally make an algorithm that is better than the previous one [5].

v. Logistic: In logistic model, the relationship is measured between the definite dependent variable and one or more independent variables by evaluating and measuring probabilities using logistic function approach i.e. cumulative logistic distribution [5].

vi. Decision Stump: It is weak classification model in which the simple tree structure consists of one split, are considered as one-level decision tree. Due to simplicity of this model, this is often considered as one of the most low predictive performance model [6].

vii. J48: J48 classification model is a predictive machine-learning model that chooses the target value of a new sample based on distinct attribute values of the available data. [6].

viii. Naïve bayes: The Naive Bayes Classifier technique works on the principal of Bayesian theorem. This technique is specifically suited when the inputs is high. Despite its simplicity, Naive Bayes can often even better than the more sophisticated classification methods [7].

ix. Bayesian Network (BN): This model works on the principle of probabilistic and directed acyclic graph theory. In this technique a graphical model is developed which shows a number of features and their various conditional dependencies through a specified acyclic graph [7].

x. Decision tree: Decision Tree Classifier is the most convent and widely used classification technique. Numbers of questions are carefully crafted about the attributes of the record or given detail has been posed in this technique [6].

Figure 1 Audit work Flow

III. EXPERIMENTAL INVESTIGATION

This section discusses about the dataset and experimental setup.

A. Dataset

The dataset was collected from 777 different firms from 6 different sectors.

B. Experimental Setting

The "Weka" is a suite of different machine learning feature like model building, feature selection and class balancing techniques. The objective of the classifier is to determine the accuracy when it is up and running and then classifying new datasets without the knowledge of true class of the firm data. The experiments are based on the principle of 10 fold cross validation technique. Firstly, the data set has been equally distributed into 10 same size subsets. Best machine learning algorithm is chosen as base classifier to train the 9 subset folds and testing is made on the basis of last folded subset. To check the robustness of framework, this process is repeated again and again and then recapitulated. To measure the quality and performance of designed framework, various parameters like accuracy, error rate, F measure, False Positive rate, Mathew's Correlation coefficient (MCC), and area under curve (AUC) are used.

IV. RESULTS AND DISCUSSION

This section discusses parameter evaluation metrics to measure the performance of various machine learning algorithms. The results of 10 fold cross validation method are presented graphically and discussed much in detail.

A. Performance Evaluation

The performance of the proposed framework is evaluated on the basis of confusion matrix as shown in the Table 1. The various evaluation metrics calculated from the Table 1 are presented in Table 2.

Table 1 Confusion Matrix

	True Reference	
Predicted Condition	Condition Positive	Condition Negative
Fraud	True Positive X	False Positive Z
No Fraud	False Negative Q	True Negative Y

Table 2 Performance Metric Formula

Performance Metric	Formula
Sensitivity	X/(X + Z)
Specificity	Y/(Q + Y)
Accuracy	(X + Y)/(X + Z + Q + Y)
F Score	(2 * X)/(2 *X) + (Q + z)
MCC	(X * Y)−(Q*Z)/SQRT((X +Q)+ (X + Z) + (Y + Q) + (Y + Z))

B. Experimental Results

This section explains results and discusses the performance of various machine learning classifiers. The results are also presented graphically and the reason different performances of classifiers are also discussed.

Table 3. Average performance comparison of machine learning methods for the prediction of an audit risk on testing dataset.

Classifier	Accuracy	Error	TP Rate	FP Rate	F-Measure	MCC	AUC
Decision tree	93.68	6.32	0.92	0.63	0.92	0.84	0.95
Adaboost	92.65	7.35	0.90	0.09	0.90	0.81	0.93
Random forest	94.78	5.22	0.92	0.06	0.93	0.84	0.96
SVM	71.58	28.42	0.69	0.38	0.68	0.32	0.66
Logistic	92.03	7.97	0.90	0.10	0.90	0.79	0.95
Neural network	77.96	22.04	0.78	0.18	0.78	0.58	0.18
Decision stump	87.68	12.32	0.84	0.09	0.84	0.73	0.85
J48	93.61	6.39	0.92	0.06	0.92	0.84	0.93
Naive bayes	82.44	17.56	0.79	0.12	0.79	0.65	0.93
Bayes net	91.86	8.14	0.91	0.07	0.91	0.82	0.95

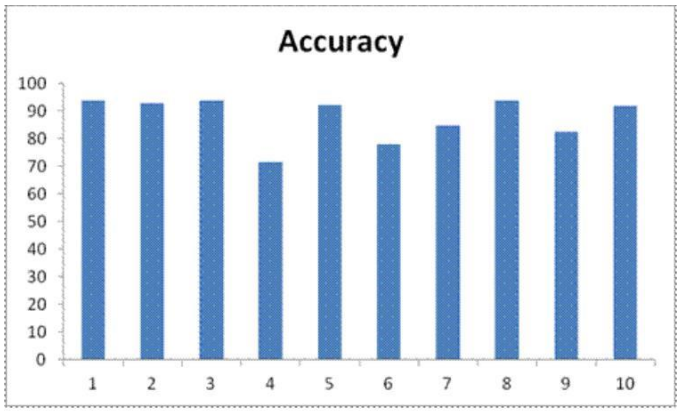

Figure 2. Comparison of accuracy of different machine learning classifiers

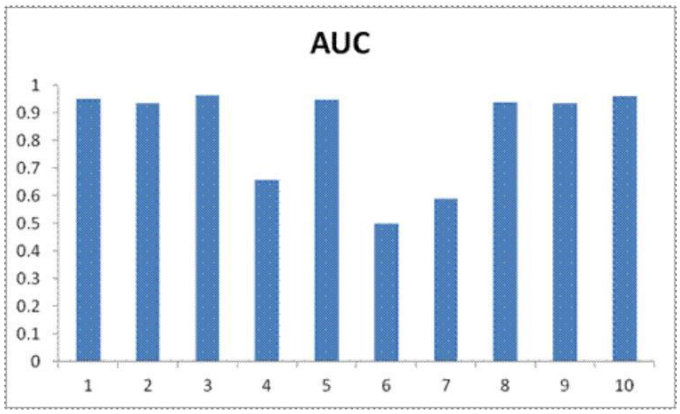

Figure 3 Comparison of AUC of different ML classifiers

It has been proved by the researchers that presently there is no such classifiers or model present that can flawlessly for all types of data issues. Similarly, there are innumerable plans to calculate the quality of classifiers and in addition to it there is no scale for all the classification problems. Keeping in mind two highly important discussions, this case study makes in us of the 10 performance metrics to differentiate 10 trending classifiers. In an endeavor to perform exhaustive performance estimation, classifiers are ranked for multiple standards.

Performance of the proposed framework has been evaluated and analyzed with 10 different performance metric parameters, and the results are summarized in Table 3. It can be observed that random forest has the highest performance for the prediction of suspicious firm. Hence, they are highlighted in the Table 3.

The results are also plotted graphically as shown in the Figure 2 and Figure 3 to compare the performance of random forest with other state-of-the-art methods. This is because the random forest being an ensemble machine learning classifier and has better performance of prediction.

V. CONCLUSION

By this paper we strive to bring forward one of the case studies of an audit company of India. The case study aims to determine the applications of Machine Learning techniques to predict and determine the fraudulent firm in the time of audit planning. The auditor will be fully equipped with a complete Audit Field Work decision Support kit to get an idea of the amount of field work required for a specific firm and to expunge visiting low risk firms. In the opening phase of an Audit development fraudulent firm prediction is vital as top risk firms are earmarked for the maximal audit investigation in the time of field meeting.

After assembling the data of different 777 firms from 6 different sectors, it is polished, transformed, and important risk factors are investigated with the help of an in-depth interview with the auditors. Various risks are identified and then the risk is evaluated in the audit dataset with the help of audit risk formula.

For future scope, we are focusing on enhancing the quality of the classifiers through various machine learning approach using best performing models.

References

[1] Hooda, Nishtha, Seema Bawa, and Prashant Singh Rana. "Fraudulent Firm Classification: A Case Study of an External Audit." *Applied Artificial Intelligence* 32.1(2018).48-64.

[2] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.

[3] Haykin, Simon, and Neural Network. "A comprehensive foundation." *Neural networks* 2.2004 (2004): 41.

[4] Hearst, Marti A., et al. "Support vector machines." *IEEE Intelligent Systems and their applications* 13.4 (1998): 18-28.

[5] Collins, Michael, Robert E. Schapire, and Yoram Singer. "Logistic regression, AdaBoost and Bregman distances." *Machine Learning* 48.1-3 (2002): 253-285.

[6] Zhao, Yongheng, and Yanxia Zhang. "Comparison of decision tree methods for finding active objects." *Advances in Space Research* 41.12 (2008): 1955-1959.

[7] Rish, Irina. "An empirical study of the naive Bayes classifier." *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. No. 22. New York: IBM, 2001.

[8] Andrieu, Christophe, et al. "An introduction to MCMC for machine learning." *Machine learning* 50.1-2 (2003): 5-43.

[9] N. Sharma, B. L. Raina, P. Rani, “Attack prevention methods for DDOS attacks in MANETs”, Asian Journal of Computer Science and Information Technology, vol. 1, no. 1, 2011.

[10]

[11] R Kaur, N Sharma -Dynamic node recovery for improved throughput in MANET … Generation Computing Technologies (NGCT), 2015 1st …, 2015

[12]

[13] R Kaur, DN Sharma Checkpointing and Trust based Recovery in MANET: A Survey- Advances in Computer Science and Information …, 2015