Audit Fraud Data Prediction using Machine Learning Algorithms

Muhammad Aroos Afzal   reg#2112115

4-1-2023

GitHub link:

https://github.com/aroos0786/CE888.git

Contents

Abstract

The purpose of this study is to visit an audit firm to investigate the applicability of machine learning to the analysis of audit data. Data on 777 different companies are gathered from six different industries. Machine learning has recently advanced and gotten much attention in predictive analytics for audit studies. The primary goal is to create a prediction model that is efficient and accurate and is a hybrid of several machine learning algorithmic characteristics. This prediction model should be able to determine whether a corporation has engaged in fraud or not. Using high dimensional audit data, these experiments produced an accuracy of 98.8% and a standard deviation of 0.08534. Support Vector Machine (SVM), Random Forest, and Logistic Regression are three classification models that are compared in terms of accurate measurement and standard deviation. Random Forest and SVM performs the best as compared to other classification model using performance metrics. Future financial fraud is expected to develop at an unprecedented rate, making machine learning imperative.

Keywords machine learning, audit dataset, fraud, fraud detection, random forest

1. Introduction

Fraud is intentionally or intentionally committing untrue crimes to benefit a particular person or group. People can commit fraud when they recognize an opportunity, when they feel pressure, when they are greedy, or when they use justification. Auditing procedures are used or responsible for finding a fraud that has been perpetrated. It examines a company's on-site financial information to see if it complies with several principles and accepted accounting practices [4]. Finding companies that can spot frauds, errors, and employees guilty of supporting an unlawful transaction is complex. When it comes to a private company's finances, audits conducted by outside organizations can be incredibly helpful in eradicating unfairness. Audits look for" material mistakes" in any statement about a specific object. Researchers are actively striving to solve the problems associated with fraud prediction using machine learning approaches [5]. As a result, creating machine learning issues using various classifiers aids in developing a prediction model to determine whether the organization has committed fraud.

The main objective of an auditor during the planning stage of an audit is to use correct analytical techniques to fairly and appropriately identify the companies that use high-risk unfair practices. Machine learning techniques are also used to build predictive analytics because it gives audit companies actionable insights. The classification of suspect firms is one of the most popular uses of predictive analytics in an audit. Studying fraud detection as a classification issue is possible. To optimize the field-testing work of high-risk enterprises that require extensive inquiry, the firms are classified during the initial stage of an audit. A study found that internal auditing has benefited from data analytics more than external auditing has in terms of advancements [1]. The following are the study's three key goals:

- To gain a thorough understanding of the company's audit risk analysis workflow through in-depth interviews with the audit staff and to suggest a framework for making decisions on risk assessment of businesses during audit planning.
- Different machine learning algorithms were used to detect the studied risk indicators and determine the Risk Audit Score for 777 target organizations. The nominated firms' Risk Audit Class (Fraud and No-Fraud) was also evaluated.
- To investigate and evaluate the viability of 3 classification models for risk class prediction and to assess the performance of the models taken for fraud prediction.

Data analytics is paying much attention to machine learning since it provides cutting-edge computational and epistemological methods for producing improved outcomes. Several techniques that come from the fields of statistics and artificial intelligence are proposed by machine learning. [9] To identify management fraud in financial statements, many academics have used algorithms such as artificial neural networks, logistic regression, decision trees, and Bayesian belief networks (Fanning & Cogger, 1998; Green & Choi, 1997; Spathis, 2002). The ensemble machine learning approach successfully improves the auditing task's categorization accuracy (Kot- siantis, 2006) [6].

In this study, multiple machine learning algorithms are trained to evaluate the categorization performance after the data has under- gone pre-processing. Modern classifier models like the random forest, logistic regression, and support vector machine (SVM), among others, are being compared to the performance and quality metrics of the best classifier. The proposed framework's performance is compared to standard classifiers like SVM, random forest, etc., utilizing several performance measures like accuracy, MCC, and standard deviation, and promising results are obtained.

2. Data

Audit Dataset is intended to assist auditors by developing a classification model that can predict fraudulent firms based on current and historical risk variables. This study dataset was generated by an audit firm that performs services for Indian government-owned firms. When organizing an audit, auditors go into the operations of many government agencies, but they specifically aim to visit those with a very high chance and significance of false assertions. This dataset is generated by evaluating the risk pertinent to financial re- porting objectives. In the audit dataset, the number of instances is 777, while the number of features is 27 [3]. Information about the firms and industries have been listed, such as Irrigation, Public Health, Buildings, Roads, etc. The dataset's data types include int, float, and decimal numerical forms. Preprocessing steps include importing all the required libraries. The median of the column has replaced missing values in the money value column in the dataset. The location ID column in the dataset has an object data type and numerical values, but three non-numeric values were removed, such as LOHARU, NUH, and SAFIDON. Duplicate rows in the dataset have been dropped.

PROB and Prob both columns have been dropped from the dataset as they had duplicate values. Unique values in each column have been determined as well. Dataset has been split into training, testing, and validation set. For training, the dataset has been split into an 80-20 ratio as 80% of the data has been used in training the model. In contrast, the remaining 20% of the data is split into 50% each for testing and validation purposes. The sections of this report are divided as follows: The dataset details are briefly described in Section 2. The approach underlying the categorization techniques utilized in the suggested framework is discussed in Section 3. The data, attributes, and experimental setting are covered in Section 4. The findings of the trials and the performance comparison are summarized in Section 4. Discussion is covered in Section 5, and finally, the conclusion and recommendations are covered in Section 6.

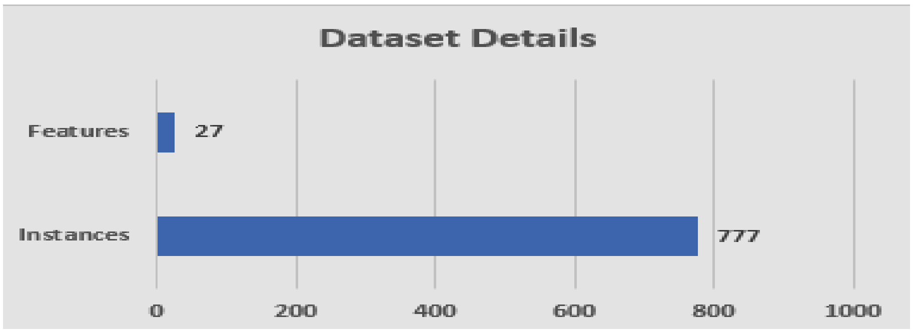

Figure 1. Dataset Details

3. Methodology

This study aims to develop and implement a prediction model for the suggested audit fieldwork decision support framework. Figure 2 presents the suggested structure, which can also function as a decision-making system. As candidates for the model's input vector, the chosen features (as listed in Table 2) are employed. A machine learning model that aids an auditor in predicting an audit risk class will be made accessible as the proposed framework's output (Fraud or No Fraud). The following are the steps performed on the dataset mentioned in the previous section.

3.1 Importing Libraries

In the first step, the relevant libraries are imported, such as pandas, NumPy, CSV, standard scaler, logistic regression, random forest, svc stratified cold, etc.

3.2 Data Cleaning and Pre-processing

The learning algorithm is rarely as strong and ideal in practical situations. Data need help with noise, replacing missing values median of the column, duplicate values, mistakes, number of unique values in each column, consistency issues, class imbalance, etc. Different kinds of risk variables are investigated once the unstructured data from various files have been cleaned and prepared. The information is arranged in 777 rows with 18 significant risk variables (columns).

3.3 Feature Scaling and Splitting Dataset

Independent and dependent variables (features) are separated from the dataset, and a feature scaling minmax scaler is applied to the data. The dataset is split into an 80-20 ratio, with 80% of the data used in training the model while 20% data is used later in testing. The remaining 20% is further divided into a 50-50 ratio as 10% of each testing data is used for validation and testing purposes.

3.4 Classifications Models

It is possible to determine whether a company is fraudulent or not as a binary classification issue utilizing an input vector (risk factors). This section discusses ten cutting-edge classification techniques used in the case study.

- **Logistic Regression** When creating machine learning models, logistic regression is a statistical technique used when the de- pendent variable is dichotomous or binary. Data and the relationship between one dependent variable and one or more independent variables are described using logistic regression [8].

- **Random Forest** an ensemble learning approach increases the classification rate by creating a forest of decision trees from random inputs (Liaw & Wiener, 2002). The accuracy and problem-solving capacity of a Random Forest Algorithm increase with the number of trees in the algorithm. To increase the dataset's predictive accuracy, a Random Forest classifier uses many decision trees on different subsets of the input data [2].
- **Support Vector Machine (SVM)** SVMs search for data points at the boundary between two classes in a space and refer to those as support vectors. It is a recommended classification method (Keerthi & Gilbert, 2002). The sorted data are output as a map by an SVM, with the margins between the two being as far away as possible. SVMs are employed in picture classification, handwriting recognition, and text categorization [7].

Stratified K Fold is used in Random-Forest with different such as several estimators, criterion, max depth, etc. Grid-Search has been applied to all the 3 machine learning models in which a base model and parameter grid are created and then, at last, fit the grid search on the data.

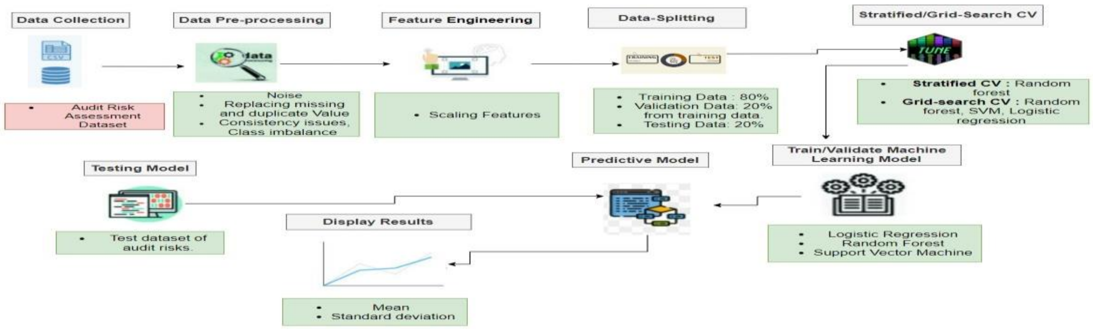
Data Collection
Audit Risk Assessment Dataset
Data Pre-processing
Noise
Replacing missing and duplicate Value
Consistency issues, Class imbalance
Feature Engineering
Scaling Features
Data-Splitting
Training Data : 80%
Validation Data: 20% from training data.
Testing Data: 20%
Stratified/Grid-Search CV
TUNE
Stratified CV : Random forest
Grid-search CV : Random forest, SVM, Logistic regression
Train/Validate Machine Learning Model
Logistic Regression
Random Forest
Support Vector Machine
Predictive Model
Testing Model
Test dataset of audit risks.
Display Results
Mean
Standard deviation

Figure 2. Proposed Methodology

Figure 2 explains the methodology in which the audit dataset is downloaded and added to the code directory. After the dataset is imported, the data is not clean. Prep-processing steps are applied to it in which noise, duplication of the data, and unique values in each column. After the pre-processing step, feature scaling is applied to the relevant features using a min-max scalar. Data is split into training, testing, and validation. For training purpose, it is 80%, while for testing, its 20%. The training 80% data is further split into

60%-20% providing 20% data to the validation set. Stratified K-Fold is applied on the random forest model while grid search cross-validation is applied on all the 3 classification models. All the machine learning classification models are further trained on the training data, while testing data is used to validate each model and display the result.

4. Results

Three machine learning models are implemented to predict an audit risk category (fraud or no-fraud). The K fold cross-validation approach (K = 10) is used to verify the robustness of the framework's design, and the models' performances are compared using 3 different performance indicators. The outcomes from the two separate measurements are listed in the hyperparameter section below. Two separate decision-making measures are used to examine these results and determine all the classifiers' overall performance scores.

4.1 Hyper Parameters

Two hyperparameters are used to evaluate the performance of the 3 different machine learning algorithms which are accuracy and standard deviation. Accuracy is obtained by dividing by the sum of true positives and negatives and the sum of false positives and false negatives as shown in Figure 3. Accuracy shows how the machine learning model correctly predicts the test data based on the trained data.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Figure 3. Accuracy Formula

The term" standard deviation" refers to a numerical measure of how dispersed the values in a dataset are present. When the standard deviation is low, most of the values are near the mean, and the numbers are dispersed over a wider range when it is large. Formula for standard deviation is mentioned below in Figure 4.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N}}$$

Figure 4. Standard Deviation Formula

Table 1 shows the classifier random forest results using 10 crosses Stratified K-Fold with a mean value of 97% and standard deviation as 0.00263.

Classifier	Mean (%)	Standard Deviation
Random Forest	97	0.00263

Table 1. Random Forest using Stratified K-Fold Cross Validation

Table 2 shows the classifier support vector machine (SVM) using a different kind of parameters such as C, gamma and kernel. The results are explained and compared in the form of mean (accuracy) and standard deviation. Maximum accuracy is achieved when C=1000, gamma: 0.03, and kernel is RBF. It is 98.8% and while stan- dard deviation is 0.08534. Minimum mean is when the value of C is 50, gamma is 0.01 and kernel is sigmoid while standard deviation is 0.0375.

Table 3 shows the classifier logistic regression using different kind of parameters such as solver, penalty and max iteration.

Classifier	Mean (%)	Standard Deviation	Tune Parameters
Support Vector Machine (SVM)	98.1	0.01848	C: 25, gamma: 0.1 kernel: ’linear’
	94	0.17531	C: 50, gamma: 0.01 kernel: ’sigmoid’
	98.8	0.08534	C: 1000, gamma: 0.03 kernel: ’rbf’

Table 2. Support Vector Machine (SVM) using 10-fold Grid- Search Cross Validation

Classifier	Mean (%)	Standard Deviation	Tune Parameters
Logistic Regression	97.6	0.0277	solver = ’newton-cholesky’, penality = ’elasticnet’ max iter = 300
	96	0.05372	solver = ’liblinear’, penality = ’l1’ max iter = 500
	84.6	0.0375	solver = ’lbfgs’, penality = ’l2’ max iter = 1000

Table 3. Logistic Regression using 10-fold Grid-Search Cross Validation

results are explained and compared in the form of mean (accuracy) and standard deviation. Maximum accuracy is achieved when solver is newton cholesky, penalty is elasticnet and max iteration is 300. It is 97.6% and while standard deviation is 0.0277. Minimum mean is when value of solver is liblinear, penalty is l1 and max iteration is 500 while standard deviation is 0.05372.

Classifier	Mean (%)	Standard Deviation	Tune Parameters
Random Forest	98.2	0.04312	n split = 3, n estimators = 25, max depth = 3
	97	0.0026	n split = 5, n estimators = 50, max depth = 5
	97.3	0.03465	n split = 10 n estimators = 100, max depth = 10

Table 4. Random Forest using 10-fold Grid-Search Cross Validation

Table 4 shows the classifier random forest using different kind of parameters such as number of splits, number of estimators and max depth. The results are explained and compared in the form of mean (accuracy) and standard deviation. Maximum accuracy is achieved when number of splits are 3, number of estimators are 25 and max depth is 3. It is 98.2% and while standard deviation is 0.04312. Minimum mean is when number of splits are 5, number of estimators are 50 and max depth is 5 which is 97%. while standard deviation is 0.03465.

5. Discussion

Researchers have established that no classifiers or models can perfectly address all kinds of data difficulties. Similar to countless plans for calculating classifier quality, there is no scale for all classification issues. This case study uses the two performance indicators to distinguish between three trending classifiers while bearing in mind two significant conversations. Classifiers are ranked for various standards to carry out extensive performance estimates. Two different performance metric factors were used to evaluate and analyze the effectiveness of the proposed framework.

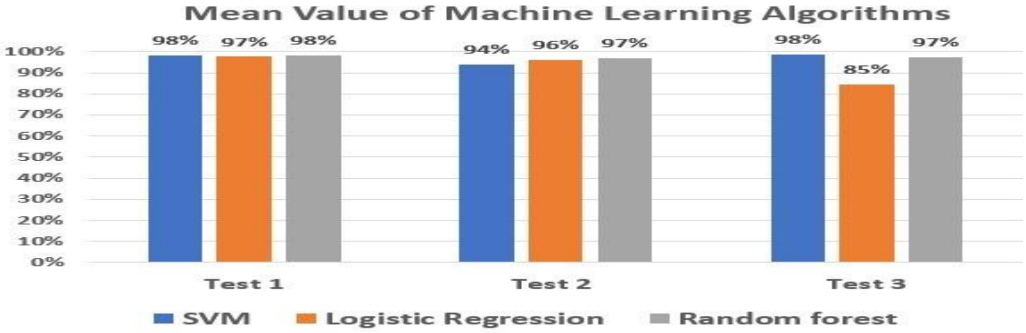

Figure 5. Mean Values of the Machine Learning Algorithms

Figure 5 shows the mean/accuracy of the different machine learning algorithms for various tests, which are hyperparameter tuning as explained in the tables of the results section. There are three machine learning algorithms, each represented by the colors blue, orange, and gray. The blue represents the support vector machine (SVM), the orange represents logistic regression, and the gray represents the random forest. Different tests are performed on these machine learning algorithms to test the accuracy and how correctly and efficiently it predicts the test data. It can be seen that the support vector machine has the same mean value as the random forest, but logistic regression lacks behind with just a 1% difference in test 1. In Test 2, Random Forest outperforms the other two classification algorithms by achieving a mean value of 97%. In the last Test 3, SVM achieves the highest mean value of 98%, while logistic regression lacks behind with 81%. It is seen that Random Forest performs the best in most of the test cases.

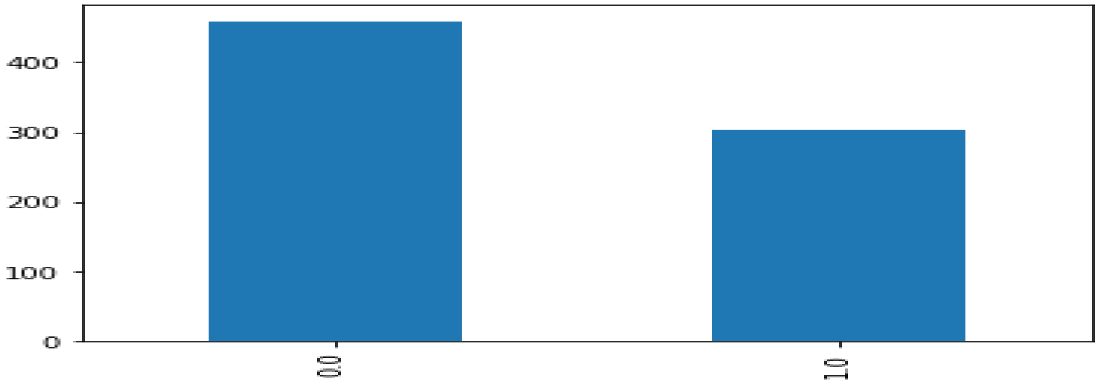

Figure 6. Fraud/Not Fraud Firms in total of 777

Figure 6 shows the number of fraud and not fraud firms. A total number of firms are 777 while more than 400 are not fraud while less than 200 are involved in different fraud activities.

Figure 7 shows the standard deviation value for different machine learning algorithms. Three different hyperparameter-tuned tests/configurations for the machine learning algorithms are used to predict the model standard deviation on the test data. Test 1 is represented in blue, Test 2 is in orange, and Test 3 is in Gray. Test 1 is the standard deviation hyperparameter tuning values in Table 4, which shows that for several splits 5, max depth five, and estimators 50, random forest performs the best on the test data.

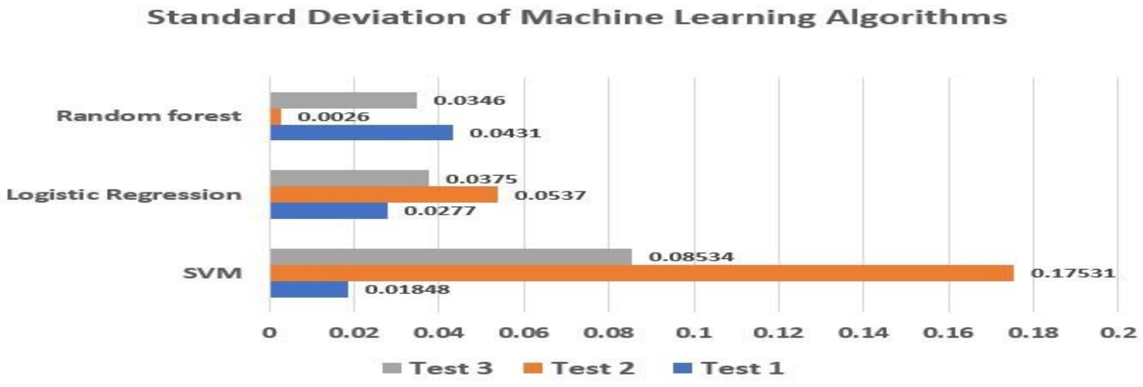

Figure 7. Standard Deviation of the Machine Learning Algorithms

Test 1 performs the best for the logistic regression as seen in the table 3. The relevant hyperparameters are solver newton Cholesky, max iteration 300, and elastic penalty net. In the last SVM, Test 1 again performed the best, having a standard deviation value of 0.01848 with c 25, gamma 0.1, and the kernel is linear.

SVM is well-known for its exceptional prediction accuracy, but RF stands out among other traditional machine learning techniques for its unique ability to combine prediction accuracy with model explicability.

6. Conclusion/Recommendations

Predicting fraudulent firms is a crucial step in an audit's initial planning stages since high-risk firms receive the most audit investigation during field engagement. This research aimed to present one of the case studies of an Indian audit organization. The case study aims to identify how machine learning techniques can be used to forecast and identify a fraudulent firm during audit preparation. The auditor will be fully prepared with a comprehensive Audit Field Work Decision Support package to eliminate visiting low-risk enterprises and estimate the fieldwork necessary for a specific firm. Fraudulent firm prediction is essential during the initial stages of audit development because top-risk firms are chosen for the most thorough audit inquiry at the field conference. The information is polished, changed, and significant risk variables are examined with an in-depth interview with the auditors after collecting it from 777 organizations in 6 distinct sectors. The dataset includes 777 rows and 18 features after gathering the information. The audit risk formula mathematically calculates various hazards for the audit dataset. Different risks are discovered, and after that, using the audit risk formula, the risk is evaluated in the audit dataset.

This work can be further enhanced by improving the quality of the classifiers through various machine-learning approaches employing best-performing models for the foreseeable future. In future projects, this work can be improved by an ensemble machine learning strategy (a hybrid of the top-performing classifiers) to in- crease the performance of the classifiers. The auditors can handle the companies' most recent ten years' worth of data using cutting-edge big data tools like Hadoop, Spark, etc., in the following stage.

References

[1] Y. Bao, G. Hilary, and B. Ke. Artificial intelligence and fraud detection. In *Innovative technology at the interface of finance and operations*, pages 223–247. Springer, 2022.

[2] H. Chen, L. Wu, J. Chen, W. Lu, and J. Ding. A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management*, 59(2):102798, 2022.

[3] N. Hooda, S. Bawa, and P. S. Rana. Optimizing fraudulent firm predic- tion using ensemble machine learning: a case study of an external audit. Applied Artificial Intelligence, 34(1):20–30, 2020.

[4] A. T. Khan, X. Cao, S. Li, V. N. Katsikis, I. Brajevic, and P. S. Sta- nimirovic. Fraud detection in publicly traded us firms using beetle antennae search: A machine learning approach. Expert Systems with Applications, 191:116148, 2022.

[5] S. Papadakis, A. Garefalakis, C. Lemonakis, C. Chimonaki, and C. Zo- pounidis. Machine Learning Applications for Accounting Disclosure and Fraud Detection. IGI Global, 2020.

[6] J. Perols. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing: A Journal of Practice & Theory, 30(2):19–50, 2011.

[7] D. A. Pisner and D. M. Schnyer. Support vector machine. In Machine learning, pages 101–121. Elsevier, 2020.

[8] K. Shah, H. Patel, D. Sanghvi, and M. Shah. A comparative analysis of logistic regression, random forest and knn models for the text classifi- cation. Augmented Human Research, 5(1):1–16, 2020.

[9] C. Shang and F. You. Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era. Engineering, 5(6):1010–1016, 2019.