## Audit Data Mining –Fraud Detection

Project Report

Submitted to Ms. Shalini Kumari

Submitted By:

Anisha

Batch no. 4866

Noida

**Table of Contents**

[Audit Data Mining –Fraud Detection 1](#_Toc38471897)

[1. Abstract 3](#_Toc38471898)

[2. Introduction 3](#_Toc38471899)

[3. Data Preparation, Exploration & Transformation 4](#_Toc38471900)

[4. Models 5](#_Toc38471901)

[4.1. Linear Classifiers 6](#_Toc38471902)

[4.1.1 Linear regression 6](#_Toc38471903)

[4.1.2 Logistic regression 6](#_Toc38471904)

[4.1.3 SGD Classifier 6](#_Toc38471905)

[4.2. Geometric Models 6](#_Toc38471906)

[4.2.1. Support vector machines (SVM) 6](#_Toc38471907)

[4.2.2. KNN using Manhattan distance 7](#_Toc38471908)

[4.2.3. KNN using Euclidean distance 7](#_Toc38471909)

[4.3. Probabilistic Model 7](#_Toc38471910)

[4.3.1. Gaussian Naive Bayes 7](#_Toc38471911)

[4.4. Tree Based 7](#_Toc38471912)

[4.5. Ensemble 7](#_Toc38471913)

[4.5.1. Voting 7](#_Toc38471914)

[4.5.2. Bagging 7](#_Toc38471915)

[4.5.3. Random Forest 8](#_Toc38471916)

[4.6. Neural Network 8](#_Toc38471917)

[4.6.1. Multilayer Perceptron (MLP) 8](#_Toc38471918)

[5. Evaluation 8](#_Toc38471919)

[6. Conclusion 10](#_Toc38471920)

# Abstract

In the auditing sector, the need for data mining is rising rapidly. As electronic systems and technology tools complicate and exploit accounting transactions, the use of data mining in the auditing profession has increased in recent years. Because auditing involves evaluating massive data in the attestation feature, data mining enables this process to be carried out easily. Data mining technology is very prevalent in the credit card fraud detection method. Data mining is popularly used because of its usefulness in combating fraud. It is a well-defined procedure that takes data as input and produces models or patterns as output.

In our project, we performed Data Preprocessing on the data set of audit data and removed null values. We did a detailed data analysis of the Risk factor (The Class label) with all other attributes. It was followed by Feature selection and finally applying various classification Models and Ensemble techniques on our data set. We performed cross-validation to determine accuracy and visualized the accuracy of various models using ROC curves and confusion matrix. After our rigorous analysis and Classification, we concluded that the random forest Feature selection reduced the training time and complexity The Gaussian Naive Bayes and support vector machine gives the highest accuracy and thus these models are best to be used for classification and further testing.

# Introduction

Audit data analytics involves the analysis of complete sets of data to identify anomalies and trends for further investigation, as well as to provide audit evidence. The goal of the research was to help the auditors by building a classification model that can predict the fraudulent firm based on present and historical risk factors. The information about the sectors of firms are listed as Irrigation, Public Health, Buildings and Roads, Forest, Corporate, Animal Husbandry, Communication, Electrical, Land, Science and Technology, Tourism, Fisheries, Industries and Agriculture. Source of the dataset was UCI Machine Learning Repository.

Exhaustive one-year non-confidential data in the year 2015 to 2016 of firms was collected from the Auditor Office of India to build a predictor for classifying suspicious firms [1]. It was found to have the following information associated with it:

Dataset Characteristics: Multivariate

Attribute Characteristics: Real

Number of instances = 777  
Number of features = 18 (including output label)  
Output label - ‘Risk’. It has only 2 possible values 0 and1. So, it is a Binary Classification Problem.

Many risk factors were examined from various areas like past records of audit office, audit-paras, environmental conditions reports, firm reputation summary, on-going issues report, profit-value records, loss-value records, follow-up reports etc. After in-depth interview with the auditors, important risk factors were evaluated and their probability of existence was calculated from the present and past records.

This report goes into detail of how the analysis of the dataset was performed. Section 3 discusses firstly how data analysis was performed to understand the nature of the data. Followed by which data pre-processing was done to make the data usable. Afterwards, feature selection was performed to do the classification based on the selected classifier. Different types of classifiers were used along with that 10-fold cross validation [2] to test the accuracy of the classifier. Section 4 describes about different models or classifiers that were used to classify the dataset into a fraud or normal result. Section 5 discusses all the observations obtained from the classifiers. Conclusion in section 6 contains the results of accuracies corresponding to each classifier followed by future direction in section5.

# Data Preparation, Exploration & Transformation

In this project, firstly data analysis was performed to understand the nature of the data. Followed by which data pre-processing was done to make the data usable. Afterwards, feature selection was performed to do the classification based on the selected classifier. Different types of classifiers were used along with that 10-fold cross validation to test the accuracy of the classifier.

It was found that Risk was equal to 1 which was approximately 2/3rd of the dataset and No Risk that is, Risk equal to 0 was 1/3rd of the dataset. Analyses ofeach feature was done against the Risk and observed how value of Risk (Output) changes with each feature.Correlation map was used to see how much each feature is correlated with another feature.

Data processing refers to the step in which the data gets transformed, or *Encoded*, to bring it to such a state that now the machine can easily parsed it. In other words, the *features*of the data can now be easily interpreted by the algorithm.

There were some missing values in the dataset, so missing values were replaced with mean value and object data type was changed to type ‘Float’, so that mathematical operations can be performed easily.

There were total 18 features in the dataset from which 10 features were selected as per the requirements. Original data was lacking unique values in most of the features. So, features were selected based on theforest method and computed the feature ranking to do the feature selection.These were the features that were finally used in further calculations: 'SCORE\_A', 'History\_Score‘, 'SCORE\_B', 'District', 'MONEY Marks', 'PARA\_A', 'Sector\_score', 'TOTAL', 'SCORE\_A', 'Money\_Value'.

# Models

In this project, different classifiers were used to classify the audit dataset into fraud had happened or not based on the classes it belongs. Results obtained from the classifiers which were used are discussed in detail in the following subsections.

For Training and Testing purpose, we split the data into 80% and 20% respectively.Also applied 10 fold cross validation to further improve the results.

Cross-validation is a statistical method used to estimate the skill of machine learning models.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

There were the model metrics that were used to evaluate the classifier:

**Accuracy Score**: Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

**Precision Score**: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

**Recall**: It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

**F1 Score**: Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. Signifies how precise classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

**Confusion Matrix**: A summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class

Gives insight of not only into the errors being made by a classifier but more importantly the types of errors that are being made.

**ROC Curve:** A receiver operating characteristic curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate and False Positive Rate.

Area under the Curve provides an aggregate measure of performance across all possible classification thresholds.

Different classifiers that were used to classify the dataset had been described below:

## Linear Classifiers

### Linear regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable [3].

### Logisticregression

It is the appropriate regression analysis to conduct when the dependent variable is binary like all regression analyses, the logistic regression is a predictive analysis.  Used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables [4].

### SGD Classifier

Stochastic gradient descent (often abbreviated SGD) is an iterative method for optimizing an objective function with suitable smoothness properties. It is a simple and efficient approach for discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression [5].

## Geometric Models

### Support vector machines (SVM)

A supervised machine learning model that uses classification algorithms for two-group classification problems.

In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well [6].

### KNN using Manhattan distance

K-Nearest Neighbour (K-NN) is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure.To determine which of the K instances in the training dataset are most similar to a new input, a distance measure is used. Manhattan distance is calculated as the distance between real vectors using the sum of their absolute difference. Also called City Block Distance [7].

### KNN using Euclidean distance

KNN is powerful because it does not assume anything about the data, other than that the distance measure can be calculated consistently between any two instances.

The usage of Euclidean distance measure is highly recommended when the data is dense or continuous. It is the best proximity measure. The Euclidean distance between two points is the length of the path connecting them [7].

## Probabilistic Model

### Gaussian Naive Bayes

The naive Bayes(NB) classifier assumes all the features are independent to each other. Even if the features depend on each other or upon the existence of the other features.

A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It’s specifically used when the features have continuous values. It’s also assumed that all the features are following a Gaussian distribution i.e., normal distribution [8].

## Tree Based

Tree based algorithms empower predictive models with high accuracy, stability and ease of interpretation.Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression). Methods like decision trees, random forest, gradient boosting are being popularly used in all kinds of data science problems [9].

## Ensemble

### Voting

Voting classifier isn't an actual classifier but a wrapper for set of different ones that are trained and valuated in parallel in order to exploit the different peculiarities of each algorithm.

We can train data set using different algorithms and ensemble then to predict the final output. The final output on a prediction is taken by majority vote [10].

### Bagging

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.

Typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it [10].

### Random Forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an Ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

Uses averaging to improve the predictive accuracy and control over-fitting [10].

## Neural Network

### Multilayer Perceptron (MLP)

A multilayer perceptron (MLP) is a deep artificial neural network. It is composed of more than one perceptron.

They consists of input layer to receive the signal, an output layer that makes a decision or prediction about the input, and in between those two, an arbitrary number of hidden layers that are the true computational engine of the MLP [10].

# Evaluation

The table below shows the results of each classifier that comes up after training and testing along with cross validation.

|  |  |  |
| --- | --- | --- |
| Classifier/Model | Model Metrices | Cross Validation Score |
| Logistic Regression |  | 0.980417 |
| SGD Classifier |  | 0.962083 |
| Support vector Machines |  | 0.987083 |
| KNN Using Manhattan distance |  | 0.904167 |
| KNN Using Euclidean Distance |  | 0.916250 |
| Gaussian Naïve Bayes |  | 0.993333 |
| Tree based model |  | NA |
| Ensemble-Voting |  | 0.923750 |
| Ensemble-Bagging |  | 0.904583 |
| Random Forest Classifier |  | 0.922917 |
| Neural Networks |  | 0.903750 |

# Conclusion

To summarize up, Dataset underwent Data Pre-processing and cleaning which includes replacing missing values and converting to proper format. Afterwards, feature selection is done to reduce the number of features of further classification.Feature selection reduced the training time and complexity. Thereafter, features were selected then apply the various classifiers to perform the classification. Different types of classifiers are applied to the dataset example: Geometric Classifiers, Linear Classifiers, Tree based, NeuralNetwork etc. Every classifier gives the good results which could be seen through confusion matrix as well as from ROC curve. But among all of the classifiers applied, Gaussian Naive Bayes and support vector machine gives the highest accuracy which is validated using 10 – fold cross validation. So these models are best to be used for classification and for further testing.