Chennai Mathematical Institute
H1, Sipcot IT Park, Siruseri, Kelambakkam, Tamil Nadu 603103

Data-Driven Prediction of Fraudulent Firm

Suman Polley

Advisor : Dr. Sourish Das

December 23, 2020

sumanp@cmi.ac.in

Abstract.

Business organisations doing fraud have direct grievous consequences on economy of a region or country based on perspective size. **Auditing** is defined as practice of examining all the business records of a company to corroborate that their financial statements are in compliance with the standard accounting laws and principles. It is very laborious and time consuming task to audit a single company .The number of companies is enormous. So, detecting fraud while necessary might not be feasible and we have to rely on internal audit of the company(Which a fraudulent company can use in their advantage to lie about their business practices.). Thus many a frauds are undetected to the point where no preventive measures can be taken resulting in bankruptcy,dis-solvency of companies and job loss ,economic output loss.

Introduction:

A firm that has large amount of discrepancy in their planned or unplanned expenditure might be doing fraud. Past record of fraud is also a indicator of doing fraud in future. Prior to in depth analysis of a firm if raw data collected (or revealed by the firm itself)about a firm can indicate fraud then fraud detection and overall auditing would become much more efficient.

Experts in the field use many analytical tool to extract information from features like discrepancy in their planned expenditure,misstatements in the past audits. Based on their assessment they assign risk scores for each feature. And based on these scores they calculate overall risk score(Audit Risk Score). If the Audit Risk Score cross a certain threshold then the firm is considered to be fraudulent.

Problem Statement:

The goal of this paper is to use only the raw data(only the collected data and not the auditors assessments based on the raw data.e.g. The feature ”Discrepancy found in the planned-expenditure in rs” would be used and not the ”risk score” based on this feature) and past data to determine fraudulent nature of a firm prior to any in depth assessment by any auditor.

This is a classification Problem i.e either a firm is fraud(Risk=1) or not fraud (Risk=0).

Objective:

The objective of this paper is to implement various machine learning algorithms (e.g. Decision Tree, SVM, Logistic Regression etc) to build models that can predict the nature of a firm(i.e. Risk = 0 or 1) given the raw data about the firm.

Data-set

The data-set is available at https://archive.ics.uci.edu/ml/datasets/Audit+Data .

The trial.csv file has been considered for this paper.
description: Exhaustive one year non-confidential data in the year 2015 to 2016 of firms is collected from the Auditor Office of India to build a predictor for classifying suspicious firms.

Data cleaning and Features :

There are total 776 data points , each representing a unique firm. The firms are situated across 45 distinct locations. Each firm is falls into thirteen different sector.

The features considered: 'Sector_score', 'LOCATION_ID', 'PARA_A', 'PARA_B', 'TOTAL', 'numbers','Money_Value', 'District', 'Loss', 'History', 'Risk'

All the other features are risk score assessments of these above features done by auditors. To meet the purpose of using only raw data the other features have been dropped.

Feature	Description
PARA_A	Discrepancy found in the planned-expenditure of inspection and summary report A in Rs (in crore)
PARA_B	Discrepancy found in the unplanned-expenditure of inspection and summary report B in Rs (in crore)
TOTAL	Total amount of discrepancy found in other reports Rs (in crore).
Money_Value	Amount of money involved in misstatements in the past audits.
Loss	Amount of loss suffered by the firm last year.
Number	Historical discrepancy score.
Sector_score	Historical risk score value of the sector the firm falls into.
History	Average historical loss suffered by firm in the last 10 years.
District	Historical risk score of the district(the firm is situated in) in the last 10 years.
LOCATION_ID	Unique ID of the city/province the firm is situated in.
Risk	Target variable indicating if a firm is fraudulent or not.

Table 1: Features and their Descriptions

Although The features Sector_score and District are not raw data ,these are used since are historical data and assumed to be available.

The features LOCATION_ID,District and Loss are categorical variables. LOCATION_ID can have 45 distinct values for each of 45 distinct locations. District have 3 distinct values: 2 for low, 4 for medium and 6 for high.Loss also have 3 different values .

The target variable Risk is binary. All the other features are numeric.(LOCATION_ID have numeric values but are in string format in the data-set,The data type has been changed to numeric format)

A snapshot of the data:

	Sector_score	LOCATION_ID	PARA_A	PARA_B	TOTAL	numbers	Money_Value	District	Loss	History	Risk
50	3.89	22	1.97	2.10	4.07	5.0	12.29	2	2	0	1
51	3.89	22	0.00	18.05	18.05	5.0	2.29	2	0	0	1
52	3.89	22	0.00	0.93	0.93	5.0	7.78	2	0	0	1
53	3.89	9	0.00	1.61	1.61	5.0	2.51	2	0	0	0
54	3.89	9	6.32	9.01	15.33	5.0	8.31	2	0	0	1

Figure 1: A snapshot of the Data-set. Data points 50 to 54 have been shown here.

Analysis of the Data:

A simple count shows that there are 486 data points with Risk = 1 and 290 data points with Risk = 0. There is one data point with missing value (i.e. index=642, Money_value = NaN).

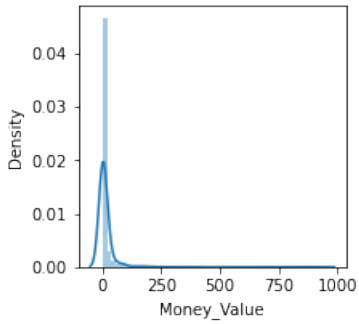

Figure 2: Density plot of Money_Value.This feature variable is extremely right skewed.

The feature Money_Value has mean = 14.137631, standard deviation = 66.606519, 25 % percentile = 0.000000, 50 % percentile = 0.090000, 75 % percentile = 5.595000. This shows that the data is right skewed. And have 328 distinct values with 0.0 value 332 times. So, the missing value can be replaced by 0.0 .(See Figure 2)

It can be seen that Money_Value has outliers and in fact quite large outliers. But for all values greater equal to 5, Risk = 1. So,the outliers can not be discarded. This makes perfect sense , since more the ”Amount of money involved in misstatements in the past audits” greater the chances of fraud .

The graph below shows the correlation between all the features:

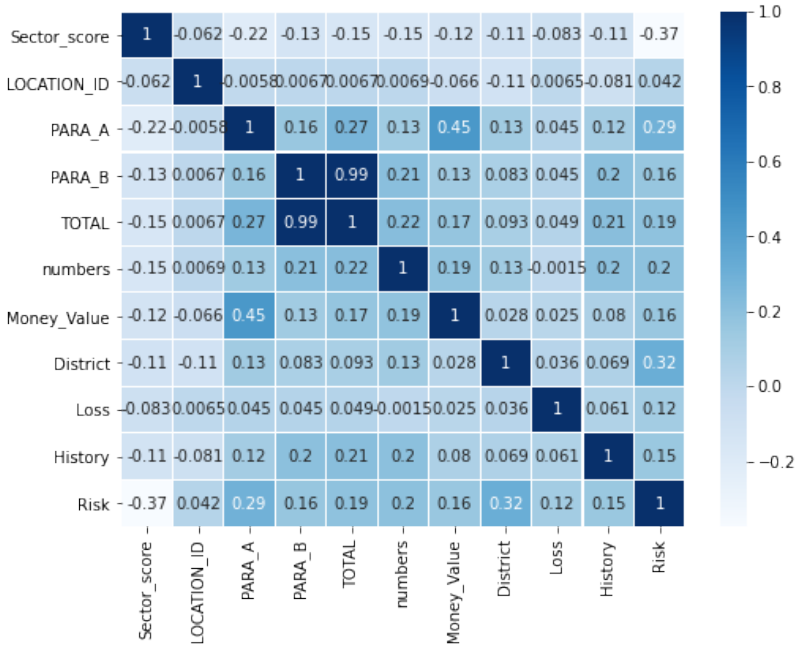

Figure 3: Heat Map of correlation coefficient between each pair of features. Darker means high positive correlation and lighter means higher negative correlation.It is observable that Sector_score has the highest correlation with Risk followed by District and PARA_A

Dark shades represents positive correlation while lighter shades represents negative correlation. It can be inferred from the figure that the features PARA_B and TOTAL have a high correlation between them .

The distribution of all the features are :

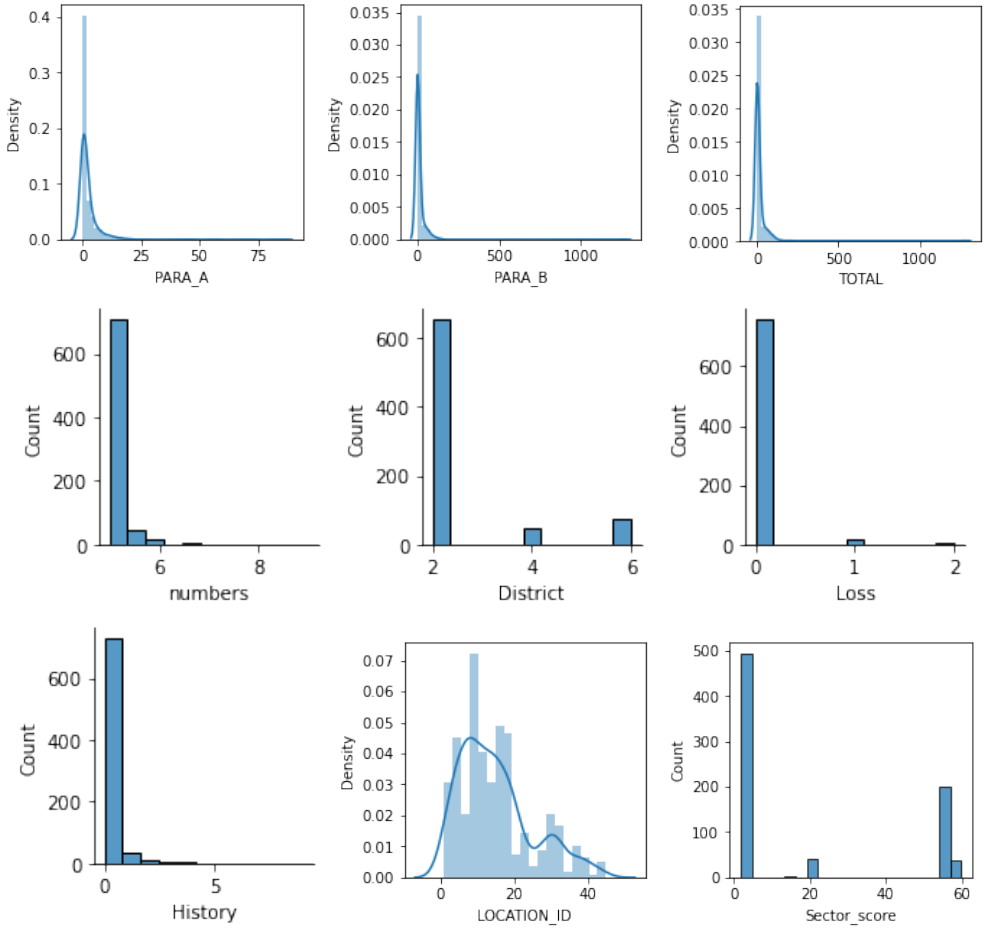

Figure 4: Distributions of features : PARA_A, PARA_B, TOTAL are continuous in nature where as numbers, District, Loss, History, LOCATION_ID and Sector_score are discrete in nature.All of the features have right skewed distribution.

It can be seen that the features PARA_A , PARA_B , TOTAL , numbers and History have positively skewed distribution and PARA_A , PARA_B , TOTAL have large outliers . However for these three features with large outliers the feature-value greater than the respective 75^{th} percentile implies Risk = 1. So, the outliers can not be discarded as the outliers are good indicators of fraud.

For the features numbers , Money_Value , District , Loss and History the above 75^{th} percentile rule woks as well. But LOCATION_ID and Sector_score do not do a good job in separating the high risk and low risk firms. This can be visualised in Figure 5

Model Selection :

There are six different types of predictive models i.e. Classification (Picking one of N labels), Regression (Numerical value prediction), Clustering (Clustering similar data points), Association rule mining, Structured output(e.g. Natural Language Processing), and Ranking.

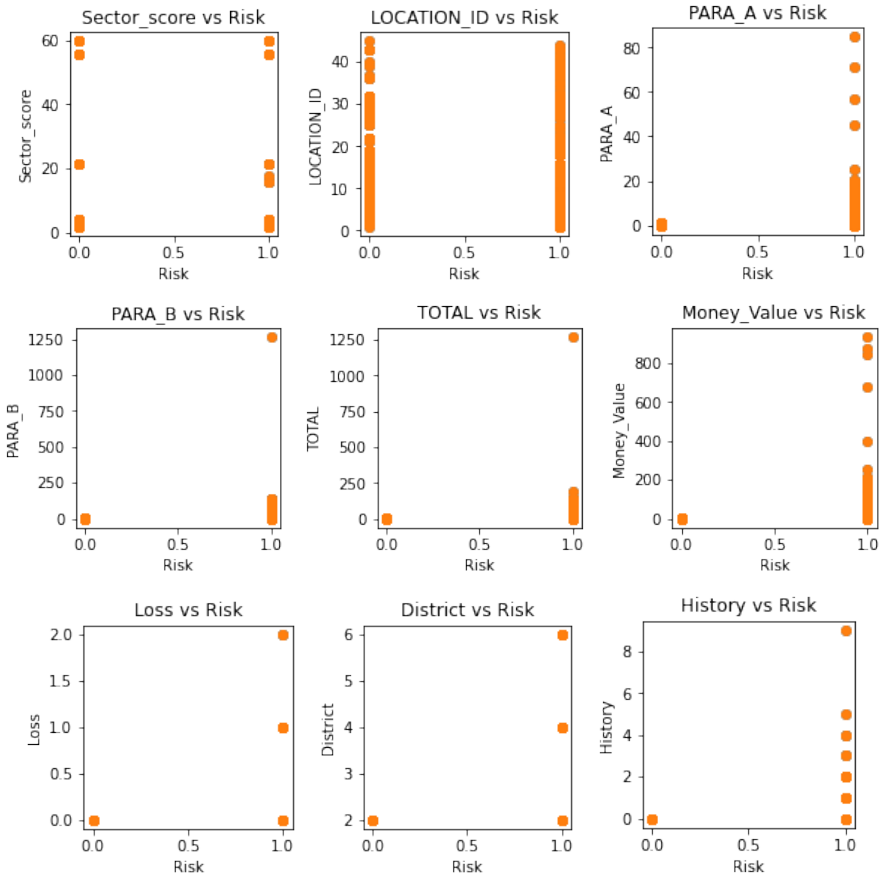

Figure 5: features vs Risk : Except for Sector_score and LOCATION_ID in all the other feature variables high value means higher risk of fraud .

The problem being discussed here is a binary classification problem(The target variable Risk takes only two values 0 and 1). The machine learning models selected for this purpose are Decision Tree, random Forest, Adaboost, Gradient Boost, Support Vector Machine, Gaussian Naive bayes, Multinomial Naive Bayes, Logistic Regression and Neural Network.

Description Of Models:

Decision Tree : A simple decision-rule may not classify all the data points properly but many decision rules together can classify with high accuracy. In Decision Tree simple decision rules acts as nodes to a tree-like structure .

Random Forest : In Decision tree the decision-rules are placed into nodes from root to leaf based on some criterion. In Random Forest decision-rules are selected in random to create multiple Decision Trees. The majority vote is considered for classification. This eliminates the over-fitting that arises from a single Decision Tree.

Adaboost : A stump is a Decision Tree with only one node. A stump is a weak learner. Adaboost or Adaptive boost creates a forest of stumps. Each stump influences the next stump in favour of the misclassified samples from the previous stump. Thus a weighted sum of output of weak learners is created to make classification more precise.

Gradient Boost : Gradient Boost creates an ensemble of weak learners . Unlike Adaboost the weights are not tweaked rather the residual error is minimized in each step.

Support Vector Machine (SVM) : In SVM all the data points are in a vector space and the hyper-plane that best divides the data points is searched. For linearly separable data points the the hyper-plane that has maximum distance from all nearest data points is searched . For linearly non-separable data points the data points are projected onto a higher-dimensional space using kernel function where the data points are linearly separable.

Naive Bayes : Naive Bayes uses the bayes theorem with the assumption of conditional independence between every pair of features. If $x_1, ...x_n$ be feature variables and y be the target variable then Naive bayes uses the Bayes' theorem :

$$P(y|x_1, ..., x_n) = \frac{P(y)P(x_1, ..., x_n|y)}{P(x_1, ..., x_n)}$$

Logistic Regression : Logistic regression ,tough seems like a regression model actually is a classification model that uses a logistic function to separate the distinct classes.The logistic function :

$$y = \frac{1}{1 + e^{-k(x-x_0)}}$$

where x_0 defines the boundary between the classes, y is the target variable , k : constant.

Neural Network : Neural Network is formed based on how our brain works . Neural Network consists of multiple layers of connected Neurons. A Neuron is similar to a single logistic regression unit that lights up(i.e. let the data pass) when data passing thorough that neuron crosses a certain threshold.

Model Fitting:

The data set has been split into two sets for training and testing purpose. The training data-set consists of eighty percent of the data points randomly chosen. Th testing data-set consists of the rest of the twenty percent of the data points.

The training data set has been further divided into ten segments for k-fold cross validation(here, k=10). K-fold cross validation ensures that a model isn't chosen that overfits the training data. The table(Table 2) below shows what parameters were chosen for each model:

Model	Python3 Library	Hyper-parameters
Decision Tree*	sklearn.tree.DecisionTreeClassifier	criterion : entropy max_depth : 8 min_samples_leaf : 26 class_weight : 2
Random Forest	sklearn.ensemble.RandomForestClassifier	n_estimators : 100
Adaboost	sklearn.ensemble.AdaBoostClassifier	Default parameters
Gradient Boost	sklearn.ensemble.GradientBoostingClassifier	Default parameters
Support Vector Machine*	sklearn.svm.SVC	kernel : linear degree : NA class_weight : 2

Gaussian Naive Bayes	sklearn.naive_bayes.GaussianNB	Default parameters
Multinomial Naive Bayes	sklearn.naive_bayes.MultinomialNB	Default parameters
Logistic Regression	sklearn.linear_model.LogisticRegression	class_weight :
Neural Network (with only one hidden layer)	tensorflow.keras.Sequential	input_layer : 20 nodes hidden_layer : 40 nodes activation_function : swish

Table 2: Models ,their python3 library and the best hyper-parameters

*These Models have too many hyper-parameters to optimize by hand and not so many that they can not be optimized efficiently. The hyper-parameters for these two models have been optimized by Bayesian Optimization.

Bayesian optimization : Bayesian Optimization is a machine-learning based optimization method that leverages on Gaussian Process to find $max(f(x))$ where f is not a convex or concave function. This makes the optimization difficult. Such functions are called $black\ box$function. f has to be continuous. While finding $max(f(x))$ using Bayesian optimization only $f(x)$ is taken into consideration, no derivative of f.

Results and Conclusion:

Except for Naive Bayes all the other models have performed extremely well. The metrics used to determine the goodness of the models are Accuracy and Recall.

True Positive(tp): No. of Positive data points classified as Positive.

False Positive(fp): No. of Negative data points classified as Positive.

True Negative(tn): No. of Negative data points classified as Negative.

False Negative(fn): No. of Positive data points classified as Negative.

Recall: $\frac{tp}{tp+fn}$ (Portion of Positive data points correctly predicted as Positive)

Precision: $\frac{tp}{tp+fp}$ (Portion of actual Positive data points in predicted Positive data points.)

F1_score(f_1): $\frac{2}{\frac{1}{Recall}+\frac{1}{Precision}}$

Since , the **Goal** is to identify every fraud company and don't want to miss even a single one, Recall is much more prominent metric than Accuracy or Precision.

In fact while using Bayesian Optimization in Decision Tree and SVM the metric that has been optimized is **Accuracy*Recall** to choose the model not only with high Accuracy but also with high Recall. Although optimizing based on only Recall seem to be a good idea. A Recall score of 1 can make the Accuracy and Precision score really low for Decision Tree and SVM. Figure 6 gives information about different models and their performances.

Cross Val Accuracy stands for average accuracy of the model in 10-fold cross validation.

Accuracy stands for accuracy of the model on test data-set.

Cross Val FRecall stands for average recall of the model in 10-fold cross validation.

Recall gives the Recall of the model on test data-set.

F1 Score stands for the F1_Score of the model on test data-set.

In terms of Accuracy and Recall on both validation and test set the best performing models Adaboost, Gradient Boost and Random Forest followed by Decision Tree, Support Vector Ma-

Model	Cross Val Accuracy	Accuracy	Cross val FRecall	Recall	F1 Score
decision_tree	0.943548	0.935897	0.971727	0.989796	0.950980
SVM	0.941935	0.961538	0.976721	0.980000	0.970297
Neural Network	0.941935	0.961538	0.976721	0.934783	0.950276
GaussianNB	0.852781	0.833333	0.770748	0.742574	0.852273
MultinomialNB	0.769355	0.717949	0.646694	0.605769	0.741176
LogisticRegression	0.956452	0.942308	0.982051	0.979167	0.954315
RandomForestClassifier	0.975806	1.000000	0.981781	1.000000	1.000000
AdaBoostClassifier	0.985484	1.000000	0.994872	1.000000	1.000000
GradientBoosting	0.985484	0.987179	0.992436	0.989130	0.989130

Figure 6: Results : Models and their Scores

chine And Logistic Regression. Neural Network not as good as these models in terms of recall is far superior than Gaussian Naive Bayes and Multinomial Naive Bayes. From Recall And F1_score it can be seen that The Naive bayes algorithms have a high Precision score.

None of the Classifiers are 100% accurate . So, it can be safely said that they can not replace extencive human evaluations . But this models aren't completely useless either,as they can speed up the process of auditing significantly. The auditors can start with the high risk firms that are already classified as high risk. This process would help in detecting the fraud companies faster.

References:

Peter I. Frazier,A Tutorial on Bayesian Optimization,https://arxiv.org/pdf/1807.02811.pdf

Gabe Dickey,Sandra Blanke and Lloyd Seaton,Machine Learning in Auditing Current and Future Applications,https://www.cpajournal.com/2019/06/19/machine-learning-in-auditing/

Leo Breiman and Adele Cutler,Random Forests,
https://www.stat.berkeley.edu/ breiman/RandomForests/cc_home.htm

Ra´ul Roja,AdaBoost and the Super Bowl of ClassifiersA Tutorial Introduction to Adaptive Boostings, http://www.inf.fu-berlin.de/inst/ag-ki/adaboost4.pdf

Jason Brownlee,Your First Deep Learning Project in Python with Keras Step-By-Step.
https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

Shaik Sameeruddin,How Gradient Boosting Algorithm Works,
https://dataaspirant.com/gradient-boosting-algorithm/

Pedregosa et al.Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.Journal of Machine Learning Research,volume 12,pages 2825–2830.

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.

Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness Correlation" (PDF). Journal of Machine Learning Technologies. 2 (1): 37–63. Archived from the original (PDF) on 2019-11-14.