Fraudulent Firm Classification

A case study of an external audit

Summary

- Internal auditing is an objective assurance and consulting activity designed to add value and improve an organization's operations. Its role includes detecting, preventing, and monitoring fraud risks and addressing those risks

- The report is specially made to explore the usefulness of Data Analytics and Machine Learning Algorithms for enhancing the quality of audit work

- Predictive Analytics provides actionable insights for the audit companies. One of the most common applications of predictive analytics in audit is the classification of suspicious firm. 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.

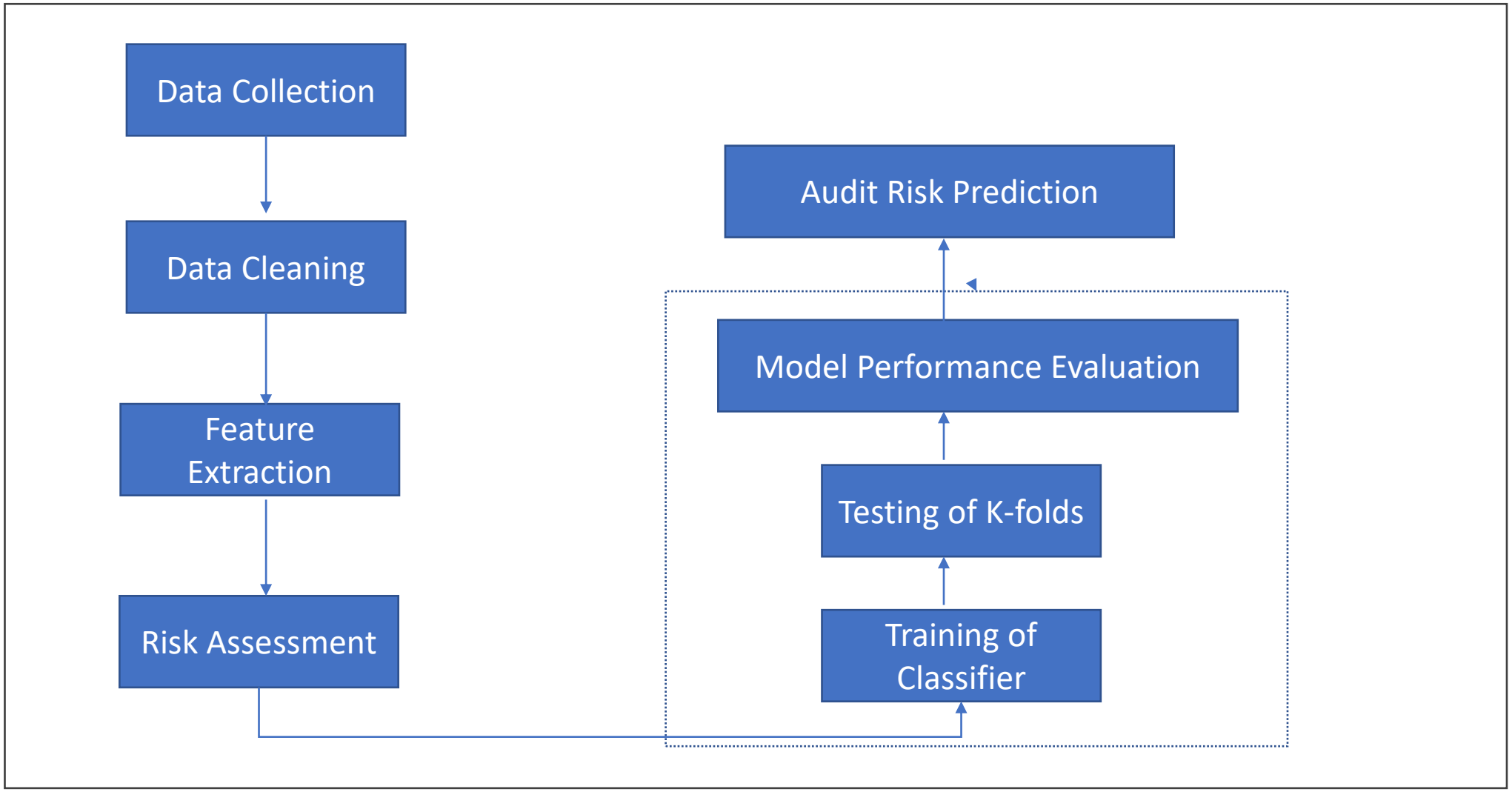
Case Study

- Case study is about 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

- Annual data of 777 firms from 14 different sectors are collected

- So, the underlying objective of this analysis is to examine the present and historical risk factors for determining the Risk Audit Score for 777 target firms and evaluating the Risk Audit Class (Fraud/ No-Fraud) of nominated firms

Prediction Model - Flow



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

- 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

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

Data Sample

Sector_sc	LOCATION	PARA_A	Score_A	Risk_A	PARA_B	Score_B	Risk_B	TOTAL	numbers	Score_B	Risk_C	Money_V	Score_MV
3.89	23	4.18	0.6	2.508	2.5	0.2	0.5	6.68	5	0.2	1	3.38	0.2
3.89	6	0	0.2	0	4.83	0.2	0.966	4.83	5	0.2	1	0.94	0.2
3.89	6	0.51	0.2	0.102	0.23	0.2	0.046	0.74	5	0.2	1	0	0.2
3.89	6	0	0.2	0	10.8	0.6	6.48	10.8	6	0.6	3.6	11.75	0.6
3.89	6	0	0.2	0	0.08	0.2	0.016	0.08	5	0.2	1	0	0.2
3.89	6	0	0.2	0	0.83	0.2	0.166	0.83	5	0.2	1	2.95	0.2
3.89	7	1.1	0.4	0.44	7.41	0.4	2.964	8.51	5	0.2	1	44.95	0.6
3.89	8	8.5	0.6	5.1	12.03	0.6	7.218	20.53	5.5	0.4	2.2	7.79	0.4
3.89	8	8.4	0.6	5.04	11.05	0.6	6.63	19.45	5.5	0.4	2.2	7.34	0.4
3.89	8	3.98	0.6	2.388	0.99	0.2	0.198	4.97	5	0.2	1	1.93	0.2
3.89	8	5.43	0.6	3.258	10.77	0.6	6.462	16.2	5	0.2	1	4.42	0.2
3.89	8	15.38	0.6	9.228	40.14	0.6	24.084	55.52	5	0.2	1	0.96	0.2
3.89	8	5.47	0.6	3.282	7.63	0.4	3.052	13.1	5	0.2	1	10.43	0.6
3.89	8	1.09	0.4	0.436	0.35	0.2	0.07	1.44	5	0.2	1	0	0.2

Data Cleaning

- Data cleaning is one of the most significant process in Data Analysis because purer the data, more accurate will be the prediction
- There are multiple ways to clean the data like Removing/Imputing the missing values, formatting the data with appropriate standards, Excluding the outliers, etc.

```
df_audit.isna().sum()
```

```
Out[12]: Sector_score      0
         PARA_A            0
         Score_A           0
         Risk_A            0
         PARA_B            0
         Score_B           0
         Risk_B            0
         numbers           0
         Score_B.1         0
         Risk_C            0
         Money_Value       1
         Score_MV          0
         Risk_D            0
         District_Loss     0
         PROB              0
         RiSk_E            0
         History           0
         Prob              0
         Risk_F            0
         Score             0
         Inherent_Risk     0
         CONTROL_RISK      0
         Detection_Risk    0
         Audit_Risk        0
         Risk              0
         dtype: int64
```

Example:

```
df_audit['Money_Value'].fillna((df_audit['Money_Value'].mean()), inplace=True)
```

```
Out[15]: Sector_score      0
         PARA_A            0
         Score_A           0
         Risk_A            0
         PARA_B            0
         Score_B           0
         Risk_B            0
         numbers           0
         Score_B.1         0
         Risk_C            0
         Money_Value       0
         Score_MV          0
         Risk_D            0
         District_Loss     0
         PROB              0
         RiSk_E            0
         History           0
         Prob              0
         Risk_F            0
         Score             0
         Inherent_Risk     0
         CONTROL_RISK      0
         Detection_Risk    0
         Audit_Risk        0
         Risk              0
         dtype: int64
```

Feature Extraction

- Many risk factors are 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 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.

- 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.

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*)		

Prediction Classifier

- There are various classification algorithms, but here we will be using simple Classification algorithms like Logistic regression and Random forest and will compare which provides better performance and proceed further with prediction, feature importance, etc.

- 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

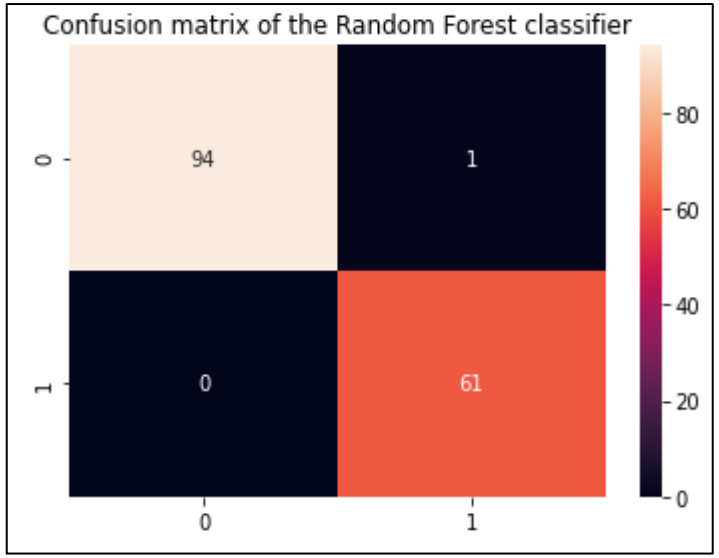
- Random Forest : It is an ensemble learning algorithm which builds a forest of decision trees using random inputs to improve the classification rate

Performance Evaluation

	Model	Accuracy	Cross Val Accuracy	Precision	Recall	F1 Score	ROC	
	0	Logistic Regression	0.967949	0.977419	0.951613	0.967213	0.95935	0.967817
	1	Random Forest	0.993590	0.998387	0.983871	1.000000	0.99187	0.994737

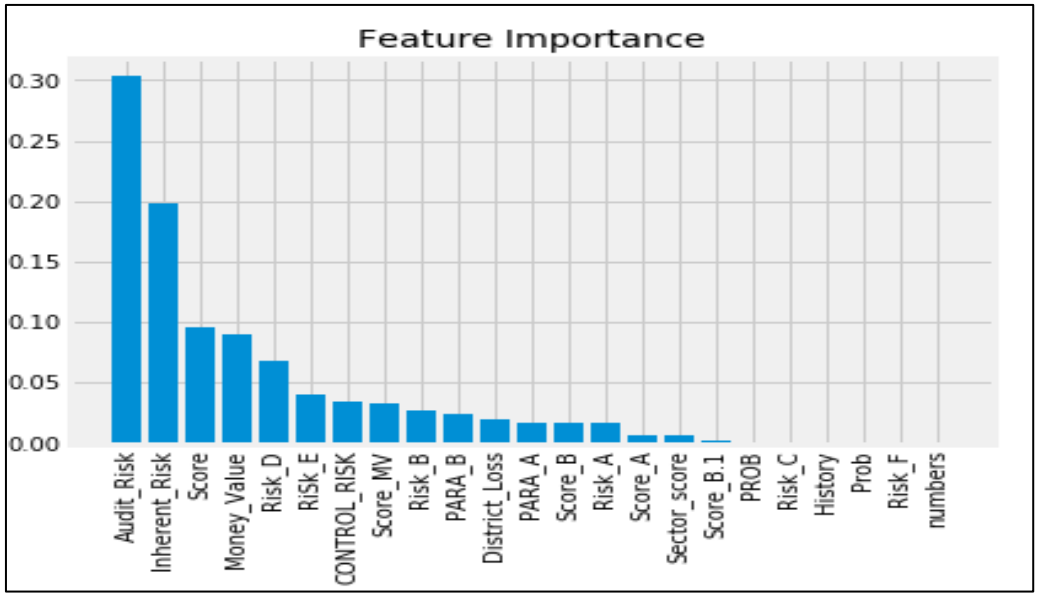
False Negative

True Positive



False Positive

True Negative



Classifier predicted one firm as potentially fraud, but it is not

Classifier ranked important features which contributes to higher risk

Conclusion/Recommendation

- Model selected: Random forest classifier
- Based on the result Investigation experts can move forward with further investigation

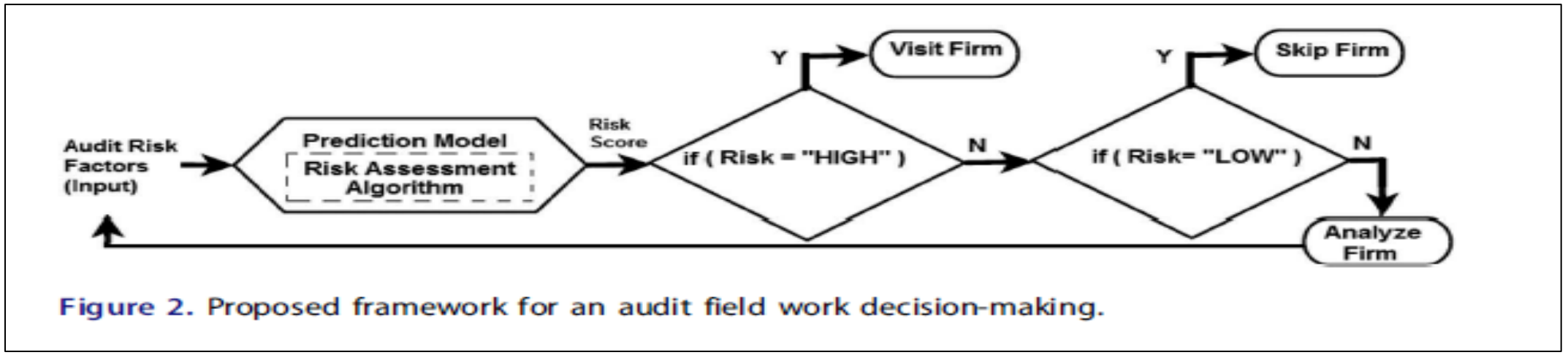


Figure 2. Proposed framework for an audit field work decision-making.

- Consider features with higher risk rank while investigating potentially fraud companies
- Limitation: a) Data was too small to provide accurate and unbiased prediction
  b) Because of lack of data due to confidentiality firm name were not provided

Thank you

Ankil Mehta

Email: ankilmehta19@gmail.com

Contact : +1 7737393518