International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

Predicting Fraudulent Firms based on Risk Factors

Omkar Narvekar[1], Yogesh Jeswani[2], Mona Deshmukh[3]

[1]PG Student, Vivekanand Education Society's Institute of Technology, Dept. of MCA, Mumbai, India
[2]PG Student, Vivekanand Education Society's Institute of Technology, Dept. of MCA, Mumbai, India
[3]Assistant Professor, Vivekanand Education Society's Institute of Technology, Dept. of MCA, Mumbai, India

---

Abstract - *This paper is a research on building a prediction model that can be used to predict fraudulent and deceitful firms. A machine learning approach is used as to make accurate predictions and to act as an basis on which further audit work can be based on. This approach is based on trying to create models with different learning algorithms to improve accuracy of the model as well as implementing feature reduction techniques for removing overfitting.*

Key Words: **Audits, Prediction, algorithms**

1. INTRODUCTION

An audit is performed usually by an independent body especially onsite to inspect or verify different aspects of a corporation such as financial records, quality control, etc. to ensure everything is done according to proper guidelines. Audits can be performed on entire organizations activities or it can remain focused on a specific function. One of the main objectives of an audit is to find discrepancies and faults in data provided by corporations and building a machine learning model can help this cause. A typical audit has a audit cycle which the auditors follow to execute these audits in a systematic manner. Therefore introducing an audit process to an automated system can help speed up things which can usually take days to implement. Automating an audit process should be the goal of an auditor as to make the audit process more efficient. As more and more audits are going on a process to become automated, classifying and predicting discrepancies during a companies activities while maintaining accuracy as a desired objective. Many attributes are analyzed during the working of machine learning model but one attribute that is helpful in predicting this faults in these companies is Risk factor. Risk factors are calculated on many other circumstantial attributes present in the dataset.

2. About the Dataset

The Dataset was acquired from UCI Machine Learning Repository which is a well known website. The dataset contains annual non-confidential knowledge on multiple corporations working in different field of expertise for the fiscal year 2015-2016. This data is collected by the auditor office (CAG) of India. The data contains information about 777 firms from 14 totally different sectors. The sectors are listed below

- Irrigation (114)
- Public Health (77)
- Buildings and Roads (82)
- Forest (70)
- Corporate (47)
- Husbandry (95)
- Communication (1)
- Electrical (4)
- Land (5)
- Science and Technology (3)
- Tourism (1)
- Fisheries (41)
- Industries (37)
- Agriculture (200)

The dataset was made accessible through a pair of csv files named as – audit_risk and trial. The audit_risk.csv contains 27 columns in total and trial.csv contains 18 columns in total.

E.g. Sector_Score, LOCATION_ID, etc. are the contents of the two csv files.

From all of these columns in this dataset, two main columns (PARA_A, PARA_B) are vital in risk calculation. Each of these columns contain inconsistencies found in planned and unplanned expenses.

2. Working on Data

After observing the data for greater understanding of the content it holds we start the actual work by processing data for removing outliers and redundancy as well as dealing with blank spaces. The following steps define how processing of data takes place

2.1 Data Processing

After preliminary evaluation plenty of data cleanup takes place, like only keeping distinctive values and removing

duplicate data. Renaming columns having similar names but having difference of letter case. Columns having values multiplied by 10 are dealt with. Dropping rows having string values where integers should be present. Replacing missing values. Removing outliers.

```
audit_risk.describe()
```

	Sector_score	PARA_A	Score_A	Risk_A	PARA_B	Score_B	Risk_B	TOTAL	numbers
count	776.000000	776.000000	776.000000	776.000000	776.000000	776.000000	776.000000	776.000000	776.000000
mean	20.184536	2.450194	0.351289	1.351029	10.799988	0.313144	6.334008	13.218481	5.067655
std	24.319017	5.678870	0.174055	3.440447	50.083624	0.169804	30.072845	51.312829	0.264449
min	1.850000	0.000000	0.200000	0.000000	0.000000	0.200000	0.000000	0.000000	5.000000
25%	2.370000	0.210000	0.200000	0.042000	0.000000	0.200000	0.000000	0.537500	5.000000
50%	3.890000	0.875000	0.200000	0.175000	0.405000	0.200000	0.081000	1.370000	5.000000
75%	55.570000	2.480000	0.600000	1.488000	4.160000	0.400000	1.840500	7.707500	5.000000
max	59.850000	85.000000	0.600000	51.000000	1264.630000	0.600000	758.778000	1268.910000	9.000000

Chart -1: Columns in audit_risk.csv

```
trial.describe()
```

	Sector_score	PARA_A	SCORE_A	PARA_B	SCORE_B	TOTAL	numbers
count	776.000000	776.000000	776.000000	776.000000	776.000000	776.000000	776.000000
mean	20.184536	2.450194	3.512887	10.799988	3.131443	13.218481	5.067655
std	24.319017	5.678870	1.740549	50.083624	1.698042	51.312829	0.264449
min	1.850000	0.000000	2.000000	0.000000	2.000000	0.000000	5.000000
25%	2.370000	0.210000	2.000000	0.000000	2.000000	0.537500	5.000000
50%	3.890000	0.875000	2.000000	0.405000	2.000000	1.370000	5.000000
75%	55.570000	2.480000	6.000000	4.160000	4.000000	7.707500	5.000000
max	59.850000	85.000000	6.000000	1264.630000	6.000000	1268.910000	9.000000

Chart -2: Columns in trial.csv

2.2 Merging Dataset

We use common columns in two data frames and merge them together. This is done to understand the data that these data frames hold. After merging we found that SCORE_A and SCORE_B are present in both datasets i.e. audit_risk.csv and trial.csv and the one present in trial.csv is multiplied by 10 thereby making them redundant. So we drop these two columns so as to remove the redundancy. Other columns which are found to be redundant are removed as well. Columns which are not important to the research and merely act as additional data about the corporations are removed.

2.2 Correlation Matrix

A correlation matrix is used to show correlation between two variables and this correlation between two variables is shown within each cell inside the table. A correlation matrix is implemented to explore patterns between variables. Correlation can be of 3 types positive, negative and neutral. The high amount of correlation can suggest that the performance of an algorithm can deteriorate and final output will be unreliable.

	Risk_A	Risk_B	Risk_C	Risk_D	RiSk_E	Prob	Score	CONTROL_RISK	Audit_Risk	Risk	MONEY_Marks	Loss
Risk_A	1.00	0.16	0.14	0.45	0.12	0.17	0.43	0.15	0.22	0.38	0.29	0.04
Risk_B	0.16	1.00	0.22	0.12	0.08	0.32	0.40	0.19	0.89	0.25	0.31	0.04
Risk_C	0.14	0.22	1.00	0.21	0.15	0.24	0.55	0.25	0.25	0.34	0.49	0.00
Risk_D	0.45	0.12	0.21	1.00	0.03	0.11	0.29	0.07	0.33	0.25	0.39	0.02
RiSk_E	0.12	0.08	0.15	0.03	1.00	0.12	0.23	0.73	0.20	0.41	0.10	0.37
Prob	0.17	0.32	0.24	0.11	0.12	1.00	0.44	0.64	0.43	0.30	0.33	0.10
Score	0.43	0.40	0.55	0.29	0.23	0.44	1.00	0.35	0.33	0.78	0.76	0.16
CONTROL_RISK	0.15	0.19	0.25	0.07	0.73	0.64	0.35	1.00	0.36	0.41	0.22	0.28
Audit_Risk	0.22	0.89	0.25	0.33	0.20	0.43	0.33	0.36	1.00	0.22	0.29	0.05
Risk	0.38	0.25	0.34	0.25	0.41	0.30	0.78	0.41	0.22	1.00	0.69	0.17
MONEY_Marks	0.29	0.31	0.49	0.39	0.10	0.33	0.76	0.22	0.29	0.69	1.00	0.12
Loss	0.04	0.04	0.00	0.02	0.37	0.10	0.16	0.28	0.05	0.17	0.12	1.00

Chart -3: Display of correlation matrix

3. Results on data

After all the data processing has taken place we apply multiple algorithms to the data and try to create a machine learning model to accurately predict fraudulent data. Some of the algorithms applied are Linear Regression, K-Nearest Neighbors, Decision Tree Classifier, Linear SVC(support vector classifier), Logistic Regression. Bagging and pasting ensemble methods are used with algorithms to sample smaller subsets of data instead of a single dataset. Bagging samples with replacement and pasting does not. Adaboost is another ensemble method that is used while creating a model. Adaboost works by combining multiple weak learners to create a strong learner which can improve the accuracy of the algorithm. Adaboost learns from multiple iterations to correct itself in predicting the final output. The following steps display results on data after applying multiple algorithms

3.1 Without implementing Principal Component Analysis

Model Name	Training Score	Testing Score
Logistic Regression with Bagging	0.980702	0.963158
KNN with Bagging	0.980702	0.963158
Decision Tree Classifier with Pasting	0.396491	0.410526
Linear SVC with Pasting	0.980702	0.963158
Logistic Regression with Adaboost	1	0.978947

Chart -4: Output for first set of algorithms

3.2 With implementing Principal Component Analysis

The main concept behind principal component analysis is to reduce the dimensionality of data, this data can highly correlated and as such its dimensionality should be reduced as to avoid the possibility of overfitting the model. Overfitting can lead to data with high amount of noise and as such the model will consider noisy data as well which will lead to an unrealistic model. The PCA reduces dimensionality by preserving the highest variation present in the dataset. PCA does dimensionality reductional also called feature reduction by transforming them into a new set called as

principal components. PCA is generally applied on large datasets. Reducing the features in a dataset can lead to less accurate results but it also results in data which is simpler in terms of computational complexity.

Model Name	Training Score	Testing Score
Logistic Regression with PCA	0.959649	0.957895
KNN Classifier with PCA	0.977193	0.942105
SVC with PCA	0.957895	0.942105
Decision Tree Classifier with PCA	0.996491	0.957895

Chart -5: Output for second set of algorithms

Algorithmic models with reduced features leads to less noisy data but as the results suggest PCA reduces the overall testing score slightly.

4. CONCLUSIONS

So after creating all the different models and implementing them on training and testing data, the best model for prediction among all the models as of now is Logistic regression. However models tends to lose their accuracy by applying Principal Component Analysis even as PCA leads to data with reduced features and more meaningful data.

REFERENCES

[1] "Fraudulent Firm Classification: A Case Study of an External Audit" 6 Apr. 2018, https://www.researchgate.net/publication/323655455_Fraudulent_Firm_Classification_A_Case_Study_of_an_External_Audit (Accessed 21 Jun. 2020).

[2] "Optimizing Fraudulent Firm Prediction Using Ensemble Machine Learning: A Case Study of an External Audit" 4 Nov.2019, https://www.researchgate.net/publication/339284563_Optimizing_Fraudulent_Firm_Prediction_Using_Ensemble_Machine_Learning_A_Case_Study_of_an_External_Audit (Accessed 21 Jun. 2020).

[3] "Audit Data Dataset - UCI Machine Learning Repository."14Jul.2018, https://archive.ics.uci.edu/ml/datasets/Audit+Data (Accessed 20 Jun.2020).

[4] "The Top 10 Machine Learning Algorithms for ML Beginners." https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners (Accessed 20 Jun, 2020).