Unsupervised Detection of Misinformation in Financial Statements

Akshada Shinde, Sushodhan Vaishampayan, Manoj Apte, Girish Keshav Palshikar
TCS Research, Tata Consultancy Services Ltd., Pune, India 411013
(akshada.shinde, sushodhan.sv, manoj.apte, gk.palshikar)@tcs.com

Abstract

Companies have incentives to hide, omit, or falsify the information reported in financial statements (FS) (e.g., Balance Sheet, Income Statement, Cash Flow Statement) to give a false impression of the company's financial health, assure investors or evade taxes. Typically, misinformation is introduced by changing FS elements e.g overstating the assets/profits or understating the liabilities/losses. Once detected, misinformation can have disastrous consequences for employees, investors, banks and government. It is important to identify such companies and the nature and extent of misinformation in their FS. Auditors or forensic accountants use complex investigative methods to detect instances of misinformation in FS. The effort intensive and subjective nature of these methods limits their capacity to effectively identify misinformation. We propose two novel unsupervised model-based anomaly detection (AD) techniques based on regression and kernel density estimates. We show they perform better than 15 standard AD techniques and data envelopment analysis for detection of suspicious FS on a real-world dataset of 4100 listed companies. Our approach provides specific suggestions regarding where the misinformation may be present, which helps in increasing the effectiveness of investigations.

Introduction

A company summarizes its financial performance annually in various standardized, structured *financial statements (FS)* such as *balance sheet (BS), profit and loss statement (P&L), cash-flow statement* etc. Accountants adhere to *generally accepted accounting principles (GAAP) and international financial reporting standards (IFRS)* when preparing FS. In many practical applications, (e.g., corporate governance, credit appraisal, risk analysis, taxation, auditing, investment decisions etc.), the contents of these FS (which are usually available to the public for listed companies) are carefully examined from different perspectives (Feldman and Libman 2007), (Drake and Fabozzi 2012).

Given the importance of FS, there are obvious incentives to hide, omit or falsify information to misrepresent the true financial health of the company; e.g., reduce tax liabilities, increase investor confidence etc. (Beasley, Carcello, and Hermanson 1999). Typical kinds of misinformation in FS include *overstating* the firm's assets, revenues and profits, or *understating* the firm's liabilities, expenses and losses[1]. It is difficult to estimate the extent to which misinformation is prevalent in FS. However, it is well-known that many large corporate frauds can be traced back to accounting misinformation; e.g., Enron in 2001 in USA (Healy and Palepu 2003), WorldCom in 2003 in USA[2] and, Satyam Computers in 2009 in India (Bhasin 2013). Misinformation in FS leads to monetary losses to customers, employees, and creditors and also loss of reputation, trust and goodwill.

Auditors and forensic accountants have developed a range of investigative techniques for verifying the sources of the numbers reported in FS and thereby identifying misinformation (Nigrini 2020). These techniques are mostly manual and depend on rich domain knowledge of the human experts. Given the effort-intensive and subjective nature of these investigations, many analytic techniques have been developed for identifying FS that are likely to contain some misinformation; we survey some of them in Related Work section.

In this paper, we use several unsupervised techniques based on regression, kernel density estimation (KDE), anomaly detection (AD) and *data envelopment analysis (DEA)* for the task of detection of suspicious FS. We report results on a public dataset of 4100 companies listed in India.

Related Work

Supervised Approaches: Each FS is usually represented as a vector of well-known financial ratios. Along with the FS, many companies also release an auditor report, which mentions any issues, or errors the auditors found in the FS (such FS are referred to as *qualified*). A qualified FS is labeled as a positive case of misinformation and an unqualified FS as negative, to create a labelled dataset of FS. (Spathis 2002) learnt a binary logistic regression classifier on a labeled training dataset of FS of 76 manufacturing firms (38 fraudulent + 38 normal), each represented using 10 financial ratios. (Kotsiantis et al. 2006) represented FS of 164 firms for 2001-02 (41 fraudulent) using 23 financial ratios, used rank influence to select 8 ratios and trained various

[1] https://www.acfe.com

[2] https://www.scu.edu/ethics/focus-areas/business-ethics/resources/worldcom/

classification models to report up to 96.7% classification accuracy for stacking ensemble. (Chen 2016) propose a two-stage system for detecting fraudulent FS. In stage 1, CART and CHAID models are used to find variables which are significant for this task. In Stage 2 CART, CHAID, Bayesian Belief Network, SVM and Artificial Neural Network classifiers are trained and then combined for detecting fraudulent FS. A similar approach is used in (Jan 2018) on 160 companies (40 fraudulent) from Taiwan Stock Exchange, yielding 90.83% classification accuracy for detecting fraudulent FS.

Unsupervised Approaches: (Mongwe and Malan 2020) used Self-Organizing Maps (SOM) on an unlabeled FS of 1560 South African municipalities represented using 3 financial ratios, clustered the resulting map using K-Means clustering and used adverse auditor comments to label some clusters as fraudulent. (Lokanan, Tran, and Vuong 2019) used a mahalanobis distance-based anomaly detection method on an unlabeled FS of 937 Vietnamese companies (each represented using 24 financial ratios) to identify anomalous FS (possibly containing misinformation).

A special form of accounting errors in FS is that numerical values of semantically equivalent facts (mentioned in different places in the FS) might be inconsistent. (Li et al. 2020) present a system to find whether two cells in a table mention the same fact using neural networks.

Misinformation is present in other financial documents (e.g., tax reports) than FS, mainly for the purpose of tax evasion where unsupervised techniques are commonly used for identifying suspicious tax reports (de Roux et al. 2018), (Matos, de Macedo, and Monteiro 2015), (Gonzalez and Velasquez 2013), (Wu et al. 2012). Broadly, the tax reports are first clustered, then patterns characterizing each cluster (e.g., association rules, in-cluster probability distributions, cluster-specific CHAID classification trees) are learnt, which are then used to identify suspicious tax reports.

Dataset

Auditors can comment about the descrepancy in the process followed by the company for reporting the numbers and also possibly the effect on the FS elements. In this paper we focus on artificial manipulation in the FS elements. FS and other financial documents (annual results, financial ratios, capital structure, annual reports, audit reports) for about 8000 Indian listed companies are available[3] for 10 years. (Maka, Pazhanirajan, and Mallapur 2020) used similar dataset for selecting significant features and identifying fraudulent FS. We web-scrapped the FS of 4100 companies which were operating in the year 2014 (we focus on balance sheets in this paper). Tables 1 and 2 describe variables and financial ratios (with summary statistics) respectively used for finding suspicious BS (values are in units of Rupees 10 million). If a ratio had value ∞ (i.e., the denominator had value 0), we replaced it with the average value of the ratio. Some of the top correlated pairs of ratios are: $(R_4, R_6, 0.83)(R_8, R_{11}, 0.68)(R_5, R_{12}, 0.51)(R_1, R_2, 0.46)$ $(R_2, R_3, -0.40)(R_2, R_{13}, -0.40)(R_9, R_{11}, 0.33)$ $(R_3, R_6, 0.32)$. The unavailability of labeled data makes using supervised techniques inappropriate.

[3]https://www.moneycontrol.com/

Table 1: Variables in BS and summary statistics.

Name	mean	stdev	Q1	Q2	Q3
V_1 *Trade Receivables*	128.6	712.7	0.2	6.4	45.6
V_2 *Current Assets*	606.6	4018.8	4.5	29.6	165.8
V_3 *Non-current Assets*	1002.5	7881.3	4.3	24.9	178.7
V_4 *Total Assets*	3471	37580.8	12.3	62.4	392.5
V_5 *Fixed Assets*	541.5	4970.1	0.5	10.1	93.9
V_6 *Inventories*	157.0	1464.6	0	4.2	37.7
V_7 *Current Liabilities*	509.2	3363.9	1.5	20.0	140.6
V_8 *Cash & Cash Eqvt.*	99.26	1007.4	0.13	1.12	9.01
V_9 *Non-current Liabilities*	470.2	4305.1	0.13	4.66	45.1
V_{10} *Shareholders Funds*	627.2	5008.7	4.8	23.3	145.2
V_{11} *Total Liabilities*	979.4	6862.0	3.1	29.8	206.8

Table 2: Financial ratios and summary statistics.

Name	Formula	mean	std	Q1	Q2	Q3
R_1	V_1/V_4	0.16	0.18	0.01	0.1	0.24
R_2	V_2/V_4	0.51	0.29	0.28	0.51	0.74
R_3	V_5/V_4	0.25	0.25	0.02	0.19	0.41
R_4	$\log(V_4)$	4.28	2.59	2.52	4.15	5.97
R_5	V_8/V_4	0.06	0.12	0.01	0.02	0.05
R_6	$\log(V_{11})$	3.16	3.09	1.31	3.33	5.25
R_7	$(V_2 - V_6)/V_7$	12.04	80.56	0.56	1.03	2.71
R_8	V_{11}/V_4	1.04	7.6	0.21	0.52	0.75
R_9	V_7/V_2	3.86	47.99	0.31	0.71	1.03
R_{10}	V_{11}/V_{10}	2	41.75	0.1	0.74	2
R_{11}	V_7/V_4	0.6	4.42	0.11	0.32	0.53
R_{12}	V_8/V_2	0.14	0.28	0.02	0.05	0.14
R_{13}	V_6/V_4	0.13	0.16	0	0.08	0.21

Detection of Suspicious FS

We use *precision@20* as the evaluation metric, which is the fraction of qualified BS in the top 20 BS reported as anomalous by any misinformation detection algorithm. Based on our study of the dataset, we treat a BS as qualified if it is qualified *either* in the current year 2014 *or* in the next year 2015, since the effects of misinformation sometimes accumulate and amplify over time, which means it may be easier for an auditor to catch it next year.

Anomaly Detection

One reasonable hypothesis is that any BS which contains some misinformation would appear *anomalous* in some sense, as compared to honest BS. To test this idea, we used several well-known AD algorithms implemented in PyOD package (Zhao, Nasrullah, and Li 2019) to identify anomalous (i.e., suspicious) BS. The *ensemble* method takes top 20 BS having the highest count of AD algorithms which marked it as anomalous. Table 3 shows the P@20 for 16 AD algorithms. As seen, Connectivity-Based Outlier Factor (COF) algorithm (Tang et al. 2001) has the highest P@20 (0.25) i.e., 5 out of 20 BS identified by it as anomalous are indeed marked as qualified by the auditors.

Table 3: Anomaly detection methods.

Algorithm	P@20	Algorithm	P@20
iForest	0.10	Mahalanobis	0.15
kNN	0.05	ABOD	0.05
LOF	0.00	CBLOF	0.10
COF	**0.25**	HBOS	0.00
PCA	0.15	OCSVM	0.05
LMDD	0.10	LODA	0.10
SOD	0.15	SOS	0.20
MCD	0.05	ensemble	0.10

Model-based Anomaly Detection

One issue with the existing approaches is that most of them do not pinpoint *where* exactly the misinformation is present in a FS, which restricts the utility of the results. To tackle this, we notice that some variables in a BS - e.g., those related to assets or liabilities - are more *susceptible* for misinformation than others (e.g., auditors adverse remarks are more often about these variables). We have considered all the variables related to liabilities as susceptible variables. We propose two novel model based approaches to identify suspicious BS. In the first approach, we build regression models and use them to identify suspicious BS. We start with a given susceptible variable, identify other variables on which that susceptible variable depends, and use only those to build best regression model(s) for that suspicious variable. Finally, we use these regression models to identify suspicious BS. The approach consists of the following steps:

1. Select a susceptible variable (say, Y);
2. On highly correlated variables, use stepwise regression to incrementally build multiple OLS regression models for Y and select the best regression model M_Y for Y having the highest adjusted R^2 value;
3. Use M_Y to identify suspicious BS as follows. Take a BS, predict the value of Y using the values (in this BS) of the independent variables used in M_Y, compute the prediction error $Y - \hat{Y}$ and mark those BS as suspicious which have the highest squared prediction error in Y using M_Y.

To illustrate, following are examples of 4 best regression models built using stepwise OLS regression for susceptible variables R_8, R_9, R_{10}, R_{11}.

M3: $R_8 \leftarrow R_5\ R_{12}\ V_4\ R_3\ V_1\ V_2\ V_3\ V_5\ V_6\ V_8\ V_{10}\ R_2\ R_1\ R_{13}\ R_4$
M4: $R_9 \leftarrow R_{12}\ R_3\ V_4\ V_1\ V_2\ V_3\ V_5\ V_6\ V_8\ V_{10}\ R_5\ R_4\ R_1\ R_{13}\ R_2$
M5: $R_{10} \leftarrow R_3\ R_4\ V_2\ V_6\ R_1\ R_{12}$
M6: $R_{11} \leftarrow R_{12}\ R_5\ R_1\ V_4\ V_6\ R_3\ V_1\ V_2\ V_3\ V_5\ V_8\ V_{10}\ R_2\ R_{13}\ R_4$

Table 4 shows the P@20 values for these models, along with P@20 for the models having the same input-output structure but built using Lasso, Support Vector Regression (SVR) and Random Forest Regression (RFR) algorithms. As seen, the ensembles of OLS, SVR and Lasso models have good P@20, significantly higher than any of the standard AD algorithms. A remarkable aspect of this approach is that even when the fitted regression model is not that good (has low adjusted-R^2 value), it is still able to detect suspicious BS rather well. A reason is that while the predictions of a poor regression model are consistently bad for all points, they are much worse for anomalous points, which means such points tend to have higher prediction errors. Importantly, if a particular model detects a BS as suspicious, the auditors can focus on the variables used by the model as candidates where misinformation may be present.

Table 4: Regression models.

Model	P@20	Model	P@20
OLS M3	0.35	Lasso M3	0.25
OLS M4	0.40	Lasso M4	0.40
OLS M5	0.25	Lasso M5	0.25
OLS M6	0.45	Lasso M6	0.45
OLS ensemble	**0.50**	**Lasso ensemble**	**0.50**
SVR M3	0.30	RFR M3	0.20
SVR M4	0.45	RFR M4	0.25
SVR M5	0.25	RFR M5	0.15
SVR M6	0.45	RFR M6	0.20
SVR ensemble	**0.50**	RFR ensemble	0.35

Our second approach for model-based anomaly detection is identical to the first approach, except that we use mutual information (MI) to identify variables on which the given susceptible variable has strong dependency. As an example, the susceptible variable R_{10} has strong dependency on R_8, R_{11}, R_9, having MI values $228.2, 75.9, 62.8$. Now we use kernel density estimation (R package `ks`) to learn a model from the data, which has the form of a joint probability distribution over these 4 variables. Then we use this distribution to compute the probability of observing each 4-tuple (in each row) and report the lowest 20 as suspicious. For example, the estimated probability density for the tuple $(R_8 = 0.98\ R_9 = 0.21\ R_{10} = 55.2\ R_{11} = 0.21)$ is 0.0179. Table 5 shows 4 models and their P@20 values.

Table 5: KDE based models.

Model	Structure	P@20
M10	$R_8 \leftarrow R_{10}\ R_{11}\ R_7$	0.20
M11	$R_9 \leftarrow R_7\ R_{10}\ R_{11}$	**0.40**
M12	$R_{10} \leftarrow R_8\ R_{11}\ R_9$	**0.40**
M13	$R_{11} \leftarrow R_8\ R_9\ R_{10}$	**0.40**

Data Envelopment Analysis

We tried an *Operation Research* technique DEA (Ramanathan 2003) for detecting suspicious BS. DEA solves an optimization problem to compute the *relative efficiency* of organizational units, called *Decision Making Units (DMUs)*, which are functionally similar to each other. Each DMU consumes some input and produces some output. The DMUs which either consume less amount of input and produce same amount of output as others or consume same amount of input and produce larger output are termed as *efficient*. We took assistance from the regression models built in previous examples to formulate the DEA optimization problem. The independent variables from these models were considered as inputs and dependent variables were considered as

outputs of each DMU for the optimization problem. We selected the 20 least efficient companies as the suspicious ones using the efficiency computed by DEA. The models didn't seem to perform as good as the regression techniques. We used 7 different DEA models (formulations). The models and their respective outputs are given in the Table 6. DEA model based on the regression model $M6$ has the highest $P@20$ of 0.25; other DEA models performed poorly.

Table 6: DEA Models.

Model	Inputs	Outputs	P@20
F1	$\{V_1, V_2, V_3, V_4, V_5, V_6, V_8, V_{10}\}$	$\{V_{11}\}$	0.05
F2	$\{V_1, V_2, V_3, V_4, V_5, V_6, V_8, V_{10}\}$	$\{V_7\}$	0.05
F3	$\{V_1, V_2, V_3, V_4, V_5, V_6, V_8\}$	$\{V_{11}\}$	0.10
F4	$\{V_1, V_2, V_3, V_4, V_5, V_6, V_8\}$	$\{V_{10}, V_{11}\}$	**0.25**
F5	$\{V_3, V_4, V_5, V_6, V_{10}\}$	$\{V_7\}$	0.00
F6	$\{V_3, V_4, V_5, V_6, V_{10}\}$	$\{V_9\}$	0.00
F7	$\{V_3, V_4, V_5, V_{10}\}$	$\{V_{11}\}$	0.00

Conclusions and Further Work

Detecting misinformation in FS is important. We proposed two novel unsupervised model-based AD techniques based on regression and KDE. We applied them to a real-world dataset of balance sheets of 4100 listed companies and showed that these techniques performed better than strong baselines of 15 standard AD techniques and AD based DEA. Each technique detects different qualified statements as anomalous. Our techniques are able to provide specific scenarios where the misinformation may be present. Our approach can help auditors to decide the focus and depth of their investigations and increase the effectiveness of audits. We are currently working on integrating the information from different FS and then detecting misinformation.

References

Beasley, M.; Carcello, J.; and Hermanson, D. 1999. *Fraudulent financial reporting (1987-1997) an analysis of US public companies*. American Institute of Certified Public Accountants.

Bhasin, M. 2013. Corporate accounting fraud: A case study of Satyam Computers Limited. *Open Journal of Accounting* 2:26–38.

Chen, S. 2016. Detection of fraudulent financial statements using the hybrid data mining approach. *SpringerPlus* 5(1):1–16.

de Roux, D.; Perez, B.; Moreno, A.; del Pilar Villamil, M.; and Figueroa, C. 2018. Tax fraud detection for underreporting declarations using an unsupervised machine learning approach. In *24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD2018)*.

Drake, P. P., and Fabozzi, F. J. 2012. *Analysis of Financial Statements*. Wiley, 3rd edition.

Feldman, M., and Libman, A. 2007. *Crash Course in Accounting and Financial Statement Analysis*. Wiley, 2nd edition.

Gonzalez, P., and Velasquez, J. 2013. Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications* 40:1427–1436.

Healy, P. M., and Palepu, K. G. 2003. The fall of Enron. *Journal of Economic Perspectives* 17(2):3–26.

Jan, C.-l. 2018. An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from Taiwan. *Sustainability* 10(2):513.

Kotsiantis, S.; Koumanakos, E.; Tzelepis, D.; and Tampakas, V. 2006. Forecasting fraudulent financial statements using data mining. *International journal of computational intelligence* 3(2):104–110.

Li, H.; Yang, Q.; Cao, Y.; Yao, J.; and Luo, P. 2020. Cracking tabular presentation diversity for automatic cross-checking over numerical facts. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2599–2607.

Lokanan, M.; Tran, V.; and Vuong, N. H. 2019. Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. *Asian Journal of Accounting Research*.

Maka, K.; Pazhanirajan, S.; and Mallapur, S. 2020. Selection of most significant variables to detect fraud in financial statements. *Materials Today: Proceedings*.

Matos, T.; de Macedo, J. A. F.; and Monteiro, J. M. 2015. An empirical method for discovering tax fraudsters: A real case study of Brazilian fiscal evasion. In *19th International Database Engineering and Applications Symposium (IDEAS 2015)*, 41–48.

Mongwe, W. T., and Malan, K. M. 2020. The efficacy of financial ratios for fraud detection using self organising maps. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1100–1106. IEEE.

Nigrini, M. J. 2020. *Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations*. Wiley, 2nd edition.

Ramanathan, R. 2003. *An Introduction to Data Envelopment Analysis: A tool for Performance Measurement*. Sage Publishing.

Spathis, C. T. 2002. Detecting false financial statements using published data: some evidence from Greece. *Managerial Auditing Journal*.

Tang, J.; Chen, Z.; chee Fu, A. W.; and Cheung, D. 2001. A robust outlier detection scheme for large data sets. In *6th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2001)*, 6–8.

Wu, R.-S.; Ou, C.; Lin, H.-Y.; Chang, S.-I.; and Yen, D. 2012. Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications* 39:8769–8777.

Zhao, Y.; Nasrullah, Z.; and Li, Z. 2019. PyOD: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research* 20:1–7.