Accounting Fraud Detection with Machine Learning

RMSC4002 Group Project
Team Member:
Leung Cheuk Wai Dominic 1155093086
Yu Wun Man Bobby 1155077997
Wong Cheuk Him Michael 1155064260

Problem

1. Introduction

In light of the book, *Financial Shenanigans*, written by Schilit in 2002, we are interested to find the underlying rules of accounting fraud detection. People had been starting to study company profiles and quality to detect financial fraud in the past. They were paying extra attention to corporate balance sheet and income statement to try to detect the odd figures or outliers, and interpreted the anomaly with knowledge from corporate finance. However, due to the advent of data science method, we are motivated to build an alternative metric to detect financial fraud from accounting statement.

2. Research on fraud detection

From previous research paper, Sharma and Panigrahi (2012)[1] conducted a review on previous method of accounting fraud detection, categorizing different techniques of machine learning, such as Neural Networks, Regression Models, Fuzzy Logic, etc. Regression can make use of both numerical and categorical as predictor variables to predict the likelihood of an accounting fraud. Neural network can learn from a larger set of financial data to identify latent patterns of fraudulent events.

In using supervised learning, we studied the approach and behaviour of company "creatively" making their accounting book. Schilit(2002) pointed out the types of accounting fraud, in which some are outlined as follows:

1. Recording Revenue Too Soon[2]
2. Boosting Income with One-Time Gains[3]
3. Shifting Current Expenses to a Later or Earlier Period[4]

There is a total of seven types of Shenanigans listed out by Schilit (2002), but we find the above three to be the most useful for data mining projects. When we are choosing financial data to be analyzed, we are following the above principles to pick relevant data for our supervised models. Hence, this project is also a good opportunity to prove the idea claimed by Schilit, which he did not justify the findings with sufficient statistics due to technical reason at that time.

For data mining, we are going to employ methods using classification tree and logistics regression to look for predictable variables. In addition, with the same set of data, we are

[1] Sharma, A., & Panigrahi, P.K. (2012). A Review of Financial Accounting Fraud Detection
[2] Schilit, H. M. (2002). *Financial shenanigans*. Warning Signal No. 1
[3] Schilit, H. M. (2002). *Financial shenanigans*. Warning Signal No. 3
[4] Schilit, H. M. (2002). *Financial shenanigans*. Warning Signal No. 4

using Artificial Neural Network to further check if there is an underlying pattern that can gives us a better prediction for fraudulent events.

Description of Dataset

3. Proxy for fraud company

For fraudulent companies, our list of companies come from the dataset from the U. S. Government Accountability Office (U. S. GAO), the legislative agency that governs the auditing service in the United States. They published a Financial Restatement Database which includes companies that were required to resubmit their financial reports in between October 2005 to June 2006.

Since the data is dated back to 2006, our time frame of the data would also be focused in 2006. Accounting principle is not changed a lot, and many lines on the balance sheet and financial statement remain the same from that time up to now, so we believe that it is still appropriate for the use of that dataset. Therefore, for the choice of non-fraud companies, we picked companies from S&P 500 Index on 1st January 2006, a representative index that all companies are assumed behave well in order to be included in that index. There are in total 812 companies including both fraudulent and non-fraudulent companies in our dataset. Many companies are listed on exchanges such as NYSE, Nasdaq, Amex, etc. The list of companies will be used as the proxy for fraud companies as it is reasonable to assume that the companies must have committed some fraud so they are forces by GAO for a resubmission.

Fraudulent Companies	312
Non Fraudulent Companies	500

Component of fraudulent companies in the list:

Exchange Market	Number of Companies
Nasdaq	164
NYSE	111
Amex	37

4. Financial data

Our approach to choose financial data is related to revenue and expense of the company. We should note that accounting fraud is not mainly furnishing the financial reports to make the companies more appealing to investors, but on the other hand, when the market is spiralling downwards, fraudulent companies would prefer to show themselves as less profitable by, for example, deferring the report of earnings so they can enjoy a higher positive market reaction when the market is bullish. Therefore, instead of focusing on data from income statement and looking for abnormal increase in profit, we should also consider its relationship with the company operation profile in our model.

Schilit (2002) mentioned different types of playing with revenue and expense in the accounting report. Hence, we are motivated to find the relationships between a set of profit

and expense data (e.g. Net Income, Gross Margin) which mainly comes from income statement and company's operation data (e.g. Cash flow, Account Receivable) which mainly comes from the balance sheet.

Financial data chosen: (Data are obtained from Bloomberg Terminal)

Financial Data	**Equivalent fields on Bloomberg**
Cashflow from Operation	CF_CASH_FROM_OPER
Net Income	NET_INCOME
Sales Growth	SALES_GROWTH
Account Receivable Growth	ACCOUNTS_RECEIVABLE_GROWTH
Inventory Growth	INVENTORY_GROWTH
Cost of Goods Sold Change	COGS_YR_GROWTH
3 Years Average Gross Margin	3YR_AVG_GROSS_MARGIN
Gross Margin	GROSS_MARGIN

Override: ("FUNDAMENTAL_DATABASE_DATE = 20060101")

*Detailed descriptions are at the back of this report

As our proxy for fraudulent companies covers a long time frame, from October 2005 to June 2006, we would therefore further extend our assumption that the fraudulent companies would show a pattern in their accounting reports, but not to detect for the reason that they are required for a restatement. Therefore, we are not detecting a certain fraud for a certain statement, but instead we are look for the general behaviour of those companies. Hence, our financial data will be overridden on a particular date, 1st January 2006, a middle point of the time frame.

5. Data cleaning part

As data are dated at a time which is quite old, many companies including those from S&P 500 show "#N/A" for some types of financial data. Imputation is not used in this occasion due to the characteristic of financial data. Lines will be deleted in our final dataset[5].

	After Cleaning
Fraudulent Companies	289
Non Fraudulent Companies	87

Some financial data are modified before putting into data mining model. Normalization of data is performed to smooth out the dataset. In addition, when we notice that one or two data from Bloomberg Terminal is so extreme that it is obviously an error (e.g. a 3000% growth in receivables), we will also change the data to NA.

[5] Source("cleanfun.R") is designed for data cleaning

The table below summarizes the modification of data cleaning.

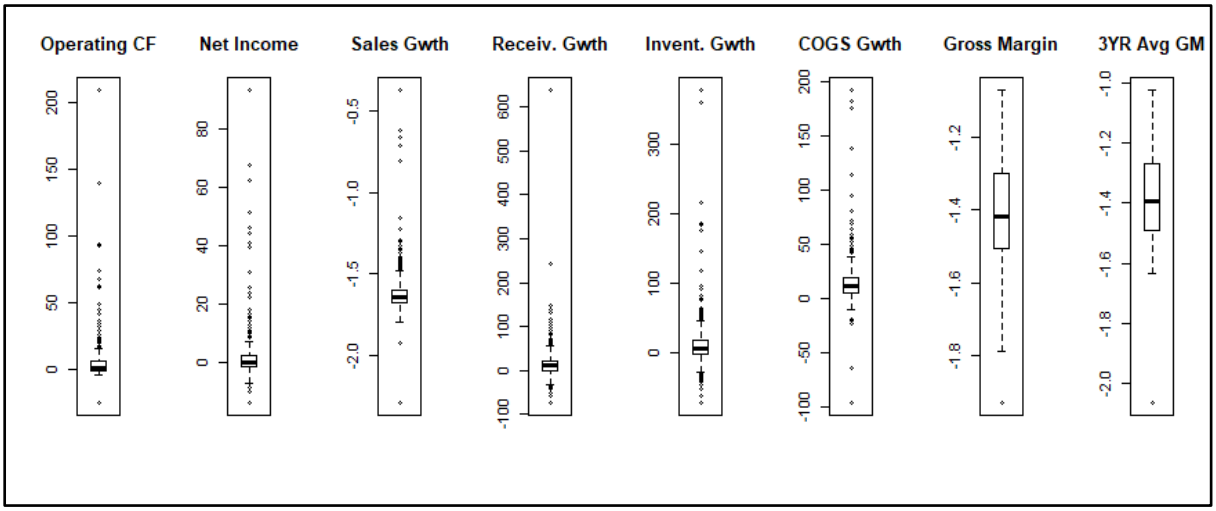
Financial Data	Modification
Cashflow from Operation	Normalized
Net Income	Normalized
Sales Growth	Normalized
Gross Margin	Normalized
3 Year Average Gross Margin	Normalized

Normalize method:

The normalization of data is before deletion of NA value in the dataset, because companies with missing values are not that missing all the items on our list. To try the best to prevent a bias in our data, all non-missing values are used in the normalization part so as to obtain a clearer picture of how the remaining data are located in the original population.

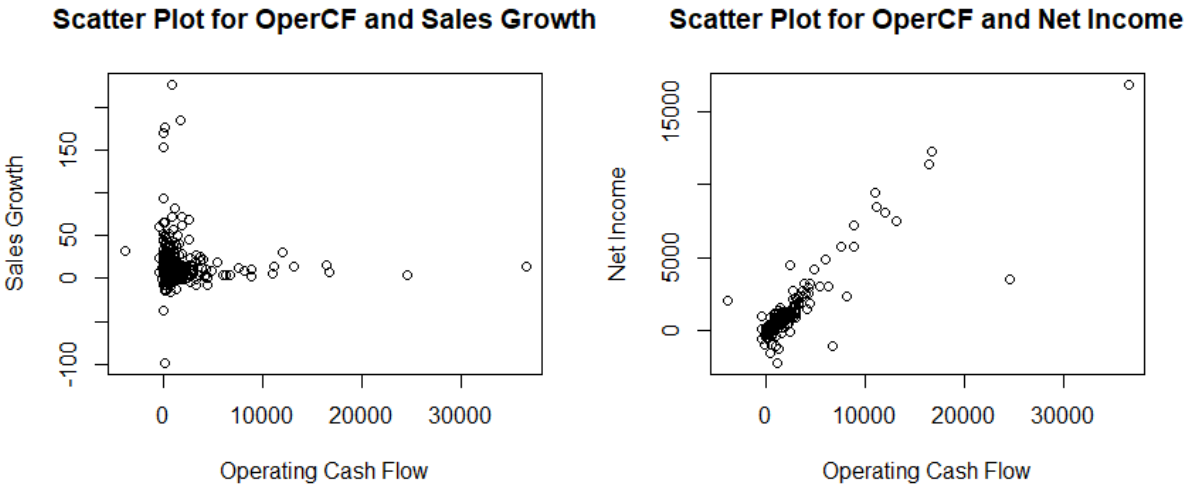
Here are the brief summary of our cleaned dataset: $\frac{(x-\mu)}{\sigma}$, where μ is the sample mean and σ is the sample standard deviation.

Box-plots to show the distribution of the cleaned data.



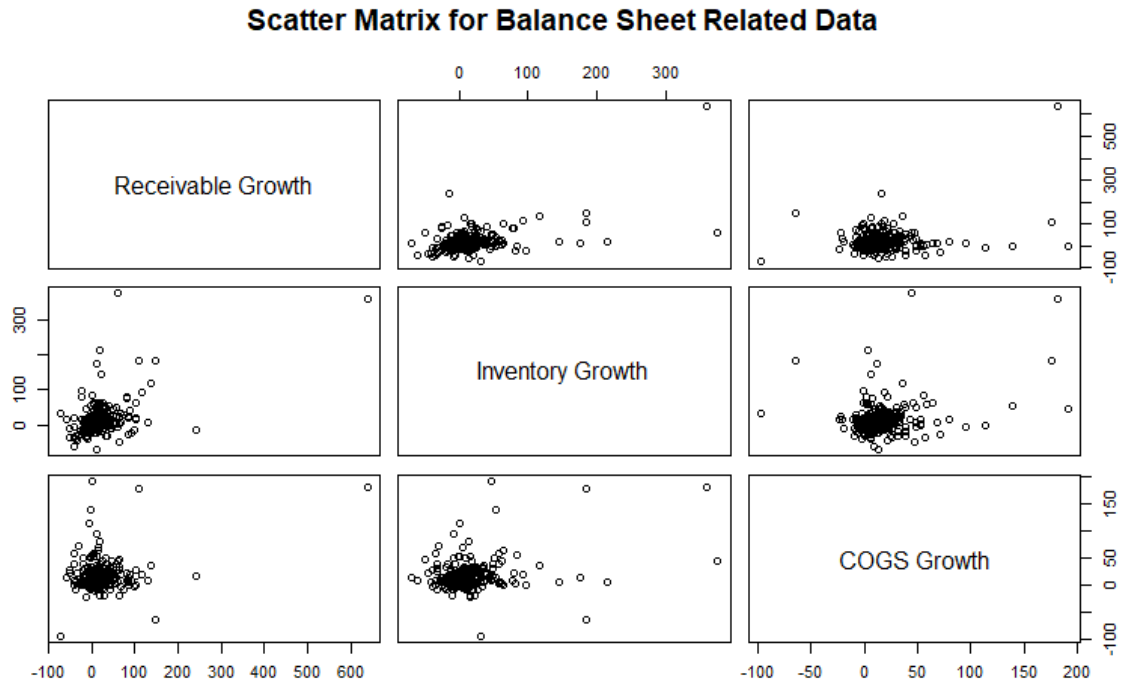
* The normalization of data takes place before the NA lines are deleted, so the median of the box plot for normalized data are not necessarily close to 1.

Cash flow from operation is the only cash flow data, below are the relationship of the cash flow to the two income statement data, Sales Growth and Net Income.



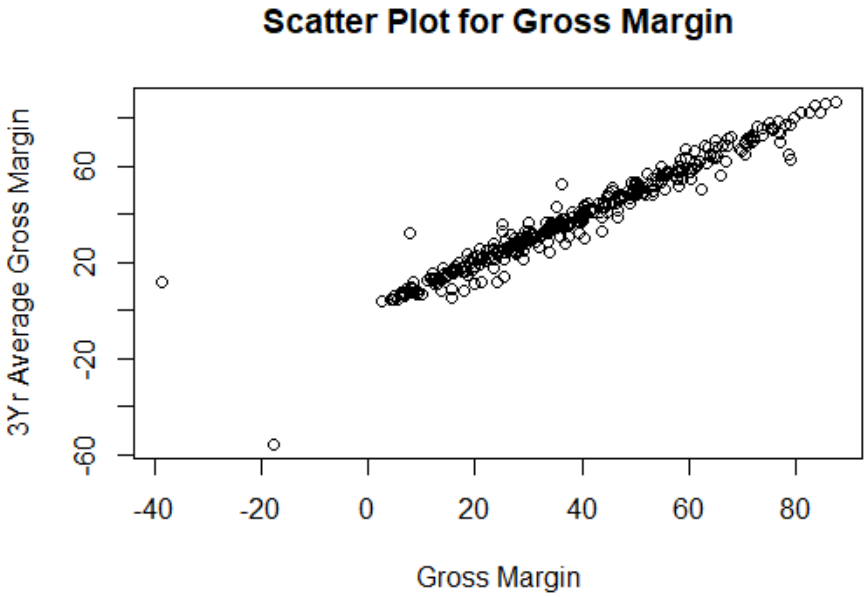
- Operating Cash Flow and Net Income has a linear relationsihp.
- There are more outlying figures in Operating Cash Flow with Sales Growth, no linear relationship is observed.

Next, we would like to discuss on the balance sheet data, Accounting Receivable Growth, Inventory Growth and Cost of Goods Sold Growth. These data are more related to the daily operation of the company which determines the revenue and expense.



- Data are dispersed within a reasonable range (-100% to 100%), with serveral outliers there.
- There are no linear relationships in the set of data
-

Gross margin and historical gross margin are collected for the model. They track the companies' profitability with respect to all the income and expenses from different sources.



- There is a linear relationship between these two gross margin data.

Data Mining

Dataset will be divided in to two types: Training Dataset: 300 (~80%) and Testing Dataset: 76 (~ 20%).

6. Classification with Artificial Neural Network

Artificial Neural Network (ANN) is first applied on the dataset. If we can get a good result (low error rate) from ANN, then it shows that there should be a pattern underlying to identify fraud companies with the normal companies. Our test parameter are set as below:

Hidden Layer	1 to 10
Number of trial for each layer	500
Maximum Iteration	2000
Reserved Testing Dataset Percentage	20%

We run through the ANN with 1 to 10 number of hidden layers, trying for 500 times to ensure that we can meet the global minimum point in either one of the trials. To choose the more appropriate size for the hidden layer, the error term and error rate of the training dataset and testing dataset are compared respectively.

The two types of error are calculated through this method.

1. Training Error: $E(w_{ij}) = [Y - V(w_{ij})]'[Y - V(w_{ij})]$

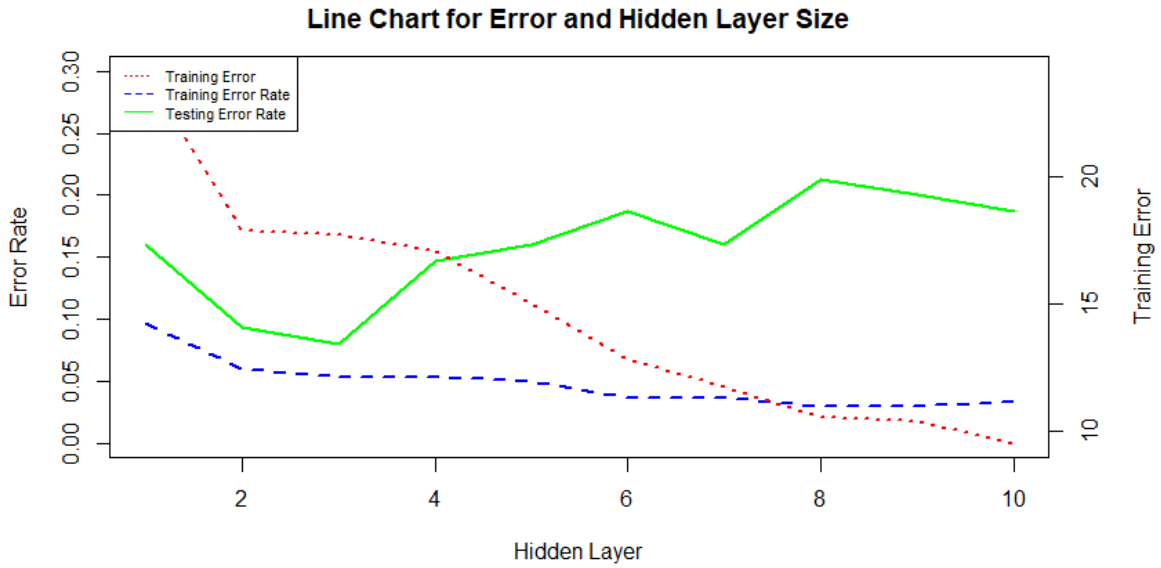
2. Training /Testing Error rate: $\frac{FP+FN}{TP+TN+FP+FN}$

The table below describes the error rate calculated from the training dataset and the testing dataset.

Maximum Iteration: 2000
Number of Trials: 500

Hidden layer	Training Error	Training Error rate	Testing Error rate
1	24.1550	0.1000	0.1600
2	17.9034	0.0633	0.0933
3	**17.7347**	**0.0667**	**0.0800**
4	14.0847	0.0600	0.1467
5	14.9859	0.0600	0.1600
6	12.8137	0.0633	0.1867
7	11.7543	0.0467	0.1600
8	10.5116	0.0400	0.2133
9	10.3465	0.0367	0.2000
10	9.5079	0.0400	0.1867

A line chart is plotted to show the movement of the error rate.



Training error keeps decreasing for the increasing hidden layer size, which means that a higher complexity structure can fit the data more easily.
However, due to overfitting, up to a certain level, we measure the error rate of the testing dataset increases when hidden layer also increases. Hence if the training error is small enough, we look for the layer which has lowest testing error, before the testing error rate

rises again due to overfitting, From our findings above, we pick the neural network with hidden layer 3.

A 8-3-1 network with 31 weights.

Classification table:

Training Dataset	Non Fraudulent	Fraudulent
ANN Non Fraudulent	223	14
Fraudulent	6	59

Testing Dataset	Non Fraudulent	Fraudulent
ANN Non Fraudulent	57	3
Fraudulent	3	13

Precision: $\frac{TP}{TP+FP} = \frac{13}{13+3} = 0.8125$

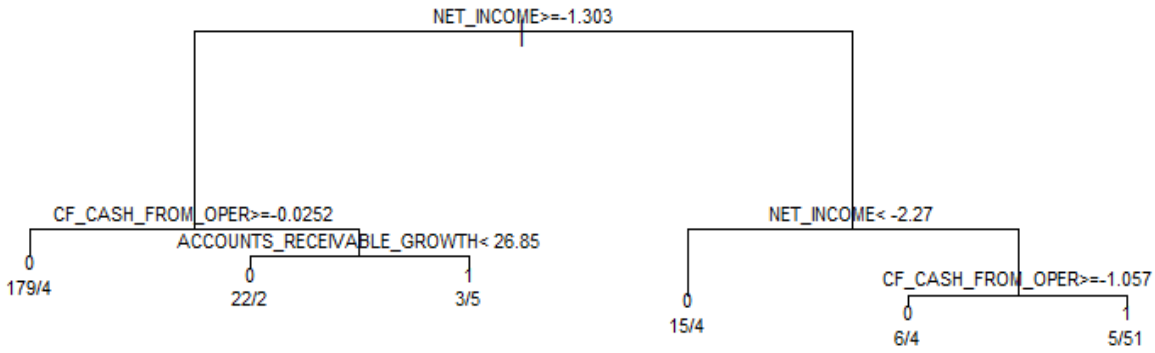
Recall: $\frac{TP}{TP+FN} = \frac{13}{13+3} = 0.8125$

F1: $2 \div (\frac{1}{Precision} + \frac{1}{Recall}) = 81.25\%$

The classification error is satisfied. It agrees with our hypothesis that data from income statement and balance sheet can have good predictive power for fraudulent companies. However, as artificial neural network lacks the explanatory power to give an explicit rules to readers, we will employ Classification Tree to tell a better story for fraudulent companies

7. Interpretation with Classification Trees

Followed by ANN, Classification Tree is applied to construct rules to determine fraudulent companies. For our test parameters, our target variable is a categorical variable, so method="class" is used; maxdepth=3 is set to limit the classification tree to a maximum of 3 layers to avoid overly complex rules.



The resulting classification tree shows three classification rules:
R1: *If* (NET_INCOME >= -1.303) and (CF_CASH_FROM_OPER >= -0.0252) *then* Result = 0 (179/4)
R2: *If* (NET_INCOME >= -1.303) and (CF_CASH_FROM_OPER < -0.0252) and (ACCOUNTS_RECEIVABLE_GROWTH < 26.85) *then* Result = 0 (22/2)
R3: *If* (NET_INCOME >= -1.303) and (CF_CASH_FROM_OPER < -0.0252) and (ACCOUNTS_RECEIVABLE_GROWTH >= 26.85) *then* Result = 1 (3/5)
R4: *If* (NET_INCOME < -2.27) *then* Result = 0 (15/4)
R5: *If* (-1.303 > NET_INCOME >= -2.27) and (CF_CASH_FROM_OPER >= -1.057) *then* Result = 0 (6/4)
R6: *If* (-1.303 > NET_INCOME >= -2.27) and (CF_CASH_FROM_OPER < -1.057) *then* Result = 1 (5/51)

The results imply that whether a company is fraudulent or not depends on its net income, operating cash flow and account receivable growth, with net income and operating cash flow being the major factors.

The table below shows the support, confidence and capture for each rule.

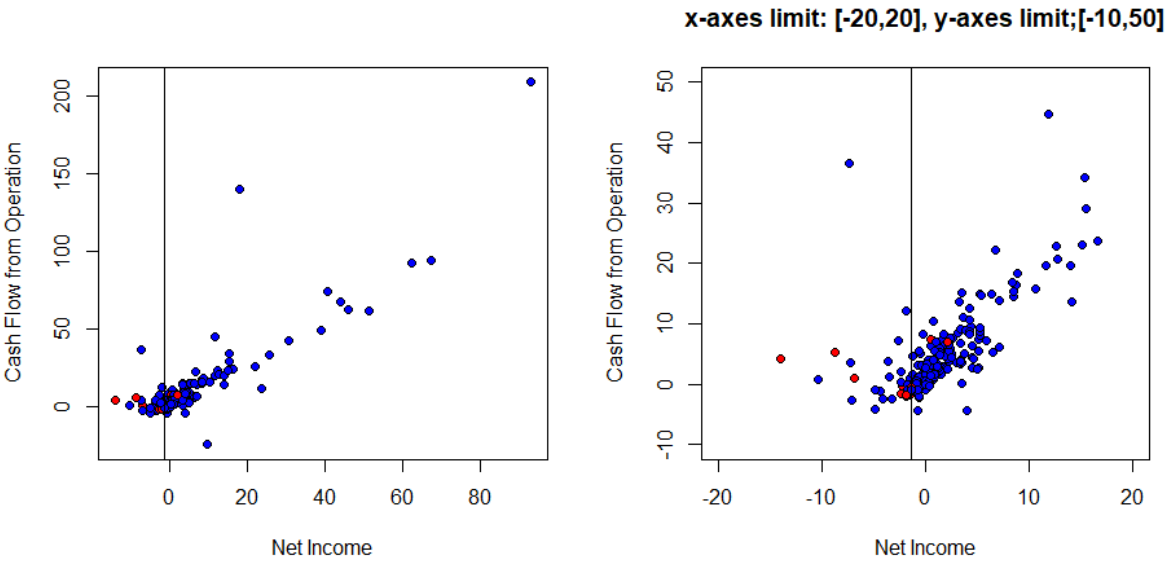
Rule	Support	Confidence	Capture
1	0.6100	0.9781	0.7783
2	0.0800	0.9167	0.0957
3	0.0267	0.6250	0.0714
4	0.0633	0.7895	0.0652
5	0.0333	0.6000	0.0261
6	0.1867	0.9107	0.7286

Classification table:

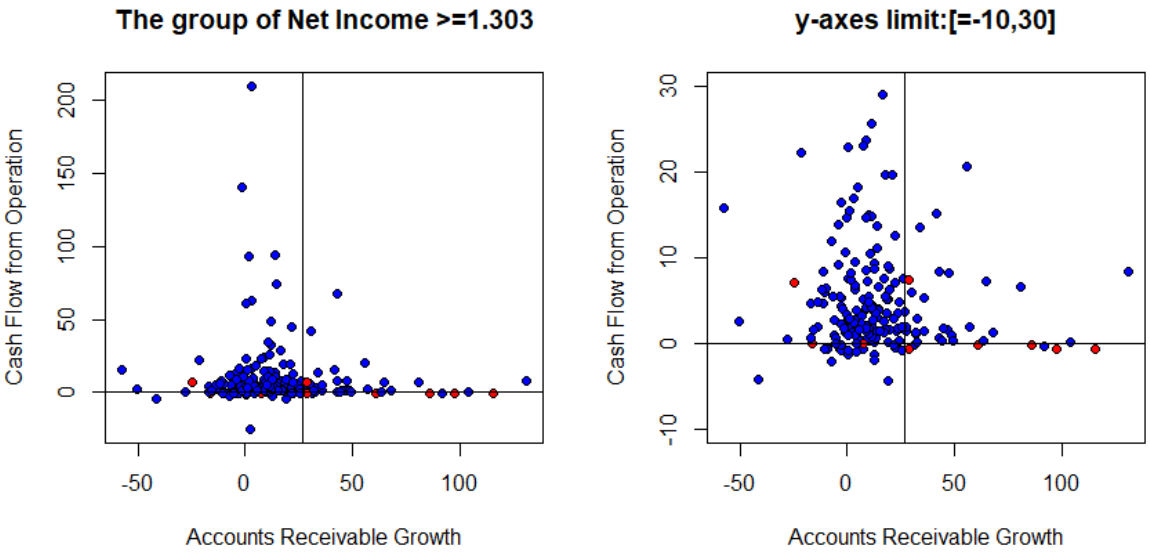
Training Dataset	Non Fraudulent	Fraudulent
CTREE Non Fraudulent	222	14
Fraudulent	8	56

Testing Dataset	Non Fraudulent	Fraudulent
CTREE Non Fraudulent	58	6
Fraudulent	1	11

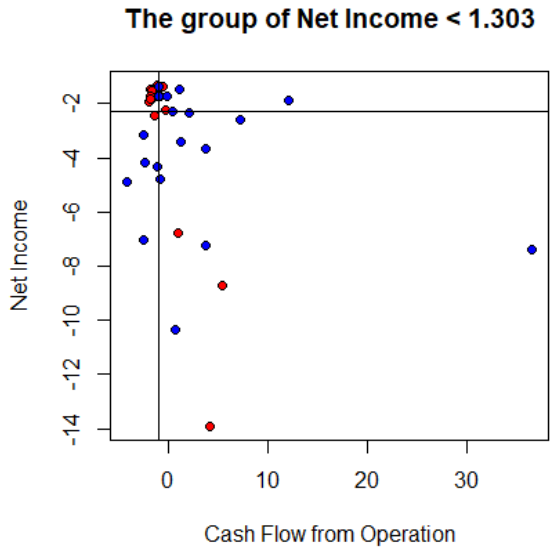
Scatter plot is provided for a better visualization of the data.
Node 1 (First Splitting Rule)

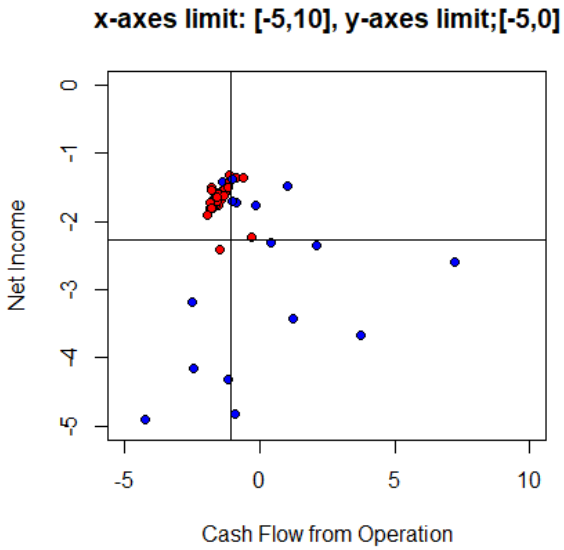


The YES group at node 1 (Net Income >=-1.303),



The NO group at node 1 (Net Income <-1.303),





From the classification table, we further tabulate the error rate, precision, recall and F1 score for training and testing dataset.

Dataset	Error Rate	Precision	Recall	F1 Score
Training	0.0733	0.8750	0.8000	0.8358
Testing	0.0921	0.9167	0.6471	0.7586

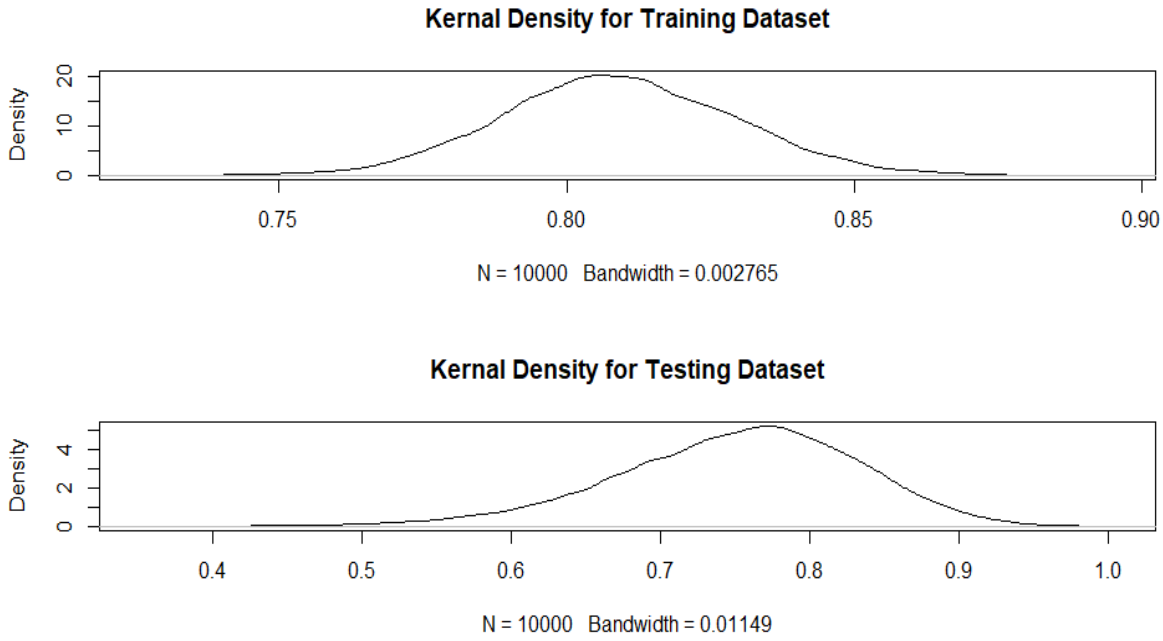
From the table, both training and testing error rate is satisfactory and comparable to that of ANN results. Testing dataset has a higher precision but higher error rate and lower recall than training dataset, thus a lower F1 score.

We run CTREE 10000 times without specifying a seed, and record all the error rate, precision, recall and F1 Score for both datasets. Below is the summary of the results.

Training	Error_Rate	Precision	Recall	F1_Score
Min	0.0500	0.7564	0.6029	0.7339
Max	0.1033	1.0000	0.9265	0.8873

Testing	Error_Rate	Precision	Recall	F1_Score
Min	0.0132	0.3846	0.2500	0.3846
Max	0.2632	1.0000	1.0000	0.9714

It is observed that the results for testing dataset varies greatly, compared to that of training dataset. Therefore, we construct the kernel density plot for both datasets.



It is shown that the testing dataset has a much longer left tail, but both datasets has a F1 Score of around 0.8, with training dataset having a peak of higher F1 Score, which is consistent with the general rule that training is explained better than testing dataset. Therefore, the difference and variability in training and testing results is mainly due to testing dataset being too small in absolute size (but not proportion). This could be improved if we could attain more data.

Both training and testing dataset has precision higher than recall, suggesting that some of the positive values are not predicted, but almost every positive value predicted are true. To balance precision and recall, the probability threshold for predicting fraudulent cases could be lowered. This is not implemented to avoid further complications, as the recall value is still within acceptable range.

8. Results and Interpretation

ANN can give a good result in classification, which signifies that the dataset must have contained some underlying patterns that can help us distinguish between fraudulent and non-fraudulent companies.

From the method of classification tree, we classify those fraudulent with Operating Cash Flow and Net Income. These also somehow agrees with what Schilit(2002) mentioned in his book - there is signal when Operating Cash Flow lags behind Net Income. In our supervised learning, we do not directly test the above warning signal, but instead, we can conclude that Operating Cash Flow and Net Income are very significant factor that helps us determine the fraudulent companies.

Also, on the other side of the tree, we can see that high Net Income companies are mostly non-fraudulent. In this category, most companies are regarded as profitable companies and they normally do not have to take the risk of furnishing their reports with fraudulent acts, but we see that this is not always the case. Interestingly, in this category, we can see that the balance sheet data, Accounting Receivable Growth, plays an important role here to distinguish companies that are fraudulent or not. Schilit(2002) pointed that if accounting receivables grow faster than sales, it is more likely that the company is fraudulent, as company wants to defer the money they receive so they can post it up in a later period to enjoy a larger surprise when the market is bullish. And from our tree, using Rule3, we can see that if a company has a high receivable growth, it is more likely to commit accounting fraud, which echoes with the point of Schilit.

In addition, the rules for fraudulent companies are companies that has both low Operating Cash Flow and Net Income. We would interpret this as company that are less profitable. This result make sense that only less profitable companies are willing to take the risk of committing accounting fraud so as to bet for a larger benefit for its own company. If companies are already in a profitable status, it is not worthwhile to take the risk of getting involved in any legal liability for a more good-looking financial report.

9. Limitation

Accounting Fraud Detection can go at a more holistic approach. The approach we used in this project is just from the perspective of financial data. However, from a financial report, apart from the data, there are approach that are potentially possible for fraud detection. Schilit(2002) pointed that footnotes, messages from the board, change of accounting policy and even change of auditor can also give us clue on accounting fraud. In the case of analyzing words in accounting statement, Neuro-Linguistic Programming (NLP) can be used to read relevant information. For this report, we are only focusing on the quantitative side of financial statement, but a **qualitative** side can also give a useful insight.

Another approach would be to get a much larger size of dataset and apply Benford's Law to look at the frequency distribution of leading digits in the financial statement. We can find out the anomaly and further verify if the company has committed any fraud.

10. Conclusion

To conclude, analysis of data from accounting report can give a good classification for fraudulent companies. The book, Financial Shenanigans, written by Mr. Howard Schilit, though it was written in 2002, makes sense to the readers that we can focus on certain financial data to detect accounting fraud. By using Artificial Neural Network, we can train a good model for accounting data, and with the use of Classification Tree, we find out that Operating Cash Flow and Net Income are the major predictable variables that help us identify accounting fraud.

Appendix:

Fraudulent company list

GAO Financial Restatement Database (GAO-06-1079sp)
https://www.gao.gov/special.pubs/gao-06-1079sp/

Detailed description for financial data

ID	Mnemonic	Calculation
CF015	Cash Flow From Operation	Net Income + Depreciation & Amortization + Other Non-cash Adjustments + Changes in Non-cash Working Capital
IS050	Net Income	N/A
RR033	Sales Growth	(Revenue from Current Period - Revenue from Same Period Prior Year) * 100 / Revenue from Same Period Prior Year
RR891	Account Receivable Growth	[(current year Accounts & Notes Receivable / prior year Accounts & Notes Receivable) - 1] *100
RR892	Inventory Growth	[(Current year Investories / Prior Year Inventories) -1] *100
RR296	Cost of Goods Sold Change	[(Current Year Cost of Goods Sold / Prior Year Cost of Goods Sold) - 1]*100
RX549	3 Years Average Gross Margin	(3-Year Average of Net Sales - Cost of Goods Sold) / Net sales *100
RR057	Gross Margin	(Net Sales - Cost of Goods Sold) * 100 / Net Sales

Reference:

Schilit, H. M. (2002). *Financial shenanigans*. New York, N.Y: McGraw-Hill.

Jonathan M. Karpoff, Allison Koester, D. Scott Lee, and Gerald S. Martin (2017) Proxies and Databases in Financial Misconduct Research. The Accounting Review: November 2017, Vol. 92, No. 6, pp. 129-163. Retrieved from https://ssrn.com/abstract=2917524

Sharma, A., & Panigrahi, P.K. (2012). A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. CoRR, abs/1309.3944.