Project Mid-Semester Report

Jason Choi, Kyle Cummings

March 11th, 2022

1 Executive Summary

1.1 Decisions to be Impacted

We train a neural network (NN) model to predict fraud in publicly traded U.S. firms using publicly available financial statement data. Building, tuning, and refining such a predictive model is of great importance to U.S. regulators and auditors, who must carefully budget limited resources, as well as investors, who may hedge their exposure to or speculatively profit, via short sales, from suspect firms.

1.2 Business Value

Accounting fraud is undeniably a worldwide problem. Such fraud inflicts widespread harm across society and therefore has been a rich topic of interest for regulators and researchers alike. Indeed, not only do investors in the fraudulent company's financial securities suffer an immediate, acute loss, but even more insidious is the chronic widening of risk premiums borne by all market participants. By tuning and validating a predictive model with cotemporaneous data, we seek to increase their predictive accuracy out of sample and thereby increase the perceived costs and risks to managers contemplating fraud. Hence, our desired future-state is a decrease in occurrences of fraud and a commensurate diminished risk premium such fraud entails.

1.3 Data Assets

Our dependent variable data – the Security and Exchange Commission's Accounting and Auditing Enforcement Releases (AAERs) – is collected from the University of Southern California Marshall School of Business and currently covers the years 1991-2014. Our explanatory variables consists of publicly available financial statement data from S&P Global Market Intelligence ("COMPUSTAT") as available through the Wharton Research Data Services (WRDS). Extensive prior literature has identified those financial statement items and ratios that have the greatest explanatory power for detecting accounting fraud (e.g., Cecchini et al. [2010] [1] and Dechow et al. [2011] [2]), and we follow this literature in identifying our data matrix.

2 Data Preprocessing

2.1 Data Description

Our data is comprised of all publicly listed U.S. firms from 1991 - 2008. In accordance with prior literature on the topic, we omit the period after 2008 as the regulatory environment changed quite drastically. Resources at the FBI shifted resources towards detecting Pomzi schemes in the wake of the Madoff fraud [4]. Furthermore, the Department of Justice spread financial fraud investigations among numerous U.S. Attorney's Offices which had little experience investigating sophisticated frauds. The effect of these changes are evident in the raw number of financial fraud AAERs issued by the SEC; after accounting for an appropriate multi-year lag, the number of financial fraud cases in 2014 drops to a mere four instances. We include summary statistics for our X matrix and dependent variable in the figures below. We also highlight the distributions of two of our 28 independent variables: (1) total assets, and (2) accounts payable. Both variables exhibit a significant rightward skew, indicative of several 'giant' firms with orders-of-magnitude greater characteristics.

Year	Instances of Accounting Fraud
1990	15
1991	27
1992	26
1993	30
1994	23
1995	22
1996	33
1997	42
1998	56
1999	73
2000	86
2001	81
2002	77
2003	69
2004	58
2005	45
2006	33
2007	30
2008	26
2009	31
2010	26
2011	21
2012	19
2013	11
2014	4

Figure 1: Fraud Detections by Year

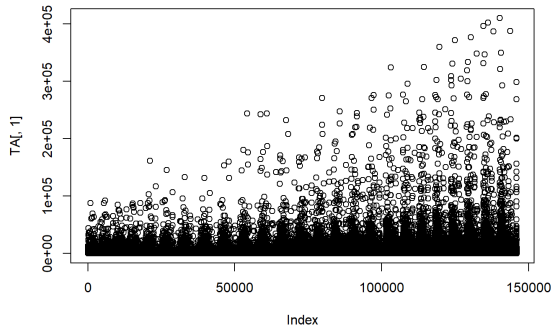

Figure 2: Scatter Plot for Total Assets

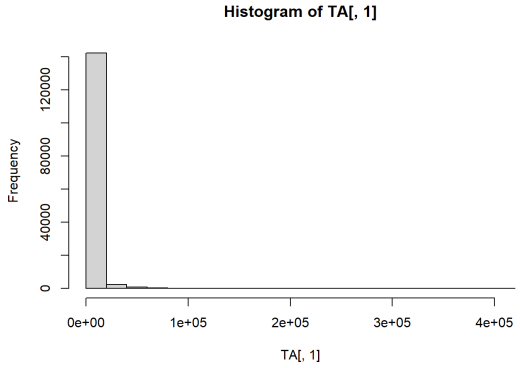

Figure 3: Histogram for Total Assets

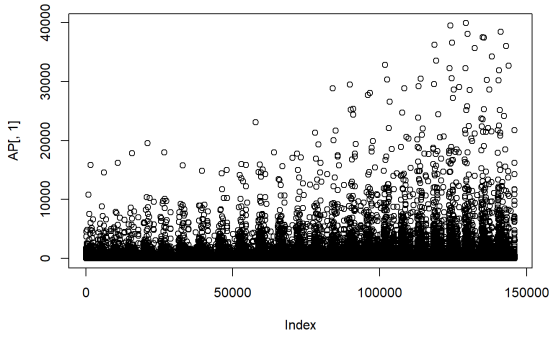

Figure 4: Scatter Plot for Accounts Payable

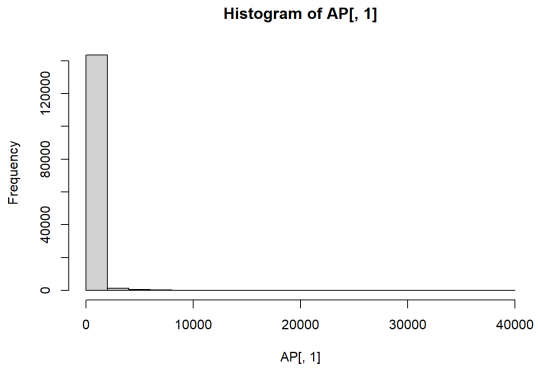

Figure 5: Histogram for Accounts Payable

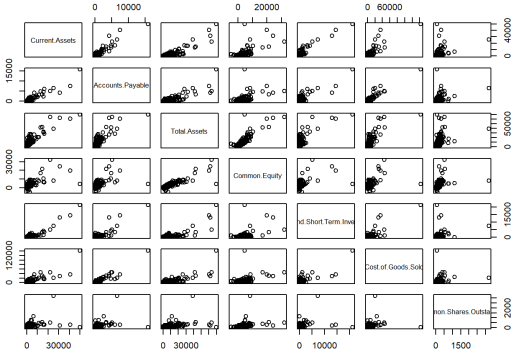

Figure 6: Pairs Charts for Variables 1-7

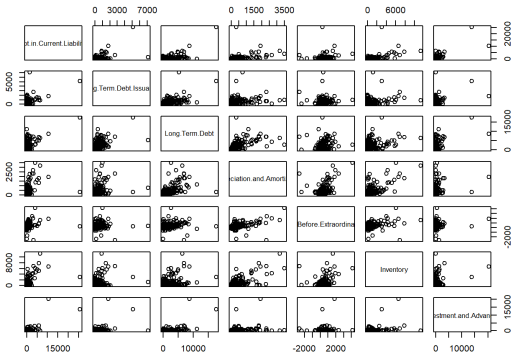

Figure 7: Pairs Charts for Variables 8-14

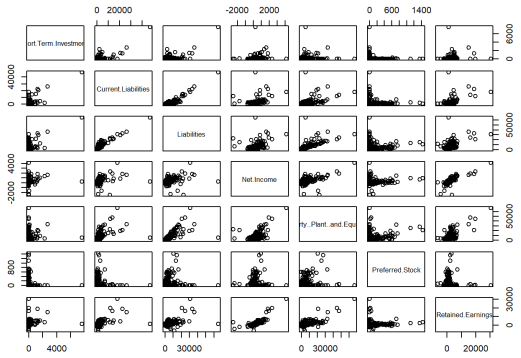

Figure 8: Pairs Charts for Variables 15-21

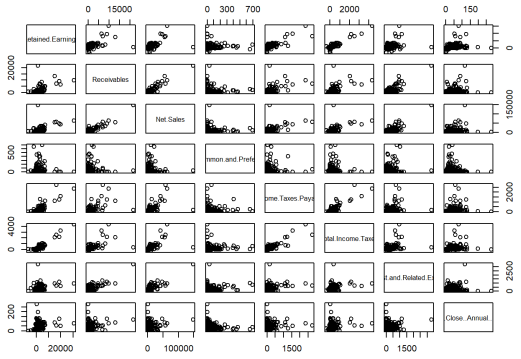

Figure 9: Pairs Charts for Variables 22-28

	Current Assets	Accounts Payable	Total Assets	Common Equity	Cash and Short-Term Investments
Min	0.00	0.00	0.00	-2624.00	0.00
25%	10.74	1.71	23.28	8.02	0.79
Median	46.25	7.59	109.09	36.77	4.96
Mean	413.54	116.38	1357.85	499.44	84.61
75%	191.88	43.53	489.47	211.15	26.39
Max	17776.00	4566.00	43775.00	18501.00	3791.95
	Cost of Goods Sold	Common Shares Outstandin	bt in Current Liabiliti	Long-Term Debt Issuance	Long-Term Debt
Min	0.00	0.00	-0.07	0.00	0.00
25%	13.99	3.88	0.50	0.00	0.97
Median	87.32	8.24	3.30	0.97	16.19
Mean	827.86	31.99	85.46	66.27	320.95
75%	410.84	24.90	25.37	22.04	108.49
Max	32113.00	1092.14	4804.00	2264.60	9118.00
	Depreciation and Amortiza	me Before Extraordinary It	Inventory	Investment and Advances	Short-Term Investments
Min	0.00	-793.46	0.00	0.00	0.00
25%	0.77	-0.71	1.90	0.00	0.00
Median	4.68	1.83	9.92	0.00	0.00
Mean	77.56	61.27	117.62	34.97	27.51
75%	21.63	19.84	61.20	1.55	0.09
Max	3440.00	3640.00	3332.00	3358.50	1929.00
	Current Liabilities	Liabilities	Net Income	roperty, Plant, and Equipmen	Preferred Stock
Min	0.00	0.00	-793.46	0.00	0.00
25%	5.02	8.76	-0.79	10.85	0.00
Median	28.66	57.95	1.85	56.76	0.00
Mean	329.90	828.17	62.15	1278.76	17.29
75%	115.09	277.08	18.52	307.10	0.00
Max	15089.00	29682.00	3640.00	44075.00	1578.40
	Retained Earnings	Receivables	Net Sales	of Common and Preferred S	Income Taxes Payable
Min	-2837.00	0.00	-1.63	0.00	-4.88
25%	-2.34	3.21	21.90	0.00	0.00
Median	12.27	17.23	131.25	0.05	0.12
Mean	314.31	182.14	1228.87	10.20	22.46
75%	115.18	86.55	589.35	1.24	2.68
Max	14082.00	12290.00	55977.00	670.00	1762.00
	Total Income Taxes	nterest and Related Expens	e Close, Annual, Fiscal		
Min	-97.00	0.00	0.02		
25%	0.00	0.34	2.12		
Median	1.03	2.97	7.56		
Mean	37.34	39.26	14.94		
75%	12.17	13.84	19.88		
Max	1741.00	1041.00	459.13		

2.2 Data Cleaning and Outlier Detection

Conveniently, our data is sourced from commercial and academic databases that have already undergone a relatively rigourous QA/QC process. Moreover, since our predictive model process is inherently a process of anomaly detection, it is inappropriate to apply standard outlier detection methods, as such outliers will undoubtedly be highly correlated with our variable of interest.

3 Model updates

3.1 Model

We aim train a feed-forward neural network (NN) model to predict fraud. Such neural networks are well-suited to supervised learning tasks. [3] Specifically, our network consist of an input layer, one or more hidden layers, and an output layer. Currently, we plan on using the rectified linear activation function (ReLU) for our hidden layers, which is a piecewise linear function that outputs positive inputs directly and otherwise converts negative outputs to 0. The output layer will consist of a single node which uses the sigmoid function, as given by

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

As a baseline, we intend on using a fairly parsimonious structure of two hidden layers consisting of 32 neurons each. As our model is a binary classifier, we use the binary cross-entropy as our loss function, as

given by

$$H_p(q) = \frac{1}{N}\sum_{i=1}^{N} y_i \log(\mathrm{p}(y_i)) + (1-y_i)\log(1-\mathrm{p}(y_i))$$

for labels in $\boldsymbol{y}$.

While the above specifications are the baseline model, we also intend to tune the hyperparamters of the network– such as the number of hidden layers, number of neurons, number of epochs, and activation function – via cross-validation. Also, we may additionally employ regularization techniques, such as random dropout of nodes, during our validation process. Per standard best-practices, we intend to employ 'early stopping' to avoid over-fitting our model to the training data.

3.2 Machine Learning Workflow as Machine Learning Morphism (MLM)

Our machine learning workflow as a machine learning morphism is comprised of the following morphisms:

- $\mathcal{ML}_0$: Standardize and center the original data matrix in $\mathcal{R}^{28}$
- $\mathcal{ML}_1$: Feed Forward Neural Network
  - An input space $\mathbb{X} \in \mathcal{R}^{28}$ of 28 standardized financial statement items,
  - an output space $\mathbb{Y} = [0,1]$ denoting the probability of a firm exhibiting fraud in a given year,
  - parameters consisting of weights $p^* \in \mathcal{R}^w$ assigned to connections between layers, optimized via backwards propagation
  - a Morphism $F = F_k(F_{k-1}(F_{k-2}...F_1(\boldsymbol{x}))))$ for the k layers in the NN,
  - binary cross-entropy as an empirical risk function $H_p(q) = \frac{1}{N}\sum_{i=1}^{N} y_i \log(\mathrm{p}(y_i)) + (1-y_i)\log(1-\mathrm{p}(y_i))$.
- $\mathcal{ML}_2$: Classification threshold
  - Input space: $\mathbb{X} \in [0,1]$
  - Output space: $\mathbb{Y} \in \{0,1\}$
  - Morphism:

$$\begin{cases} y = 1 & \text{if} \quad x \geq T \\ y = 0 & \text{else} \end{cases}$$

  - Parameter Prior: prior information of threshold $T \in [0,1]$
  - Risk Function: Optimal threshold based on ROC curve over cross validation sets.

Then, the full workflow is:

$$\mathcal{M} : \mathbb{X} \longrightarrow \{0,1\} = \mathcal{ML}_2 \circ \ \mathcal{ML}_1 \circ \mathcal{ML}_0$$

4 Next Steps

- Build our neural network and begin cross-validation to tune hyperparameters.
- Collect financial accounting data for publically traded U.S. firms from COMPUSTAT for recent years 2020, 2021 in order to make contemporary fraud predictions.
- Possibly hand-collect data from the SEC's website to cover the AAER years 2015-2021. Further research is required to ascertain if the regulatory environment has sufficiently 'recovered' i.e. resourced relative to the post-2008 dropoff to make this effort worthwhile.
- Highlight the model's performance for high-profile fraud cases, such as Enron and Worldcom.

References

[1] Mark Cecchini et al. "Detecting Management Fraud in Public Companies". In: *Management Science* 56.7 (2010), pp. 1146–1160.

[2] Patricia Dechow et al. "Predicting Material Accounting Misstatements". In: *Contemporary Accounting Research* 28 (Apr. 2010).

[3] "Detection of financial statement fraud and feature selection using data mining techniques". In: *Decision Support Systems* 50.2 (2011), pp. 491–500. ISSN: 0167-9236.

[4] Yang Bao et al. "Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach". In: *Journal of Accounting Research* 58.1 (2020), pp. 199–235.