Using Machine Learning to Detect Accounting Fraud: Mid-semester Presentation

Kyle Cummings and Jason Choi

March 21st, 2022

Introduction

- Managers and owners of companies have conflicts of interest
- Fraud: intentional misrepresentation of the financial statement information
  - Overstate assets
  - Understate liabilities
- E.g. Enron (2001)
- Auditors are tasked with QA/QC of financial statement information
- Regulators (i.e., SEC and DOJ) enforce Generally Accepted Accounting Principles (GAAP)

Executive Summary

- We train neural network (NN) model to predict fraud in publicly traded U.S. firms using publicly available financial statement data.
- Auditors and regulators have limited resources.
- Knowing which firms are ‘suspect’ can help guide auditors and regulators target resources more efficiently.
- Investors can steer clear of suspect firms, or speculatively short them.
- X ~ Publicly available financial statement data.
- Y~ Probability a firm will commit fraud in a given year.

Data Asset Description

- X: publicly available financial statement data from S&P Global Market Intelligence ("COMPUSTAT")
- Y: SEC's Accounting and Auditing Enforcement Releases (AAERs)
- Covers the years 1991-2014
  - Research indicates the regulatory environment drastically changed in 2009
  - Resources were shifted away from accounting fraud

Data Asset Description (Cont.)

- X:
  - 28 financial statement items
  - 18444 unique firms
  - 146045 total observations
- Y:
  - 964 instances of fraud

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

	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 Outstanding	Debt in Current Liabilities	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 Amortization	Income Before Extraordinary Items	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	Property, Plant, and Equipment	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	Sale of Common and Preferred Stock	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	Interest and Related Expense	Price 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		

Data Preprocessing

- As has been touched on, with the change in regulatory environment we will stop looking at data after 2008.
- With that adjustment, we see 112981 observations with 852 instances of fraud

Data Preprocessing - Outlier Detection

- A model such as ours is inherently anomaly detection–through investigation we've found that outlier detection methods may be inappropriate.
- In our data set of 112981 observations, there were 852 instances of fraud.
  - Occurrence rate of .754%
  - Again worth pointing out the majority of fraud goes undetected

Data Preprocessing - Outlier Detection

- We used several outlier detection methodologies on the dataset (K-Nearest Neighbor, Kernel Density, Generalised Dispersion)
  - Attained a set of 2036 common outliers. 62 of which were instances of fraud
  - This would mean removing over 7% of our data points with fraud positively detected.

Data Preprocessing - Outlier Detection

- Also important to point out:
  - Fraud Occurrence Rate in Outlier Set:

    62/2036 = 3.045%
  - Fraud Occurrence Rate in Whole Set:

    852/112981 = .754%
- Represents over a 300% increase

Our Model

- Feed-forward neural network (NN)
- Two hidden layers consisting of 32 neurons each Use rectified linear activation function (ReLU) for our hidden layers
- Output layer is a single node using the sigmoid function:
- Loss func $\sigma(x) = \frac{1}{1+\exp(-x)}$ nary cross-entropy:

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

Next Steps

- Build our neural network and begin cross validation and tuning of model parameters

- Pull raw accounting data from compustat for firms in 2020 and 2021. Using this data we'll additionally test our model on contemporaneous data.

Next Steps (Continued)

- Potentially hand collect SEC AAER years 2015-2021. With the desire to see if the regulatory environment has ‘recovered’.

- Highlight model performance on high-profile fraudulent firms (Enron, Worldcom, etc.)

Questions?