Journal of Forensic & Investigative Accounting
Volume 9: Issue 2, July–December, 2017

Detecting Fraud Using Validated and Specifically-identified Metrics

Natalie Tatiana Churyk

Danny Lanier, Jr.[*]

Introduction

Accounting researchers have long sought stronger models and tools to detect fraudulent activity. Given the continuing incidence of reported fraud and resultant increased Securities and Exchange Commission (SEC) enforcement filings, the need is great (Eaglesham and Rapoport 2015). Unfortunately, while many such models have achieved significant accuracy, they generally fail to provide adequate notice to avoid resultant catastrophic financial effects. Thus, this article: 1) examines the fraud prediction literature for model variables and to create variables when needed (e.g., Smith 2013); 2) determines the association of the metrics with fraud vs. non- fraud firms; and 3) develops a fraud detection model using logistic regression.

The three most promising fraud detection models include: 1) Lee et al.,'s (2013 a, b) validated linguistics-based fraud detection model; 2) a contemporaneous risk factors model based on *Statement on Auditing Standards (SAS) No. 99* (Skousen and Wright 2008); and 3) Dechow, Ge, Larson, and Sloan's (2011)[1] comprehensive financial, non-financial, and off balance sheet data model. The first two could be early fraud detection models, while the last model depends upon published financial statements.

Lee et al., (2013a, b) validated an early fraud detection model that they designed in Churyk et al., (2008, 2009), which used content analysis to analyze the Management's Discussion and Analysis (MDA) sections of annual reports to identify potential deception indicators. These works provided more timely likelihood of fraud detection than did current quantitative models. By examining narratives (asynchronous communication) from firms with SEC-mandated company restated annual reports, they found key fraud detection, linguistic characteristics. Lee et al., (2013 a) successfully created a model that a holdout sample validated to achieve an overall early fraud prediction accuracy of about fifty-nine percent.

Skousen and Wright's (2008) early fraud detection model identified firm-related factors for financial statement fraud. Their empirical SAS No. 99 factors focused on pressure, opportunity, and rationalization—popularly called the fraud triangle. Results reportedly classified fraud and non-fraud firms accurately about 69.8 percent of the time. But this classification uses only one sample and a cross-validation procedure, rather than a holdout sample.

Dechow et al., (2011) based their "material accounting misstatement" prediction model upon twenty-eight total variables categorized into five areas (accruals quality, financial performance, nonfinancial measures, off-balance sheet activities, and market-based measures). Depending on their included variables, comparing SEC-required firms to restate financial statements to those not required to restate financial statements, they accurately classified a 61.7 percent holdout sample, whose accuracy increased to 65.3 percent when adding industry variables to the model.

Smith (2013) identified several accounts and footnotes manipulated to deceive investors, users, auditors, and the SEC into believing that HealthSouth was a profitable going concern. We call Smith's (2013) metrics "specifically-identified metrics." The areas of manipulation relate to accounts receivable, goodwill including impairment, mergers and acquisitions, cash flows, related parties, and one-time charges.

Acknowledgement**:** We would like to thank 2014 and 2016 Midwest Region AAA participants and Alan Reinstein of Wayne State University for comments on earlier versions of this manuscript.

[1] Dechow et al., (2011) use the term 'material accounting misstatements' versus fraud, whereas we use the term 'fraud.'

The current paper builds on the aforementioned papers' statistical model building techniques. Using specifically-identified metrics from Smith (2103), in conjunction with validated hypothesized metrics from Lee et al., (2013 a) and Dechow et al., (2011), should produce a more accurate fraud detection model than using any one set of metrics alone.

Our research contributes to the fraud literature by providing a model that provides greater fraud prediction/accuracy than prior published fraud prediction models. The rest of the article is organized as follows: we briefly review the literature used to support and develop our research questions, then present our methodology and analysis, followed by presenting our results, and then conclude.

Literature and Research Question Development

We primarily seek to determine if specifically identified metrics and validated hypothesized metrics will provide managers, accounting professionals, creditors, and shareholders with a tool indicating whether the information under investigation is unreliable and potentially fraudulent. Lee (2005), building on Zhou et al.,'s (2004) cues, finds deceivers' written communications contain higher levels of rhetoric, word sophistication, sentence sophistication, and self-reference. Deceivers seek to manage information through quantity manipulations (Buller and Burgoon, 1996 and Zhou et al., 2004). This manipulation leads to testing complexity and diversity. Burgoon et al., (1996) explain that, on average, deceivers tend to use depersonalization. However, Lee (2005) explains that deceivers' written communications attempt to mimic truth tellers and thus use more self-reference or non-immediacy, rather than depersonalization. If the deceivers always successfully mimic truth tellers, they would find no cues to differentiate both groups. However, since deceivers may want to sound more truthful than truth tellers because they have more at stake, they may use more words, more complex sentences, reference management more often, and describe events in more detail than truth tellers to increase their credibility.

Churyk et al., (2009) provide more detailed theoretical support for the above factors. Based on prior research Churyk et al., (2009) identified ten narrative variables that significantly differ between fraud and non-fraud companies; fraudulent firms' MDA contained: 1) more words; 2) fewer unique words; 3) fewer colons; 4) fewer semicolons; 5) fewer explanatory phrases (e.g., "for example," "such as"); 6) fewer positive emotion terms; 7) fewer optimism and energy terms; 8) greater anxiety terms; 9) fewer causation terms; and 10) fewer present tense verbs.

Extending the above referenced research, Lee et al., (2013a), used backward stepwise regression on the ten above variables to determine which factors best predicted deception. They found that four of these variables contributed positively to the stepwise model. The variables included in the model to suggest the presence of deception were: 1) fewer terms indicating positive emotion; 2) fewer present tense verbs; 3) the presence of an increased number of words; and 4) fewer colons.

Dechow et al., (2011) examine an extensive set of firm characteristics to develop a model to predict accounting misstatements, by classifying these characteristics into five categories: 1) accrual quality; 2) financial performance; 3) nonfinancial measures; 4) off-balance-sheet activities; and 5) market-based measures. They find that measures of accruals are unusually high (income-increasing) in both the years preceding and the year of misstatement, consistent with managers' optimism or overinvestment prior to engaging in more aggressive reporting tactics. The authors also find evidence of declining financial performance (e.g., return on assets) and nonfinancial performance (e.g., change in employee headcount relative to the change in total assets) among their sample of misstating firms.

Regarding off-balance-sheet activities, Dechow et al., (2011) find: 1) unusually higher use of operating leases to accelerate earnings and reduce long-term debt; and 2) higher expected returns on pension plan assets (to reduce pension expense) among misstating firms. Finally, they find unusually high market-based metrics (e.g., price-earnings and market-to-book ratios) for misstating firms, consistent with managers acting on incentives to avoid market penalties associated with earnings disappointments (Skinner and Sloan 2002).

Dechow et al., (2011) use backward logistic regression to build three additive models: 1) financial statement variables (i.e., accruals quality and financial performance); 2) Model I plus non-financial statement and off-balance-sheet variables; and 3) Model II plus market-based measures.[2] The backward regression procedure yielded a final model (Model III) that

[2] Dechow et al., (2011) use this model sequencing to reflect the relative ease of obtaining information about the firm characteristics within each category.

included these variables: 1) working capital accruals; 2) change in receivables; 3) change in inventory; 4) percentage of "soft" assets; 5) change in cash sales; 6) change in return on assets; 7) issuance of securities; 8) abnormal change in employees; 9) existence of operating leases; 10) market-adjusted stock returns; and 11) lagged market-adjusted stock returns. The Dechow et al., (2011) variables, along with all potential variables (forty-one total), are defined in Table I. [see Table I, pg 756]

As mentioned above, Smith (2013) identified several manipulated accounts and footnotes created to deceive investors, users, auditors, and the SEC into believing that HealthSouth was a profitable going concern. Areas of manipulation include: accounts receivable, goodwill, and one-time charges. Smith notes that firms (particularly those in the healthcare industry) are prone to manipulate earnings by understating allowances for uncollectible accounts (AFUA). This claim suggests AFUA relative to total receivables would be lower for fraud firms compared to non-fraud firms. He also describes the practice of understating the net assets of acquirees to create "sock" accounts that could be used to "lever" future understatements, which would overstate goodwill. Thus, we measure each firm's changes in goodwill to examine for patterns of unusually high values.

To capture Smith's "one-time" charges to clean up effects of past earnings manipulations, we develop two proxies: *neg_special_dum*, and indicator variable set equal to one (zero otherwise) if the firm reported a negative special item during year t; and *neg_special_a*, measured as special items scaled by beginning assets (zero otherwise) for observations reporting negative special items in year t. Since measures of the latter proxy are coded as either negative or zero, the predicted sign on *neg_special_a* is negative to indicate that larger, negative one-time charges are associated with a greater likelihood of fraud. The Compustat variables used to measure these proxies appear in Table I.

If our study provides expected findings, we will develop a useful and potentially powerful tool for managers, accounting professionals, creditors, and shareholders concerned with examining whether companies are potentially misstating information to commit frauds.

> RQ1: Specifically identified metrics, along with narrative disclosures, accruals quality, financial performance, nonfinancial measures, off-balance sheet activities, and market-based measures contain information content useful for classifying firms into fraudulent and non-fraudulent groups.
>
> RQ2: Combining specifically identified metrics along with narrative disclosures, accruals quality, financial performance, nonfinancial measures, off-balance sheet activities, and market-based measures in a stepwise regression will result in a model that will provide greater predictive accuracy for early fraud detection than otherwise obtainable earlier models.

Methodology and Analysis

Sample

We develop our prediction model using data provided by Churyk et al., (2009). Their sample was constructed from the population of SEC-defined "fraudulent" firms (i.e., firms receiving an Accounting and Auditing Enforcement Release (AAER) from 2000–2003). From this list, the authors found the originally filed financial statements (i.e., not the restated statements) to identify earnings in the year before the fraud to find matching companies for comparison. Next, they located and matched financial statements of companies in the same industry and of a similar size to produce a non-fraud sample. To test the validity of our model, we use a holdout sample provided by Lee et al., (2013a), that consists of AAER issues from 2004–2006, and a corresponding set of non-fraud matches.[3]

We modified the model and holdout samples to include Dechow et al., (2011) and Smith (2013) variables. Table II provides a summary of how the samples were collected. Per Table II, our final model building and holdout samples include seventy-seven and fifty-seven respective matched pairs of fraud/non-fraud firms. Table III provides each group's industry composition, and denotes a variety of SIC codes, indicating the samples covering many areas. [see Tables II and III, pg 758]

[3] We collected a more recent random holdout sample using AAERs issued from 2016 and 2015. We describe this in the Conclusion section.

Analysis

Churyk et al., (2009) used the Linguistic Inquiry and Word Count Program 2007 (LIWC2007) to content analyze the MDA sections of annual reports for both the fraud and matched (non-fraud) samples.[4] Upon performing that operation, the frequencies of the word occurrences were analyzed and the software factored groups of interrelated words into themes (combinations of word patterns). The themes were also statistically analyzed. We combined their output data with our additional data. As shown in Table I, we used the Churyk et al., (2009) LIWC output for the content analysis variables and for both the specifically identified metrics and Dechow et al., (2011) metrics, we used Compustat data to develop proxies. Once we had values for all variables, we determined which variables were statistically different between fraud/non-fraud firms and then entered the significant correctly signed variables into a logistic regression analysis to construct and analyze the combined model.

Logistic regression can transform the linear probabilities into logit probabilities leading to equation (1):

$$\text{P} = \frac{1}{1+e^{-\alpha-\beta' x'}} \qquad (1)$$

Where x is a vector of input variables, β is a vector of coefficients, and α is a constant.

The logistic maximum likelihood procedure estimates coefficients of factors corresponding to each independent variable. The β coefficients represent the effect of the independent variable on the probability of the misstatement divided by the probability of non-misstatement. The coefficient values reflect the input variables' significance to help identify important cues of fraudulent activity. We extrapolate this approach to the variables from the validated and specifically identified metrics using logistic regression to form a combined and integrated model of early fraud detection.

Results

Table IV reports the results of RQ1; specifically, variables that statistically differ between fraud and non-fraud firms. Results show that fourteen of the forty-seven total variables are statistically different between fraud and non-fraud firms. However, only ten of the fourteen variables are signed correctly. We use these ten variables (*cff, ch_emp, ep, leasedum, neg_special_a, neg_special_dum, colons, semi, posemo,* and *present*) to build the logistic regression model. [see Table IV, pg 759]

We conducted backward stepwise regression to choose from a variable set that included Churyk et al.,'s (2009) and Dechow et al.,'s (2011) theoretically- and empirically-supported metrics, as well as newly-developed proxies based on Smith (2013), deriving our final model's five variables. Per Lee et al., (2013), "backward stepwise regression is the preferred stepwise method because forward stepwise regression will more likely exclude predictors, and forward regression has a higher risk of making a Type II error." [5]

The five variables included in our model to suggest the presence of fraud are: 1) abnormal change in employees; 2) existence of operating leases; 3) negative special items relative to beginning assets; 4) fewer present tense verbs; and 5) fewer semicolons. The coefficients produced by the stepwise model for the original data sample appear in the following equation and, when used collectively, provide overall fraud prediction accuracy rates for the original sample, the cross-validation using the discriminant method, and the holdout sample, respectively, of 73.6 percent, 71.0 percent, and 68.5 percent, each exceeding the model accuracy in prior published papers.

$$\begin{aligned}\text{FRAUD}_i = {}&-19.791 - 1.134\,(\text{ch\_emp}_i) + 21.586\,(\text{lease\_dum}_i)\\ &- 8.818\,(\text{neg\_special\_a}_i) - 0.509\,(\text{Present}_i) - 2.690\,(\text{Semi}_i) + e_i\end{aligned}$$

Thus, applying the equation to the original and holdout samples confirmed our research question. Using Lee et al.,'s (2013a) format, Table V provides descriptive statistics for the five variables, and Table VI provides a matrix of data

[4] Content analysis is a methodology where a computer program parses narrative documents and identifies, for example, parts of speech and syntax.

[5] For validity, we conducted a forward stepwise regression. As suggested by Lee et al., (2013), two important predictor variables (present tense and semi-colons) were excluded from the forward stepwise regression model. This resulted in a lower overall accuracy rate of 55.3% when compared to that of the backward stepwise regression model with an overall accuracy rate of 68.5%.

revealing the fraud prediction statistics to include: 1) results for the original model building sample; 2) the cross-validation of that sample; and 3) the holdout sample. We report results using both a cross-validation method and using a holdout sample for comparison to prior studies.[6] Our original model-building sample contained seventy-seven matched pairs (reduced to sixty-eight during the logistic regression) and our holdout sample contained fifty-seven matched pairs.[7] As shown in Table VI in testing the individual fraud/non-fraud prediction rates (RQ2), our cross-validated method accurately predicted fraud 67.6 percent of the time and non-fraud 74.3 percent of the time. Comparatively, application of the model to the holdout sample accurately predicted fraud 82.5 percent of the time and non-fraud 54.4 percent of the time resulting in an overall accuracy of 68.5 percent—an accuracy amount that exceeds most conventional, quantitative financial models. [see Tables V and VI, pg 760–761]

Conclusion

Since implementing Sarbanes-Oxley legislation, professional managers of publicly held companies became accountable to higher fraud and risk assessment standards, making our results important to them and to the auditors. We extend the fraud prediction literature by combining separately previously examined metrics and newly created proxies based on the HealthSouth fraud described in Smith (2013) into one model. Results show that fraud prediction model accuracy improves with this extension. A limitation of our study is that our sample used Churyk et al., (2009) and Lee et al., (2013 a) data and a very small sample from 2015 and 2016 to provide evidence of model predictability. The data collection process involves hand collection of MDA on our part, thereby imposing significant time and resource constraints. However, agencies such as the SEC have recently developed programs that do not require hand collection. For instance, the SEC (2016) is currently analyzing thousands of narratives using topic modeling (concepts) and sentiment analysis (tone). These concepts are like the LIWC program we used but created explicitly for SEC registrant filing analysis.

[6] Skousen and Wright (2008) use cross-validation.

[7] To test our model on a more recent sample, we examined a random sample of 11 AAERs from the years 2015–2016. Due to data availability, the sample was reduced to nine fraud firms. We matched on revenue in the year preceding the fraud year to develop our sample of non-fraud firms. Results from this holdout sample accurately predicted fraud 77.8 percent of the time and non-fraud 33.3 percent of the time resulting in an overall accuracy of 55.6% percent.

Table I: Variable Definitions

Variable	**Abbreviation**	**Predicted Sign**	**Calculation**
Variables from Churyk et al., (2009)			
Total words	Words	+	Count of a written character or combination of characters representing a word
Lexical diversity	Unique	-	Total number of different words or terms/total number of words or terms x 100, which is the percentage of unique words or terms in all words or terms x 100
Colons	colons	-	Count colons/total number of words or terms x 100
Semicolons	semi	-	Count semicolons/total number of words or terms x 100
For example	example	-	Count the term "for example"/total number of words or terms x 100
Positive emotion	posemo	-	Total number of words or terms indicating positive emotion/total number or words or terms x 100; examples include "happy, pretty, good"
Optimism	optim	-	Total number of words or terms indicating optimism/total number or words or terms x 100; examples include "certainty, pride, win"
Anxiety	anx	+	Total number of words or terms indicating anxiety/total number or words or terms x 100; examples include "nervous, afraid, tense"
Causation	cause	-	Total number of words or terms indicating causation/total number or words or terms x 100; examples include "because, effect, hence"
Present tense	present	-	Total number of present tense verbs/total number of words or terms x 100; examples include "walk, is, be"
Variables from Dechow et al., (2011)			
Accruals quality related variables			
WC accruals	WC_acc	+	[[Current Assets (ACT) – Δ Cash and Short-term Investments (CHE)] - [Current Liabilities (LCT) –Δ Taxes Payable (TXP)]/Average total assets
RSST accruals	rsst_acc	+	(ΔWC + Δ NCO + Δ FIN)/Average total assets, where WC = [Current Assets (ACT) - Cash and Short-term Investments (CHE)] - [Current Liabilities (LCT) - Debt in Current Liabilities (DLC)]; NCO = [Total Assets (AT) - Current Assets (ACT) - Investments and Advances (IVAO)] - [Total Liabilities (LT) - Current Liabilities (LCT) - Long-term Debt (DLTT)]; FIN = [Short-term Investments (IVST) + Long-term Investments (IVAO)] - [Long-term Debt (DLTT) + Debt in Current Liabilities (DLC) + Preferred Stock (PSTK)]; following Richardson et al., (2005).
Change in receivables	ch_rec	+	Δ Accounts Receivable (RECT)/Average total assets
Change in inventory	ch_inv	+	Δ Inventory (INVT)/Average total assets
% Soft assets	soft_assets	-	(Total Assets (AT) - PP&E (PPENT) - Cash and Cash Equivalents (CHE))/Total Assets (AT)
Modified Jones model discretionary accruals	da	+	Measured residuals from estimating this model cross-sectionally using all firm-year observations in the same two-digit SIC code: WC Accruals = $\alpha + \beta$ (1/Beginning assets) + γ (ΔSales – Δ Rec)/Beginning assets + ρ(Change inPPE/)Beginning assets + ε
Performance-matched discretionary accruals	dadif	+	Difference between the modified Jones discretionary accruals for firm *i* in year *t* and the modified Jones discretionary accruals for the matched firm in year t, following Kothari et al.,'s (2005) procedure.
Mean-adjusted absolute value of Dechow and Dichev (2002) residuals	resid	+	Estimate this regression cross-sectionally by industry (2-digit SIC): Δ WC = $\beta_0 + \beta_1$*CFO_{t-1} + β_2*CFO_t + β_3*CFO_{t+1} + ε. Compute the mean absolute value of the residual for each industry and then subtract it from the absolute value of each firm's residual.

Performance variables			
Change in cash sales	ch_cs	-	Percentage change in cash sales [Sales (SALE) - ΔAccounts Receivable (RECT)]
Change in cash margin	ch_cm	-	Percentage change in cash margin, which measures cash margin as 1 - [(Cost of Goods Sold (COGS) - ΔInventory (INVT) + ΔAccounts Payable (AP))/(Sales(SALE) - ΔAccounts Receivable (RECT))]
Change in return on assets	ch_roa	+	[Earnings$_t$ (IB)/Average total assets$_t$] - [Earnings$_{t-1}$/Average total assets$_{t-1}$]
Change in free cash flows	ch_fcf	-	Δ[Earnings (IB) - RSST Accruals]/Average total assets
Deferred tax expense	def_tax	+	Deferred tax expense for year *t* (TXDI)/Total assets for year *t-1* (AT)
Nonfinancial variables			
Abnormal change in employees	ch_emp	-	Percentage change in number of employees (EMP) - percentage change in assets (AT)
Abnormal change in order backlog	ch_backlog	-	Percentage change in order backlog (OB) - percentage change in sales (SALE)
Off-balance-sheet variables			
Existence of operating leases	leasedum	+	Indicator variable set equal to 1 (0 otherwise) if future noncancellable operating lease obligations > zero
Change in operating lease activity	oplease	+	Change in present value of future noncancellable operating lease obligations (MRC1, MRC2, MRC3, MRC4, MRC5)/Average total assets
Expected return on pension plan assets	pension	+	Expected return on pension plan assets (PPROR)
Change in expected return on pension plan assets	ch_pension	+	ΔExpected return on pension plan assets [(PPROR at *t*) - (PPROR at *t-1*)]
Market-related incentives			
Ex ante financing need	exfin	+	Indicator variable coded 1 (0 otherwise) if [(CFO - past three-year average capital expenditures)/Current Assets] < -0.5
Actual issuance	issue	+	Indicator variable coded 1 (0 otherwise) if firm issued securities during the year t (SSTK > 0 or DLTIS > 0)
CFF	cff	+	Level of finance raised (FINCF)/Average total assets
Leverage	leverage	+	Long-term debt (DLTT)/Total Assets (AT)
Market-adjusted stock return	rett	+	Annual buy-and-hold returns inclusive of delisting returns minus annual buy-and hold value-weighted market return
Lagged market-adjusted stock return	rett-1	+	Previous year's annual buy-and-hold returns inclusive of delisting returns minus annual buy-and hold value-weighted market return
Book-to-market	bm	-	Common equity (CEQ)/Market value (CSHO x PRCC)
Earnings-to-price	ep	-	Earnings (IB)/Market value (CSHO x PRCC)
Specifically-identified variables based on Smith (2013)			
Allowance for uncollectible accounts as a percentage of receivables	afua_to_netrec	-	Allowance for doubtful accounts (RECD), divided by total receivables (RECT)
Change in goodwill	ch_gdwl	+	ΔGoodwill (GDWL)/Average total assets
Existence of negative special items	neg_special_dum	+	Indicator variable set to 1 (0 otherwise) if Special items (SPI) < 0 in year t
Negative special items relative to beginning assets	neg_special_a	-	Special items (SPI), deflated by opening total assets (AT), for observations reporting negative special items during year t, and zero otherwise

Table II: Sample Selection

SEC Accounting and Audit Enforcement Releases

	Original	Holdout
SEC AAER's (Revenue related)	311 firms	288 firms
Duplicate firms (multiple AAER's for same offense)	66	145
Subtotal	245	143
Firms without financials or firms without identifiable prior year revenue	152	52
Potential sample firms	93	91
No matching firms (match amended, etc.) (Firms are matched on revenue in the year before the fraud occurred)	25	23
Final individual firm count	68	68
Final sample including multiple years for restated firms for Lee et al., (2013)	88	68
Less matches without our Table I variables	(11)	(11)
Final sample	77	57

Table III: Sample Industry Composition

		Original		Holdout	
SIC Code	SIC Description	Frequency	%	Frequency	%
<1000	Agriculture, forestry, and fishing	0	0.0%	0	0.0%
1000-1999	Mining and construction	1	1.3	0	0.0
2000-2999	Manufacturing – food, tobacco, textile, apparel, lumber, furniture, paper, printing, chemicals and refining.	10	13.0	7	12.3
3000-3999	Manufacturing – rubber, leather, stone, metal, machinery, electronic, transportation, controlling instruments, miscellaneous	24	31.2	12	21.1
4000-4999	Transportation, communications, electric, gas and sanitary	4	5.2	2	3.5
5000-5999	Retail trade	6	1.3	11	19.3
6000-6999	Finance, insurance, and real estate	10	13.0	7	12.3
7000-7999	Services – hotels, personal, business automotive repair, motion picture, amusement	17	22.1	15	26.3
8000-8999	Services – health, legal, educational, social, museums, membership, accounting, engineering, research	4	5.2	3	5.2
9000-9999	Public Administration	1	1.3%	0	0.0
	Total	77	100%	57	100%

Table IV: Statistically Significant Variables (one tailed)

Variable	Predicted Sign	Means	Significance	Correct Direction
afua_to_netrec	-		0.0765	No
Fraud		0.1046		
Non-fraud		0.0649		
cff	+		0.052	yes
Fraud		0.1471		
Non-fraud		0.0740		
ch_emp	-		0.0635	yes
Fraud		(0.1722)		
Non-fraud		(0.0176)		
def_tax	+		0.0145	no
Fraud		(0.01)		
Non-fraud		(.0005)		
ep	-		0.0535	yes
Fraud		(0.2464)		
Non-fraud		(0.0630)		
exfin	+		0.052	no
Fraud		.019		
Non-fraud		0.31		
leasedum	+		0.001	yes
Fraud		0.94		
Non-fraud		0.75		
neg_special_a	-		<0.000	yes
Fraud		(0.06)		
Non-fraud		(0.01)		
neg_special_dum	+		0.001	yes
Fraud		0.53		
Non-fraud		0.29		
ret	+		0.0155	no
Fraud		(0.1944)		
Non-fraud		0.0506		
colons	-		0.026	yes
Fraud		0.1029		
Non-fraud		0.1400		
semi	-		0.078	yes
Fraud		0.12		
Non-fraud		0.15		
posemo	-		0.007	yes
Fraud		2.0700		
Non-fraud		2.3458		
present	-		0.0075	yes
Fraud		2.8032		
Non-fraud		3.0594		

Table V: Descriptive, Univariate, and T-Tests

Original Sample
Non-Fraud

Variable	Mean	Median		SD	SE
Ch_emp	(0.02)	(0.03)	0.34	0.04	
leasedum	0.05	1.00		0.43	0.05
neg_special_a	(0.01)	(0.00)	0.05	0.01	
Present	3.06	3.04		0.68	0.08
Semi-colon	0.15	0.09		0.15	0.02

Fraud

Variable	Mean	Median		SD	SE
Ch_emp	(0.17)	(0.04)	0.80	0.09	
leasedum	0.94	1.00		0.25	0.03
neg_special_a	(0.06)	(0.00)	0.12	0.01	
Present	2.80	2.74		0.62	0.07
Semi-colon	0.12	0.10		0.11	0.01

Holdout Sample
Non-Fraud

Variable	Mean	Median		SD	SE
Ch_emp	(0.04)	(0.05)	0.39	0.05	
leasedum	0.84	1.00		0.34	0.05
neg_special_a	(0.02)	0.00		0.05	0.01
Present	2.72	2.82		0.76	0.10
Semi-colon	0.20	0.20		0.23	0.03

Fraud

Variable	Mean	Median		SD	SE
Ch_emp	(0.27)	(0.08)	0.92	0.13	
leasedum	0.84	1.00		0.37	0.05
neg_special_a	(0.04)	(0.00)	0.08	0.01	
Present	2.76	2.73		0.72	0.10
Semi-colon	0.19	0.11		0.23	0.03

Table VI: Model Prediction Accuracy

Original Sample:

	Non-Fraud	Fraud	Correct %
Non-Fraud	48	20	70.6
Fraud	16	52	76.5
Overall %			73.6

Cross-validated:

	Non-Fraud	Fraud	Correct %
Non-Fraud	55	19	74.3
Fraud	24	50	67.6
Overall %			71.0

Holdout Sample:

	Non-Fraud	Fraud	Correct %
Non-Fraud	31	26	54.4
Fraud	10	47	82.5
Overall %			68.5

References:

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, 6, 203–242.

Burgoon, J. K. (2005). The future of motivated deception and its detection. *Communication Yearbook*, 29, 49–95.

Churyk, N. T., Lee, C. C., & Clinton, B. D. (2009). Early detection of fraud: Evidence from restatements, *Advances in Accounting Behavioral Research*, 12, 25–40.

Churyk, N. T., Lee, C. C., & Clinton, B. D. (2008). "Can we detect fraud earlier?" *Strategic Finance* (October), 50–54.

Dechow, P. M., & Dichev, I. (2002). The quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review*. 77 (Supplement): 35–59.

Dechow, P. M., Ge, W., Larson, C. R., & Clinton, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research*. 28 (1): 17–82.

Eaglesham, J., & Rapoport, M. (2015, January 20). SEC gets busy with accounting investigations: Number of cases and investigations surges at agency. *The Wall Street Journal*. http://www.wsj.com/articles/sec-gets-busy-with-accounting-investigations-1421797895

Kothari, S. P., Leone, A., & Wasley, C. (2005). Performance-matched discretionary accruals measures. *Journal of Accounting and Economics* 39 (1): 163–97.

Lee, C., Churyk, N. T., & Clinton, B. D. (2013) (a) Validating early fraud prediction using narrative disclosures, *Journal of Forensic and Investigative Accounting*. 5(1).

_____ b. Confirming early fraud detection: Validating management accounting's role, *Strategic Finance*.

Lee, C. (2005). Credibility-enhancing displays as a source of cues for the detection in text-based, asynchronous, computer-mediated communication. Working paper.

Securities and Exchange Commission. (2016). Has Big Data made us lazy? https://www.sec.gov/news/speech/bauguess-american-accounting-association-102116.html

Securities Exchange Act of 1934. 2002. §21(d).

Skinner, D. J., & Sloan, R. G. (2002). Earnings surprises, growth expectations, and stock returns or do not let an earnings torpedo sink your portfolio. *Review of Accounting Studies*, 7 (2–3): 289–312.

Skousen, C. J., & Wright, C. J. (2008). Contemporaneous risk factors and the prediction of financial statement fraud. *Journal of Forensic Accounting*, 9, 37–62.

Smith, W. (2013). Lessons from the HealthSouth fraud: An insider's view. *Issues in Accounting Education*. 28(4): 901–912.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T. & Nunamaker, J. F. (2004a). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20 (4): 139–165.

_____ Nunamaker, J. F. Jr., & Twitchell, D. P. (2004b). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13: 81–106.

_____, D. Zhang, & J. F., Nunamaker. (2004c). language dominance in interpersonal deception in computer-mediated communication. *Computers in Human Behavior*, 20 (3): 381–402.