Title
Machine Learning and Corporate Fraud Detection

Author
WALKER, STEPHEN

Publication Date
2021

Peer reviewed|Thesis/dissertation

Machine Learning and Corporate Fraud Detection

by

Stephen Walker

A dissertation submitted in partial satisfaction of the

Requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Panos Patatoukas, Chair
Professor Christine Parlour
Professor Steven Soloman
Professor James Wilcox

Spring 2021

Machine Learning and Corporate Fraud Detection

Abstract

Machine Learning and Corporate Fraud Detection

by

Stephen Walker

Doctor of Philosophy in Business Administration

University of California, Berkeley

Professor Panos Patatoukas, Chair

The purpose of this dissertation was to study why corporate fraud detection models are often met with skepticism by industry practitioners despite a vast literature supporting their use. This dissertation examined the parsimonious standards in the academic literature for corporate fraud detection and included the latest studies that introduced ideas from Benford's Law and machine learning algorithms. The study of corporate fraud detection models is important because academic literature is relied upon by industry practitioners and government regulators including the Securities and Exchange Commission. This paper starts with a critique that was recently published in *Econ Journal Watch*. This critique examined the results of a paper recently published in the *Journal of Accounting Research* applying machine learning to the detection of accounting fraud. Afterwards, I applied the most popular ensemble boosting algorithm in machine learning known as XGBoost to a comprehensive sample of financial ratios and variables. In addition to this model, I ran a horserace with the other models from the extant literature. Results showed that the F-Score (Dechow, et al. 2011) stood up quite well against the machine learning models. Interestingly, a univariate screen on sales growth performed about as well as more complicated methodologies at the top of the probability distribution. Finally, I provided a discussion based on a Bayesian analysis that illustrated why practitioners find fraud detection difficult.

Contents

List of Figures

List of Tables

1. Introduction

The purpose of this dissertation was to understand what works in corporate fraud detection. This subject is highly relevant to industry regulators and practitioners and has generated a large volume of research published in the top academic journals. The primary motivation for this study was driven by the anecdotal evidence whispered amongst industry practitioners that corporate fraud detection models found in the academic journals fall short of expectations, even though this literature reported significant results. I wrote this paper to understand what the key differences were between academics and practitioners, and to contribute to the literature bridging the gap between the two groups.

The corporate fraud literature is vast and covers decades of work in the social sciences including accounting, finance, and law. The extant literature has contributed to understanding how fraud is operationalized inside corporations, and through which observables it might be detected. The last wave of corporate fraud scandals occurred twenty years ago, particularly made famous by the collapses of Enron and Worldcom. These accounting scandals were widely publicized, and the executives earned well-deserved jail terms. These scandals motivated the Congress to prevent future fraud through the passage of new legislation including Sarbanes-Oxley which required improvements in controls and internal processes for large companies in addition to requiring a new level of personal accountability for the senior corporate officers who would become criminally liable for what was presented in financial statements.

Earlier corporate fraud detection models relied upon econometrics including, for example, the logistic and probit regressions. The logistic regression, or the logit, was developed in the 1940s and can be used to model binary dependent variables (e.g., whether a company is fraud or not a fraud). Recent additions to the corporate fraud detection literature applied machine learning algorithms and reported significantly better outcomes over the econometric models. These papers followed a call from Hal Varian nearly seven years ago who implored economists to search for applications for machine learning, which provided for superior capabilities including the ability to handle large variable sets and the ability to model complex nonlinear relationships (Varian 2014). Varian is chief economist at Google and was the founding dean of the School of Information at the University of California at Berkeley. While

artificial intelligence and machine learning are often seen as relatively new phenomena, there have been numerous artificial intelligence cycles over the last 60 years. For example, the Sunday edition of the New York Times on July 13th, 1958 featured a small story about an artificial intelligence experiment ("Electronic 'Brain' Teaches Itself" 1958). In this article, a Navy computer named "Perceptron" was described, which at the time cost $100,000, or about $1 million today. The article reported that this device learned the difference between a right and left orientation following 40 training datapoints. The perceptron was shown to be 97 percent accurate and Navy officers noted that they "hesitated to call it a machine because it is so much like a 'human being without life'". Machine learning and artificial intelligence have experienced numerous cycles and advancements since the perceptron. In the 1960s, economist Herbert Simon said that "machines will be capable, within twenty years of doing any work a man can do," (Simon 1965). Since then, machine learning advanced to rules-based expert systems, to decision trees and support vector machines, to neural networks and ensemble methods that combine multiple trees for prediction. This dissertation applied the latest popular boosting algorithm known as XGBoost, which is short for extreme gradient boosting (Chen and Guestrin 2016). In the last decade, machine learning and artificial intelligence have found success with perception tasks including reverse image search and facial recognition to medical diagnosis using scans (Narayanan 2020). In recent years, Accounting researchers have answered Varian's call and have published papers applying machine learning, including to the task of corporate fraud detection.

Either through the application of older econometric models, or through modern machine learning algorithms, the summary statistic generated from the modeling exercise represents a probability, or a risk score for fraud. These scores can be ranked, and from this list the relevant gatekeepers can go to work. For example, these risk scores could provide early warning to auditors to scrutinize their higher-risk clients, and potentially increase their audit efforts. Other gatekeepers include short-seller activists who could use this information in their search process to short potential frauds. Government regulators and law enforcement professionals could apply them in their investigative work. Given the plethora of models available to researchers, one goal of this dissertation was to benchmark them to provide guidance to practitioners as to which models work best today.

While machine learning models can potentially improve prediction results, they often come at a cost to interpretability. Increasingly, researchers are questioning the social consequences from black box models. An associate professor of computer science at Princeton entitled a presentation "How to recognize AI snake oil" arguing that "much of what's being sold as AI today is snake oil—it does not and cannot work."(Narayanan 2020). He argued that the commercialization of these tools has overly promoted their benefits and that they overshadow the real progress made in AI which included tasks such as reverse image search, facial recognition, medical diagnosis, speech to text, and deep fakes, which are "perception" tasks. Other tasks that were not described as perfect but are improving included automated tasks such as detection of copyrighted material and spam detection. However, the author noted that predicting social outcomes was "fundamentally dubious" which included tasks such as predicting criminal recidivism, job performance, policing, terrorist risk, and at-risk children. In addition to ethical concerns for these prediction models, he wrote that they were amplified by inaccuracy noting

that the problem most significant was the "lack of explainability." The following quotation from that presentation provided a clear illustration to his point.

"Imagine a system in which every time you get pulled over, the police enters your data into a computer. Most times you get to go free, but at some point, the black box system tells you you're no longer allowed to drive."

Currently, governments typically employ a points-based system where there are clear penalties for traffic infractions and the sum of these penalties would lead to license suspension. A black-box system could make a better prediction of your future driving capability, but at the cost of explaining the why behind it. In the last chapter of this dissertation, I will explore a points-based system for detecting fraud and results were interestingly comparable to more complex statistical methods. Relative to machine learning, the econometric models provide for easy interpretation of marginal effects from the observables. In machine learning, "importances" for the variables are available, but these do not provide much insight into the why. Ultimately, exploring causal relationships is best left to a well-designed casual study (Zhao & Hastie 2019). Increasingly, the computer sciences are moving towards causal tools of the economists rather than vice versa.

This dissertation is organized as follows. Chapter 2 provides a thorough review of the literature covering accounting fraud in addition to useful background on the topic. Chapter 3 presents my recently published paper in *Econ Journal Watch* that critiques a machine learning paper published in the *Journal of Accounting Research*. This paper was originally a chapter in this dissertation, but I was able to get it published prior to the completion of this document. Chapter 4 reviews the research design and describes a critical design feature for applying machine learning techniques. Unlike prior models, the time dimension becomes crucial, and researchers should proceed with caution when constructing these models to prevent out-of-sample biases. Chapter 5 applies the XGBoost algorithm to a "kitchen sink" of financial variables and ratios and compares results to the top models in the literature. Chapter 6 dives deeper into which metrics matter. One metric matters the most from the perspective of the fraud investigator, which is *positive predictive value*, or *precision*. A Bayesian analysis shows why the best models still underperform expectations in practice. Chapter 7 explores additional items that did not fit well in the other chapters including variable importance. Chapter 8 concludes with final thoughts.

2. Background

The literature on corporate fraud is multidisciplinary and this paper primarily focused on the detection models within the accounting literature. For a more detailed review of the literature from the fields of accounting, finance, and law, please see Amiram, et al. (2018). This study focused on the most severe form of financial reporting misconduct that culminated into an action by the Securities and Exchange Commission (SEC), specifically the Accounting & Auditing Enforcement Release (AAER). The first multivariate model applying financial statement ratios to detecting AAERs was the M-Score (Beneish 1999), which is commercially available in AuditAnalytics. AuditAnalytics is a data and analytics provider to the Accounting industry. The M-Score is also included in the curriculum for the Chartered Financial Analyst (CFA), an important designation for Wall Street professionals. The F-Score (Dechow, et al. 2011) followed the M-Score, which was estimated with a larger and more comprehensive set of AAER cases. Both models were benchmarked in this dissertation.

In addition to AAERs, there are other potential dependent variables for financial misreporting, many of which would not lead to an SEC action. For example, these outcome variables include shareholder lawsuits and financial restatements. Shareholder lawsuits may occur for many reasons in addition to corporate fraud. Kim & Skinner (2012) studied shareholder lawsuits and created the first multivariate prediction model based on Stanford's shareholder lawsuit database. The other potential dependent variable, financial restatements, covers a wide range of financial reporting mistakes not necessarily related to fraud. The available database for financial restatements is AuditAnalytics' non-reliance financial restatement database. Larcker & Zakolyukina (2012) studied restatements and found that the F-Score and the M-Score performed poorly when applied to restatements relative to AAERs. Unreported results are consistent with this finding that shareholder lawsuits and financial restatements made poor dependent variables for detecting fraud, which is why the AAER was applied in this dissertation.

Returning to the financial-ratios based models, Beneish estimated the M-Score using a probit regression that was based on a limited matched sample of problematic financial statements that included Accounting and Auditing Enforcement Releases (AAERs). Dechow, et al. (2011) studied a much larger and comprehensive database of AAERs starting in the early 1980s and produced a detection model based on a logistic regression analysis of seven key variables that

can be easily derived from financial reports. The authors currently support this database for research and graciously gave me access to use it in this dissertation. The AAER database can be obtained from the USC Marshall School of Business (Dechow, et al. 2011). Regarding the advanced machine learning tools, early literature included Perols (2011), which compared the performance of various machine learning and statistical models and Cecchini, et al. (2010), which applied a support vector machine to a custom financial kernel mapping of financial statement variables. However, these studies involved matched samples. Inferences from matched samples (e.g., 50:50) can be significantly different when applied to real world prevalence (e.g., 1:200). In fact, Beaver previously observed this issue in the context of bankruptcy (Beaver 1966). A potential downside to studying rare events with machine learning is that machines may have difficulty in training. Perols, et al. (2017) explored random undersampling procedures for the training data and showed that they improved model effectiveness. In fact, it is standard procedure in machine learning to perform either a random over-sampling or under-sampling to balance the training data. Note that out-of-sample test sets are unchanged so that inferences would remain valid. Bao, et al. (2020) applied a boosting model with random undersampling called RUSBoost using sample data with realistic rare prevalence rates. One innovation of their paper was that they did not apply financial ratios, unlike the M-Score or the F-Score, and instead relied upon raw financial statement variables. The benefit from this approach was to provide a simpler and more direct way to generate risk scores based on what was available from the financial reports without requiring additional transformations. The tradeoff was interpretability since machine learning explored nonlinear paths to generate predictions. In their publication, they reported a 70 percent improvement over F-Score, which was a large margin. Since the *Journal of Accounting Research* required the publication of the code supporting their published paper, I took a closer look that led to writing a critique that was recently featured in *Econ Journal Watch.* This critique was included in this dissertation in the next chapter. Outside of traditional machine learning, Amiram, et al. (2015) introduced Benford's law which is based on the distribution of the first digits. While Benford's law has been in the toolkit for fraud investigators outside of corporate fraud detection, the authors were the first to bring it to the detection of AAERs, and their summary measure was included in the benchmarking analysis in Chapter 5. An additional discussion on Benford's law is included in Chapter 7.

In addition to prediction modeling, another area of research focused on how fraud is discovered. Dyck, Morse and Zingales (2010) explained that most fraud was typically uncovered by other means including investigative journalism or the results of criminal investigations writing that fraud detection "takes a village." This village involved outside parties including employees, media, and industry regulators. They described different views for detecting fraud. For example, "The legal view claims fraud detection belongs to the auditors and securities regulators" (Coffee 1986 and Dyck, et al. 2010). The finance view (Fama 1990 & Dyck, et al 2010) said that debt and equity holders would do the heavy lifting including their analyst and auditor agents. In contrast to these points of view, Dyck, Morse, and Zingales ultimately found that employees, other industry regulators, and the media were largely responsible for detecting fraud while auditors and the SEC only accounted for ten percent and seven percent of detection, respectively.

Another recent publication applied textual analysis to "assess whether the thematic content of financial statement disclosures is incrementally informative in predicting intentional misreporting" (Brown, et al. 2020). Like Bao, et al. (2020), the authors noted an improvement at the top one percent of the probability distribution, though these results did not stand out relative to the results reported by the Bao, at al. paper. Both papers were featured in the same March 2020 journal issue. In fact, Bao and coauthors argued, "one interesting question future researchers may explore is whether the usefulness of textual data continues to hold if the information from the readily available raw financial data is more efficiently extracted using advanced data mining techniques." They further noted that "our results raise the bar for this line of text-mining research because we show that the commonly used Dechow, et al. ratio-based logistic regression model significantly understates that value of financial data in fraud prediction" (Bao et al. 2020).

A quick take from practitioners reveals something interesting about the usefulness of these models which the accounting literature should be aware. For example, short seller Carson Block of Muddy Waters Research stated publicly in a recent interview that "my issue with running screens is…you get a lot of false positives." (Block 2020). I also spoke with Dr. Schilit, author of the book *Financial Shenanigans*, which provided the first popular field study of SEC enforcement actions. He advised me that "you've got to get your hands dirty," and understand the "why?" When asked directly whether he applied advanced statistical models to fraud detection, his answer was simple: "No" (Schilit 2020).

3. Critique of Recent Paper in Detecting Accounting Fraud

The following critique was published in *Econ Journal Watch* on March 31st, 2021 (Walker 2021)[1]. This critique covered an article in the *Journal of Accounting Research* (Bao, et al. 2020). This critique was originally included in this chapter but was submitted and approved for publication prior to the filing of this dissertation. Proper permissions were obtained both from this author, in addition from the editor of the journal Professor Daniel Klein.

The authors of the critiqued article (Bao, et al. 2020) replied to this critique in the same journal issue (Bao, et al. 2021). Their reply is not included here but is available online at no cost to the reader[2]. I included my rebuttal following this article.

[1] https://econjwatch.org/1231
[2] https://econjwatch.org/1232

3.1 Critique

This critique treats an article in *Journal of Accounting Research* entitled "Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach" by authors Yang Bao, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang (Bao et al. 2020). In addition to the published paper, the authors provide their Matlab code with an associated dataset in a CSV file[3]. This paper applies their code and dataset to replicate the results and study the key assumption driving those results.

Within the fields of accounting and finance, corporate fraud detection models have been the subject of a significant volume of work. The literature follows a long line of prediction and detection models found in the literature on capital markets. Parties with interest in these models include the investing public and regulatory bodies such as the Securities and Exchange Commission. Previous corporate frauds including Enron and Worldcom left significant damage in their wake, affecting not only their employees and investors but also the public's trust and faith in capital-market institutions. The great hope is that an early warning system can alert the Securities and Exchange Commission and investors to potential fraud and act before the fraud grows too large.

The previous standard in the accounting literature for detecting accounting fraud is known as the F-Score, which is based on a seven-variable logistic regression model published by Patricia Dechow and collaborators (2011). For modeling purposes, the best proxy for accounting fraud is the SEC-issued Accounting and Auditing Enforcement Release (AAER), an enforcement action that describes the fraud and typically orders a restatement of previously issued financial reports (e.g., 10-Ks). The observable covariates to these fraud models are financial statement ratios that might include changes in sales, accounts receivables, and inventories, in addition to indicator variables for capital-markets activity including share or debt issuances. These ratios are based on a long line of theoretical and empirical work. A novel innovation of the Bao et al. (2020) paper is that they do not use financial ratios, but rather apply raw financial variables taken directly from the financial statements.

The authors provide a dataset that includes a total of 146,045 firm-year observations from 1991–2014. The data comes from the CompuStat database. AAER data is sourced from the USC Marshall School of Business (previously the Haas School of Business). Unique AAER cases total 413 (each of which may last multiple years), and the sample's total fraud-case firm-years is 964 firm-years. Taking the 964 AAER-affected firmyears and dividing by the total of 146,045 firm-years gives an approximation for the unconditional probability of finding fraud for any firm in any given year as 0.7 percent. Fraud is a rare event, and comparing detection rates against this unconditional expectation is important within accounting research.

Replicating the paper is relatively simple. The software Matlab is required. The Matlab code file is called "run_RUSBoost28." The dataset is a CSV file called "uscecchini28.csv." The column headers are shown in Figure 1.

[3] https://github.com/JarFraud/FraudDetection

Figure 1: CSV file (the dataset)

Position	Column	Description
1	fyear	Fiscal Year
2	gkvey	Compustat firm identifier
3	sich	4-digit Standard Industrial Classification Code (SIC)
4	insbnk	An indicator variable for financial institutions between SIC 6000-6999
5	understatement	An indicator variable if the misstate indicator involved an understatement
6	option	Not used
7	p_aaer	Identifier for AAER
8	new_p_aaer	New Identifier for AAER
9	misstate	Indicator variable for misstatement
10	act	Current Assets - Total
11	ap	Accounts Payable - Trade
12	at	Assets - Total
13	ceq	Common/Ordinary Equity - Total
14	che	Cash and Short-Term Investments
15	cogs	Cost of Goods Sold
16	csho	Common Shares Outstanding
17	dlc	Debt in Current Liabilities
18	dltis	Long-Term Debt Issuance
19	dltt	Long-Term Debt Total
20	dp	Depreciation and Amortization
21	ib	Income Before Extraordinary Items
22	invt	Inventories - Total
23	ivao	Investment and Advances Other
24	ivst	Short-Term Investments - Total
25	lct	Current Liabilities – Total
26	lt	Liabilities – Total
27	ni	Net Income (Loss)
28	ppegt	Property, Plant and Equipment - Total (Gross)
29	pstk	Preferred/Preference Stock (Capital) - Total
30	re	Retained Earnings
31	rect	Receivables Total
32	sale	Sales/Turnover (Net)
33	sstk	Sale of Common and Preferred Stock
34	txp	Income Taxes Payable
35	txt	Income Taxes - Total
36	xint	Interest and Related Expense - Total
37	prcc_f	Price Close - Annual - Fiscal

The dependent variable is an indicator variable equaling 1 if the AAER covered the firm-year in the data, and zero otherwise which is in the dataset's column 9, labeled *misstate*. The independent variables are 28 raw financial statement variables reported by the company in their annual report and shown in columns 10–37, which include items such as total assets and ending price per share for the period. In the Matlab code, the dataset will be divided into a training and test set. For example, the first looped-trained model was based on data covered by the period from 1991 through 2001. The model was then applied out of sample, e.g., to the year 2003, and that application generated a probabilistic score for each firm in that year. The top 1 percent of the probability scores were taken from this selection and if there is a firm in this subset with an actual AAER for that year, it is counted as a correctly identified positive hit. The fraction of correct hits is the positive predictive value. The model was run iteratively for each year in the study's test period, 2003 through 2008.

Machine learning requires measuring results using a hold-out test sample because machine learning can overfit training datasets and produce results that are too good to be true. An iterative approach is preferable because it shows results as it steps through time, which is what would be experienced in the real world, and thus adds validity to the model. A two-year (or longer) gap between the training sample and test sample is required because AAERs are not immediately known when financial reports are issued. In fact, many years can pass between the financial report and the AAER issuance. A modeler must ask (in the spirit of Senator Howard Baker): What can the model know, and when can the model know it?

One issue related to that question involves serial frauds. Some serial frauds may traverse both training and test periods since they cover more than the gap period. To address this issue, the readme file that accompanies the data and code[4] notes:

> "The variable new_p_aaer is used for identifying serial frauds as described in Section 3.3 (see the code in "RUSBoost28.m" for more details)."

Section 3.3 from their paper is reported in its entirety below, with boldface added to emphasize the action described.

> 3.3 SERIAL FRAUD (Bao, et al. 2020)
> Accounting fraud may span multiple consecutive reporting periods, creating a situation of so-called "serial fraud." In our sample, the mean, median, and 90th percentile of the duration of the disclosed accounting fraud cases is two years, two years, and four years, respectively, suggesting that it is **common for a case of fraud to span multiple consecutive reporting periods.** Such serial fraud may overstate the performance of the ensemble learning method if instances of fraudulent reporting span both the training and test periods. This is because ensemble learning is more flexible and powerful than the logistic regression model, and may therefore be better able to fit a fraudulent firm than a fraudulent firm-year. **Hence, enhanced performance of the ensemble learning method may result from the fact that both the training and test samples contain the same**

[4] https://github.com/JarFraud/FraudDetection/blob/master/README.md

> **fraudulent firm**; the ensemble learning model may not perform as well when the sample contains different firms. **To deal with this concern, we break up those cases of serial fraud that span both the training and test periods. Because we have a small number of fraudulent firm-years relative to the number of nonfraudulent firm-years in any test year, we recode all the fraudulent years in the training period to zero for those cases of serial fraud that span both the training and test periods.** Although this approach helps us avoid the problems associated with serial fraud, it may also introduce measurement errors into the training data. (Bao et al. 2020, 211–212, my emphases)

In summary, serial fraud concerns AAER cases that span multiple reporting periods. However, the section does not directly address why the column *new_p_aaer* was created. Returning to the Matlab code for an explanation, Figure 2 shows the code for the model.

Figure 2: The Matlab Code

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% run RUSBoost model with 28 raw accounting variables as features %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% set parameters
iters = 3000; % the number of iterations/trees of RUSBoost model
gap = 2; % the gap between training and testing periods, 2-year gap by default

%% train and test models
for year_test = 2003:2008
    rng(0,'twister'); % fix random seed for reproducing the results
    % read training data
    fprintf('Running RUSBoost (training period: 1991-%d, testing period: %d, with %d-year gap)...\n',year_test-gap,year_test,gap);
    data_train = data_reader('uscecchini28.csv','uscecchini28',1991,year_test-gap);
    y_train = data_train.labels;
    X_train = data_train.features;
    newpaaer_train = data_train.newpaaers;
    data_test = data_reader('uscecchini28.csv','uscecchini28',year_test,year_test);
    y_test = data_test.labels;
    X_test = data_test.features;
    newpaaer_test = unique(data_test.newpaaers(data_test.labels~=0));
    % handle serial frauds as described in our paper
    num_frauds = sum(y_train==1);
    y_train(ismember(newpaaer_train,newpaaer_test))=0;
    num_frauds = num_frauds - sum(y_train==1);
    fprintf('Recode %d overlapped frauds (i.e., change fraud label from 1 to 0).\n',num_frauds);

    % train model
    t1 = tic;
    t = templateTree('MinLeafSize',5); % base model
    % fit RUSBoost model (default parameters: learning rate: 0.1, RatioToSmallest: [1 1])
    rusboost = fitensemble(X_train,y_train,'RUSBoost',iters,t,'LearnRate',0.1,'RatioToSmallest',[1 1]);
    t_train = toc(t1);
    % turn on the following line of code if you want to get feature importance
    % [imp,ma] = predictorImportance(rusboost);

    % test model
    t2 = tic;
    [label_predict,dec_values] = predict(rusboost,X_test); % predict frauds in the testing year
    dec_values = dec_values(:,2); % get fraud probability
    t_test = toc(t2);

    % print evaluation results
    fprintf('Training time: %g seconds | Testing time %g seconds \n', t_train, t_test);
    metrics = evaluate(y_test,label_predict,dec_values,0.01); % topN=0.01
    fprintf('Performance (top1%% as cut-off thresh): \n');
    fprintf('AUC: %.4f \n', metrics.auc);
    fprintf('NCDG@k=top1%%: %.4f \n', metrics.ndcg_at_k);
    fprintf('Sensitivity: %.2f%% \n', metrics.sensitivity_topk*100);
    fprintf('Precision: %.2f%% \n', metrics.precision_topk*100);
    % fprintf('Importance of predictors:%d\n', output.imp);

    % turn on the following lines of code if your want to save prediction results in a file
    output_filename = ['prediction_rusboost28_',num2str(year_test),'.csv'];
    dlmwrite(output_filename,[data_test.years, data_test.firms, y_test, dec_values],'precision','%g');

end
```

Line 10 starts the loop that runs the model iteratively stepping through each year of the test period from 2003–2008. Line 21 creates a list of unique values of AAER identifiers where the *misstate* column is not equal to zero (equal to 1) for the test set. Line 24 performs the action described in Section 3.3 and sets the *y_train* indicator values to zero where there is a match in the AAER identifiers in the training sample to the previously created list from the test sample.

The intention of Section 3.3 appears to be correctly coded in Matlab. However, what is the *new_p_aaer* field? In Figure 1, the 7th position contains another field called *p_aaer*. The *p_aaer* field is the AAER number that matches the SEC issued number, which can be searched on the SEC website.[5] When comparing these two columns, it appears that *new_p_aaer* takes the original AAER number and adds a '1' or '2.' In fact, all but 17 AAER cases take the original AAER number and add a '1.'

I sent an email to the authors of the paper copying their editor and asked specifically about this issue. Professor Ke Bin sent the following response on behalf of the author group to all recipients of the original email (boldface added):

> As we discussed in Section 3.3 of our paper, "we recode all the fraudulent years in the training period to zero for those cases of serial fraud that span both the training and test periods." **Our serial frauds have two requirements: (1) have the same AAER id, and (2) are consecutive in our sample. "1" and "2" are suffix to distinguish serial frauds with the same AAER id but not consecutive in our sample.**

I understand the first part of the requirement. However, I do not understand the second part to the requirement—which was not described in the paper or in the online supporting documents. The serial fraud issue is a problem with the span of the fraud itself, not whether it is consecutive in their sample.

The reason that some cases are not consecutive in the sample was provided by the next explanation, given by Professor Ke when I asked why there were a few missing firm-year observations in the sample.

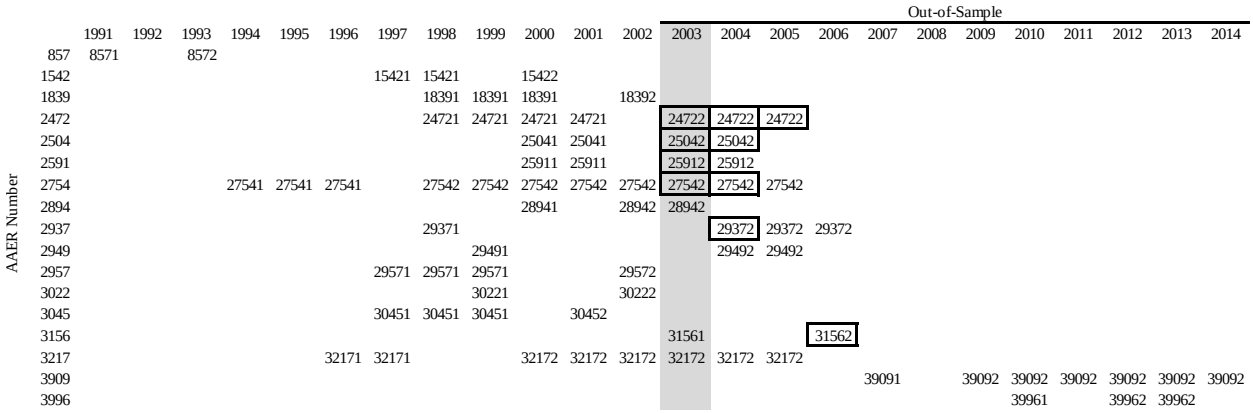
> We require all observations to contain non-missing values for the 28 raw accounting variables, consistent with prior studies cited in our paper. Those observation [related to the 17 AAERS] are dropped because one of the 28 raw variables are missing in WRDS COMPUSTAT database. For example, firm-years of AAER No. 2472, 2504, 2591, and 2894 are missing DLTIS (Long-term debt issuance) and firm-years of AAER No. 2754 and 3217 are missing XINT (Interest and related expense, total).

To show what Professor Ke is speaking to, Figure 3 shows the AAERs at issue. There are only 17 AAERs where *new_p_aaer* changes values because of the "not consecutive in our sample" issue out of a total of 413 unique AAERs in their sample. Additionally, a large fraction of correct cases identified by the model are related to these 17 AAERs. The number of firm-years correctly

[5] https://www.sec.gov/divisions/enforce/friactions.htm

identified by the AAERs from 2003–2008 total 10 firm-years and are shown in the bolded boxes. The total correct cases identified by their model are 16 firm-years. So, 63 percent of the correct cases are associated with this issue.

Figure 3: Seventeen AAER cases with two different new AAER identifiers



Professor Ke's explanation is not consistent with how other variables are handled in the dataset. The statement suggests a rule that an observation is dropped if it has a missing CompuStat variable. According to the "SAS coding.pdf" file[6], the authors recoded *txp*, *ivao*, *ivst*, and *pstk* to 0 if they were missing. If done for these four variables, why are variables *dltis* and *xint* inconsistently handled?

However, the real issue is not these missing observations per se. Rather, it is the additional requirement that a consecutive sample be required for serial fraud identification. Section 3.3 of their paper describes the bias in machine learning related to serial fraud occurring when "both the training and test samples contain the same fraudulent firm" (Bao et al. 2020, 211–212). To illustrate, take for example AAER No. 2504. This AAER affected Delphi Corporation for the years 2000–2004 and was issued by the SEC in 2006. Summarizing Delphi in context of the Matlab code,

- If an AAER identifier from the test set matches the same identifier from the training set, the Matlab model recodes AAER's *misstate* = 1 in the training set to 0.
- As shown in Figure 3, the AAER identifier for Delphi **changes** to 25041 in the training set and to 25042 in the test set.
- Because Delphi 25042 is not in the training set, the Matlab code **will not** recode Delphi 25041's *misstate* = 1 to 0.

Because the Matlab code treats Delphi AAER No. 2504 as two different AAERs 25041 and 25042, the same fraudulent firm is contained in both the training and test samples. Therefore, the

[6] https://github.com/JarFraud/FraudDetection/blob/master/SAS%20coding.pdf

Bao et al. (2020) results are still susceptible to the problem they addressed in Section 3.3. In fact, if Delphi's AAER had not been changed, their machine learning model would not have identified the fraud for the year 2003 or 2004 contributing significantly to the published results.

I investigated how the authors' AAER identifier change affected the results. I return the AAER identifiers to their original values by replacing the column *new_p_aaer* with data from the *p_aaer* column in the CSV file. This avoids making any code changes within Matlab. Running their original code on this modified dataset excludes from training the additional firm-years associated exactly with these 17 unique AAER cases, but changes nothing else.

Figure 4: Three model scenarios

Panel A. Correct cases predicted to be positive

Year	(1) Published Model	(2) Re-run Model	(3) Recoded Model
2003	8	7	4
2004	4	4	3
2005	2	2	1
2006	1	1	0
2007	1	1	1
2008	0	0	0
Total	**16**	**15**	**9**

Panel B. Positive predictive values (correct cases / # predicted positive)

Year	(1) Published Model	(2) Re-run Model	(3) Recoded Model
2003	13.3 percent	11.7 percent	6.7 percent
2004	6.7 percent	6.7 percent	5.0 percent
2005	3.4 percent	3.4 percent	1.7 percent
2006	1.7 percent	1.7 percent	0.0 percent
2007	1.7 percent	1.7 percent	1.7 percent
2008	0.0 percent	0.0 percent	0.0 percent
Total	**4.5 percent**	**4.2 percent**	**2.5 percent**

The updated results are reported in Figure 4. The first column reports the results by year from the Github supporting documents. Correct cases total 16 for the 2003–2008 out-of-sample test, corresponding to a 4.5 percent positive predictive value, matching the reported values published. Positive predictive value, also known as precision, is calculated as the proportion of

correct AAER firm-years out of the cases predicted to experience an AAER. The second column reports the results I obtain when running their original code on their original dataset, showing 15 correct cases corresponding to a 4.2 percent positive predictive value (I'm not sure why it is 15 rather than 16 as in the published paper). The third column reports the results I obtain when running their original code on the dataset with the AAER identifiers replaced by their original values, showing only 9 correct cases corresponding to a 2.5 percent positive predictive value. This value is critical because their published model compared the machine learning result with the result from a parsimonious logit model based on prior literature, which their paper reports to be 2.63 percent for positive predictive value. The updated result shows that the prior model in the literature outperforms this machine learning approach.

The crucial issue in the present critique is to address whether it is appropriate to give new identifiers to the AAER because there is a break in the series resulting from missing data. Since the serial fraud issue concerns the span of the AAER itself and not the sample data, there does not appear to be a logical purpose for the recoding done by the authors. Giving a new AAER identifier to these 17 unique cases out of a total of 413 disproportionately improved their reported results. Without the change, results do not improve upon the prior literature.

3.2 Walker Rejoinder to Bao, et al. 2021

In their reply published in the same edition of *Econ Journal Watch* (Bao, et al. 2021), the authors chose not to respond to the central issue raised in the critique, which was: What was the justification for relabeling AAER identifier values? The authors responded to an initial email inquiry writing that these relabelings were necessary because of a previously undisclosed requirement that a consecutive sample was needed. This explanation made no sense, which was the motivating factor in writing the critique, and the authors provided no further justification in their reply. AAER identifier values were used to identify the AAERs. A recoding would imply that there were two different AAERs issued, which did not occur. I showed that recoding identifiers overstated reported results—so much so, that the logit-based regression from prior literature outperformed their machine learning model.

To illustrate once again, take, as an example, a fraud case that covered fiscal years 2000 through 2002. When making predictions for the year 2002 (the test sample), the authors argued in their original paper that that the fraud indicator variable in the year 2000 should be recoded to zero because of this serial-fraud issue. In essence, they did not want the model to learn from its own case. However, an additional requirement was added within the code that was not explained in the paper: Do this procedure only if *all* observations for 2000-2002 are contained in the sample. Note that training ends in 2000 for the prediction year 2002 because of a two-year gap requirement, which is a separate issue from this discussion. So, the observation for year 2001 is irrelevant here. In cases where all observations are available, the AAER indicator variable would be recoded to zero for year 2000, which is consistent with the description written in their original paper. Consider the alternative scenario, which introduces the controversy. If the observation for the irrelevant year 2001 was dropped because of a missing values problem, then the authors would relabel the fraud identifier for any fraud year that occurred thereafter. So, for this example, the fraud identifier for 2002 would be labeled differently from the fraud identifier for the year 2000. Since the two identifiers no longer match (for the same AAER), the fraud indicator variable would not be recoded to zero and would therefore be included in training, which contradicts Section 3.3 of their original paper. Without this relabeling, the model performed no better than a logit-based regression that existed previously in the literature.

Rather than give a rational explanation to relabeling identifiers, the authors chose to write extensively on other topics, and concluded with one that was irrelevant to the critique. Their reply started by addressing the missing values problem. In their original email responding to my initial questions, the authors stated that observations with missing values were dropped. However, some observations with missing values were not dropped, and instead were filled with zeros. When I asked why this was the case in the critique, the reply explained that the treatment was consistent with prior literature and that they applied this logic consistently. So, CompuStat

variables *txp*, *ivao*, *ivst*, and *pstk* were recoded to zero because it followed "common practice in the prior accounting literature." Observations that had missing values for other variables, including debt issuance, were dropped from the sample. While one would think that these cases could also be reasonably recoded to zero, the authors chose not to do so and provided no further explanation. However, I encourage the reader to move beyond this concern because it is not the real issue. Rather, it is why would AAER identifiers need relabeling at all?

The next section of their reply was entitled "Walker's approach to dealing with serial fraud." I do not know why the authors chose to attribute my name to their own written methodology as described in Section 3.3 of their original paper. They started with the argument that I did not recalibrate the number of trees in the RUSBoost parameters. To provide some background on this issue, machine learning algorithms contain several parameters which could be changed so that a better fit can be obtained for the training sample. However, it is unknown a priori how this would affect the out-of-sample test. Specifically, the authors wrote that I did not properly tune the number of trees parameter to optimize performance, and, when they did so, their performance improved from what I reported in my critique. First, upon inspection of their results in the reply shown as their Table 1, taking their "improved" parameters at face value, results were virtually identical to the results I showed in my critique making it obvious that tuning does not matter. For example, for the main sample period of 2003-2008, I reported 9 hits in the critique while they reported 10 hits in the reply. So, there was one additional hit from this "tuning." Furthermore, this result is far from the purported improvement where they reported 16 hits in the original publication. Second, the authors do not state how their parameter tuning was implemented, nor do they provide the code for this process either with their published paper or with this reply. This is typically a requirement since parameter tuning must be done only on the training sample, and if it was done to maximize out-of-sample performance, then the procedure would be invalid. Generally, tuning does not alter machine learning performance significantly. In fact, recent literature in the computer sciences reported that leaving models at their default parameter values was non-inferior to optimization (Weerts, Mueller, and Vanshoren 2020).

The second issue raised in this section was that I only published Table 3 from their original paper for the years 2003-2008, while ignoring the results from the following Table 5, which included three alternative test samples. The implication was that I cherry-picked results. The reason I chose Table 3 was that it was their *main result*. In their original paper, they stated "we use the years 2003-2008 as our primary test sample" (Bao, et al. 2020). Table 5 was included for robustness in a section entitled "Supplemental Analyses." Regardless, equivalent comparisons for the alternate period 2003-2005 could be easily calculated since I provided cross-sectional results by year in the analysis, whereas they only report the overall average. Incremental results beyond 2008 reported by them were essentially the same between the logit-based model and the RUSBoost model. The authors concluded this section by writing that the RUSBoost "always outperforms." This statement contradicted their own reported table because they show the value for NDCG@k for the logit-based model outperformed the value for their

RUSBoost model (0.0273 vs. 0.0237). We know that this was their preferred metric because, in the original paper, the authors wrote that relative to the AUC (Area-under-the-curve), the "NDCG@k is more useful to regulators and other monitors." In the reply, the authors also concluded that results did not alter inferences. How could this be true? This new table showed results far from the purported 70 percent improvement shown in their original publication.

The last section was entitled "What is the optimal approach for dealing with serial fraud?" They concluded this section by saying, "Walker's approach of relying solely on p_AAER ID to define serial fraud could be inappropriate." Again, why is this my approach? What other way is there to identify the AAER except with the AAER identifier? Their Figure 1 was entitled "a serial fraud example with a key fraud revelation event during the training period." While timing of fraud revelation might make an interesting discussion, it was never addressed in my critique. In fact, what they label as "Walker's approach" is precisely their approach applied in the paper to all observations without a missing intervening variable.

In summary, the authors provided no justification for relabeling identifiers with their reply. In my critique, I made a clear case that their code was not consistent with how their paper described the implementation of the solution to the serial fraud problem as described in Section 3.3 of their publication. Without a reasonable explanation, results do not hold up to scrutiny and I can only conclude that their RUSBoost model does not outperform the logit-based model for the detection of corporate fraud. While the authors wrote otherwise, their data supported this conclusion, which showed that the logit-based model from prior literature outperformed their RUSBoost model for the main sample period 2003-2008 in terms of their preferred metric NCDG@k.

4. Research Design: What Can the Model Know, and When Can the Model Know It?

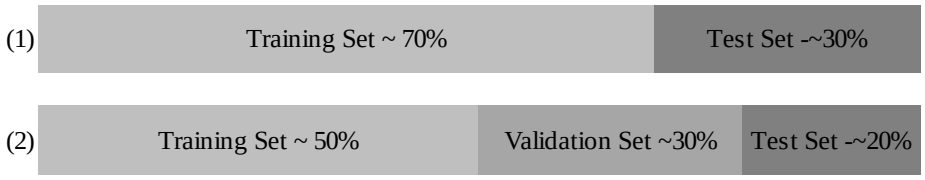
This chapter provides detail into the research design for implementing a reproduceable and believable machine learning process. Replicability is a significant problem in the economic sciences and out-of-sample testing is considered a gold standard to reporting results. Econometric models historically were estimated and reported using in-sample results. Until recently, dataset and computational power limitations largely prevented the use of out-of-sample data. With big data, out-of-sample testing becomes increasingly possible. With machine learning, it is necessary. Because machine learning maps closer to the nonlinearities in the data, overfitting the in-sample dataset is a significant concern, which could lead to poor out-of-sample (real-world) performance. To avoid this problem, it is a standard procedure to measure and report test statistics with out-of-sample data. With time-series data, this means out-of-time.

Up until the most recent studies, testing out-of-sample meant that the researcher kept a hold-out sample separate from the estimation process. This was typically done through a random selection of the sample splitting it into two or three groups. Under the two-group method, the first group was assigned to the in-sample training group used for model estimation. This is known as the training set. The second group was the hold-out sample used to measure results, called the test set. A third group can be added to improve robustness by inserting a validation group in between the training and test sets, which is the typical design for deep learning problems since estimation of deep learning models requires iteration over the validation sample. Optimally, the test sample will be analyzed once in the entire research process. In practice, this might occur a few times, though with each iteration, a feedback loop can be intentionally or unintentionally created that would lead to overstated results. In addition to the splitting of the

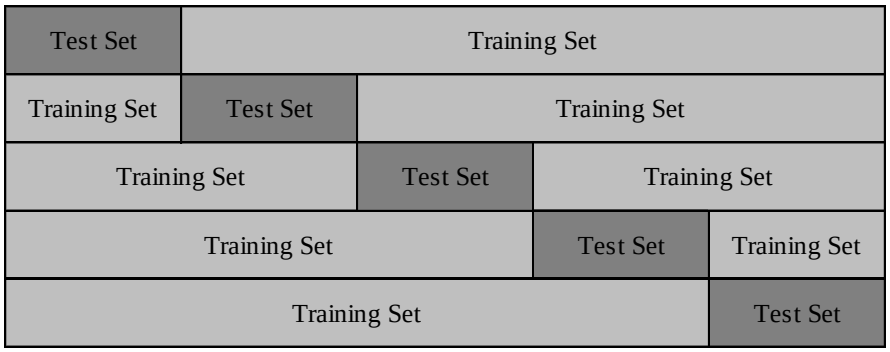
data, cross validation is a technique that reduces the dependence of the estimates on the random selection of the training and test groups. In cross validation, the results from multiple models using different subsamples are averaged. The following example describes a leave-on-out cross validation procedure. However, this method pools the time dimension, which implies covariate stability across time, which is often not the case with real-world data. Therefore, results could still be over-stated even with this cross-validation procedure.

Figure 5: Illustrations of Validation and Cross-validation

Panel A: Examples of Sample Splitting



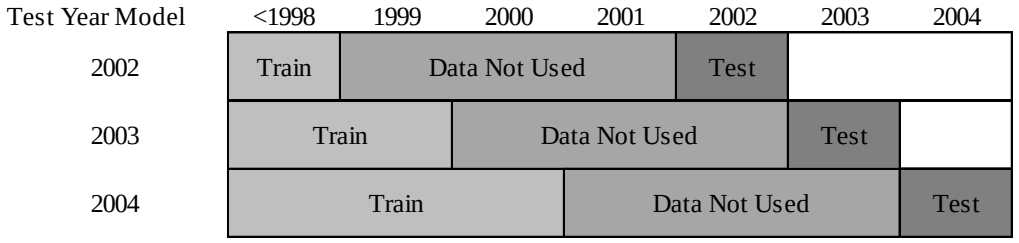
Panel B: Example of Five-fold Cross Validation using leave-one-out



Since our data includes a time dimension, a recursive out-of-sample process was implemented. This method improves the believability of the out-of-sample results because it measures how the model would have performed had it been implemented at the time. Roger Stein of Moody's Investor Service described this as the "walk-forward" approach that his firm employed (Stein 2007). Recent studies including Bao, et al. (2020) implemented this walk-forward approach. However, prior studies avoided it perhaps because it was computational

expensive. Results from recent models applying machine learning are therefore not comparable to earlier studies. In this dissertation, there are five models that were re-estimated recursively (the other four are static models, and do not require re-estimation). For a 20-year rolling period, this translates to 100 uniquely trained models. A decade ago, this might have been intractable. Today, this estimation can be done on a personal laptop with a graphics processing unit. An example figure showing this approach is shown below.

Figure 6: Recursive Design or the "Walk-forward"



In this design, we are attempting to make a prediction for a single year. Take the year 2004, for example. At some point in early 2005 we would have access to the financial detail for firms in 2004 and would like to make a prediction as to those that might stand out for further inspection. My AAER database would only be up to date with known AAERs through some fiscal year in the past because it takes time for these to come to light. While, in 2021, I already know about those cases that occurred in 2003, I could not have known about them then. In fact, as I stand in 2021, I have reasonable confidence that I know about most cases through 2017. The AAER database applied in this paper unfortunately does not include dates of issuance. However, I hand collected a sample post 1999 from the SEC website and found that both the mean and median time between the fiscal year and year of issuance was four years. Arguably, a news story might break that would occur before the AAER issuance. Karpoff, et al. (2017) reported that the average number of years between the fiscal year and the year of first revelation was closer to three years. In this study, I chose four years because it was the more conservative of the two and I cannot know before-hand if a news story would result in an AAER, which is the dependent variable studied. This intervening period is known as the "gap" assumption.

This gap is a critical assumption and involves the question—What can the model know and when can the model know it? Since Bao, et al. (2021) brought up the issue in their reply to my critique, which I had not originally examined in the critique, I can address it here. In their original paper, they chose a two-year gap window. So, the prediction for year 2004 would have known about *all* AAER cases through the year 2002, which was a strong assumption. I took

their data and code and changed one parameter within the code, which was the gap-year assumption. In Table 1 in Columns 2 and 3, I change this figure to 3 years and 4 years, respectively. Observe how performance dramatically drops by the third column to half what it was in the published paper. By this point, the logit-based model would have outperformed. After returning the 17 relabeled AAERs to their original values, the logit-based model certainly outperformed.

Of course, this gap assumption is not foolproof. Arguably, since the mean number of years between the fiscal year involved and the year of AAER issuance was four years, up to half of the AAERs would potentially enter the training set potentially biasing results. To guard against machine learning finding its own cases for frauds that span multiple years, the same methodology described by Bao, et al. (2020) was implemented where the indicator variable for fraud was recoded to zero wherever there was a match in the test sample.

Table 1: Changing Gap Year Assumptions from Bao, et al. (2020)

Panel A: Top 1 percent: AAERs Captured

Year	(1) Published Model: 2yr	(2) Same Data Gap: 3yr	(3) Same Data Gap: 4Yr	(4) Return 17 AAERs to Original Nos. Gap: 4Yr
2003	**8**	**3**	**2**	**1**
2004	4	6	4	2
2005	2	2	2	1
2006	1	1	0	0
2007	1	1	0	0
2008	0	0	0	0
Total	**16**	**13**	**8**	**4**

Panel B: Top 1 percent: Positive Predicted Values (AAERs captured / number in top 1 percent)

Year	(1) Published Model: 2yr	(2) Same Data Gap: 3yr	(3) Same Data Gap: 4Yr	(4) Return 17 AAERs to Original Nos. Gap: 4Yr
2003	**13.3%**	**5.0%**	**3.3%**	**1.7%**
2004	6.7%	10.0%	6.7%	3.3%
2005	3.4%	3.4%	3.4%	1.7%
2006	1.7%	1.7%	0.0%	0.0%
2007	1.7%	1.7%	0.0%	0.0%
2008	0.0%	0.0%	0.0%	0.0%
Total	**4.5%**	**3.7%**	**2.3%**	**1.1%**

5. Model Horserace: A Comprehensive Benchmark

The purpose of this chapter is to provide a comprehensive benchmark for the best models in the literature in the detection of AAERs. This benchmarking will apply the walk-forward methodology as described in the previous chapter. The benefit of this approach is that it gives visibility into the year-to-year performance of the model whereas previous literature has averaged these out-of-sample years into one statistic. Second, this benchmarking runs for 20 years from fiscal years 1993 through 2012, which is the longest out-of-sample test period in the literature. This maps into how practitioners of these models would experience them in the real world. Year-to-year model results matter particularly as the performance of investigators and investors are measured on an annual basis. Statistical significance of the models were made using a Wilcoxon rank sum test.

5.1 Description of Models

An overview of the models benchmarked include the following.

(1) Financial Ratios (XGBoost)

This model applies the XGBoost algorithm applying an exhaustive set of financial variables and ratios as compiled by Perols, et al. (2017). The definitions of these variables are provided in Appendix A.

(2) Raw Vars (XGBoost)

This model uses the XGBoost algorithm and applies 28 raw financial variables sourced directly from the financial statements as originally proposed by Bao, et al. (2020), but with a new gap assumption of four years to be comparable with the first model. The definitions of these variables are provided in Appendix B.

(3) Raw Vars (RUSBoost)

This model is sourced directly from the code provided by the authors of Bao, et al. (2020). The new extensive variable set is applied to the model, but with a new gap assumption of four years to be comparable to the XGBoost model.

(4) 4-Year Sales Growth (Screen)

As suggested by Dr. Schilit, this screen is a simple four-year geometric growth rate on sales.

(5) F-Score

This score is based on the Dechow, et al. (2011) and applies the logit coefficients estimated in the original paper. Unlike the previous models, this model is not re-estimated through time.

(6) M-Score

This score is based on Beneish (1999) and applies the original estimated coefficients to the custom variables described in the paper. Similar to the F-Score, this model is not re-estimated through time.

(7) FSD Score

The Financial Statement Divergence Score (FSD Score) was created by Amiram, et al. (2015). The FSD Score is based on Benford's law, which examines the distribution of first digits of financial variables. Benford's law has been applied in other domains of fraud research, which has found an association between fraud and the deviation from empirical

distributions relative to the theoretical distributions. The actual FSD Scores applied were downloaded from coauthor Professor Bozanic's website.

(8) 7 Vars (Logit)

This logistic regression is the same logistic regression from Dechow, et al. (2011) that generates the F-Score. However, in this scenario, the model is re-estimated iteratively across the twenty-year sample period with the four-year gap requirement.

(9) 7 Vars + FSD (Logit)

This logistic regression is the same logistic regression from the previous model adding the FSD Score as an additional variable.

When I attempted to examine prior literature and compare how the models performed relative to each other, I found the exercise difficult due to a variety of issues. Samples and prevalence were different with each study. These differences affected comparability. Reported statistics varied depending on the classification threshold. Since models only output some measure of fraud probability, the classification threshold is a choice left to the researcher. Some models included the threshold agnostic "area-under-the-curve" measure, but others did not. Recent papers emphasized accuracy at the top of the probability distribution, where researchers care most about the highest fraud-risk cases. Accuracy at the top is an idea that comes from the information retrieval literature. For example, a user of Google cares most about the accuracy in the first page of results and is much less concerned about the accuracy across the other pages. The next chapter dives into these metrics in more detail because it is a much larger issue and critical to properly measuring fraud detection. I followed the most recent literature and measured accuracy at the top of the probability distribution. I applied a threshold examining the top 1 percent and the top 10 percent and uses *positive predictive value*, which is also known as *precision*, as the preferred metric. Positive predictive value is the proportion of true positives out of the number of predicted positives in the sample. Bao, et al. (2020) chose the top 1 percent level since it roughly matched the unconditional expectation. However, I choose both the top 1 percent and the top 10 percent to illustrate key issues with examining solely the top 1 percent.

Machine learning models have parameters that include choice of loss or objective function, the number of leaves and trees that can be created, and so on. Many out of the box implementations of these machines come with defaults based either on keeping computational intensity low, or on other best practices recommended by the literature. Unreported tests show

very little sensitivity from varying these parameters, and much of the success in machine learning is driven by the improvements in data rather than through the tuning of model parameters. Furthermore, recent computer science literature suggested that leaving these parameters at their default values was not inferior to optimizing them (Weerts, et al. 2020). For the XGBoost parameters, the tree method selected was 'gpu_hist', which activated the onboard NVIDIA graphics processing unit significantly improving training time (i.e., this option makes estimation blazingly fast relative to a CPU, and it is one of the reasons for its popularity and success). The objective function was set to optimize 'rank:ndcg" since this measure maximizes accuracy at the top of the probability distribution. I will discuss what NDCG is in the next chapter. Unreported tests using the logistic objective function did not change inferences. The max_bin setting was 1000, and the random seed was set to 42 to allow for replicability of results. As discussed in the previous chapter, the gap assumption, or the time between the final training sample and the test year predicted is four years. For the RUSBoost MATLAB model form Bao, et al., the parameters were unchanged, other than to change the gap assumption to four years for consistency across all models.

5.2 Sample Selection

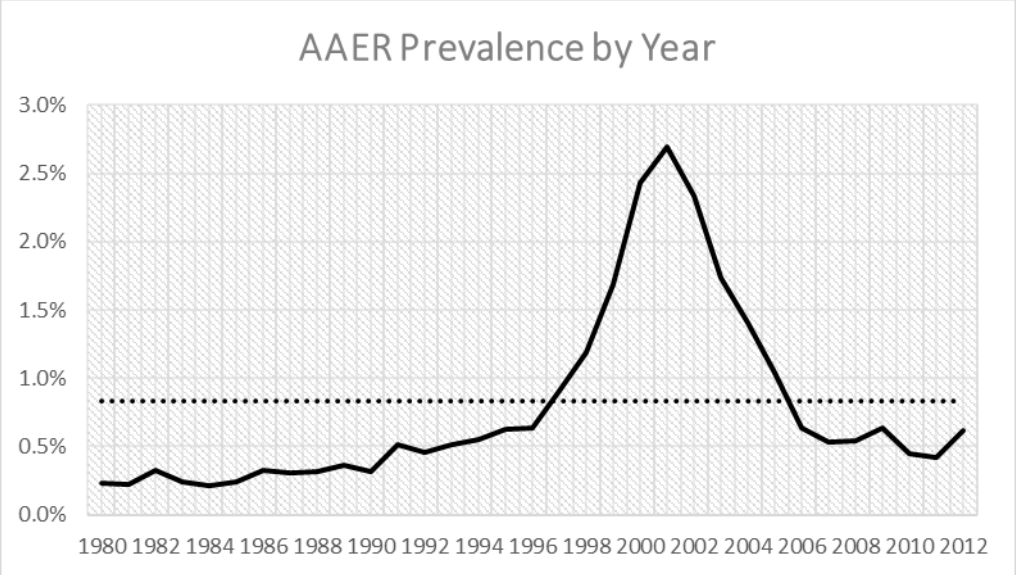
The sample selection process is shown in Table 2. Financial data was sourced from the CompuStat database for the years 1979-2012, which provided sufficient data for models to start the out-of-sample 20-year period in 1993. This design drops financials and penny stocks, or where stock prices are missing. Stocks are required to trade on major exchanges or over the counter, and must have non-missing or positive sales, total assets, income, and cash & equivalent balances. The proxy for corporate fraud is the Accounting and Auditing Enforcement Release (AAER) issued by the SEC sourced from the AAER database that can be obtained from the USC Marshall School of Business (Dechow, et al. 2011). For this sample, there were a total of 112,113 observations and 939 AAERs for a prevalence across the entire sample of 0.84 percent.

Figure 7 shows a chart for the AAER prevalence by year. AAERs are rare events, and the dotted line represents the unconditional expectation of the sample, or 0.84 percent. However, there are large year-to-year differences in AAER prevalence, notably in the early 2000s. Towards the end of the sample period, the prevalence declined significantly from its previous peak.

Table 2: Sample Selection

Steps	**Firm-years**
Compustat 1979-2012	352,092
Drop GICS Codes 40 (Financials) and 60 (REITs) or missing	(84,766)
Drop penny stocks (Stock price <$5), or missing stock price	(142,222)
Require stocks listed on NYSE, AMEX, NASDAQ, OTC	(9,228)
Drop missing or negative/zero Sales, Total Assets; and, missing income or cash & equivalent balances	(1,823)
Keep years 1980 - 2012	(1,940)
Total Sample Observations	112,113
Total SEC Auditing and Accounting Enforcement Releases (AAER)	939
AAER Prevalence	0.84%

Figure 7: AAER Prevalence by Year



While there is an open question whether the drop in cases in the later years were driven by enforcement choices, this drop could also have been driven by improvements in accounting systems implemented through Sarbanes-Oxley compliance. However, Judge Rakoff opined in 2014 that the government had failed to prosecute senior executives in connection with the financial crisis of the late 2000s (Rakoff 2014). He wrote that this decision may have been driven by diversion of resources related to the Madoff scandal amongst other reasons including the government's role in the financial scandal. While the overall rates have declined, regulators still pursue accounting fraud cases. One recent example involves the case of MiMedx Group, a Biotech company based in Marietta, Georgia. The SEC charged the company and the executives with Accounting Fraud and the Department of Justice unsealed indictments against former CEO Parker Petit. In November 2020, a jury found Mr. Petit guilty of securities fraud and, in February 2021, Judge Rakoff sentenced Mr. Petit to one year in federal prison. In the virtual sentencing hearing, Judge Rakoff opined "Mr. Petit did indeed intentionally, knowingly and willingly commit securities fraud. This is much less than I would have given, but for his health, age and good deeds." (Kanell & Quinn 2021).

Note that this fraud was uncovered by short-sellers and not through financial modeling applied by the SEC. While not the topic of this dissertation, Twitter is becoming increasingly valuable as a tool to share ideas amongst the short-selling community. A must-see Twitter video showing the confrontation between the short-seller Marc Cohodes and Parker Petit at a 2018 JP Morgan conference was posted by Edwin Dorsey (Dorsey 2020). Dorsey achieved fame for exposing Care.com's shoddy business practices that led to the resignation of their leadership team (Grind, et al. 2020).

5.3 Empirical Results

The following tables report the empirical results of this horserace. Aggregations included the first 10 years of the sample spanning 1993-2002, the second 10 years spanning 2003-2012, a subset spanning 2003-2008 to match the out-of-sample period from Bao, et al. for comparison purposes, and a full 1993-2012 summary. The results from the top 1 percent are presented in Table 3.

Analyzing this table, the XGBoost model applying the comprehensive set of financial ratios outperformed the models applying raw financial variables alone. Recall that the authors of Bao, et al. argued that machine learning can learn from raw variables alone with no need for further calculations or ratio transformation. This analysis refutes this argument. The application of financial ratios is consistent with how a human researcher with an organic brain would analyze financial statements through calculating margins and growth rates and making comparisons.

For the second out-of-sample period 2003-2012, the XGBoost beats all other models in the sample. Though, for the 20-year period, the re-estimated F-score called the 7-Vars (Logit) model in Column 9 beat all models implying that the logistic model is at least as good, or

perhaps even better than machine learning. The shaded grey boxes reveal the best model by year, and it lights up like a Christmas tree with no model beating any other model consistently. In the first 10 years, the F-Score (Column 6) and the re-estimated F-Score (Column 9) perform the best. One argument could be made that the F-Score performed well because the original paper was estimated using a similar dataset with a similar time frame. It is possible that there was some overfitting of results though they also held up relative to the other models in the second ten-year period.

One problem with examining the top one percent is just how few cases get discovered. Table 4 reports the raw number of true positive hits for each model revealing that few make it to the list. For those models that worked, only one or two cases were captured. Since there are roughly 38 firms at the top 1 percent each year, finding only 1 case out of these 37 would make the search futile. Even if the search is done, marginal results would be obtained.

The motivation for examining the top 1 percent is to reduce the number of false positives in the sample, which comes at the expense of missing much of the known cases in the sample. Increasing the proportion classified positive will increase the false positives but will increase the number the number of discovered cases as well. Therefore, for comparison purposes, I expanded this set to examine the top 10 percent of known cases, as shown in Table 5. Positive predictive values declined, but the models showed consistent results year-to-year, unlike the top 1 percent. Again, the XGBoost model performed best overall, but results do not appear much better than the simpler logistic-based models using far fewer variables. The F-Score performed about as well as the most advanced machine learning algorithms. Perhaps more surprising is how well a simple univariate screen on sales growth performed relative to the advanced statistical models. For the 20-year period, the four-year geometric growth rate on sales roughly matched the F-Score and the XGBoost models. Table 6 shows the number of hits for each model and as expected, there were substantially more hits. However, this came at a cost: the number of firms increased 10x from the previous amount to roughly 380 firms per year on average.

Overall, machine learning with XGBoost added a marginal benefit, but did not significantly improve upon the F-Score. If running a simple screen, a screen on sales works about as well any other methodology. From there, knowledge about industry and other firm-specific knowledge could be added by the researcher to make investigations worthwhile. However, launching investigations from this high-risk list alone would likely be cost prohibitive given the resources it takes to perform a well-researched analysis.

If a logit (or probit) based model is applied, the F-Score superceded the M-Score, which made sense given that it was estimated using a larger sample of AAERs. This is an important finding because the M-Score continues to be part of the CFA curriculum in addition to being commercially available through AuditAnaltyics.

Regarding Benford's Law, the evidence for the FSD Score showed that it does not add much value in the detection of AAERs either as a stand-alone or in combination with the seven F-Score variables. This is a somewhat surprising finding given that the original paper said that the FSD Score "predicts material misstatements as identified by SEC Accounting and Auditing

Enforcement Releases" (Amiram, et al. 2015). Their paper received much fanfare in the popular press and led to the 2017 Deloitte Foundation Wildman Medal Award presented at the American Accounting Association's annual conference. AuditAnalytics also makes the Benford's Law Analysis commercially. The results call into question the commercial viability of both Benford's Law and the M-Score as measures to detect fraud.

Table 3: Positive Predictive Values for the Top 1 percent (highest risk)

		Positive Predicted Value									
Year	Prevalence (1)	Financial Ratios (XGBoost) (2)	Raw Vars (XGBoost) (3)	Raw Vars (RUSBoost) (4)	4-Yr Sales Gr. (Screen) (5)	F-Score (Dechow) (6)	M-Score (Beneish) (7)	FSD Score (Benford) (8)	7 Vars (Logit) (9)	7 Vars + FSD (Logit) (10)
1993	0.5%	-	-	7.5%	-	5.0%	2.5%	-	na	na
1994	0.6%	-	2.5%	2.5%	2.5%	-	-	2.6%	-	-
1995	0.6%	6.7%	2.2%	2.2%	4.4%	2.2%	4.4%	-	7.0%	7.0%
1996	0.6%	-	2.0%	2.0%	2.0%	2.0%	2.0%	2.2%	4.3%	2.1%
1997	0.9%	4.0%	2.0%	6.0%	-	2.0%	2.0%	-	2.1%	4.2%
1998	1.2%	4.5%	-	4.5%	4.5%	4.5%	4.5%	2.4%	4.8%	4.8%
1999	1.7%	2.1%	6.4%	4.3%	1.9%	6.4%	-	2.3%	2.3%	2.3%
2000	2.4%	5.1%	10.3%	2.6%	3.4%	12.8%	2.6%	2.8%	13.5%	13.5%
2001	2.7%	8.3%	11.1%	5.6%	8.3%	5.6%	13.9%	3.0%	11.8%	5.9%
2002	2.3%	12.5%	6.3%	-	3.1%	12.5%	-	3.6%	10.0%	10.0%
2003	1.7%	5.3%	2.6%	-	-	2.6%	2.6%	2.9%	2.8%	2.8%
2004	1.4%	2.5%	5.0%	-	-	2.5%	5.0%	2.8%	2.7%	2.7%
2005	1.0%	-	-	-	2.6%	2.6%	-	2.9%	2.8%	2.8%
2006	0.6%	2.5%	-	-	-	-	-	-	-	-
2007	0.5%	2.6%	-	-	-	2.6%	2.6%	-	2.9%	2.9%
2008	0.5%	3.6%	-	-	-	-	-	-	3.8%	3.8%
2009	0.6%	3.1%	-	-	-	-	-	-	-	-
2010	0.5%	2.9%	-	-	-	-	-	-	-	3.2%
2011	0.4%	3.2%	-	-	-	3.2%	3.2%	-	-	-
2012	0.6%	3.2%	-	-	-	3.2%	-	-	-	-
'93-'02	1.3%	4.0%	4.0%	3.8%	2.9%	5.0%	3.1%	1.8%	5.3%	4.8%
'03-'12	0.8%	2.8%	0.9%	-	0.3%	1.7%	1.4%	1.0%	1.7%	2.1%
'03-'08	1.0%	2.7%	1.3%	-	0.4%	1.8%	1.8%	1.5%	2.4%	2.4%
1993-2012	1.1%	3.5%	2.6%	2.1%	1.8%	3.5%	2.3%	1.4%	3.8%	3.6%

Table 4: Total Number of AAERs Captured for the Top 1 percent (highest risk)

	AAERs Captured by Model								
Year	Financial Ratios (XGBoost) (1)	Raw Vars (XGBoost) (2)	Raw Vars (RUSBoost) (3)	4-Yr Sales Gr. (Screen) (4)	F-Score (Dechow) (5)	M-Score (Beneish) (6)	FSD Score (Benford) (7)	7 Vars (Logit) (8)	7 Vars + FSD (Logit) (9)
1993	-	-	3	-	2	1	-	na	na
1994	-	1	1	1	-	-	1	-	-
1995	3	1	1	2	1	2	-	3	3
1996	-	1	1	1	1	1	1	2	1
1997	2	1	3	-	1	1	-	1	2
1998	2	-	2	2	2	2	1	2	2
1999	1	3	2	1	3	-	1	1	1
2000	2	4	1	2	5	1	1	5	5
2001	3	4	2	3	2	5	1	4	2
2002	4	2	-	1	4	-	1	3	3
2003	2	1	-	-	1	1	1	1	1
2004	1	2	-	-	1	2	1	1	1
2005	-	-	-	1	1	-	1	1	1
2006	1	-	-	-	-	-	-	-	-
2007	1	-	-	-	1	1	-	1	1
2008	1	-	-	-	-	-	-	1	1
2009	1	-	-	-	-	-	-	-	-
2010	1	-	-	-	-	-	-	-	1
2011	1	-	-	-	1	1	-	-	-
2012	1	-	-	-	1	-	-	-	-
'93-'02	17	17	16	13	21	13	7	21	19
'03-'12	10	3	-	1	6	5	3	5	6
'03-'08	6	3	-	1	4	4	3	5	5
1993-2012	27	20	16	14	27	18	10	26	25

Table 5: Positive Predictive Values for the Top 10 percent (high-risk)

		Positive Predicted Value									
Year	Prevalence (1)	Financial Ratios (XGBoost) (2)	Raw Vars (XGBoost) (3)	Raw Vars (RUSBoost) (4)	4-Yr Sales Gr. (Screen) (5)	F-Score (Dechow) (6)	M-Score (Beneish) (7)	FSD Score (Benford) (8)	7 Vars (Logit) (9)	7 Vars + FSD (Logit) (10)
1993	0.5%	2.0%	1.3%	2.5%	0.3%	1.5%	1.3%	0.3%	na	na
1994	0.6%	1.3%	1.3%	1.0%	1.3%	1.5%	1.0%	0.8%	0.3%	0.3%
1995	0.6%	0.7%	0.7%	1.4%	1.8%	2.3%	1.4%	0.9%	2.4%	2.4%
1996	0.6%	1.2%	0.4%	1.6%	1.2%	2.0%	1.4%	0.6%	1.9%	1.3%
1997	0.9%	2.6%	1.4%	2.0%	1.6%	2.0%	0.8%	0.4%	1.5%	1.9%
1998	1.2%	1.8%	1.8%	2.1%	3.2%	3.4%	3.0%	1.0%	3.1%	2.9%
1999	1.7%	3.9%	4.1%	3.7%	4.8%	3.5%	1.7%	2.3%	3.2%	3.4%
2000	2.4%	6.3%	3.4%	6.0%	6.3%	6.0%	3.7%	0.8%	6.6%	6.1%
2001	2.7%	7.0%	5.6%	8.1%	7.6%	5.6%	5.3%	3.9%	8.0%	7.4%
2002	2.3%	8.0%	5.8%	4.2%	5.1%	4.8%	2.9%	4.8%	5.5%	5.5%
2003	1.7%	4.0%	4.3%	3.7%	3.2%	3.5%	2.9%	1.4%	3.1%	3.1%
2004	1.4%	2.0%	1.8%	3.3%	2.3%	2.0%	1.3%	1.9%	2.2%	2.2%
2005	1.0%	2.6%	1.0%	2.1%	2.3%	1.6%	1.3%	2.0%	1.4%	1.4%
2006	0.6%	1.3%	1.5%	0.5%	1.0%	1.8%	0.3%	0.6%	1.4%	1.7%
2007	0.5%	0.5%	0.3%	0.3%	0.5%	0.8%	0.8%	0.3%	0.6%	0.6%
2008	0.5%	0.4%	-	0.4%	0.7%	0.7%	0.4%	0.4%	1.2%	0.8%
2009	0.6%	0.6%	0.3%	1.9%	0.6%	-	0.6%	1.0%	-	-
2010	0.5%	1.2%	0.9%	0.9%	0.3%	1.2%	-	1.0%	1.0%	1.0%
2011	0.4%	1.0%	1.0%	0.6%	0.6%	0.6%	0.3%	0.7%	0.7%	0.7%
2012	0.6%	1.6%	0.7%	1.6%	0.3%	1.3%	1.6%	0.7%	0.7%	1.1%
'93-'02	1.3%	3.2%	2.4%	3.1%	3.1%	3.1%	2.1%	1.4%	3.1%	2.9%
'03-'12	0.8%	1.6%	1.2%	1.6%	1.3%	1.4%	1.0%	1.0%	1.4%	1.5%
'03-'08	1.0%	1.9%	1.5%	1.8%	1.7%	1.8%	1.2%	1.1%	1.7%	1.7%
1993-2012	1.1%	2.5%	1.9%	2.4%	2.3%	2.4%	1.6%	1.3%	2.4%	2.3%

Table 6: Total Number of AAERs Captured for the Top 10 percent (high-risk)

	AAERs Captured by Model									
Year	Financial Ratios (XGBoost) (1)	Raw Vars (XGBoost) (2)	Raw Vars (RUSBoost) (3)	4-Yr Sales Gr. (Screen) (4)	F-Score (Dechow) (5)	M-Score (Beneish) (6)	FSD Score (Benford) (7)	7 Vars (Logit) (8)	7 Vars + FSD (Logit) (9)
1993	8	5	10	1	6	5	1	na	na
1994	5	5	4	5	6	4	3	1	1
1995	3	3	6	8	10	6	4	10	10
1996	6	2	8	6	10	7	3	9	6
1997	13	7	10	8	10	4	2	7	9
1998	8	8	9	14	15	13	4	13	12
1999	18	19	17	22	16	8	10	14	15
2000	24	13	23	24	23	14	3	24	22
2001	25	20	29	27	20	19	13	27	25
2002	25	18	13	16	15	9	14	16	16
2003	15	16	14	12	13	11	5	11	11
2004	8	7	13	9	8	5	7	8	8
2005	10	4	8	9	6	5	7	5	5
2006	5	6	2	4	7	1	2	5	6
2007	2	1	1	2	3	3	1	2	2
2008	1	-	1	2	2	1	1	3	2
2009	2	1	6	2	-	2	3	-	-
2010	4	3	3	1	4	-	3	3	3
2011	3	3	2	2	2	1	2	2	2
2012	5	2	5	1	4	5	2	2	3
'93-'02	135	100	129	131	131	89	57	121	116
'03-'12	55	43	55	44	49	34	33	41	42
'03-'08	41	34	39	38	39	26	23	34	34
1993-2012	190	143	184	175	180	123	90	162	158

On the topic of statistical significance between the models, Table 7 reports a Wilcoxon matched-pair signed-rank test with two-sided z-stats and p-values for each model comparison at both the top 1 percent and the top 10 percent levels. As expected, most models were statistically significant relative to a random draw out of the hat. However, the FSD Score as a stand-alone (Benford's Law) was not statistically significant at the p=0.10 level for any of the observations. The M-Score was only significant for the first 10-year period for the top 10 percent.

Table 7: Wilcoxon Matched-Pair Signed-Rank Tests (Two-sided) For Models Compared to the Unconditional Expectation

	Top 1%			Top 10%		
1993-2002	Z	Pr |z|	PPV	Z	Pr |z|	PPV
vs. Unconditional Expectation						
(1) Financial Ratios (XGBoost)	1.89	0.06	4.0%	2.80	0.01	3.2%
(2) Raw Variables (XGBoost)	2.40	0.02	4.0%	2.60	0.01	2.4%
(3) Raw Variables (Matlab RUSBoost)	2.29	0.02	3.8%	2.80	0.01	3.1%
(4) 4-year Sales Growth (Screen)	2.19	0.03	2.9%	2.70	0.01	3.1%
(5) F-Score (Dechow)	2.70	0.01	5.0%	2.80	0.01	3.1%
(6) M-Score (Beneish)	1.38	0.17	3.1%	2.60	0.01	2.1%
(7) FSD Score (Benford Law)	1.48	0.14	1.8%	0.76	0.44	1.4%
(8) Logistic - 7 Variables	2.55	0.01	5.3%	2.55	0.01	3.1%
(9) Logistic - 7 Variables + FSD Score	2.55	0.01	4.8%	2.55	0.01	2.9%
2003-2012						
vs. Unconditional Expectation						
(1) Financial Ratios (XGBoost)	2.70	0.01	2.8%	2.19	0.03	1.6%
(2) Raw Variables (XGBoost)	(0.97)	0.33	0.9%	1.33	0.18	1.2%
(3) Raw Variables (Matlab RUSBoost)	(2.80)	0.01	0.0%	2.09	0.04	1.6%
(4) 4-year Sales Growth (Screen)	(1.89)	0.06	0.3%	1.58	0.11	1.3%
(5) F-Score (Dechow)	1.78	0.07	1.7%	2.29	0.02	1.4%
(6) M-Score (Beneish)	0.56	0.58	1.4%	0.25	0.80	1.0%
(7) FSD Score (Benford Law)	(0.05)	0.96	1.0%	1.38	0.17	1.0%
(8) Logistic - 7 Variables	1.27	0.20	1.7%	2.19	0.03	1.4%
(9) Logistic - 7 Variables + FSD Score	1.78	0.07	2.1%	2.09	0.04	1.5%

Note: Positive Z favors the in-row variable; negative values favor the bolded variable.

Table 8 reports the Wilcoxon matched-pair signed-rank tests when compared to the F-Score, which is the current standard in the literature. A positive Z value means that it favors the in-row model, while a negative Z value favors the F-Score. For the top 1 percent, no model beats the F-Score at the 5 percent level and models are generally insignificant either way relative to the F-Score. At the top 10 percent level, inferences are similar. Year-to-year differences in model performance vary too much for any one model to beat out the other.

Table 8: Wilcoxon Matched-Pair Signed-Rank Tests (Two-sided) For Models Compared to the F-Score

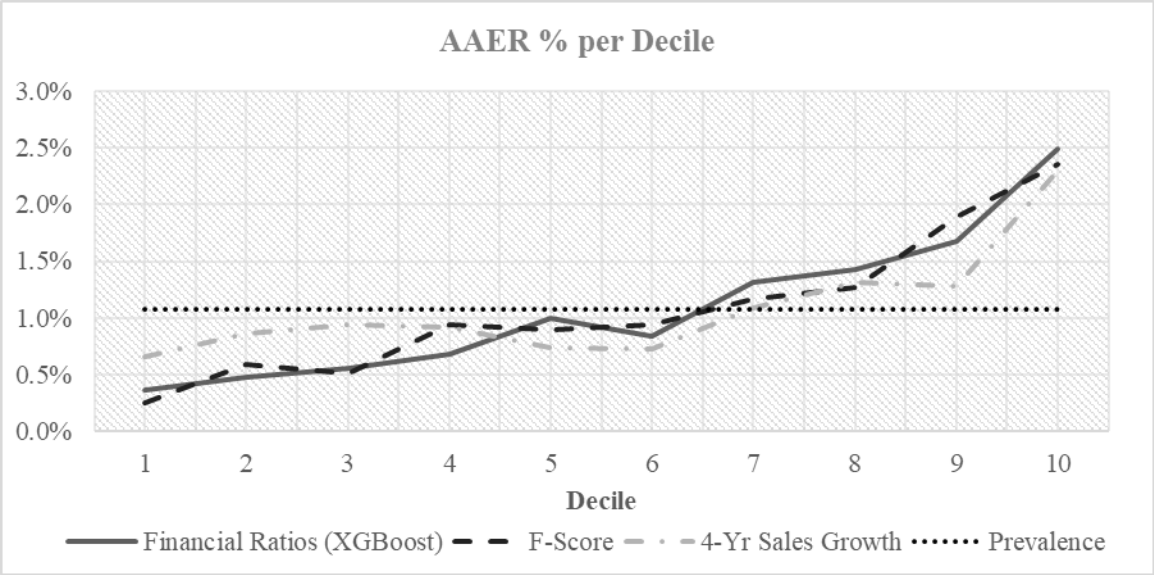
	Top 1%			Top 10%		
1993-2002	Z	Pr |z|	PPV	Z	Pr |z|	PPV
vs. F-Score (Dechow)						
(1) Financial Ratios (XGBoost)	(0.67)	0.50	4.0%	1.48	0.14	3.2%
(2) Raw Variables (XGBoost)	(0.90)	0.37	4.0%	0.36	0.72	2.4%
(3) Raw Variables (Matlab RUSBoost)	(0.16)	0.87	3.8%	(1.63)	0.10	3.1%
(4) 4-year Sales Growth (Screen)	(1.13)	0.26	2.9%	(0.51)	0.61	3.1%
(6) M-Score (Beneish)	(1.01)	0.31	3.1%	(0.15)	0.88	2.1%
(7) FSD Score (Benford Law)	(2.09)	0.04	1.8%	(2.80)	0.01	1.4%
(8) Logistic - 7 Variables	0.61	0.54	5.3%	(2.80)	0.01	3.1%
(9) Logistic - 7 Variables + FSD Score	0.31	0.76	4.8%	(0.36)	0.72	2.9%
2003-2012						
vs. F-Score (Dechow)						
(1) Financial Ratios (XGBoost)	1.75	0.08	2.8%	0.82	0.41	1.6%
(2) Raw Variables (XGBoost)	(1.54)	0.12	0.9%	1.03	0.30	1.2%
(3) Raw Variables (Matlab RUSBoost)	(2.39)	0.02	0.0%	(0.87)	0.39	1.6%
(4) 4-year Sales Growth (Screen)	(2.21)	0.03	0.3%	0.41	0.68	1.3%
(6) M-Score (Beneish)	(0.70)	0.48	1.4%	(0.92)	0.36	1.0%
(7) FSD Score (Benford Law)	(0.48)	0.63	1.0%	(1.63)	0.10	1.0%
(8) Logistic - 7 Variables	0.78	0.44	1.7%	(1.38)	0.17	1.4%
(9) Logistic - 7 Variables + FSD Score	1.03	0.30	2.1%	(1.33)	0.18	1.5%

Note: Positive Z favors the in-row variable; negative values favor the bolded variable.

Another way to think about how models improve classification is to consider the next figure. Figure 8 shows the mapping of positive predictive values for three models considered in this analysis including the XGBoost model, the F-Score, and the sales growth screen. The dotted line represents the unconditional prevalence for the out-of-sample period between 1993-2012, which is 1.1 percent. The goal of a classification algorithm is to tilt this line downward for the least likely cases (the lowest deciles) and upward for the highest risk cases. Note that for the

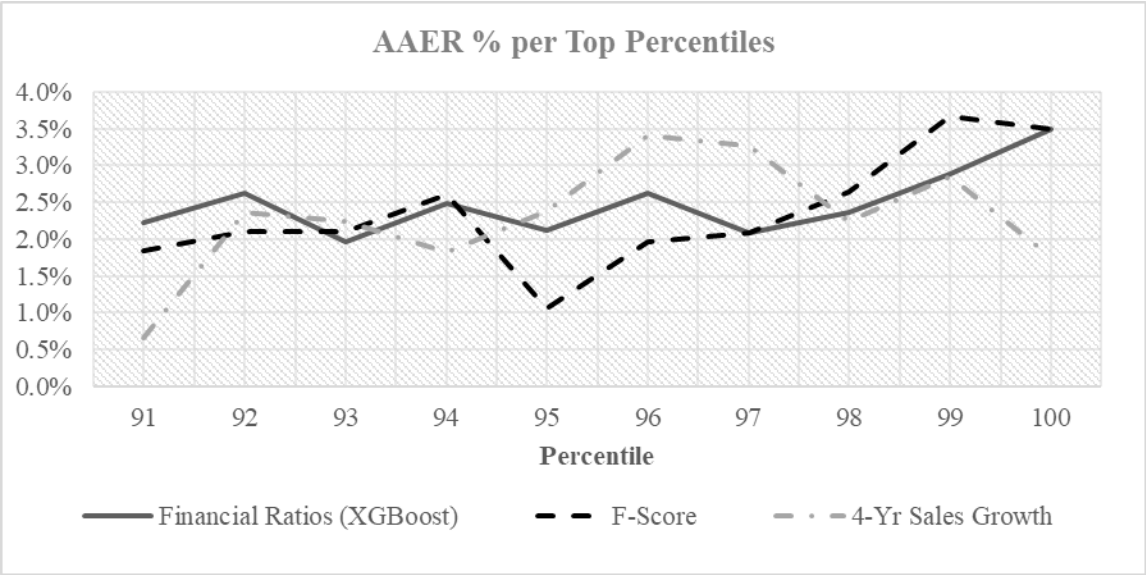
20-year period, both the Financial Ratios (XGBoost) model and the F-Score model performed similarly with a slight improvement at the top decile. The 4-Yr Sales Growth performs worse on the low end, but if measuring only the top matters, then it performs similarly to the other models.

Figure 8: Positive Predictive Values by Decile



However, when analyzing performance within the top decile, the relationships were no longer monotonic (Figure 9). It is not clear that we should be examining only the 100th percentile from this graph. For example, the F-Score does better if one were to examine only the 99th percentile. The 4-yr Sales growth peaks at the 96th percentile.

Figure 9: Positive Predictive Values within the top 10 percent

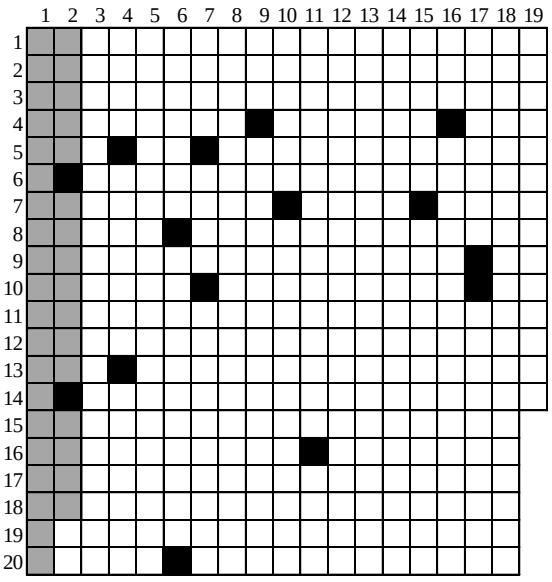


Another picture showing the difficulty of finding fraud, particularly at the top 1 percent is shown in Figure 10. This picture reports the location of the AAER firms for the year 2003 in a 20 x 19 matrix representing the top 10 percent of the probability distribution, or the top 375 observations. The grey section represents the top 1 percent or the first 38 of the observations. The rank order of the probabilities starts with the highest value in the upper left and are ranked in descending order working down the rows in the first column and continuing across columns one through nineteen. The actual AAERs are shown in black. First, this image illustrates the sparsity of hits in the sample that is predicted to be positive. Even in 2003, which is an above average year for AAER issuance, the Financial Variables (XGBoost) model captured only 15 cases leaving 360 false positives in the sample at the top 10 percent. Even at the top 1 percent only two cases would have been discoverable out of 38 companies classified positive.

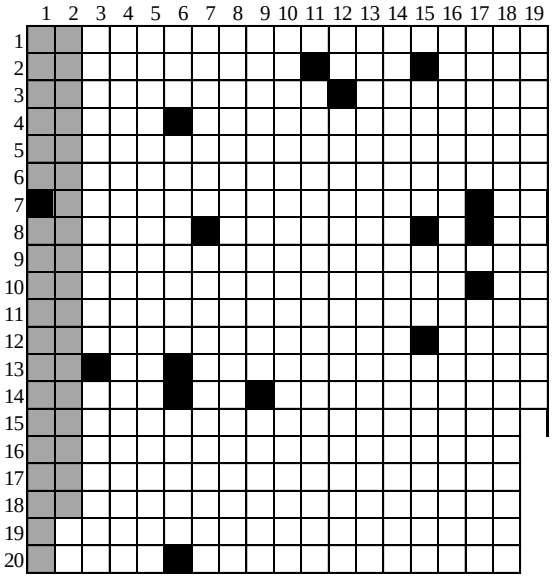
Figure 10: Location of AAERs in the Top Decile for Year 2003

Note that the order starts in the Col 1, Row 1 and goes down, then right

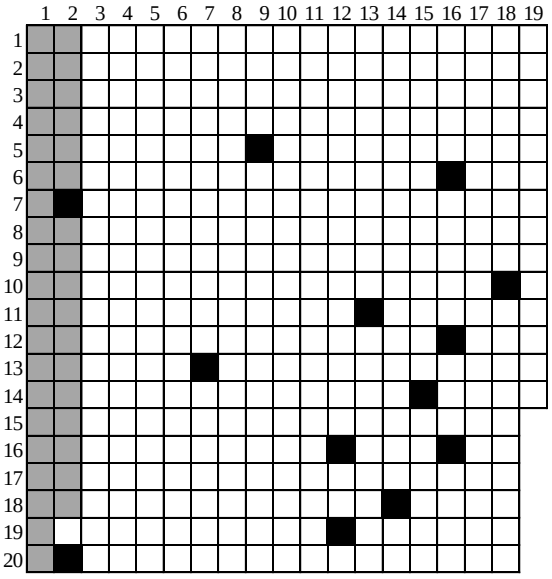
Financial Variables (XGBoost)



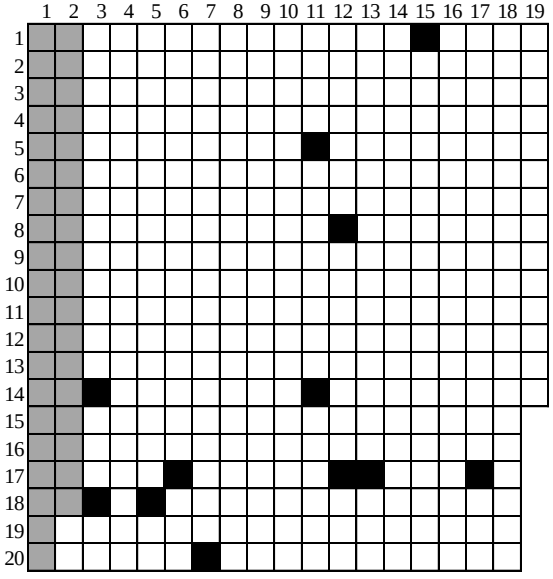
28 Raw Variables(XGBoost)



F-Score (Dechow)



4 Yr. Sales Growth



Another analysis shows how little overlap in the probability space there is between the models. Figure 11 shows a Venn Diagram analysis for the firms in the top 1 percent. Very few firms were identified by all three models. Only 6 firm-years over 20-years overlapped for predicted positives. For models that overlapped twice, there was more shared space between the advanced methods with 67 cases sharing the same space versus any two-way combination with sales growth. This made sense because the advanced statistical methods were utilizing more information relative to the univariate screen on sales growth.

Figure 11: Total Positive Firm-Year Classifications in the top 1 percent

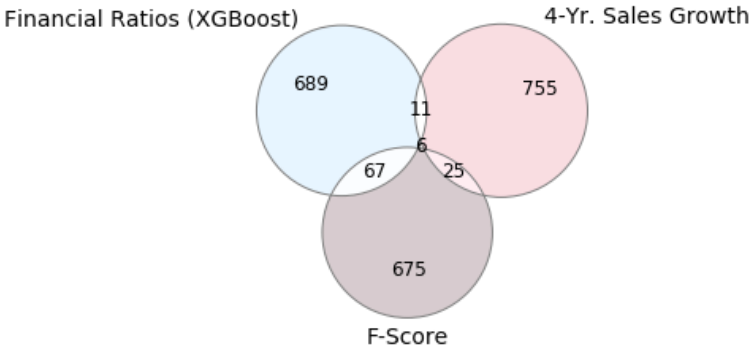
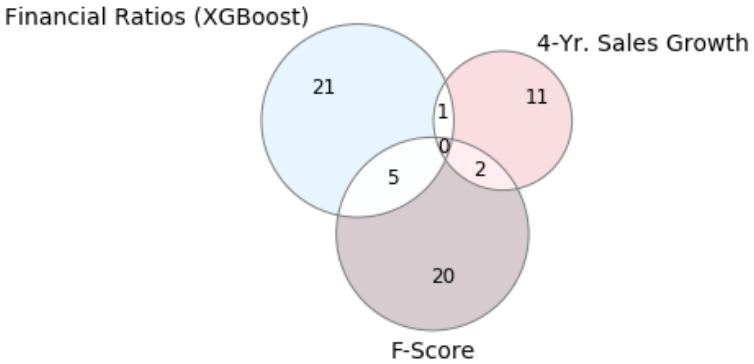


Figure 12 shows shared space for the known AAERs in the top 1 percent. Zero AAERs over 20 years were captured by all models in the top one percent. While these models attempted to detect AAERs, even the advanced statistical methods were picking up on different characteristics.

Figure 12: Total AAERs in the top 1 percent



Next, we want to repeat this analysis, but at the top 10 percent level. Figure 13 shows the Venn diagram for these firms. Even at this wider range, there was still little agreement amongst all three models, though the F-Score shared the most space for predicting positives with the XGBoost model.

Figure 13: Total Positive Firm-Year Classifications in the top 10 percent

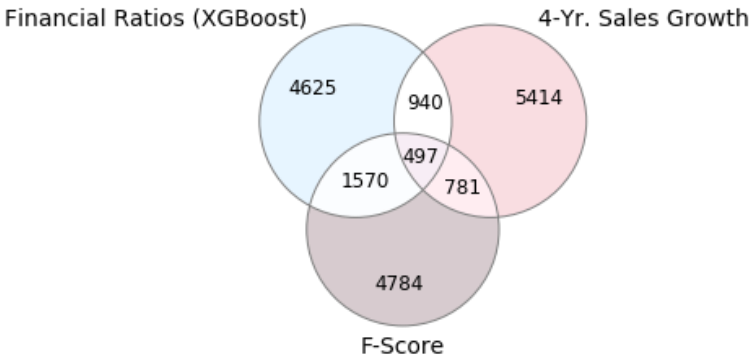
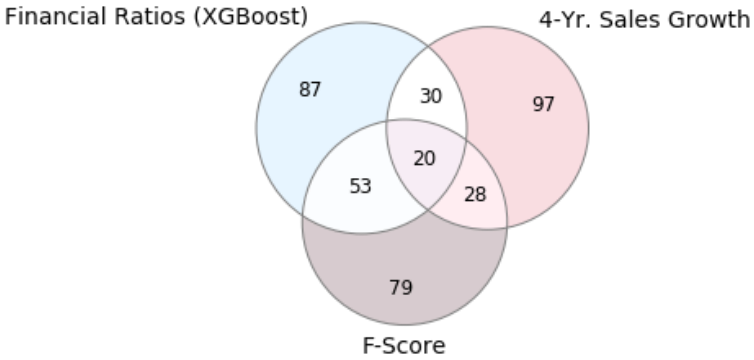


Figure 14 reports the shared space of the hits, or the AAERs in the top 10 percent. Similarly, the AAERs captured by the models shared little overlap. Only 20 AAERs over the 20-year period were predicted positive by these three models.

Figure 14: Total AAERs in the top 10 percent



The summary table for the Venn diagrams are reported in Table 9 and in Table 10. Interestingly, this analysis provides an opportunity to describe how ensemble learning works, in general. Because differing models share different probability spaces, combining their predictions could potentially reduce false positives thereby improving positive predictive values. However, as this table shows, positive predictive values were not promising for any of the combinations. The highest value was 9.1 percent for a combination from the top 1 percent of the models, but that measured 1 hit out of 11 cases predicted positive over the entire 20-year sample. Waiting 20 years to find one case is not practical. Combining all three models at the top 10 percent levels generated a 4.0 percent positive predictive value with 0.7 percent of the original sample remaining, which was roughly in line with the 3.5 percent positive predictive value for the financial ratios XGBoost model at the top one percent. Overall, an ad-hoc combination of models did not contribute significantly to the detection task.

Table 9: Venn Diagram Analysis Summary for the top 1 percent

	# of Firm-years	% of Total	AAERs	PPV
Financial Ratios (XGBoost) Only	773	1.0%	27	3.5%
F-Score Only	773	1.0%	27	3.5%
4-Yr Sales Growth Only	797	1.0%	14	1.8%
Either of the three	2,228	2.9%	60	2.7%
Include only (not the others)				
Financial Ratios (XGBoost)	689	0.9%	21	3.0%
F-Score	675	0.9%	20	3.0%
4-Yr Sales Growth	755	1.0%	11	1.5%
Include two (not the other)				
Financial Ratios (XGBoost) + F-Score	67	0.1%	5	7.5%
Financial Ratios (XGBoost) + 4-Yr Sales Growth	11	0.0%	1	9.1%
F-Score + 4-Yr Sales Growth	25	0.0%	2	8.0%
All Three	6	0.0%	0	0.0%

Table 10: Venn Diagram Analysis Summary for the top 10 percent

	# of Firm-years	% of Total	AAERs	PPV
Financial Ratios (XGBoost) Only	7,632	10.0%	190	2.5%
F-Score Only	7,632	10.0%	180	2.4%
4-Yr Sales Growth Only	7,632	10.0%	175	2.3%
Either of the three	18,611	24.4%	394	2.1%
Include only (not the others)				
Financial Ratios (XGBoost)	4,625	6.1%	87	1.9%
F-Score	4,784	6.3%	79	1.7%
4-Yr Sales Growth	5,414	7.1%	97	1.8%
Include two (not the other)				
Financial Ratios (XGBoost) + F-Score	1,570	2.1%	53	3.4%
Financial Ratios (XGBoost) + 4-Yr Sales Growth	940	1.2%	30	3.2%
F-Score + 4-Yr Sales Growth	781	1.0%	28	3.6%
All Three	497	0.7%	20	4.0%

In addition to sales growth, I tested the other variables applied in this paper from Appendix A as stand-alone univariate screens. Results for the top 15 variables are reported in Table 11. Prior to 1993, the sales growth screen was in the middle of the pack, but it was in the top position from 1993-2002 and remained at the top from 2003-2012. Overall, for the entire sample period, sales growth proved to be the best univariate screen amongst the variables tested in this paper. The empirical results supported Dr. Schilit's experienced-based research.

Table 11: Top decile positive predictive values for the univariate screens

Variable	<1993	1993-2002	2003-2012	1980-2012
Inventory to sales	1.0%	1.6%	0.6%	1.1%
Accounts receivable to sales	0.9%	1.9%	0.9%	1.2%
Abnormal % change in expenses	0.9%	2.5%	1.0%	1.5%
% Change in expenses	0.8%	2.4%	1.0%	1.4%
Fixed assets to total assets	0.8%	2.1%	1.0%	1.3%
Abnormal % change in assets	0.8%	2.4%	0.9%	1.4%
Debt-to-equity	0.8%	1.8%	0.6%	1.1%
Four-year geometric sales growth rate	**0.8%**	**3.2%**	**1.1%**	**1.7%**
% Change in assets	0.8%	2.5%	0.9%	1.4%
Change in inventory	0.8%	2.1%	0.9%	1.3%
Level of finance raised	0.8%	1.8%	0.8%	1.1%
Percentage change in cash sales	0.8%	2.7%	1.0%	1.5%
WC accruals	0.8%	2.3%	0.9%	1.3%
% Change in liabilities	0.8%	2.2%	0.8%	1.2%
Total debt to total assets	0.8%	1.5%	0.9%	1.1%
AAER Prevalence	0.3%	1.3%	0.7%	0.8%

6. Measuring Machine Learning Performance

This chapter, in my view, is the most important in this dissertation because it concerns how to properly evaluate model performance. There is a plethora of statistics to measure classification performance and this chapter explains why positive predictive value matters the most from the perspective of the fraud analyst. Prior literature most emphasized the AUC (area under the curve), which is described as the de-facto standard (Fawcett 2006) for measuring machine learning performance. Other literature cited test statistics such as specificity, sensitivity, and classification accuracy, which I will define shortly. All classification models, directly or indirectly, generate probability estimates for the predicted outcome. Unlike the logistic regression, machine learning generally does not produce probabilities that are directly interpretable, but these values can be transformed by applying a logistic regression to them. One benefit of directly examining the top of the probability distribution is that it avoids the problem of this 'model calibration' needed to obtain a directly comparable probability statistic.

The classification models do not inform at which point along the probability distribution that the values should be cut to make positive and negative predictions. In the original paper for the F-Score, the F-Score was normalized by the sample prevalence so that 1.0 equaled the unconditional expectation. Test statistics reported were measured at this point. Differing cutoff choices make comparison across published papers difficult. One advantage of the AUC is that it is agnostic to a cutoff choice since it measures across all possible cutoffs. However, does an analyst really care about how well the model performed at the low end of the probability distribution, or for the least risky cases? Alternatively, the researcher could determine an optimal cutoff by weighting the costs of the test errors (false positives and false negatives). Beneish took this approach with the expected cost of misclassification (Beneish 1999). In real-time analytics, such as preventing fraud in e-commerce transactions, decisions on whether to classify a transaction as fraudulent are weighted accordingly (e.g., shipping a product and not

getting paid for it versus losing a valid sale because the system declined the transaction). However, fraud investigation is not like e-commerce. From the point of view of the fraud investigator, incremental investigations are costly since existing resources are tied up in other investigations. Therefore, only the highest risk cases are worth examining.

To start this discussion, let us review the classification matrix as shown in Figure 15, which maps the sample into four buckets following a classification exercise.

Figure 15: Example of a Classification Matrix

		Prediction Based on Threshold (e.g. top 1%)		
		YES: AAER	**NO: AAER**	**Total**
Actual	**AAER**	True Positive (TP)	False Negative (FN)	Total AAERs
	Non-AAER	False Positive (FP)	True Negative (TN)	Total Non-AAERs
	Total	Total Predicted Positive	Total Predicted Negative	Total Sample (N)

From this classification matrix, the following statistics can be calculated:

- Classification accuracy ((TP+TN)/Total Sample)
- Sensitivity (TP/Total AAERs); Also known as the true positive rate
- Specificity (TN/Total Non-AAERs); Also known as the true negative rate
- Type I Error (FP/Total Non-AAERs); Also known as the false positive rate (1-specificity)
- Type II Error (FN/Total AAERs); Also known as the false negative rate (1-sensitivity)
- Positive Predictive Value (TP/Total Predicted Positive); Also known as "precision"

In the fraud detection literature, the F-Score paper reported classification accuracy, sensitivity, Type I and Type II errors. Bao, et al. (2020) reported AUC, NDCG@k, sensitivity, and precision. Brown, Crowley, and Elliott (2020) emphasized AUC, but also reported NDCG@k and sensitivity. AUC and NDCG@k are not statistics that can be directly measured from the classification matrix. These measures will be described momentarily. Classification accuracy takes the correct classifications (true positives and true negatives) and divides by the number in the sample. However, classification accuracy is only potentially meaningful for balanced classification tasks that have an equal number of positive and negative cases. For severely imbalanced classification involving rare events, classification accuracy is a rough approximation for specificity, which measures the true negatives. To see why, consider the following decomposition. Classification accuracy represents the number of true positives and true negatives divided by the total number in the sample.

$$Classification\ Accuracy\ = \frac{TP + TN}{N}$$

Separating the terms gives the following expression.

$$Classification\ Accuracy\ = \frac{TP}{N} + \frac{TN}{N}$$

The first term for true positives in classification accuracy can be decomposed into prevalence multiplied by sensitivity.

$$\frac{TP}{N} = Prevalence\ x\ Sensitivity$$

To see why, consider the following. TP+FN will cancel out leaving this term.

$$Prevalence\ = \frac{TP + FN}{N}$$

$$Sensitivity\ = \frac{TP}{TP + FN}$$

$$\frac{TP}{N} = \frac{TP + FN}{N} \text{ x } \frac{TP}{TP + FN}$$

The second term for true negatives can be decomposed into specificity multiplied by one minus prevalence.

$$\frac{TN}{N} = Specificity\ x\ (1 - prevalence)$$

To see why, consider the following. Similarly, TN+FP cancel out leaving this term.

$$Specificity\ = \frac{TN}{TN + FP}$$

$$1 -\ Prevalence\ = \frac{TN + FP}{N}$$

$$\frac{TN}{N} = \frac{TN}{TN + FP} \text{ x } \frac{TN + FP}{N}$$

We can know re-write classification accuracy in terms of prevalence, sensitivity, and specificity.

$$Classification\ Accuracy = [Prevalence * Sensitivity] + [(1 - Prevalence) * Specificity]$$

Classification accuracy is calculated by weighting sensitivity and specificity according to prevalence. In a balanced sample, it is the simple average of the two. Given that fraud detection involves rare events, classification accuracy will become heavily weighted by the true negatives, rather than the true positives. For samples with extremely low prevalence rates, classification accuracy essentially measures specificity.

Furthermore, an uninformative rule for rare events could be applied to maximize classification accuracy. For rare events, the strategy would be to classify all cases in the negative. In this case sensitivity would be zero as no events would be captured, but specificity would be 100 percent as all negatives would be classified correctly. Therefore, for a rare event occurring 0.5 percent of the time, classification accuracy would equal 99.5 percent. Occasionally, specificity is reported by its complement, which is the false positive rate (Type I error) and is measured by one minus specificity. Likewise, sensitivity can be reported by its complement, which is the false negative rate (Type II error), or one minus sensitivity.

In his seminal work on bankruptcy risk (Beaver 1966), Beaver wrote a section entitled "Likelihood Ratios." He noted that likelihood ratios are "essentially a Bayesian approach" and that the "posterior probability is the probability of failure after the ratio analysis." To understand what the posterior probability is in this context, it is worth re-examining the eponymous Bayesian formula.

$$P(A|B) = \frac{P(A) \text{ x } P(B|A)}{P(B)}$$

This formula states that the prior probability P(A) can be transformed into a posterior probability P(A|B) through knowing P(B) and P(B|A). In the context of the AAER sample, we can think about the unconditional probability of an AAER as the prior P(A) while the probability of a positive classification can be written as the P(B). Thus, we can rewrite this equation accordingly.

$$P(AAER|Positive) = \frac{P(AAER) \text{ x } P(Positive\ |AAER)}{P(Positive)}$$

This equation says that the probability of an AAER given a positive classification is equal to the unconditional probability of an AAER (prevalence) multiplied by the probability of a positive classification given that it is a true AAER (sensitivity). The numerator represents the true positive proportion. The denominator can be decomposed into the true and false positives.

$$P(AAER|Positive) = \frac{P(AAER) \text{ x } P(Positive\ |AAER)}{P(AAER) \text{ x } P(Positive|AAER) + P(No\ AAER) \text{ x } P(Positive|No\ AAER)}$$

The true positive term repeats from the numerator. The false positive term is the proportion of negatives that are misclassified in the positive.

Finally, we can rewrite these terms to values that are typically reported with classification modeling including sensitivity and specificity. The posterior probability is called positive predictive value (PPV), which is a function of three parameters.

$$PPV = \frac{Prevalence * Sensitivity}{Prevalence * Sensitivity + \ (1 - Prevalence) * (False\ Positive\ Rate)}$$

Before continuing with the positive predictive value analysis, I want to describe to additional measures including NDCG@k and AUC. NDCG@k is a recent addition to the fraud detection literature and comes from the information retrieval literature for measuring "accuracy at the top" (Boyd, et al 2012). It is worth illustrating how this metric works by example. Formally, the function is defined in the following terms.

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

$$NDCG@k = \frac{DCG@k}{Ideal\ DCG@k}$$

$$rel_i = Relevance\ Score\ in\ rank_i$$

While relevance offers a way to weight webpages, for fraud detection this value simplifies to the binary indicator for fraud in the dependent variable. Therefore, the numerator $2^{rel_i} - 1$ simplifies to zero or one matching the AAER value.

Figure 16 shows a toy example for NDCG@k where there are 10 observations and 3 positive cases. Column 1 ranks the probability outcome in rank order from 1 to 10. Column 2 identifies the AAERs, which are shown to be in the second, fourth and fifth rank. DCG@k sums the discounted cumulative gain values applying the discount factor of $1/\log_2(\mathrm{i}+1)$ as shown in

the formula above. To normalize this value, an ideal ranking must be computed. To do so, the cumulative gain column is sorted in descending order. The final column discounts this ranking applying the same factor as before. With the DCG@k and the Ideal DCG@k values, the normalized value NDCG@k can be calculated. In this example, the value is 0.68. A perfect classifier would have the value of 1.0. However, real NDCG@k values for fraud detection are reported much lower in recent literature as the ideal rank column fills up with AAERs from the entire sample. For more theoretical background on NDCG@k, see Wang, et al. (2013).

Figure 16: Normalized Discounted Cumulative Gain Example (NDCG@k)

$Rank_i$	Cumulative Gain	Discount Factor[1]	Discount Cumulative Gain (DCG@k)	Ideal Ranking	Ideal DCG@k
(1)	(2)	(3)	(2) x (3)	(4)	(3)*(4)
1	0	1.00	0.00	**1**	1.00
2	**1**	0.63	0.63	**1**	0.63
3	0	0.50	0.00	**1**	0.50
4	**1**	0.43	0.43	0	0.00
5	**1**	0.39	0.39	0	0.00
6	0	0.36	0.00	0	0.00
7	0	0.33	0.00	0	0.00
8	0	0.32	0.00	0	0.00
9	0	0.30	0.00	0	0.00
10	0	0.29	0.00	0	0.00
Total	**3**		**1.45**	**3**	**2.13**

1. Discount Factor is $1/\log_2(rank_i+1)$

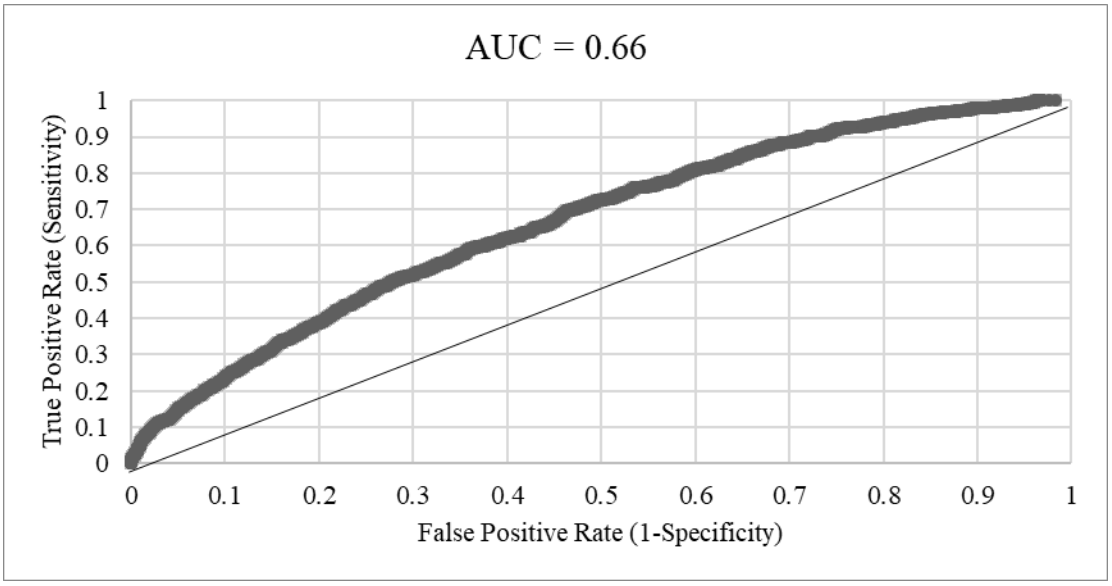
NDCG@k	**0.68**

Since our use of NDCG@k involves binary relevance, the incremental informativeness of this metric over positive predictive value is lower relative to cases where relevance weighting matters. Returning to the previous figure, Column 2 shows Cumulative Gain, which sums the total number of true positives above the cutoff threshold *k*. This value divided by the total observations in the subgroup is 30 percent. This figure is the same as positive predictive value. NDCG@k simply adds a discount factor to the rank position. Given the loss of interpretability with the use of this measure and the fact that Figure 10 showed how scattered these rare events were, the incremental informativeness of NDCG@k is not much greater than positive predictive value.

Regarding the AUC, this measure maps the area under the curve with sensitivity on the y-axis and the false positive rate (1-specificity) on the x-axis for all cutoff points. The upside is that it provides a summary statistic that does not depend on a cutoff point. The downside is interpretability, and it is particularly problematic with rare event problems (Saito and

Rehmsmeier 2015). For example, the AUC for the F-Score tested in this paper is shown in Figure 17**.**

Figure 17: F-Score Area Under the Curve



The AUC measures the entire area below the curve on the chart. If the curve went up in a straight line to where sensitivity is one and the false positive rate is zero, the AUC value would equal 1.0 which would imply a perfect classifier. This makes sense because all AAERs would be detected at 100 percent specificity with zero false positives. The minimum AUC possible is 0.5, which is represented by the diagonal. Where sensitivity equals the false positive rate, the classifier contributes no information and is not different from a random guess in the sample. To see why, let us substitute the false positive rate with sensitivity from the positive predictive value formula since the two would now equal. Since sensitivity is now multiplied to every term, it cancels out leaving only prevalence in the numerator and denominator. The denominator reduces to one leaving behind prevalence only, which is the unconditional expectation.

$$\frac{Prevalence * \boldsymbol{Sensitivity}}{Prevalence * \boldsymbol{Sensitivity} + \ (1 - Prevalence) * (\boldsymbol{Sensitivity})}$$

$$\frac{Prevalence}{Prevalence + \ (1 - Prevalence)}$$

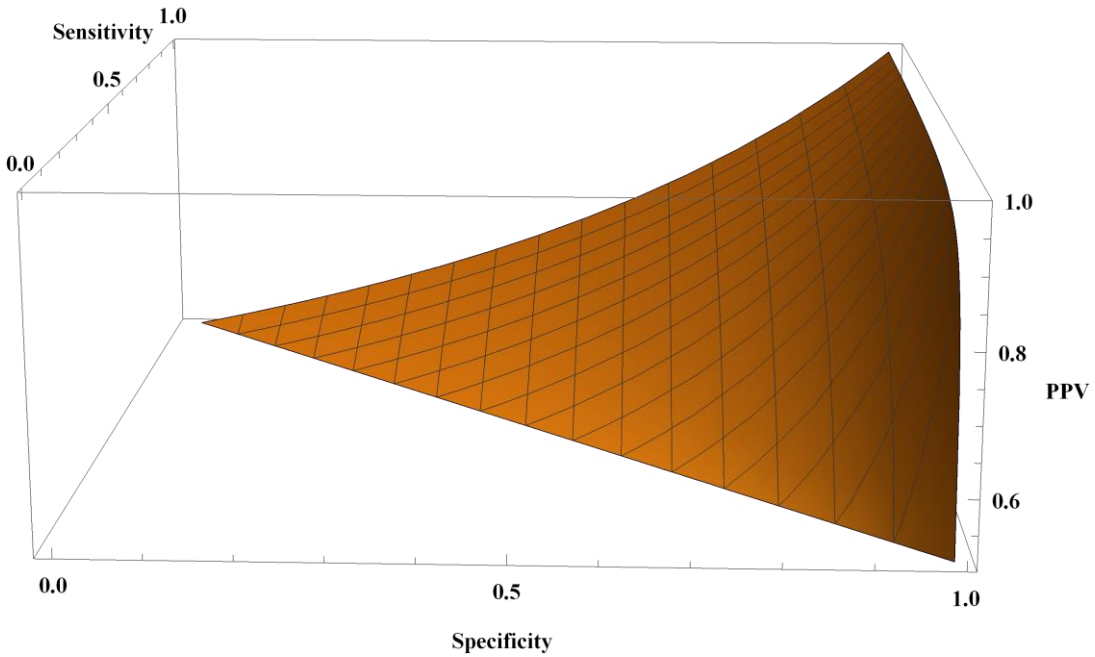
$$\frac{Prevalence}{1}$$

The reason why the AUC cannot be below the diagonal is that the decision rule where the classifier does worse than the unconditional expectation could be flipped. This is like how Wall Street veterans poke fun at terrible analysts because those that are consistently wrong give opportunities to trade opposite their recommendations. In summary, sensitivity must be greater than the false positive rate for a classifier to improve the odds of detection.

The missing third dimension to the AUC is positive predictive value, which can be calculated by considering prevalence. For rare events, this enters the positive predictive value equation in a nonlinear way. In his seminal work on firm failure, Beaver made an interesting observation. He observed that the posterior odds for firm failure would be affected "by the probability of failure for this sample (i.e., 0.50), which is vastly different from the probability for all firms in the economy (i.e., less than 0.01)" (Beaver 1966). This is a critical point to this analysis. Interestingly, Altman, following Beaver, published the famous Z-Score paper (Altman 1968) that applied a balanced (50:50) sample of bankrupt and non-bankrupt firms reporting a 95 percent classification accuracy and made no mention of odds ratios or positive predictive values. As shown earlier, classification accuracy is not the same as the posterior odds, particularly for imbalanced real-world situations. Beaver concluded his study by saying that developing a multivariate model was not encouraging because "the best single ratio appears to predict about as well as the multi-ratio models". Interestingly, the previous chapter found that a univariate screen on sales growth performed about as well at the top decile of risk relative to advanced statistical methodologies. While "significant" improvements can be made in various test statistics, when it comes to rare events, the needle does not move that much in terms of posterior odds.

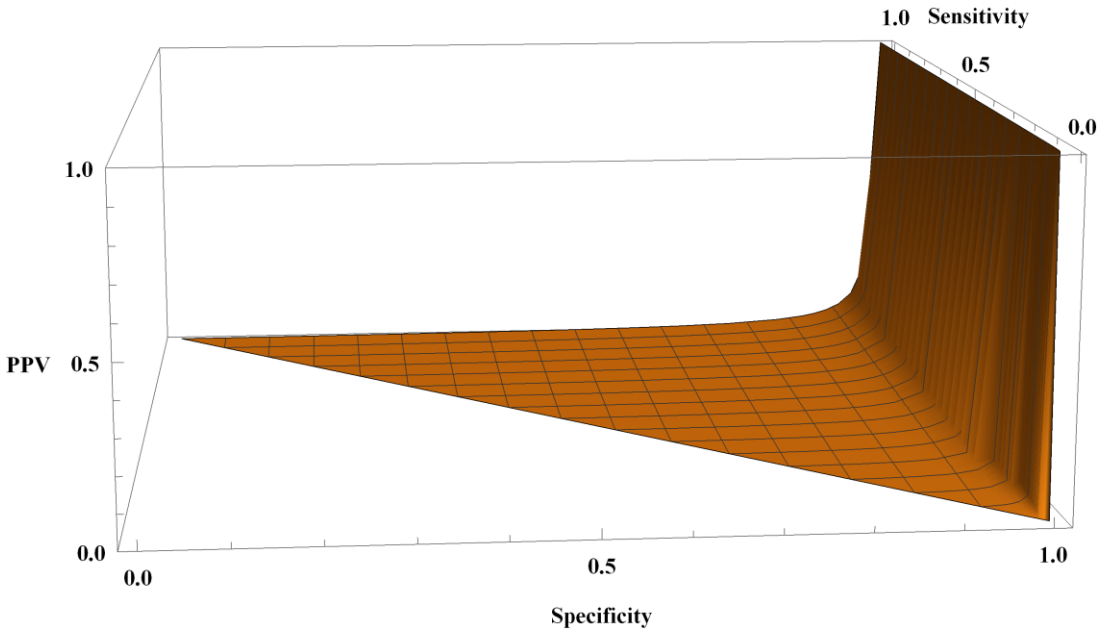
To illustrate this point, Figure 18 shows how positive predictive value changes along sensitivity and false positive rates for a balanced (50:50) sample.

Figure 18: Positive Predictive Values for a Balanced Sample



Observe that the diagonal is the same from the AUC two-dimensional chart. In this example, the rotation of this graph shows specificity instead of the false positive rate (1-specificity), but they represent the same concept. Positive predictive value is steeper at the higher values for specificity. In some sense, this graph shows why false positive rates matter so much in fraud detection. The odds of detecting fraud are maximized in the region where false positives are also the lowest. Figure 19 reduces the prevalence to a value reflecting a rare sample. In this case, the figure is 0.5 percent, which is like the real-world prevalence of AAERs.

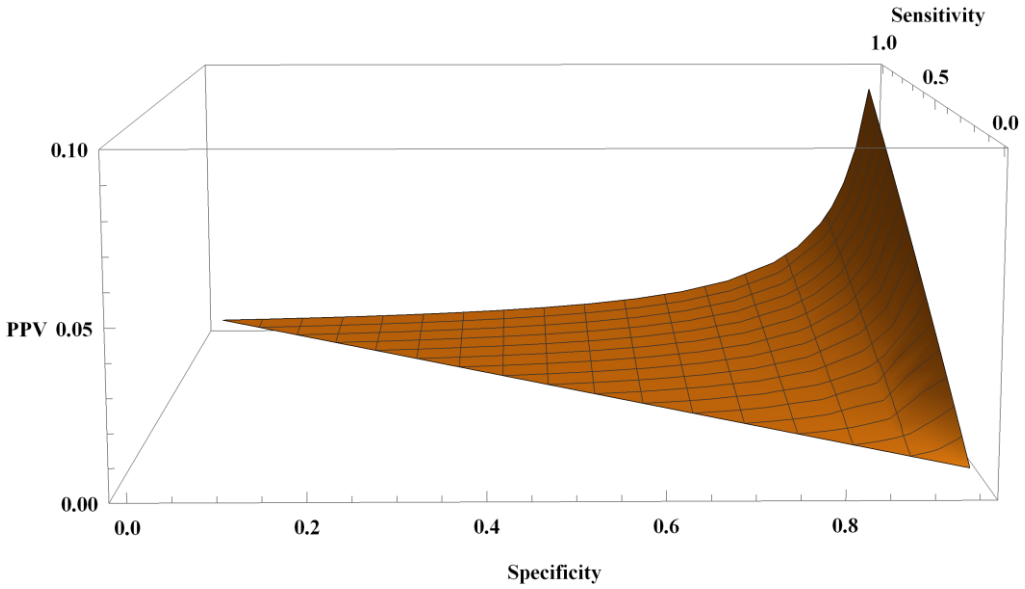
Figure 19: Positive Predictive Values for a Rare Sample



This graph is nearly unreadable at this scale. What is clear, however, is that positive predictive values only begin to rise at extreme values of specificity. To analyze this graph better, I impose real-world constraints on both sensitivity and specificity limiting the low end for sensitivity to 5 percent. The aggressiveness of this value will become clear shortly. To provide some detail on this value, for the twenty-year sample, there are 825 AAERs, or about 40 per year on average. If sensitivity is 5 percent, then there would be 2 AAERs per year discovered. Specificity was chosen to be at most 95 percent because it is an aggressive assumption given the sensitivity constraint.

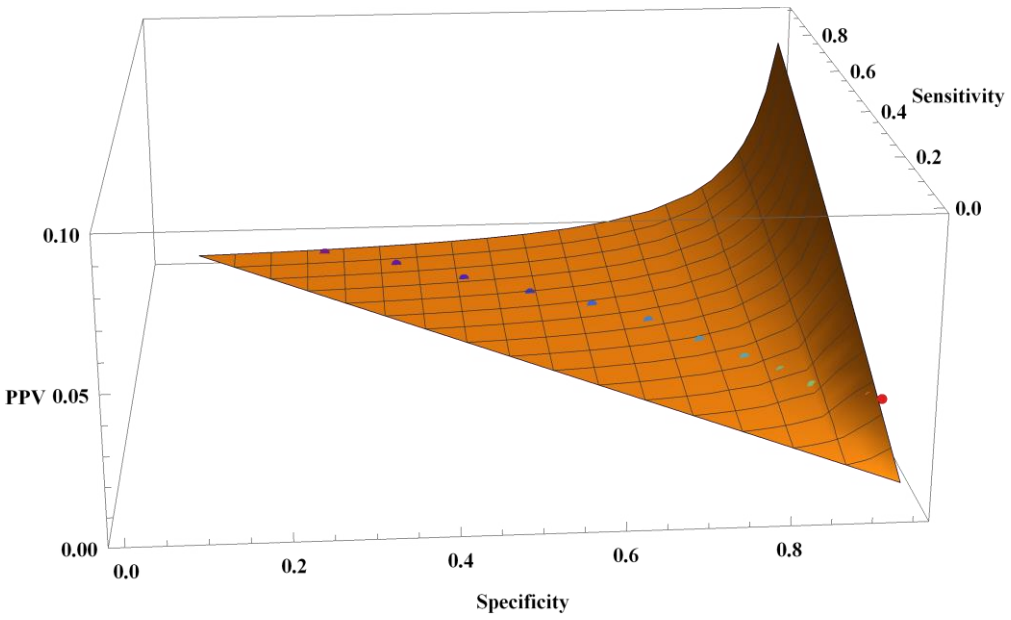
Given these constraints, Figure 20 illustrates that the maximum positive predictive value obtainable is 8.7 percent, a point where the model captures 95 percent of known cases and experiences a low false positive rate of 5 percent.

Figure 20: Positive Predictive Values for a Rare Sample with Constraints



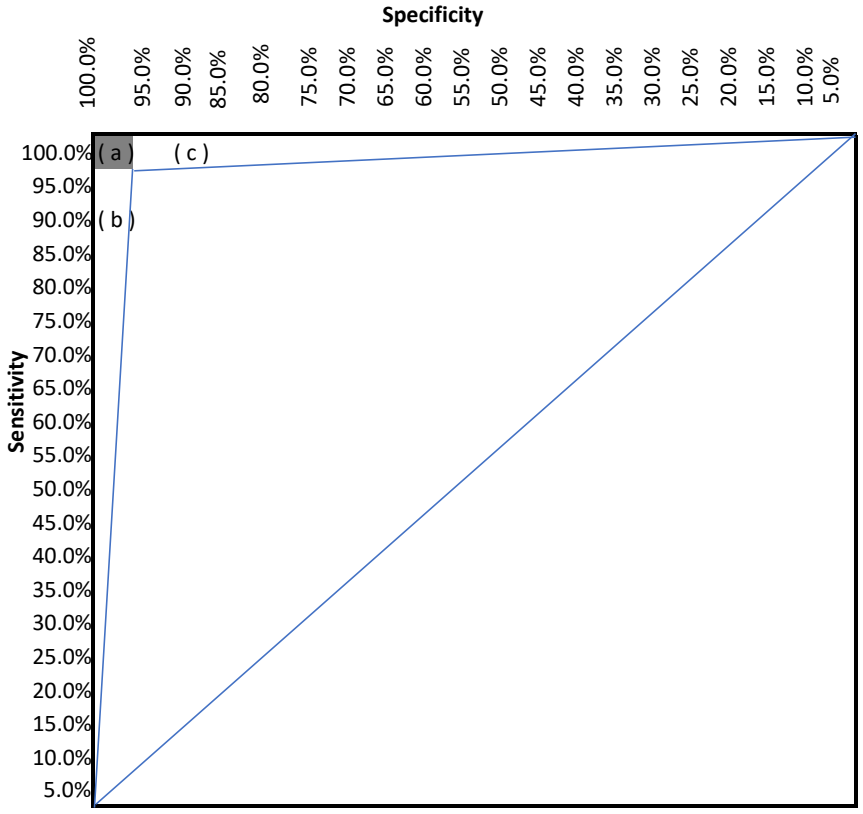
In addition to this graph, we can trace the AUC curve to find where positive predictive value is maximized. Figure 21 shows this projection and the positive predictive value was maximized at the specificity constraint. The problem is that very few of the AAERs will be captured as sensitivities are low at this point.

Figure 21: Mapping the F-Score AUC to the Graph



Finally, the aggressiveness of these assumptions becomes clear in Figure 22. In fact, these two constraints at 95 percent would imply an AUC close to 0.95 while current models in the published literature are in the 0.70 range.

Figure 22: Hypothetical AUC with Constraints



Value of Grey Square (a) = 0.05*0.05	0.25%
Value of Triangles (b) and (c) = 0.95*0.05*0.5*2	4.75%
Total Area	5.00%
AUC = 1- Total Area	95.00%

In summary, since fraud detection involves a rare event, it is mathematically difficult to move posterior probabilities beyond the single digits for this research question. The top models applying financial statement variables are far away from this optimistic scenario. Fraud detection is difficult because the event is rare and classification models remain far too noisy to automate the fraud detection task.

7. Additional Analyses

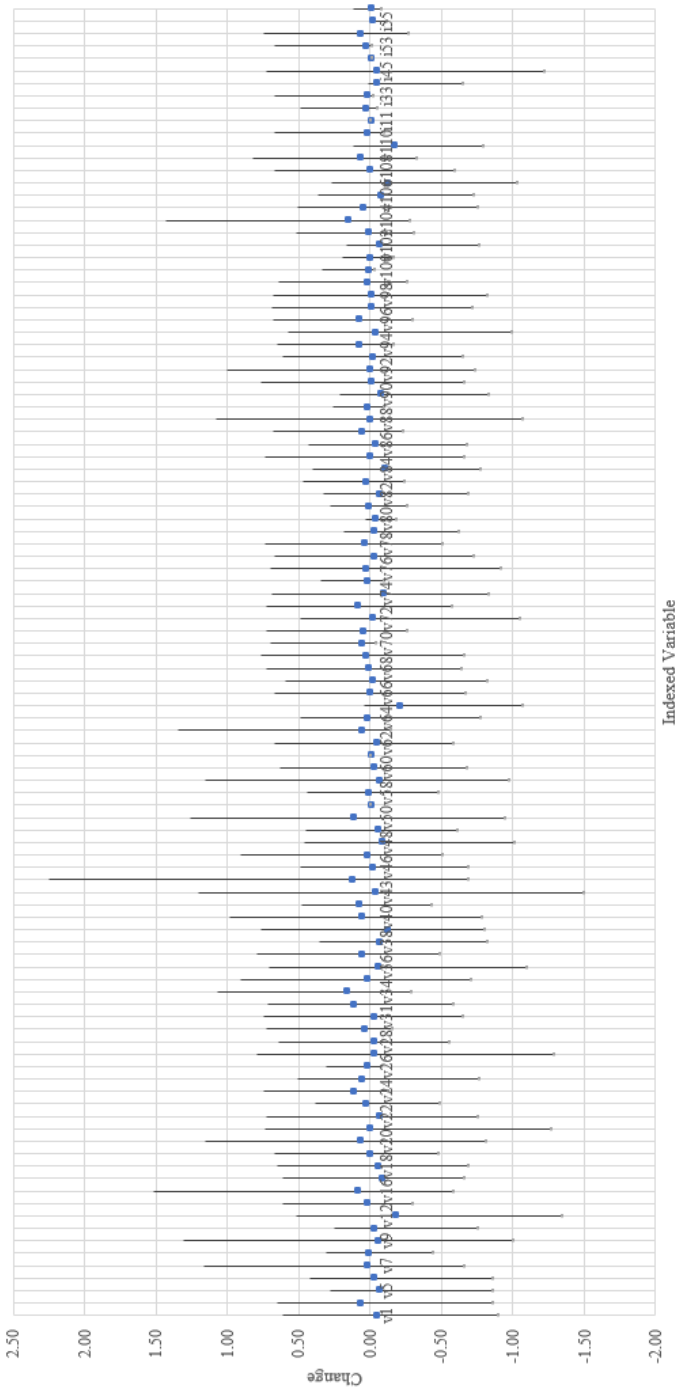
This chapter includes additional analyses that did not fit neatly within the other chapters but are worth discussing with regards to the topic of accounting fraud. While fraud detection was the focus of this dissertation where the goal was to maximize the odds of detecting fraud, other topics of interest include variable importance, which provides a rank order to the variables applied in machine learning algorithms. While the goal of examining variable importance might be to identify which variables most affect the outcome of the prediction exercise, these measures of importance do not provide much insight into their causal nature. The standard econometric toolkit is best suited to answer these types of questions.

In addition to variable importance, I also performed a returns test for select models to examine whether a trading strategy would be profitable. Unfortunately, results were not promising. Earlier years showed positive strategies, but with increasing efficiency in the capital markets, these profits have been competed away. Earlier in this dissertation, I introduced Professor Narayanan and his "How to Recognize AI Snake Oil" presentation where he suggested that A.I. does not perform substantially better than a manual scoring rule when used for predicting social outcomes (Narayanan 2020). Since this is an empirical question, I created a manual score with the seven F-Score variables to see how well they performed relative to the more advanced statistical techniques and results were interesting. The last analysis in this chapter takes a closer look at Benford's law from Amiram, et al. (2015).

For variable importance, a technique known as permutation importance was applied. Permutation importance analyzes variable importance by taking the column vector for each variable after training each model and randomizing it so that it becomes noise. Then the out-of-sample test was reapplied to measure the change to positive predictive value at the top decile in the probability distribution. These values were scaled by in-year prevalence. The maximum, minimum, and averages are reported for the 20-year period. What can be immediately observed in Figure 23 is the volatility in importance of the variables, with some that cross over the zero line implying that they worsen model performance in some years.

Figure 23: Variable Importance for the Machine Learning Models

The y-axis measures change in positive predictive value for the top 10 percent of fraud risk scaled by in-year prevalence for the XGBoost model.



More advanced techniques such as partial dependence plots are available, but Hastie recently wrote that these plots "should not replace a randomized controlled experiment or a carefully designed observational study to establish casual relationships." (Zhao & Hastie 2019). With the large set of variables, not much can be ascertained as they also suffer from the same issues in econometric models including collinearity and multicollinearity. Pearl spoke to the causal revolution coming to machine learning as prediction tasks are increasingly asking cause and effect type questions (Pearl and Mackenzie 2018). In contrast to Varian's call to for econometricians to add machine learning to their toolkit, Pearl is calling on computer scientists to add causal inference to theirs.

Regarding trading strategies, the evidence showed that they may have worked in the past, but these opportunities have largely been traded away (Table 12). Twelve month holding periods started after the fourth month following the end of the fiscal year. Returns were sourced from CRSP inclusive of dividends. Delisting returns were included, but proceeds were held as cash through the twelve-month window. Group-adjusted returns represent the difference between the raw returns and a matched portfolio return based on quintiles of size, market-to-book, and price-to-earnings ratios.

Table 12: 12 Month Holding Period Returns

Years 1993-2002

Group-Adjusted Returns

Deciles	XGBoost (All)	F-Score	SalesG	XGBoost (28 Raw Vars)
1	1.8%	1.1%	0.1%	1.0%
2	1.1%	1.0%	0.3%	0.5%
3	1.6%	1.4%	(0.5%)	0.8%
4	0.9%	2.7%	(2.8%)	0.2%
5	0.2%	1.5%	(1.8%)	0.3%
6	0.1%	0.8%	(5.2%)	(1.1%)
7	(0.4%)	1.1%	0.4%	0.1%
8	(0.6%)	(0.2%)	3.6%	(0.1%)
9	(0.8%)	(3.9%)	3.0%	1.1%
10	(3.8%)	(5.3%)	2.6%	(2.6%)
1 minus 10	**5.6%**	**6.4%**	**(2.5%)**	**3.6%**

Raw Returns

Deciles	XGBoost (All)	F-Score	SalesG	XGBoost (28 Raw Vars)
1	15.6%	10.8%	11.1%	13.8%
2	13.5%	9.3%	12.3%	12.5%
3	14.4%	10.3%	9.8%	10.8%
4	12.5%	13.6%	7.4%	10.2%
5	10.2%	12.3%	7.9%	10.7%
6	10.4%	11.7%	3.2%	8.5%
7	10.0%	11.8%	8.1%	8.4%
8	7.1%	10.2%	13.2%	8.5%
9	5.6%	6.3%	13.0%	9.3%
10	0.4%	3.1%	12.9%	6.7%
1 minus 10	**15.2%**	**7.7%**	**(1.8%)**	**7.1%**

Years 2003-2012

Group-Adjusted Returns

Deciles	XGBoost (All)	F-Score	SalesG	XGBoost (28 Raw Vars)
1	0.7%	0.9%	(0.6%)	0.9%
2	(0.5%)	0.3%	0.4%	0.5%
3	(0.0%)	(0.2%)	1.5%	(0.1%)
4	0.6%	(0.1%)	0.0%	0.1%
5	0.4%	0.9%	0.5%	(0.5%)
6	0.1%	(0.6%)	(1.5%)	0.2%
7	0.0%	(0.3%)	0.1%	(1.1%)
8	(0.1%)	1.0%	(0.9%)	0.7%
9	(0.4%)	0.1%	(0.2%)	(0.5%)
10	(0.6%)	(1.9%)	0.7%	0.0%
1 minus 10	1.3%	2.8%	(1.4%)	0.9%

Raw Returns

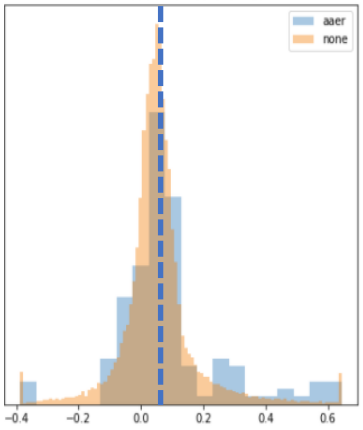
Deciles	XGBoost (All)	F-Score	SalesG	XGBoost (28 Raw Vars)
1	11.1%	10.0%	9.8%	10.2%
2	9.6%	9.6%	11.7%	11.0%
3	10.1%	9.8%	12.8%	10.0%
4	10.7%	10.6%	11.8%	10.2%
5	11.0%	11.3%	11.3%	10.4%
6	9.8%	9.8%	8.5%	11.1%
7	10.1%	11.5%	10.1%	8.9%
8	10.0%	11.7%	8.0%	10.1%
9	10.0%	10.9%	8.7%	10.6%
10	10.6%	7.8%	10.4%	10.5%
1 minus 10	**0.4%**	**2.2%**	**(0.6%)**	**(0.4%)**

For the manual scoring analysis inspired by Narayanan's presentation, I first produced histogram plots of the distribution of the six continuous variables from the F-Score for cases that experienced AAERs and for the remaining sample. These histograms are shown in Figure 24. I created the following rules based on drawing a line based on a simple visual inspection. The last variable in the F-Score remains the same because it is an indicator variable for security issuance where those that issue securities are more likely to experience an AAER relative to those that do not.

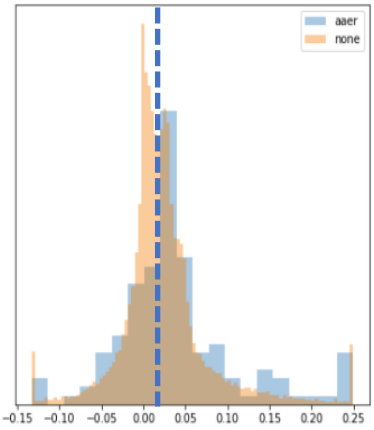
- Indicator 1: If RSST Accruals (v19) were greater than 0.05, then 1, otherwise 0.
- Indicator 2: If change in receivables (v8) is greater than 0.02, then 1, otherwise 0.
- Indicator 3: If change in inventories (v6) is greater than 0.03, then 1, otherwise 0.
- Indicator 4: If soft assets (v20) is greater than 0.55, then 1, otherwise 0.
- Indicator 5: If change in cash sales (v18) is greater than 0.3, then 1, otherwise 0.
- Indicator 6: If change in cash sales (v8) is less than 0.05, then 1, otherwise 0
- Indicator 7: Same as issuance (i2)

Figure 24: Histograms for F-Score Variables

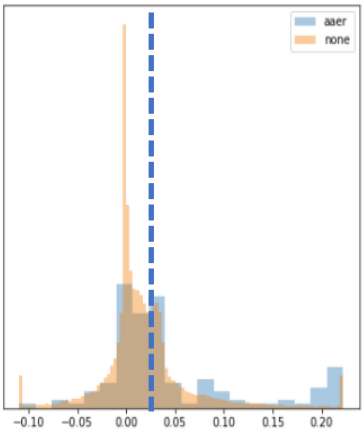
RSST Accruals (v19)



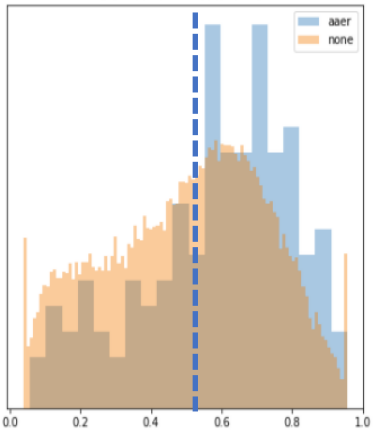
Change in Receivables (v8)



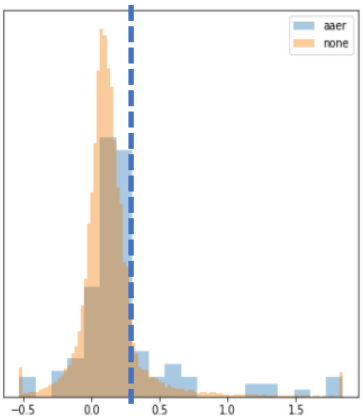
Change in Inventory (v6)



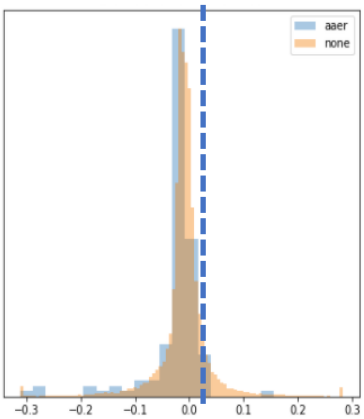
Soft Assets (v20)



RSST Accruals (v19)



Change in Receivables (v8)



The results of this scoring method are shown in Figure 25. In this score, the zero had the lowest risk and 7 had the highest risk for fraud. For a cutoff score of 5 or higher, the positive predictive value was 2.11 percent, which included about 12 percent of the total sample from 1993-2002. For 2003-2012, the positive predictive value was 1.29 percent for roughly 7 percent of the sample. While the samples were not directly comparable to the top 10 percent levels given the lack of granularity with this manual score, these results were remarkably comparable to the statistical models reported in Table 5.

Figure 25: Manual Score Method

1993-2002 Sample

Simple Score	Total Obs.	% of Total	AAERs	PPV
7	4	0.0%	0	0.00%
6	594	1.4%	21	3.54%
5	4,418	10.6%	85	1.92%
4	9,149	22.0%	145	1.58%
3	13,621	32.7%	168	1.23%
2	9,934	23.9%	89	0.90%
1	3,571	8.6%	28	0.78%
0	350	0.8%	2	0.57%
Total	41,641	100.0%	538	1.29%
Score >=5	5,016	12.0%	106	2.11%
PPV of Score>=5 / Prevalence				1.6x

2003-2012 Sample

Simple Score	Total Obs.	% of Total	AAERs	PPV
7	0	0.0%	0	0.00%
6	207	0.6%	3	1.45%
5	2,192	6.3%	28	1.28%
4	6,656	19.2%	93	1.40%
3	12,380	35.8%	97	0.78%
2	9,616	27.8%	51	0.53%
1	3,249	9.4%	15	0.46%
0	288	0.8%	0	0.00%
Total	34,588	100.0%	287	0.83%
Score >=5	2,399	6.9%	31	1.29%
PPV of Score>=5 / Prevalence				1.6x

Finally, while Amiram, et al. said that their model applying Benford's law predicted AAERs, they did not quantify this result, nor did they report classification statistics typical in classification literature. The only analysis provided was the output from a logistic regression. One slight modification was that Amiram, et al. applied a slightly different model specification where they encoded the dependent variable only for the initial year of the AAER. Model 1 in Table 10 (of their original paper) reported the logistic regression estimated for their sample of AAERs with the FSD Score controlling for other measures of accrual quality and fraud risk including the F-Score. The coefficient reported in model 1 for the FSD Score was the largest coefficient of any of the other variables at -40.691 with three stars showing its significance. However, did it really move the needle in terms of probabilities? The results from the horse race proved poor overall. However, a careful reading of the original paper could have determined the back-of-the-envelope effect, which I describe below.

This calculation is only an approximation because the data was not provided in exact detail in the paper, but it gives a directional magnitude for the effect of a deviation in Benford's law. The descriptive statistics are based on the full sample of 43,332 observations for 2001-2011, and their logistic regression applied only 27,805 observations. The results of this analysis are shown in Table 13. While FSD_Score was shown at the top of the logistic regression in the paper with the largest coefficient relative to the other variables, the evaluations in log odds reveals that the absolute probability change was quite small, or around 0.08 percent for an interquartile change in the FSD_Score. Note also that an increase in divergence reduced the likelihood of an AAER, which is different from what the authors originally thought the relationship should be. They attributed this finding to companies running out of space to manipulate their earnings. Regardless, the magnitude of the deviation in Benford's law to the detection of fraud is quite small, and in the previous horserace, shown to be of little significance to the results.

Table 13: Evaluation of Log Odds from the FSD Score Logistic Regression (Amiram, et al. 2020)

	Model 1 Coefficients from Amiram, et al.	Q1 Value for FSD_Score; else, at means.	Q3 Value for FSD_Score; else, at means.	Calculate Log Odds	Calculate Log Odds
	(1)	(2)	(3)	(1) x (2)	(1) x (3)
FSD_Score	-40.691	0.023	0.035	-0.952	-1.428
ABS_JONES_RESID	-1.078	0.184	0.184	-0.198	-0.198
STD_DD_RESID	0.011	0.123	0.123	0.001	0.001
MANIPULATOR	0.122	0.143	0.143	0.017	0.017
F_SCORE	1.980	0.401	0.401	0.793	0.793
ABS_WCACC	-1.233	0.054	0.054	-0.067	-0.067
ABS_RSST	0.401	0.138	0.138	0.055	0.055
CH_CS	0.004	0.146	0.146	0.001	0.001
CH_ROA	1.339	-0.002	-0.002	-0.003	-0.003
SOFT_ASSETS	-0.121	0.545	0.545	-0.066	-0.066
ISSUE	-0.341	0.915	0.915	-0.312	-0.312
MTB	0.166	1.360	1.360	0.226	0.226
AT	0.000	3228.380	3228.380	0.000	0.000
Constant	-5.686			-5.686	-5.686
Sum of Log Odds				-6.189	-6.666
Odds Ratio - exp(log odds)				0.002	0.001
Probability (odds ratio / 1+odds ratio)				0.205%	0.127%
Change in Probability				0.077%	

8. Conclusion

Throughout these chapters, I hope to have conveyed to the reader the stark difficulty of detecting fraud through statistical means alone. A vast literature contributed significantly to understanding how financial misconduct occurs and through which channels these might be observable to outsiders. Despite this work, the evidence in this study showed how difficult detection remains. Machine learning did not appear to provide a black box solution to this task. The horse race presented in Chapter 5 provided evidence that raw variables do not perform better than financial ratios. AuditAnalytics makes both the M-Score (Beneish) and Benford's law analysis commercially available and the results from this analysis suggested that the F-Score would be of better use to their clients. At the top decile of risk, a univariate screen on sales growth performed about as well as the F-Score and machine learning models. When the kitchen sink of financial variables was thrown into the XGBoost model, results improved, but slightly in absolute terms. Given the complexity of implementing a machine learning approach, the logit-based measures may be easier to implement overall. Chapter 6 analyzed classification metrics in detail and showed why positive predictive value matters the most from the perspective of the fraud investigator. The analysis showed that the third dimension of an AUC is positive predictive value which is driven in a nonlinear way based on the prevalence of the underlying sample. This value is maximized where false positive rates are the lowest, which is at the top of the probability distribution.

This is why the false positive rate matters so much. However, it comes at the cost of the sensitivity. Therefore, few actual cases will be captured at the top of the probability distribution. Even when maximizing posterior probabilities at the top of the probability distribution, results remained quite low. Finally, Chapter 7 included a few oddities that did not quite fit in the flow of the other chapters. Variable importance from machine learning offers almost no real information to understanding the why behind it. As Hastie acknowledged, this is the work of well-designed causal studies. One key insight is just how noisy variables proved to be including the counterintuitive observation that some variables hurt the detection task depending on the year. Of course, this is not knowable in advance and variables were included because they improved the model on average. A returns analysis showed that advanced models would have

outperformed during a time when few had access to advanced models. However, by the 2000s, published academic research and increased use of machine learning models competed away these opportunities. Today, there is little benefit from trading based on the output of these models analyzed in this paper. The overall message of this dissertation follows Professor Schilit's advice, who wrote the book on financial fraud. Get your hands dirty and understand the why. Shoe-leather, and not machine learning, is still required for this work.

References

Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), pp.589-609.

Amiram, D., Bozanic, Z. and Rouen, E., 2015. Financial statement errors: Evidence from the distributional properties of financial statement numbers. *Review of accounting studies*, *20*(4), pp.1540-1593.

Amiram, D., Bozanic, Z., Cox, J.D., Dupont, Q., Karpoff, J.M. and Sloan, R., 2018. Financial reporting fraud and other forms of misconduct: a multidisciplinary review of the literature. Review of Accounting Studies, 23(2), pp.732-783.

Bao, Y., Ke, B., Li, B., Yu, Y.J. and Zhang, J., 2020. Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach. *Journal of Accounting Research*, *58*(1), pp.199-235.

Bao, Y., Ke, B., Li, B., Yu, Y.J. and Zhang, J., 2021. A Response to" Critique of an Article on Machine Learning in the Detection of Accounting Fraud". Econ Journal Watch, 18(1), p.71.

Beaver, W.H., 1966. Financial ratios as predictors of failure. *Journal of accounting research*, pp.71-111.

Beaver, W.H., 1968. Market prices, financial ratios, and the prediction of failure. *Journal of accounting research*, pp.179-192.

Beneish, M.D., 1997. Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *Journal of accounting and public policy*, *16*(3), pp.271-309.

Beneish, M.D., 1999. The detection of earnings manipulation. *Financial Analysts Journal*, *55*(5), pp.24-36.

Beneish, M.D. and Vorst, P., 2020. The cost of fraud prediction errors. *Available at SSRN 3529662*

Block, Carson. “Interview with Carson Block on China, GSX, Tesla, Activist Shorts, And More” *YouTube video, 57:41. July 12, 2020. https://youtu.be/XkWwSgKX9bs*

Boyd, S., Cortes, C., Mohri, M. and Radovanovic, A., 2012. Accuracy at the top. In *Advances in neural information processing systems* (pp. 953-961).

Brown, N.C., Crowley, R.M. and Elliott, W.B., 2020. What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, *58*(1), pp.237-291.

Cecchini, M., Aytug, H., Koehler, G.J. and Pathak, P., 2010. Detecting management fraud in public companies. Management Science, 56(7), pp.1146-1160.

Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Coffee Jr, J.C., 1986. Understanding the plaintiff's attorney: The implications of economic theory for private enforcement of law through class and derivative actions. *Colum. L. Rev.*, *86*, p.669.

Dechow, P.M., Ge, W., Larson, C.R. and Sloan, R.G., 2011. Predicting material accounting misstatements. *Contemporary accounting research*, *28*(1), pp.17-82.

Dorsey, E. (@StockJabber). 19 November 2020, 9:29 a.m. Tweet. *https://twitter.com/StockJabber/status/1329476925300027395?s=20*

Dyck, A., Morse, A. and Zingales, L., 2010. Who blows the whistle on corporate fraud?. *The journal of finance*, *65*(6), pp.2213-2253.

"Electronic 'Brain' Teaches Itself". *New York Times*, 13 July 1958, Accessed on 7 July 2020, https://nyti.ms/3eJklE3

Fama, E.F., 1990. Contract costs and financing decisions. *Journal of Business*, pp.S71-S91.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), pp.861-874.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189-1232.

Grind, K., Zuckerman, G., Shifflet, S. 2019. "Care.com Puts Onus on Families to Check Caregivers' Backgrounds—With Sometimes Tragic Outcomes". *Wall Street Journal, March 8th, 2019. https://www.wsj.com/articles/care-com-puts-onus-on-families-to-check-caregivers-backgroundswith-sometimes-tragic-outcomes*

Kim, I. and Skinner, D.J., 2012. Measuring securities litigation risk. Journal of Accounting and Economics, 53(1-2), pp.290-310.

Kanell, M.E., Quinn, C. 2021. Parker Petit sentenced to one year in prison, $1 million fine. *Atlanta Journal-Consitution, February 23, 2021. https://www.ajc.com/ajcjobs/parker-petit-sentenced-to-one-year-in-prison-1-million-fine/PYFVC7323VDALG4MFIDIPEB7R4/*

Karpoff, J.M., Koester, A., Lee, D.S. and Martin, G.S., 2017. Proxies and databases in financial misconduct research. The Accounting Review, 92(6), pp.129-163.

Kedia, S. and Rajgopal, S., 2011. Do the SEC's enforcement preferences affect corporate misconduct?. *Journal of Accounting and Economics*, *51*(3), pp.259-278.

Kuhn, M. and Johnson, K., 2013. Applied predictive modeling (Vol. 26). New York: Springer.

Larcker, D.F. and Zakolyukina, A.A., 2012. Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50(2), pp.495-540.

Lundegaard, K. "Delphi Discloses Accounting Problems". *Wall Street Journal, March 7th, 2005. https://www.wsj.com/articles/SB110994509329670632* (accessed November 28th, 2020).

Narayanan, A. "How to recognize AI snake oil". . https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf (accessed April 15th, 2020).

Niculescu-Mizil, A. and Caruana, R., 2005, August. Predicting good probabilities with supervised learning. In Proceedings of the 22nd international conference on Machine learning (pp. 625-632).

O'glove, T.L., 1987. Quality of earnings.

Pearl, J. and Mackenzie, D., 2018. The book of why: the new science of cause and effect. Basic books.

Perols, J., 2011. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, *30*(2), pp.19-50.

Perols, J.L., Bowen, R.M., Zimmermann, C. and Samba, B., 2017. Finding needles in a haystack: Using data analytics to improve fraud prediction. *The Accounting Review*, *92*(2), pp.221-245.

Rakoff, J.S., 2014. The financial crisis: why have no high-level executives been prosecuted?. *The New York Review of Books*, *9*(2), p.7.

Saito, T. and Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), p.e0118432.

Schilit, H., 2018. *Financial shenanigans*, 4th ed. McGraw-Hill Education.

Schilit, H., 2020. Personal Interview. September 4th, 2020.

Simon, H.A., 1965. The shape of automation for men and management (Vol. 13), New York: Harper & Row, pp.96.

Stein, R.M., 2007. Benchmarking default prediction models: Pitfalls and remedies in model validation. Journal of Risk Model Validation, 77-113 (Spring)

Varian, H.R., 2014. Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), pp.3-28.

Vashisth, S., Linden, A., Hare, J., Krensky, P. 2019 Hype Cycle for Data Science and Machine Learning, 2019. *Gartner Research,* Accessed on 19 July 2020.

Walker, S., 2021. Critique of an Article on Machine Learning in the Detection of Accounting Fraud. Econ Journal Watch, 18(1), p.61.

Wang, Y., Wang, L., Li, Y., He, D., Chen, W. and Liu, T.Y., 2013, June. A theoretical analysis of NDCG ranking measures. In Proceedings of the 26th annual conference on learning theory (COLT 2013) (Vol. 8, p. 6).

Weerts, H.J., Mueller, A.C. and Vanschoren, J., 2020. Importance of tuning hyperparameters of machine learning algorithms. arXiv preprint arXiv:2007.07588.

Zelnick, Brad. Credit Suisse Initial Report on Snowflake. October 12, 2020

Zhao, Q. and Hastie, T., 2019. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, pp.1-10.

Appendix A: Feature List for Financial Ratios (XGBoost model)

Variables are labeled sequentially and start with "v" for continuous variables and "i" for binary.

Variable	Variable Name	Definition
v1	Abnormal change in order backlog	$(OB - OB_{t-1})/OB_{t-1} - (SALE - SALE_{t-1})/SALE_{t-1}$
i2	Actual issuance	if SSTK > 0 or DLTIS > 0 then 1, else 0
v3	Book-to-market	CEQ/(CSHO* PRCC_F)
v5	Change in free cash flows	(IB - RSST Accruals)/Average total assets - (IB_{t-1} - RSST $Accruals_{t-1}$)/Average total $assets_{t-1}$
v6	Change in inventory	$(INVT - INVT_{t-1})$/Average total assets
v7	Change in operating lease activity	((MRC1/1.1 + MRC2/1.1^2 + MRC3/1.1^3 + MRC4/1.1^4 + MRC5/1.1^5) -(MRC1t-1/1.1 + MRC2t-1/1.1^2 + MRC3t-1/1.1^3 + MRC4t-1/1.1^4 +MRC5t-1/1.1^5))/Average total $assets_{t-1}$
v8	Change in receivables	$(RECT - RECT_{t-1})$/Average total assets
v9	Change in return on assets	IB/Average total assets - IB_{t-1}/Average total $assets_{t-1}$
v10	Deferred tax expense	$TXDI/AT_{t-1}$
i11	Demand for financing (ex ante)	if (OANCF - (CAPXt-3 + CAPXt-2 + CAPXt-1)/3)/ACT < -0.5, then 1, else 0
v12	Earnings to price	IB/(CSHO * PRCC_F)
i13	Existence of operating leases	if MRC1 > 0, or MRC2 > 0, or MRC3 > 0, or MRC4 > 0, or MRC5 > 0,then 1, else 0
v15	Level of finance raised	FINCF/Average total assets
v16	Leverage	DLTT/AT
v17	Percentage change in cash margin	$((1 - (COGS + (INVT - INVT_{t-1}))/(SALE - (RECT - RECT_{t-1}))) - (1-(COGS_{t-1} + (INVT_{t-1} - INVT_{t-2}))/(SALE_{t-1} - (RECT_{t-1} - RECT_{t-2}))))/(1- (COGS_{t-1} + (INVT_{t-1} - INVT_{t-2}))/(SALE_{t-1} - (RECT_{t-1} - RECT_{t-2})))$
v18	Percentage change in cash sales	$((SALE - (RECT - RECT_{t-1})) - (SALE_{t-1} - (RECT_{t-1} - RECT_{t-2})))/(SALE_{t-1} - (RECT_{t-1} - RECT_{t-2}))$
v19	RSST accruals	RSST Accruals= (DWC + DNCO + DFIN)/Average total assets, where:
		WC = (ACT - CHE) - (LCT- DLC)
		NCO = (AT - ACT - IVAO) - (LT - LCT - DLTT)

		FIN = (IVST + IVAO) - (DLTT + DLC + PSTK)
v20	Soft assets	(AT - PPENT - CHE)/Average total assets
v21	Unexpected employee productivity	(SALE/EMP - $SALE_{t-1}/EMP_{t-1}$)/($SALE_{t-1}/EMP_{t-1}$) - INDUSTRY((SALE/EMP - $SALE_{t-1}/EMP_{t-1}$)/($SALE_{t-1}/EMP_{t-1}$))
v22	WC accruals	(((ACT - ACT_{t-1}) - (CHE - CHEt-1)) - ((LCT - LCT_{t-1}) - (DLC - DLC_{t-1}) - (TXP - TXP_{t-1})) - DP)/Average total assets
v23	Accounts receivable to sales	RECT/SALE
v24	Accounts receivable to total assets	RECT/AT
v25	Allowance for doubtful accounts	RECD
v26	Allowance for doubtful accounts to accounts receivable	RECD/RECT
v27	Allowance for doubtful accounts to net sales	RECD/SALE
v28	Altman Z-score	3.3 * (IB + XINT + TXT)/AT + 0.999 * SALE/AT + 0.6 * CSHO – 1.0*PRCC_F/LT + 1.2 *WCAP/AT + 1.4 * RE/AT
v29	Big Four auditor	if 0 < AU < 9, then 1, else 0
v30	Current minus prior year inventory to sales	INVT/SALE - $INVT_{t-1}/SALE_{t-1}$
v31	Days in receivables index	(RECT/SALE)/($RECT_{t-1}/SALE_{t-1}$)
v32	Debt-to-equity	LT/CEQ
v33	Declining cash sales dummy	if SALE - (RECT - $RECT_{t-1}$) , $SALE_{t-1}$ - ($RECT_{t-1}$ - $RECT_{t-2}$) then 1, else 0
v34	Fixed assets to total assets	PPEGT/AT
v35	Four-year geometric sales growth rate	(SALE/$SALE_{t-4}$) ^ (1/4) - 1
v36	Gross margin	(SALE - COGS)/SALE
v37	Holding period return	(PRCC_F - $PRCC\_F_{t-1}$)/$PRCC\_F_{t-1}$
v38	Industry ROE minus firm ROE	NI of industry/CEQ of industry - NI/CEQ
v39	Inventory to sales	INVT/SALE
v40	Net sales	SALE

i41	Positive accruals dummy	if (IB - OANCF) > 0 and (IB_{t-1} - $OANCF_{t-1}$) > 0, then 1, else 0
v42	Prior-year ROA to total assets current year	(NI_{t-1}/AT_{t-1})/AT
v43	Property, plant, and equipment to total assets	PPENT/AT
v44	Sales to total assets	SALE/AT
v45	The number of auditor turnovers	if AU ,> AU_{t-1}, then 1, else 0 + if AU_{t-1} ,> AU_{t-2}, then 1, else 0 + if AU_{t-2},> AU_{t-3} then 1, ELSE 0
v46	Times interest earned	(IB + XINT + TXT)/XINT
v47	Total accruals to total assets	(IB - OANCF)/AT
v48	Total debt to total assets	LT/AT
v49	Total discretionary accrual	RSST $Accruals_{t-1}$ + RSST $Accruals_{t-2}$ + RSST $Accruals_{t-3}$
v50	Value of issued securities to market value	if CSHI > 0, then CSHI - PRCC_F/(CSHO - PRCC_F) else if (CSHO - $CSHO_{t-1}$) > 0, then ((CSHO - $CSHO_{t-1}$) PRCC_F)/(CSHO - PRCC_F), else 0
i51	Whether accounts receivable > 1>1 of last year's	if $RECT/RECT_{t-1}$ > 1.1, then 1, else 0
v52	Whether firm was listed on AMEX	if EXCHG = 5, 15, 16, 17, 18, then 1, else 0
v53	Whether gross margin percent > 1>1 of last year's	if ((SALE - COGS)/SALE)/(($SALE_{t-1}$ - $COGS_{t-1}$)/$SALE_{t-1}$) > 1>1, then 1, else 0
v54	Whether LIFO	if INVVAL = 2, then 1, else 0
v55	Whether new securities were issued	if (CSHO - $CSHO_{t-1}$) > 0 or CSHI > 0, then 1, else 0
v56	Whether SIC code between 2999 and 4000	if 2999 < SIC < 4000, then 1, else 0
v57	Sales	SALE
v58	Change in sales	SALE - $SALE_{t-1}$
v59	% Change in sales	(SALE - $SALE_{t-1}$)/$SALE_{t-1}$
v60	Abnormal % change in sales	(SALE - $SALE_{t-1}$)/$SALE_{t-1}$ - INDUSTRY(SALE - $SALE_{t-1}$)/$SALE_{t-1}$
v61	Sales to assets	SALE/AT
v62	Change in sales to assets	SALE/AT - $SALE_{t-1}/AT_{t-1}$

v63	% Change in sales to assets	$(SALE/AT - SALE_{t-1}/AT_{t-1})/(SALE_{t-1}/AT_{t-1})$
v64	Abnormal % change in sales to assets	$(SALE/AT - SALE_{t-1}/AT_{t-1})/(SALE_{t-1}/AT_{t-1})$ - INDUSTRY$(SALE/AT - SALE_{t-1}/AT_{t-1})/(SALE_{t-1}/AT_{t-1})$
v65	Sales to employees	SALE/EMP
v66	Change in sales to employees	$SALE/EMP - SALE_{t-1}/EMP_{t-1}$
v67	% Change in sales to employees	$(SALE/EMP - SALE_{t-1}/EMP_{t-1})/(SALE_{t-1}/EMP_{t-1})$
v68	Sales to operating expenses	SALE/XOPR
v69	Change in sales to operating expenses	$SALE/XOPR - SALE_{t-1}/XOPR_{t-1}$
v70	% Change in sales to operating expenses	$(SALE/XOPR - SALE_{t-1}/XOPR_{t-1})/(SALE_{t-1}/XOPR_{t-1})$
v71	Abnormal % change in sales to operating expenses	$(SALE/XOPR - SALE_{t-1}/XOPR_{t-1})/(SALE_{t-1}/XOPR_{t-1})$ - INDUSTRY$(SALE/XOPR - SALE_{t-1}/XOPR_{t-1})/(SALE_{t-1}/XOPR_{t-1})$
v72	Return on assets	NI/AT
v73	Change in return on assets	$NI/AT - NI_{t-1}/AT_{t-1}$
v74	% Change in return on assets	$(NI/AT - NI_{t-1}/AT_{t-1})/(NI_{t-1}/AT_{t-1})$
v75	Abnormal % change in return on assets	$(NI/AT - NI_{t-1}/AT_{t-1})/(NI_{t-1}/AT_{t-1})$ - INDUSTRY$(NI/AT - NI_{t-1}/AT_{t-1})/(NI_{t-1}/AT_{t-1})$
v76	Return on equity	NI/CEQ
v77	Change in return on equity	$NI/CEQ - NI_{t-1}/CEQ_{t-1}$
v78	% Change in return on equity	$(NI/CEQ - NI_{t-1}/CEQ_{t-1})/(NI_{t-1}/CEQ_{t-1})$
v79	Abnormal % change in return on equity	$(NI/CEQ - NI_{t-1}/CEQ_{t-1})/(NI_{t-1}/CEQ_{t-1})$ - INDUSTRY$(NI/CEQ - NI_{t-1}/CEQ_{t-1})/(NI_{t-1}/CEQ_{t-1})$
v80	Return on sales	NI/SALE
v81	Change in return on sales	$NI/SALE - NI_{t-1}/SALE_{t-1}$
v82	% Change in return on sales	$(NI/SALE - NI_{t-1}/SALE_{t-1})/(NI_{t-1}/SALE_{t-1})$
v83	Abnormal % change in return on sales	$(NI/SALE - NI_{t-1}/SALE_{t-1})/(NI_{t-1}/SALE_{t-1})$ - INDUSTRY$(NI/SALE - NI_{t-1}/SALE_{t-1})/(NI_{t-1}/SALE_{t-1})$

v84	Accounts payable to inventory	AP/INVT
v85	Change in accounts payable to inventory	AP/INVT - $AP_{t-1}/INVT_{t-1}$
v86	% Change in accounts payable to inventory	(AP/INVT - $AP_{t-1}/INVT_{t-1}$)/($AP_{t-1}/INVT_{t-1}$)
v87	Abnormal % change in accounts payable to inventory	(AP/INVT - AP_{t-1}/INVTt-1)/($AP_{t-1}/INVT_{t-1}$) - INDUSTRY(AP/INVT - AP_{t-1}/ $INVT_{t-1}$)/($AP_{t-1}/INVT_{t-1}$)
v88	Liabilities	LT
v89	Change in liabilities	LT - LT_{t-1}
v90	% Change in liabilities	(LT - LT_{t-1})/LT_{t-1}
v91	Abnormal % change in liabilities	(LT - LT_{t-1})/LT_{t-1} - INDUSTRY(LT - LT_{t-1})/LT_{t-1}
v92	Liabilities to interest expenses	LT/XINT
v93	Change in liabilities to interest expenses	LT/XINT - $LT_{t-1}/XINT_{t-1}$
v94	% Change in liabilities to interest expenses	(LT/XINT - $LT_{t-1}/XINT_{t-1}$)/($LT_{t-1}/XINT_{t-1}$)
v95	Abnormal % change in liabilities to interest expenses	(LT/XINT - $LT_{t-1}/XINT_{t-1}$)/($LT_{t-1}/XINT_{t-1}$) - INDUSTRY(LT/XINT -$LT_{t-1}/XINT_{t-1}$)/($LT_{t-1}/XINT_{t-1}$)
v96	Assets	AT
v97	Change in assets	AT - AT_{t-1}
v98	% Change in assets	(AT - AT_{t-1})/AT_{t-1}
v99	Abnormal % change in assets	(AT - AT_{t-1})/AT_{t-1} - INDUSTRY(AT - AT_{t-1})/AT_{t-1}
v100	Assets to liabilities	AT/LT
v101	Change in assets to liabilities	AT/LT - AT_{t-1}/LT_{t-1}
v102	% Change in assets to liabilities	(AT/LT - AT_{t-1}/LT_{t-1})/(AT_{t-1}/LT_{t-1})
v103	Abnormal % change in assets to liabilities	(AT/LT - AT_{t-1}/LT_{t-1})/(AT_{t-1}/LT_{t-1}) - INDUSTRY(AT/LT - AT_{t-1}/LT_{t-1})/(AT_{t-1}/LT_{t-1})

v104	Expenses	XOPR
v105	Change in expenses	$XOPR - XOPR_{t-1}$
v106	% Change in expenses	$(XOPR - XOPR_{t-1})/XOPR_{t-1}$
v107	Abnormal % change in expenses	$(XOPR - XOPR_{t-1})/XOPR_{t-1}$ - INDUSTRY$(XOPR - XOPR_{t-1})/XOPR_{t-1}$
v108	Cash Liquidity	CHE / AT

Appendix B: Raw Variable List (Bao, et al. 2020)

Compustat Code	**Compustat Description**
ACT	Current Assets - Total
AP	Accounts Payable - Trade
AT	Total Assets
CEQ	Common / Ordinary Equity - Total
CHE	Cash and Short-term equivalents
COGS	Cost of Goods Sold
CSHO	Common Shares Outstanding
DLC	Debt in Current Liabilities
DLTIS	Long-term Debt Issuance
DLTT	Long-term Debt Total
DP	Depreciation and Amortization
IB	Income before extraordinary items
INVT	Inventories - Total
IVAO	Investment and Advances - Other
IVST	Short-Term Investments - Total
LCT	Current Liabilities - Total
LT	Liabilities - Total
NI	Net Income (Loss)
PPEGT	Property, Plant & Equipment - Total (Gross)
PSTK	Preferred / Preference Stock (Capital) - Total
RE	Retained Earnings
RECT	Receivables - Total
SALE	Sales / Turnover (Net)
SSTK	Sale of Common and Preferred Stock
TXP	Income Taxes Payable
TXT	Income Taxes - Total
XINT	Interest and Related Expense - Total
PRCC_F	Price Close - Annual - Fiscal