ENHANCING CORPORATE CRIME ENFORCEMENT WITH MACHINE LEARNING—A MULTIDISCIPLINARY RISK FACTOR APPROACH

By

Fiona Chan

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Criminal Justice—Doctor of Philosophy

2022

ABSTRACT

ENHANCING CORPORATE CRIME ENFORCEMENT WITH MACHINE LEARNING—A MULTIDISCIPLINARY RISK FACTOR APPROACH

By

Fiona Chan

Despite its severe and lasting social and financial ramifications, corporate financial crime remains one of the most understudied crime types, as it is often hindered by two challenges. First, its multidisciplinary nature requires both financial and criminological expertise among others to conduct proper investigations. Second, corporate crime data is fraught with constraints such as high dimensionality, complex interactions, and nonlinear functional forms that are ill-suited for classical statistical modeling. The lack of research coupled with the limited resources in corporate crime enforcement represent a great impediment to the advancement of fraud interventions. This dissertation seeks to overcome these specific challenges by unifying cross-disciplinary financial fraud research under a risk factor framework, and by leveraging recent advancements in artificial intelligence. The goal is to examine whether two machine learning algorithms—random forest and neural network—can be used to enhance corporate fraud risk detection/prediction beyond more commonly employed analytical techniques.

Findings from the analysis showed that the random forest algorithm outperformed logistic regression and a naïve classifier in a 1:1 matched sample. The neural network performed better than a naïve classifier but slightly worse than logistic regression. Feature selection improved the algorithms' predictive accuracy

and ability to distinguish between classes even further. Despite promising results from the 1:1 matched sample, both machine learning algorithms struggled with a heavily imbalanced 1: many dataset, which represents a more realistic setting. With the implementation of an oversampling strategy and feature selection, the algorithms improved substantially in identifying the rare fraud cases, and showed promise of improvement with further research on imbalanced classification.

Feature importance from the random forest classifier identified risk factors that are consistent with findings from prior studies. Measures of financial distress ranked lower in importance than measures of financial health, suggesting future research can build on prior findings on corporate strain to examine specific mechanisms. The analysis also identified auditor independence as a key concept of guardianship and opportunity structure that warrants further study. Findings from this research also have important methodological implications for corporate crime studies—namely, the need to improve measurements of organizational-level fraud risk factors.

In memory of my lost lemon.

ACKNOWLEDGEMENTS

My long journey of pursuing a Ph.D. comes with an equally long list of individuals to whom I owe my sincerest gratitude. I express my deepest thanks to my dissertation committee: Drs. Carole Gibbs, Michael Benson, Steven Chermak, Chris Melde, and Wenjuan Ma, for your invaluable insights on this dissertation project, and for your immense contribution to my professional growth.

Carole, thank you for being the best mentor a PhD student can ask for. I can honestly say I would not have survived this program without your guidance and encouragement. Thank you for all the time and effort you have invested into being such a genuine mentor to me, and for always having my best interest at heart. Mike, thank you for letting this random public accountant show up at your office door years ago, and for inspiring her to pursue a career in academia. My life as an academic began with you, and I am forever grateful for your continuous support. Rae, thank you for being the most wonderful colleague, friend, and human being. You make the hardest parts of this journey so much more tolerable.

Thank you, mom and dad, for your unconditional love and care, even when faced with life-altering crises of your own. Porro, for keeping my mental health in check. Holmes and Adie, for the sacrificed walkies and fetch time. Mr. and Mrs. K, for all the crisis management. Matthew, for grad school rants and reminders of life beyond career. Dear friends, for the much-needed mandatory fun. Most importantly, thank you, Kyle, for being the one constant that makes everything good possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1| Problem Statement

Corporate financial fraud broadly refers to the misrepresentation of an enterprise's financial condition through accounting practices that deviate from the Generally Accepted Accounting Principles (GAAP) and/or omission of pertinent information from mandated disclosures.[1] Despite being one of the least prevalent forms of white-collar crime, corporate financial fraud has consistently been shown to be the costliest (see ACFE's Reports to the Nation 2010-2018). Dyke et al. (2013) estimated annual cost of corporate financial fraud to be upwards of $181 billion in the U.S. The substantial monetary losses associated with corporate financial fraud not only devastate businesses, they also threatens the livelihood of public investors and employees, jeopardizing retirement savings and job securities. It further impacts the economy as a whole, as investors lose trust and confidence in the capital markets (FBI, 2018). Thus, these consequences of corporate financial fraud can be enduring and far-reaching.

Despite these social consequences, corporate crime research and enforcement are each faced with its own difficulties. Resources dedicated to the regulation and enforcement of corporate financial fraud is relatively limited in comparison to conventional street crime (Karpoff, Koester, Lee and Martin, 2017). Perceptions of lax enforcement (Unnever, Benson and Cullen, 2008; Holtfreter, Van Slyke,

[1] GAAP represent a standard set of accounting principles set forth by the Financial Accounting Standards Board (in the U.S.), with which all publicly listed corporations must adhere to. See below for further discussion on definitions of corporate financial fraud.

Bratton, and Gertz, 2008) creates a challenge with achieving deterrence, as effective deterrence requires swift and certain detection and proportionate punishment (Bentham, 1962; Becarria, 1764).

There is also a lack of evidence-based research to inform corporate crime policies and prevention efforts. Scholars have attributed this paucity of corporate crime research to resource constraints, lack of centralized, longitudinal official data, conceptual ambiguity of the definitions of white-collar and corporate crime, and offenders' resourcefulness in avoiding prosecution (Rorie et al., 2018; Paternoster, 2016; Simpson & Yeager, 2015; Braithwaite, 2016; Simpson, 2013). In addition to these commonly cited drawbacks, two additional factors may have contributed to this problem.

First, corporate crime is an inherently multidisciplinary problem, requiring domain-specific knowledge from disparate technical fields of study. Research interest in different aspects of corporate financial fraud has generated scattered pockets of knowledge in disciplines such as criminal justice, accountancy/finance, information systems, organizational psychology, linguistics and communications (discussed in further detail below). While scant in quantity, lessons learned from this body of work, when combined, may help advance current prevention and intervention efforts targeting corporate financial fraud. Recognizing this need, scholars have called for more interdisciplinary collaboration (Simpson, 2013; Trompeter, Carpenter, Jones and Riley, 2012).

Second, corporate crime research is fraught with methodological challenges related to statistical modeling, due to the complex interactions of cross-level antecedents and insufficient degrees of freedom (Perols et al., 2016). Exacerbating the problem of weak data, many commonly employed statistical models suffer from an underfitting problem, as they fall short in capturing the complexity of the data structure and relationships between corporate crime variables (Simpson, 2013; Paternoster, 2016). Operating under the frequentist philosophy on probability, these models also impose strict assumptions that are ill-suited for corporate crime data (Perols et al., 2016). As a result, researchers face tremendous difficulties in establishing baseline relationships between potential antecedents and corporate crime that can inform enforcement practices (Schell-Busey et al., 2016).

1.2.| Goals and Merit of Current Research

By combining criminological theories, domain knowledge in accounting and linguistics/ communication, and technological and analytical advancements in the information systems disciplines, this dissertation aims to initiate a first step in overcoming some of the challenges described above. The goal of this project is to develop machine learning fraud prediction algorithms and assess whether recent developments in this subfield of artificial intelligence can aid in corporate financial fraud detection. This goal can be further broken down into three objectives—first is to synthesize cross-disciplinary knowledge on corporate financial fraud under a risk factor framework; next is to use the identified risk factors as inputs to develop two

machine learning risk assessment tools and compare them against standard benchmarks for financial fraud detection; last is to further our understanding on one of the identified risk factor groups—deception cues—that has received very little attention in corporate crime research both theoretically and empirically.

There is much merit to accomplishing these goals. By adopting a risk factor approach to fraud detection, the current project helps unify cross-disciplinary literature and brings cohesion to the understudied field of corporate crime. And by leveraging the most recent advancements in computational data analytics, this dissertation project seeks to overcome the data scarcity and methodological challenges that hamper corporate crime research. Risk data compiled for the project represents the first step toward a comprehensive multidisciplinary corporate financial fraud risk database with combined explanations that aid in interdisciplinary theory development. Data on the fraud and matched firms will also represent the most recent empirical data set for corporate crime gathered from original open-source data. Together, these datasets will constitute a step forward in overcoming the data deficiency problem in white-collar crime research. The final machine learning models will contribute to crime prevention by identifying the key risk factors that are most predictive of corporate financial fraud. This will not only pinpoint specific areas most in need of interventions, but will also form the quantitative basis for subsequent qualitative causal mechanism inquiries. Finally, the machine learning models will also contribute to risk assessment methods for other crime types or for research on reoffending.

In terms of practical significance, given the limited resources afforded to corporate financial fraud regulatory agencies, the machine learning models may provide an efficient tool in screening and detecting fraud cases. Since the algorithms are adaptive and scalable, it will be able to accommodate new data and efficiently produce the predictive results that swift enforcement requires. Improvement in the celerity and certainty of detection will in turn directly improve deterrence of corporate financial fraud. It may also advance our understanding on the true scope of corporate financial fraud. Since corporate crime tends to accumulate in severity the longer it remains undetected, more efficient detection tools can aid in reducing government spending on rectifying corporate misconduct in the long run. Furthermore, higher enforcement levels can raise investor awareness (Brazel et al., 2015), which can introduce a different form of crime control mechanism.

CHAPTER 2. BACKGROUND AND RELEVANT LITERATURE

2.1| Challenges in the Study of White-Collar and Corporate Crime

2.1.1| *Definition of Corporate Financial Fraud*

White-collar and corporate crime research is difficult to synthesize in part due to definitional inconsistencies. Over seven decades after Edwin Sutherland coined the term "white-collar crime" (1949), its definition continued to remain a source of debate amongst white-collar and corporate crime scholars. Most research have now adopted the approach of defining the crime based on the goal of the research (Friedrichs, 1992). For the purposes of this dissertation, corporate crime is defined as "conduct of a corporation, or of employees acting on behalf of a corporation, which is proscribed and punishable by law" (Braithwaite, 1984; p.6). A corporation refers to a legal entity formed under the laws of its state of incorporation. It is considered a "legal person" distinctive from its owners. But even with an offense-based definition, what have been referred to as corporate financial fraud so far could mean financial misreporting, financial misrepresentation, fraudulent interstate transactions, and/or securities fraud involving manipulative and deceptive devices, depending on the data used in the research.[2] These subtypes of corporate financial fraud are enforced under various sections of the Securities Exchange Act of 1934.

Financial misreporting refers violation of Section 13(a), which requires timely filing of financial reports (including annual 10-K reports and quarterly 10-Q

[2] Language and definitions presented here are adopted from Amiram, Bozanic, Cox, Dupont, Karpoff and Sloan (2018).

reports) from publicly listed firms. Financial misrepresentation refers to violation of Section 13(b), which requires listed firms to keep accurate books and records and maintain an effective internal control system that ensures accurate reporting. In particular, financial records, statements and disclosures must be in conformity to the Generally Accepted Accounting Principles (GAAP). GAAP represents a standard set of accounting rules and principles set forth by the Financial Accounting Standards Board (FASB) and are codified under the Accounting Standards Codification (ASC). Section 17(a) prohibits the use of interstate commerce for the purpose of fraud and deceit, and finally Section 10(b) prohibits the use of manipulative and deceptive device with the purchase and sale of any security.[3]

The present study focuses on financial misrepresentation, and more specifically Sections 13(b)(2)(A) and 13(b)(2)(B) of the Securities Exchange Act (1934). While cases involving these sections can also involve financial misreporting, fraudulent interstate transactions or securities fraud, it is important to make the distinction between misstatements and omission from financial reports and disclosure (financial misrepresentation) and a late filing (financial misreporting), for instance, as one would expect the risk factors associated with these two subtypes of corporate financial fraud to be distinctive due to the different motivations and mechanisms of offending.

[3] Rule 10(b)-5 is frequently levied in private enforcement against alleged fraud firms through class action lawsuits.

2.1.2| *Methodological and Data Challenges*

In addition to definitional debates that may hamper empirical corporate crime research, more technical difficulties with data quality and statistical modeling are often the culprits for null findings and uninterpretable results. Unlike the more readily available and systematically collected FBI official data for offense-based white-collar crimes, corporate crime regulation agencies are far from consistent in records of corporate crime instances. Antecedents of corporate crime are often high in quantity and multilevel in nature, and are characterized by complex interactions across industry, firm and individual levels. This high dimensionality coupled with small datasets often lead to insufficient degrees of freedom (Perols et al., 2016). Exacerbating the problem with cross-sectional data, rigid functional form and strict assumptions, many empirical analyses exhibit signs of underfitted models (Simpson, 2013; Paternoster, 2016; Perols et al., 2016). Scarce replicable empirical results are likely due to these research difficulties pertaining to data and methods. Scarce replicable baseline associations between a predictor and corporate crime makes informing policies and practices with evidence-based research near impossible.

2.1.3| *Enforcement Challenges*

Enforcement of corporate crime has often been criticized as lax in public perception (Unnever, Benson and Cullen, 2008; Holtfreter, Van Slyke, Bratton, and Gertz, 2008). For example, Cohen et al. (2015) reported slippage in SEC's mandatory disclosure requirements; Cox et al. (2016), found substantial percentage

of unenforced joinders of multiple unconnected items in proxy resolutions. Some corporate crime scholars have attributed this limited enforcement to the lack of resources dedicated to the regulation and enforcement, especially when compared to conventional street crime (Karpoff, Koester, Lee and Martin, 2017; Feroz et al. 1991). Others have placed blame on regulatory capture (Bozanic et al., 2012; Vaughan, 2002). When accounting irregularities are uncovered and come into the attention of the SEC, it often takes 36 months for the SEC's Enforcement Division to open, investigate and file a case (Woodcock, Shipchandler, McKown and Day, 2019). As such, the enforcement agencies often only take on investigations and prosecutions when there is a high probability of conviction. The cases that are processed typically result in a class-action civil suits against senior executives of the firm itself (COSO, 1999).

The SEC seldom prosecute audit firms associated with the corporations accused of financial misconducts (Brennan and McGrath, 2007).[4] Corporations that attribute their financial misrepresentation to audit failure often have to hold the auditing firms accountable via private litigation.[5] In addition to budgeting issues, researchers have also pointed out other struggles that plague the SEC—such as its ineffectiveness in collecting disgorgement, constitutional challenges to its administrative procedures, and the growth of digital asset offerings. These

[4] Publicly traded firms are required to have their financial statements opined on by external auditors. An unqualified opinion meant that external auditors have deemed the company's financial reports to be free of material misstatements.

[5] Palmros (1987) found that auditor litigations are often associated with the economic climate. That is, auditor litigations tend to increase during economic recessions.

challenges have been exacerbated by recent political climate, with hiring freezes and staffing decreases (by 400 positions compared to 2016) and enforcement personnel reduced by 10%, fewer cases involving public companies are filed (Woodcock et al., 2019).

The phenomenon described above creates tremendous challenge in achieving deterrence, as enforcement is neither swift, certain, nor severe (Bentham, 1962; Becarria, 1764). The lack of evidence-based policies derived from research coupled with the lack of consistent enforcement effort limits the effectiveness of corporate crime control efforts, which is evident in the recurring corporate financial scandals every few years or so.

2.2| The Risk Factor Approach

2.2.1| *Overview*

A risk factor refers to any attribute or characteristic of an organization that increases its likelihood of corporate financial fraud. Identification of risk factors plays an important role in both enforcement and research of corporate financial fraud. With regards to enforcement, all publicly traded companies are required to disclose self-assessed risk factors in their annual financial report (Form 10-K) to the Securities Exchange Commission (SEC). External auditors are also required to consider a clients' fraud risk factors as part of the annual audit procedures in accordance to the Statements of Auditing Standards No.99. With regards to research, risk-focused studies have contributed to numerous crime prevention

efforts in other criminal justice domains (e.g., Farrington, 2000). Risk-based research lays the groundwork in establishing basic patterns and relationships between relevant antecedents and the crime of interest, from which future research can be built. Some scholars (e.g., Bernard and Snipes, 2016) have even argued for a risk factor approach to theory integration, as the traditional approach of theory falsification does not appear to have made much progress in theory reduction.

Given the importance of risk factors and the research interest evident in various disciplines, the proposed project aims to aggregate research on corporate financial fraud to compile a comprehensive set of corporate financial fraud risk factors. However, identification alone is insufficient, as unexplained risk factors give regulators, enforcement agents and researchers very little guidance in planning relevant provisions and interventions to mitigate those risks. Thus, I would also like to document the domain specific explanations associated with each risk factor and how it may be linked to crime prevention and criminal justice theories broadly.

2.2.2| *Corporate Crime Risk Factors*

White-collar crime scholars have linked performance pressure, organization size, and organizational structure and complexity to various types of corporate crime, although empirical findings are not consistent. For example, some studies found a weak but significant negative relationship between multiple corporate offense types and firm profit (Clinard and Yeager, 1980; McKendall et al., 2002), profitability trends such as declining financial performance (Keane, 1993), low

growth (Alexander and Cohen, 1996; Clinard and Yeager, 1980), substantial savings for the firm (Paternoster and Simpson, 1996), and financial distress such as bankruptcy risk (Schwartz et al., 2021). Schuchter and Levi's (2013) interviews with high-profile fraudsters provided further support on the role of performance pressure as a salient factor in their crime decisions and rationalizations. Yet, other research has indicated that firm profits are unrelated to financial (e.g. Simpson, 1986) and non-financial crimes (Baucus and Near, 1991; Hill et al., 1992); some even found profits (McKendall and Wagner, 1997) and growth (Simpson, 2002; Wang & Holtfreter, 2012) to be associated with increased violations.

Though inconsistent (McKendall and Wagner, 1997; Paternoster and Simpson, 1996; Simpson, 2002), overall the literature suggests a positive relationship between organization size and financial crimes (Schwartz, 2021; Baucus and Near, 1991; Simpson, 1986) as well as non-financial crimes such as discrimination (Baucus and Near, 1991) and environmental cases (Alexander and Cohen, 1996). While complexity (McKendall and Wagner, 1997) is theoretically relevant, diversification is unrelated to multiple types of corporate misconduct (Clinard and Yeager, 1980; Hill et al., 1992). However, more criminogenic industries can exacerbate financial strains and accordingly, violation rates (Wang & Holtfreter, 2012). In addition to these empirical findings, scholars have also theorized extensively on the criminogenic properties of organizational structures. Certain cultural mandates, political environments and departmentalization within an organization can engender corporate misconducts, by facilitating acts of

normalized deviance, concerted ignorance and/or structural secrecy (e.g., Vaughan, 2002; Prechel and Morris, 2010). For instance, knowledge of misconducts can be compartmentalized within subunits of a complex organization such that detection by another subunit is difficult.

Accounting and organizational research is more specific to corporate financial fraud. In addition to case studies of major "creative accounting" scandals (e.g., Cohan, 2002; Bhasin 2013; Jones, 2011), this body of research also identified a variety of risk factors associated with corporate financial fraud. For example, various financial metrics such as those measuring rapid growth or financial instability (e.g., cash flow, debt, and sales-related indices) and operational efficiency (e.g., assets-related indices and turnover ratios) are linked to higher likelihood of fraudulent financial reporting. Firm characteristics are also relevant; the lack outside blockholders[6] (Dechow et al., 1996), high latitude of managerial discretion (e.g., discretionary accrual estimates) (Beneish, 1999; Bell and Carcello, 2000), and firms in certain industries are also more likely to engage in specific types of financial misstatements (Beasley et al., 2000).

Other studies focus on the individual incentives/motivations for committing corporate financial fraud. Executive equity incentives such as stock options have received mixed support in relation to financial fraud (e.g., Burns and Kedia, 2006; Erickson et al., 2006; Efendi et al., 2007). Meeting analyst expectations was

[6] Owners of large blocks of company shares and/or bonds with special voting rights.

associated with an increase in accounting scandals at the macro level (Koh et al., 2008) and the organizational level (Perols and Lougee, 2011).[7]

Then there are studies that have examined the guardianship dimension of opportunities for corporate financial fraud. These studies link weak internal controls, corporate governance and audit quality to the likelihood of financial misstatements. Specifically, board composition and/or characteristics such as committee independence (Beasley, 1996; Abbott et al., 2004), the presence of financial experts (Farber, 2005), CEO's tenure and his/her dual roles as Chairman of the board (Dechow et al., 1996) are board characteristics that significantly predict financial fraud. Loebecke et al. (1989) also attributed corporate financial fraud to weak internal controls that are dominated by motivated management. Smith et al. (2000) raised concern for the effectiveness of external auditors in their guardianship role when they found that risk assessments only alter the allocation of control versus substantive testing, but did not increase the likelihood of fraud detection. Auditor characteristics, including auditing firms' size, industry specialization, tenure with the client, and individual experience of the audit team, are also associated with likelihood of fraud (Carcello and Nagy, 2004; Myers et al., 2003).

Linguistics and communications scholars have identified deception-related risk factors that represent verbal, visual and audible cues of deceitful content (e.g. Dyer et al., 2016; Throckmorton et al., 2015; Humphreys, Moffit, Burns, Burgoon

[7] Financial analysts specializing in specific public corporations/industries make periodical forecast and prediction on how the companies perform prior to actual filing of company financial statements to the SEC. These forecasts are generally viewed by the public as expert advice for investment purposes. As such, companies have an incentive to meet analyst expectations in order to preserve the trend of their stock prices.

and Felix, 2011). Humphreys et al. (2011) examined linguistic cues and found significant differences between fraud and non-fraud firms. According to these authors these linguistic-based risk factors may be management's deliberate attempt to deceive (explained by Bloomfield's management obfuscation hypothesis and McCornack's information manipulation theory) (Bloomfield, 2002; McCornack, 1992), or they may be attempts to hide unintentional "leakage" of deception cues that stem from being dishonest (explained by interpersonal deception theory and four factor theory) (Buller and Burgoon, 1996; Zuckerman, DePaulo and Rosenthal, 1981).

Finally, organizational psychologists have examined organizational justice related risk and protective factors that may facilitate corporate misconduct or encourage whistleblowing (e.g. Young, 2013; Seifert et al., 2010; Lewicki et al., 2005; Cropanzano et al., 2001). Other factors such as corporate culture and perceived likelihood of retaliation are linked to intention to whistleblow (Commers, 2004; Mesmer-Magnus and Viswesvaran, 2005), which may serve as protective factors for corporate financial crime. The concept of relational governance (Poppo & Zenger, 2002; Kramer, 1999; Zajac & Olsen, 1993) similarly represents protective factors that are akin to the informal crime control mechanisms in the street crime literature.

As one may have observed, research from diverse disciplines have applied their respective interests to cases of corporate crime. Despite the diversity of findings and conclusions, they provide unique and valuable insights when

combined. Systematic organization of these risk factors under a criminological framework will provide a multi-disciplinary understanding of the understudied area of corporate financial fraud.

CHAPTER 3. CURRENT RESEARCH

3.1| Research Questions

In light of the project's objectives to synthesize cross-disciplinary knowledge and to overcome current methodological challenges by ways of machine learning, the research questions that will be addressed in this dissertation are as follows:

1. Can the multi-disciplinary risk factors identified by research be used to predict corporate financial fraud with the use of a random forest classifier (i.e., does the algorithm perform better than a naïve classifier[8])?
2. How does the random forest classifier perform in comparison to commonly employed prediction tools (e.g., logistic regression)?
3. Which of the multi-disciplinary risk factors are most important in predicting corporate financial fraud?
4. Can the multi-disciplinary risk factors be used to predict corporate financial fraud with the use of a deep neural network classifier? (i.e., does the algorithm perform better than a naïve classifier)?
5. How does a neural network classifier perform in comparison to logistic regression and the random forest classifier?

3.2| Identification of Corporate Financial Fraud Risk Factors

In the previous chapter, I reviewed the risk factors associated with corporate crime in general. To focus on empirically measured and tested risk factors

[8] A naïve classifier refers to one that predict the classes randomly (i.e., predicts no better than random chance) or predict the same class invariably (e.g., predicts every case as fraud).

associated with corporate financial fraud specifically, I systematically reviewed the literature by searching in a list of academic databases (see Appendix A) for variants of the following search terms: "accounting fraud", "financial fraud", "financial reporting fraud", "financial statement fraud", "management fraud", "earnings management", "financial misstatement", "earnings quality", and "audit quality". As fraud detection research has notoriously been associated with high dimensionality (Perols, Bowen, Zimmerman and Samba, 2016), I limited my documentation of risk factors to organizational-level ones only. Appendix B represents the extensive list of risk factors that resulted from this identification process. Although the current study focuses on fraud prediction, I believe it is necessary to understand the roles these risk factors play in influencing corporate financial fraud in order to shed light on potential future interventions and provisions. As such, I also documented the conceptual explanation of the hypothesized relationships between these risk factors and corporate financial fraud. Many of the identified risk factors pertain to management's motivation and opportunity to commit corporate financial fraud.

The motivation-based risk factors identified can be conceptually subdivided into pressure-related or incentive-related motivations, explained by different sets of theories. They are also consistent with the accounting regulations' (namely, Sarbanes-Oxley Act section 404 and Statement of Auditing Standards No. 99) adaptation to Cressey's (1953) fraud triangle. Prior studies have explained the origins of pressure by applying Merton's (1983) strain theory to the corporate context (Gross, 1980; Vaughan, 1983; Simpson and Koper, 1997; McKendall and

Wagner, 1997; Clinard and Yeager, 2006; Wang and Holtfreter, 2012). These studies suggested that financial strain is constantly present in the corporate environment, and sources of pressure can stem from multiple levels. Industry-level strain may exist when an industry is declining and common resources are scarce, leading to diminished legitimate means to achieve financial goals (Clinard and Yeager, 2006). Driven by profit maximization and meeting earnings expectations, organizational strain may be generated both internally by management and externally by investors and creditors. Finally, potential offenders may experience personal-level strain that supplies motivation for fraud. Another related white-collar crime construct is Wheeler's (1992) "fear of falling". Applying the same logic as Piquero (2012) to an organizational context, corporations may be susceptible to financial pressure in fear of losing competitiveness in addition to personal losses.

Incentives refer to motivation that is driven by rewards from perpetrating the fraud. Rewards can be financial gain or intrinsic to the offender (such as reputation). A classic incentive-driven risk factor for corporate financial fraud is whether management was granted stock options as part of the compensation structure. The conflict of interest that arises between agent (management) and principal (corporation) when their goals no longer align is described in agency theory (Benson & Simpson, 2018; Shapiro, 1990).

The opportunity-based risk factors identified from literature primarily focused on guardianship. Unlike conventional street crime, corporate crime offenders often have legitimate access to targets; they also need not converge

physically in time and space (Cohen and Felson, 1979; Benson, Madensen and Eck, 2009). Due to these specific characteristics of corporate crime, target hardening has little applicability in corporate financial fraud scenarios (especially since management is afforded broad-based access). Thus, corporate financial fraud guardianship focuses primarily on the discovery of fraud. In other words, for management to successfully perpetrate corporate financial fraud, an available mechanism to conceal the fraud and avoid detection is a crucial part of the risk and reward calculus (Clarke and Cornish, 1985; Simpson and Paternoster, 2017). There are two possible ways that management may conceal corporate financial fraud —1) they may limit guardians' access to information, and 2) they may disguise manipulations as legitimate transactions (Chan & Gibbs, 2021). These tactics are not mutually exclusive and are often employed in combination. Risk factors falling into these categories represent the mechanisms through which corporate financial fraud is perpetrated.

Most risk factors identified represent the second type of concealment, where transactions are manipulated to appear legitimate. Concealment is a particularly important aspect of corporate financial fraud, as deception rather than physical threat (street crime) is used to perpetrate the crime (Benson and Simpson, 2014). By the time regulatory agencies are investigating the misconduct, the fraudulent financial statements have already successfully deceived external auditors and circumvented corporate oversights.

Management is subjected to relatively limited oversight; the board of directors and external auditors are the only primary guardians. As previously mentioned, their role is not to bar management from access to certain operational processes, but rather to grant themselves access to the same processes in order to facilitate the detection of irregularities. However, consistent with agency theory, information asymmetry exists between management and these guardians. Withholding information or limiting access to information, therefore, is a strategy of concealment that can be used by motivated offenders. Risk factors regarding guardianship generally attempt to capture guardianship effectiveness.

3.3| Machine Learning

3.3.1| *Overview*

Driven by big data and the advancement in computing power, machine learning has quickly become a popular choice for predictive analyses in scientific research (Jordan and Mitchell, 2015). As a subfield of artificial intelligence (AI), machine learning processes data to make decisions through training from examples rather than explicit programming (Chollet, 2018; Goodfellow, Bengio and Courville, 2016). This is typically done by bifurcating a sample into a training set that is supplied to the computer as examples to learn from, and a test set that is reserved for assessing how well the algorithm performs post-learning. Broadly speaking, machine learning can be classified into three categories—supervised learning, unsupervised learning and reinforcement learning. Supervised learning involves

predicting a target variable (analogous to response/ dependent variable) given a set of features (analogous to predictor/independent variables). It is particularly suited for classification and regression tasks where the target variable is labeled (Müller and Guido, 2016; Sullivan, 2017). That is, we know the classification of the target. Unsupervised learning is used to uncover hidden patterns from unlabeled data (Müller and Guido, 2016; Sullivan, 2017). In other words, it is suited for clustering analyses that help group uncategorized data into meaningful categories. Finally, reinforcement learning involves decision making through interacting with the environment at real time (Sullivan, 2017). The computer learns how to optimize their decisions given a reward and punishment system.

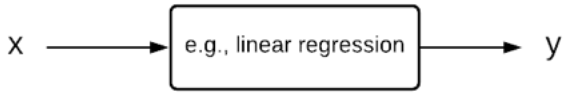
The current project employs supervised learning. In the present context, we know whether each annual financial filing is fraudulent or not. Deviating from traditional computer science programming where the programmer supplies explicit rules to classify a fraud firm from a non-fraud firm (e.g., if financial performance of the company declined more than 5% when compared to previous year, classify as potential fraud firm), machine learning allows the computer to learn and modify its algorithm based on a training set of fraud and non-fraud firm examples we provide. The researcher is able to manipulate the parameters of the algorithm to maximize or minimize any performance metrics we choose to favor, and test the algorithm's generalizability with a holdout sample of fraud and non-fraud cases that is not used in the training process. More detailed descriptions of these procedures are described under each algorithm below.

3.3.2| *Differences between Machine Learning and Inferential Statistics*

In his seminal work on the two "cultures" of statistical modeling, Leo Breiman (2001) described the differences between the statistical modeling culture and the statistical learning culture. In criminal justice and most other social sciences, the modeling culture has dominated the analytical methods. Data modelling involves the assumption of a stochastic data model (e.g., linear regression, logistic regression, Cox model) that explains the data generation process between predictor x and response y (Figure 1). Selection of the model is based on model assumptions that seem to reflect the data generating process the most (i.e., what we believe to represent reality). Parameters are then estimated from the data and inferences are drawn with regards to the population parameters in the stated hypotheses. Model validation is typically done through residual analyses or goodness-of-fit tests and generalization are based on inferential statistics such as frequentist/classical inference or Bayesian inference. There are certain drawbacks of the data modelling approach. One is that model validity is uncertain despite the goodness-of-fit tests, as predictive accuracy is not considered. Thus, if the selected model emulated the data generation process/reality poorly, conclusions made based on the model's mechanisms may be faulty. Another drawback is directed particularly to the frequentist approach of hypothesis testing that has dominated empirical research in criminal justice—namely, that p-values and confidence intervals do not provide the probability of whether the tested hypotheses were true,

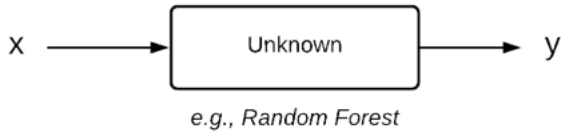
only how incompatible the data are with a specified statistical model, and only if the data generation process were to be repeated infinite times.

Figure 1. Statistical Modelling



On the other hand, the learning/algorithmic modeling culture considers the mechanism with which *x* predicts *y* as complex and unknown (Breiman, 2001) (Figure 2). Since the mechanism is treated as a black box, the focus of statistical learning is to find an algorithm that results in the best predictions based on observed data. Different sets of algorithms may be used for such an endeavor, including random forest, support vector machines (SVM) and neural networks. A model is validated by maximizing predictive accuracy and generalization is made via the training and testing processes. The main drawback of the machine learning approach is that the algorithm is less transparent than a known stochastic data model, which is well studied and understood. In other words, there is a trade-off between prediction accuracy (emphasized by machine learning) and model interpretability (emphasized by statistical modelling) (Chollet, 2018; Müller and Guido, 2016). Both techniques are adopted in this project to address two sets of research questions; one a classification task that requires accurate prediction, and another a regression task that requires certain degree of interpretability.

Figure 2. Statistical/Machine Learning



3.3.3| *Suitability for White-collar and Corporate Crime Data*

I chose to explore machine learning for several reasons. First, the two machine learning techniques I have chosen are suitable for supervised, classification tasks with a discrete binary outcome (likelihood of financial misstatement). Second, these models can accommodate complex modeling of non-linear relationships and complex interactions without the need for a priori specification (Chollet, 2018; Gromping, 2009; Hartshorn, 2016; Hastie, Tibshirani and Friedman, 2017). Since most corporate crime theories have suggested a series of complex interactions across risk factors within and between levels of analyses (e.g., Rorie, 2016; Shover and Hochstetler, 2005), this avoids the underfitting problem encountered in corporate crime research where the model falls short in capturing the complexity of the data structure and relationships (Simpson, 2013; Paternoster, 2016). Third, machine learning does not impose strict assumptions to the data. This is particularly important in modeling financial ratio risk factors, as they are likely to be highly correlated due to the nature of double-entry accounting and the theoretical underpinning of their inclusion. Lax assumptions also reduce the threat caused by the lack of degrees of freedom and statistical power that have historically prohibited the empirical analysis of corporate crime data, as they often suffer from high dimensionality and small sample size (Perols et al., 2016; Bellman,

1961). Fourth, since motivation and opportunity are both ubiquitous to corporations, understanding characteristics of firms that engaged in fraudulent acts require comparison to similar non-offending firms through designs such as case-control studies (Benson et al., 2009). Thus, the training and testing process is particularly well-suited to the task at hand.

Furthermore, machine learning has shown some promise in its application in both accounting and the criminal justice arenas. A comparison study by Duwe and Kim (2016) has shown that machine learning algorithms outperform the Burgess methodology in predicting offender recidivism. The National Institute of Justice (NIJ) has dubbed random forest models to be a new risk prediction tool that "shows great promise" in helping to prioritize probation and parole decisions when resources are scarce (Ritter, 2013). Various machine learning algorithms have been shown to successfully predict financial fraud in transaction-level data (e.g. Chan and Stolfo, 1998; Bolton and Hand, 2002).

There are many machine learning algorithms that are suitable in addressing the goal of fraud prediction. In addition to random forest and neural network, support vector machines, naïve Bayes, or various boosting algorithms are also appropriate choices for supervised, classification tasks like fraud prediction. Since part of this project is to explore how machine learning methods can help overcome some existing research challenges, I opted to explore one of the most interpretable algorithms in machine learning (random forest) and one of the least interpretable one (neural network) for comparative purposes.

CHAPTER 4. DATA AND ANALYTIC STRATEGY

4.1| Sampling Methodology

4.1.1| *Fraud Sample*

To identify fraudulent firms, data are hand collected from published SEC's Audit and Enforcement Releases (AAERs). AAERs represent enforcement records including civil lawsuits brought about by the SEC in federal courts, notices and orders and any settlement of administrative proceedings towards an individual or an organization. The target sample of fraud firms consists of ones that had violated sections 13(b)(2)(A) and 13(b)(2)(B) of the Securities Exchange Act between the years of 2002 and 2018. I began by gathering all enforcement records published between 2002 and 2018, which yielded 2,442 AAERs. As enforcement actions can be directed at internal employees and the external parties independently from the fraudulent firms, I have collapsed the AAERs involving individuals and external parties to the corresponding enforcement action. This process resulted in 643 AAERs spanning 17 years.

Since there is often a lag between crime commitment and detection by the SEC, and between detection and the enforcement publication, I examined each enforcement record and only included cases in which the crime is stated to occur after 2002. It is necessary to distinguish the period when the fraud is committed (a.k.a. the relevant period) from the date of the enforcement action in order mitigate any inconsistencies resulting from changes in disclosure, accounting, and corporate governance rules after the enactment of Sarbanes Oxley Act (2002). This resulted in

421 enforcement cases. Cases pertaining to private firms (such as CPA firms) are excluded, as financial and organizational risk factors are not readily available. This resulted in 357 publicly traded companies that can be matched to a Central Index Key. [9] Cases pertaining to fraudulent quarterly filings are also excluded to ensure fair comparison of financial information from the income statement or the cash flow statement, which are period-based statements unlike the balance sheet (a point-in-time statement). The procedure identified a fraud sample of 191 unique companies, with 450 fraudulent annual financial filings. Note that unit of analysis for the current study is company-year (i.e., filing per firm per year). The discrepancy between the number of companies and the number of fraudulent annual financial filings is explained by serial or repeat offending. Each of the 191 firms have at least one fraudulent filing; most firms have two. The maximum number of fraudulent financial reports a single firm had filed with the SEC is 10. The 191 companies are scattered across a wide range of industries, with 118 unique Standard Industrial Classification (SIC) codes. Table 1 shows that the majority of the fraudulent filings fall in the transportation and public utilities and the manufacturing sectors.

[9] Central Index Key (CIK) is a unique key to identify corporations that have filed disclosures with the SEC.

Table 1. Fraud Sample Industries *(n=450)*

Industry	**Percentage**
Wholesale & Retail Trade	1.30%
Services	4.40%
Transportation & Public Utilities	43.30%
Finance, Insurance & Real Estate	5.80%
Agriculture	8.70%
Public Administration	5.30%
Manufacturing	30.20%
Mining & Construction	0.20%

Step are taken to ensure we adhere to the definition of corporate financial misrepresentation defined in the previous section. Cases pertaining to the Foreign Corrupt Practice Act (FCPA) are prosecuted under Section 13(b) of the Securities Exchange Act (1934), and typically have substantial impact on the financial statements. Thus, they are included in the current sample of fraud firms. As corporate crimes often result in financial impacts, the sample of fraud firms will include a diverse range of offenses—including earnings management, foreign corruption, material weaknesses internal control deficiencies, misappropriation, and embezzlement. However, other forms of corporate crime such as environmental crime, securities fraud, tax fraud and racketeering that are prosecuted under different provisions or under different enforcement entities are not included in the current sample, despite how these types of corporate misconduct may also impact the accuracy of financial statements (as most corporate crime involves financial transactions to carry out and conceal the crime).

To ensure each firm that faced enforcement actions resulted in financial restatement, I cross-referenced relevant periods to restatement records from the SEC's EDGAR system. Restatement records document alterations to a corporation's financial records after its initial publication (alterations can be prompted by unintentional errors or intentional fraud). This cross-referencing process helps to identify and exclude cases in which the defending corporation had won its case against the SEC's allegations.

The examination of the relevant period also showed that lag of enforcement is indeed substantial. The latest fraudulent filing prosecuted in 2018 AAERs occurred in 2014. Amongst the 2008 AAERs, only six cases that were perpetrated after 2002 had received an enforcement action; other cases enforced in 2008 dated back to as far as 1997. Consistent with enforcement trends of other white-collar and corporate crime, enforcement of corporate financial fraud, when broken down to the firm-year level, have shown a steep decline over the recent decades (Garrett, 2020). As shown in Figure 3, enforcement have dropped from 77 cases in 2002 to 4 cases in 2014, with a brief increase in enforcement (31 filings in 2009) just after the 2008 financial crisis.

Figure 3. Corporate Financial Fraud Enforcement Trend

4.1.2| *Non-Fraud Samples*

Consistent with prior studies on fraud classification (e.g. Fanning and Cogger, 1998; Beneish, 1999; Kaminski et al., 2004), the initial non-fraud sample represents a 1:1 match, where the 450 fraudulent filings were matched to their non-fraud counterparts based on fiscal year of the fraud, industry, and company size. For each fraud filing, I first identified all the non-fraud annual filings in the same fiscal year and industry (identified by the Standard Industrial Classification (SIC) code). [10] I then selected the company that matched as closely to the fraud filing as possible in terms of company size, measured by total assets. I also matched fiscal

[10] The North American Industry Classification System (NAICS) is not used because regulatory bodies are slow in transitioning to the new standard and updating their data.

year-ends to ensure a fair comparison in reporting periods. Various proxy measures have been used to account for firm size in corporate studies, including total assets, number of employees and market cap. Yet, to date, there is no empirical analysis available to shed light on which proxy is more appropriate for the different types of corporate and organizational research questions. I have opted to use total assets here because it has been shown to be more relevant to governance measures and capital structure, which are identified risk factors of corporate fraud (Dang, Li and Yang, 2017). Its correlation with sales data is also generally weaker (Al-Khazali and Zoubi, 2005), thus having less potential risk for multicollinearity issues, given financial crime is often committed to inflate income (a difference of sales and costs).

The above matching procedures was performed without replacement—that is, once a non-fraud firm has been identified to match a fraud firm, the next fraud firm in the same fiscal year and industry will be matched with the next non-fraud firm that has not already been selected. This resulted in 447 non-fraud filings, and a total 1:1 sample of 894. Matched non-fraud filings provide benchmarks for training the machine learning algorithms to compare our fraud filings against. It is important to hold these factors constant across control (non-fraud firms) and treatment (fraud firms) groups as variations in risk factors can be specific to industry, company size and their reporting cycle. Matched non-fraud filings therefore allow us to take into consideration macro-economic conditions, seasonal financial patterns and other characteristics that are unique to specific industries, company size and reporting cycle. A 1:1 match in this manner is also akin to the

random under-sampling of the majority class (non-fraud cases), which is a frequently employed method to account for class imbalance characteristic of rare events such as fraud (Perols, 2011; Perols and Bowen, 2016). That is, it allows the computer to learn from as many fraud firms as non-fraud firms. In sum, this 1:1 matching provides the same benefits as case-control studies of rare events, which has been advocated for the study of white-collar crime (Benson, Madensen and Eck, 2009; Shadish, Cook and Campbell, 2002).

While a 1:1 matching is commonplace in corporate crime studies and has its own merits, in reality, the proportion of fraudulent to non-fraudulent annual financial filings is likely to be much smaller. Much like other forms of crime, the dark figure of corporate financial fraud is elusive. Yet, if enforcement were any indication of reality, fraudulent financial filings comprised a minuscular fraction of the total annual filings of all publicly traded companies, even at peak enforcement years. Therefore, if the goal of a machine learning algorithm is to detect fraud, it must be effective at distinguishing fraud from non-fraud even when the proportion is not 1:1. To simulate this more realistic scenario, I also created a 1:many sample, where each fraud filing was matched to all the annual filings in the same industry at the given fraud year. This resulted in 13,015 non-fraud filings company-years, and a total 1:many sample of 13,465.

4.2| Risk Factors Data Collection

Once the fraud and non-fraud firm-years were identified, annual reports (Form 10-K) filed with the SEC were obtained for both fraud and non-fraud firms for each year of the relevant period in question. For each company-year, I extracted the relevant financial risk factors from the 10-K reports using the Python scripting language.[11] Table 2 provides the list of financial risk factors used in this dissertation. These risk factors represent line items from the three financial statements in the annual 10-k report—Balance Sheet, Income Statement and Cash Flow Statement. They are required components of the annual financial report and therefore no missing data is associated with these risk factors. Zeros in these financial statement risk factor represent a true value, indicating the lack of the corresponding financial item during or as of that reporting period.

Table 3 contains the list of organizational risk factors used in the analysis. Many of these motivation-related organizational risk factors are measured with financial ratio proxies. For example, return on assets is a commonly used as a proxy measure for a company's financial health or profitability and some forms of liquidity measure involving working capital is often used as a proxy for financial distress when testing strain in the corporate setting (e.g., Wang & Holtfreter, 2012; Swchartz et al., 2021). Other opportunity-related organizational risk factors pertain to corporate governance. For instance, a CEO also serving as director of the board may indicate higher risk of conflict of interest, should there be any misalignment

[11] Financial risk factor data is not obtained from COMPUSTAT because the database backfills restated figures when they are issued.

between management self-interest and firm interest (e.g., Simpson & Koper, 1997). Companies that employ one of the Big 4 auditing firms are hypothesized to have lower risk of fraud because Big 4 firms have more resources, tend to specialize in specific industries and are more concerned with audit quality due to their need to maintain reputation (e.g., Farber, 2005).

These motivation and opportunity measures are obtained from the WRDS COMPUSTAT database or calculated from items from the 10-ks. More detailed description of each risk factor can be found in Appendix B. While they do not embody the comprehensive list of risk factors identified in the literature review, they represent those that are accessible to the public and to law enforcement, and will serve adequately in this exploratory project. Consistent with previous machine learning studies, observations with missing theoretical risk factors are list-wise removed and the remaining normalized. Sample size for the 1:1 matched sample was reduced to 760, and to 10,972 for the 1:many sample. Descriptive statistics for each of the 26 financial risk factors and 20 organizational risk factors can be found in Appendix B, along with point-biserial correlations with the binary dependent variable—fraud.

Table 2. Financial Risk Factors

FINANCIAL STATEMENT RISK FACTORS

Accounts Payable	Investments and Equivalents
Capital Stocks	Long Term Debt
Cash and Equivalents	Net Sales
Common Shares Outstanding	Property Plant Equipment
Cost of Goods Sold	Retained Earnings
Current Assets	Short Term Investments
Current Liabilities	Taxes Payable
Debt in Current Liabilities	Total Assets
Debt Issuance	Total Common Equity
Depreciation and Amortization	Total Current Liabilities
Income Before Extraodinaries	Total Inventories
Income Taxes	Total Liabilities
Interest Expense	Total Receivables

Table 3. Organizational Risk Factors

THEORETICAL RISK FACTORS	FRAUD ELEMENT	RESEARCH EXAMPLE
Audit Fees	Opportunity	*Ferguson et al. (2003)*
Auditor Change	Opportunity	*Myers et al. (2003)*
Big Four Auditors	Opportunity	*Farber (2005)*
Book to Market Ratio	Motivation	*Dechow et al. (2011)*
Cash Margin	Motivation	*Green and Choi (1997)*
CEO Duality	Opportunity	*Simpson & Koper (1997)*
Change in Cash Sales	Motivation	*Beneish (1997)*
Change in Free Cash Flows	Motivation	*Dechow et al. (2011)*
Change in Non Cash Operating Assets	Motivation	*Dechow et al. (1996)*
Change in Reveivables	Motivation	*Green and Choi (1997)*
Depreciation Index	Motivation	*Beneish (1999)*
Nonaudit Fees	Opportunity	*Frankel et al. (2002)*
Officer Change	Opportunity	*Simpson & Koper (1997)*
Retained Earnings on Assets	Motivation	*Dechow et al. (2011)*
Return on Assets	Motivation	*Wang & Holtfreter (2012)*
Sale of Stock	Motivation	*Beneish (1999)*
Soft Assets Ratio	Motivation	*Dechow et al. (2011)*
Stock Price at Year End	Motivation	*Dechow et al. (2011)*
Total Fees to Public Accounting Firms	Opportunity	*Frankel et al. (2002)*
Working Capital	Motivation	*Perols & Lougee (2009)*

4.3| Machine Learning: Application

4.3.1| *Overview of Machine Learning Procedures*

This dissertation will concentrate on using supervised learning to solve a classification problem. In supervised learning, each observation of input measures is associated with a corresponding outcome measure. The values or labels of the outcome measure guide the learning of the algorithm. The goal is to fit a model that generates the best predictions of the outcome measures based on a number of input measures. For the current research question of fraud prediction, the outcome labels (or classes) are binary, and the goal is to fit a model that most accurately classify the fraud filings into the fraud class or the non-fraud class. The following paragraphs in this section describe the general workflow of the analysis performed in this dissertation. This workflow represents the conventional procedures in machine learning, and are applied to both the random forest and neural network classifiers used in this current analysis.

All machine learning applications in this dissertation are coded in Python (version 3.8.12), using the scikit-learn library for random forest models and the Keras library for neural network models. Recall that in machine learning, we assess the generalizability of a model by applying it to unseen data that is not used to train the algorithms. To do so, the total sample is randomly split into a training set and a holdout test set. The training set was used in training the algorithm for pattern recognition, hyperparameter tuning, and cross validation. Only when the training was completed was the holdout test set used to evaluate the

generalizability of how well the learners did in classifying fraud cases against non-fraud cases. This train-test split procedure in machine learning ensures that the test set data does not influence model selection. I adopted a 75/25 split, as consistent with range of train/test split ratios in prior fraud detection literature (e.g., Lin et al., 2003). The split is stratified on the outcome measure (fraud/non-fraud) to ensure the holdout test set has the same fraud to non-fraud ratio. A seed is assigned to record the exact split to ensure future replicability.

Training the classifiers was an iterative process. As with any form of modeling, we strive to develop a well-generalized model that minimizes prediction error. Prediction error of a model can be expressed in the following equation:

$$Error(x) = \left(E\left[\hat{f}(x)\right] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right] + \sigma_e^2$$

, where the first term $\left(E\left[\hat{f}(x)\right] - f(x)\right)^2$ represents the squared bias, the second term $E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right]$ represents the variance, and σ_e^2 represents the irreducible error. Since irreducible error is unlikely to avoid and remains constant, the goal is to minimize bias and variance. In a layperson's terms, we strive to produce an algorithm that can model the true relationships between the predictors and outcome accurately (low bias), and yield consistent or precise results across different randomly drawn samples (low variance). However, for any given prediction error, we can see that there will be a trade-off between bias and variance. A model with high bias is said to be underfitting the data predictions may not be very accurate, whereas a model with high variance is said to be overfitting the data such that predictions may vary greatly across samples. Thus, striking a balance in the

bias-variance tradeoff is one of the components in the training process that is at the judgment and discretion of the researcher.

There are many factors that can impact the performance of a machine learning algorithm. One such factor is the number and quality of the input measures (or "features" in machine learning jargon). In the next chapter of this dissertation, I compare models using all the risk factors previously mentioned in Tables 2 and 3, against models that used only a subset of the risk factors that appeared to be the most important in the classification process. Another factor that can impact the performance of a machine learning algorithm is how the parameters of the algorithm are specified. Each machine learning algorithm has a different set of parameters that I can manipulate (or "tune") to determine the best combination of parameter specifications. For example, with a random forest classifier, one can specify the number of trees in the forest, how large (deep) the trees can be, how many features each tree can consider, and other characteristics discussed in further detail below. Hyperparameter tuning is the iterative process of finding the best combination of values specified for each of these parameters that control the learning process of the algorithms. To find these combinations, I performed randomized searches and grid searches on the optimal hyperparameters that will maximize the area under the receiver operating characteristic (ROC) curve in the training data. While the most common performance evaluation metric is accuracy, I favored the area the ROC curve (AUC) as a metric in the tuning process for purpose

of model comparison, especially with imbalanced classes.[12] Essentially, these searches allowed me to take advantage of the processing power of the computer to evaluate hundreds and thousands of models with different combinations of hyperparameters.

In order to ensure that the model is performing consistently, and that any variation in performance metrics is not an artifact of the train/test split, I performed 10-fold cross-validation with all the models presented in this dissertation. This is consistent with prior studies on fraud detection (e.g. Liu, Chan, Kazmi and Fu, 2015; Throckmorton, Mayew, Venkatachalam and Collins, 2015). K-fold cross-validation is a technique that involves randomly dividing the data into k groups (or "folds"; 5 or 10 folds are common choices), of approximately equal sizes (James et al., 2013). Each fold is treated as a test set while the remaining $k - 1$ folds are used to fit the model. Essentially, we are fitting and testing the model 10 times with 10 subsets of the data. We thus obtain 10 performance metrics (in our case, the AUC), one for each subset of data, allowing us to examine the variation across folds scores and providing us with an overall average.

The cross-validation technique is used throughout the iterative training phase both in hyperparameter random and grid searches and in validating the AUC scores obtained. Note that the holdout test set that is set aside at the beginning of the analysis is not used in this cross-validation procedure and remains completely unseen by the algorithm during training. Only when performance of the algorithm

[12] Evaluation metrics will be further discussed in the Results section of the dissertation.

is optimized and validated, did I finally test it on the test set. The test set AUC score for each model is compared to the corresponding average validation score. This provides added assurance that the performance metrics obtained from the test set are valid and not a result of how the data is split.

4.3.2| *Random Forest Classifier*

This section of the dissertation will address why the random forest algorithm is chosen and describe how the algorithm is trained to perform a classification task. I elected to implement random forests for the risk assessment of corporate financial fraud for several reasons. First, it is a simple to understand algorithm that has shown success in predicting offender recidivism in prior studies (e.g. Neuilly, Zgoba, Tita and Lee, 2011; Pflueger, Franke, Graf & Hachtel, 2015), but has yet been applied to other criminal justice inquiries. It has also been implemented by the Pennsylvania Board of Probation and Parole to reduce re-arrests for both violent or non-violent crime (Berk, 2017; Barnes, Hyatt, Ahlman and Kent, 2012). Second, tree-based algorithms are flexible due to their ability to create decision regions rather than a linear decision boundary (Hartshorn, 2016). Fraud detection, being "cursed" with data dimensionality problem (Bellman, 1961), can also benefit from their ability to accommodate a large number of risk factors with respect to sample size (Breiman, 2001; 2002). Finally, while random forest has once been referred to as a black box model in some criminal justice research, recent interpretation aids have been developed to address this criticism. Not only does it provide rank order feature importance, allowing us to assess whether certain risk factors are more

predictive of corporate financial fraud risk than others, it also provides decision paths that allows us to break down how decisions are made. Therefore, for both practical and educational reasons, I believe random forest to be an appropriate choice for a fraud risk assessment tool.

To understand random forest, we must first understand how decision trees work. Figure 4 is a graphical representation of what a simplified classification tree may look like in the present context. It takes each observation in our data and follows the decision arrows beginning at the root node at the top, through the branches (or internal nodes), and ends with the leaf nodes. The figure used a binary theoretic risk factor at the root as example, but continuous variables such as the financial risk factors is split by using a decision threshold (e.g., total assets < $300 million). Leaf nodes are said to be "impure" when there are a mixture of fraud and non-fraud observations. The impurity of a leaf node can be quantified with several different methods, including Gini impurity, entropy and information gain. Models from this dissertation used the Gini method, which is the default with the random forest classifier from the Scikit Learn library. The Gini impurity for a single leaf node can be computed as such:

$$1 - (probability\ of\ fraud)^2 - (probability\ of\ nonfraud)^2$$

Figure 5 shows an example of this calculation. It also shows the total Gini impurity of the risk factor, which is the weighted average of the two leaves. The goal in general, is to minimize the impurity. Therefore, if one risk factor is insufficient in creating a pure classification of fraud and non-fraud filings, the node is split again

with a second risk factor, and a third and so on until we are satisfied with the level of impurity. However, what constitute a satisfactory level of impurity is an example of the bias-variance tradeoff issue discussed previously. A large tree with many risk factors may produce clear-cut classifications with pure leaves at the end, but may overfit the data and generalize poorly on unseen data. Even with hyperparameters tuning such as limiting tree depth and maximum number of splits, decision trees are said to be prone to overfitting. Random forest, which consists of many decision trees instead of just one, helps with this problem (Breiman, 2001).

Figure 4. A Single Classification Tree

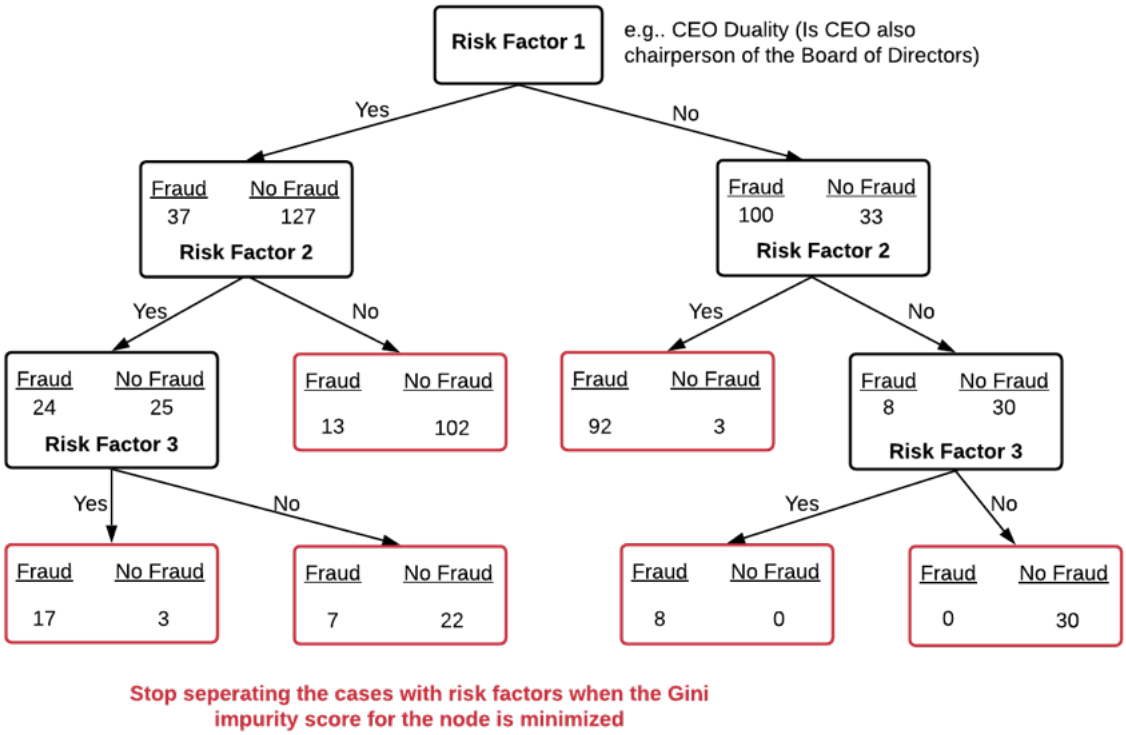
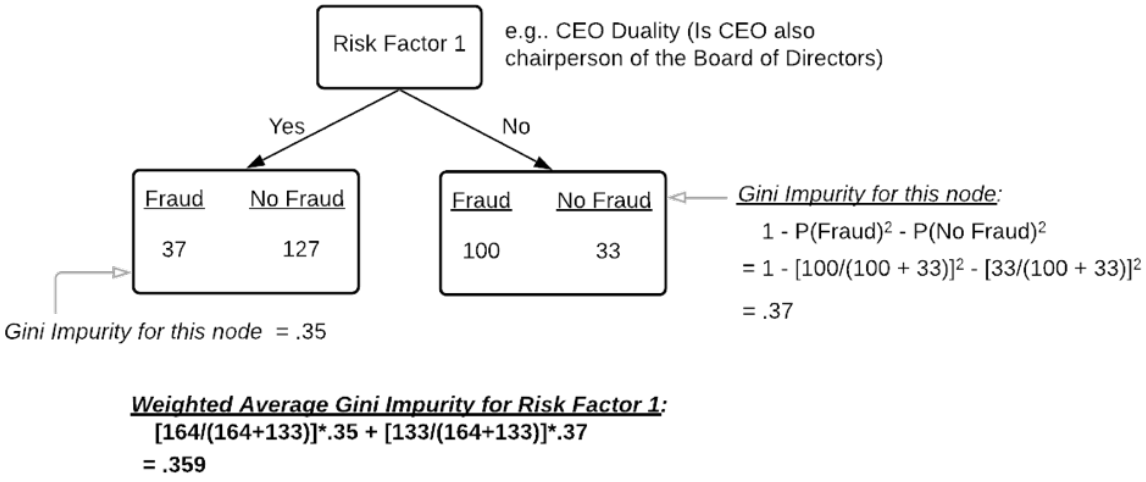


Figure 5. Gini Impurity Calculation



Random forest is an extension of an ensemble learning method known as bootstrap aggregation (a.k.a. bagging). The main idea behind bagging is to combine multiple high in variance but relatively unbiased learners to reduce overall prediction error, with each learner using a bootstrap sample (randomly drawn with replacement) of the training set (Hastie et al., 2017). In the case of a random forest, a number of decision trees are created using bootstrap samples that is the same size as the training set. To avoid having overly similar trees within the ensemble, random forest further introduces randomness by only allowing a subset of the input features to be considered at each node. Each individual tree in the ensemble is considered independent of each other, and will operate as it would singly, by determining features and split points to minimize impurity. Random forest then produces the final prediction by aggregating the results from each tree through a majority voting mechanism for classification tasks (or by averaging for regression tasks).

There are many types of decision trees—e.g., ID3, C4.5, CART, MARS. The base learner I used in the random forest classifier is CART (classification and regression tree), with impurity computed using the Gini criterion. CART is chosen for its ability to accommodate the diverse distributional characteristics of financial data (including the accommodation of outliers), and it's ability to predict financial distress in prior studies (e.g., Salehi & Fard, 2013, Chen, 2011). Random Forest has quite a few hyperparameters to optimize—the number of CARTs in the forest, the maximum number of input features considered at each node, the minimum observations required in a leaf, the number of observations required to and the split a node further. To bring some structure to the hyperparameter tuning process, I first obtained a baseline level of each model's performance using the default settings for these hyperparameters from the Scikit Learn random forest classifier. Then to narrow down the search, I performed a randomized search using a parameter grid with a range of values for each hyperparameter. A randomized search attempts to optimize an evaluation metric by testing random combinations of the range of values for each parameter.[13] Then, based on the best parameters of the randomized search, I further fine-tuned the hyperparameters with a manual grid search. A manual grid search requires the researcher to stipulate a parameter grid of specific values for the hyperparameters. Both of these searches are performed using 5-fold

[13] For example, I can specify the randomized search to generate 5 random numbers between 20 to 200 for the number of trees in the random forest. Then, specify it to generate 5 random numbers between 1 to 25 for the number of features considered in each tree. The random search will consider the $5^5 = 3125$ models to find the best combination of these two hyperparameters. Something that is difficult to achieve without leveraging computational power of modern processors.

cross validation, with AUC being the metric to be optimized. The out of bag error rate generated from the bootstrap sampling procedures was also consulted as our guide to determine the optimal number of CARTs to include in the ensemble. The best combination of hyperparameters from these searches is used for a final 10-fold cross validation of the AUC score to assess its reliability. The final models are then tested on the holdout test set and detailed decomposition of the confusion matrix along with various metrics are reported in the results section of this dissertation. Despite of the availability of the out of bag sample, I elected to separately retain a holdout test set for model evaluation as other algorithms used for comparison (e.g. logistic regression) do not necessarily have an out of bag sample from the bootstrap aggregation process.

4.3.3| *Neural Network Classifier*

This section of the dissertation will address why the neural network algorithm is chosen and describe how the algorithm is trained to perform a classification task. The neural network was chosen for the fraud detection tool for several reasons. First, most white-collar criminologists have hypothesized corporate crime as a result of complex interactions between opportunity-based risk factors and between opportunities and motivations (Coleman, 1987; Shover & Hochstetler, 2006), many of which cannot be directly observed. The hidden layers of deep networks are apt to capture these unobservable interactions that are present in the real world. Second, previous studies have shown success in varying forms of neural network in predicting fraud (e.g. Fanning & Cogger, 1998; Lin, Hwang & Becker,

2003), but the risk factors included were limited to selected financial ratios, the samples used were prior to substantial regulatory changes, and the networks were limited to one hidden layer. Third, deep networks have great scalability both in terms of sample size and model complexity (Chollet, 2018). While our analysis may be limited to the data available to us currently, it is easily adaptable to incorporate larger scale data, more detailed level of analysis and any other risk factors identified in the future. Finally, neural network provides an excellent contrast against random forest, the former being one of the least interpretable and the latter being the one of the most interpretable machine learning algorithms. Comparing these algorithms will serve as an exploratory analysis into how diverse forms of machine learning methods compare to our standard approach of logistic regression in the context of fraud detection.

Figure 6 represents an illustration of a neural network in the present context. An artificial neural network is made up of many neurons arranged in layers. The far most column of neurons represents the input layer, where each neuron represents the value of a risk factor. The middle column of neurons represents a hidden layer, capturing the interactions between the input measures. The right most column of neurons represents the output layer, which is made up of the two classes of our outcome measures. The arrows between neurons carry weights that are to be estimated. Neural networks make predictions by computing the dot products of each neuron, and then applying an activation function to capture the nonlinearity in the data. Figure 7 demonstrates the calculation for a single

neuron in the hidden layer. This process of taking values from the input layer and moving through the hidden layers to make a prediction at the output later is known as forward propagation.

Figure 6. Neural Network Structure

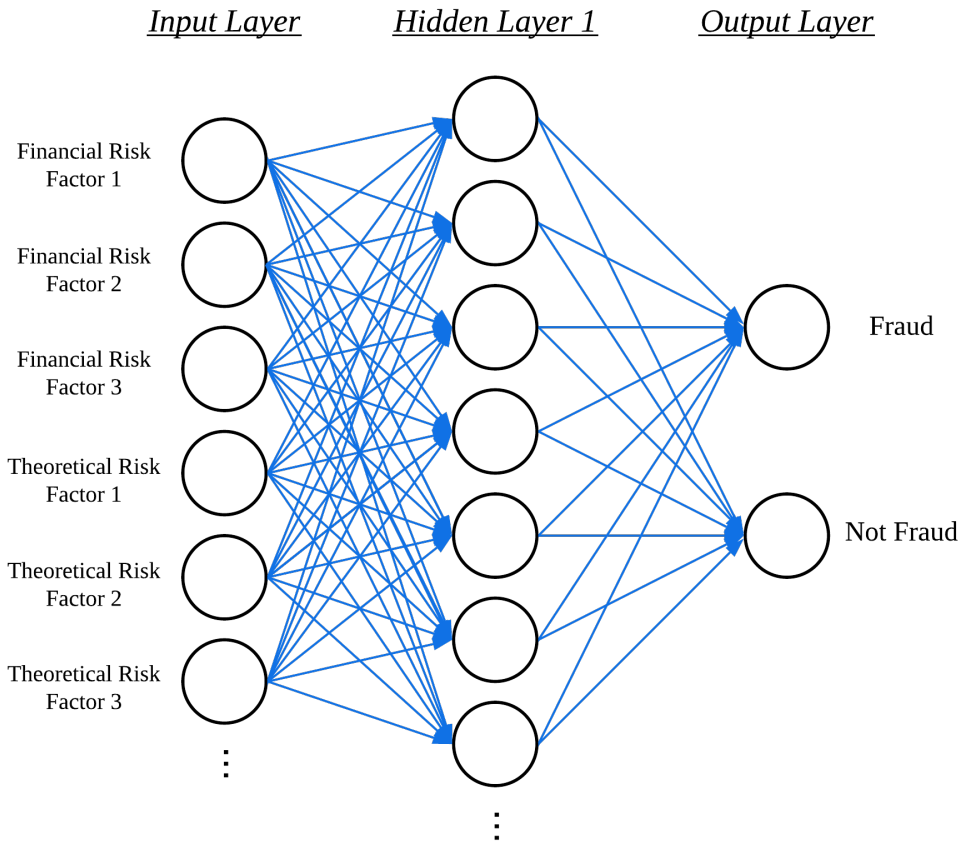
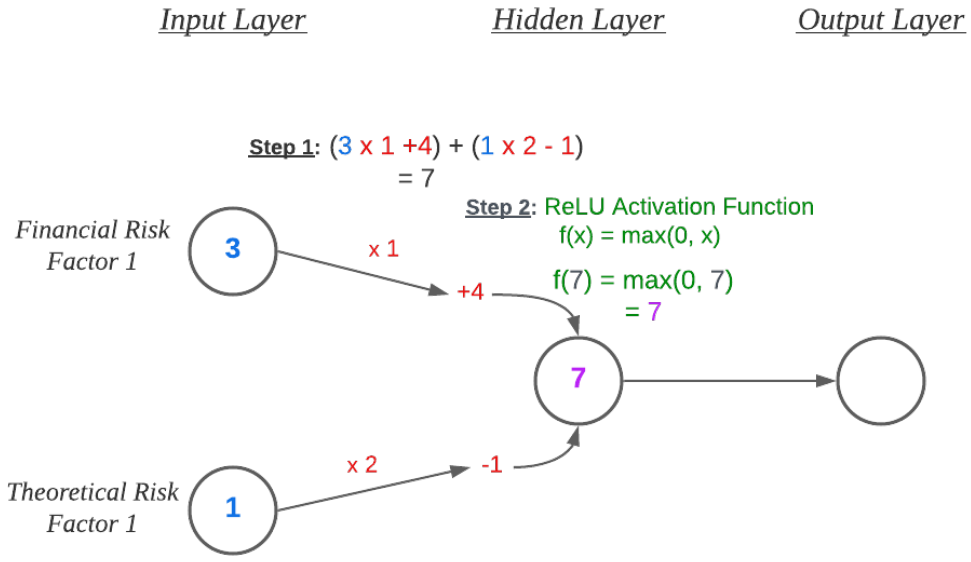


Figure 7. Forward Propagation



The goal of training is to determine the combination of weights and biases associated with each arrow in the diagram to produce the best prediction for the fraud and non-fraud class. Since we are capturing many interactions between many input features and hidden nodes, one can see that a neural network must estimate many weights and biases. Much like how weights are optimized in linear regression by minimizing the sum of squared residuals, a neural network estimates the weights by minimizing a loss function. Different loss functions are used with different data types, but the way a neural network optimizes any loss functions is with an algorithm called gradient descent.[14] In introductory calculus courses, we were taught to find the minima of functions by taking the derivative and setting

[14] Technically, the neural network models in this dissertation uses stochastic gradient descent, which performs the same procedures as gradient descent, but in batches of randomly selected samples from all the observations. To simplify the explanation, I left out the detail of how batches maximize efficiency of the algorithm in terms of convergence time, but the batch sizes used are reported for each model in the results section.

that to zero. Gradient descent takes a similar but slightly different approach; it makes an initial guess of the value of the local minimum and take steps towards it (larger steps when the guess is further away from and smaller steps as the guess is closer to zero). This small difference makes gradient descent a powerful optimizer in a wide range of scenarios where derivative equals zero is not possible to solve.

The slope for a single node can be computed by multiplying three components: 1) the slope of the loss function with respect to the value of the node of interest, 2) the value of the nodes that feeds into our weight, and 3) the slope of the activation function with respect to the value of the node of interest. To move this slope towards the lowest point of the loss function, we compute a new slope by subtracting the old slope by a small fraction of itself. This small fraction is known as the learning rate. In other words, $New\ Slope = Old\ Slope - (Old\ Slope \times Learning\ Rate)$. A small learning rate prevents us from missing the slope. This is done iteratively until we minimized the loss function. However, recall that a neural network must estimate many weights and biases simultaneously. A gradient represents the same concept as derivatives/slopes, but to a function of multidimensional inputs. A neural network uses an algorithm called backpropagation to compute said gradients with an efficient implementation of the chain rule in calculus. This is what makes deep learning (i.e., neural networks with multiple hidden layers) feasible. To compute the gradients, backpropagation takes the prediction error from the output layer obtained from the previous procedures, and propagates it backwards through the hidden layers to the input layer. Note that

backpropagation is often mistaken for the algorithm used to train the neural networks, but it is the automatic differentiation algorithm used to compute the gradients that are then used by an optimization algorithm like stochastic gradient descent in the learning process (Goodfellow, Bengio & Courville, 2016).

All neural networks in this dissertation are built with the Keras library, which is an extension to a popular open-source machine learning platform named Tensorflow. The optimizer used is "Adam", which is a form of stochastic gradient descent that has an adaptive instead of a fixed learning rate. The hyperparameters in a neural network are either related to the network structure or the training algorithm. Network structure hyperparameters include the number of layers, number of neurons in the layers and the activation function. Hyperparameters related to the training algorithm include the learning rate, number of epochs and batch size. The input layer of the neural network in this dissertation will correspond to the number of input measures used in the model. The output layer will contain two nodes, one of each class of the outcome measure. The two nodes in the output layer, as well as the softmax activation function corresponds to the loss function I opted for my neural networks—sparse categorical cross-entropy. Sparse categorical cross-entropy is appropriate for any multiclass classification problems (Géron, 2019). It computes the same error as the cross-entropy loss function (which is similar to that of the log loss function) and yields the predicted probability for each class in the output layer. The number of hidden layers of the neural networks presented in this dissertation are limited to two for feasibility and because two

hidden layers are theoretically all that is necessary to represent any functional forms (Heaton, 2008). Note that unlike random forest, the structures of the hidden layers along with learning rate, number of epochs and batch size are tuned to minimize the loss function, not a specific performance metric.

Validation for neural networks is often done with a validation set that is separate of the training and the test set. That is because 1) neural networks or deep learning are often applied to a large dataset with hundreds of thousands of observations and a k-fold cross validation for every model will be extremely computationally expensive and time consuming, 2) practically speaking, the number of epochs and batch size used in k-fold cross-validation would be different, as those are a function of the total number of observations. Nevertheless, I performed 10-fold cross validation for the neural network models presented in this dissertation for consistency reasons.

CHAPTER 5. RESULTS

This chapter reports findings to the research questions of this dissertation project:

1. Can the multi-disciplinary risk factors identified by research be used to predict corporate financial fraud with the use of a random forest classifier (i.e., does the algorithm perform better than a naïve classifier[15])?
2. How does the random forest classifier perform in comparison to commonly employed prediction tools (e.g., logistic regression)?
3. Which of the multi-disciplinary risk factors are most important in predicting corporate financial fraud?
4. Can the multi-disciplinary risk factors be used to predict corporate financial fraud with the use of a deep neural network classifier? (i.e., does the algorithm perform better than a naïve classifier)?
5. How does a neural network classifier perform in comparison to logistic regression and the random forest classifier?

In other words, I sought to determine whether the risk factors identified in previous research can be used to predict corporate financial fraud with the use of machine learning. Specifically, I wished to examine how a random forest classifier and a neural network classifier performed compared to the more commonly employed logistic regression. I will first present my findings in this chapter, and

[15] A naïve classifier refers to one that predict the classes randomly (i.e., predicts no better than random chance) or predict the same class invariably (e.g., predicts every case as fraud).

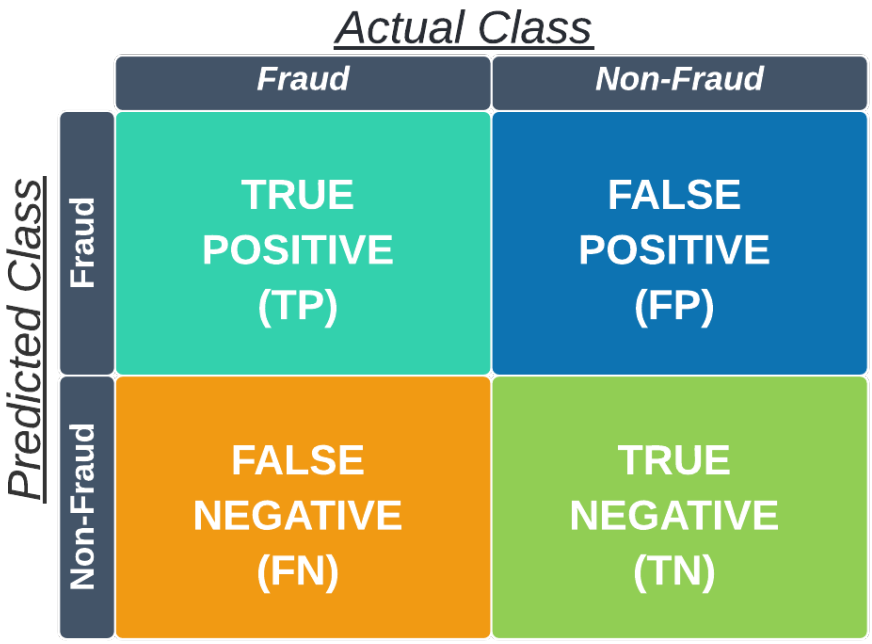
discuss the meanings and implications of the findings in more detail in the following chapter (Chapter 6).

5.1| Performance Evaluation Metrics

Before we examine results from the different algorithms, it may be prudent to discuss the various forms of performance evaluation metrics used in machine learning. If one were to examine the metrics documentation of the Scikit Learn library, one would discover a list of almost 40 metrics[16]; almost half of them relate to classification tasks and the rest regression and clustering tasks. Despite the overwhelming number of classification metrics, almost all of them can be traced back to the confusion matrix. A confusion matrix is a matrix that compares the predicted classes from the algorithm to the true classes, as shown in Figure 8. In the present context of fraud detection, a true positive represents a fraud case that has been correctly classified by the algorithm as such. A true negative represents a non-fraud case that has also been correctly classified as such. A false positive represents a non-fraud case that has been incorrectly classified as a fraud case; in other words, it is a false alarm that is synonymous with type I error in hypothesis testing. A false negative represents a fraud case that is incorrectly classified as a non-fraud case; in other words, it is a missed detection that is synonymous with type II error in hypothesis testing.

[16] https://scikit-learn.org/stable/modules/classes.html#sklearn-metrics-metrics

Figure 8. Confusion Matrix for Fraud Classification



From the four values in the confusion matrix, we can compute some of the most widely use classification metrics in machine learning:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

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

Accuracy, perhaps the most commonly reported metric, allows us to assess how well an algorithm does overall at correctly identifying the truth (both positives and negatives). If we refer to the confusion matrix, the denominator of the sensitivity

(recall) formula is made up of the two cells in left column of the matrix (i.e., the actual fraud cases). In other words, it tells us how well an algorithm does in identifying the total positive cases (as false negatives are actual positives). The denominator of the precision formula speaks to the top two cells of the confusion matrix (i.e., the predicted positives). In other words, it tells us how correct an algorithm is when it predicts a positive outcome. The denominator of the specificity formula refers to the two cells on the right of the confusion matrix (i.e., the actual non-fraud cases), thus telling us how well an algorithm does in identifying the total negative cases.

The ROC curve summarizes much of this information in the form of a graph. It plots an algorithm's false positive rate (x-axis) against its true positive rate (y-axis). The false positive rate can be computed by $1 - Specificity$. The true positive rate is synonymous with sensitivity/recall. The ROC curve shows the trade-off between correctly predicting the positive class (fraud) and incorrectly predicting the negative class (non-fraud). If a naïve classifier were to predict the classes randomly or to predict the same class invariably, it would be represented by the diagonal line stemming from the origin of the graph, and the area under that diagonal line would equal .5. We refer to it as a naïve classifier because it shows no ability in discriminating between positive and negative classes. A skilled classifier would perform better at distinguishing the two classes, with the fraction of correct positive predictions close to 1 and the fraction of incorrect negative predictions close to 0,

thus generating a curve above the diagonal line, and having an AUC of greater than .5.

These classification metrics are not only helpful in model comparison, but also in hyperparameter tuning and in the consideration of real-world applications. Aside from a perfect system, error is unavoidable and thus there must be a trade-off between type I and type II error. Even within the domain of fraud detection, users of the algorithm may have different goals and different resource constraints. If a fraud detection algorithm is designed to flag credit card fraud such that bankers can follow up with owners of the credit cards, a higher false positive rate may not be problematic or may even be desired. On the contrary, given how our data has shown how long it takes to investigate and prosecute a corporate financial fraud case, a law enforcement agency such as the SEC that has high resource constraints may not be able to afford having a lot of false alarms.

5.2| Results from the 1:1 Sample

Table 4 shows the random forest and the neural network models in comparison to logistic regression. These models are trained with all the input measures listed in Table 1 and 2, and the ones with the best AUC after the iterative training process are reported.[17] The hyperparameters that are used to optimize the AUC scores are also reported. Recall that these reported metrics are based on

[17] Note that company central index key is not included in the models as an input measure despite the nested structure. Most firms only have two fraud filings or less, and the intraclass correlation at the firm-level is less than .08 per the unconditional (null) model.

testing the tuned and cross-validated models on the completely unseen test data from the 25% holdout sample. This holdout sample contains 190 observations, 99 of which are fraud cases. 10-fold cross validation results for the AUC are reported in parenthesis. The other classification metrics discussed above are reported along with the deconstructed confusion matrices.

Table 4. Classification Results from Random Forest and Neural Network *(n=760)*

	Random Forest	**Neural Network**	**Logistic Regression**
Model Specification	max depth: 60 max features: 5 min samples leaf: 1 min samples split: 3 n: 120	hidden layer 1: neurons=30 activation=softmax batch size: 50 epochs: 250	n/a
TP	76	76	79
TN	66	27	28
FP	25	64	63
FN	23	23	20
Accuracy	0.747	0.542	0.563
Sensitivity	0.768	0.768	0.798
Precision	0.752	0.543	0.556
Specificity	0.725	0.297	0.308
AUC (CV)	0.815 (.812)	0.614(.602)	.688(.695)

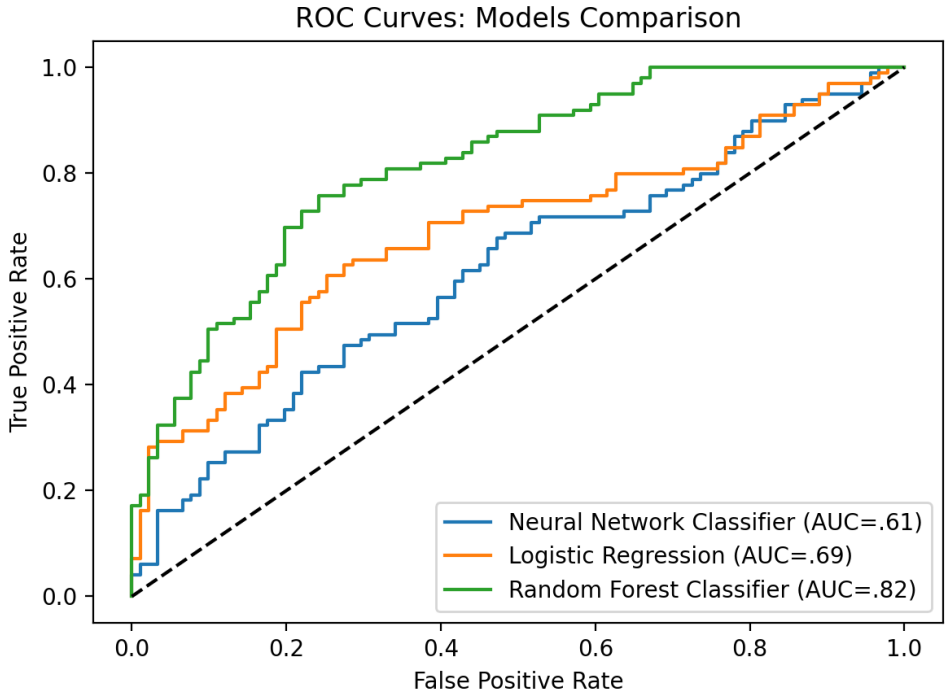
With respect to the research questions 2 and 5, both the random forest and the neural network algorithms performed more effectively than a naïve classifier in this 1:1 matched sample, with AUCs of .815 and .614 respectively. In comparison, the logistic regression model resulted in an AUC of .688, performing slightly more effectively than the neural network model but not as well as the random forest model, per research questions 3 and 6. The overall accuracy scores of these models are in alignment with the AUC—the random forest algorithm performed best in

correctly classifying 75% of the unseen test cases, while neural network (with a single hidden layer and 512 trainable parameters) and logistic regression correctly classified 54% and 56% of all the test cases.

When the random forest algorithm predicts a test case to be fraudulent, it is correct 75% of the time, as compared to 54% of the time by the neural network and 56% by logistic regression. This precision metric is particularly important to consider when implementing a fraud detection algorithm in real life. Since there is often a finite amount of resources, users of the algorithm must decide how many false positives is tolerable depending on their goals. Given the nature of a binary outcome, users of a classifier must weigh the cost of lower precision against the benefit of higher sensitivity/recall, as there is always a trade-off. Despite scoring the best at overall accuracy, precision and specificity, the random forest algorithm is outperformed by logistic regression in recalling fraud cases. Specifically, logistic regression was able to identify 80% of the fraud cases in the test set, whereas the random forest and the neural network algorithms were only able to identify 77% of them. Since we know that logistic regression performed well in identifying the fraud cases but was only marginally better than random chance in overall accuracy, we can expect that it performed poorly in identifying the non-fraud cases, as confirmed by its specificity score of .308. This was also the case with the neural network model, which only identified less than 30% of the non-fraud cases. In contrast, the random forest algorithm had a more balanced tradeoff between type I (false positive) and type II errors (false negative).

The overlayed ROC curves of all three algorithms shown in Figure 9 represent a graphical summary of the results reported above. The random forest classifier performed the best overall, yielding a ROC curve that is furthest away from the random chance (represented by the black dotted line) and yielding the greatest area under the curve. Since it was able to distinguish fraud and non-fraud cases proportionately well, it also exhibited a more symmetrical ROC curve, whereas the ROC curves for logistic regression and neural network were slightly skewed in comparison. The average 10-fold cross-validation scores for the AUCs are less than .02 from the test scores, suggesting results presented here are relatively consistent and that the model did not overfit the data. Not only do the AUCs allow us to conveniently compare between models, they ROC curves also allow us to explore the optimal thresholds for classification, depending on the application and the ultimate goal of implementing the algorithms. To ensure consistent comparison, the threshold for classification is set at .5; that is cases of assigned probabilities of greater or equal to .5 were classified as fraud, and cases with probabilities below .5 were classified as compliant. The changing of threshold level will be discussed further in Section 5.3 when dealing with imbalanced data.

Figure 9. ROC Curve and AUC for Models Comparison



As mentioned in the previous chapters, one of the reasons random forest was chosen for this project was due to its reputation for being one of the more interpretable machine learning algorithms. Once the algorithm has been trained, "feature importance" can be extracted from the trained model. Figure 10 represents a bar chart of the top 25 features that are considered most "important" in the random forest classification process. Feature importance is operationalized as the reduction in the Gini impurity index when the feature is used to split a node, averaged across all trees in the forest[18]. This average reduction in impurity identified *common equity,* as the financial statement (line item on the balance

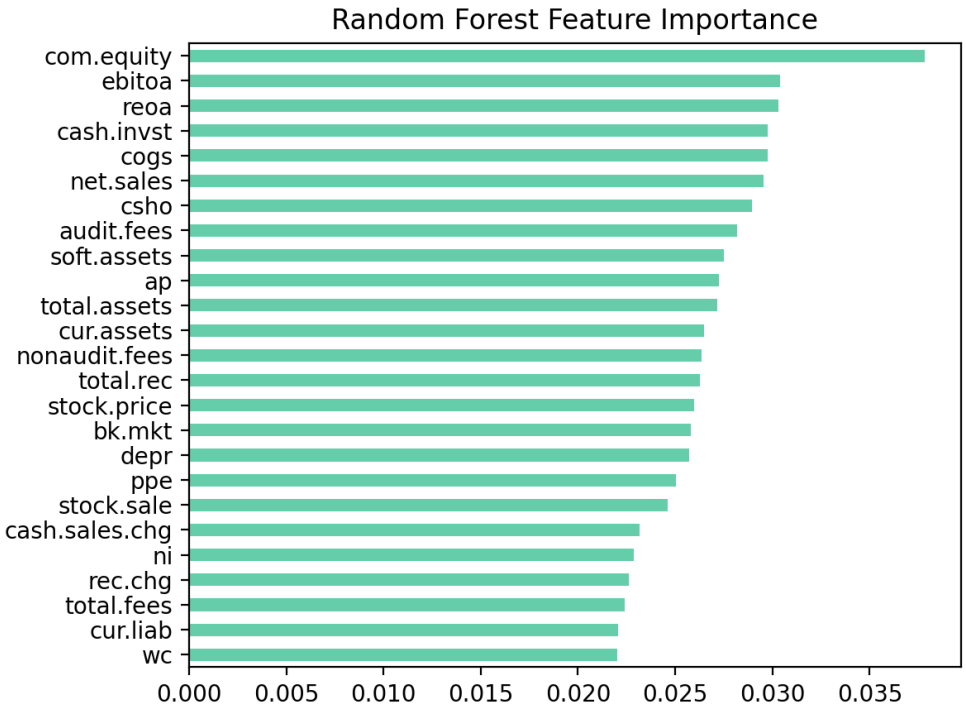
[18] The source code for how the Sciki Learn library computes feature importance function can be found here: https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/tree/_tree.pyx#L1056

sheet) risk factor that was used most frequently in the classification process. *Return on assets* was identified as the most important motivational proxy measure; it represents a financial ratio used in financial statement analysis to assess a company's performance in generating profits with its resources. The most important opportunity proxy measure was *audit fees*, which is a proxy measure for audit quality.

Upon examining the point biserial correlations of these features with outcome variable (fraud), only return on assets had a statistically significant positive association and even then, a very weak one ($r = .098$, p-value $< .05$). Common equity and audit fees have even weaker positive associations (.045 and .070, respectively) that are not statistically significant. This provides some insights as to why *return on assets* is frequently used in empirical studies of corporate crime as a proxy indicator for profitability (e.g. Wang & Holtfreter, 2012; Schwartz et al., 2021). Being a special case of Pearson's correlation for a binary variable, point biserial correlation is based on the assumption that the two variables have a linear relationship, whereas the other two variables may have a more complex relationship with fraud. For example, *audit fees*, as a proxy measure for audit quality is often associated with auditor's capability of detecting fraud (i.e., a protective factor), yet it can also be interpreted as more complex audit requiring more staff-hours (i.e. a risk factor). The random forest algorithm may be more attuned to identifying non-linear relationships such as this. Yet it is important to note that computation of feature importance does not take into consideration

collinearity among input measures. As shown in Figure 10, the second and third most relied upon measures are return on assets and retained earnings on assets, two highly related numerators divided by the same denominator. Since only a subset of features are considered in every tree in the random forest, two highly correlated features can both be identified as important. Put differently, even features that are considered important can be redundant in its contribution to the algorithm's ability to make a classification.

Figure 10. Feature Importance from Random Forest (n=760)



5.3| Feature Selection & Importance

In addition to hyperparameter tuning, one of most effective ways to improve an algorithm's classification performance is through feature selection. That is, to reduce the number of input measures by removing those that either do not

contribute too much to the algorithm's ability to classify, or are redundant in relation to another input measure. As briefly demonstrated earlier, redundancy is an especially prevalent issue with the use of financial ratios as proxy measures, as many of the financial ratios are computed with the same financial statement line items. While multicollinearity issues do not violate any assumptions in machine learning unlike traditional statistical methods, they do impact an algorithm's performance. In this section, I answer research question 4 of this dissertation, by exploring whether more parsimonious subsets of our input measures can help improve performance of the algorithms.

Since random forest appeared to be the overall best performing algorithm in the fraud classification task, I used it to compare the five different subsets of risk factors. Predictive results using the same tuning procedures and evaluation metrics are outlined in Table 5. The first subgroup of input measures (subset 1) contains the financial statement line items only. A comparison of the overall performance metrics showed that using a subset of financial features only improved the performance of the random forest classifier; AUC increased by 5 points (from .815 to .865) while overall accuracy increased by 2 points (from 75% to 77%). Without the organizational proxy measures, the financial statement line items were able to improve identification of the non-fraud cases (i.e., increased specificity) without compromising the ability to identify fraud cases (i.e., little change in sensitivity), thereby also improving precision and overall accuracy.

I then investigated the remaining organizational proxy measures only in subset 2. The results in Table 4 showed these proxy measures are not as powerful predictors as the financial ones and have reduced the classification performance of the original model with all the features. AUC declined 2 points (from .815 to .792) and overall accuracy also declined by 2 points (from 75% to 73%). The organizational proxy measures did not perform as well at identifying fraud cases, but was especially wanting in its identification of non-fraud cases (with specificity dropped by 3 points compared to the full model and 9 points compared to the financial measures only model). I attributed this decline in performance to the financial ratios used as proxies for much of the motivational measures. As such, I tested a subset of selected financial statement line items and the opportunity measures in subset 3, excluding all financial ratio variables.[19] Prediction results from subset 3 showed that the combination of financial line items and opportunity measures without ratio proxies yielded even greater improvement in accuracy, sensitivity, and precision to the original model than with financial measures only. This mixed group of selected financial and opportunity measures improved upon identifying non-fraud cases, but not as much as the comprehensive set of financial measures alone. Hence, AUC improved 4 points from .815 to .859 as opposed to 5 points with subset 1. Taken as a whole, these three subsets lend support to the need

[19] Subset 3 contains the following input measures: common equity, interest expense, long term debt, accounts payable, total liabilities, total assets, total inventories, current assets, current liabilities, sale of stock, total receivables, property plant and equipment, cash and equivalents, net sales, cost of goods sold, common stock outstanding, debt in current liabilities, audit fees, non-audit fees, big four external auditor, auditor change, officer duality and officer change.

to reexamine the use of financial ratios as motivational proxy measures, which will be discussed further in the following chapter (Chapter 6).

Whereas the first three subsets of features examined above were guided by theory and domain knowledge, the next couple of subsets aimed to investigate common features selection methods used in machine learning. Variables used in subset 4 is produced by a feature selection tool from the same Scikit Learn library used for training and testing the machine learning algorithms. First, chi-squared tests were used to assess the dependencies between the categorical input measures and the outcome measure (fraud), and f-tests to assess the variances for continuous financial measures. Based on those tests, the feature selection tool assigns a score of range 0 to 1 to each feature. I selected all the input measures that received a score of greater than .5. I further removed 3 input measures that may be redundant from an accounting perspective. Note that variables identified by this features selection tool consist primarily of financial statement line items and opportunity measures with only a few financial ratios. [20] Most financial ratio measures received scores of less than .5. Aside from selecting features based on dependencies and covariances, another popular method to prune the variables is to use the feature importance scores from the original model as a guide. For comparison purposes, I constructed subset 5 to contain the top 25 features as shown in Figure 10.

[20] Subset 4 variables: soft assets ratio, interest expense, long term debt, big four auditors, accounts payable, total liabilities, total assets, auditor change, total inventories, audit fees, current assets, sale of stocks, current liabilities, total receivables, non-audit fees, property plant and equipment, cash and equivalents, debt in current liabilities, net sales, cost of goods sold, retained earnings on assets, common shares outstanding.

AUCs for both subsets are greater than that of the full model, once again lending support to the premise that a more parsimonious set of features can help reduce noise and improve predictions. Subset 4 yielded the best overall accuracy of all the subsets and was able to identify over 80% of all fraud cases in the test set with the precision of 78%. It fell slightly short on specificity (.758) but still improved upon the full model and the model with all the organizational proxy measures (subset 2). Subset 5 did similarly better at identifying fraud cases but not at non-fraud cases, identifying 78% of the former and 73% of the latter. The overall accuracy of subset 5 (75%) and AUC (.83) suggests that sole reliance of the feature importance as a selection method may not yield the most effective improvement in classification algorithm. Nonetheless, once the algorithm is trained, it offers some helpful insights.

Figure 11 shows which risk factors are most favorable by the algorithm when making predictions from Subset 4 (model with highest accuracy). Many of the top predictors appeared to lend support to existing empirical studies of financial fraud. For example, *total assets* was identified as the top predictor once *common equity* has been removed, which is consistent with prior empirical studies of corporate crime (e.g. Beasley, 1996), but also raise questions about its validity as a proxy measure for firm size. Net sales (the second most important predictor) and retained earnings over assets (the fifth predictor) are also consistent with findings regarding firm profitability and likelihood of committing financial fraud (e.g. Erickson, Hanlon & Maydew, 2006). *Non-audit fees*, being identified as the topmost important

opportunity proxy measure, lends support to continuous criticisms for the lack of auditor independence (Davis, Soo & Trompeter, 2003). However, it is important to be reminded that feature importance merely tells us how much class discriminatory information each feature contains, it does not seek to establish directional relationships between these input features and fraud. In other words, while these features are predictive of fraud, the mechanisms and reasons (i.e., the how and why) remain unknown. It is also noteworthy to point out that many of financial distress related measures (such as debt or liquidity risk factors) are ranked lower in importance than financial health measures (such as assets and profitability risk factors), as prior studies have used financial ratios as proxy measures to indicate financial distress and health in testing corporate strain (e.g. Schwartz, 2021).

Table 5. Random Forest Models with Risk Factor Subsets[21]

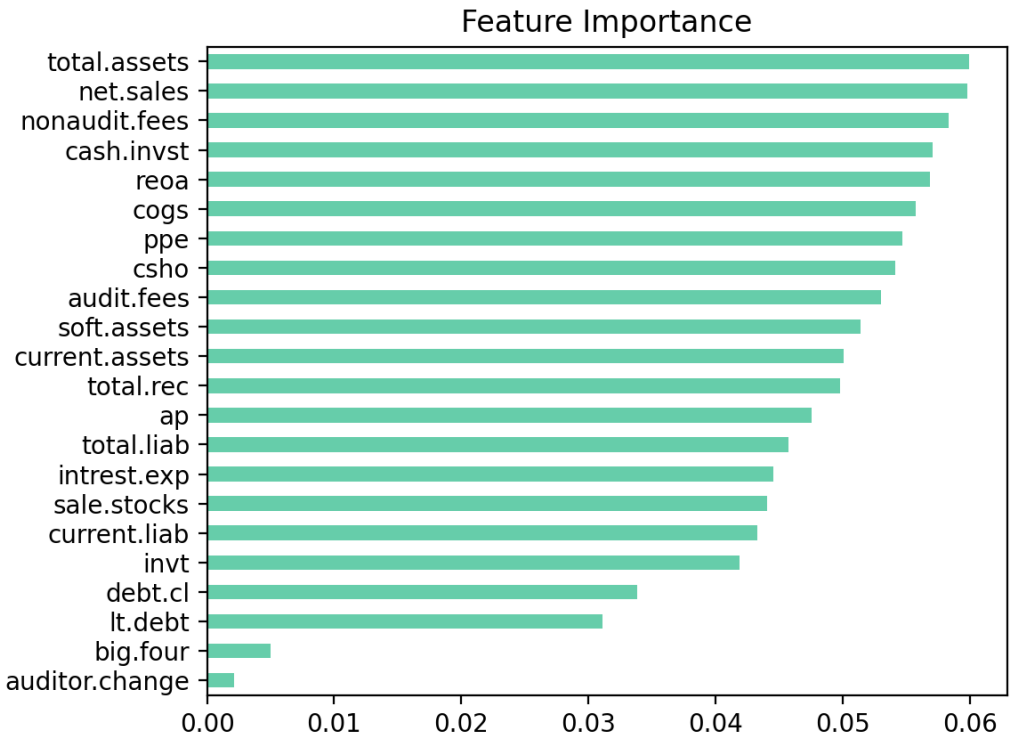
	Random Forest		
	Subset 1	**Subset 2**	**Subset 3**
Model Specification	max depth: 70, max features: 6, min samples leaf: 1, min samples split: 2, n: 120	max depth: 50, max features: 5, min samples leaf: 2, min samples split: 4, n: 200	max depth: 95, max features: 3, min samples leaf: 1, min samples split: 3, n: 273
TP	86	75	78
TN	87	63	70
FP	24	28	21
FN	27	24	21
Accuracy	0.772	0.726	0.779
Sensitivity	0.761	0.758	0.788
Precision	0.782	0.728	0.788
Specificity	0.784	0.692	0.769
AUC (CV)	0.865(.842)	0.792(.729)	0.859(.840)

[21] Subset 1 is trained and tested on a sample of 894 while subsets 2-5 is trained and tested on a sample of 760. This is due to financial measures being more reliably available for public traded companies than organizational risk factors (please refer back to Chapter 4 for the handling of missing data).

Table 5 (cont'd)

	Random Forest	
	Subset 4	**Subset 5**
Model Specification	max depth: 10, max features: 3, min samples leaf: 2, min samples split: 3, n: 277	max depth: None max features: 3, min samples split: 4, min samples leaf: 3, n: 282
TP	80	77
TN	69	66
FP	22	25
FN	19	22
Accuracy	0.784	0.753
Sensitivity	0.808	0.778
Precision	0.784	0.755
Specificity	0.758	0.725
AUC (CV)	0.856(.820)	0.830(.820)

Figure 11. Feature Importance from Subset 4[22]



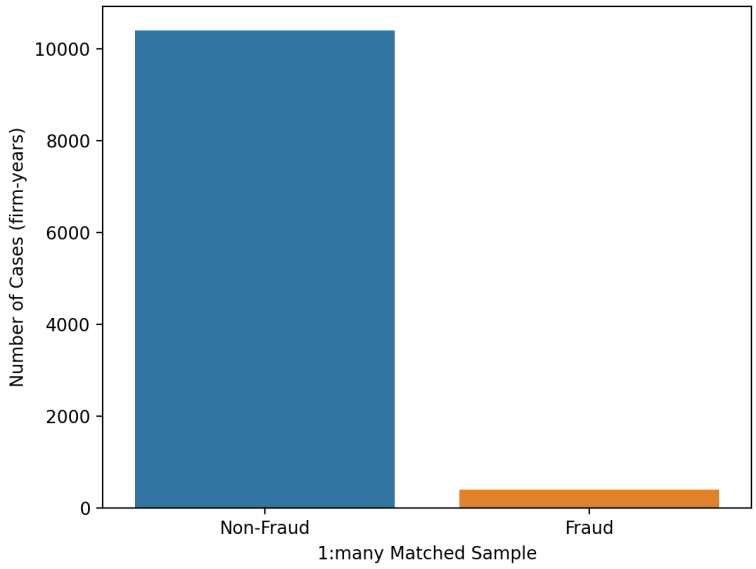
[22] Refer to Appendix D for key to input features.

5.4| Results from the 1:many Sample

5.4.1| *The Challenge of Classifying an Imbalanced Sample*

The above investigation provided valuable initial insights into how machine learning methods can aid in fraud detection. However, as discussed in the previous chapter, a 1:1 match is far from ideal. The proxy measures for company size alone is wanting, and 1:1 match is likely not an accurate representation of the reality a classifier will face. Figure 12 shows a class decomposition that is more likely to reflect reality. The initial 1:many sample consist of 13,015 filings that are matched by fiscal year and industry. Excluding cases with missing data, the final sample used in the analysis presented in this section consists of 10,397 non-fraud filings and 395 fraud filings. This represents a highly imbalanced but more realistic scenario where the minority class accounts for less than 4% of the total dataset.

Figure 12. 1:many Industry-Matched Sample (n= 10,972)



Results in Table 6 shows how this class imbalance impact classification performance in comparison to the previous sample. Consistent with the prior

analysis of the 1:1 matched sample, I used AUC as the model selection metric and compared a random forest and a neural network model to a logistic regression model. The reported metrics are again based on testing done on a completely unseen holdout sample where the train/test split ratio remained the same at 75/25. Due to the class imbalance, it is important that I stratify the sample to keep the ratio of fraud and non-fraud cases in the holdout test set. The algorithms are trained and cross-validated on 8,094 observations, with 7,798 non-fraud cases and 296 fraud; they are then tested on the holdout sample containing 2,698 observations, with 2,599 non-fraud cases and 99 fraud cases. 10-fold cross validation results for the AUC scores are reported in parenthesis.

Table 6. Classification Results *(n=10,792)*

	Random Forest	**Neural Network**	**Logistic Regression**
Model Specification	max_depth: 25, max_features: 6, min_samples_leaf: 2, min_samples_split: 3, n_estimators: 325	*hidden layer 1* neurons: 10 activation function: softmax *hidden layer 2* neurons: 5 activation function: relu	n/a
TP	3	2	8
TN	2599	2599	2579
FP	0	0	20
FN	96	97	91
Accuracy	0.964	0.964	0.959
Sensitivity	0.030	0.020	0.081
Precision	1.000	1.000	0.286
Specificity	1.000	1.000	0.992
AUC (CV)	0.515(.509)	0.510(.500)	.537(.526)

At first examination, all three models boasted very optimistic scores in overall accuracy, correctly classifying over 95% of cases. However, a closer examination of the confusion matrices revealed that the class imbalance posed a significant challenge to the machine learning algorithms. None of the algorithms performed much better than a naïve classifier at distinguishing between fraud and non-fraud cases, with logistic regression having a slight edge. Not only do the algorithms lack training instances or "density" from fraud cases (Fernandez et al., 2018), but when the minority class comprise less than 4% of the total training sample (i.e., when the imbalance between classes is extreme), the algorithms learned the rarity of the event. In other words, they predicted the majority class for almost all cases, as doing so automatically meant they are correct 96% of the time (as seen in the accuracy metric). This highlights the limited utility of the accuracy metric and the usefulness of the sensitivity metric in data with imbalanced class distributions. Even when model complexity is increased, the random forest and the neural network algorithms only predicted 3 and 2 fraud cases, respectively, with no false positives at all. Thus, these models suffered from poor recall (low sensitivity). Logistic regression predicted 8 fraud cases but also generated 20 false positives, yielding poor sensitivity and precision, but overall a slightly better AUC than random forest and neural network. In sum, to answer our research questions, even though the machine learning algorithms yielded high overall accuracy, they did not perform much better than a naïve classifier in a real-world setting. In comparison,

logistic regression outperformed both algorithms but also only marginally better than a naïve classifier.

The extreme class imbalance also generated another complication with the cross-validation process. When the sample is split into 10 folds, some folds may not contain any fraud cases at all. As a result, validations scores among folds varied widely, some with zero AUC scores, yielding averages that are drastically different from the test scores. To remedy this, I stratified the data to ensure close to equal ratio of fraud to non-fraud cases in each cross-validation fold. Once that is accomplished, the 10-fold cross-validation produced consistent scores compared to the test set and between folds.

5.4.2| *Synthetic Minority Oversampling Technique (SMOTE)*

The imbalance class problem is not unique to fraud detection, as rare events have long been the subjects of interest in academic research and real-world applications alike. With the expansion of data sources and the increased popularity of big data analytics, research on methods to tackle imbalanced data have also increased (He & Garcia, 2009). Yet, since each dataset has its own unique characteristics that can exacerbate the classification challenge associated with imbalance data (e.g., size, label noise, data distribution), there is no one-size-fits-all strategy (Fernandez et al., 2018). There are many different schools of thoughts when it comes to handling imbalanced data, but the most dominant approach is to

target the sampling methodology.[23] This section represents an exploratory analysis on the effectiveness of one such technique on corporate financial fraud detection. Specifically, I applied a widely implemented resampling technique called Synthetic Minority Over-sampling Technique (SMOTE) (Chowla, Bowyer & Kegelmeyer, 2011) to the 1:many imbalanced data above, and evaluated its performance using the same random forest, neural network algorithms and logistic regression.

Sampling methods used to handle imbalance data have a straightforward goal—to balance the class frequencies either through under-sampling, or over-sampling, or both. The most simplistic methods would be to randomly duplicate observations in the minority class (random over-sampling) or randomly remove observations from the majority class (random under-sampling). SMOTE differs from random over-sampling by generating new observations based on feature space similarities between existing observations of the minority class (Chowla et al., 2011), instead of simply duplicating existing data. Using k-nearest neighbor, SMOTE identifies observations that are close in the feature space, forms a line through those observations and create new observations along that line. This helps build larger decision regions surrounding those fraud cases (Chowla et al., 2011).

Since the 1:1 matched sample already resembles an under-sampling strategy, I chose to investigate how an over-sampling strategy would compare. The SMOTE algorithm I used came from the Imbalanced Learn library. Recall that our

[23] Other schools of handling imbalanced data include cost-sensitive methods, kernel-based learning methods, active learning methods and one-class learning methods. He and Garcia (2009) offers a more detailed survey of these methods.

preprocessed training data consist of 7,798 non-fraud observations and 296 fraud observations. After the implementation of SMOTE, the minority class have been over-sampled to match the 7,798 majority class, yielding a total training sample of 15,596. Note that SMOTE is only applied to the training data and not the holdout test set. This way, performance evaluation of the algorithms is not impacted by resampling of the test data. To better visualize the effect of this over-sampling procedure, I created scatterplots of two financial variables before and after resampling (Figure 13).

Table 7 shows the random forest and the neural network models in comparison to logistic regression after SMOTE is implemented. Recall that these reported metrics are based on the same holdout sample from the previous section, which contains 2,698 observations (2,599 non-fraud and 99 fraud cases). All three models improved in identifying fraud cases. The random forest algorithm improved the least, only identifying 19% of all the fraud cases. In comparison, the neural network and the logistic regression models with resampled data were able to identify over 60% of the fraud cases. However, in the process of doing so, both models generated a lot of false positives, hence producing extremely low precision scores. Interestingly, despite the low recall, random forest scored higher in precision. It generated fewer false positives but made very little positive predictions to begin with, suggesting that SMOTE may have only improved the algorithm marginally by not predicting all cases as non-fraud. In contrast, while neural network and logistic regression scored lower in specificity (.650 and .631,

respectively) in comparison to random forest (.978), it may be an artifact of the high false positives. In other words, SMOTE appeared to have made the neural network and logistic regression models predict fraud more frequently[24].

The 10-fold cross validation process is slightly more complicated when a resampling method is introduced. If one were to over-sample the entire training set, then split it into 10 folds, information about data from one fold would likely appear in another fold and the test set for each fold would also have contained resampled data. This data leakage problem would ultimately cause the cross-validated metric to generalize poorly to unseen data, as the algorithm that was supposed to be learning from data within one fold would have also learned from data leaked from another fold and from the test set, nullifying the validation process. Therefore, to obtain a more robust cross-validation result, I first split the sample into 10 folds, stratifying to ensure equal proportions of fraud to non-fraud cases in each fold, then split each fold into training and test sets, and performed SMOTE on only the training data of each fold prior to training and testing. This resulted in a more generalizable average AUC score, as shown in the relatively low discrepancies between cross-validation and test results.

[24] To account for potential variations between relevant periods, I also performed supplemental analysis with the fiscal year variable included in the random forest and logistic regression models with SMOTE. Results are reported in Appendix E.

Figure 13. Training Sample Before and After SMOTE

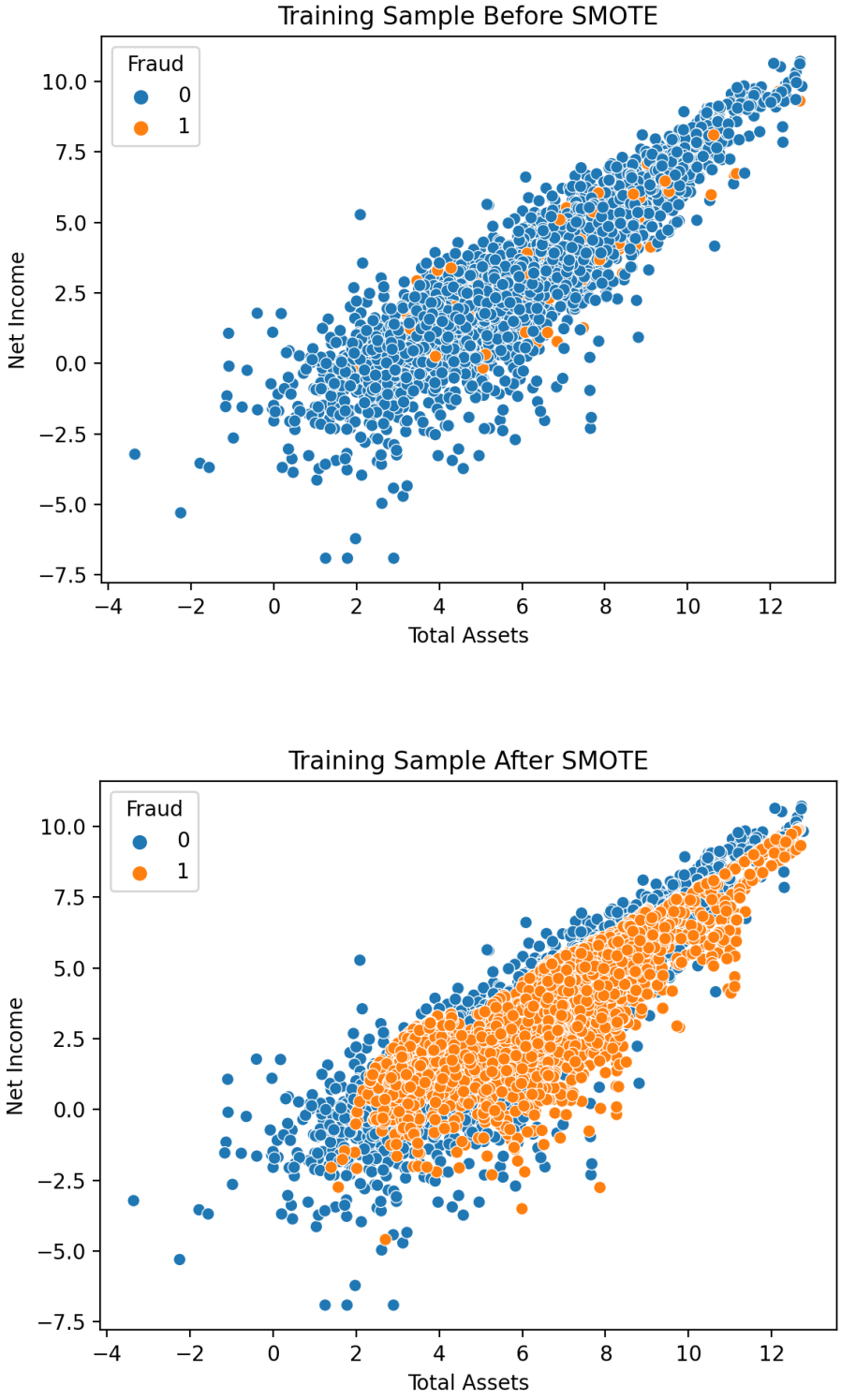


Table 7. Results of 1:Many Sample After SMOTE

	Random Forest	Neural Network	Logistic Regression
Model Specification	n estimators: 183, min samples split: 4, min samples leaf: 3, max leaf nodes: 48, max features: 5, max depth: 110	*hidden layer 1:* neurons= 10 activation=softmax *hidden layer 2:* neurons= 5 activation=relu batch size: 50 epochs: 150	n/a
TP	19	48	64
TN	2543	2098	1641
FP	56	501	958
FN	80	51	35
Accuracy	0.950	0.795	0.632
Sensitivity	0.192	0.485	0.646
Precision	0.253	0.087	0.063
Specificity	0.978	0.807	0.631
AUC (CV)	0.846(.867)	0.702(.681)	.659(.661)

To explore whether feature selection in conjunction with oversampling can help counterbalance the challenges brought about by class imbalance, I used the combination of financial and organizational risk factors in subset 4 above to train the three different algorithms with the 1:many dataset. The same performance evaluation metrics as the previous sections are reported in Table 7. Feature selection improved the random forest algorithm's ability to identify fraud cases from 19% to 59%. However, this improvement did come at the cost of precision (from 25% of the time correct when predicting fraud, to 12%). In other words, it had generated many more false positives, as the other algorithms did in the previous section with over-sampling alone. Specificity also declined for the random forest algorithm, from being able to identify 98% of the non-fraud case to 84%. Nonetheless, overall accuracy and the ability to distinguish between classes are still superior to the other

two models. With a more parsimonious subset of input features, neural network experienced slight improvement in overall accuracy, precision, specificity, and AUC scores. However, sensitivity in fraud case have decreased from 62% to 53%, and precision only improved marginally (from 6% to 8%). Finally, pruning some features appeared to have slightly impacted the classification performance of logistic regression in a negative manner. Note that for equal comparisons, I have not changed the classification thresholds to optimize sensitivity or precision, as each algorithm will have a different optimal thresholds per their different ROC curves. Further optimization of these algorithms for applications will be discussed in the following chapter.

Table 8. Results with SMOTE and Feature Selection

	Random Forest	**Neural Network**	**Logistic Regression**
Model Specification	n_estimators: 277, min_samples_split: 3, min_samples_leaf: 2, max_features: 3, max_depth: 10	*hidden layer 1* neurons: 18 activation function: softmax *hidden layer 2* neurons: 10 activation function: relu	n/a
TP	58	52	67
TN	2174	1983	1527
FP	425	616	1072
FN	41	47	32
Accuracy	0.827	0.754	0.591
Sensitivity	0.586	0.525	0.677
Precision	0.120	0.078	0.059
Specificity	0.836	0.763	0.588
AUC (CV)	0.802(.711)	0.708(.716)	0.650(.632)

5.5| Key Findings

- Both the random forest and the neural network algorithms performed well above a naïve classifier in a balanced sample (research questions 1 and 4)
- Random forest outperformed logistic regression by a clear margin in terms of overall predictive accuracy and ability to distinguish between the two classes; its performance surpassed logistic regression in all metrics except sensitivity (research question 3)
- Neural network's classification ability is subpar compared to logistic regression, performing slightly worse across all metrics (research question 5)
- A random forest model with only financial statement line items as input measures yielded the highest AUC score (.865), whereas model using financial ratios as proxy measures performed the worst (research question 3)
- A random forest model with a mixture of financial, motivational and opportunity measures yielded the highest accuracy (78%) in the classification of unseen fraud and non-fraud cases (research question 3)
- Feature importance identified several financial measures that are consistent with prior empirical studies in financial fraud (research question 3)
- Auditor independence measured by non-audit fees was identified as a key concept for guardianship and opportunity structures that is predictive of fraud (research question 3)

- Measures of financial distress (debt and liquidity related risk factors) rank lower in importance than measures of financial health (performance and asset based risk factors) (research question 3)
- Both machine learning algorithms and logistic regression performed poorly in a heavily imbalanced dataset, but was able to improve identification of fraud cases to above 50% with the use of an oversampling strategy (SMOTE) and a more parsimonious set of features. Random Forest remained the best performing algorithm.
- Imbalanced data cautioned the use of single metrics to evaluate a classifier in rare events
- Since the 1:1 demonstrated promising performance and essentially represents an under-sampling strategy, further improvement is hopeful with the help of analytic strategies specifically designed to handle imbalanced data

CHAPTER 6. DISCUSSION

Corporate financial crime research is fraught with challenges; not only is financial crime a subject that requires interdisciplinary expertise, complex financial data is also associated with a myriad of methodological difficulties. Yet, without more cross-disciplinary, empirical research to guide regulations, or some form of alleviation to combat the limited resources to enforce these crimes, there is little hope to curtail the continued recurrence of large-scale corporate scandals. The overarching goal of this dissertation is to investigate whether some of these research and enforcement related challenges can be overcome with the help of recent developments in artificial intelligence. I sought to develop and train two machine learning algorithms and assess their ability in detecting corporate financial fraud with a set of publicly accessible financial and organizational risk factors. Results discussed in the previous chapter have shown promising results in a controlled setting, but have also demonstrated that more future research is required in order to implement the machine learning algorithms as a real-world decision aid in fraud detection.

6.1 | Limitations of the Present Study

Before discussing the results further, I must first point out the limitations of the current study. To identify cases of corporate financial fraud, I have primarily relied on official enforcement releases from the SEC. As discussed in Benson, Kennedy, and Logan (2016), there are no standards for systematically reporting

violations of any specific corporate violations. The AAERs used to identify fraud cases comprised a variety of civil, administrative, and criminal proceedings that is deemed to be related to accounting and auditing. Some AAERs pertain to a single company with multiple years of violations and multiple individuals charged; other companies have multiple AAERs across different years. The unit of analysis used in this dissertation (firm-years or filings) partially mitigated this inconsistency, but enforcement releases are by no means systematic or comprehensive. It is also part of the reason the true ratio of fraud to non-fraud cases is unknown. The sample of fraudulent filings I collected suggest a downward enforcement trend of corporate financial fraud (Chapter 4). How much of this decrease can be attributed to the deterrence effect of regulations such as the Sarbanes-Oxley Act or the Dodd Frank Act, and how much can be attributed to the reduction in enforcement effort is an empirical question. While I have taken the precautionary measure of checking for restatements in the non-fraud group to ensure their filings are reasonably compliant, it is possible that some corporate financial fraud went undetected and uncorrected and therefore resulted in no restatements. An uncaught fraud filing labeled as non-fraud impact our analysis by introducing noise to the data that may be irreducible. Fraud cases also impact our interpretation of the analysis, as the fraud sample used in this dissertation represent the likelihood that a firm has committed fraud *and* are prosecuted. Therefore, it is important to improve our understanding of the decision processes used to prosecute a fraud case. Our limited understanding of the fraud to non-fraud distribution also impacts our sampling

effort and our evaluation of the models. As demonstrated in the previous chapter, training algorithms with a 1:1 sample and a 1:many sample yielded very different results.

Related to the reliability of corporate crime data sources is the accessibility of corporate crime risk factors. Aside from financial risk factors and motivational proxy measures (most often measured with financial ratios) that can be extracted from the 10-k, other organizational variables are not consistently reported or easily accessible. While I was able to obtain some information about the corporation's officers and auditors, much of the missing values from the dataset pertain to these organizational risk factors. Board and committee information that was once readily available through the various databases is no longer available or maintained. The omission of these variables may have substantial impact on the prediction of fraud, as they shed light on compensation/incentive and guardianship structures that may be key factors in offenders' decision processes.

One common criticism of machine learning research pertains to the lack of interpretability of the trained models. This weakness of machine learning is also its strength. In traditional statistical modelling, we are assuming the reality fits a certain function that can be expressed in an easily understood mathematical formula. We use goodness of fit tests to assess said assumption. In contrast, machine learning makes the assumption that a model that predicts the best is most representative of reality, and reality may not always fit neatly into a mathematical equation (Breiman, 2001). It is because of this tradeoff between interpretability and

predictive ability that I chose to compare machine learning methods with more commonly employed method in corporate crime research. The hope is to examine whether foregoing interpretability will help improve generalizability of our findings. However, this tradeoff is also why it is especially important for machine learning researchers to communicate what the algorithms can and cannot achieve. Results presented in Chapter 5 of this dissertation speaks to prediction and prediction only; we must be careful when interpreting parts of the analysis such as feature selection and feature importance. While some of the results showed support for existing findings, we should take caution to not ascribe a causal link or a relationship direction between the risk factors and corporate financial fraud. There is also a spectrum of interpretability among machine learning algorithms. I chose to test one of the most interpretable algorithms against one of the least interpretable one to examine whether they provide different advantages in different use cases. If two algorithms were to perform similarly in the metrics we wished to prioritize, the more interpretable one would likely be more desirable.

Finally, as only publicly traded corporations have annual financial reporting requirements, by our definition of fraud in violation of section 13(b) of the Securities Exchange Act, no privately held firms was examined. Yet, Benson et al. (2016) pointed out that publicly held companies represent only 1% of all corporations. Although financial fraud in those privately held firms may have less impact on public investors, the harm they impose on employees, consumers and other

stakeholders can be equally devastating, as exemplified by the recent case of Theranos (Straker, Nusem & Wrigley, 2021).

6.2| Discussion and Implications

6.2.1| *General Discussion and Methodological Implications*

In a balanced 1:1 sample, both the random forest and the neural network algorithms performed more effectively than a naïve classifier that would either classify cases at random or consistently predict only one of the classes. Random forest performed especially well across all categories, suggesting that in a controlled setting, the full set of risk factors do contain sufficient class discriminatory information to classify fraud cases from non-fraud cases with 75% accuracy. While the current analysis cannot answer how or why these risk factors predict fraud with this level of accuracy, this gives grounds for future theory development and testing. It is especially true when feature engineering appeared to be able to effectively improve the algorithm's learning ability by reducing unwanted noise. With more parsimonious subsets of input measures, random forest was able to achieve accuracy as high as 78%, showing promise for further improvements should the quality of measures improve, or should more features become accessible. In comparison to logistic regression's overall performance (accuracy of 56%), the random forest algorithm may be a case where the substantial increase in predictive accuracy outweighs the concerns of interpretability. In contrast, neural network, while performing adequately, represents the least interpretable and least effective

classifier. Therefore, there is little justification for using neural network for fraud prediction, especially since the balanced sample is not large enough in size to play to its strength.

Financial accounting rules are often not black and white as commonly misunderstood. Much of financial accounting requires judgment and estimations that are based on assumptions agreed upon by industry standards or between management and external auditors. As such, the task of predicting corporate financial fraud is an inherently challenging endeavor. However, this also means that using official enforcement data has its merits, as it implies that the identified cases possess characteristics that make a seemingly gray area less so. That is, there must be some characteristics embedded in these fraud cases that make them unambiguously fraudulent. The question at hand is therefore, whether we can capture and measure these characteristics adequately.

One of the more important findings of this research pertains to such measurement issues in the study of corporate crime. As shown in the feature selection analysis (Table 4), the random forest algorithm with the highest AUC score was trained with financial statements line items only (subset 1), and the model with the lowest AUC score was trained by the organizational proxy measures only. This discrepancy casts doubt as to the validity of financial ratios as proxy measures for motivational risk factors. It is noteworthy because financial ratios are very frequently used in corporate crime research as proxy measure to a corporation's financial health or financial distress. Yet, the adoption of financial

ratios often make little theoretical sense and, as shown in the comparisons of models trained by subsets 1, 2 and 3, also make very little sense analytically. Financial accounting uses a double entry system. Roughly speaking, that is when a transaction occurs, it is recorded as a debit to one account on the financial statements, as well as a credit to another account. This is why there exists an inherent risk of multicollinearity when using figures from the financial statement line items. Financial ratios are used by accountants, auditors and analysts in the financial statement analysis process, and often used to assess period to period changes or trends. However, since they are often computed with two or more financial statement line items, when used in the same statistical model, the multicollinearity issue becomes magnified, especially when multiple ratios are used in the same analysis.

For example, tests of corporate strain often include a ratio for financial health and another for financial distress. A popular proxy for financial health is *return on assets*, which is computed by dividing net income (or some form of earnings before or after extraordinary items) by total assets. A frequently used financial ratio to measure financial distress is the Altman's z-score, which consists of five components (all ratios in and of themselves), two of which contains earnings/income and total assets, and the remaining three components are often included in the analysis as proxies for some other theoretical concept. Doing so is essentially capturing the same few financial statements line items multiple times in an analysis, while also artificially imposing a restriction range due to the nature of

a quotient. More importantly, even if the use of financial ratios is theoretically appropriate and does not violate assumptions of the statistical methods used, the analysis presented in the previous chapter suggests that they do not make very good predictors of fraud. Put simply, there is little justification for the use financial ratios when raw financial statement line items can be used just as easily and with greater classification power.

Overall, machine learning represents a helpful tool to investigate how well a set of risk factors can predict fraud. The underlying premise of machine learning techniques is that if an algorithm can predict/generalize well on unseen data, it is more likely to be reflective of reality, regardless of whether the model can be specified in a neat mathematical formula with closed-form solutions. Therefore, it provides a basis for future research to further explore *how* said set of risk factors is related fraud. While more commonly used statistical models such as logistic regression can specify a function that explains the relationship between the risk factors and the fraud outcome, the results presented in this dissertation have shown that it may possess less predictive power. In other words, the logistic regression model, despite being more transparent, may be flawed in reflecting the true relationships between the risk factors and the fraud outcome. Statistical modeling and machine learning thus provide different utilities in the scholarly pursuits of understanding corporate crime—machine learning allows us to examine how predictive a risk factor is, whereas statistical models allow us to examine how a risk factor impact crime.

6.2.2| *Theoretical Implications*

Despite the necessary precaution against overinterpreting results from feature importance, the findings do shed some light on theory and direction for future work. Due to the improvement feature selection analysis made to the learning of the random forest algorithm, the combinations of risk factors used in the subsets warrant further theoretical investigation, especially subsets 3 and 4 where particular subsets of financial and organizational risk factors were able to increase classification accuracy. Ranked order feature importance for the best the subset 4 model identified several financial features—particularly total assets and net sales—that are consistent with existing with empirical studies (e.g., Beasley, 1996, Erickson, Hanlon & Maydew, 2006). However, this consistency is only limited to these risk factors being predictive of fraud and non-fraud filings; as to their causal direction, relationship signs (positive or negative) or strength, the random forest algorithm is silent.

Prior studies on corporate strain have included financial health and distress indicators, primarily measured by financial ratios (e.g. Schwartz et al., 2021). Results from the random forest classifier suggest that financial health risk factors are more class discriminatory than financial distress risk factors, which supports prior findings on fraud firms being more profitable and exhibit lower bankruptcy risk (Schwartz et al., 2021). When applying the anomie-strain perspective to corporate crime, prior studies have proposed that crime does not always arise from negative circumstances (e.g., Benson, 2010). Rather, motivations to commit fraud

can arise from pressure in maintaining performance and growth relative to peer firms. Risk factors relating to debt and liabilities exhibited less class discriminatory value in the random forest models than risk factors relating to sales and assets, but are still relevant in the prediction process. This suggests possible differential mechanisms through which strain can impact a corporation's likelihood to offend. Future research should explore this more thoroughly. With regards to opportunity risk factors, some external auditor related measures are consistent with extant research and theory on non-independence (Davis, Soo & Trompeter, 2003). Yet, other opportunity measures such as officer change, officer duality, and auditor change are not identified as important, in contrast to previous studies (e.g., Simpson & Koper, 1997). Further investigation is necessary to determine how much data cardinality played a role in this, but it may support further inquiries on guardianship capability in the corporate setting (e.g., Chan & Gibbs, 2022).

Other less commonly tested risk factors identified by feature importance bar charts also suggest that random forest may be able to capture nonlinear relationships that may not be captured by general or generalized linear models. A sequential explanatory mixed method study may help shed light on the mechanisms with which these risk factors relate to fraud prediction. However, to truly capture each individual risk factor's impact on prediction, it may require the exclusion of each risk factor from a model to compare the differences in classification results. This is an especially labor-intensive task with machine learning, as each model needs to be tuned and cross-validated to ensure generalizability.

6.2.3| *Practical Implications*

Since the project is partly motivated by evaluating possible decision aid tools to combat resource constraints in corporate crime enforcement, I trained and tested the machine learning algorithms in a 1:many sample that might better reflect real-life fraud detection scenarios. Both machine learning algorithms and logistic regression trained with the heavily imbalanced sample performed poorly in the test set, but showed improvements after over-sampling with SMOTE and in conjunction with feature selection. This investigation cautioned the generalization of a 1:1 matched sample. It also cautioned the reliance of evaluation metrics without thoroughly consulting the confusion matrix. When working with imbalanced data, a classifier that heavily favored the majority class in its prediction will naturally yield unrealistically high accuracy and specificity scores. In such cases, it is more helpful to examine the precision-sensitivity tradeoff, which brings us to the importance of defining the goal for a decision aid prior to its actual implementation. If the goal were to detect new fraud cases, one might choose to optimize sensitivity when training the algorithms, as one might be willing to accept more false positives, as long as it helps to identify more fraud cases. However, if the concern is to conserve resources, then optimizing precision may be of higher import. The algorithms presented in Table 7 are optimized based on AUC for consistency in comparison. All three models appeared to favor sensitivity at the expense of precision. However, in practice, one would examine the ROC curve or the precision-recall curve to determine a threshold that best suit the goal of the decision aid.

As this portion of the dissertation merely represents an initial exploration to the practical application of a fraud detection algorithm, many potential solutions to the class imbalance issue have yet been explored. For example, I have discussed the different over- and under-sampling techniques in the previous chapter. Since the 1:1 matched sample have shown promising results, an under-sampling technique or a combined sampling strategy may be explored. Aside from resampling, there are also cost-sensitive learning algorithms that have been shown to be superior in classifying imbalanced data in comparison to resampling (McCarthy, Zabar & Weiss, 2005). These algorithms do not assume all misclassification errors to be equal. For example, in medical research, a missed diagnosis of a lethal disease is a more serious error than a false positive diagnosis. To reflect this differential consequence in classification error, each class is assigned a misclassification cost, and optimizing the algorithm is about minimizing classification cost rather than optimizing a certain evaluation metric (He & Ma, 2013). There is also ensemble learning, which combines multiple algorithms to create better prediction. For instance, while our results have shown random forest to be a better overall classifier, logistic regression performs better at identifying fraud cases (sensitivity) in certain instances. Ensemble learning can yield better prediction by leveraging each algorithm's strength.

6.2.4| *Directions for Future Research*

As this project merely signifies a first step in exploring the use of machine learning to study corporate crime, the discussion above has revealed a long list of

potential future inquiries. With regards to improving fraud prediction with machine learning, future research should consider other machine learning algorithms such as support vector machines or incorporate other ensemble or boosting methods. Future research should also explore ways to combat class imbalance in the fraud detection setting, including the use of resampling and cost-sensitive learning algorithms. Future corporate crime research should seek to examine class discriminatory risk factors identified in this research and assess how they may differentially impact different types of corporate crime or corporations in different industries. Findings on financial distress ranking lower financial health in feature importance also suggest the need to explore differential mechanisms with which corporate strain affect likelihood to offend. The analysis also identified auditor independence as a key concept of guardianship and opportunity structure that warrants further study. Finally, to improve robustness of corporate crime research findings, we should aim to explore better measurements to capture motivational risk factors for theoretical inquiries. We should also better define and measure firm size or employ more rigorous matching techniques (such as propensity scores).

6.2.5 | *Conclusion*

In this dissertation project, I set out to investigate whether machine learning algorithms, trained by a set of multidisciplinary risk factors, can be used in lieu of more commonly employed statistical models to detect corporate financial fraud. Overall, the results have shown random forest to be a promising algorithm in the fraud classification. Although more work needs to be done for a real-world

application, the fraud detection ability of the random forest algorithm has surpassed that of logistic regression. Applications of algorithms have been criticized in the criminal justice realm for perpetuating biases on vulnerable populations. It is my hope that with more rigorous, transparent, and reproducible research procedures, algorithms can be used to address a crime type that is understudied, underenforced and perpetrated predominantly by powerful corporations.

APPENDICES

APPENDIX A. SEARCH TERMS & DATABASES

Search Terms. Variants of the following keywords are used to search the databases to identify risk factor research related to corporate financial fraud:

- "accounting fraud"
- "financial fraud"
- "financial reporting fraud"
- "financial statement fraud"
- "management fraud"
- "earnings management"
- "financial misstatement"
- "earnings quality"
- "audit quality"

Databases. The following databases are used to search for empirically measured and tested, organizational-level risk factors related to corporate financial fraud:

- Criminal Justice Abstracts
- ProQuest Criminal Justice
- NCJRS Abstracts, Sociological Abstracts
- ABI/INFORM Collection
- Business Source Complete
- Business Economic and Theory Collection
- Business Insights
- JSTOR
- SAGE Journals Online
- CSA Linguistics and Language Behavior
- Interdisciplinary Science Reviews
- American Journal of Sociology
- Social Science Research
- SciFinder
- Web of Science
- IEEE Transactions
- Communication Abstracts

APPENDIX B. CORPORATE FINANCIAL FRAUD RISK FACTORS

Table 9. Financial Fraud Risk Research and Synthesis

Risk Factor	Fraud Element	Explanation	Empirical Research
Accounts Receivable to Sales	Motivation	Proxy measure for liquidity	Kaminski et al. (2004)
Accounts Receivable Turnover	Motivation	Proxy measure for efficiency in debt collection	Kaminski et al. (2004)
Altman Z Score	Motivation	Low scores indicate financial distress; bankruptcy risk	Fanning & Cogger (1998); Brazel et al. (2009); Perols et al. (2011); Perols et al. (2017); Wang & Holtfreter (2012); Schwartz et al. (2021)
Asset Quality Index	Motivation	Proportion of total assets with less certain future benefits (> 1 indicates potential for cost deferral by capitalizing)	Beneish (1999)
Audit Committee Independence	Opportunity	Proxy for board member independenece	Klein (2003)
Audit Committee Meetings	Opportunity	More effective governance	Abbott et al. (2000); Uzun et al. (2004); Farber (2005)
Audit Fees	Opportunity	Proxy measure for audit quality	Ferguson et al. (2003)
Auditor Change	Opportunity	Proxy measure for audit quality; new auditors are less familiar with business	Myers et al. (2003)
Auditor Tenure	Opportunity	Can represent both risk and protective factor--complacency decrease audit quality and turnover may indicate opinion shopping; auditor knowledge of client improve audit quality	Myers et al. (2003); Carcello & Nagy (2004)

Table 9 (cont'd)

Risk Factor	Fraud Element	Explanation	Empirical Research
Auditor is Big4 Firm	Opportunity	Proxy for audit quality; Big 4 firms have more resources and are more concerned with reputation should an audit fail	Fanning and Cogger (1998); Carcello and Nagy (2004); Farber (2005)
Board Size	Opportunity	Less effective governance	Uzun et al. (2004)
Book to Market Ratio	Motivation	Proxy for under- or over-valuation for a firm	Carcello and Nagy (2004); Efendi et al. (2007)
Cash Margin	Motivation	Proxy measure for financial health/profitability	Green & Choi (1997)
CEO Duality	Opportunity	CEOs who are also chairpersons/ presidents of the board have potential conflict of interest in corporate governance	Carcello and Nagy (2004); Uzun et al. (2004); Farber (2005)
CEO Duality or Board Independence	Opportunity	Officer (non)independence	Simpson & Koper (1997); Klein (2003)
CEO is Founder	Motivation	CEOs who are also founders have potential conflict of interest in corporate governance	Dechow et al. (1996)
CEO Turnover	Opportunity	Can signal internal detection of misconduct	Dechow et al. (1996); Fanning & Cogger (1998); Feroz et al. (2000); Arthaud-Day et al. (2006); Simpson &Koper (2007)
Change in Free Cash Flow	Motivation	Proxy measure for financial distress/lack of liquidity	Dechow et al. (1996); Dechow et al. (2011)
Change in Cash Sales	Motivation	High growth firm are associated with higher external pressure on achieving earning targets	Jones (1991); Fanning and Cogger (1998); Beneish (1999); Bell and Carcello (2000); Erickson et al. (2006); Efendi et al. (2007); Brazel et al. (2009)

Table 9 (cont'd)

Risk Factor	Fraud Element	Explanation	Empirical Research
Change in Non-Cash Operating Assets	Motivation	Proxy measure for risk diversification	Dechow et al. (1996)
Change in Receivables	Motivation	More receivables signal less cash flow	Green & Choi (1997)
Compensation Committee	Motivation	Protective Factor--Provides oversights on executive compensation to reduce potential conflict of interest	Klein (2003); Uzun et al. (2004)
Compensation Committee	Opportunity	More oversight	Uzun et al. (2004)
Compensation Committee Independence	Opportunity	More oversight	Klein (2003)
Days in Receivables	Opportunity	Revenue inflation	Beneish (1997); Chen and Sennetti (2005)
Debt to Equity Ratio	Motivation	Measures firm's financial leverage	Kaminski et al. (2004); Fanning & Cogger (1998)
Decentralized Organization	Opportunity	Decentralized organizational structure are more crime conducive	Simpson & Koper (1997)
Depreciation Index	Motivation	Signals upward revision of asset useful lives or change in depreciation methods	Beneish (1999)
Disclosure Complexity	Opportunity	Firms may intentionally complicate financial disclosure to conceal financial misconduct	Humphreys et al. (2011)

Table 9 (cont'd)

Risk Factor	Fraud Element	Explanation	Empirical Research
Discretionary (Abnormal) Accruals	Opportunity	The unobservable portion of total accruals that is subjected to managerial discretion	Jones (1991); DeFond & Jiambalvo (1994); Dechow et al., (1995); Marrakchi et al., (2001); Perols and Lougee (2009); Dechow et al. (2012)
Equity Compensation Committee	Motivation	Officer (non)independence	Gillett & Uddin (2005); Erickson et al. (2006)
External Board Members	Opportunity	Higher proportion of external board members represent more oversight and less chance for collusion	Beasley (1996); Beasley (2000); Abbott et al. (2000); Carcello and Nagy (2004); Farber (2005); Crutchley et al. (2007)
Financial Expert in Audit Committee	Opportunity	Protective Factor--Financial expert in audit committee represent stricter guardianship and higher risk of fraud discovery	Farber (2005)
Fixed to Total Assets	Motivation	Proxy for efficient management of assets	Kaminski et al. (2004)
In-the-Money Options	Motivation	Higher conflict of interest for officers	Efendi et al. (2007)
Industry Competition	Motivation	Low ratio represents higher competition amongst firms in the industry	Rasheed & Prescott (1992); Palmer & Wiseman (1999); Ndofor et al. (2015)
Inventory Related Measures/Ratios	Opportunity	There are many ways to manipulate inventories as management have discretion over accounting and valuation methods	Fanning & Cogger (1998); Summers & Sweeney (1998); Kaminski et al. (2004)

Table 9 (cont'd)

Risk Factor	Fraud Element	Explanation	Empirical Research
Inventory to Sales	Motivation	Proxy for efficient management of inventories	Kaminski et al. (2004)
Meeting/ Beating Analyst Consensus	Motivation	Management may experience pressure or incentives to meet/ beat analyst forecast on performance	Coleman (1987); Finney and Lesieur (1987); Kagan et al. Gunningham et al. (2004); Perols & Lougee (2009)
Non-Audit Fees	Opportunity	Proxy for auditor (non)independence	Frankel et al. (2002)
Officer Change	Opportunity	Change in top management disrupts social control	Simpson & Koper (1997)
Officers are Accountants	Opportunity	More effective governance	Albrecht et al. (2018)
Operating Leases	Opportunity	Operating leases allow firm to recognize earnings early and reduce reported debt	Dechow et al. (2012)
Operating Leases	Opportunity	Proxy for off-balance shee financing (i.e. way to front load earnings)	Krische, Sanders & Smith (2012)
Outside Audit Committee Member	Opportunity	Proxy for board member independenece	Abbott et al. (2000); Crutchley wt al. (2007)
Retained Earnings on Asset	Motivation	Measures relianace debt or equity financing	Dechow et al. (2011); Kaminski et al. (2004)
Return on Asset	Motivation	Proxy measure for financial health/profitability	Wang & Holtfreter (2012); Schwartz et al. (2021)
Return on Equity	Motivation	Financial health/growth	Feroz et al. (2000), Wang & Holtfreter (2012)
Sale of Stock	Motivation	Measures incentive to maintain high stock prices	Dechow et al. (2011); Beneish (1999)

Table 9 (cont'd)

Risk Factor	Fraud Element	Explanation	Empirical Research
Sales Growth	Motivation	High growth firm tend to experience more external pressure on achieving earning targets	Jones (1991); Green & Choi (1997); Fanning & Cogger (1998); Beneish (1999); Myers et al. (2003); Bell & Carcello (2000); Lin et al. (2003); Erickson et al. (2006); Efendi et al. (2007); Brazel et al. (2009)
Soft Assets Ratio	Motivation	Financial health; ability to manage short term earnings	Dechow et al. (2011)
Stock Price at Year End	Motivation	Proxy for financial health	Dechow et al. (2011)
Total Accruals to Total Assets	Motivation	High proportion of accrual as opposed to cash are associated with sales	Beneish (1997); Dechow et al. (1996); Beneish (1999); Lee et al. (1999); Crutchley et al. (2007); Bayley & Taylor (2007)
Total Fees to Public Accounting Firms	Opportunity	Higher fees associated with less independence	Frankel et al. (2002)
Unexpecter Revenue per Employee	Opportunity	Firms that artificially manage earning will experience inflated revenue per employee	Perols & Lougee (2009)
Volatile Industries	Motivation	Volatile industries may be more prone to earnings smoothing in order to convey a signal of stability to investors	Rasheed and Prescott (1992); Palmer and Wiseman (1999); Ndofor et al. (2015)
Working Capital	Motivation	Measure for financial liquidity	Perols & Lougee (2009); Kaminski et al. (2004)

APPENDIX C. DESCRIPTIVE STATISTICS AND CORRELATIONS

Table 10. Risk Factor Descriptives and Point-Biserial Correlations *(n=760)* [25]

Risk Factor	Fraud Mean	Fraud SD	Non-Fraud Mean	Non-Fraud SD	Point-Biserial Corrrelation
Accounts Payable	607.37	607.37	220.00	1,549.17	0.10*
Audit Fees	1,980.14	1,980.14	1,488.95	3,545.03	0.07
Auditor Change	0.08	0.08	0.08	0.28	-0.01
Big Four Auditors	0.79	0.79	0.71	0.46	0.10*
Book to Market Ratio	0.58	0.58	0.51	0.92	0.05
Capital Stocks	3.51	3.51	1.43	16.41	0.05
Cash and Equivalents	644.45	644.45	570.86	2,864.84	0.01
Cash Margin	0.10	0.10	-0.09	2.11	0.05
CEO Duality	0.96	0.96	0.94	0.24	0.06
Change in Cash Sales	0.24	0.24	0.16	0.95	0.04
Change in Free Cash Flows	-0.03	-0.03	-0.04	0.41	0.02
Change in Non Cash Operating Assets	0.06	0.06	0.06	0.31	0.00
Change in Reveivables	0.03	0.03	0.01	0.07	0.08*
Common Shares Outstanding	195.44	195.44	196.37	844.79	0.00
Cost of Goods Sold	2,881.01	2,881.01	1,521.45	9,538.42	0.07
Current Assets	1,899.70	1,899.70	1,235.21	5,863.88	0.05
Current Liabilities	1,305.78	1,305.78	672.18	3,571.74	0.08*
Debt in Current Liabilities	234.33	234.33	81.54	672.91	0.09*
Debt Issuance	552.61	552.61	243.11	1,601.23	0.07
Depreciation and Amortization	198.47	198.47	176.46	844.77	0.01
Depreciation Index	1.09	1.09	1.06	0.45	0.03

[25] SD = Standard Deviation; Point-Biserial Correlations with Fraud measure (1=Fraud, 0=Non-Fraud); * $p < .05$

Table 10 (cont'd)

Risk Factor	Fraud		Non-Fraud		Point-Biserial
	Mean	SD	Mean	SD	Corrrelation
Income Before Extraodinaries	106.12	106.12	279.05	1,518.99	-0.04
Income Taxes	105.17	105.17	157.42	1,096.50	-0.03
Interest Expense	116.56	116.56	30.30	101.65	0.09*
Investments and Equivalents	241.92	241.92	124.19	763.61	0.06
Long Term Debt	1,531.45	1,531.45	518.36	2,076.32	0.09*
Net Income	-36.18	-36.18	294.80	1,567.09	-0.04
Net Sales	4,062.69	4,062.69	2,593.35	12,862.22	0.05
Nonaudit Fees	826,349.19	826,349.19	325,463.92	642,561.38	0.15*
Officer Change	0.39	0.39	0.38	0.49	0.01
Property Plant Equipment	3,769.06	3,769.06	2,021.05	11,627.64	0.04
Retained Earnings	579.25	579.25	1,072.26	6,141.83	-0.03
Retained Earnings on Assets	-0.22	-0.22	-1.88	8.35	0.14*
Return on Assets	0.00	0.00	-0.13	0.91	0.10*
Sale of Stock	124.06	124.06	43.56	186.60	0.04
Short Term Investments	163.41	163.41	197.82	1,166.01	-0.02
Soft Assets Ratio	0.62	0.62	0.54	0.24	0.16*
Stock Price at Year End	22.21	22.21	18.57	39.78	0.06
Taxes Payable	23.23	23.23	43.32	252.27	-0.05
Total Assets	6,346.42	6,346.42	3,032.66	13,974.58	0.07
Total Common Equity	2,708.94	2,708.94	1,548.41	7,152.81	0.05
Total Fees to Public Accounting Firms	2,806.49	2,806.49	1,814.41	3,894.78	0.11*
Total Inventories	395.63	395.63	228.85	1,328.49	0.06
Total Liabilities	3,529.50	3,529.50	1,462.22	7,241.99	0.09*
Total Receivables	683.84	683.84	358.64	2,182.63	0.07
Working Capital	0.01	0.01	-0.01	0.15	0.09*

APPENDIX D. KEY TO FEATURE IMPORTANCE (FIGURE 11)

- total.assets = Total Assets
- net.sales = Net Sales
- nonaudit.fees = Nonaudit Fees
- cash.invst = Cash and Equivalents
- reoa = Retained Earnings on Assets
- cogs = Cost of Goods Sold
- ppe = Property, Plant and Equipment
- csho = Common Shares Outstanding
- audit.fees = Audit Fees
- soft.assets = Percentage of Soft Assets
- current.assets = Current Assets
- total.rec = Total Receivables
- ap = Accounts Payable
- total.liab = Total Liabilities
- interest.exp = Interest Expense
- sale.stocks = Sale of Stocks
- current.liab = Current Liabilities
- invt = Total Inventories
- debt.cl = Debt in Current Liabilities
- lt.debt = Long Term Debt
- big.four = Big Four Auditor
- auditor.change = Change in Auditors

APPENDIX E. SUPPLEMENTAL ANALYSIS

Table 11. Classification Results with Years Included

	Random Forest (without years)	Random Forest (with years)	Logistic Regression (without years)	Logistic Regression (with years)
Model Specification	n estimators: 183, min samples split: 4, min samples leaf: 3, max leaf nodes: 48, max features: 5, max depth: 110	n estimators: 183, min samples split: 4, min samples leaf: 3, max leaf nodes: 48, max features: 5, max depth: 110	n/a	n/a
TP	19	18	64	42
TN	2543	2549	1641	2235
FP	56	50	958	364
FN	80	81	35	57
Accuracy	0.950	0.951	0.632	0.844
Sensitivity	0.192	0.182	0.646	0.424
Precision	0.253	0.265	0.063	0.103
Specificity	0.978	0.981	0.631	0.860
AUC (CV)	0.846(.867)	0.855(.844)	.659(.661)	.642(.631)

The above analysis shows the changes in prediction results from the random forest algorithm compared to logistic regression when the years of the relevant period are included as an input measure. Random forest showed minimal changes in classification performance, while logistic regression saw a trade-off between specificity and precision. There was an increase in accuracy but decrease in overall AUC. Since these changes are relatively small, we can conclude that year to year variations do not impact predictive power in the classification of fraud cases.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abbott, L. J., Parker, S., & Peters, G. F. (2004). Audit committee characteristics and restatements. *Auditing: A journal of practice & theory*, *23*(1), 69-87.

ACFE. (2018). Report to the Nations on Occupational Fraud and Abuse: 2018 Global Study on Occupational Fraud and Abuse, Association of Certified Fraud Examiners.

Alexander, C. R., & Cohen, M. A. (1996). New evidence on the origins of corporate crime. *Managerial and Decision Economics*, *17*(4), 421-435.

Al-Khazali, O. M., & Zoubi, T. A. (2005). Empirical testing of different alternative proxy measures for firm size. *Journal of Applied Business Research (JABR)*, *21*(3).

Barak, G. (2012). *Theft of a nation: Wall street looting and federal regulatory colluding*. Rowman & Littlefield Publishers.

Barnes, G., & Hyatt, J. (2012). Using random forest risk prediction in the Philadelphia probation department. *Washington, DC: National Criminal Justice Reference Service, Document NCJ241346*.

Baucus, M. S., & Near, J. P. (1991). Can illegal corporate behavior be predicted? An event history analysis. *Academy of management Journal*, *34*(1), 9-36.

Beasley, M. S. (1996). An empirical analysis of the relation between the board of director composition and financial statement fraud. *Accounting review*, 443-465.

Beasley, M. S., Carcello, J. V., Hermanson, D. R., & Lapides, P. D. (2000). Fraudulent financial reporting: Consideration of industry traits and corporate governance mechanisms. *Accounting horizons*, *14*(4), 441-454.

Beccaria, C. (1764). On crimes and punishment. Trans by Paolucci. H. IN: Bobbs-Merrill.

Bell, T. B., & Carcello, J. V. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory*, *19*(1), 169-184.

Bellman, R. (1961). Curse of dimensionality. *Adaptive control processes: a guided tour. Princeton, NJ*.

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

Benson, M. (2010). Evolutionary ecology, fraud, and the global financial crisis. In R. Rosenfeld, K. Quinet, and C. Garcia (Eds.), Contemporary Issues in Criminological Theory and Research: The Role of Social Institutions (pp. 299–306). Belmont, CA: Wadsworth.

Benson, M., Madensen T.D., & Eck, J.E. (2009). White Collar Crime from an Opportunity Perspective. In S. S. Simpson & D. Weisburd (Eds.), *The criminology of white-collar crime* (Vol. 228). New York, NY: Springer.

Benson, M. L., & Simpson, S. S. (2009). *White collar crime: An opportunity perspective*. New York, NY: Routledge.

Bentham,J. (1962). Principles of penal law. In John Bowring (Ed.), The Works of Jeremy Bentham (p.396). NewYork: Russell and Russell.

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, *13*(2), 193-216.

Bernard, T. J., & Snipes, J. B. (1996). Theoretical Integration in Criminology (From Crime and Justice: A Review of Research, Volume 20, P 301-348, 1996, Michael Tonry, ed.--See NCJ-161959).

Bhasin, M. L. (2013). Corporate accounting fraud: A case study of Satyam Computers Limited. *Open Journal of Accounting*, *2*, 26-38.

Bloomfield, R. J. (2002). The 'incomplete revelation hypothesis' and financial reporting. Cornell University Working Paper

Bloomfield, R. (2008). Discussion of "annual report readability, current earnings, and earnings persistence". *Journal of Accounting and Economics*, *45*(2-3), 248-252.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, 235-249.

Bozanic, Z., Dirsmith, M. W., & Huddart, S. (2012). The social constitution of regulation: The endogenization of insider trading laws. *Accounting, Organizations and Society*, *37*(7), 461-481.

Braithwaite, J. (2016). In search of Donald Campbell: Mix and multimethods. *Criminology & Public Policy*, *15*, 417.

Brazel, J. F., Jones, K. L., Thayer, J., & Warne, R. C. (2015). Understanding investor perceptions of financial statement fraud and their use of red flags: Evidence from the field. *Review of Accounting Studies*, *20*(4), 1373-1406.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199-231.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brennan, N. M., & McGrath, M. (2007). Financial statement fraud: Some lessons from US and European case studies. *Australian accounting review*, *17*(42), 49-61.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication theory*, *6*(3), 203-242.

Burns, N., & Kedia, S. (2006). The impact of performance-based compensation on misreporting. *Journal of financial economics*, *79*(1), 35-67.

Carcello, J. V., & Nagy, A. L. (2004). Audit firm tenure and fraudulent financial reporting. *Auditing: a journal of practice & theory*, *23*(2), 55-69.

Chan, F., & Gibbs, C. (2022). When guardians become offenders: Understanding guardian capability through the lens of corporate crime. *Criminology*.

Chan, P. K., & Stolfo, S. J. (1998). Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. In *KDD*, 98, 164-168.

Chen, M. Y. (2011). Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Computers & Mathematics with Applications*, *62*(12), 4514-4524.

Chollet, F. (2018). Keras: The python deep learning library. *Astrophysics Source Code Library*.

Clinard, M. B., & Yeager, P. C. (2006). Corporate crime. New Brunswick, NJ.

Cohan, J. A. (2002). " I didn't know" and" I was only doing my job": has corporate governance careened out of control? A case study of enron's information myopia. *Journal of Business Ethics*, *40*(3), 275-299.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American sociological review*, 588-608.

Coleman, J. W. (1987). Toward an integrated theory of white-collar crime. *American journal of Sociology*, *93*(2), 406-439.

Commers, M. S., & Vandekerckhove, W. (2004). Whistle blowing and rational loyalty. *Journal of Business Ethics*, *53*(1), 225-233.

Cressey, D. R. (1953). Other people's money; a study of the social psychology of embezzlement. New York, NY, US: Free Press.

Cropanzano, R., Byrne, Z. S., Bobocel, D. R., & Rupp, D. E. (2001). Moral virtues, fairness heuristics, social entities, and other denizens of organizational justice. *Journal of vocational behavior*, *58*(2), 164-209.

Dang, C., Li, Z. F., & Yang, C. (2018). Measuring firm size in empirical corporate finance. *Journal of banking & finance*, *86*, 159-176.

Davis, L. R., Soo, B. S., & Trompeter, G. M. (2007). Auditor tenure and the ability to meet or beat earnings forecasts. *Contemporary Accounting Research, 26(2), 517-548*.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1996). Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC. *Contemporary accounting research*, *13*(1), 1-36.

Dechow, P. M., Hutton, A. P., Kim, J. H., & Sloan, R. G. (2012). Detecting earnings management: A new approach. *Journal of accounting research*, *50*(2), 275-334.

Duwe, G., & Kim, K. (2016). Sacrificing accuracy for transparency in recidivism risk assessment: The impact of classification method on predictive performance. *Corrections*, *1*(3), 155-176.

Dyck, I. J., Morse, A., & Zingales, L. (2013). How pervasive is corporate fraud? *Rotman School of Management Working Paper*, (2222608).

Efendi, J., Srivastava, A., & Swanson, E. P. (2007). Why do corporate managers misstate financial statements? The role of option compensation and other factors. *Journal of financial economics*, *85*(3), 667-708.

Erickson, M., Hanlon, M., & Maydew, E. L. (2006). Is there a link between executive equity incentives and accounting fraud?. *Journal of accounting research*, *44*(1), 113-143.

Fanning, K. M., & Cogger, K. O. (1998). Neural network detection of management fraud using published financial data. *Intelligent Systems in Accounting, Finance & Management*, *7*(1), 21-41.

Farber, D. B. (2005). Restoring trust after fraud: Does corporate governance matter?. *The accounting review*, *80*(2), 539-561.

Farrington, D. P. (2000). Explaining and preventing crime: The globalization of knowledge—The American Society of Criminology 1999 presidential address. *Criminology*, *38*(1), 1-24.

Farrington, D.P. (2005). Integrated developmental and life-course theories of offending. Transaction Publishers: New Brunswick, NJ.

FBI. (2018). White-Collar Crime. US Department of Justice. Retrieved from https://www.fbi.gov/investigate/white-collar-crime

Feroz, E. H., Park, K., & Pastena, V. S. (1991). The financial and market effects of the SEC's accounting and auditing enforcement releases. *Journal of accounting research*, *29*, 107-142.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, *63*(4), 308-319.

Gross, E. (1980). Organizational structure and organizational crime. *White collar crime: Theory and research*, 52-76.

Hartshorn, S. (2016). *Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners*. Kindle Edition.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S* (pp. 249-307). Routledge.

He, H., & Ma, Y. (Eds.). (2013). Imbalanced learning: foundations, algorithms, and applications.

Hill, C. W., Kelley, P. C., Agle, B. R., Hitt, M. A., & Hoskisson, R. E. (1992). An empirical examination of the causes of corporate wrongdoing in the United States. *Human Relations*, *45*(10), 1055-1076.

Holtfreter, K., Van Slyke, S., Bratton, J., & Gertz, M. (2008). Public perceptions of white-collar crime and punishment. *Journal of Criminal Justice*, *36*(1), 50-60.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, *50*(3), 585-594.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

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

Koh, K., Matsumoto, D. A., & Rajgopal, S. (2008). Meeting or beating analyst expectations in the post-scandals world: Changes in stock market rewards and managerial actions. *Contemporary Accounting Research*, *25*(4), 1067-1098.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual review of psychology*, *50*(1), 569-598.

Kranacher, M. J., Riley, R. A, Jr, & Wells, J. T. (2011). *Forensic accounting and fraud examination*. New York: Wiley.

Lin, J. W., Hwang, M. I., & Becker, J. D. (2003). A fuzzy neural network for assessing the risk of fraudulent financial reporting. *Managerial Auditing Journal*, *18*(8), 657-665.

Liu, C., Chan, Y., Alam Kazmi, S. H., & Fu, H. (2015). Financial fraud detection model: based on random forest. *International journal of economics and finance*, *7*(7).

Lewicki, R. J., Wiethoff, C., & Tomlinson, E. C. (2005). What is the role of trust in organizational justice. *Handbook of organizational justice*, 247-270.

McCarthy, K., Zabar, B., & Weiss, G. (2005, August). Does cost-sensitive learning beat sampling for classifying rare classes?. In *Proceedings of the 1st international workshop on Utility-based data mining* (pp. 69-77).

McCornack, S. A. (1992). Information manipulation theory. *Communications Monographs*, *59*(1), 1-16.

McKendall, M. A., & Wagner III, J. A. (1997). Motive, opportunity, choice, and corporate illegality. *Organization Science*, *8*(6), 624-647.

Mesmer-Magnus, J. R., & Viswesvaran, C. (2005). Whistleblowing in organizations: An examination of correlates of whistleblowing intentions, actions, and retaliation. *Journal of business ethics*, *62*(3), 277-297.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.".

Myers, J. N., Myers, L. A., & Omer, T. C. (2003). Exploring the term of the auditor-client relationship and the quality of earnings: A case for mandatory auditor rotation?. *The accounting review*, *78*(3), 779-799.

Neuilly, M. A., Zgoba, K. M., Tita, G. E., & Lee, S. S. (2011). Predicting recidivism in homicide offenders using classification tree analysis. *Homicide studies*, *15*(2), 154-176.

Paternoster, R. (2016). Deterring Corporate Crime: Evidence and Outlook. *Criminology & Public Policy*, *15*, 383.

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

Perols, J. L., Bowen, R. M., Zimmermann, C., & Samba, B. (2016). Finding needles in a haystack: Using data analytics to improve fraud prediction. *The Accounting Review*, *92*(2), 221-245.

Perols, J. L., & Lougee, B. A. (2011). The relation between earnings management and financial statement fraud. *Advances in Accounting*, *27*(1), 39-53.

Pflueger, M. O., Franke, I., Graf, M., & Hachtel, H. (2015). Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. *BMC psychiatry*, *15*(1), 62.

Poppo, L., & Zenger, T. (2002). Do formal contracts and relational governance function as substitutes or complements?. *Strategic management journal*, *23*(8), 707-725.

Prechel, H., & Morris, T. (2010). The effects of organizational and political embeddedness on financial malfeasance in the largest US corporations:

Dependence, incentives, and opportunities. *American Sociological Review*, *75*(3), 331-354.

PwC. (2018). Global Economic Crime and Fraud Survey. London, U.K.: PwC.

Rigano, C. (2018). Using Artificial Intelligence to Address Criminal Justice Needs. *National Institute of Jusitice Journal*, 280, 37-46.

Ritter, N. (2013). Predicting Recidivism Risk: New Tool in Philadelphia Shows Great Promise. *National Institute of Jusitice Journal*, 271, 4-13.

Rorie, M., Alper, M., Schell-Busey, N., & Simpson, S. S. (2018). Using meta-analysis under conditions of definitional ambiguity: the case of corporate crime. *Criminal Justice Studies*, *31*(1), 38-61.

Salehi, M., & Fard, F. Z. (2013). Data mining approach to prediction of going concern using classification and regression tree (CART). *Global Journal of Management and Business Research Accounting and Auditing*, *13*.

Schell-Busey, N., Simpson, S. S., Rorie, M., & Alper, M. (2016). What works? A systematic review of corporate crime deterrence. *Criminology & Public Policy*, *15*(2), 387-416.

Schuchter, A., & Levi, M. (2016). The fraud triangle revisited. *Security Journal*, *29*(2), 107-121.

Seifert, D. L., Sweeney, J. T., Joireman, J., & Thornton, J. M. (2010). The influence of organizational justice on accountant whistleblowing. *Accounting, Organizations and Society*, *35*(7), 707-717.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shover, N., & Hochstetler, A. (2005). *Choosing white-collar crime*. Cambridge University Press.

Simpson, S. S. (2013). White-collar crime: A review of recent developments and promising directions for future research. *Annual Review of Sociology*, *39*, 309-331.

Simpson, S. S., Garner, J., & Gibbs, C. (2007). Why do corporations obey environmental law. *National Institute of Justice. Washington, DC: US Department of Justice*.

Simpson, S. S., & Koper, C. S. (1997). The changing of the guard: Top management characteristics, organizational strain, and antitrust offending. *Journal of Quantitative Criminology*, *13*(4), 373-404.

Simpson, S., Rorie, M., Alper, M. E., Schell-Busey, N., Laufer, W., & Smith, N. C. (2014). Corporate crime deterrence: A systematic review. *Campbell systematic reviews*, *10*(4).

Simpson, S. S., & Yeager, P. C. (2015). Building a Comprehensive White-Collar Violations Data System. *Washington, DC: Bureau of Justice Statistics*.

SEC. (2018). Accounting and Auditing Enforcement Releases No. 4012. Retrieved from https://www.sec.gov/litigation/admin/2018/33-10601.pdf

Schwartz, J., Steffensmeier, D., Moser, W. J., & Beltz, L. (2020). Financial Prominence and Financial Conditions: Risk Factors for 21st Century Corporate Financial Securities Fraud in the United States. *Justice Quarterly*, 1-29.

Smith, J. R., Tiras, S. L., & Vichitlekarn, S. S. (2000). The interaction between internal control assessment and substantive testing in audits for fraud. *Contemporary Accounting Research*, *17*(2), 327-356.

Sullivan, W. (2017). *Machine Learning For Beginners Guide Algorithms: Supervised & Unsupervsied Learning. Decision Tree & Random Forest Introduction*. Healthy Pragmatic Solutions Inc.

Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., & Collins, L. M. (2015). Financial fraud detection using vocal, linguistic and financial cues. *Decision Support Systems*, *74*, 78-87.

Trompeter, G. M., Carpenter, T. D., Desai, N., Jones, K. L., & Riley Jr, R. A. (2012). A synthesis of fraud-related research. *Auditing: A Journal of Practice & Theory*, *32*(sp1), 287-321.

Trompeter, G. M., Carpenter, T. D., Jones, K. L., & Riley Jr, R. A. (2014). Insights for research and practice: What we learn about fraud from other disciplines. *Accounting Horizons*, *28*(4), 769-804.

Unnever, J. D., Benson, M. L., & Cullen, F. T. (2008). Public support for getting tough on corporate crime: Racial and political divides. *Journal of Research in Crime and Delinquency*, *45*(2), 163-190.

Vaughan, D. (2005). Organizational rituals of risk and error. *Organizational encounters with risk*, 33-66.

Wang, X., & Holtfreter, K. (2012). The effects of corporation-and industry-level strain and opportunity on corporate crime. *Journal of Research in Crime and Delinquency*, *49*(2), 151-185.

Wolfe, D. T., & Hermanson, D. R. (2004). The fraud diamond: Considering the four elements of fraud. *The CPA Journal*, 74(12), 38-42.

Woodcock, R. A. (2019). The Antitrust Case for Consumer Primacy in Corporate Governance. *UC Irvine L. Rev.*, *10*, 1395.

Young, R. (2013). The role of organizational justice as a predictor of intent to comply with internal disclosure policies. *Journal of Accounting and Finance*, *13*(6), 29-44.

Zajac, E. J., & Olsen, C. P. (1993). From transaction cost to transactional value analysis: Implications for the study of interorganizational strategies. *Journal of management studies*, *30*(1), 131-145.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In *Advances in experimental social psychology* (Vol. 14, pp. 1-59). Academic Press.