공학박사 학위논문

Building Financial Misstatement Detection Models using Multi-class Cost-sensitive Learning and Feature Generation from CFO survey

다중 클래스 비용 의존 학습 방법과 CFO 설문조사를 활용한 변수생성을 통한 분식회계 예측 모형의 개발

2016 년 8 월

서울대학교 대학원
협동과정 기술경영경제정책 대학원

김 연 국

Building Financial Misstatement Detection Models using Multi-class Cost-sensitive Learning and Feature Generation from CFO survey

다중 클래스 비용 의존 학습 방법과 CFO 설문조사를 활용한 변수생성을 통한 분식회계 예측 모형의 개발

지도교수 조 성 준

이 논문을 공학박사 학위논문으로 제출함
2016 년 7 월

서울대학교 대학원
협동과정 기술경영경제정책 전공
김 연 국

김연국의 공학박사 학위논문을 인준함
2016 년 7 월

위 원 장 박 종 헌 (인)
부위원장 조 성 준 (인)
위 원 백 복 현 (인)
위 원 조 재 희 (인)
위 원 장 우 진 (인)

Abstract

Building Financial Misstatement Detection Models using Multi-class Cost-sensitive Learning and Feature Generation from CFO survey

Yeonkook J. Kim
Technology Management Economics Policy Program
The Graduate School
Seoul National University

Material misstatements are such omissions and misstatements of financial information included in the financial statements that can affect the economic decisions of the users of financial statements. When building material misstatement detection or prediction models, researchers should consider two important issues: first, financial misstatements can be classified as involving either errors (i.e., unintentional misapplications of accounting rules) or irregularities (i.e., intentional misreporting). Second, there is selection bias in target variables. Many firms that manipulate earnings are likely to go unidentified and there

could be selection biases in cases pursued by financial authorities.

I develop financial misstatement detection models using machine learning and statistical models addressing the two issues present in target variables. First, to address the issue of fraud intention in the target variable, I develop multi-class financial misstatement detection models to detect misstatements by fraud intention. Hennes, Leone and Miller (2008) performed post-event analysis on financial restatements and classified restatements by fraud intention (i.e. intentional vs. unintentional misstatements). Using their result (along with non-misstated firms) as a three-class target variable, I develop three multi-class classifiers, multinomial logistic regression, support vector machine, and Bayesian networks as predictive tools to detect and classify misstatements by fraud intention. To deal with class imbalance and asymmetric misclassification costs, I perform cost-sensitive learning using MetaCost.

Second, one way to reduce the effect of selection bias in the target variable is to perform domain expert guided feature selection. I propose to utilize the earnings management survey to public company CFOs by Dichev et al. (2016) for feature selection and feature generation. To detect material misstatements, I create features using the

survey result and build binary detection models. I compare the performance of the new models with the existing scoring models from accounting and finance literature.

Keywords: Financial misstatement detection; Fraud intention; Multi-class cost sensitive learning, CFO survey, Earnings Management
Student Number: 2011-30999

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Can we detect accounting fraud? How are intentional financial misstatements different from accounting irregularities without managerial intent? Answering these questions is of critical importance to the efficient functioning of capital markets and to increase our understanding of financial statement fraud. Fraudulent financial statements affect not just shareholders, but also lenders, creditors and employees. Perols (2011) estimates the cost of financial statement fraud in the U.S. to be $572 billion per year.

Due to the significance of this topic, academics have performed post-event studies extensively to understand the causes, motivations, and consequences of financial misstatements and earnings manipulation (Beneish, 1999; Dechow, Ge, & Schrand, 2010; Dechow, Sloan, & Sweeney, 1995, 1996; DeFond & Jiambalvo, 1994; Ettredge, Scholz, Smith, & Sun, 2010; Gillett & Uddin, 2005; Hennes, Leone, & Miller, 2013; Jones, Krishnan, & Melendrez, 2008; Palmrose & Scholz, 2004; Schrand & Zechman, 2012). Building upon these studies, various prediction/detection models have been proposed in

the accounting and data mining literature (Abbasi, Albrecht, Vance, & Hansen, 2012; Beneish, 1999; Cecchini, Aytug, Koehler, & Pathak, 2010; Dechow, Ge, Larson, & Sloan, 2011; Huang, Tsaih, & Yu, 2014; Kirkos, Spathis, & Manolopoulos, 2007; Kotsiantis, Koumanakos, Tzelepis, & Tampakas, 2006; Lin, Chiu, Huang, & Yen, 2015; Pai, Hsu, & Wang, 2011; Ragothaman, Carpenter, & Buttars, 1995).

To develop a detection/prediction model, databases such as financial restatements, lawsuits, and financial regulatory investigation data such as U.S. Securities and Exchange Commission (SEC) enforcement samples have been often used as a binary target variable. However, there are two problems with using these target variables as a binary target variable. The first problem is that financial misstatements can be classified as involving either errors (i.e., unintentional misapplications of accounting rules) or irregularities (i.e., intentional misreporting). The second problem is that there is some apparent selection bias in the target variables. For example, the most popular target variable is the SEC enforcement samples known as Accounting and Auditing Enforcement Releases (AAER) dataset. As Dechow et al. (2011) have pointed out, predicting "the AAER firms" (i.e. firms that the SEC identifies as having misstated earnings) is predicting the likelihood of engaging in an accounting misstatement and receiving an enforcement action from the SEC. There are factors such as politically connectedness that influence the likelihood of receiving an enforcement action from the SEC (Correia,

2014).

To deal with the first problem of fraud intention in financial misstatements, two approaches have been adopted when building detection models in accounting and datamining literature. For an illustration purpose, I summarize the Hennes et al. (2008) classification result in Figure 1.1.

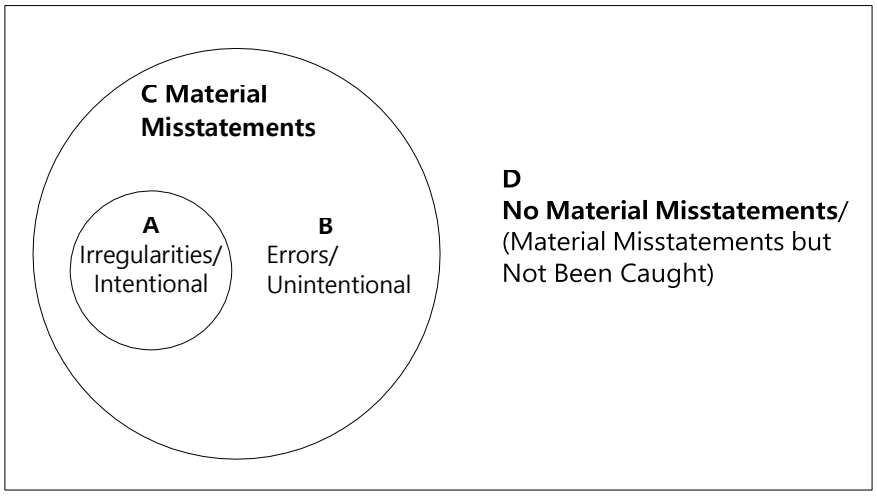

Figure 1.1: Hennes et al. (2008) classification result

The first approach is to classify the misstatement class (C=A+B) and the non-misstatement class (D). The second approach is to classify the intentional misstatement/fraudulent financial statement class (A) and the non-fraudulent financial statements (D). The shortcoming of the first approach is that we cannot make inferences about managerial misconducts. It may not be called as a fraud detection model since unintentional errors contaminate the data. On

the other hand, the weakness of the second approach is that information is lost by deleting unintentional misstatement class data. Thus, the second type of models underutilize the data and may not be effective when used to detect material but unintentional misstatements.

To resolve the problems of the existing binary detection models, I propose multi-class detection models in chapter 3 of this thesis. I classify data into three classes: 1. intentional misstatements/ fraudulent financial statements, 2. Unintentional misstatements, 3. No misstatements. To deal with class imbalance and asymmetric mis-classification costs, I undertake cost-sensitive learning using MetaCost.

The second problem is much more challenging problem with no direct solution. But it is also the problem that is prevalent in many social science and business study settings. In financial material misstatement detection modelling, proper feature selection process is critical to minimize the effect of selection bias. For example, misstated firms (AAERs) are much larger than average (or medium) size firms in Compustat, a database of financial, statistical and market information. This may be due to the fact that larger firms tend to have more institutional investors and more analyst covering that their financial reporting and internal controls are more likely to be scrutinized and analyzed. Also, the SEC may be more likely to pursue cases where stock price declines rapidly after the manipulation is

revealed, because the identifiable losses to investors are greater (Dechow et al., 2011). Therefore, the use of features like market capitalization or total asset value will improve classification accuracy but it may not improve general misstatement detectability.

One way to deal with this problem is to do domain expert-guided feature selection. What I propose in chapter 4 is to do feature selection and generation based on the CFO survey result compiled and analyzed in Dichev et al. (2016, 2013). Dichev et al. (2016) conduct a survey with 375 CFO's of private and public companies on earnings manipulation. Based on the survey, they list twenty "red flags" of earnings misrepresentation. Using these red flags, I generate features and build misstatement detection models. Then I compare the performance of the new models against the existing scoring models in accounting and finance literature.

Chapter 2

Literature Review

2.1 Fraud detection

West and Bhattacharya (2016) group common types of financial fraud into three groups: bank fraud (e.g. credit card fraud, Mortgage fraud, money laundering), corporate fraud (e.g. financial statement fraud, securities and commodities fraud) and insurance fraud (e.g. automobile insurance fraud, health card fraud). Researchers have analyzed various types of financial fraud (Fawcett and Provost, 1997; Bolton and Hand, 2002) and have proposed statistical and machine learning methods to detect fraud effectively (Ngai et al., 2011; West and Bhattacharya, 2016). For example, Dal Pozzolo, Caelen, Le Borgne, Waterschoot, and Bontempi (2014) and Van Vlasselaer et al. (2015) propose credit card fraud detection models.

Among various types of financial fraud, accounting researchers have performed post-event studies extensively to understand the causes, motivations, and consequences of financial statement fraud (Beasley, 1996; Dechow et al., 1995, 1996; Beneish, 1999b; Erickson et al., 2006; Jones et al., 2008; Badertscher, 2011). Many of these post-event studies use the binary target variable made of fraud (misstatement) firms and non-fraud (non-misstatement) firms.

2.2 Classification of material misstatements according to fraudulent intention (post-event studies)

Researchers recently began to investigate and classify financial material misstatements according to management intent to mislead, manipulate or defraud, rather than to simply tag them all as examples of fraud. For example, Beasley (1996) searches Accounting and Auditing Enforcement Releases (AAERs) which as issued by the SEC during or at the conclusion of an investigation against a company, an auditor, or an officer for alleged accounting and/or auditing misconduct. He identifies fraud firms by removing AAERs not involving financial statement fraud (e.g., unintentional misapplication of GAAP).

In an award winning study, Hennes et al. (2008) formally propose the following three rules to classify financial restatements as errors (unintentional) or irregularities (intentional):

1. Classify any restatements using variants of the words "fraud" or "irregularity" in reference to the misstatement in 8-K filings as irregularities
2. Classify restatements with related SEC or Department of Justice investigations as irregularities
3. Presence or absence of other investigations into accounting matter (e.g., the audit committee hires a forensic accounting firm): classify restatements with related independent investigations as irregularities.

They perform three validity tests to support their classification approach. The first validity test shows a significant difference in the stock market reactions between the two groups: the mean (median) cumulative abnormal return for the unintentional-misstatement sample was -1.93 %(0.90%) compared to -13.64 %(-19.4%) for the intentional-misstatement sample. The second validity test compares the frequency of securities class-action lawsuits, showing that 84 of the 105 intentional misstatements in their sample had contemporaneous class-action lawsuits while one of the 83 unintentional-misstatement samples had a related lawsuit. The third validity test shows that the percentage of restating firms experiencing CFO/CEO turnover in the 13 months surrounding the restatements (six months before to six months after) was 49%(64%) for CEOs(CFOs) in their intentional-misstatement sample but only 8%(12%) in the unintentional-misstatement sample. In the analysis for CFO/CEO turnover, they showed that the power of the hypotheses test on accounting restatements significantly improved either by limiting restatement samples to intentional misstatements or by including a control variable distinguishing unintentional misstatements from intentional misstatements. In a later study, Plumlee and Yohn (2010) classify financial restatements into four groups: intentional manipulation, internal company error, transaction complexity, and accounting standards.

Hayes (2014) proposes a simple text-search approach to classify

financial restatements as unintentional errors or intentional misstatements, or as unclassified. She analyzes how intentional misstatements and unintentional errors differ, finding first that unintentional errors are associated with a market reaction of a smaller magnitude, are more likely to be associated with weak internal controls, occur more frequently, affect a broader range of accounts, and are likely to lead to auditor turnover if a restatement occurs (Plumlee and Yohn, 2010; Hennes et al., 2008; Palmrose and Scholz, 2004). Moreover, whereas external users and audit committee members can rationally infer the presence of intentional misstatements by managerial incentives to manage earnings or commit fraud to meet certain targets, unintentional errors are less transparent and thus may be more likely to result in decision errors.

Second, unintentional errors may reflect the (lack of) competence or ability rather than the integrity of management and auditors. While intentional misstatements may be partially due to management incentives, unintentional errors can indicate either management's inability (Demerjian et al., 2012b) or a lack of incentive to implement and maintain effective controls over financial reporting (Hoitash et al., 2012).

Third, management may react differently to an auditor's detection of intentional vs. unintentional misapplications of GAAP. Management motivated by concerns such as shareholder reactions to missing analysts' forecasts is more likely to resist correcting an

intentional misstatement. Non-fraud-intended management, on the other hand, should be less resistant to correcting material unintentional errors. Therefore, restatements that correct unintentional errors indicate that management has failed to implement and maintain effective controls over financial reporting and that the auditor lacks the requisite skills to plan and conduct an effective audit.

In short, these studies show that intentional misstatements and unintentional misstatements are two different types of events and that distinction between the two is important to increase the power of the tests. Moreover, these studies are ex-post studies, suggesting that ex-ante studies are warranted. According to a report by the Committee of Sponsoring Organizations of the Treadway Commission on the SEC investigation between 1998 and 2007, the median fraud period was two years. This implies that it takes approximately two years for financial statement fraud to be identified, investigated and announced. For example, on November 8th, 2001, Enron Corp. announced that it would restate earnings for the period 1997-2001. Then, on December 2nd, 2001, Enron and 13 of its subsidiaries filed for Chapter 11 bankruptcy protection (GAO, 2002). This suggests that financial fraud was revealed approximately three years after its first misstated financial statements had been issued. My goal is to detect Enron's misstatement in the first year it takes place and predict if Enron intended to defraud and thus mislead readers of its financial statement.

2.3 Binary Prediction/Detection models

Most financial misstatement detection/prediction models deal with a binary problem. I reprint the classification result of Hennes et al. (2008) in Figure 2.1. Studies either classify intentional-misstatement firms (or a subset of A as fraudulent firms[1]) and non-misstatement firms (i.e. A vs. D, e.g. Cecchini et al. (2010)) or classify misstatement firms and non-misstatement firms (C vs. D, where C=A+B, e.g. Dechow et al. (2011)).

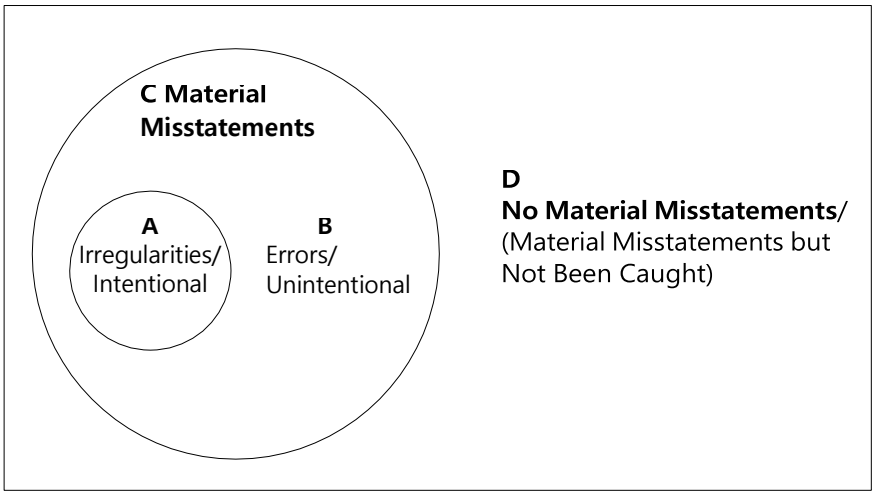

Figure 2.1: Hennes et al. (2008) classification result.

I describe several previous studies. First, Beneish (1999a)

[1] Although the distinction between fraud and irregularities has become blurred over the years, the two terms are technically not identical (Hennes et al., 2008). Studies such as that by Erickson et al. (2006) make a strict distinction between the two. However, auditing guidelines (e.g. SAS No. 82, AICPA 1997) use the term "fraud" and the term "intentional misstatements" interchangeably (Hennes et al., 2008).

classifies intentional earnings manipulators and non-manipulators (A vs D) and develops a probit model termed the M-score model using eight financial ratios to predict cases of upward earnings manipulation. The study explains that the model may make two types of errors; it can classify a company as a non-manipulator when it manipulates (a Type I error), or it can classify a company as a manipulator when it does not manipulate (a Type II error). The probability cutoffs that minimize the expected costs of misclassification depend on costs associated with the relative cost of making an error of either type. He compares the expected costs by increasing relative costs of Type I to Type II errors from 1:1 to 100:1.

Extending the study of Beneish (1999a), Dechow et al. (2011) classify misstated and non-misstated firms (C vs. D) and develop F-score models using logit models. Unlike the M-score model, which uses only financial statement variables, the F-score models use financial statement variables, market-related variables, off-balance sheets and other nonfinancial variables.

Applying a more flexible modelling approach, Cecchini et al. (2010) classify intentionally misstated (fraudulent) and non-misstated firms (A vs. D) using support vector machine classifier that incorporates a custom financial kernel. The financial kernel is a graph kernel that uses input financial variables to derive implicitly numerous financial ratios (i.e., 24 financial statement variables into 1,518 features in the study). They argue that because fraud tactics change over the years,

a method that utilized exhaustive combinations of potential fraud variables has a better chance of effectively catching fraud as compared to methods that restrict themselves to a few possible constructs. They show that their model outperforms a logit model (F-score model) and the neural network model by Green and Choi (1997) and argue that the superior performance stems more from the additional features used than from the specific induction technique used. To counter class imbalance and non-symmetric mis-classification costs, they apply different weightings for fraud and non-fraud classes and achieve their best result at 200:1.

Abbasi et al. (2012) also classify fraudulent firms and non-fraud firms (A vs D) using ensemble and adaptive learning. Similar to Cecchini et al. (2010), the authors argue that prior financial fraud-detection studies utilize feature sets that are too small in number and exclusively from annual statements, leading to feature sets lacking representational richness and that are simply not large enough to generate appropriate hypothesis spaces for the classification methods utilized. To alleviate this problem, they incorporate the following refinements:

(1) The inclusion of organizational and industry-level context information and

(2) The utilization of data based on quarterly and annual statements.

The study uses twelve financial statement ratios as seed variables and constructs, first, the organizational context features by computing

the difference between (−) and the ratio of (/) the firms' current period seed financial ratios and previous time period ratios. Second, they construct the industry−level con− text features using two industry−representative models; Top−5 (industry leaders) models are created by averaging the data from the five largest companies in each industry−year (in terms of sales), creating what Whiting et al. (2012) refer to as "centroid" firms. The twelve seed financial ratios from centroid firms are then generated and compared to misstated/non−misstated firm ratios. Whiting et al. (2012) argue that industry−representative centroid models measure a majority influence on what is normal for each industry. Also, by averaging the top five firms in each industry, the individual differences between companies are smoothed, providing a better industry−representative model. Further−more, given that single comparison companies cannot be guaranteed not to be fraudulent or misstated, this method provides a more robust and theory−driven context.

Similarly, closest−5 models (industry peers) are created for each firm by averaging the data from the five companies in the same industry−year that are most similar in terms of sales. Organizational context features and industry−level context features are created using both annual data and quarterly data. Therefore, 84 yearly context−based feature sets and 336 quarterly context−based feature sets are created and used to train 14 base classifiers. They perform stacking and semi−supervised learning and achieved a fraud−

detection rate of 88 percent and an area-under-the-receiver operating characteristic curve which exceeded 0.9.

In summary, two-class models that classify C and D (as in Figure 2.1) measure the patterns of both intentional misstatements and unintentional misstatements. On the other hand, binary models that classify A and D only measure the difference between intentional misstatements and non-misstatements. Unless the patterns of intentional misstatements and unintentional misstatements are ordinal in the measurements, this approach may not be effective when used to detect material but unintentional misstatements.

Chapter 3

Detecting financial misstatements by fraud intention using multi-class cost-sensitive learning

An intentional misstatement arises when management has incentives to manage earnings or commit fraud to meet certain objectives, such as maximizing personal gain through stock-based compensation (Erickson et al., 2006). When a restatement, the revision and publication of one or more of a company's previous financial statements with a material inaccuracy, is announced, investors show different responses depending on the presence of fraud intention. For example, Palmrose, Richardson and Scholz (2004) report that the market reacts to restatement announcements differently showing an average abnormal return of -20% for financial restatements caused by deliberate misreporting as opposed to an average abnormal return of -6% for non-fraud restatements. Also, when intentional misstatements are announced, a higher CEO/CFO turnover rate and more frequent securities class-action lawsuits follow compared to when unintentional restatements are announced (Hennes et al., 2008).

In comparison to intentional misstatements, unintentional misstatements are more likely to result from weak internal controls, to

occur more frequently, to affect a broader range of accounts, and are likely to lead to auditor turnover (Hayes, 2014). Moreover, unintentional errors indicate either management is less reputed (Demerjian et al., 2012b) or less incentivized to implement and maintain effective controls over financial reporting (Hayes, 2014).

Due to the difference between intentional and unintentional misstatements, researchers testing hypotheses involving managerial misconduct are at risk of making incorrect inferences regarding their hypotheses if they do not specifically distinguish intentional misstatements from unintentional errors (Hennes et al., 2008). This is especially critical given that the relative frequency of error- related misstatements has increased due to the tighter regulation in the post-Enron regulatory environment. Furthermore, if researchers limit their samples to only fraudulent misstatements for their detection models, they are underutilizing information by throwing away more commonly occurring unintentional misstatements. As a result, their models may not effectively detect more frequent but less egregious misstatements or discriminate between intentional misstatements and unintentional errors.

To resolve the problems of the existing binary detection models, I propose multi-class detection models in this chapter. Specifically, I classify data into three classes: 1. Intentional misstatement (Irregularity); 2. Unintentional misstatement (Error); and 3. No misstatement. To deal with asymmetric misclassification costs in

building misstatement detection models, I undertake cost-sensitive learning using MetaCost, a wrapper approach developed by Domingos (1999).

Table 3.1: Summary of literature review

	Binary Target Variable	Tertiary Target Variable
Post-event Analyses	Beasley (1996); Dechow et al. (1995, 1996); Beneish (1999b); Erickson et al. (2006); Jones et al. (2008); Badertscher (2011)	Hennes et al. (2008); Plum- lee & Yohn (2010); Hayes (2014)
Predictive Studies	Green & Choi (1997); Beneish (1999a); Kotsiantis et al. (2006); Kirkos et al. (2007); Cecchini et al. (2010); Dechow et al. (2011); Pai et al. (2011); Perols (2011); Abbasi et al. (2012)	Chapter 3

As shown in Table 3.1, previous studies have focused on building binary detection/prediction models. Also, even though there are a number of post-event studies classifying restatements into multiple classes, there has not been multi-class detection models classifying material misstatements according to the presence of fraud intention.

In the following section, I describe overall research methodology. Then I describe the data used and the testing methodology employed. I present the results in section 3.4.

3.1 Research Methodology

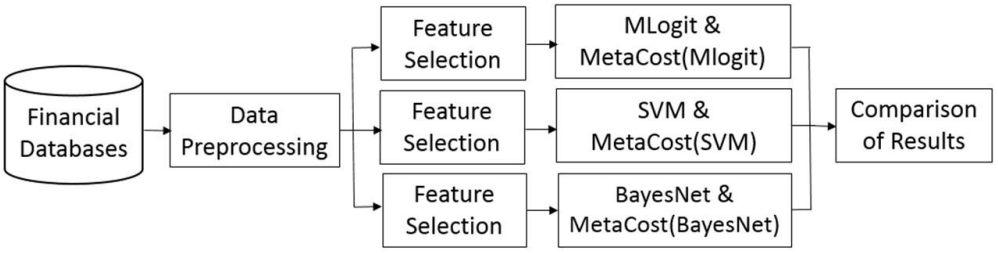

Figure 3.1: Research Framework

As shown in Figure 3.1, my proposed research framework consists of several steps. First, I gather restatement and financial data from various databases such as Compustat. Second, in the data preprocessing step, I create variables/ratios based on the past accounting and datamining studies. Regarding missing values and outliers, I follow the guidelines in the literature that use the variables. For example, the Asset Quality index variable (the first variable in Table 3.2) that measures year-to-year changes, I winsorize the data at the 1 percent and 99 percent percentiles. Then, I set the missing values to 1 as described in Beneish (1999a). Third, I perform feature selection process for each model separately. Using Weka 3.7.12, a popular machine learning software, I test both the filter method (e.g. Info gain attribute evaluator) and the wrapper method (e.g. Bayesian Network model with the genetic algorithm search). I select the feature selection process that gives the best performance for each classifier. For example, for the Bayesian Network model, I use the wrapper

method with the Particle Swarm Optimization algorithm to do feature selection. Lastly, I build the models and compare the result.

3.2 Variables

I use the post-event classification data on financial restatements from Hennes et al. (2008) as a three-class target variable. My misstated dataset contains 788 instances with 214 instances of irregularities and 355 as errors for the period of 1992 to 2005. The non-misstated dataset contains 2,156 instances selected based on industry and fiscal year.

I select and test various features from prior post-event studies and detection studies. Financial statement fraud evolves with time as perpetrators find new and clever ways to circumvent implemented procedures and controls. At the same time, new regulations or weak internal control systems in rapidly growing firms may cause unintentional misstatements. In order to effectively detect and discriminate between intentional misstatements and unintentional errors, it is necessary to use not just financial statement data but also various other features that are found to have some predictability in detecting financial misstatements. Researchers have found that off-balance sheet variables (e.g., operating leases), nonfinancial measures (e.g., abnormal changes in employees), market variables (e.g. market-adjusted stock returns) and governance measures (e.g. CEO power) may signal the presence of possible conditions of accounting

manipulation. I list the final selection of 49 features included in my experiment and the corresponding references in Table 3.2. The Y/Q column indicates whether a feature is an annual ratio/measure or a quarterly ratio/measure.

Table 3.2: Feature list

No	Variable	Definition	Y/Q	Reference
1	Asset Quality Index	Ratio of non-current assets other than property, plant, and equipment (PP&E), to total assets, for time period t relative to time period t-1	Y, Q	Abbasi et al. (2012); Beneish (1999)
2	Inventory Growth	Inventory at period t/Inventory at period t-1	Y, Q	Abbasi et al. (2012)
3	Accounts Receivable to Total Assets	Accounts Receivable /Total Assets	Y, Q	Dechow et al. (2011)
4	Soft assets	(Total Assets - PP&E – Cash and Cash Equivalent) / Total Assets	Y, Q	Dechow et al. (2011)
5	PP&E to Total Assets	PP&E / Total Assets	Y, Q	Perols (2011)
6	Receivable	Accounts Receivable	Y, Q	Perols(2011); Green & Choi (1997)
7	RSST Accrual	Accrual measure following Richardson et al. (2005)	Y	Dechow et al. (2011)
8	Total accruals to total assets	Change in working capital accounts other than cash less depreciation/ Asset	Q	Beneish (2004)
9	Cash Flow Earnings Correlation	Correlation between earnings and cash flow from operation	Q	Dichev, Graham, Harvey & Rajgopal (2013)
10	Earnings smoothness	Standard deviation of earnings divided by standard deviation of cash flow from operation	Q	Dichev et al. (2013)
11	Asset Turnover	Net sales/Total assets	Y, Q	Abbasi et al. (2012);
12	Operating Margin	Net income/Net sales	Y, Q	Abbasi et al. (2012)
13	Depreciation Index	Ratio of the rate of depreciation in period t-1 to the corresponding measure in period t	Y, Q	Abbasi et al. (2012); Beneish (2000)
14	Days Sales in Receivables	Ratio of day sales in receivables in period t to the corresponding measure in period t-1	Y, Q	Abbasi et al. (2012); Beneish (2000))
15	Gross Margin Index	Ratio of the gross margin in period t-1 to the gross margin in period t	Y, Q	Abbasi et al. (2012); Beneish (2000)

16	SGE Expense	Ratio of selling and general administrative expenses to net sales in period t by the same ratio in period t-1	Y, Q	Abbasi et al. (2012); Beneish (2000)
17	Sales Growth	Net sales in period t divided by net sales in period t-1	Y, Q	Abbasi et al. (2012); Beneish (2000)
18	Change in Cash sales	Percentage change in cash sales where cash sales = (Sales- △Accounts Receivable)	Y, Q	Dechow et al. (2011)
19	Change in return on assets	Earnings⁄ Average total assets in period t divided by the same ratio in period t-1	Y, Q	Dechow et al. (2011)
20	Bloat	Net operating assets in year t divided by total sales at the end of year t-1	Y, Q	Ettredge (2010); Badertscher(2011)
21	Inverse of the firm's interest coverage ratio	Interest expense in year t, divided by operating income before depreciation in year t-1;	Y, Q	Badertscher(2011)
22	Special Items as a Percentage of Sales	Special Items / Sales	Y, Q	McVay (2006)
23	Deferred tax expense	Deferred tax expense for period t／ total assets for t-1	Y, Q	Dechow et al. (2011)
24	Debt to Equity Ratio	Debt/Equity	Y, Q	Perols (2011)
25	Firm efficiency	Revenue generating ability given level of resources such fixed assets and R&D expenses	Y	Demerjian, Lev & McVay (2012); Demerjian et al.(2013)
26	Unexplained Discretionary expenses	Error term from industry discretionary expense regression equation	Y	Roychowdhury (2006)
27	Unexplained Production costs	Error term from industry production cost regression equation	Y	Roychowdhury (2006)
28	Cash flow from operations over asset	Cash flow from operations at period t scaled by asset at period t-1	Y	Roychowdhury (2006)
29	Unearned Revenue Long Term over Revenue	Long-term Deferred Revenue / Revenue	Y, Q	GMI Ratings (2013)
30	Accounts Payable over Operating Expenses	Accounts Payable /Operating Expenses	Y, Q	GMI Ratings (2013)
31	Abnormal change in employees	Percentage change in the number of employees /percentage change in assets	Y	Dechow et al. (2011)
32	Change in operating lease activity	Change in the present value of future noncancelable operating lease obligations deflated by average total assets	Y	Dechow et al. (2011)

33	Existence of operating leases	Indicator variable coded 1 if future operating lease obligations are greater than zero	Y	Dechow et al. (2011)
34	Book-to-market ratio	(Asset-Liabilities) /Market capitalization	Y	Badertscher(2011); Ettredge (2010)
35	Market value	Natural log of the market value of equity	Y	Ettredge (2010)
36	Actual issuance	Indicator variable coded 1 if the firm issued securities during year t	Y	Dechow et al. (2011)
37	Acquisition	Indicator variable equal to 1 if firm had an acquisition that contributed to sales in the prior year is greater than 0, zero otherwise	Y	Ettredge (2010)
38	Meeting Analyst forecast	Indicator variable equal to 1 if the firm meets or beats the median annual analyst earnings forecast; 0 otherwise	Y	Badertscher(2011); Perols (2011)
39	Analyst forecast error	Median analyst earnings forecast - actual EPS	Y	Badertscher(2011)
40	CEO Bonus	Ratio of the CEO's bonus divided by the total compensation received by the CEO in year t	Y	Badertscher(2011)
41	CEO Chairman	indicator variable that equals 1 if the CEO also holds the title of chairperson in year t, and 0 otherwise	Y	Badertscher(2011)
42	Owner	Sum of restricted stock grants in the current period and the aggregate number of shares held by the executive (excluding stock options) scaled by total outstanding shares	Y	Badertscher(2011)
43	CEO Salary	Amount of CEO's base salary in year t;	Y	Badertscher(2011)
44	Stock options	Number of unexercised options (including options grants in the current period) that the executive held at year-end t-1 scaled by total outstanding shares of the firm	Y	Badertscher(2011)
45	Short interest	Short interest ratio at year end	Y	Dechow et al. (2011)
46	Cash flow from financial activities over asset	Level of finance raised ⁄ Average total assets	Y, Q	Dechow et al. (2011)
47	Change in leverage Ratio	Ratio of sum of short-term debt and long-term debt in period t, scaled by total assets in period t-1 divided by the same ratio one period later	Y, Q	Abbasi et al. (2012);
48	Market-adjusted stock return	Annual buy-and-hold return inclusive of delisting returns minus the annual buy-and-hold value-weighted market return	Y	Dechow et al. (2011)

49	Lagged market-adjusted stock return	Previous year's annual buy-and-hold return inclusive of delisting returns minus the annual buy-and-hold value-weighted market return	Y	Dechow et al. (2011)

Following Abbasi et al. (2012), I create top−5 models (industry leaders) and closest−5 models (industry peers) for the variables that have sufficient data. For a yearly financial ratio, for an example, I compare it with the top−five firm (sales−weighted) average value and with the average value from five peer firms. In addition, I calculate the changes by dividing (or subtracting) the current year ratio by the prior year ratio (as in Table 3.3), creating six additional variables to be used in my experiments. I repeat the process with quarterly variables. In Table 3.3, R1Q1 signifies the quarterly ratio number (R1 means the ratio number one) and the number of quarter (Q1 means the first quarter). Since there are four quarters in a given fiscal year, I have four quarterly ratios. For the top−5 model, I compute the difference between (−) and the ratio of (/) the firms' current quarterly ratios and the same ratio of the average of five largest companies in the same period. I repeat the same procedure with the averages of five industry peers (firms with the most similar sales amounts in the same industry) to create the closest−5 model variables. The organizational context variables are created by computing the changes of the firms' current quarterly ratios from previous quarterly ratios. I generate in total 1,086 features. Following Abbasi et al. (2012), I create top−5 models (industry leaders) and closest−5 models (industry peers) for the variables that have sufficient data. For a yearly financial ratio, for an

example, I compare it with the top-five firm (sales-weighted) average value and with the average value from five peer firms. In addition, I calculate the changes by dividing (or subtracting) the current year ratio by the prior year ratio (as in Table 3.3), creating six additional variables to be used in my experiments. I repeat the process with quarterly variables. In Table 3.4, R1Q1 signifies the quarterly ratio number (R1 means the ratio number one) and the number of quarter (Q1 means the first quarter). Since there are four quarters in a given fiscal year, I have four quarterly ratios. For the top-5 model, I compute the difference between (-) and the ratio of (/) the firms' current quarterly ratios and the same ratio of the average of five largest companies in the same period. I repeat the same procedure with the averages of five industry peers (firms with the most similar sales amounts in the same industry) to create the closest-5 model variables. The organizational context variables are created by computing the changes of the firms' current quarterly ratios from previous quarterly ratios. I generate in total 1,086 features.

Table 3.3: Yearly Context-Based Feature Set Type

Type	Description
Yearly financial ratio	R1
Industry-level context: Top-5 Model	R1-T1, R1/T1
Industry-level context: Closest-5 Model	R1-C1, R1/C1
Organizational context	R1-P1, R1/P1

Table 3.4: Quarterly Context−Based Feature Set Type

Type	Description
Quarterly financial ratios	R1Q1, R1Q2, R1Q3, R1Q4
Industry-level context: Top-5 Model	R1Q1-T1Q1, R1Q2-T1Q2, R1Q3-T1Q3, R1Q4-T1Q4 R1Q1/T1Q1, R1Q2/T1Q2, R1Q3/T1Q3, R1Q4/T1Q4
Industry-level context: Closest-5 Model	R1Q1-C1Q1, R1Q2-C1Q2, R1Q3-C1Q3, R1Q4-C1Q4 R1Q1/C1Q1, R1Q2/C1Q2, R1Q3/C1Q3, R1Q4/C1Q4
Organizational context	R1Q2-R1Q1, R1Q3-R1Q2, R1Q4-R1Q3, R1Q1-R1Q4 R1Q2/R1Q1, R1Q3/R1Q2, R1Q4/R1Q3, R1Q1/R1Q4

3.3 Methods

To perform the classification, I employ three multi−class classifiers: Multinomial logistic regression, Support Vector Machines, and Bayesian Networks. I select the three models based on prior studies. According to Abbasi et al. (2012), various statistical and machine learning models are used in developing financial misstatement detection models. However, the most popular model is the logistic regression model: 10 out of 15 financial statement fraud detection studies in their review employ logistic regression models. Logistic regression models can be used in binary and multi−class classifications. Also, logistic regression models are often used in related accounting literatures because the models are interpretable and statistical inference can be easily made (Dechow et al., 1995; Beneish, 1999a; Jones et al., 2008; Ettredge et al., 2010; Dechow et

al., 2011).

On the other hand, in more recent predictive studies, support vector machines are a more popular choice due to their high classification accuracy (Cecchini et al., 2010; Pai et al., 2011; Perols, 2011; Abbasi et al., 2012). For example, Perols (2011) compares the performance of six popular statistical and machine learning models in detecting financial statement fraud under different assumptions of misclassification costs and ratios of fraud firms to non-fraud firms. He shows that logistic regression and support vector machines outperform an artificial neural network, bagging, C4.5, and stacking.

Kirkos et al. (2007) also investigate the usefulness of decision trees, neural networks and Bayesian networks in the fraudulent financial statement detection and show that the Bayesian network model achieves the best result. Bayesian network models can do multi-class classification, directly produce class probability estimates (which is useful for the cost-sensitive learning later), and performs well even when the data is highly class-imbalanced (Leong, 2015).

3.3.1 Multinomial logistic regression (MLogit)

Multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems. It predicts the probabilities of different possible outcomes of a categorically distri-buted target variable given a set of independent features. More

specifically, it models the posterior class probabilities Pr(G = j | X = x) for J classes via linear functions in x while at the same time ensuring that they sum to one and remain between 0 and 1. The model has the form:

$$Pr(G = j|X = x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}}, \quad \sum_{k=1}^{J} F_k(x) = 0, \tag{3.1}$$

where $F_j(x) = \beta_j^T * x$ are linear regression functions. The model is usually fit by finding maximum likelihood estimates for the parameters β_j. Among the variants of logistic regression models, I use SimpleLogistic classifier implemented in Weka 3.7.12, which uses the LogitBoost algorithm to fit the logistic models. LogitBoost performs the forward stage-wise fitting of additive logistic regression models by generalizing the Equation 3.1 to $F_j(x) = \Sigma_m f_{mj}(x)$, where f_{mj} can be arbitrary functions of the input variables that are fit by least squares regression. The SimpleLogistic classifier determines the best number of Logit-Boost iterations by cross-validation. During the cross-validation process, only those attributes that improve the performance are included. In this manner, automatic feature selection is performed (Landwehr et al., 2005).

3.3.2 Support Vector Machines (SVM)

SVM use a linear model to implement nonlinear class boundaries by mapping input vectors nonlinearly into a high-dimensional feature

space. In the new space, an optimal separating hyperplane is constructed. The training examples that are closest to the maximum margin hyperplane are known as support vectors. All other training examples are irrelevant with regard to defining the binary class boundaries. Good separation is achieved by the hyperplane with the greatest distance to the nearest training-data point of any class, as in general the larger the margin, the lower the generalization error of the classifier (Han et al., 2011).

SVM is directly applicable for two-class tasks. Therefore, it is necessary to apply algorithms that reduce a multi-class task to several binary problems. The most common approach is to build binary classifiers using one of the classes and the rest (one-versus-all) or with every pair of classes (one-versus-one).

Unlike a probabilistic classifier (such as logistic regression) that is able to predict, given a sample input, a probability distribution over a set of classes, SVM produces only scores as the output. To obtain proper probability estimates, a common approach in the binary case is to apply Platt scaling, which learns a logistic regression model with the scores (Platt, 1999). In the multi-class case, the predicted probabilities can be coupled using the pairwise coupling algorithm by Hastie and Tibshirani (Hastie et al., 1998; Witten and Frank, 2005).

3.3.3 Bayesian Networks (BayesNet)

Bayesian classification is based on the Bayes theorem, a method to determine the probability that a given hypothesis is true. The theorem

states that for a hypothesis H (such as whether an object X can be classified within a given class), the probability P is given as follows:

$$P(H|X) = \frac{(P(X \mid H) \times P(H))}{P(X)} \qquad (3.2)$$

A Bayesian classifier calculates $P(C_i|X)$ for all possible classes and inserts X into the class with the highest conditional probabilities. Bayesian networks are probabilistic graphical models that allow the representation of dependencies among subsets of attributes. More specifically, a Bayesian Network is a directed acyclic graph, where each node represents an attribute and each arrow represents an instance of probabilistic dependence. If an arrow is drawn from node A to node B, then A is the parent of B and B is a descendent of A. In a Bayesian Network, each variable is conditionally independent of its non-descendents, given its parents (Han et al., 2011). For each node X, there exists a conditional probability table, which specifies the conditional probability of each value of X for each possible combination of the values of its parents. The probability of an instance having m attributes is expressed as shown in Equation 3.3 below.

$$P(x_1, x_2, \ldots x_m) = \prod P(x_i \mid \text{Parents}(x_i)) \qquad (3.3)$$

The network structure can be defined in advance or can be inferred from the data. To perform the classification, one of the nodes can be defined as the class node, with the network then calculating the probability of each alternative class.

3.3.4 Cost-sensitive learning

Cost-Sensitive Learning is a type of learning in data mining that takes the misclassification costs (and possibly other types of cost) into consideration. Its goal is to minimize the total cost (Ling and Sheng, 2010). Cost-sensitive learning methods, such as the MetaCost procedure, deal with class-imbalance by incurring different costs for different classes (Ling and Sheng, 2010). It is feasible to handle unequal misclassification costs and class-imbalance in a unified framework using cost-sensitive learning as long as the data is not very severely class-imbalanced (Liu and Zhou, 2006). Prior research proposed various cost-sensitive learning methods for multi-class classification tasks (Domingos, 1999; Witten and Frank, 2005; Sun et al., 2006; Zhou and Liu, 2006; Bourke et al., 2008; Xia et al., 2009; Zhou and Liu, 2010; Bermejo et al., 2011).

Ling and Sheng (2010) categorize cost-sensitive learning into two categories. The first category is to design classifiers that are cost-sensitive in themselves. One example is a cost-sensitive decision tree (Drummond and Holte, 2000).

The second category is to design a "wrapper" that converts any existing cost-insensitive classifiers into a cost-sensitive classifier. They further classified the wrapper method into sampling and thresholding categories. Sampling modifies the class distribution of training data based on misclassification costs and then applies

cost-insensitive classifiers to the sampled data directly. Weighting can also be viewed as a sampling method. It assigns a normalized weight to each instance according to the misclassification costs. Examples of a rare class with a higher misclassification cost are assigned proportionally higher weights. This can be viewed as example duplication that increases the sample sizes of the rare class.

To explain thresholding, I briefly introduce a theory of cost-sensitive learning using a simple binary example. However, this example can be easily extended to multiple classes. In a binary classification, classification costs can be represented in a cost matrix where two types of correct classification, true positives (TP) and true negatives (TN), and two types of error, false positives (FP) and false negatives (FN), have different costs and benefits. As shown in Table 3.5, C(i, j) represents the cost of classifying an instance belonging to class j into class i. Cost(0,0) and Cost(1,1) are usually considered benefits, while the other two cases are costs.

Table 3.5: Two-class cost matrix (2x2)

	Predicted Misstated	Predicted Not Misstated
Actual Misstated (= 0)	C(0,0) or TN	C(1,0) or FN
Actual Not Misstated (= 1)	C(0,1) or FP	C(1,1) or TP

Given the cost matrix, an instance should be classified into a class that generates the minimum expected cost. The expected cost (or conditional risk) of classifying an instance x into class i, R(i/x), can be expressed as

$$R(i|x) = \sum_j \mathrm{P(j|x)} * C(i,j) \qquad (3.4)$$

where P(j|x) is the probability of class j being the actual class of instance x (Kim et al., 2012). The classifier will then classify instance x into the "not misstated" class (where class 1 is not misstated) if

$$P(0|X)C(1,0) + P(0|X)C(1,0) \leq P(0|X)C(1,0) + P(0|X)C(1,0) \qquad (3.5)$$

If I rearrange the Equation 3.5:

$$P(0|X)(C(1,0) - C(0,0) \leq P(0|X)(C(0,1) - C(0,0)) \qquad (3.6)$$

If I assume C(0,0)=C(1,1), and because P(0|X)=1−P(1|X), I can determine the threshold p* for the classifier to classify an instance x as non−misstated if P(1|X) ≥ P*, where

$$P^* = \frac{c(1,0)}{C(1,0)+C(0,1)} = \frac{FP}{FP+FN} \qquad (3.7)$$

If a cost−insensitive classifier can produce a posterior probability estimation p(1|x) for test example x, I can make it cost−sensitive by simply choosing the classification threshold using Equation 3.7 and classify any example to be non−misstated whenever P(1|x) ≥ P*. Thresholding uses Equation 3.7 as a threshold to classify examples into positive or negative if the cost−insensitive classifiers can produce probability estimations.

The cost−sensitive learning method I employ in this paper is a wrapper method known as MetaCost (Domingos, 1999). MetaCost is

a thresholding method. It first uses bagging on the base classifier to obtain reliable probability estimations of training examples and relabels the classes of training examples according to Equation 3.7 (Ling and Sheng, 2010). The advantages of using MetaCost are that it is applicable to multi-class classification problems and can even be used with classifiers that do not produce class probabilities directly. In brief, MetaCost works by (1) sampling multiple bootstrap replicates of the training set, (2) learning the classifier of each set, (3) producing the average of the class probabilities that are directly yielded by the classifiers or calculating the fraction of votes received from the ensemble, (4) using expected costs to relabel each training instance, and (5) reapplying the classifier to the relabeled training set. Although class probabilities from the base classifier are not required for MetaCost to work, the use of base classifier class probabilities appears to improve the result.

3.4 Results

3.4.1 Feature importance

To assess the impact of features in the classification process, I list features selected by multinomial logistic regressions to compute logistic regression functions when estimating the probability of each class. Following Abbasi et al. (2012), I report the features for the multinomial logistic regression model in Table 3.4. Under each class

in Table 3.4, the two columns present a description of the features and the coefficient values. In the description columns, the letters R, T, C, Q, and P represent the ratio, top−5 industry model, closest−5 industry model, quarter, and previous year, respectively, while the numbers indicate the ratio or quarter. For example, under the intentional−misstatement class, the first feature, R2, signifies the second feature, Inventory Growth in the feature table in section 3.2. The ratio number corresponds to the number in the feature table. The second example is the sixth feature under the intentional−misstatements class, R5Q1−R5Q4. R5 signifies the fifth feature, the ratio of property, plants, and equipment to the total assets, and Q1 signifies the first quarter. Thus, R5Q1−R5Q4 signifies the difference between the ratio in the first quarter and the same ratio in the fourth quarter of the previous year.

As shown in Table 3.4, numerous variables related to accruals quality such as changes in inventory (R2) and receivables (R3) are selected along with industry− level and organizational context−based measures. The table provides insight into how the context−based measures supplement the financial ratios, resulting in enhanced financial fraud−detection capabilities.

One of the features selected for the intentional−misstatement class is the firm−efficiency rank (R25), developed by Demerjian et al. (2012a). The firm−efficiency measure attempts to estimate the revenue−generating ability of a firm with a given level of resources,

such as fixed assets and R&D expenses, by means of a data envelopment analysis, a nonparametric method for the estimation of production frontiers to measure empirically the productive efficiency of decision-making units (Baik et al., 2013). The feature has a positive coefficient of 0.46, showing that intentionally misstating firms tend to show higher income-generating abilities given their resources, though not according to the efficient use of their resources but by accounting manipulations. As shown in Figure 3.2, while the no misstatement class is evenly distributed in terms of the proportion of firms in each decile (represented by a line graph), a higher proportion of the intentional-misstatement class is found in the top deciles (0.8 to 1) compared to the unintentional classes

Table 3.6: Features and coefficients by multinomial logistic regression model

No	Class: Intentional Misstatements		Class: Unintentional Misstatements		Class: No Misstatements	
	Descr.	Coeff.	Descr.	Coeff.	Descr.	Coeff.
1	R2	3.45	R3-P3	2.69	R2	-4.96
2	R3-P3	2.9	R2Q4-R2Q3	-1.5	R3-P3	-4.85
3	R4	2.21	R4	0.97	R4	-2.75
4	R36	2.02	R36	0.67	R3Q3-R3Q2	2.06
5	R45	2.02	R3Q4	-0.6	R23-P23	1.55
6	R5Q1-R5Q4	1.85	R28Q3-R28Q2	-0.39	R4-P4	-1.12
7	R4Q2	0.7	R11Q1	0.28	R7	-0.78
8	R3	0.64	R46-P46	0.23	R36	-0.74
9	R37	0.48	R49	0.12	R19	0.59
10	R30Q3/R30Q2	-0.47	R47Q3-T47Q3	0.11	R46	-0.45
11	R25	0.46	R5Q2/R5Q1	0.11	R4Q3	-0.33
12	R46Q1-R46Q4	-0.34	R16Q2-C16Q2	0.11	R4-T4	-0.33
13	R46Q3	0.29	R13Q3/C13Q3	0.1	R46-T46	-0.2
14	R38	-0.18	R47Q1/R47Q4	0.06	R9	0.18
15	R16Q2-C16Q2	-0.15	R7-P7	0.05	R11-C11	0.17
16	R49	-0.07	R7-T7	0.03	R19-T19	0.17
17	R17-C17	0.07	R20Q4-R20Q3	0.01	R5/C5	0.1
18	R13Q2/R13Q1	0.02	R47Q4/R47Q3	0.01	R5/P5	-0.07
19	R13-P13	-0.01	R14Q4-T14Q4	0.01	R32/P32	-0.04
20			R22-C22	-0.01	R30/P30	0.03
21					R31-P31	0.02

Another interesting selected feature is the short interest ratio (SIR; R45). Short interest refers to the total number of shares of a particular stock that have been sold short by investors but have not yet been covered or closed out. This indicator is used by both fundamental and technical traders to identify the prevailing sentiment held by the market for a specific stock. The coefficient of the SIR for intentional misstated firms is 2.02, suggesting that the higher the SIR is, the more likely the firm is intentionally misstating. This shows that negative sentiment in the financial market may signal the presence of possible conditions of accounting manipulation in the firm.

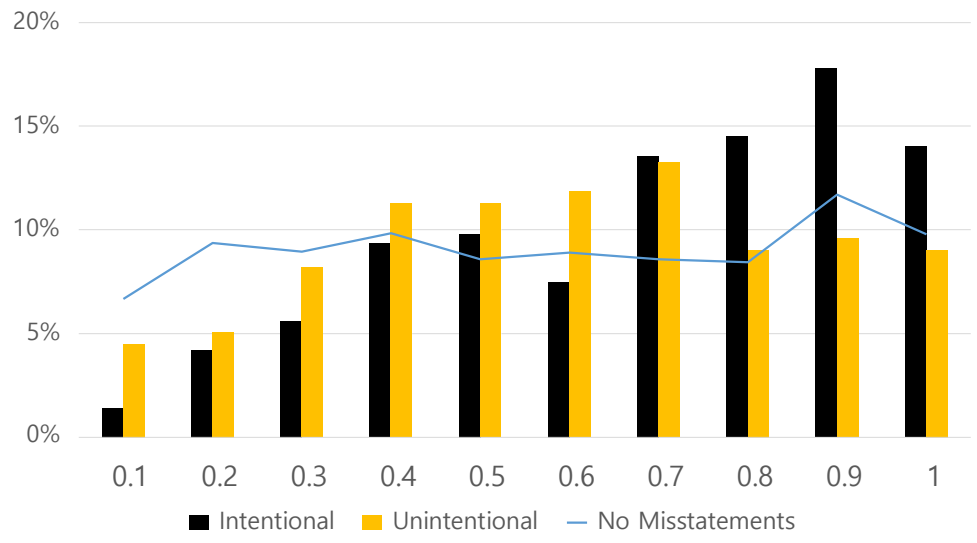

Figure 3.2: Firm efficiency deciles for each class

3.4.2 Classification result

I undertake classification with three multi-class classifiers,

multinomial logistic regression, support vector machine (with a linear kernel), and Bayesian networks, using stratified ten-fold cross validation. As a sanity check, I test how the classifiers classify Enron Corporation, a notorious example of financial fraud. All three models classify the instance correctly, assigning highest class probability levels to the intentional-misstatement class, as shown in Table 3.7.

Table 3.7: Class probabilities for Enron Inc.

	Intentional	Unintentional	No Misstatement
MLogit	0.73	0.13	0.14
SVM-Lin	0.66	0.15	0.20
BayesNet	0.92	0.03	0.05

To assess the classification performance, I select evaluation measures from prior studies. For example, Sun et al. (2006) use accuracy and the G-mean to compare the performance of cost-sensitive boosting algorithms in multi-class classification problems with imbalanced class distribution. On the other hand, Zhou and Liu (2006) use the total misclassification costs to evaluate the performance of sampling and threshold-moving methods in training cost-sensitive neural networks for both binary and multi-class classification problems. Since each measure used in the past studies has advantages and disadvantages, I use these three measures as a more balanced set of measures of the classification performance. Moreover, since the first objective in building misstatement detection models is to detect material misstatements, I use a measure that shows how well each model detects material misstatements, whether

intentional or unintentional, as my fourth measure.

I describe my four evaluation measures. The first measure is accuracy, which is the ratio of the number of instances correctly classified to the total number of instances. The second measure is the G-mean which is calculated as follows,

$$G-mean=\left(\prod_{i=1}^{k} R_i\right)^{1/k} \tag{3.8}$$

where R_i is a recall value for class i. As each recall value representing the classification performance of a specific class, G-mean measures the balanced performance among multiple classes of a classification output (Sun et al., 2006). However, this measure does not reflect non-symmetric misclassification costs.

In order to address the issues of non-symmetric misclassification costs in my performance evaluation, I use two other measures. The third measure is the percentage of misstatements detected. Because it would be more costly to misclassify an intentional misstatement as a non-misstatement than to misclassify an intentional misstatement as an unintentional misstatement, I compare what portion of intentional and unintentional misstatements are not classified as non-misstatements by each classifier.

Because in a multi-class task, a performance measure such as the cost curve, which directly incorporates misclassification costs, is not available (Drummond and Holte, 2000), following Zhou and Liu (2006), I create three types of cost matrices and calculate the total

misclassification costs as the fourth measure of my performance evaluation. Given that actual misclassification costs are different for different users, I select three cost sets within the cost range of previous studies on two-class classification such as Beneish (1999a) (see Table 3.8).

Table 3.8: Three cost matrices

		Predicted								
		Cost Matrix 1			Cost Matrix 2			Cost Matrix 3		
		I	U	N	I	U	N	I	U	N
Actual	Intentional (I)	0	10	50	0	10	70	0	20	100
	Unintentional (U)	10	0	30	10	0	20	10	0	50
	No Misstated (N)	5	5	0	10	10	0	10	10	0

The classification result is presented in Table 3.9. For ease of comparison, costs are normalized by costs produced by MLogit. Multinomial logistic regression and support vector machine show comparable results. I also undertake a cost-sensitive learning using the cost matrices in Table 3.8. The result is shown in Table 3.10, where the first three columns (accuracy, G-mean and % Misstatements detected) show the maximum value under the three different cost scenarios.

Table 3.9: Classification result

Classifier	Accuracy	G-mean	% Misstatements detected	Cost 1	Cost 2	Cost 3
MLogit	0.884	0.605	75.2%	100.00	100.00	100.00
SVM-Lin	0.877	0.580	75.4%	100.22	98.10	99.83
BayesNet	0.821	0.550	68.7%	130.02	136.90	129.17

Table 3.10: Classification result using MetaCost

Classifier	Accuracy	G-mean	% Misstatements detected	Cost 1	Cost 2	Cost 3
MLogit	0.869	0.698	92%	69.3	85.9	66.1
SVM-Lin	0.854	0.656	90%	73.8	89.9	70.8
BayesNet	0.825	0.584	76%	113.7	94.8	120.5

With the MetaCost procedure, I document that G-mean measures and misstatement detection rates improved significantly. As shown earlier, multinomial logistic regression and support vector machine show fairly comparable results when compared under each cost scenario.

The G-mean values are not very satisfactory because it is difficult to classify unintentional-misstatement instances correctly. As shown in Table 3.9, the unintentional-misstatement class consists of smaller firms in terms of sales and total assets. They are less leveraged (less debt and thus a safer investment). Yet, in terms of accruals, a measure of earnings quality, the unintentional class shows higher accrual values than non-misstated firms but lower values than intentional-misstatement firms. However, what actually causes low G-mean score is variability. The unintentional group shows a much higher standard deviation for many features (not all variables reported here), making classification much more challenging. This suggests that it is necessary to break down the unintentional class further into a number of separate classes with similar characteristics, such as misstatement causes, to improve the classification results.

Table 3.11: Feature values by class

Class	Sales	Assets	Leverage		Accruals	
	Median	Median	Median	Std. Dev.	Median	Std. Dev.
Intentional	247.86	352.18	0.31	0.68	0.06	0.33
Unintentional	206.66	209.95	0.21	**1.99**	0.04	**3.53**
No Misstated	261.36	257.46	0.23	0.46	0.00	0.38

Another possible explanation for low G-mean score is that my target variable is not a perfect measure of managerial intent. As Hennes et al. (2008) noted in their study, management intent is impossible to observe in actuality. In this study, I assume that the target variable was correctly labeled but, due to the inability to measure human intention perfectly, some instances may have been mislabeled, causing less accurate detection rates.

Moreover, it is noteworthy that a multiclass classification problem is intrinsically more complex than a binary problem, since the generated classifier must be able to separate the data into a higher number of categories, which increases the chances of classification errors. As a result, its complexity increases for more classes (Liu et al., 2008; Lorena et al., 2008). Especially when the data are of high dimensionality and the sample size is small, the classification accuracy degrades very rapidly as the number of class increases (Li et al., 2004). My tertiary models are theoretically better construct than the binary models employed by the past studies but G-means (accuracy per class) may be lower compared to the binary models.

Chapter 4

Building financial misstatement detection models using features generated from CFO survey contents

To build a financial misstatement detection model, modelers need a proper target variable. Three types of database have been used extensively for building accounting misstatement detection models (Dechow et al., 2011; Price et al., 2011). The first type of data source is restatement databases such as the Government Accountability Office (GAO) financial statement restatement database. The second type of source is shareholder lawsuit database such as Stanford Law Database on Shareholder Lawsuits. The third type of source is the SEC enforcement samples. The SEC issues Accounting and Auditing Enforcement Releases (AAERs) during or at the conclusion of an investigation against a company, an auditor, or an officer for alleged accounting and/or auditing misconduct (Dechow et al., 2011).

Dechow et al. (2011) list two advantages of using AAERs relative to other types of samples. The first advantage is that the use of AAERs as a proxy for manipulation is a straightforward. The governmental body has investigated and confirmed the accounting misconduct. The

second is that AAERs are also likely to capture a group of economically significant manipulations as the SEC has limited resources and likely pursues the most significant cases. However, as they further point out, the SEC's such investigation practice may introduce selection bias.

Recently, researchers have found that political and physical factors are at play worsening the SEC selection bias. For example, Correia (2014) shows that politically connectedness can influences the likelihood of receiving an enforcement action from the SEC. On the other hand, Kedia & Rajgopal (2011) show that the SEC is more likely to investigate firms located closer to its offices.

Due to this selection bias, Dechow et al. (2011) point out that predicting "the AAER firms" is predicting the likelihood that a firm is engaging in an accounting misstatement and receiving an enforcement action from the SEC at the same time. However, researchers cannot avoid the selection bias issue by simply switching to other types of misstatement databases. This is due to the fact that selection bias exists in other types of database and it is a general concern when analyzing the determinants of earnings manipulation (Dechow et al. 2010).

This bias concern is especially relevant to datamining researchers since recent high performing machine learning based detection models use hundreds or even thousands of variables to achieve high classification accuracy.

Table 4.1 Financial statement fraud/misstatement detection models

	Study	Feature Set	Method	Test Data Set	Results	% Minority	AUC
1	Beneish (1999)	8 financial measures	Probit regression	2,406 firm-years: 74 fraud(F), 2,332 non-fraud(NF)	Overall: 89.5%; Fraud: 54.2%	0.3%	0.49*
2	Dechow et al. (2011)	7 financial measures	Logistic regression	79,651 firm-years: 293F/79,358NF	Overall: 63.7%; Fraud: 68.6%	0.4%	0.76*
3	Cecchini et al. (2010)	24 financial ⇒ 1,518 ratios	SVM with custom financial kernel	3,319 firm-years: 132F/3,187NF	Overall: 90.4%; Fraud: 80.0%	4%	0.88*
4	Abbasi et al. (2011)	12 financial ratios ⇒ 420 ratios	Stacking, Semi-supervised	3,862 firm-years: 406F/ 3,456NF	Fraud: 88%	10%	0.93**

Sources: * Cecchini et al. (2010) and ** Abbasi et al. (2012)

As shown in Table 4.1, machine learning models (3 and 4) are superior to two highly cited regression-based models from accounting and Finance literature (1 and 2) in terms of prediction accuracy. For example, the model by Cecchini et al. (2010) achieves an AUC of 0.88 whereas the F-score model by Dechow et al. (2011) achieves an AUC of 0.76. However, while Dechow et al. (2011) use a logistic regression with only 7 explanatory variables, Cecchini et al. (2010) utilizes a custom financial kernel that produces 1,518 variables. Thus, it is very difficult to interpret the result in Cecchini et al. (2010) and it is also difficult to assess if their superior result is due in part to capturing the effect of selection bias in the target variable.

One way of building financial misstatement detection models while minimizing the selection bias risk in target variable is to build interpretable models with domain expert guided feature selection process. Especially, proper feature selection process is critical to

minimize the effect of selection bias. For example, as shown in Table 4.2, misstated firms (AAERs) are much larger than average (or medium) size firms in Compustat. This may be due to the fact that larger firms tend to have more institutional investors and more analyst covering that their financial reporting and internal controls are more likely to be scrutinized and analyzed. Also, the SEC may be more likely to pursue cases where stock performance declines rapidly after the manipulation is revealed, because the identifiable losses to investors are greater due to the (Dechow et al., 2011). Therefore, the use of features like market capitalization or total asset value in misstatement detection models will improve classification accuracy but it may not improve general misstatement detectability.

Table 4.2 Characteristics of misstatement firms (AAERs) vs. non misstatement firms

	Misstatement years			Non misstatement years			Misstate- Nonmisstae		
							Pairwise Diff.	Pairwise T-test	Wilcoxon test
Variable	N	Median	Mean	N	Median	Mean	in mean	P-value	P-value
Assets - Total ($MM)	770	$534	$8,341	89,826	$103	$1,521	$6,820	<.0001	<.0001
Market Cap. ($MM)	761	$588	$4,282	84,227	$90	$1,196	$3,086	<.0001	<.0001

In this chapter, I develop binary financial misstatement detection models with features generated from a CFO survey. More specifically, my research objectives are, first, to create features to detect material misstatements using earnings management survey to public company

CFOs by Dichev et al. (2016). The second objective is to build detection models using survey-based features and features from related academic literature. The third is to compare the performance of the new models with the existing scoring models from accounting and finance literature.

What differentiates this study from other related studies is that I develop financial material misstatement detection models based on features that practitioners deem important in detecting earnings management. This way the feature selection process is performed by domain experts and this, in turn, may reduce the possible effect of selection bias in the target variable. Also, using new features, I improve prediction accuracy of the existing detection models.

In the following section, I describe overall research methodology. I describe the data used and the testing methodology employed. I present the results in section 4.2.

4.1. Research Methodology

4.1.1 Feature generation

Dichev et al. (2016) conduct a survey with 375 CFO's of private and public companies on earnings manipulation. Based on their survey, they list twenty "red flags" of earnings misrepresentation. I create variables based on the twenty red flags as shown in Table 4.3.

Using these variables I develop two new misstatement detection models. The first model is, what I call, the survey model since I use only the features generated from the survey result. The second model is the combined model that uses variables from the survey, existing models, and related literature. I build the two logistic regression models using SAS enterprise miner 13.1.

4.1.2 Data

I use the data from Compustat between the year 1989 and 2007 (101,042 firm years (840 misstated firm years and 100,202 firm years for non-misstated firms)). For target variable, I use the AAER list provided by the Center for Financial Reporting and Management. I follow the procedure in Dechow et al. (2011) for the target firm year selection. Following Perols (2011), I eliminate financial firms and cross-listed foreign firms using American depositary receipt ratios. I use both annual and quarterly financial statement data to calculate variables.

For performance comparison, I select two highly cited scoring models from accounting and finance literature: the M-score model by Beneish (1999) and the F-score model by Dechow et al. (2011).

Table 4.3 20 Red flags of earnings misrepresentation and generated variables

No	Red Flag	% of Responses	Variables	Definition
1	GAAP earnings do not correlate with cash flow from operations; weak cash flows; earnings and cash flow from operations move in different direction for six to eight quarters; earnings strength with deteriorating cash flow.	34%	Cash Flow Earnings Correlation	Corr(Net Income (NI), cash flow from operation(CFO))
2	Deviations from industry (or economy, peers') norms/experience (cash cycle, volatility, average profitability, revenue growth, audit fees, growth of investments, asset impairment, accounts payable, level of disclosure).	24	Unexplained Production costs/Discretionary expenses	Error terms from industry production cost/ discretionary expense/CFO regression equation
4	Large/frequent one-time or special items (restructuring charges, write-downs, unusual or complex transactions, gains/losses on asset sales).	17	Special Items as a Percentage of Sales	Special Items / Sales
5	Lots of accruals; large charges in accruals; jump in accruals/sudden changes in reserves; insufficient explanation of such changes; significant increase in capitalized expenditures; changes in asset accruals; high accruals liabilities.	15	RSST Accrual	Accrual measure following Richardson et al. (2005)
6	Too smooth/too consistent of an earnings progression (relative to economy, market); earnings and earnings growth are too consistent (irrespective of economic cycle and industry experience); smooth earnings in a volatile industry.	14	Earnings smoothness; Sales/Asset volatility	$\sigma(NI)/\sigma(CFO)$; $\sigma(Sales_t/AT_{t-1})$
7	(Frequent) changes in (significant) accounting policies.	10	Accounting Change effect	Accounting Changes - Cumulative Effect divided by AT
9	High executive turnover; sudden change in top management; change in financial management; sudden director turnover; employee (non-management) turnover.	8	Abnormal change in employees	% Δ(Number of employees) – % Δ(Total Assets(AT))

Table 4.3 (cont' d)

No	Red Flag	% of Responses	Variables	Definition
10	Inventory buildup/age of raw materials; buildup in work in progress; mismatch between inventory/cost of goods sold/reserves.	7	1. Change in inventory; 2. Inventory over Cost of good sold(COGS)	1. Δ(Inventory)/ Avg. AT; 2. Inventory/COGS
11	Large volatility (wide swings) in earnings, especially without real change in business.	7	Unexpected employee productivity	% Δ(SALE/Employees)
12	Buildups of receivables; deterioration of receivables days outstanding; accounts receivable balance inconsistent with cash cycle projections/allowance for doubtful accounts.	5	1. Change in A/R; 2. Allowance for doubtful accounts	1. % Δ(Accounts receivables/AVG. AT); 2. Allowance for doubtful accounts/Sale
15	Major jumps or turnarounds; break with historical performance; unexplained volatility in margins.	15	Operating performance margin	NI/SALE
18	Accruals/assets/working capital growing faster or slower than revenue.	3	Change in Asset turnover	Δ(Sales/AT)
19	Increased debt/high liabilities.	3	Leverage	Long-Term Debt /AT
20	Weak sales growth vs. industry/declining performance (e.g., ROA or weakened cash flows, current ratio, working capital).	3	1. Sales growth; 2. Change in return on assets	1. $Sale_t$/ $Sale_{t-1}$; 2. Δ(Earnings/ Avg. AT)

4.2. Result

4.2.1 Models

I present the four models (two new models: the survey model and the combined model in in Tables 4.4 and 4.5, and two existing models: the M-score model and the F-score model in Tables 4.6 and 4.7). Histograms of model scores are in figures 4.1, 4.2, 4.3, and 4.4. I recalibrate the M-score model and the F-score models. However, when I recalibrate the M-score model and the F-score models, some of the variables become statistically insignificant (in terms of p value). I drop those insignificant variables when I compare prediction accuracy.

Table 4.4 The M-score model

Definition	Parameter	Estimate	p-value
	Intercept	-5.4701	<.0001
Asset quality index	AQI	0.0179	0.4921
Depreciation index	DEPI	0.0387	0.6565
Days' sales in receivables index	DSRI	0.0511	0.3032
Gross margin index	GMI	0.0401	0.3162
Leverage index	LVGI	-0.0191	0.7299
Sales, general, and administrative expenses index	SGAI	0.2975	0.0073
Sales growth index	SGI	0.1897	<.0001
Total accruals to total assets	TATA	-0.2622	0.0437

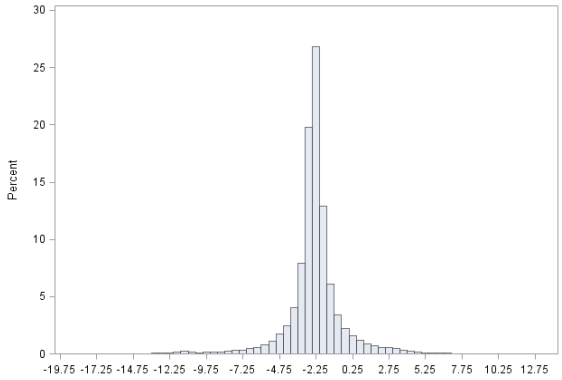

Figure 4.1 Histogram of M-score

Table 4.5 The F－score model

Definition	Parameter	Estimate	p-value
	Intercept	-6.5216	<.0001
Change in cash sales	CH_CS	0.0318	0.3856
Change in A/R	CH_INV	1.2359	0.0532
Change in Receivables	CH_REC	2.1555	<.0001
Change in return on assets	CH_ROA	-0.4758	0.0012
An indicator variable coded 1 if the issuance firm issued securities during year t	ISSUE	-0.6145	<.0001
RSST accruals	RSST_ACC	0.3973	0.0003
% Soft Asset	SOFT_ASSTS	1.7398	<.0001

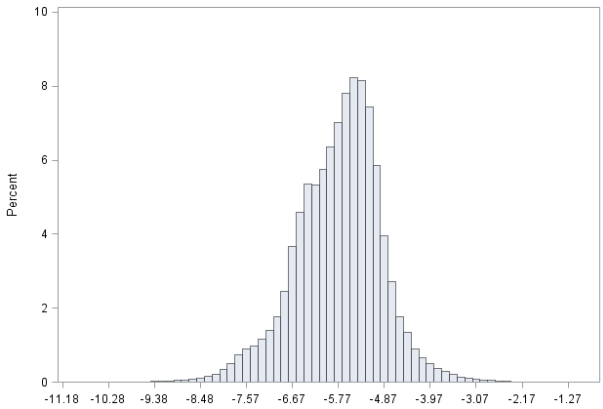

Figure 4.2 Histogram of F-score

Table 4.6 The survey model

Definition	Parameter	Estimate	p-value
	Intercept	-5.0934	<.0001
2. Error term from industry CFO regression equation	CFO3	-0.047	0.0406
10. Change in inventory	CH_INV	1.9684	0.012
12. Change in A/R	CH_REC	2.6625	<.0001
20. Change in return on assets	CH_ROA	-0.4517	0.0302
5. RSST Accrual	RSST_ACC	0.5498	0.0003
20. Sales growth	SGI	0.188	0.0174
6. Sales/Asset volatility	$\sigma(Sales_t/AT_{t-1})$	-1.0397	0.0743
18. Change in Asset turnover	a_at_sub	-0.238	0.101
15. Operating performance margin	a_opm	0.2495	0.0002
12. Allowance for doubtful accounts	p_allow_sale	8.7594	<.0001
11. Unexpected employee productivity	p_unexp_prod	-0.1955	0.0629
4. Special Items as a Percentage of Sales	si	-1.2026	<.0001

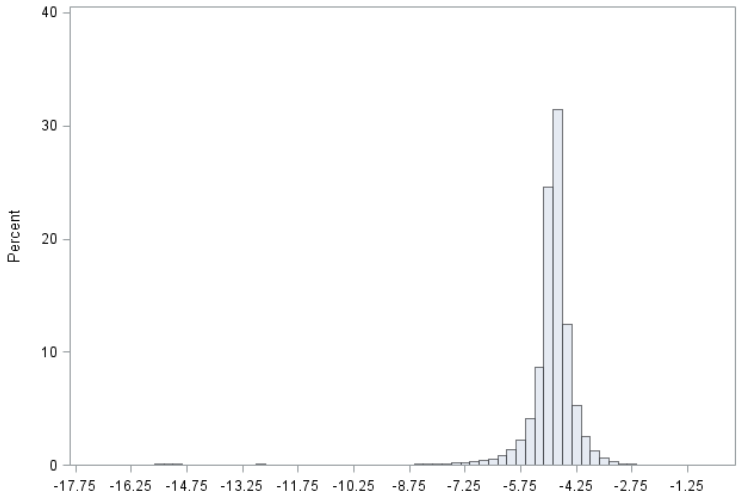

Figure 4.3 Histogram of Survey Model score

Table 4.7 The Combined model

Definition	Parameter	Estimate	p-value
	Intercept	-7.56	<.0001
Change in cash sales (F-score)	CH_CS	0.1083	0.0476
12. Change in A/R	CH_REC	1.649	0.0037
Operational Efficiency measure (in decile)	FIRM_EFFICIENCY_2013_RANK		<.0001
An indicator variable coded 1 if the issuance firm issued securities during year t (F-score)	ISSUE	0.458	0.0002
An indicator variable coded 1 if future operating lease obligations are greater than zero (F-score)	LEASEDum	0.4411	0.0001
2. Error term from industry production cost regression equation	PROD3	0.0686	0.0627
Sales, general, and administrative expenses index (M-score)	SGAI	0.446	0.0041
% Soft Asset (F-score)	SOFT_ASSTS	1.6278	<.0001
18. Change in Asset turnover	a_at_sub	-0.287	0.0377
15. Operating performance margin	a_opm	0.136	0.0116
Number of quarters of EPS increase (over 2 years)	eps_up_ct		<.0001
Account Payable/operating expenses	g_ap_opexp	1.198	0.0068
12. Allowance for doubtful accounts	p_allow_sale	9.1012	0.0001
4. Special Items as a Percentage of Sales	si	-1.1053	<.0001
Short Interest ratio	sir	5.7581	<.0001

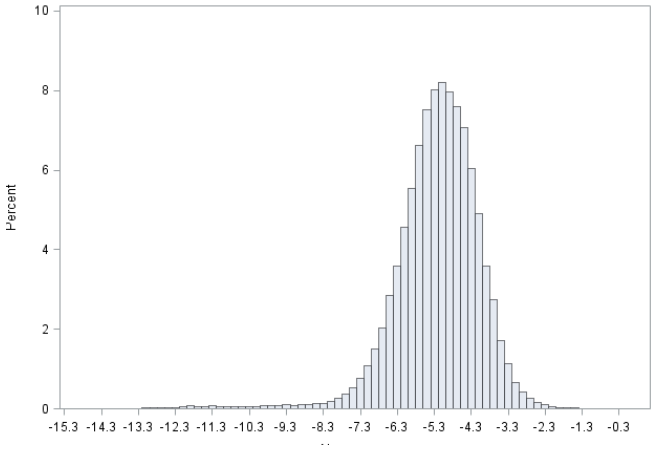

Figure 4.4 Histogram of Combined score

I calculate the Pearson correlations and Spearman correlations among four scores. Table 4.8 reports the correlation matrix for the four scores (Spearman correlations are above the diagonal and Pearson correlations are below the diagonal). Due to the fact that some

variables are used in different models, some scores show higher correlation than others.

Table 4.8 Correlation Matrix

Correlation Coefficients: Prob >|r| under H_0:ρ=0

	Combined Model Score	Survey Model Score	M-SCORE	F-SCORE
Combined Model Score	1	0.48476 <.0001	0.14098 <.0001	0.60487 <.0001
Survey Model Score	0.6699 <.0001	1	0.3601 <.0001	0.49675 <.0001
M-SCORE	0.11254 <.0001	0.18991 <.0001	1	0.3031 <.0001
F-SCORE	0.53138 <.0001	0.29469 <.0001	0.2891 <.0001	1

Note: This table reports Spearman (above diagonal) and Pearson (below diagonal) correlations for sample variables.

4.2.2 Prediction Accuracy

I compare the prediction accuracy of the four models using the receiver operating characteristic (ROC) curve. As shown in Figure 4.5, the combined model achieves the highest classification accuracy, while the M-score model performs the worst. The survey model achieves comparable accuracy with the F-score model. This result show that the CFO survey result can help researchers better detect the material misstatements. When I combine the survey content based features with the features from related literature, the best result is achieved.

4.2.3 Equity return predictability

To test if these scores can be used as a financial risk management tool,

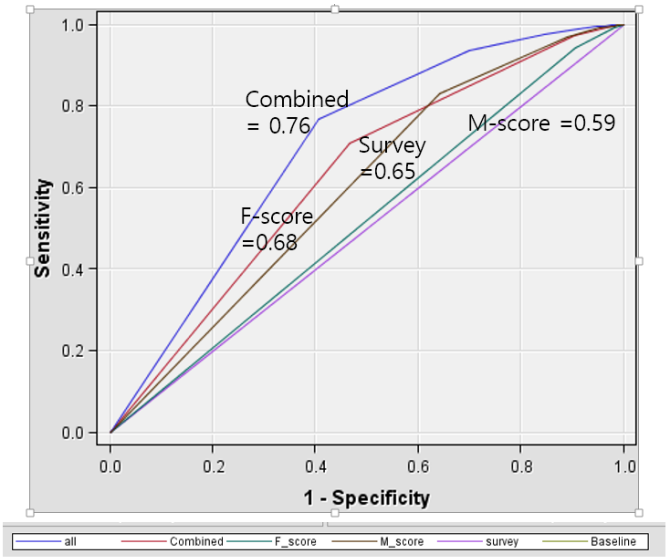

Figure 4.5 ROC Curves for the four model scores

I also test how well each model score predicts significant negative stock returns one-year ahead. I create a binary target variable where the variable (called extreme event) equals one if a firm experiences stock return of negative 70 per cent or more in a fiscal year. As shown in Table 4.9, negative stock returns of 70 percent of more have occurred,

Table 4.9 %Stock returns of negative 70% or more in a year

Years	% (70%+ stock return drops)
1989~1994	3.5%
1995-1999	4.3%
2000~2005	8.5%
2006~2010	6.4%
2011~2012	3.6%
24-year avg.	5.34%

on average, about 5.34% of the time between 1989 and 2012 in my data sample of 76,107 firm years between 1989 and 2012.

Then I test the relationship between current model scores and extreme events a year ahead (as a target variable). Using chi-square attribute ranking, I rank the model scores as a predictor as shown in Table 4.10.

Table 4.10 Chi-square attribute ranking for model scores

Chi-squared ranking	Scores
1008.05	Survey Model Score
683.86	M-SCORE
370.08	F-SCORE
231.70	Combined Model Score

The survey model score has the highest score meaning that it is the best attribute in predicting the extreme events of negative stock return of 70 per cent or more one year ahead. This also show that the survey content based features may be useful in predicting huge negative future stock returns.

Chapter 5

Conclusion

Financial statement fraud has substantial economic consequences for any economy. Management that intentionally misleads users of their financial statements will invoke considerable financial and social costs. Due to the significance of this topic, academics have proposed various prediction and detection models in the accounting and data mining literature. In this thesis, I attempt to resolve two challenges in financial misstatement detection modelling. The first challenge is that a target variable can be classified as involving either errors (i.e., unintentional misapplications of accounting rules) or irregularities (i.e., intentional misreporting). When testing hypotheses involving managerial misconduct, researchers could make incorrect inferences regarding their hypotheses if they do not specifically distinguish intentional misstatements from unintentional errors. If researchers limit their samples to only fraudulent misstatements for their detection models, they are underutilizing information by throwing away more commonly occurring unintentional misstatements. As a result, their models may not effectively detect more frequent but less egregious misstatements or discriminate between intentional misstatements and unintentional errors.

To deal with the shortcomings of the existing binary detection

models, I develop multi-class detection models. Using the post-event analysis in Hennes et al. (2008), I develop three-class financial misstatement detection models. The models are developed to detect financial misstatements and classify misstatements according to fraud intention. I also apply multi-class cost-sensitive learning using MetaCost to deal with class imbalances and asymmetric misclassification costs.

I find that variables related to accruals quality, such as changes in inventory along with industry-level and organizational context-based measures, have discriminatory power. The firm-efficiency measure and market variables such as the short interest ratio are also found to be useful to detect misstatements and deliberate fraud.

I acknowledge that management intention is not observable; thus, my target variable may not be a perfect measure of managerial intent. Also, building multi-class detection models is more difficult than building binary class models because of higher complexity in the definition of the decision boundaries associated with larger number of classes. Nonetheless, understanding and detecting fraud intention is a crucial step toward preventing misstatements effectively. Also, my approach is one way to resolve the problem of fraud intention in existing binary detection models. Moreover, the contributions of this study go further than filling a void in the literature by developing the first multi-class predictive models alone. This study provides a quantitative tool to detect fraud intention of senior management of public firms. This should benefit

academics and practitioners in financial regulation and capital markets. More specifically, regulators such as the SEC would benefit from my work because they could focus their investigation efforts on cases that are more likely to involve fraudulent intention. Also, investors and financial institutions would benefit from appropriately adjusting their levels of exposure to suspected firms in advance. Moreover, auditors can tailor their audit processes accordingly and minimize their possible legal risks.

To improve the tertiary models further, a breakdown of unintentional-misstatement firms may be necessary. Future studies can incorporate audit-related variables such as audit fees and internal control variables to improve detectability even further.

The second challenge in financial misstatement detection modelling is selection bias in target variables. Selection bias is an issue that is prevalent in all the target variables but it is also the issue that researchers tend to ignore. One way to reduce the effect of selection bias in target variable is to do feature selection with domain experts. I generate features using the CFO survey by Dichev et al. (2016) as a domain expert-guided feature selection process. I build two new detection models using the features based on the survey contents. The new models show better or comparable classification accuracy when compared with existing scoring models, the M-score model and the F-score model. Also, I show that the survey content-based features have potential for predicting the extreme events of negative stock return of

70 per cent or more one year ahead. A detailed analysis needs to be conducted to better understand the relationship between the likelihood of misstatement and future equity return. I leave this as a prospect for future study.

I present a quick summary of the two main chapters of this thesis in Table 5.1. First, to deal with the issue of fraud intention in target variables, I build tertiary models using the GAO restatement database. The main contribution of my approach is that my tertiary models are the first predictive models that classify misstatements according to fraud intention. Second, to reduce the possible effect of selection bias in target variable, I do feature selection and generation based on the CFO survey result compiled and analyzed in Dichev et al. (2016, 2013). Using the survey contents, I generate features and build misstatement detection models. Then I compare the performance of the new models against the existing scoring models in accounting and finance literature. The main contribution of my approach is that it is a simple but effective way to utilize practitioners' expertise in feature selection and feature generation to better detect the material misstatements. More importantly, my analysis should provide useful insights and quantitative tools to auditors, regulators, investors, and other financial statement users to better identify misstating firms. The efficient functioning of capital markets depends crucially on the quality of the financial information provided to capital market participants. Curtailing misstatement activity should lead to improved financial information and

hence improved returns for investors and more efficient allocation of capital (Dechow et al., 2011).

Table 5.1 Summary of the thesis

	Chapter 3	**Chapter 4**
Model	A vs. B. vs. D Three-class model	C vs D Two-class scoring model
Target Variable	Restatements	AAER's
Challenges	Presence of fraud intention in target variable	Selection bias in target variable
Contributions	Detect fraud Intention using tertiary classification models	Utilize practitioners' expertise in feature selection process

Bibliography

Abbasi, A., Albrecht, C., Vance, A., and Hansen, J. (2012). Metafraud: a meta-learning framework for detecting financial fraud. Mis Quarterly, 36(4):1293 – 1327.

Badertscher, B. A. (2011). Overvaluation and the choice of alternative earnings management mechanisms. The Accounting Review, 86(5):1491 – 1518.

Baik, B., Chae, J., Choi, S., and Farber, D. B. (2013). Changes in operational efficiency and firm performance: a frontier analysis approach. Contemporary Accounting Research, 30(3):996 – 1026.

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

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

Beneish, M. D. (1999b). Incentives and penalties related to earnings overstatements that violate gaap. The Accounting Review, 74(4):425 – 457.

Bermejo, P., Gámez, J. A., and Puerta, J. M. (2011). Improving the performance of naive bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets. Expert Systems with Applications, 38(3):2072 – 2080.

Bolton, R. J. and Hand, D. J. (2002). Statistical fraud detection: A review. Statistical science, pages 235 – 249.

Bourke, C., Deng, K., Scott, S. D., Schapire, R. E., and Vinodchandran, N. (2008). On reoptimizing multi-class classifiers. Machine Learning, 71(2- 3):219 – 242.

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

Correia, M. M. (2014). Political connections and sec enforcement. Journal of Accounting and Economics, 57(2):241 – 262.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., and Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. Expert systems with applications, 41(10):4915 – 4928.

Dechow, P., Ge, W., and Schrand, C. (2010). Understanding earnings quality: A review of the proxies, their determinants and their consequences. Journal of Accounting and Economics, 50(2):344 – 401.

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

Dechow, P. M., Sloan, R. G., and Sweeney, A. P. (1995). Detecting earnings management. Accounting review, pages 193 – 225.

Dechow, P. M., Sloan, R. G., and 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.

DeFond, M. L. and Jiambalvo, J. (1994). Debt covenant violation and manipulation of accruals. Journal of accounting and economics, 17(1):145 – 176.

Demerjian, P., Lev, B., and McVay, S. (2012a). Quantifying managerial ability: A new measure and validity tests. Management Science, 58(7):1229 – 1248.

Demerjian, P. R., Lev, B., Lewis, M. F., and McVay, S. E. (2012b). Managerial ability and earnings quality. The Accounting Review, 88(2):463 – 498.

Dichev, I., Graham, J., Harvey, C. R., and Rajgopal, S. (2016). The misrepresentation of earnings. Financial Analysts Journal, 72(1):22 – 35.

Dichev, I. D., Graham, J. R., Harvey, C. R., and Rajgopal, S. (2013). Earnings quality: Evidence from the field. Journal of Accounting and Economics, 56(2):1 – 33.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 155 – 164. ACM.

Drummond, C. and Holte, R. C. (2000). Explicitly representing expected cost: An alternative to roc representation. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 198 – 207. ACM.

Erickson, M., Hanlon, M., and Maydew, E. L. (2006). Is there a link between executive equity incentives and accounting fraud? Journal of Accounting Research, pages 113 – 143.

Ettredge, M., Scholz, S., Smith, K. R., and Sun, L. (2010). How do restatements begin? Evidence of earnings management preceding restated financial reports. Journal of Business Finance & Accounting, 37(3-4):332 – 355.

Fawcett, T. and Provost, F. (1997). Adaptive fraud detection. Data Mining and Knowledge Discovery, 1(3):291 – 316.

GAO (2002). Financial statement restatements: Trends, market impacts, regulatory responses, and remaining challenges. US General Accounting Office.

Gillett, P. R. and Uddin, N. (2005). CFO intentions of fraudulent financial reporting. Auditing: A Journal of Practice & Theory, 24(1):55 – 75.

Green, B. P. and Choi, J. H. (1997). Assessing the risk of management fraud through neural network technology. Auditing, 16(1):14.

Han, J., Kamber, M., and Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques. Elsevier.

Hastie, T., Tibshirani, R., et al. (1998). Classification by pairwise coupling. The annals of statistics, 26(2):451 – 471.

Hayes, L. (2014). Identifying unintentional error in restatement disclosures. Available at SSRN 2269086.

Hennes, K. M., Leone, A. J., and Miller, B. P. (2008). The importance of distinguishing errors from irregularities in restatement research: The case of restatements and ceo/cfo turnover. The Accounting Review, 83(6):1487 – 1519.

Hennes, K. M., Leone, A. J., and Miller, B. P. (2013). Determinants and market consequences of auditor dismissals after accounting restatements. The Accounting Review, 89(3):1051 – 1082.

Hoitash, R., Hoitash, U., and Johnstone, K. M. (2012). Internal control material weaknesses and cfo compensation. Contemporary Accounting Research, 29(3):768 – 803.

Huang, S.-Y., Tsaih, R.-H., and Yu, F. (2014). Topological pattern discovery and feature extraction for fraudulent financial reporting. Expert Systems with Applications, 41(9):4360 – 4372.

Jones, K. L., Krishnan, G. V., and Melendrez, K. D. (2008). Do models of discretionary accruals detect actual cases of fraudulent and restated earnings? an empirical analysis. Contemporary Accounting Research, 25(2):499 – 531.

Kedia, S. and Rajgopal, S. (2011). Do the sec' s enforcement preferences affect corporate misconduct? Journal of Accounting and Economics, 51(3):259 – 278.

Kim, J., Choi, K., Kim, G., and Suh, Y. (2012). Classification cost: An empirical comparison among traditional classifier, cost-sensitive classifier, and metacost. Expert Systems with Applications, 39(4):4013 – 4019.

Kirkos, E., Spathis, C., and Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, 32(4):995 – 1003.

Kotsiantis, S., Koumanakos, E., Tzelepis, D., and Tampakas, V. (2006). Forecasting fraudulent financial statements using data mining. International Journal of Computational Intelligence, 3(2):104 – 110.

Landwehr, N., Hall, M., and Frank, E. (2005). Logistic model trees. Machine Learning, 59(1-2):161 – 205.

Leong, C. K. (2015). Credit risk scoring with bayesian network models. Computational Economics, pages 1 – 24.

Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15):2429 – 2437.

Lin, C.-C., Chiu, A.-A., Huang, S. Y., and Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. Knowledge-Based Systems, 89:459 – 470.

Ling, C. X. and Sheng, V. S. (2010). Cost-sensitive learning. In Encyclopedia of Machine Learning, pages 231 – 235. Springer.

Liu, J., Ranka, S., and Kahveci, T. (2008). Classification and feature selection algorithms for multi-class cgh data. Bioinformatics, 24(13):i86 – i95.

Liu, X.-Y. and Zhou, Z.-H. (2006). The influence of class imbalance on cost-sensitive learning: An empirical study. In Data Mining, 2006. ICDM' 06. Sixth International Conference on, pages 970 – 974. IEEE.

Lorena, A. C., De Carvalho, A. C., and Gama, J. M. (2008). A review on the combination of binary classifiers in multiclass problems. Artificial Intelligence Review, 30(1-4):19 – 37.

Ngai, E., Hu, Y., Wong, Y., Chen, Y., and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3):559 – 569.

Pai, P.-F., Hsu, M.-F., and Wang, M.-C. (2011). A support vector machine-based model for detecting top management fraud. Knowledge-Based Systems, 24(2):314 – 321.

Palmrose, Z.-V., Richardson, V. J., and Scholz, S. (2004). Determinants of market reactions to restatement announcements. Journal of accounting and economics, 37(1):59 – 89.

Palmrose, Z.-V. and Scholz, S. (2004). The circumstances and legal consequences of non-gaap reporting: Evidence from restatements. Contemporary Accounting Research, 21(1):139 – 180.

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.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers, 10(3):61 – 74.

Plumlee, M. and Yohn, T. L. (2010). An analysis of the underlying causes attributed to restatements. Accounting Horizons, 24(1):41 – 64.

Ragothaman, S., Carpenter, J., and Buttars, T. (1995). Using rule induction for knowledge acquisition: An expert systems approach to evaluating material errors and irregularities. Expert systems with Applications, 9(4):483 – 490.

Schrand, C. M. and Zechman, S. L. (2012). Executive overconfidence and the slippery slope to financial misreporting. Journal of Accounting and Economics, 53(1):311 – 329.

Sun, Y., Kamel, M. S., and Wang, Y. (2006). Boosting for learning multiple classes with imbalanced class distribution. In Data Mining, 2006. ICDM' 06. Sixth International Conference on, pages 592 – 602. IEEE.

Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., and Baesens, B. (2015). Apate: A novel approach for automated credit

card transaction fraud detection using network-based extensions. Decision Support Systems, 75:38 – 48.

West, J. and Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57:47 – 66.

Whiting, D. G., Hansen, J. V., McDonald, J. B., Albrecht, C., and Albrecht, W. S. (2012). Machine learning methods for detecting patterns of management fraud. Computational Intelligence, 28(4):505 – 527.

Witten, I. H. and Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Xia, F., Yang, Y.-w., Zhou, L., Li, F., Cai, M., and Zeng, D. D. (2009). A closed-form reduction of multi-class cost-sensitive learning to weighted multi-class learning. Pattern Recognition, 42(7):1572 – 1581.

Zhou, Z.-H. and Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. Knowledge and Data Engineering, IEEE Transactions on, 18(1):63 – 77.

Zhou, Z.-H. and Liu, X.-Y. (2010). On multi-class cost-sensitive learning. Computational Intelligence, 26(3):232 – 257.

국문초록

기계 학습 및 통계 모델을 사용하여 재무재표상의 심각한 오류를 감지하는 여러 개의 예측 모형을 개발한다. 첫째, 고의성 여부 포함하는 다중 클래스 모델을 개발한다. 기존의 이분류 모형의 단점을 보안하고자 Hennes 외(2008)의 고의성의 여부에 따라 정정 공시 데이터를 분류하는 사후 연구 결과를 활용하여 사기 의도성에 의한 정정 공시 기업, 과실로 인한 정정 공시 기업, 정정공시를 하지 않은 기업으로 삼분류하는 예측 모형을 로지스틱 회귀 모델, 서포트 벡터 머신, 베이즈 네트워크 모형을 활용하는 예측 모형을 계발한다. 클래스 불균형과 비대칭 오 분류 비용문제를 처리하기 위해, MetaCost 방법론을 사용하는 다중 클래스 비용 의존 학습 방법을 수행하여 모형의 성능을 향상시킨다.

둘째, Dichev 외(2016)의 연구에서 발표한 재무최고책임자 설문 조사의 결과를 바탕으로 변수를 생성하고 재무제표의 심각한 오류를 감지하는 스코어링 모형을 구축한다. 단순히 통계적 기법이나 알고리즘을 활용한 변수 선택법을 활용한 모형들은 타겟 변수에 존재하는 선택변견에 취약할 수 있다. 따라서 변수 선택 시 회계전문가의 참여가 필수적이다. 본 연구에서는 Dichev 외(2016)의 연구에서 발표한 재무최고책임자들의 설문 조사를 활용하여 변수를 생성하고 변수 선택에 활용하는 방법을 제안하고 이러한 변수들을 활용하여 만든 스코어링 모델과 기존 연구에서 제시된 스코어링 모델과 비교를 통해 새로운 변수와 모델의 성능을 평가한다.

주요어: 분식회계 감지 모형; 사기 의도성; 다중 클래스 비용 의존 학습법;
수익조정; 최고 재무 책임자 설문조사
학번: 2011-30999