

Identifying Informative Audit Quality Indicators (IAQI) Using Machine Learning

Chanyuan (Abigail) Zhang
PhD candidate
Rutgers Business School
abigail.zhang@rutgers.edu

Soohyun Cho
Assistant Professor
Rutgers Business School
scho@business.rutgers.edu

Miklos Vasarhelyi
KPMG Distinguished Professor of AIS
Rutgers Business School
miklosv@business.rutgers.edu

Feb 2021

Acknowledgments: We thank Efrim Boritz, Louise Hayes, Aleksandr Kogan, Ann Medinets, Andrea Rozario, Helen Brown-Liburd, Michael Alles, Vernon Richardson, Guangyue Zhang, Yu Gu, Dong Gil Kim, Fangbing Xiong, Huaxia Li, Kitae Kum, Lanxin Jiang, Meehyun Kim, participants from 2018 Rutgers accounting seminars, anonymous reviewers and participants from 2019 Accounting Horizons Conference on “Data Analytics in Accounting”, anonymous reviewers and participants from 2019 American Accounting Association annual meeting, and anonymous reviewers and participants from 2020 American Accounting Association midyear meeting for their valuable suggestions.

Identifying Informative Audit Quality Indicators (IAQI) Using Machine Learning

ABSTRACT

Researchers have tested a wide array of theories and developed a broad set of explanatory variables for audit quality. However, little is known about which of these audit-related variables are the most predictive of audit failure (i.e., low-quality audits). We devote this study to provide researchers and regulators with a portfolio of informative audit quality indicators (IAQI) - publicly available theory-driven audit-related variables that can best predict audit failure. We adopt machine learning, a computational method that can navigate through a long list of variables and identify a subset that can make the best *out-of-sample* predictions. By applying various machine learning algorithms, we identify 11 audit-related variables as IAQI with their predictive power validated. These IAQI reflect auditor competence, independence, effort, incentive, and the quality of the audited financial reports. Our study provides researchers, regulators, audit firms, and investors with a list of IAQI that are predictive of audit failure.

JEL Codes: C53; M41; M42

Keywords:

Audit Quality, Audit Failure, Audit Quality Indicators, Machine Learning, Material Restatement

I. INTRODUCTION

Which aspects of an audit can best predict audit failure? Investors, auditors, academics, and regulators around the world are increasingly searching for answers to this question, especially after the recent accounting scandals of Wells Fargo (Wall Street Journal 2016) and Wirecard (Financial Times 2020). Following prior literature (e.g., Francis 2004; Francis and Michas 2013; Li, Qi, Tian, and Zhang 2017), we define an audit failure, or a low-quality audit, as the failure to issue a modified or qualified audit report when there is a material misstatement in the audit client's financial statements. Although extant research has tested a wide array of theories and developed a broad set of explanatory variables for audit quality, little is known about which of these audit-related variables (ARV) are the most predictive of audit failure. Answering this question can help investors assess the credibility of financial reports, guide auditors in identifying poor-quality audits, suggest regulators a starting point for investigations, and provide researchers a refined list of variables in audit quality research.

Most of existing audit quality research makes causal (or association) inferences on a factor's impact on audit quality. These studies adopt *explanatory modeling*, in which researchers establish hypotheses based on theories and then collect data and create statistical models to test against these hypotheses (Shmueli 2010). In explanatory modeling, researchers use *within-sample* measures, which are calculated based on the dataset on which the model is constructed, to evaluate a model's or a variable's explanatory power, such as R-squared, effect size, and statistical significance (Rapach and Wohar 2006; Shmueli 2010; Shmueli and Koppius 2011). While explanatory modeling plays an essential role in accounting and auditing research, it provides limited utility in prediction-related problems (Shmueli 2010; Kleinberg, Ludwig, Mullainathan, and Obermeyer 2015; Bertomeu 2020), such as "which aspects of audit can best predict audit

failure?” These prediction-related problems can be better answered by *predictive modeling* (Shmueli 2010; Shmueli and Koppius 2011). Predictive modeling is “the process of applying a statistical model or mining algorithm to data for the purpose of predicting new or future observations” (Shmueli 2010). In evaluating predictive modeling, researchers use *out-of-sample* measures, which are derived from holdout samples that are different from the sample on which the model is built (Shmueli 2010; Bao, Ke, Li, Yu, and Zhang 2020; Bertomeu 2020). Out-of-sample measures can assess a model’s performance on unseen situations, thus reflecting the predictive power of the model (Shmueli 2010; Shmueli and Koppius 2011). While predictive and explanatory modeling are fundamentally different, they complement each other to advance scientific inquiry (Shmueli 2010; Bertomeu 2020).

Predictive modeling applies machine learning, a computational method that can identify hidden patterns from large and high-dimensional datasets and can select a subset of variables that can make the best out-of-sample predictions (Alpaydin 2014; Cecchini, Aytug, Koehler, and Pathak 2010; Bertomeu 2020). With a long list of theory-driven audit-related variables and a sizable amount of historical data, machine learning is suitable for uncovering which aspects of audit are the most predictive of audit failure. Our study builds upon prior literature and leverages predictive modeling with machine learning to identify a portfolio of informative audit quality indicators (IAQI): theory-driven and publicly available audit-related variables that can best predict audit failure.

We use material restatements of annual financial reports, generated due to Generally Accepted Accounting Principles (GAAP) violations or frauds (hereafter, material annual restatement or MAR), as the proxy for audit failure, guided by earlier examples of auditing literature (e.g., Lobo and Zhao 2013; Kinney, Palmrose, and Scholz 2004; Newton, Wang, and

Wilkins 2013). MAR indicate that auditors signed off on materially misstated financial statements (Defond and Zhang 2014; Lobo and Zhao 2013; Center of Audit Quality 2013; Tan and Young 2015; Aobdia 2019; Audit Analytics 2020) because auditors are responsible for expressing opinions on whether the financial statements are presented in conformity with GAAP and obtaining reasonable assurance about whether the financial statements are free of material misstatement, whether caused by error or fraud (AS 1001; Kinney et al. 2004; Stanley and DeZoort 2007; Newton et al. 2013; Francis, Michas, and Yu 2013; Eshleman and Guo 2014).

To achieve our research objective of identifying IAQI, we first collect theory-driven audit-related variables (ARV) that are publicly available from the literature. Next, we apply each of the five popular machine learning algorithms identified in accounting research (Logistic Regression or LR, Random Forest or RF, Support Vector Machine or SVM, Artificial Neural Network or ANN, and AdaBoost or AB) (Perols 2011; Perols, Bowen, Zimmermann, and Samba 2016; Bao et al. 2020; Brown, Crowley, and Elliott 2020; Ding, Lev, Peng, Sun, and Vasarhelyi 2020), and adopt feature subset selection (FSS) to select a subset of the ARV that are the most predictive of MAR. Using a voting mechanism, we identify IAQI as the ARV that are determined to be the most predictive variables by the majority of the machine learning algorithms examined.

In this study, we collect 31 publicly available and theory-driven ARV that represent audit inputs, audit process, and audit output, and we create a dataset for U.S. public firms spanning the years 2005 to 2017. We find that variables that can best predict audit failure are those that reflect auditor competence, independence, effort, incentive, and the quality of the audited financial reports. Table 1 summarizes the IAQI identified from this study. We further examine the predictive power of IAQI for audit failure by establishing a cost-sensitive learning (CSL) and rolling-window prediction (RWP) mechanism where IAQI are inputs in each of the five popular machine learning

algorithms to predict MAR. After measuring overall predictive ability by area under the curve or AUC (Bao et al. 2020; Brown et al. 2020) based on holdout samples, we find that IAQI per se can reasonably predict MAR with the highest AUC reaching 64.5%,¹ comparable with that reported in relevant research (e.g., Bertomeu, Cheynel, Floyd, and Pan 2020) and validating the predictive power of IAQI.

[Insert Table 1 here]

The findings from this study are subject to robustness checks, including adopting alternative measures of audit failure and algorithm performance. Moreover, to further the understanding of IAQI, we perform two additional analyses. In the first additional analysis, we compare the predictive power of IAQI with financial variables and find that IAQI are better than financial variables in predicting MAR. This finding suggests future studies to examine whether audit-related factors or clients' financial conditions are the driving force in producing MAR. In the second additional analysis, we aggregate IAQI into a forward-looking index using machine learning, and perform a series of statistical tests to show that this index can provide incremental information that is associated with MAR. This additional analysis provides further validity towards IAQI and suggests a potential way of using IAQI in audit quality research.

Although studies exist that use machine learning to predict restatements/misstatements, the majority of them aim to forecast irregularities, such as severe frauds or accounting misconducts (e.g., Perols 2011; Perols et al. 2016; Cecchini et al. 2010; Brown et al. 2020; Bao et al. 2020). Therefore, they use misstatements announced on Accounting and Auditing Enforcement Releases

¹ AUC ranges from 0 to 1. AUC of 0.5 means random prediction. AUC below 0.5 means the prediction is worse than a random guess and AUC above 0.5 means the prediction is better than a random guess. AUC of 1 means perfect prediction. The best practice in accounting research that predicts restatements/misstatement produces AUC in the range of 60% to 70%, depending on the specific research designs and datasets (e.g., Dechow et al. 2011; Perols et al. 2016; Bao et al. 2020; Bertomou et al. 2020).

(AAER) that are results of SEC investigations for securities law violations (SEC 2017). In contrast, this study uses material restatements disclosed on Item 4.02 of Form 8-K, which underscore the non-reliance of past financial reports (Center of Audit Quality 2013). Compared to AAER, MAR or restatements in general can capture a wider range of potentially low-quality audits than the extreme cases in AAER that received SEC enforcement actions or lawsuits (Panel on Audit Effectiveness 2000; Francis 2004). In sensitivity analysis, we use MAR instances that are also included in AAER as an alternative proxy for audit failure and obtain similar findings. The closest research to ours is Bertomeu et al. (2020). While Bertomeu et al. (2020) use machine learning to predict Form 8-K material misstatements using a broad set of variables from accounting, capital markets, governance, and auditing, they focus on the effective detection of material misstatement.² In comparison, this study is audit-oriented and aims to identify IAQI, a portfolio of theory-driven audit-related variables that are the most predictive of material misstatement, a proxy for audit failure. Furthermore, this study includes a more complete list of ARV (31 variables) than Bertomeu et al. (2020) (8 variables). Results from this study can shed light on the relevance of existing explanatory models in audit quality research by examining the distance between theory and practice (Shmueli 2010; Shmueli and Koppius 2011). Appendix A presents a detailed comparison between our paper and relevant literature.

This study makes several important contributions. First, it provides researchers and regulators with a list of theory-driven audit-related variables that are the most predictive of audit failure out-of-sample. Previous research has identified many audit-related variables. However, it is unknown which of them are the most predictive of audit failure and, therefore, should be

² Bertomeu et al. (2020) refer to prediction of new observations as “detection”, and prediction of future observations as “prediction”. This study follows information system and accounting literature (e.g., Shmueli and Koppius 2011; Bao et al. 2020) and uses the term “prediction” to mean prediction of new observations. Therefore, the term “prediction” in this study is equivalent to “detection” in Bertomeu et al. (2020).

included in a prediction model. Although Aobdia (2019) has investigated the degree of agreement between fifteen measures of audit quality used in academia and two proprietary measures of audit process quality, he focuses on the within-sample correlations, rather than the out-of-sample predictive power of audit-related variables. Even though Rajgopal, Srinivasan, and Zheng (2021) have evaluated how well existing audit quality proxies predict specific allegations related to audit deficiencies, they still construct within-sample evaluation metrics. Thus, this study furthers researchers' and regulators' understanding about which audit-related variables are the most predictive and can be used for forecasting audit failure out-of-sample.

Second, our work adds to the growing stream of accounting research that adopts predictive modeling (e.g., Cecchini et al. 2010; Perols 2011; Perols et al. 2016; Bao et al. 2020; Brown et al. 2020; Ding et al. 2020; Bertomeu et al. 2020; Hunt, E., Hunt, J., and Richardson 2019). In particular, it identifies a salient prediction issue in audit quality studies: namely, which audit-related variables that have been widely studied in the existing literature are the most predictive of audit failure and how effective they can forecast audit failure? In a domain such as audit quality in which researchers have tested a wide array of theories and accumulated a broad knowledge of explanatory factors, it is vital to understand the predictive power of audit-related variables that are operationalized based on theoretical constructs. Understanding the predictive power can spur comparisons of competing theories, different operationalizations of constructs, and different measurement instruments (Shmueli 2010; Shmueli and Koppius 2011). This study builds upon the findings (i.e., which variables are associated with audit quality) from previous literature and goes one step further to adopt predictive modeling to explore the predictive power of these variables.

Lastly, this study has practical implications for regulators and other stakeholders, including audit committees, audit firms, and investors. In particular, this study echoes PCAOB's call for

research on audit quality indicators (AQI) using public source information (PCAOB 2015). When interpreted within specific contexts, IAQI can assist regulators and stakeholders in assessing audit quality and can facilitate relevant decision-making processes (PCAOB 2015). Additionally, these measures can aid audit firms in risk assessment and management while helping investors to more effectively assess firms' reporting risks as well (PCAOB 2015). Furthermore, this study lends regulators and other stakeholders the methodologies on how to identify predictive variables and establish predictive scores using proprietary data (PCAOB 2019).

II. MACHINE LEARNING

Machine learning is a computational method that can identify hidden patterns from data and make predictions (Alpaydin 2014). Compared to traditional approaches of data analysis, machine learning works better with large or high-dimensional datasets and requires fewer underlying assumptions (Alpaydin 2014; Cecchini et al. 2010; Bertomeu 2020). Machine learning is used in predictive modeling, whose purpose is to predict new or future observations (Shmueli 2010).³ Predictive modeling is evaluated by out-of-sample measures that are derived from holdout samples, such AUC (Shmueli 2010; Bao et al. 2020). Out-of-sample metrics can assess a model's performance on unseen situations, thus reflecting the model's predictive power (Shmueli 2010; Shmueli and Koppius 2011).

A main subset of machine learning is called supervised learning, in which the algorithm learns from available examples or experiences with known positive or negative "labels" and then

³ When the objective is to predict the outcomes of new observations given their input values, the type of predictive modeling is called non-stochastic prediction (Shmueli 2010). Non-stochastic prediction is commonly used in accounting research that adopts machine learning (e.g., Perols 2011; Perols et al. 2017; Bao et al. 2020; Bertomeu et al. 2020). In contrast to non-stochastic prediction is the temporal prediction, where observations until time t are used to forecast future values at time $t+k$, $k>0$. In this study, we use the term "prediction" to mean non-stochastic prediction, following most prior accounting literature (e.g., Perols 2011; Perols et al. 2017; Bao et al. 2020). We do not adopt temporal prediction because audit quality can only be assessed when the audit is finished.

makes predictions about future instances (Alpaydin 2014). For example, after being provided with examples of known fraudulent and legitimate transactions, a supervised learning algorithm can be trained to extract identifiable patterns. The trained algorithm will then be able to predict whether a new transaction is fraudulent. Since labeled data is available (i.e., with known MAR or not), supervised learning is adopted for this study. Popular supervised learning algorithms used in accounting research include Artificial Neural Networks (ANN), Logistic Regressions (LR), Random Forest (RF), Support Vector Machine (SVM), and AdaBoost (AB) (Alpaydin 2014; Cecchini et al. 2010; Perols 2011; Perols et al. 2016; Bao et al. 2020; Brown et al. 2020).⁴

Accounting researchers have explored the use of machine learning to predict fraud (Perols 2011; Perols et al. 2016; Cecchini et al. 2010; Bao et al. 2020), bankruptcy (Gentry, Shaw, Tessmer, and Whitford 2002), restatement/misstatement (Dutta, I., Dutta, S., and Raahemi 2017; Bertomeu et al. 2020; Hunt et al. 2019), and accounting estimates (Ding et al. 2020). They also use machine learning as a tool to make automatic classifications or to extract or generate variables that can be used in explanatory-modeling studies (e.g., Li 2010; Sun and Sales 2018; Sun 2018; Hayes and Boritz 2019; Brown et al. 2020). This study adds to the accounting literature by using machine learning to identify audit-related variables that are the most predictive of material restatements.

III. MATERIAL ANNUAL RESTATEMENTS AND AUDIT-RELATED VARIABLES

Material Annual Restatements (MAR)

Audit quality is either unobservable or can only be observed when there are known errors or deficiencies (PCAOB 2015; Causholli and Knechel 2012). Consequently, researchers, regulators, and professionals often describe what high audit quality “is not” (i.e., in terms of errors

⁴ Detailed descriptions for each of the popular supervised learning algorithm will be provided upon request.

or deficiencies that reduce audit quality) rather than defining audit quality for what “it is” (Knechel, Krishnan, Pevzner, Shefchik, and Velury 2013). In academic research, various proxies are used to measure audit quality.⁵ There is no consensus on what the best measure of audit quality is because different measures capture discrete aspects of an audit (Defond and Zhang 2014). Overall, compared to less direct measures of audit quality (e.g., accruals quality), more direct measures (e.g., restatements) can capture egregious audit failures and have higher consensus on measurement, but they are less able to capture the continuous nature of audit quality (Defond and Zhang 2014).

In this study, material restatement of annual reports due to GAAP violations or frauds (or MAR) is used as the measure of audit failure based on the following reasons. First, we capture the aspect of materiality in audits (Christensen, Glover, Omer, and Shelley 2016) by focusing on the material restatements (a.k.a., “Big R” or re-issuance) that are disclosed in Item 4.02 of Form 8-K (Center of Audit Quality 2013; Audit Analytics 2020).⁶ Restatements announced in 8-Ks address a material error that requires re-issuance of past financial statements, and these “Big Rs” are the primary type of restatements to garner concern (Center of Audit Quality 2013; Audit Analytics 2020). In contrast, restatements disclosed in periodic reports (e.g., 10-Ks, 10-K/As, 10-Qs) are immaterial changes that are considered ongoing adjustments made in the ordinary course of business (Audit Analytics 2020). Therefore, material restatements are a tacit admission that auditors signed off on materially misstated financial statements (Defond and Zhang 2014; Lobo

⁵ Audit quality proxies can be grouped into output-based and input-based audit quality measures (Defond and Zhang 2014). Examples of output-based audit quality measures include material misstatements, going concern opinions, financial reporting quality, perceptions of audit quality (e.g., market reaction and cost of capital), and auditor-client contracting features (e.g., audit fees; Defond and Zhang 2014). Input-based measures, on the other hand, focus on auditor characteristics, such as Big 4 and industry specialization (Defond and Zhang 2014).

⁶ The SEC stipulates that material restatements must be filed in section 4.02 “Non-Reliance on Previously Issued Financial Statements or a Related Audit Report or Completed Interim Review” on Form 8-K or Form 8-K/A. See <https://www.sec.gov/rules/final/33-8400.htm>

and Zhao 2013; Center of Audit Quality 2013; Tan and Young 2015; Aobdia 2019; Audit Analytics 2020).

Second, to reflect the expected responsibilities of auditors, we limit the reasons of material restatement to GAAP violations or frauds because auditors are held responsible for expressing opinions on whether the financial statements are presented in conformity with GAAP and obtaining reasonable assurance about whether the financial statements are free of material misstatement, whether caused by error or fraud (AS 1001; Kinney et al. 2004; Stanley and DeZoort 2007; Newton et al. 2013; Francis et al. 2013; Eshleman and Guo 2014). Furthermore, we restrict material restatements to those of annual reports because only the annual financial reports of public firms are required to be audited (Kinney et al. 2004; Stanley and DeZoort 2007; Cao, Myers, and Omer 2012; Lobo and Zhao 2013; Bills, Cunningham, and Myers 2016).

Third, there is a consensus in the literature that restatements can capture actual audit failure with little measurement error and that they are a relatively direct output-based measure of audit failure as compared to other proxies (e.g., Romanus, Maher, and Fleming 2008; Defond and Zhang 2014; Francis et al. 2013; Knechel and Sharma 2012; Newton et al. 2013; Ettredge, Fuerherm, and Li 2014; Eshleman and Guo 2014; Bills et al. 2016; Lennox 2016; Aobdia 2018; Bhaskar, Schroeder, and Shepardson 2019; Cunningham, Li, Stein, and Wright 2019; Ahn, Hoitash, and Hoitash 2020; Rajgopal et al. 2021).

Fourth, both audit professionals and investors also identify financial statement restatements as the most readily available and outcome-based signal of audit failure since the existence of a restatement indicates that an improved audit process could have identified the error (Christensen et al. 2016; Gaynor, Kelton, Mercer, and Yohn 2016).

Lastly, material restatements have been found to be strongly associated with the audit process quality measured by PCAOB's Part I Findings and with the internal assessments of audit quality from audit firms (Aobdia 2019).

Limitations of Using MAR as a Proxy for Audit Failure

While this paper uses MAR as a proxy of audit failure, it also acknowledges the limitations of using this proxy. First, not all poor audit quality incidences produce a material restatement because a MAR is a joint outcome acknowledging that a material misstatement happened, that auditors failed to identify the misstatement, and that the misstatement was eventually unveiled and disclosed (Gaynor et al. 2016). Therefore, MAR measures known audit failure and the absence of MAR does not indicate high audit quality. Second, MAR cannot capture the procedure-related characteristics of audit quality, such as the extent and appropriateness of evidence supporting the auditor's opinion, and the degree of correspondence between the auditor's procedures and auditing standards (Bell, Causholli, and Knechel 2015). Third, evidence from lawsuits data of whether auditors are held accountable for restatements is mixed.⁷

Despite its limitations, MAR remains the most readily accessible indicator of audit failure from public source information (Christensen et al. 2016) and it is commonly used in audit quality research (Romanus et al. 2008; Defond and Zhang 2014; Francis et al. 2013; Knechel and Sharma 2012; Newton et al. 2013; Ettredge et al. 2014; Eshleman and Guo 2014; Bhaskar et al. 2019).

⁷ Recently, Lennox and Li (2020) find that auditors are rarely blamed when there are allegations of financial reporting failures by comparing auditor lawsuits to a sample of accounting lawsuits in which audit firms are not sued. However, the infrequency of lawsuits against auditors when there are restatements does not necessarily indicate that auditors are generally not responsible for materially misstated financial reports. Instead, it could be because the liability standards imposed on auditors have been elevated by Supreme Court's rulings in *Tellabs v. Makor* and *Janus v. First Derivative* and that many lawsuits against auditors are dismissed at the initial stage of the pleading procedure (Honigsberg, Rajgopal, and Srinivasan 2019).

Audit-Related Variables

This research utilizes public source information to derive IAQI that are the most predictive of audit failure. Specifically, we first collect publicly available variables that are found to be associated with audit quality (audit-related variables, or ARV) from the literature. Then, machine learning is adopted to use different subsets of ARV to predict MAR. The subset of ARV that can best predict MAR are IAQI.

By examining research on audit quality, we identify 31 theory-driven ARV that are publicly accessible from the literature. Following the frameworks of Gaynor et al. (2016), Francis (2011), and PCAOB (2013), we classify these ARV into three categories: audit input, audit process, and audit output. Then, we further classify them into sub-categories, including auditor characteristics, task characteristics, environmental characteristics, auditor-client contracting features, auditor communication, and the quality of the audited financial statements. Based on the literature, we also summarize different aspects of an audit each ARV captures. Table 2 provides the list of ARV and their classifications.

[Insert Table 2 here]

IV. DATA AND RESEARCH DESIGN

Data and Sample

We start with the COMPUSTAT population from 2000 to 2019.⁸ After removing firm-year observations that do not have Central Index Key (CIK) numbers, that do not file 10-K, and that are duplicates, we match the remaining data with the Audit Analytics (AA) dataset by fiscal year-

⁸ We download the data from 2000 so that we can calculate tenure (i.e., the length of auditor-client relationship) accurately.

end.⁹ After removing observations with missing values and keeping observations from 2005 to 2017, there are 26,339 firm-year observations. The details of sample derivation are provided in Table 3. We choose the starting year to be 2005 because the Form 8-K disclosure requirement came into effect in August of 2004.¹⁰ The data ends in 2017 because there is an average of 2-year lag between the restatement filing date (the most recent filing date before this study was 2019) and the restated date (Lobo and Zhao 2013; Eshleman and Guo 2014).

[Insert Table 3 here]

We obtain restatement data from the AA database, which houses a complete population of restatements records originated from Form 8-Ks or periodic reports (Karpoff, Koester, Lee, and Martin 2017; Lobo and Zhao 2013; Audit Analytics 2020).¹¹ To derive MAR, we first consult the “Non-Reliance Restatements” section of the AA database and download restatement data from 2000 to 2019. Then, we obtain the material restatements by limiting the sources of disclosure to be 8-K (Center of Audit Quality 2013; Audit Analytics 2020).¹² Next, we keep material restatements that can be ascribed to GAAP violations or fraud.¹³ Lastly, we exclude the interim restatements and keep the annual restatements.¹⁴ The resulting restatement instances are MAR. For the restated firm-year observations in the sample, the “year” denotes the fiscal year in which

⁹ COMPUSTAT and Audit Analytics have different ways of deciding fiscal year. Therefore, we use the field “Data Date” in COMPUSTAT and “Fiscal year ended” in Audit Analytics for matching.

¹⁰ See <https://www.sec.gov/rules/final/33-8400.htm>

¹¹ Other databases such as the one associated with the Center for Financial Reporting and Management (CFRM) mainly provide misstatements disclosed on AAERs that were generated due to SEC investigations for accounting or auditing misconduct or that led to lawsuits (Karpoff et al. 2017). Accordingly, such databases are more often used in research related to fraud/misconduct prediction.

¹² In the “Non-Reliance Restatements” section of the Audit Analytics database, there is one data field that provides information about the source document in which the restatement has been announced.

¹³ In the “Non-Reliance Restatements” section of the Audit Analytics database, there are data fields indicating whether a restatement is related to GAAP violations, fraud (financial fraud, irregularities and misrepresentations), or clerical errors.

¹⁴ The regular module of Audit Analytics only provides a time range of the restated financial statements, which comprise both interim restatements and annual restatements. We use a Python script to derive the annual restatements and will provide the Python codes by request. This way of identifying annual restatements is consistent with Audit Analytics 2020.

the firm’s annual report ultimately received material restatement and not the year the restatement was announced.

Additionally, some MAR instances in our sample occur across consecutive years. For example, a firm may have materially restated its 10-K in 2011, 2012, and 2013. According to the data provider, most consecutive instances occur because the material misstatement from the starting year carries over to subsequent years; therefore, a MAR’s starting year is usually the year when the material misstatement originated. To more accurately capture audit failure, we retain only the starting year of the MAR instances in our sample (Stanley and Dezoort 2007).¹⁵ It is necessary to remove repeated MAR also because serial MAR that span both the training and testing periods could overstate the performance of ensemble learning algorithms (e.g., AdaBoost and Random Forest) (Bao et al. 2020). In un-tabulated analyses, we use the sample without correcting for consecutive MAR and obtain similar results.

Our sample contains 446 starting-year MAR observations, which account for around 1.69% of the entire population. Table 4 presents the starting-year MAR distribution by fiscal year. The distribution of our sample is consistent with Audit Analytics’ 2020 restatement report (Audit Analytics 2020). Table 5 provides the descriptive statistics of the 31 ARV, most of which are comparable to previous studies. The pairwise correlations of ARV are provided in Appendix B.

[Insert Table 4 here]

[Insert Table 5 here]

Research Design

To identify IAQI, we perform feature subset selection (FSS) using five popular machine learning algorithms to select a subset of ARV that can best predict MAR. Then, we examine the

¹⁵ For example, if a firm restated its 10-K in consecutive years from 2011 to 2013, we only keep the observation for 2011 and delete those for 2012 and 2013.

predictive power of IAQI by inputting IAQI into multiple machine learning algorithms to predict MAR via a cost-sensitive learning (CSL) and rolling-window prediction (RWP) mechanism. Figure 1 presents our overall research design.

[Insert Figure 1 here]

Feature Subset Selection (FSS)

In real-world situations, the most predictive features for a target outcome are often unknown a priori (Dash and Liu 1997; Tang, Alelyani, and Liu 2014). To identify those predictive features, one can start with a list of candidate features that are identified from domain knowledge (Dash and Liu 1997; Tang et al. 2014). In this study, the candidate features are the 31 ARV identified from prior audit quality literature. In many applications, including all candidate features to predict the target outcome does not necessarily generate better performance than including only a selected subset of the candidate features (Dash and Liu 1997; Hocking and Leslie 1967; Perols 2011; Hastie, Tibshirani, and Tibshirani 2017a; Hastie, Tibshirani, and Friedman 2017b; Bao et al. 2020; Bertomeu et al. 2020). This is because some candidate features may be redundant or irrelevant in predicting the target outcome, thus causing model overfitting (Dash and Liu 1997; Tang et al. 2014; Hastie et al. 2017a; Hastie et al. 2017b). Removing those irrelevant/redundant features can create a parsimonious model with enhanced algorithm performance as well as reduced computational complexity (Hall and Smith 1998; Tang et al. 2014; Bao et al. 2020).

FSS is a technique to select a subset of features that can maximize the performance of a learning algorithm (Dash and Liu 1997; Tang et al. 2014). In this study, we adopt backward stepwise selection. Backward stepwise selection starts with all candidate features, then it iteratively removes the feature that has the least impact on the pre-defined performance metrics, and it stops when there is no significant improvement to the performance (Hastie et al. 2017b). We

use backward stepwise selection in this study because it not only makes feature selection computationally feasible (Hastie et al. 2017a; Hastie et al. 2017b), but also avoids potential omission of predictive candidate features (Guyon and Elisseeff 2003).

Depending on the learning mechanism, the best subset of features will differ across discrete learning algorithms (Perols 2011). To reduce any potential bias arising from the choice of a particular machine learning algorithm, this study refers to Perols (2011) and adopts a “voting mechanism” in which an ARV is considered as an IAQI if it is selected as the best subset of features by the majority of the machine learning algorithms adopted in this research. Following Perols (2011), we implement FSS using 5-fold stratified cross-validation on the entire dataset.¹⁶

Performance Evaluation Metric

Following prior accounting literature (e.g., Dechow, Ge, Larson, and Sloan 2011; Bao et al. 2020; Brown et al. 2020), we use area under the Receiver Operating Characteristic (ROC) curve, or AUC, to evaluate the overall predictive ability of a machine learning algorithm. ROC curve is a plot of the true positive rate (i.e., the percentage of audit failure accurately classified as audit failure) on the y-axis against the false positive rate (i.e., the percentage of non-audit-failure firms incorrectly classified as audit failure) on the x-axis for different possible classification thresholds (Bradley 1997). AUC ranges from 0 to 1: AUC of 0.5 means random prediction; AUC below 0.5 means the prediction is worse than a random guess; AUC above 0.5 means the prediction is better than a random guess; and AUC of 1 means perfect prediction (Bradley 1997). Extant accounting research that predicts restatements/misstatement produces AUC in the range of 60% to 70%

¹⁶ Details of the operationalization of FSS will be provided upon request. We adopt cross-validation because the main objective here is to identify patterns from all instances regardless of when they have happened. Furthermore, we do not adjust the cost imbalance in FSS because the objective here is not to make the best prediction but to identify the most relevant features. We will adjust the cost imbalance when we evaluate the predictive power of these features in the next section of cost-sensitive learning and rolling window prediction.

depending on specific research design and dataset (e.g., Dechow et al. 2011; Perols et al. 2016; Bao et al. 2020; Bertomou et al. 2020).

Previous studies have also used estimated relative cost of misclassification (ERC) or expected cost of misclassification (ECM) to evaluate algorithm performance (Perols 2011; Perols et al. 2016). However, ERC and ECM are only informative when the prior probability of positive instances and the misclassification costs can be reasonably estimated (Perols 2011; Perols et al. 2016). In contrast, AUC does not require such estimates, representing an average comparison of classifiers in domains with class and cost imbalances (Perols 2011). Since there is scant research relating to estimates of the prior probability of an audit's deficiency or the range of misclassification costs for "Big R" predictions, we consider the use of AUC to measure the overall performance of the machine learning algorithms in audit quality prediction an appropriate application. In the sensitivity analysis, an alternative measure is used to evaluate the algorithm performance.

Examining the Predictive Power of IAQI

After IAQI are identified from the FSS procedure, we examine their predictive ability by inputting them into each of the five common machine learning algorithms to predict MAR via cost-sensitive learning (CSL) and the rolling-window prediction (RWP) mechanism. This section introduces CSL and RWP.

Cost-Sensitive Learning (CSL)

In predicting audit failure, a machine learning algorithm can make two types of mistakes: false-positive errors (i.e., Type 1 errors) and false-negative errors (i.e., Type 2 errors). In this study, "positive" means that observations have known MAR and "negative" otherwise. A false negative error happens when an algorithm mistakenly classifies a positive instance as negative, while a false

positive error happens when the algorithm classifies a negative instance as positive. In the context of this research, a false negative error is a more severe mistake than a false positive error because the investigation costs incurred from false positives are usually much lower than the financial costs (e.g., financial losses for investors and litigation costs for the issuer and audit firm) arising from audit failure (Beneish and Vorst 2020). Similar cost imbalances also occur in other domains like fraud detection and loan default prediction (e.g., Perols 2011; Perols et al. 2016; Beneish and Vorst 2020). We define a *misclassification cost* as the cost ratio between false negatives and false positives (Perols 2011). For example, a misclassification cost of 20 indicates that a false negative is 20 times as costly as a false positive. Since the actual misclassification cost of audit quality prediction is unknown, we will test different misclassification costs ranging between 1 and 100 (Perols 2011).

To sufficiently consider the cost imbalance issue in our research setting, we adopt a CSL mechanism that can adjust the ratio between the number of positive and negative instances in the training dataset (Elkan 2001). In particular, we follow Perols et al. (2016) and adopt the multi-subset Observation Undersampling (OU) method (Chan and Stolfo 1998) to implement CSL. Details of CSL and OU method are provided in Appendix C.

Rolling-Window Prediction (RWP)

To mimic the decision-making process of learning from the past and predicting the future, and to ensure that the prediction model adapts to the changing environments through time, we adopt a rolling-window prediction in which the machine learning model is trained with five years of historical data, and then the trained model is used to predict outcomes following two years (Bao et al. 2020; Brown et al. 2020).¹⁷ Specifically, we use seven sets of training and testing data in the

¹⁷ The main findings hold when we also set out a validation sample to tune the hyper-parameters of the model.

rolling-window prediction with the testing years spanning from 2011 to 2017.¹⁸ Our predictions target periods that are two years into the future because the lag between financial report filing and misstatement identification is on average two years (Lobo and Zhao 2013; Eshleman and Guo 2014; Bao et al. 2020). In this way, we can ensure that most of the past MAR have already been revealed at the time the prediction is made, reducing look-ahead biases (Brown et al. 2020). In untabulated analyses, predictions targeting periods three years in the future or training machine learning models using eight years of historical data show similar results. Appendix D provides the overall experiment procedure to evaluate the predictive power of IAQI.

V. RESULTS

Identifying IAQI

Table 6 displays the results of FSS for each machine learning algorithm. For each row in Table 6, the value “1” indicates that an algorithm selects this variable from FSS. 11 ARV obtain the majority of the “votes” from 5 algorithms; we consider these ARV to be IAQI.

[Insert Table 6 here]

The 11 IAQI represent the variables that are the most predictive of MAR, which we use as a proxy for audit failure. Since all of the ARV are theory-driven variables selected from relevant literature, the IAQI generated from the predictive modeling process serve as a validation on the predictive power of the audit-related variables examined in previous studies. However, variables that are not selected as IAQI are not irrelevant to understanding audit quality. Instead, the implication is that when the objective is to *predict* MAR, IAQI are more predictive than other

¹⁸ These seven sets are: 1) training using 2005 to 2009 data in order to predict the outcome in 2011; 2) training using 2006 to 2010 data in order to predict the outcome in 2012; 3) training using 2007 to 2011 data in order to predict the outcome in 2013; 4) training using 2008 to 2012 data in order to predict the outcome in 2014; 5) training using 2009 to 2013 data in order to predict the outcome in 2015; 6) training using 2010 to 2014 data in order to predict the outcome in 2016; and 7) training using 2011 to 2015 data in order to predict the outcome in 2017.

variables. Besides, unlike in explanatory modeling where it is crucial to examine the direction in which a variable affects audit quality, in predictive modeling, the direction of effect is of less concern because pursuing a correctly specified model may compromise its predictive power (Bao et al. 2020; Shmueli 2010; Hastie et al. 2017b). Based on our categorization of ARV in Table 2, we present the 11 IAQI together with their categories, sub-categories, and aspects of audit captured in Table 1.

The 11 IAQI are composed of variables from the whole cycle of an audit engagement: audit input, audit process, and audit output. Specifically, these IAQI represent auditor characteristics (competence and resource measured by office size), audit task characteristics (informational advantage and independence measured by tenure; audit efficiency measured by audit report lag; workload and comprehensiveness measured by integrated audit), auditor-client contracting features (incentive measured by auditor resignation; effort and budget measured by audit fees), auditor communication (independence and competence measured by internal control weakness report), and the quality of the audited financial statements (within-GAAP manipulations measured by accruals and discretionary accruals).¹⁹

Examining the Predictive Power of IAQI

Table 7 (Figure 2) provides the descriptive statistics (plot) of the (average) AUC values for each algorithm at different misclassification costs.²⁰ We find that IAQI can reasonably predict MAR with an average AUC of 0.621 and a maximum of 0.645, comparable with that reported from Bertomeu et al. (2020).²¹

¹⁹ Detailed discussions for each of the IAQI will be provided upon request.

²⁰ Since the objective of this section is to examine the predictive power of IAQI, rather than constructing models that can outperform existing ones, we simply present the AUC of using IAQI to predict MAR under different popular machine learning algorithms, instead of comparing the performance with benchmark models.

²¹ The results reported in Perols (2011), Perols et al. (2016), and Bao et al. (2020) are not comparable to ours because they have different research settings from ours: they use AAER as the dependent variable and financial variables as

To formally examine the relative performance of the algorithms at different levels of misclassification costs, we apply a one-way ANOVA test to the AUC values of different algorithms for each misclassification cost. The ANOVA test uses the algorithm factor as the main effect on AUC. The un-tabulated results indicate a significant difference among the AUC values of the five algorithms at each misclassification cost. Therefore, we further perform post-hoc analysis using Tukey’s Honest Significance Difference (i.e., Tukey’s HSD test; Perols 2011). Table 8 reports the Tukey HSD results in the format of connected letters to rank the relative AUC values of algorithms at each misclassification cost. We rank the average AUC alphabetically, with “A” indicating the lowest value. The difference in AUC between any two algorithms is significant when they are ranked by different letters.²² The Tukey HSD results show that the AUC for AB is among the highest in most misclassification costs (1 and 10 - 70). When the misclassification cost is low (below 10), AB, ANN, and LR are equivalent. When the misclassification cost is high (above 70), ANN, LR, and RF have equivalently high AUC. Overall, within the five algorithms examined in this study, AB performs the best in using IAQI to predict MAR as it has high AUC in a wide range of misclassification costs.

[Insert Table 7 here]

[Insert Figure 2 here]

[Insert Table 8 here]

the predictors. Bertomeu et al. (2020) adopt a similar setting to ours in one of their tests and they report an AUC of 0.617 using only audit variables to predict material misstatement.

²² For example, in Table 8, when the misclassification cost is 50, LR, RF, and SVM have significantly lower AUC than AB. ANN has a lower (higher) AUC than AB (LR, RF, and SVM), but the difference is not statistically significant.

VI. SENSITIVITY ANALYSIS

Alternative Measure of Audit Failure

In this section, we adopt an alternative measure of audit failure: MAR that also receive Accounting and Auditing Enforcement Releases (AAER) from the SEC. AAER are results of SEC investigations for securities law violations (SEC 2017). Prior literature adopts AAER instances to identify severe frauds or accounting misconducts (e.g., Perols 2011; Perols et al. 2016; Cecchini et al. 2010; Brown et al. 2020; Bao et al. 2020). The alternative measure of audit failure used here is a subset of MAR that are also investigated by SEC and eventually received AAER, indicating that these material misstatements involve severe frauds or accounting misconducts. We use this alternative measure of audit failure and perform the same analysis of FSS as stated in prior section. Under this alternative audit failure measure, six out of the 11 IAQI from the main analysis are also identified as the most predictive variables: office size, tenure, audit fee, Disc. Accruals, Abs(Accruals), and Abs(Accruals/CFO). Most of the IAQI overlap under this alternative measure of audit failure, strengthening the robustness of the main analysis.

Alternative Measure of Algorithm Performance

In terms of algorithm performance measure, besides using AUC, which is area under the ROC curve, we adopt an alternative evaluation metric, area under the Precision-Recall Curve (PR-AUC). Precision-Recall curve is similar to ROC curve but with one axis changed from false positive rate to precision (Jeni, Cohn, and De La Torre 2013; Saito and Rehmsmeier 2015). PR-AUC balances precision (i.e., fraction of true audit failures among the predicted audit failures) and recall (i.e., the fraction of true audit failures predicted), and it is an alternative metric to evaluate the algorithm when the data is imbalanced (Jeni et al. 2013; Saito and Rehmsmeier 2015). Some prior studies use F-score to measure algorithm performance (e.g., Bertomeu et al. 2020). Similar

to PR-AUC, F-score also weighs precision and recall (Rijsbergen and Joost 2004). However, unlike F-score whose value depends on specific classification thresholds, PR-AUC is an aggregated value under different possible classification thresholds, thus better reflecting the overall predictive power (Jeni et al. 2013; Saito and Rehmsmeier 2015). We use PR-AUC as an alternative measure of algorithm performance and perform the same FSS as stated before. Under this alternative algorithm performance measure, seven out of 11 IAQI from the main analysis are also found to be the most predictive variables: office size, tenure, audit fee, auditor resignation, internal control weakness, Abs(Accruals), and DD Residual, further supporting the robustness of the main analysis.

Overall, across different measures of audit failure and algorithm performance, office size, tenure, audit fee, and Abs(Accruals) are constantly identified as the most predictive variables for audit failure.

VII. ADDITIONAL ANALYSIS

Comparing Predictive Powers of Audit-Related Variables and Financial Variables

In the main analysis of this study, we only utilize audit-related variables as the predictors of MAR in consideration of our research objectives. In contrast, other studies have mainly used financial variables as predictors of restatement/misstatement (e.g., Cecchini et al. 2010; Dechow et al. 2010; Perols 2011; Perols et al. 2016; Dutta et al. 2017; Bao et al. 2020). Although Bertomeu et al. (2020) compare the predictive powers of different groups of variables (accounting, capital markets, governance, and auditing) for material misstatement, they include limited categories of audit-related variables. Therefore, we still know relatively little about which group of variables can better predict MAR: financial/accounting variables or audit-related variables. This section is dedicated to a comparison of the predictive powers of audit-related variables and financial

variables, similar to the process conducted in Jones (2017) and Bertomeu et al. (2020). We follow the research on fraud prediction using financial variables (e.g., Bao et al. 2020; Perols 2017; Dechow et al. 2011; Cecchini et al. 2011) and collect 33 raw financial variables from the COMPUSTAT database.²³ Panel A of Table 9 documents these variables. Recalling how FSS is used to identify IAQI in the main analysis, here, the same method is adopted to select the most informative financial variables (IFV) from the 33 raw financial variables. Out of these financial variables, we identify eight as IFV.²⁴

[Insert Table 9 here]

Next, we perform the same cost-sensitive learning and rolling-window prediction procedures as described in the previous section with inputs of only IAQI, only IFV, and IAQI combined with IFV, respectively. In this process, we use AdaBoost, the best-performing algorithm based on overall predictive ability measured by AUC from our main analysis. Panel B of Table 9 presents a comparison of AUC values of different groups of variables. The AUC using only IFV to predict MAR is comparable to the experiment performed by Brown et al. (2020), which uses the F-score from Dechow et al. (2010) to predict fraudulent restatements extracted from the AA database, and that reported in Bertomeu et al. (2020). Consistent with Bertomeu et al. (2020), our comparison results show that IAQI have significantly higher predictive power than IFV. Additionally, IAQI combined with IFV do not provide significant incremental predictive power as compared to using IAQI alone, unless the misclassification cost is high (above 60). The findings

²³ For company's data prior to a restatement due to an SEC investigation, COMPUSTAT gives the original numbers with the "PRE_AMENDS" tag. Following Ding, Peng, and Wang (2019), we replace restated account values with the original, non-restated values if the "PRE-AMENDS" tag exists.

²⁴ They are: Accounts Payable - Trade, Cash and Short-Term Investments, Short-Term Investments - Total, Sales/Turnover (Net), Sale of Common and Preferred Stock, Income Taxes - Total, Working Capital (Balance Sheet), Interest and Related Expense – Total.

from this section suggest further studies to examine whether audit-related factors or clients' financial conditions are the driving force in producing MAR.

Aggregating IAQI into PAQI

Inspired by existing research that establishes a score based on a list of financial variables to reg flag earnings management and misstatement (Dechow et al. 2011), we further use machine learning to aggregate IAQI into a forward-looking index, predictive audit quality index (PAQI), to enhance the processing fluency (Reber, Schwarz, and Winkielman 2004) of IAQI.

In aggregating IAQI into PAQI, AdaBoost is used because it is the best-performing algorithm based on overall predictive ability measured by AUC from our main analysis; then, we input IAQI into this chosen algorithm to obtain probability prediction via CSL and RWP. Specifically, for each rolling-window prediction set, we tune the hyper-parameters of AdaBoost via 5-fold stratified cross-validation on the training data;²⁵ We then utilize the AdaBoost model with the tuned hyper-parameters in the cost-sensitive learning and rolling-window prediction, setting the misclassification cost as 20; Lastly, we standardize the probability prediction output from the algorithm to obtain PAQI.²⁶ Panel A of Figure 3 provides summary statistics of the PAQI. The average and the standard deviation of the Predictive MAR Score are 0 and 1, respectively, because the scores have been transformed into a standardized distribution. Panel B shows that the observations in the test set that have actual MAR feature PAQI that are significantly higher than

²⁵ Hyper-parameters are parameters whose values are set manually instead of being “learned” from training data (one example of a hyper-parameter would be the type of loss function used). We set the misclassification cost here at 20 just for illustration purpose since the accurate misclassification cost for “Big R” prediction is unknown. We also generate PAQI under other cost levels and the results are similar.

²⁶ Standardization transforms the original distribution of the predicted probability into one with a mean of 0 and a standard deviation of 1. For example, if an observation in the 2016 test set has a final probability prediction of 0.45, the mean and standard deviation of the final probability prediction for all observations in the 2016 test set is 0.32 and 0.12 respectively, then the Predictive MAR Score for this observation is 1.08 $((0.45-0.32)/0.12)$. We adopt standardization as a rescaling method because it is relatively easy to interpret scores in a standardized distribution and to identify extreme values of the scores.

those that do not have MAR. In un-tabulated results, the PAQI generated under other misclassification costs are similar.

[Insert Figure 3 here]

Since PAQI is forward-looking in nature, scores are created for observations in the test sets (i.e., data from 2011 to 2017). PAQI indicates the likelihood of an audit engagement having a MAR based on information from IAQI. To assess whether the PAQI are indeed associated with actual MAR and whether the predictive scores have appropriately captured audit quality two validation tests are performed. First, in order to assess whether the PAQI has incremental information that is associated with MAR, we establish the following model:

$$MAR_{it} = \beta_0 + \beta_1 * PAQI_{it} + \mathbf{Controls}_{it} * \boldsymbol{\beta} + \varepsilon$$

MAR_{it} equals one if the financial statement for firm i in fiscal year t has submitted a material restatement due to GAAP violations or fraud.²⁷ $PAQI_{it}$ represents the PAQI generated from the research performed for firm i in fiscal year t . Meanwhile, we obtain control variables from related literature (See the footnotes of Table 10). If the PAQI can provide incremental information that is associated with an actual MAR, we expect parameter β_1 to be significantly positive (we use logistic regression to perform this test). We also control for year and industry fixed effects and calculate the standard errors using firm clusters. The results documented in Table 10 validate the postulation that the PAQI provide incremental information that is associated with an actual MAR and that the higher the score, the higher the likelihood of the existence of an actual MAR. *Ceteris paribus*, a 1-point increase in the PAQI increases the odds of having an actual MAR by 1.51 times.²⁸

²⁷ Consistent with the machine learning experiment, only starting-year MAR is used.

²⁸ The odds ratio is calculated as $e^{0.413}$.

To assess whether PAQI can capture audit quality, we establish the following model based on the well-recognized conclusion that Big 4 auditors provide higher audit quality (Eshleman and Guo 2014; DeFond, Erkens, and Zhang 2017; Jiang, Wang and Wang 2019).²⁹ We also expect the parameter γ_1 to be significantly negative.

$$PAQI_{it} = \gamma_0 + \gamma_1 * Big4_{it} + \mathbf{Controls}_{it} * \boldsymbol{\gamma} + \varepsilon$$

$PAQI_{it}$ represents the PAQI generated from the research performed for firm i in fiscal year t . The value of $Big4_{it}$ is one if firm i in fiscal year t was audited by a Big 4 auditor. Similar control variables are used in this model (See the footnotes of Table 10). We use OLS (Ordinary Least Squares) to perform this test while also controlling for year and industry fixed effects and calculating the standard errors using firm clusters. The results in Table 10 show that holding other factors constant, financial reports audited by Big 4 auditors have significantly lower PAQI compared to other engagements, thus confirming our expectations. From an economic perspective, *ceteris paribus*, financial reports audited by Big 4 auditors feature Predictive MAR Scores that are 0.315 points lower than those assessed by non-Big 4 auditors. The 0.315 points difference is economically significant, given that the difference of PAQI between observations that have MAR and those that do not is 0.49 (Panel B in Figure 3).

In summary, in this section, we aggregate IAQI into PAQI, and we show that PAQI provides incremental information that is associated with actual MAR.

[Insert Table 10 here]

²⁹ Although Big 4 auditor status is not chosen as a *predictive* factor of MAR, it is an important *explanatory* factor of audit quality. Please refer to Section 1 and Section 2.2 for further discussion about differences between predictive modeling and explanatory modeling. Additional details on these differences can be found in Shmueli (2010) and Shmueli and Koppius (2011).

VIII. DISCUSSION AND CONCLUSION

Over the past decades, mainstream accounting research has adopted explanatory modeling and identified a broad set of variables based on theories that are associated with audit quality. However, little is known about how predictive these theory-driven audit-related variables are in forecasting audit failure out-of-sample. Learning about the predictive power of audit-related variables can help assess the relevance of existing explanatory models by examining the distance between theory and practice (Shmueli 2010; Shmueli and Koppius 2011). Thus, by understanding the predictive power of audit-related variables, we can narrow down a list of audit-related variables that are the most relevant in forecasting audit failure. Understanding the predictive power can also facilitate comparisons of competing theories, different operationalizations of constructs, and different measurement instruments (Shmueli 2010; Shmueli and Koppius 2011). Our study builds upon previous literature to explore the predictive power of audit-related variables.

We contribute to the accounting and auditing literature by identifying a fundamental prediction question in audit quality research: which publicly available audit-related variables are the most predictive of audit failure? In seeking to provide solutions to this issue, we adopt predictive modeling with machine learning and use material annual restatements as the proxy for audit failure. We collect 31 publicly available audit-related variables as predictors for public U.S. firms spanning the period between 2005 and 2017. Based on feature subset selection results from five popular machine learning algorithms, we identify 11 audit-related variables as IAQI. We further validated their predictive power via cost-sensitive learning and the rolling-window prediction. The 11 IAQI are composed of variables that represent auditor characteristics, audit task characteristics, auditor-client contracting features, auditor communication, and the quality of the audited financial statements. These IAQI serve as a validation on the predictive power of the audit-

related variables examined in previous studies. When the objective is to predict MAR, IAQI are more predictive than other audit related variables. Although IAQI are predictive of material restatements, they cannot be treated as an absolute and definitive prediction of whether an audited financial report will eventually be materially restated. Instead, IAQI should be interpreted within specific contexts to assist stakeholders in predicting audit failure and facilitate relevant decision-making processes.

Overall, we set an example for future accounting research by providing methodologies that guide the adoption of predictive modeling in audit quality studies. A limitation for this paper is the usage of MAR as a proxy for audit failure, which may have measurement bias since not all audit failures produce a material restatement (Gaynor et al. 2016; Suresh and Guttag 2019). However, we try to mitigate such measurement bias by using a relatively large sample because the parameters in the model will converge with the expected correct value (Suresh and Guttag 2019). Furthermore, we cross validate the main findings by using alternative measures of audit failure and algorithm performance. Future research may seek to adopt other proxies of audit quality, such as those outlined in the Part 1 Findings of the PCAOB inspections and proposed in audit firms' internal assessments of audit quality. However, these proxies are inherently limited by their small sample size and representativeness. Therefore, researchers must balance the trade-offs of using alternative proxies of audit quality. Future research can also explore developing innovative measures of audit failure and utilize audit-related variables generated from unorthodox sources, such as social media and online forums. Our research also suggests future studies to examine whether material annual restatements are driven more by audit-related factors or clients' innate financial conditions.

REFERENCES

- AS 1001: Responsibilities and Functions of the Independent Auditor. Available at: <https://pcaobus.org/Standards/Auditing/Pages/AS1001.aspx>
- Anantharaman, D., and Wans, N. (2019). Audit office experience with SOX 404 (b) filers and SOX 404 audit quality. *The Accounting Review*, 94(4), 1-43.
- Asthana, S. C., and Boone, J. P. (2012). Abnormal audit fee and audit quality. *Auditing: A Journal of Practice and Theory*, 31(3), 1-22.
- Audit Analytics. (2020). 2019 Financial Restatements: A Nineteen Year Comparison. Available at: <http://auditanalytics-2019restatementreport-j.pagedemo.co/>
- Ahn, J., Hoitash, R., and Hoitash, U. (2020). Auditor task-specific expertise: The case of fair value accounting. *The Accounting Review*, 95(3), 1-32.
- Alpaydin, E. (2014). *Introduction to Machine Learning* (3rd ed.). The MIT Press.
- Aobdia, D. (2018). The impact of the PCAOB individual engagement inspection process—Preliminary evidence. *The Accounting Review*, 93(4), 53-80.
- Aobdia, D. (2019). Do practitioner assessments agree with academic proxies for audit quality? Evidence from PCAOB and internal inspections. *Journal of Accounting and Economics*, 67(1), 144-174.
- Balsam, S., Krishnan, J., and Yang, J. S. (2003). Auditor industry specialization and earnings quality. *Auditing: A journal of practice and Theory*, 22(2), 71-97.
- Bao, Y., Ke, B., Li, B., Yu, Y. J., and Zhang, J. (2020). Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach. *Journal of Accounting Research*.
- Bell, T. B., Causholli, M., and Knechel, W. R. (2015). Audit firm tenure, non-audit services, and internal assessments of audit quality. *Journal of Accounting Research*, 53(3), 461-509.
- Beneish, M. D. and Vorst, P., The Cost of Fraud Prediction Errors (January 31, 2020). Kelley School of Business Research Paper No. 2020-55, Available at SSRN: <https://ssrn.com/abstract=3529662> or <http://dx.doi.org/10.2139/ssrn.3529662>
- Bertomeu, J. (2020). Machine learning improves accounting: discussion, implementation and research opportunities. *Rev Account Stud*. <https://doi.org/10.1007/s11142-020-09554-9>
- Bertomeu, J., Cheynel, E., Floyd, E., and Pan, W. (2020). Using Machine Learning to Detect Misstatements. *Review of Accounting Studies*, Forthcoming, Available at SSRN: <https://ssrn.com/abstract=3496297>
- Bills, K. L., Cunningham, L. M., and Myers, L. A. (2016). Small audit firm membership in associations, networks, and alliances: Implications for audit quality and audit fees. *The Accounting Review*, 91(3), 767-792.
- Blankley, A. I., Hurtt, D. N., and MacGregor, J. E. (2012). Abnormal Audit Fees and Restatements. *Auditing: A Journal of Practice and Theory*, 31(1), 79–96.
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), 1145-1159.
- Brown, N. C., Crowley, R. M., and Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, 58(1), 237-291.
- Bhaskar, L. S., Schroeder, J. H., and Shepardson, M. L. (2019). Integration of internal control and financial statement audits: Are two audits better than one?. *The Accounting Review*, 94(2), 53-81.
- Carey, P., and Simnett, R. (2006). Audit partner tenure and audit quality. *The accounting review*, 81(3), 653-676.

- Causholli, M., and Knechel, W. R. (2012). An examination of the credence attributes of an audit. *Accounting Horizons*, 26(4), 631-656.
- Cao, Y., Myers, L. A., and Omer, T. C. (2012). Does company reputation matter for financial reporting quality? Evidence from restatements. *Contemporary Accounting Research*, 29(3), 956-990.
- Cao, J., Chen, F., and Higgs, J. L. (2016). Late for a very important date: financial reporting and audit implications of late 10-K filings. *Review of Accounting Studies*, 21(2), 633-671.
- Cecchini, M., Aytug, H., Koehler, G. J., and Pathak, P. (2010). Detecting Management Fraud in Public Companies. *Management Science*, 56(7), 1146–1160.
- Center of Audit Quality. (2013). Financial Restatement Trends in the United States: 2003-2012. Available at: <https://www.thecaq.org/wp-content/uploads/2019/03/financial-restatement-trends-in-the-united-states-2003-2012.pdf>
- Chan, P. K., and Stolfo, S. J. (1998, August). Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. In *KDD* (Vol. 1998, pp. 164-168).
- Choi, J. H., Kim, C., Kim, J. B., and Zang, Y. (2010). Audit office size, audit quality, and audit pricing. *Auditing: A Journal of practice and theory*, 29(1), 73-97.
- Christensen, B. E., Glover, S. M., Omer, T. C., and Shelley, M. K. (2016). Understanding audit quality: Insights from audit professionals and investors. *Contemporary Accounting Research*, 33(4), 1648-1684.
- Cohen, J. R., Hoitash, U., Krishnamoorthy, G., and Wright, A. M. (2014). The effect of audit committee industry expertise on monitoring the financial reporting process. *The Accounting Review*, 89(1), 243-273.
- Cunningham, L. M., Li, C., Stein, S. E., and Wright, N. S. (2019). What's in a name? Initial evidence of US audit partner identification using difference-in-differences analyses. *The Accounting Review*, 94(5), 139-163.
- Dash, M., and Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, 1(1-4), 131-156.
- DeAngelo, L. E. (1981). Auditor size and audit quality. *Journal of accounting and economics*, 3(3), 183-199.
- Dechow, P. M., and Dichev, I. D. (2002). The quality of accruals and earnings: The role of accrual estimation errors. *The accounting review*, 77(s-1), 35-59.
- Dechow, P. M., Sloan, R. G., and Sweeney, A. P. (1995). Detecting earnings management. *Accounting review*, 193-225.
- Dechow, P. M., Ge, W., Larson, C. R., and Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary accounting research*, 28(1), 17-82.
- DeFond, M., and Zhang, J. (2014). A Review of Archival Auditing Research. *Journal of Accounting and Economics*, 58(2–3), 275–326.
- DeFond, M., Erkens, D. H., and Zhang, J. (2017). Do client characteristics really drive the Big N audit quality effect? New evidence from propensity score matching. *Management Science*, 63(11), 3628-3649.
- Ding, K., Peng, X., and Wang, Y. (2019). A machine learning-based peer selection method with financial ratios. *Accounting Horizons*, 33(3), 75-87.
- Ding, K., Lev, B., Peng, X., Sun, T., and Vasarhelyi, M. A. (2020). Machine learning improves accounting estimates: evidence from insurance payments. *Review of Accounting Studies*, 1-37.

- Dutta, I., Dutta, S., and Raahemi, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems with Applications*, 90, 374–393.
- Eshleman, J. D., and Guo, P. (2014). Do Big 4 Auditors Provide Higher Audit Quality After Controlling for the Endogenous Choice of Auditor? *Auditing: A Journal of Practice and Theory*, 33(4), 197–220.
- Elkan, C. (2001, August). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence (Vol. 17, No. 1, pp. 973-978)*. Lawrence Erlbaum Associates Ltd.
- Ettredge, M., Fuerherm, E. E., and Li, C. (2014). Fee pressure and audit quality. *Accounting, Organizations and Society*, 39(4), 247-263.
- Ferguson, A., Francis, J. R., and Stokes, D. J. (2003). The effects of firm-wide and office-level industry expertise on audit pricing. *The accounting review*, 78(2), 429-448.
- Financial Times (2020). After Wirecard: is it time to audit the auditors? Available at: <https://www.ft.com/content/b220719a-edca-4ebf-b6bc-5f7a67078745>
- Francis, J. R. (2004). What do we know about audit quality?. *The British accounting review*, 36(4), 345-368.
- Francis, J. R., and Yu, M. D. (2009). Big 4 Office Size and Audit Quality. *The Accounting Review*, 84(5), 1521–1552.
- Francis, J. R. (2011). A framework for understanding and researching audit quality. *Auditing: A journal of practice & theory*, 30(2), 125-152.
- Francis, J. R., and Michas, P. N. (2013). The contagion effect of low-quality audits. *The Accounting Review*, 88(2), 521-552.
- Francis, J. R., Michas, P. N., and Yu, M. D. (2013). Office size of Big 4 auditors and client restatements. *Contemporary Accounting Research*, 30(4), 1626-1661.
- Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119-139.
- Gaynor, L. M., Kelton, A. S., Mercer, M., and Yohn, T. L. (2016). Understanding the relation between financial reporting quality and audit quality. *Auditing: A Journal of Practice & Theory*, 35(4), 1-22.
- Gentry, J. A., Shaw, M. J., Tessmer, A. C., and Whitford, D. T. (2002). Using inductive learning to predict bankruptcy. *Journal of Organizational Computing and Electronic Commerce*, 12(1), 39-57.
- Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182.
- Hastie, T., Tibshirani, R., and Tibshirani, R. J. (2017a). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692.
- Hastie, T., Tibshirani, R., and Friedman, J. (2017b). *The elements of statistical learning*. New York: Springer series in statistics.
- Hall, M. A., and Smith, L. A. (1998). Practical feature subset selection for machine learning.
- Hayes, B. L. and Boritz, E. (2019). *Classifying Restatements: An Application of Machine Learning and Textual Analytics*. Working paper. University of Guelph. Available at SSRN: <https://ssrn.com/abstract=2716166> or <http://dx.doi.org/10.2139/ssrn.2716166>
- Hocking, R. R., and Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics*, 9(4), 531-540.
- Hunt, E., Hunt, J., Richardson, V. (2019). *Predicting Accounting Misstatements Using Machine Learning*. Mississippi State University. Working paper.

- Huang, Y., and Scholz, S. (2012). Evidence on the association between financial restatements and auditor resignations. *Accounting Horizons*, 26(3), 439-464.
- Honigsberg, C., Rajgopal, S., and Srinivasan, S. (2019). The Changing Landscape of Auditor Liability. *Journal of Law and Economics*, Forthcoming.
- Jeni, L. A., Cohn, J. F., and De La Torre, F. (2013, September). Facing imbalanced data-- recommendations for the use of performance metrics. In 2013 Humaine association conference on affective computing and intelligent interaction (pp. 245-251). IEEE.
- Jiang, J., Wang, I. Y., and Wang, K. P. (2019). Big N auditors and audit quality: New evidence from quasi-experiments. *The Accounting Review*, 94(1), 205-227.
- Johnson, V. E., Khurana, I. K., and Reynolds, J. K. (2002). Audit-firm tenure and the quality of financial reports. *Contemporary accounting research*, 19(4), 637-660.
- Jones, S. (2017). Corporate bankruptcy prediction: a high dimensional analysis. *Review of Accounting Studies*, 22(3), 1366-1422.
- Karpoff, J. M., Koester, A., Lee, D. S., and Martin, G. S. (2017). Proxies and databases in financial misconduct research. *The Accounting Review*, 92(6), 129-163.
- Kinney, W. R., Palmrose, Z.-V., and Scholz, S. (2004). Auditor Independence, Non-Audit Services, and Restatements: Was the U. S. Government Right? *Journal of Accounting Research*, 42(3), 561-588.
- Knechel, W. R., and Vanstraelen, A. (2007). The relationship between auditor tenure and audit quality implied by going concern opinions. *AUDITING: A journal of practice and theory*, 26(1), 113-131.
- Knechel, W. R., and Sharma, D. S. (2012). Auditor-provided nonaudit services and audit effectiveness and efficiency: Evidence from pre-and post-SOX audit report lags. *Auditing: A Journal of Practice and Theory*, 31(4), 85-114.
- Knechel, W. R., Krishnan, G. V., Pevzner, M., Shefchik, L. B., and Velury, U. K. (2013). Audit quality: Insights from the academic literature. *Auditing: A Journal of Practice and Theory*, 32(sp1), 385-421.
- Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5), 491-95.
- Kothari, S. P., Leone, A. J., and Wasley, C. E. (2005). Performance matched discretionary accrual measures. *Journal of accounting and economics*, 39(1), 163-197.
- Krishnan, J., and Krishnan, J. (1997). Litigation risk and auditor resignations. *Accounting Review*, 539-560.
- Lambert, T. A., Jones, K. L., Brazel, J. F., and Showalter, D. S. (2017). Audit time pressure and earnings quality: An examination of accelerated filings. *Accounting, Organizations and Society*, 58, 50-66.
- Lennox, C. S. (2016). Did the PCAOB's restrictions on auditors' tax services improve audit quality?. *The Accounting Review*, 91(5), 1493-1512.
- Lennox, C., and Li, B. (2020). When Are Audit Firms Sued for Financial Reporting Failures and What Are the Lawsuit Outcomes?. *Contemporary Accounting Research*.
- Leuz, C., Nanda, D., and Wysocki, P. D. (2003). Earnings management and investor protection: an international comparison. *Journal of financial economics*, 69(3), 505-527.
- Lobo, G. J., and Zhao, Y. (2013). Relation between audit effort and financial report misstatements: Evidence from quarterly and annual restatements. *The Accounting Review*, 88(4), 1385-1412

- López, D. M., and Peters, G. F. (2012). The effect of workload compression on audit quality. *Auditing: A Journal of Practice and Theory*, 31(4), 139-165.
- Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research*, 48(5), 1049-1102.
- Li, L., Qi, B., Tian, G., and Zhang, G. (2017). The contagion effect of low-quality audits at the level of individual auditors. *The Accounting Review*, 92(1), 137-163.
- Lim, C. Y., and Tan, H. T. (2008). Non-audit service fees and audit quality: The impact of auditor specialization. *Journal of accounting research*, 46(1), 199-246.
- Lim, C. Y., Tan, H. T., and Cheng, Q. (2010). Does Auditor Tenure Improve Audit Quality? Moderating Effects of Industry Specialization and Fee Dependence. *Contemporary Accounting Research*, 27(3), 923.
- McNichols, M. F. (2002). Discussion of the quality of accruals and earnings: The role of accrual estimation errors. *The accounting review*, 77(s-1), 61-69.
- Myers, J. N., Myers, L. A., and Omer, T. C. (2003). Exploring the term of the auditor-client relationship and the quality of earnings: A case for mandatory auditor rotation?. *The accounting review*, 78(3), 779-799.
- Newton, N. J., Wang, D., and Wilkins, M. S. (2013). Does a lack of choice lead to lower quality? Evidence from auditor competition and client restatements. *Auditing: A Journal of Practice and Theory*, 32(3), 31-67.
- Rapach, D. E., and Wohar, M. E. (2006). In-sample vs. out-of-sample tests of stock return predictive power in the context of data mining. *Journal of Empirical Finance*, 13(2), 231-247.
- Rajgopal, S., Srinivasan, S. and Zheng, X. (2021). Measuring audit quality. *Rev Account Stud.* <https://doi.org/10.1007/s11142-020-09570-9>
- Reber, R., Schwarz, N., and Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?. *Personality and social psychology review*, 8(4), 364-382.
- Reichelt, K. J., and Wang, D. (2010). National and office-specific measures of auditor industry expertise and effects on audit quality. *Journal of Accounting Research*, 48(3), 647-686.
- Rijsbergen, V. and Joost, C. (2004). *The geometry of information retrieval*. Cambridge University Press.
- Romanus, R. N., Maher, J. J., and Fleming, D. M. (2008). Auditor Industry Specialization, Auditor Changes, and Accounting Restatements. *Accounting Horizons*, 22(4), 389-413.
- Ruddock, C., Taylor, S. J., and Taylor, S. L. (2006). Nonaudit services and earnings conservatism: Is auditor independence impaired?. *Contemporary Accounting Research*, 23(3), 701-746.
- Paterson, J. S., and Valencia, A. (2011). The effects of recurring and nonrecurring tax, audit-related, and other nonaudit services on auditor independence. *Contemporary Accounting Research*, 28(5), 1510-1536.
- Panel on Audit Effectiveness. 2000. Report and recommendations: Public oversight board of the American Institute of Certified Public Accountants. Available at: <http://www.pobauditpanel.org/download.html>
- PCAOB. (2013). Standing Advisory Group Meeting, Discussion-Audit Quality Indicators. Available at:

- https://pcaobus.org/news/events/documents/05152013_sagmeeting/audit_quality_indicators.pdf
- PCAOB. (2015). Concept Release on Audit Quality Indicators. Available at: https://pcaobus.org/Rulemaking/Docket%20041/Release_2015_005.pdf
- PCAOB. (2019). Strategic Plan 2019-2023. Available at: <https://pcaobus.org/About/Administration/Documents/Strategic%20Plans/Strategic%20Plan-2019-2023.pdf>
- Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice and Theory*, 30(2), 19-50.
- Perols, J. L., Bowen, R. M., Zimmermann, C., and Samba, B. (2016). Finding needles in a haystack: Using data analytics to improve fraud prediction. *The Accounting Review*, 92(2), 221-245.
- Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), e0118432.
- Schroeder, J. H. (2016). The impact of audit completeness and quality on earnings announcement GAAP disclosures. *The Accounting Review*, 91(2), 677-705.
- SEC. (2017). How Investigations Work. Available at: <https://www.sec.gov/enforce/how-investigations-work.html>
- Stanley, J. D., and DeZoort, F. T. (2007). Audit firm tenure and financial restatements: An analysis of industry specialization and fee effects. *Journal of Accounting and Public Policy*, 26(2), 131-159.
- Shmueli, G. (2010). To explain or to predict?. *Statistical science*, 25(3), 289-310.
- Shmueli, G., and Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS quarterly*, 553-572.
- Suresh, H., and Gutttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.
- Sun, T. 2018. The incremental informativeness of the sentiment of conference calls for internal control material weaknesses. *Journal of Emerging Technologies in Accounting* 15 (1): 11–27. <https://doi.org/10.2308/jeta-51969>
- Sun, T., and L. J. Sales. 2018. Predicting public procurement irregularity: An application of neural networks. *Journal of Emerging Technologies in Accounting* 15 (1): 141–154. <https://doi.org/10.2308/jeta-52086>
- Tan, C. E., and Young, S. M. (2015). An analysis of “Little r” restatements. *Accounting Horizons*, 29(3), 667-693.
- Tang, J., Alelyani, S., and Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, 37.
- Thai-Nghe, N., Gantner, Z., and Schmidt-Thieme, L. (2010, July). Cost-sensitive learning methods for imbalanced data. In *The 2010 International joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE.
- Wall Street Journal (2016). Wells Fargo: Where Was the Auditor? Available at: <https://www.wsj.com/articles/wells-fargo-where-was-the-auditor-1478007838>
- Zhao, Y., Bedard, J. C., and Hoitash, R. (2017). SOX 404, auditor effort, and the prevention of financial report misstatements. *Auditing: A Journal of Practice and Theory*, 36(4), 151-177.

TABLES AND FIGURES

Table 1. Informative Audit Quality Indicators (IAQI)

Category	Sub-category	Aspects Captured	Variable	Measurement
Audit input	Auditor characteristics	Competence, Resource, Independence	Office Size	Natural logarithm of one plus total annual audit fees of an audit office (Aobdia 2019)
		Informational advantage, Independence	Tenure	Number of years that the company is audited by the same audit firm (Bell et al. 2015)
Audit process	Task characteristics	Audit effort, Audit efficiency	Audit Report Lag	Natural logarithm of the number of days between fiscal year-end and the signature date of audit opinion (Lobo and Zhao 2013)
		Audit effort, Workload	Integrated Audit	Indicator variable equal to 1 when the audit engagement is an integrated audit of financial statements and internal controls, and 0 otherwise (Aobdia 2019)
Audit output	Auditor-client contracting features	Incentive	Auditor Resignation	Indicator variable equal to 1 if the current auditor will resign (instead of being dismissed by the company) from the next fiscal year, 0 otherwise (Krishnan and Krishnan 1997)
		Audit effort	Audit Fees	Natural logarithm of 1 plus the audit fees charged to the auditee (Aobdia 2019)
		Competence, Independence	Internal Control Weakness	Indicator variable equal to 1 if a material weakness is reported for the year, 0 otherwise (Aobdia 2019)
		Disc. Accruals		Residual from the cross-sectional modified Jones model in Aobdia (2019)
Audit output	Quality of the audited financial statements	Within-GAAP manipulation	Abs (Disc. Accruals)	The absolute value of Disc. Accruals (Aobdia 2019)
			Abs (Accruals)	The absolute value of accruals deflated by beginning assets (Aobdia 2019)
			Abs (Accruals/CFO)	The absolute value of accruals deflated by cash flow from operations (Aobdia 2019)

Table 2. Theory-Driven Audit-Related Variables (ARV)

No.	Variable	Description	Main source	Category	Sub-Category	Aspects Captured
1	Industry Specialization_National	Auditor's annual market share based on audit fees within a two-digit SIC category (Aobdia 2019)	Aobdia (2019); Balsam et al. (2003); Reichelt and Wang (2010); Romanus et al. (2008)	Audit input	Auditor characteristics	Competence, Informational advantage
2	Industry Specialization_MSA	Auditor's annual market share based on audit fees within a two-digit SIC category for a particular Metropolitan Statistical Area (MSA) ³⁰ (Reichelt and Wang 2010)	Reichelt and Wang (2010); Ferguson et al. (2003); Ahn et al. (2020)	Audit input	Auditor characteristics	Competence, Informational advantage
3	Office Size	Natural logarithm of one plus total annual audit fees of an audit office (Aobdia 2019)	Aobdia (2019); Choi et al. (2010); Francis and Yu (2009)	Audit input	Auditor characteristics	Competence, Resource, Independence
4	Big 4	Indicator variable equal to one if the audit firm is a Big 4, and 0 otherwise (Aobdia 2019)	Lobo and Zhao (2013); Newton et al. (2013); Eshleman and Guo (2014); Defond et al. (2017)	Audit input	Auditor characteristics	Competence, Resource, Independence
5	New Client (or Auditor Change)	Indicator variable equal to 1 if the auditor-client relationship is in its first year, and 0 otherwise (Aobdia 2019)	Aobdia (2019); Francis et al. (2013); Cohen et al. (2014); Schroeder (2016)	Audit process	Task characteristic	Informational advantage, Independence
6	Tenure	Number of years that the company is audited by the same audit firm (Bell et al. 2015)	Myers et al. (2003); Johnson et al. (2002); Knechel and Vanstraelen (2007); Lim et al. (2010); Bell et al. (2015); Stanley and DeZoort (2007); Francis and Yu (2009); Lobo and Zhao (2013)	Audit process	Task characteristic	Informational advantage, Independence
7	Local Auditor_MSA	Indicator variable equal to 1 if the audit engagement office is located in the same MSA where audit clients are headquartered, and 0 otherwise (Choi et al. 2012)	Choi et al. (2012); Francis et al. (2013)	Audit process	Task characteristic	Informational advantage
8	Integrated Audit	Indicator variable equal to 1 when the audit is an integrated audit of financial statements and internal controls, and 0 otherwise (Aobdia 2019)	Aobdia (2019); Zhao et al. (2017); Bhaskar et al. (2019)	Audit process	Task characteristic	Audit effort, Workload
9	Accelerated Filer	Indicator variable equal to 1 for firms that are accelerated filers, and 0 otherwise (Newton et al. 2013)	Lambert et al. (2017); Newton et al. (2013)	Audit process	Task characteristic	Time pressure, Workload
10	Busy	Indicator variable equal to 1 if a company has a fiscal year-end date of December, and 0 otherwise (Lopez and Peters 2012)	Lopez and Peters (2012); Lobo and Zhao (2013)	Audit process	Environmental characteristics	Time pressure, Workload

³⁰ According to Reichelt and Wang (2010), the geographical city available from Audit Analytics is not the MSA. MSA information is available from the U.S. Census Bureau.

11	Workload Compression	The relative level of workload compression of an auditor office during the fiscal year-end month of the auditee ³¹ (Lopez and Peters 2012)	Lopez and Peters (2012)	Audit process	Environmental characteristics	Workload
12	Auditor Competition_MS A	MSA-level auditor concentration based on Herfindahl index. Details are provided in (Newton et al. 2013)	Newton et al. (2013)	Audit process	Environmental characteristics	Incentive
13	Auditor Resignation	Indicator variable equal to 1 if the current auditor will resign (instead of being dismissed by the company) from the next fiscal year, 0 otherwise (Krishnan and Krishnan 1997)	Krishnan and Krishnan (1997); Huang and Scholz (2012);	Audit process	Auditor-client contracting features	Incentive
14	Audit Fees	Natural logarithm of 1 plus the audit fees charged to the auditee (Aobdia 2019)	Aobdia (2019); Paterson and Valencia (2011); Cao et al. (2012); Francis et al. (2013); Lobo and Zhao (2013); Newton et al. (2013); Eshleman and Guo (2014); Cohen et al. (2014)	Audit process	Auditor-client contracting features	Audit effort
15	Tax Fee	Natural logarithm of 1 plus the total tax fees charged to the auditee	Kimney et al. (2004); Paterson and Valencia (2011); Lennox (2016)	Audit process	Auditor-client contracting features	Independence, Informational advantage
16	Audit-Related Fee	Natural logarithm of 1 plus the audit-related fees charged to the auditee	Kimney et al. (2004); Paterson and Valencia (2011)	Audit process	Auditor-client contracting features	Independence, Informational advantage
17	Other Fees	Natural logarithm of 1 plus the other fees charged to the auditee	Kimney et al. (2004); Paterson and Valencia (2011)	Audit process	Auditor-client contracting features	Independence, Informational advantage
18	Non-Audit Fee Ratio	Non-audit fees deflated by total fees paid (audit plus non-audit fees). Non-audit fee equals to the sum of benefit fee, IT fee, Tax fee, audit related fee, and other fees. (Ruddock et al. 2006)	Lim and Tan (2008); Shindih and Gul (2007); Ruddock et al. (2006); Cao et al. (2012); Newton et al. (2013); Cohen et al. (2014)	Audit process	Auditor-client contracting features	Independence, Informational advantage
19	Influence	The ratio of a company's total fees (i.e., audit fees plus non-audit fees) relative to the aggregate annual total fees generated by the local office that audits the company (Lopez and Peters 2012)	Lopez and Peters (2012); Francis and Yu (2009)	Audit process	Auditor-client contracting features	Independence
20	Abnormal Audit Fee	The unscaled residual from the audit fee model used in Blankley et al. (2012) ³²	Blankley et al. (2012); Asthana and Boone (2012); Lobo and Zhao (2013); Schroeder (2016)	Audit process	Auditor-client contracting features	Abnormal audit effort
21	Audit Report Lag	Natural logarithm of the number of days between fiscal year-end and the signature date of audit opinion (Lobo and Zhao 2013)	Knechel and Sharma (2012); Lobo and Zhao (2013)	Audit process	Task characteristic	Audit effort, Audit efficiency

³¹ “For each month, we add the audit fees charged to clients with the same fiscal year-end month in each local office; we then divide each monthly sum by the total audit fees collected by the local office for the year.” (Lopez and Peters 2012)

³² We base the model on Blankley et al. (2012) but use the definition of IC_Weak from Newton et al. (2013).

22	Non-timely Issuance of 10-K Due to Audit	Indicator variable equal to 1 if the company filed 10-K late and the lateness is due to audit, 0 otherwise	Cao et al. (2016); Lambert et al. (2017); Wang et al. (2013)	Audit process	Task characteristic	Audit effort, Audit efficiency
23	Going Concern	Indicator variable equal to 1 if auditor gave a going concern opinion (Aobdia 2019)	Aobdia (2019); Carey and Simnett (2006); Lobo and Zhao (2013); Ettredge et al. (2014); Lemox (2016)	Audit output	Auditor communication	Independence, Competence
24	Internal Control Weakness	Indicator variable equal to 1 if a material weakness is reported for the year, 0 otherwise (Aobdia 2019) ³³	Aobdia (2019); Anantharaman and Wans (2019)	Audit output	Auditor communication	Independence, Competence
25	Disc. Accruals	Residual from the cross-sectional modified Jones model in Aobdia (2019)	Aobdia (2019); Dechow et al. (1995); Kothari et al. (2005); Reichelt and Wang (2010)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
26	Abs (Disc. Accruals)	Absolute value of Disc. Accruals (Aobdia 2019)	Aobdia (2019); Dechow et al. (1995); Kothari et al. (2005); Reichelt and Wang (2010); Bills et al. (2016); Krishnan, Krishnan, and Song (2017).	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
27	DD Residual	Residual from the Dechow and Dichev model in Aobdia (2019)	Aobdia (2019); Dechow and Dichev (2002); McNichols (2002)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
28	Abs (Accruals)	The absolute value of accruals deflated by beginning assets (Aobdia 2019)	Aobdia (2019); Leuz et al. (2003)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
29	Abs (Accruals/CFO)	The absolute value of accruals deflated by cash flow from operations (Aobdia 2019)	Aobdia (2019); Leuz et al. (2003)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
30	Small Profit	Indicator variable equal to 1 if ROA is between 0% and 3%, 0 otherwise (Aobdia 2019)	Aobdia (2019); Francis and Yu (2009)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
31	Prior ROA Meet	Indicator variable equal to 1 if the year-on-year change in ROA is between 0% and 1%, 0 otherwise (Aobdia 2019)	Aobdia (2019)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation

³³ Here, a value of 1 indicates that the audit engagement is an integrated audit and the auditor expressed weakness in ICFR.

Table 3. Sample Determination

	Number of firm-year observations
COMPUSTAT population from 2000 to 2019	224,047
Less: observations without CIK	-25,279
Less: observations that do not file 10K	-16,615
Less: duplicated CIK and Fiscal Year-End	-23,761
Match with Audit Analytics by fiscal year-end	
Less: missing audit fee records	-48,604
Less: audit fees reported in foreign currencies	-4,723
Less: observations without going concern opinion records	-4,105
Less: observations with foreign auditors or business	-10,933
Less: observations without SIC code	-657
Less: observations without MSA information	-9,497
Less: observations without abnormal audit fee	-31,696
Less: observations without audit fee lag	-59
Less: missing auditor resigned data	-520
Less: observations with infinite values of Industry Specialization_MSA	-3
Less: observations with infinite values of workload compression	-3
Less: observations without discretionary accruals	-5,167
Less: observations with missing DD residual variable	-6,609
Less: observations outside 2005 to 2017	-8,975
Less: observations that are not first-year restatements in serial restatements	<u>-502</u>
Final sample size	26,339

Table 4. Sample Distribution by Years

Fiscal Year	Number of Firms	Number of Starting year MAR	Percentage
2005	2188	69	3.15%
2006	2066	48	2.32%
2007	1864	31	1.66%
2008	2054	44	2.14%
2009	1894	27	1.43%
2010	1981	45	2.27%
2011	1946	35	1.80%
2012	1963	35	1.78%
2013	2076	31	1.49%
2014	2064	26	1.26%
2015	2049	25	1.22%
2016	2087	11	0.53%
2017	2107	19	0.90%
Total	26339	446	1.69%

*Note: MAR is material annual restatement due to GAAP violations or fraud.

Table 5. Descriptive Statistics*

Variable	Mean	Sfd. Dev.	Min	Max	Comparable research, if any
Industry	0.17	0.16	0.00	1.00	Aobdia (2019)
Specialization_National	0.44	0.37	0.00	1.00	Choi et al. (2012); Francis et al. (2013); Lopez and Peters (2012)
Industry					
Specialization_MSA	0.44	0.37	0.00	1.00	Choi et al. (2012); Francis et al. (2013); Lopez and Peters (2012)
Office Size	16.19	2.14	8.01	20.22	Aobdia (2019); Newton et al. (2013); Lopez and Peters (2012)
Big 4	0.63	0.48	0.00	1.00	Aobdia (2019); Newton et al. (2013); Choi et al. (2012)
New Client	0.11	0.31	0.00	1.00	Aobdia (2019)
Tenure	6.19	4.26	1.00	18.00	Lobo and Zhao (2013); Bell et al. (2015)
Local Auditor_MSA	0.68	0.47	0.00	1.00	Choi et al. (2012); Francis et al. (2013)
Integrated Audit	0.63	0.48	0.00	1.00	Aobdia (2019); Zhao et al. (2017)
Accelerated Filer	0.63	0.48	0.00	1.00	Newton et al. (2013)
Busy	0.75	0.43	0.00	1.00	Aobdia (2019); Lopez and Peters (2012); Blankley et al. (2012)
Workload Compression	0.67	0.33	0.00	1.00	Lopez and Peters (2012)
Auditor Competition_MSA	0.28	0.14	0.09	1.00	Competition level higher than Newton et al. (2013), maybe because of differences in the sample period.
Auditor Resignation	0.02	0.12	0.00	1.00	Huang and Scholz (2012)
Audit Fees	13.43	1.54	0.00	18.23	Aobdia (2019); Newton et al. (2013); Francis et al. (2013)
Tax Fee	7.17	5.67	0.00	16.86	Pateron and Valencia (2011)
Audit-Related Fee	5.91	5.73	0.00	17.91	Pateron and Valencia (2011)
Other Fees	2.73	4.39	0.00	16.97	Pateron and Valencia (2011)
Non-Audit Fee Ratio	0.13	0.14	0.00	1.00	Lower non-audit fee ratio compared to Newton et al. (2013) maybe because of differences in the sample period.
Influence	0.16	0.24	0.00	1.00	Asthana and Boone (2012); Lopez and Peters (2012)
Abnormal Audit Fee	0.01	0.62	-15.86	2.76	Asthana and Boone (2012); Blankley et al. (2012); Lobo and Zhao (2013)
Audit Report Lag	4.23	0.30	0.00	7.34	Asthana and Boone (2012); Lopez and Peters (2012)
Non-timely Issuance of 10K_Due to Audit	0.01	0.10	0.00	1.00	Lower value compared to Cao et al. (2016) due to our revised definition of the variable
Going Concern	0.11	0.32	0.00	1.00	Minutti-Meza (2013)
Internal Control Weakness	0.03	0.16	0.00	1.00	Aobdia (2019); Newton et al. (2013)
Disc. Accruals	0.00	0.87	-12.62	12.22	Aobdia (2019); Reichelt and Wang (2009); Francis and Yu (2009)
Abs (Disc. Accruals)	0.29	0.83	0.00	12.62	Francis and Yu (2009)
Abs (Accruals)	0.34	1.35	0.00	10.99	Lower compared to Aobdia (2019) maybe because of sample composition
Abs (Accruals/CFO)	1.78	3.72	0.00	24.39	Aobdia (2019)
DD Residual	0.06	0.12	0.00	1.41	Aobdia (2019)
Small Profit	0.07	0.25	0.00	1.00	Lower compared to Aobdia (2019) maybe because of sample composition
Prior ROA Meet	0.01	0.11	0.00	1.00	Lower compared to Aobdia (2019) maybe because of sample composition

*Note: the number of observations for each variable is 26339.

Table 6. Feature Subset Selection Results

Variables	AB	ANN	SVM	RF	LR	Total
Industry Specialization_National	0	0	0	1	0	1
Industry Specialization_MSA	0	0	1	1	0	2
Office Size	0	1	1	1	1	4
Big 4	0	0	0	1	0	1
New Client	0	0	0	1	1	2
Tenure	1	1	1	1	1	5
Local Auditor_MSA	0	0	0	1	0	1
Integrated Audit	1	1	0	1	1	4
Accelerated Filer	0	1	1	0	0	2
Busy	0	1	0	1	0	2
Workload Compression	1	0	0	0	0	1
Auditor Competition_MSA	0	0	1	0	1	2
Auditor Resignation	1	1	0	1	1	4
Audit Fees	1	0	0	1	1	3
Tax Fee	1	0	0	0	0	1
Audit-Related Fee	0	1	0	0	0	1
Other Fees	0	0	0	1	0	1
Non-Audit Fee Ratio	0	0	0	1	1	2
Influence	1	0	0	1	0	2
Abnormal Audit Fee	0	0	0	1	0	1
Audit Report Lag	1	0	0	1	1	3
Non-timely Issuance of 10K_Due to Audit	0	1	0	0	0	1
Going Concern	0	1	0	0	1	2
Internal Control Weakness	1	1	0	1	1	4
Disc. Accruals	1	1	0	1	0	3
Abs (Disc. Accruals)	1	1	0	0	1	3
Abs (Accruals)	1	1	1	0	1	4
Abs (Accruals/CFO)	1	1	0	1	1	4
DD Residual	0	1	0	1	0	2
Small Profit	0	0	0	0	0	0
Prior ROA Meet	0	0	0	1	0	1

Note:

AB is AdaBoost. ANN is Artificial Neural Network. SVM is Support Vector Machine. RF is Random Forest. LR is Logistic Regression. The definition of variables is provided in Table 1. The variables that are selected by more than or equal to 3 algorithms in the feature subset selection are highlighted in bold.

Table 7. Descriptive Statistics of Average AUC

	AB	ANN	LR	RF	SVM	Overall
Mean	0.631	0.629	0.627	0.616	0.601	0.621
Median	0.635	0.630	0.626	0.624	0.610	0.626
Min	0.610	0.625	0.625	0.542	0.507	0.507
Max	0.645	0.631	0.628	0.627	0.629	0.645
Std. Dev.	0.010	0.001	0.001	0.021	0.035	0.021

Note:

AB is AdaBoost. ANN is Artificial Neural Network. SVM is Support Vector Machine. RF is Random Forest. LR is Logistic Regression.

Table 8. Tukey HSD Connected Letter Report for Average AUC

Misclassification Cost	AB	ANN	LR	RF	SVM
1	C (0.627)	C (0.626)	C (0.628)	B (0.542)	A (0.507)
5	BC (0.621)	C (0.629)	C (0.628)	B (0.609)	A (0.536)
10	C (0.633)	C (0.630)	C (0.628)	B (0.608)	A (0.581)
15	C (0.636)	BC (0.629)	BC (0.627)	AB (0.619)	A (0.608)
20	C (0.638)	BC (0.628)	BC (0.627)	AB (0.618)	A (0.607)
25	C (0.644)	B (0.629)	B (0.627)	B (0.624)	A (0.608)
30	B (0.645)	A (0.629)	A (0.627)	A (0.626)	A (0.622)
35	B (0.640)	A (0.630)	A (0.627)	A (0.624)	A (0.629)
40	B (0.639)	A (0.630)	A (0.626)	A (0.621)	A (0.623)
45	B (0.640)	A (0.630)	A (0.626)	A (0.624)	A (0.623)
50	B (0.637)	AB (0.630)	A (0.626)	A (0.622)	A (0.626)
60	A (0.630)	A (0.631)	A (0.626)	A (0.627)	A (0.628)
70	A (0.625)	A (0.630)	A (0.626)	A (0.624)	A (0.627)
80	AB (0.622)	B (0.630)	B (0.626)	B (0.625)	A (0.612)
90	B (0.615)	C (0.630)	BC (0.626)	BC (0.624)	A (0.593)
100	B (0.610)	C (0.630)	C (0.625)	C (0.625)	A (0.592)

Note:

The average AUC is presented in parentheses. The Tukey HSD connected letter report ranks the average AUC of algorithms alphabetically, with A indicating the lowest value. The difference in AUC between two algorithms is significant when they are ranked by different letters. AB is AdaBoost. ANN is Artificial Neural Network. SVM is Support Vector Machine. RF is Random Forest. LR is Logistic Regression.

Table 9

Panel A. Raw Financial Variables*

28 Raw financial variables from Bao et al. (2020), Dechow et al. (2011) and Cecchini et al. (2010)	
<i>Cash and Short-Term Investments</i>	<i>Common/Ordinary Equity - Total</i>
Receivables - Total	<i>Preferred/Preference Stock (Capital) - Total</i>
<i>Inventories - Total</i>	<i>Retained Earnings</i>
Short-Term Investments - Total	<i>Sales/Turnover (Net)</i>
<i>Current Assets - Total</i>	<i>Cost of Goods Sold</i>
Property, Plant and Equipment - Total (Gross)	<i>Depreciation and Amortization</i>
<i>Investment and Advances - Other</i>	<i>Interest and Related Expense - Total</i>
<i>Assets - Total</i>	<i>Income Taxes - Total</i>
Accounts Payable - Trade	<i>Income Before Extraordinary Items</i>
Debt in Current Liabilities - Total	<i>Net Income (Loss)</i>
Income Taxes Payable	<i>Long-Term Debt - Issuance</i>
Current Liabilities - Total	<i>Sale of Common and Preferred Stock</i>
Long-Term Debt - Total	<i>Price Close - Annual - Calendar</i>
Liabilities - Total	<i>Common Shares Outstanding</i>
5 Raw financial variables from Perols (2011) and Perols et al. (2017)	
Common Shares Issued	<i>Property, Plant and Equipment - Total (Net)</i>
Operating Activities - Net Cash Flow	<i>Working Capital (Balance Sheet)</i>
Operating Income Before Depreciation	

*Note: Informative Financial Variables (IFVs) are highlighted in bold.

Panel B. Comparison of Predictive Power using IAQI and IFV

Misclassification cost	IAQI	IFV	IAQI + IFV	Misclassification cost	IAQI	IFV	IAQI + IFV
1	C (0.627)	A (0.608)	B (0.618)	40	B (0.639)	A (0.594)	B (0.646)
5	B (0.621)	A (0.609)	C (0.638)	45	B (0.640)	A (0.592)	B (0.644)
10	B (0.633)	A (0.609)	B (0.638)	50	B (0.637)	A (0.593)	B (0.642)
15	B (0.636)	A (0.604)	B (0.640)	60	B (0.630)	A (0.596)	B (0.637)
20	B (0.638)	A (0.601)	B (0.642)	70	B (0.625)	A (0.595)	C (0.639)
25	B (0.644)	A (0.599)	B (0.644)	80	B (0.622)	A (0.593)	C (0.634)
30	B (0.645)	A (0.601)	B (0.640)	90	B (0.615)	A (0.595)	C (0.634)
35	B (0.640)	A (0.597)	B (0.641)	100	B (0.610)	A (0.594)	C (0.631)

Note:

The average AUC is presented in parentheses. The Tukey HSD connected letter report ranks the average AUC of algorithms alphabetically, with A indicating the lowest value. The difference in AUC between two algorithms is significant when they are ranked by different letters. The average AUC is presented in parentheses. IAQI include Office Size, Tenure, Audit Report Lag, Integrated Audit, Auditor Resignation, Audit Fees, Internal Control Weakness, Disc. Accruals, Abs (Disc. Accruals), Abs (Accruals), and Abs (Accruals/CFO). The definition of IAQI are provided in Table 1. IFV include Accounts Payable - Trade, Cash and Short-Term Investments, Short-Term Investments - Total, Sales/Turnover (Net), Sale of Common and Preferred Stock, Income Taxes - Total, Working Capital (Balance Sheet), and Interest and Related Expense – Total.

Table 10. Assessment Tests Results

	MAR			PAQI		
	Coef.	z score	P-value	Coef.	t statistic	P-value
PAQI	0.413	2.36***	0.009			
Big4	-1.271	-3.05	0.002	-0.315	-8.60***	0.000
Delta_Rec	0.000	-1.67*	0.094	0.000	1.59	0.111
Delta_INV	0.000	1.11	0.269	0.000	-1.82*	0.068
soft_asset	2.048	2.95***	0.003	0.066	1.00	0.315
LNASSET	0.332	3.03***	0.002	-0.096	-7.72***	0.000
ATURN	0.014	0.07	0.943	0.023	1.40	0.161
ROA	0.281	0.56	0.578	0.045	1.55	0.121
Leverage	0.139	0.89	0.372	0.012	0.53	0.597
CURR	0.051	1.10	0.272	-0.014	-3.07***	0.002
MKBK	-0.004	-0.63	0.528	0.000	-0.38	0.705
EPSGrowth	-0.066	-2.25**	0.024	0.001	0.26	0.793
EPS	-0.092	-0.38	0.706	-0.056	-2.04**	0.041
SalesGrowth	0.205	2.08**	0.038	0.012	0.88	0.380
MA	0.335	1.19	0.234	0.054	2.08***	0.038
Restructure	-0.202	-0.65	0.516	-0.005	-0.20	0.842
FirmAge	0.000	-0.57	0.566	0.000	-12.14***	0.000
Going concern	-0.330	-0.51	0.613	-0.113	-2.05**	0.041
Auditor Change	0.903	2.27**	0.023	0.409	7.56***	0.000
Influence	0.340	0.56	0.574	0.096	1.65*	0.098
FREEC	-0.409	-1.96*	0.050	-0.018	-0.46	0.643
Abnormal Audit Fee	0.107	0.35	0.724	0.062	2.25**	0.024
Busy	-0.376	-1.13	0.258	0.037	1.11	0.268
Auditor Competition_MSA	0.463	0.44	0.663	0.166	1.59	0.111
Industry						
Specialization_MSA	0.440	0.99	0.320	0.021	0.52	0.600
Non-Audit Fee Ratio	0.028	0.03	0.976	0.042	0.46	0.648
Local Auditor_MSA	0.100	0.35	0.727	-0.024	-0.95	0.343
Year and industry fixed effects	Yes			Yes		
Pseudo R2	15.64%					
R2				23.57%		
P value for model	0.000			0.000		
Number of Observations	5326			5946		

Note: ***, **, and * indicate significance level at 1%, 5%, and 10% respectively (one-sided p-value for the test variable and two-sided p-value for control variables). All standard errors are estimated by clustering firms. MAR is material annual restatement due to GAAP violations or financial fraud. PAQI is the predictive index generated from IAQI and the machine learning process. Big 4 is 1 if the auditor is one of the Big 4 audit firms. 0 otherwise. Delta_Rec = change in accounts receivable scaled by total assets: $RECT_t/AT_t - RECT_{t-1}/AT_{t-1}$ (Lobo and Zhao 2013). Delta_INV = change in inventory scaled by total assets: $INVT_t/AT_t - INVT_{t-1}/AT_{t-1}$ (Lobo and Zhao 2013) Soft asset = soft assets as a percentage of total assets: $(AT_t - PPENT_t - CHE_t)/AT_t$ (Lobo and Zhao 2013). LNASSET = log of total assets: $\log(AT_{t-1})$ (Eshleman and Guo 2014). ATRUN = total sales divided by lagged total assets: $REVT_t/AT_{t-1}$ (Eshleman and Guo 2014). ROA = income before extraordinary items divided by average total assets: $IB_t/((AT_t + AT_{t-1})/2)$ (Eshleman and Guo 2014). Leverage = financial leverage, defined as long-term debt plus debt in current liabilities, all scaled by total assets: $(DLTT_t + DLC_t)/AT_t$ (Eshleman and Guo 2014). CURR = the current ratio, defined as current assets divided by current liabilities: ACT_t/LCT_t (Eshleman and Guo 2014). MKBK = the market-to-book ratio, defined as market value at fiscal year-end scaled by book equity: $(PRCC_{F_t} * CSHO_t)/CEQ_t$ (Eshleman and Guo 2014). EPS = the earnings-to-price ratio, defined as income before extraordinary items, scaled by market value at fiscal year-end: $IB_t/(PRCC_{F_t} * CSHO_t)$ (Eshleman and Guo 2014). EPSGrowth = the growth rate of EPS: $(EPS_t - EPS_{t-1})/EPS_{t-1}$ (Eshleman and Guo 2014). TtotalAccruals = change in noncash assets (noncash total assets minus total liabilities and

preferred stocks) from year $t-1$ to year t scaled by average total assets: $\frac{\{(AT_t - CHE_t) - (LT_t + PSTK_t)\} - \{(AT_{t-1} - CHE_{t-1}) - (LT_{t-1} + PSTK_{t-1})\}}{(AT_t + AT_{t-1})/2}$ (Lobo and Zhao 2013). SalesGrowth = change in sales from the prior year to the current year: $(SALE_t - SALE_{t-1})/SALE_{t-1}$ (Lobo and Zhao 2013). MA = 1 if involved in merger activity (Stanley and DeZoort 2007). Restructure = 1 if the firm has restructuring changes during the year (Newton et al. 2013). FirmAge = The natural logarithm of the number of years the firm has been listed on COMPUSTAT (Eshleman and Guo 2014): $\ln(\text{fiscal year} - \text{IPODATE})$. GoingConcern = 1 if the company received a going concern modified opinion in year t , zero otherwise (Ettredge Emeigh Fuerherm Li 2014). AuditorChange = 1 if an audit engagement occurs within the first year of an auditor change, and 0 otherwise (Francis et al. 2013). This is equivalent to the variable NewClient defined in Table 1. Influence = is the ratio of a specific client's total fees (audit fees plus nonaudit fees) relative to annual fees of SEC registrants generated by the practice office in a given year (Francis et al. 2013), Eshleman and Guo 2014). FREEC = Demand for external financing, defined as operating cash flows (OANCF) less capital expenditures (CAPX), all scaled by lagged assets (Eshleman and Guo 2014). Abnormal Audit Fee = The unscaled residual from the audit fee model used in Blankley et al. (2012). Busy = 1 if a company has a fiscal year-end date of December, and 0 otherwise (Lopez and Peters 2012). Auditor Competition_MSA = MSA-level auditor concentration based on Herfindahl index. Details are provided in (Netown et al. 2013). Industry Specialization_MSA = auditor's annual market share of audit fees within a two-digit SIC category for a particular city. A city is defined as a Metropolitan Statistical Area (MSA) (Reichelt and Wang 2010). Non-Audit Fee Ratio = Non-audit fees deflated by total fees paid (audit plus non-audit fees). Non-audit fee equals to the sum of benefit fee, IT fee, Tax fee, audit related fee, and other fees. (Ruddock et al. 2006) Local Auditor_MSA = 1 if the audit engagement office is located in the same MSA where audit clients are headquartered, 0 otherwise (Choi, Kim, Qiu, and Zhang 2012). We did not include qualified opinion from Eshleman and Guo (2014) because observations with qualified opinion were removed due to missing other variables. We did not include Leases from Lobo and Zhao (2013) because the remaining observations all have future operating lease obligations that are greater than 0 ($MRCT > 0$).

Figure 1. The Overall Research Design

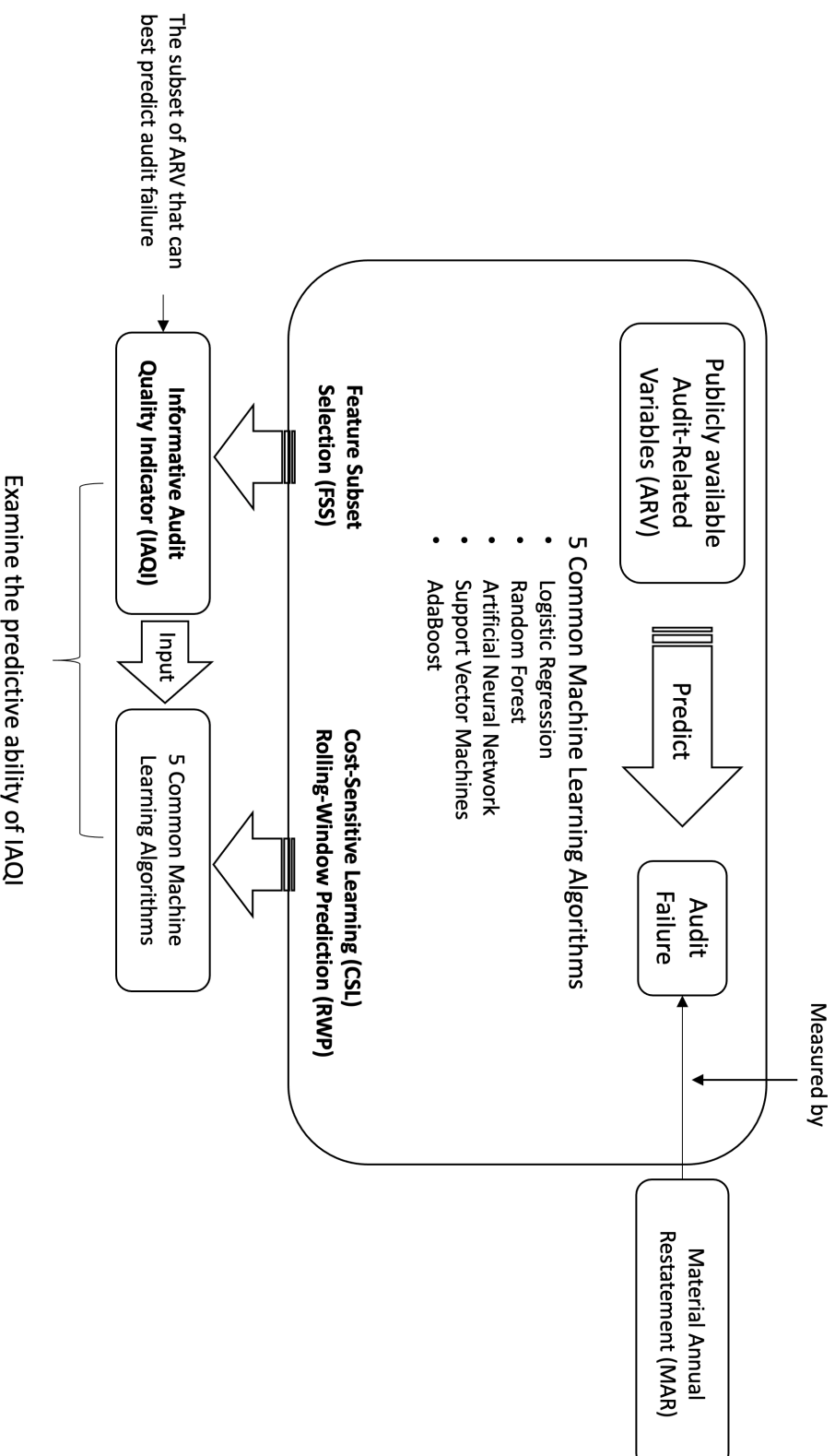
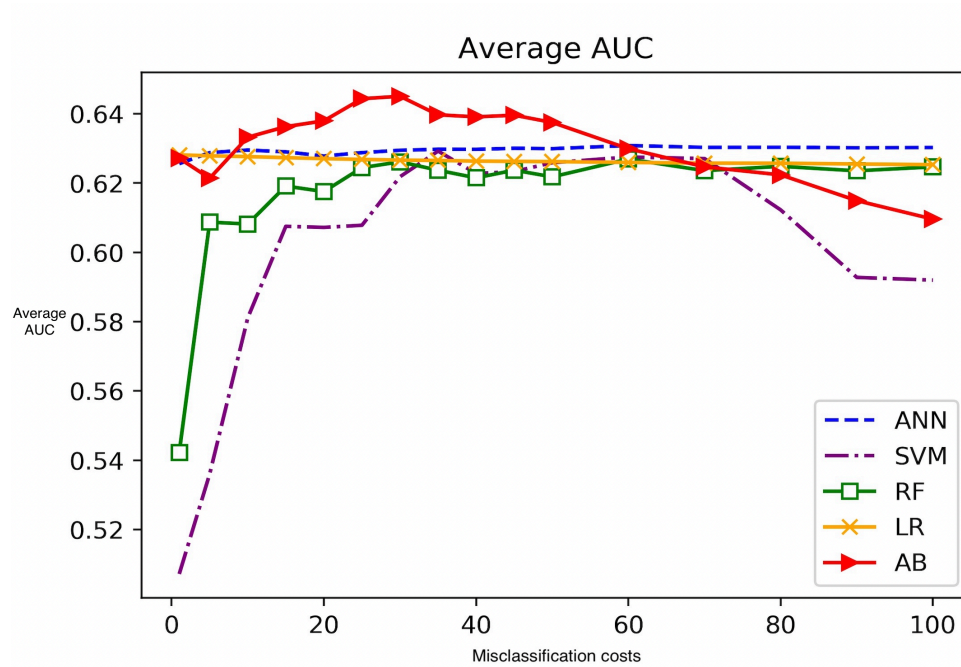


Figure 2. Algorithm Performance



Note:

AB is AdaBoost. ANN is Artificial Neural Network. SVM is Support Vector Machine. RF is Random Forest. LR is Logistic Regression.

Figure 3.

Panel A. Summary Statistics of PAQI

Mean	Median	Min	Max	Std. Dev
0.00	0.00	-7.17	5.89	1.00

Panel B. T-tests of PAQI

Actual MAR (Mean)	Actual Non-MAR (Mean)	Difference
0.48	-0.01	0.49***

Note: *** indicates significance level at 1% (based on two-sided p value). MAR is Material Annual Accounting Restatement, or material annual restatement due to GAAP violations. “Actual MAR” represents observations that actually have MAR. “Actual Non-MAR” represents observations that do not have MAR. PAQI is the predictive audit quality index generated from IAQI and the machine learning process.

APPENDIX

Appendix A. Comparison Between This Paper and the Most Related Literature

Literature	Objectives	Target Variable (Dependent Variable)	Predictors (Independent Variable)	Methodology	Evaluation Method
Cecchini et al. (2010)	<ol style="list-style-type: none"> Predict management fraud using basic financial data Develop a kernel specific to the domain of finance 	Fraud investigations enforced by the SEC and disclosed in AAER	Financial variables that have been used in fraud prediction	Predictive modeling (Support Vector Machine)	Out-of-sample evaluation (Recall rate, AUC)
DeChow et al. (2011)	<ol style="list-style-type: none"> Predict material misstatements Develop a comprehensive database of financial misstatements 	Material misstatements enforced by the SEC and disclosed on the AAER	Accrual quality Financial performance Nonfinancial measures Off-balance-sheet activities	Explanatory modeling (Regression)	In-sample evaluation (Recall rate, Specificity rate, AUC)
Perols (2011)	<ol style="list-style-type: none"> Identify which classification algorithms provide the most utility in predicting fraud Identify the best prior fraud probability and misclassification cost when training classifiers Identify useful predictors for classification algorithms 	Fraud investigations enforced by the SEC and disclosed in AAER	Mostly financial variables that have been identified to be significant in fraud research	Predictive modeling (148, SMO, Multilayer Perceptron, Logistics, stacking, and bagging)	Out-of-sample evaluation (Estimated Relative Costs, ERC, of misclassification)
Perols (2016)	<p>Introduce and evaluate three data analytics preprocessing methods to address challenges related to (1) the rarity of fraud observations, (2) the relative abundance of explanatory variables identified in the prior literature, and (3) the broad underlying definition of fraud.</p>	Fraud investigations enforced by the SEC and disclosed in AAER	Mostly financial variables that have been identified to be significant in fraud research	Predictive modeling (Support Vector Machine)	Out-of-sample evaluation (Estimated Relative Costs, ERC, of misclassification)
Dutta et al. (2017)	Predict restatements	All types of restatements from Audit Analytics database (fraudulent or erroneous, disclosed in all sources)	Financial variables that are related to fraud/restatement	Predictive modeling (Decision Tree, Artificial Neural Network, Naïve Bayes, Support Vector Machine, and Bayesian Belief Network)	Out-of-sample evaluation (Recall rate Specificity rate AUC)
Aobdia (2019)	Investigate the degree of concordance between fifteen measures of audit quality used in academia and two measures of audit	Part I Findings and internal inspection ratings from PCAOB	15 measures of audit quality used in academia	Explanatory modeling (Regression)	Statistical significance

	process quality determined either by audit firms' internal inspections or by PCAOB inspections of individual engagements.			
Bao et al. (2020)	Predict accounting fraud using machine learning using financial data	Accounting frauds from SEC's AAER in CFRM database	28 raw financial variables from Dechow et al. (2011) and Cecchini et al. (2010)	Predictive modeling (Ensemble learning) Out-of-sample evaluation (AUC and NDCCG@k)
Brown et al. (2020)	Use a machine learning technique to assess whether the thematic content of financial statement disclosures is incrementally informative in predicting intentional misreporting.	Accounting frauds from SEC's AAER in CFRM database Fraud-related restatements from Audit Analytics database Fraud-related restatements from 10K/A	F-score from Dechow et al. (2011) Thematic content variables (topic and style)	Use Bayesian topic modeling algorithm to determine and quantify the topic content of a large collection of 10-K narratives Use predictive modeling (logistic regression) to examine the incremental predictive power of thematic contents Out-of-sample evaluation (AUC)
Bertomeu et al. (2020)	Use machine learning to predict material misstatements.	Material restatements disclosed in Form 8-K from Audit Analytics database	Variables from accounting, capital markets, governance, and auditing datasets	Predictive modeling (Gradient Boosted Regression Tree and other common algorithms) Out-of-sample evaluation (AUC)
Rajgopal et al. (2021)	Provide detailed descriptive analyses of how poor audits are perceived in both public and private litigation settings. And evaluate how well existing audit quality proxies predict detailed allegations related to how auditors actually performed in specific engagements.	Audit deficiencies in specific engagements alleged by the SEC or private law firms	14 frequently used audit quality proxies	Explanatory modeling (Logistic Regression) In-sample evaluation (Statistical significance, AUC)
This paper	Identify IAQI, which are theory-driven audit-related variables that are the most predictive of audit failure.	Material annual restatements due to GAAP violations or financial fraud disclosed in Form 8-K from Audit Analytics database	Theory-driven audit-related variables	Predictive modeling (Random Forest, Artificial Neural Network, Support Vector Machine, Logistic Regression, AdaBoost) Out-of-sample evaluation (AUC)

Appendix B. Pairwise Correlation of Audit-Related Variables

Industry	Industry Specialization_MSA	Industry Specialization_MSA	Office Size	Big-4	New Client	Turnover	Local Auditor_MSA	Integrated Audit	Accidental Filter	Busy	Workload Compression	Auditor Competition_MSA	Auditor Resignation	Audit Fees	Tax Fee	Audit-Rated Fee	Other Audit Fee	Non-Audit Fee Ratio	Influence	Abnormal Audit Fee
1	0.4511*	0.2237*	0.7870*	0.2267*	0.4276*	0.1105*	0.1582*	0.9690*	0.0620*	0.7846*	0.768*	0.0172*	0.1383*	0.4464*	0.3338*	0.1418*	0.2475*	0.0536*	0.0364*	0.0743*
Specialization_MSA	-0.4264*	-0.1085*	-0.2037*	-0.4066*	-0.0556*	0.3735*	0.1639*	0.0613*	0.0620*	0.7846*	0.0964	0.0946	0.0462	0.0892	0.0617	0.0158*	0.0563*	-0.0991*	0.1033*	0.0191*
Office Size	0.7786*	0.3397*	0.7870*	0.2267*	-0.4276*	0.1105*	0.1582*	0.9690*	0.0620*	0.7846*	0.768*	0.0172*	0.1383*	0.4464*	0.3338*	0.1418*	0.2475*	0.0536*	0.0364*	0.0743*
Big-4	-0.1903*	0.01627*	0.3397*	0.2066*	-0.0556*	0.3735*	0.1639*	0.0613*	0.0620*	0.7846*	0.0964	0.0946	0.0462	0.0892	0.0617	0.0158*	0.0563*	-0.0991*	0.1033*	0.0191*
New Client	0.3581*	0.0171*	0.2323*	0.2066*	-0.0556*	0.3735*	0.1639*	0.0613*	0.0620*	0.7846*	0.0964	0.0946	0.0462	0.0892	0.0617	0.0158*	0.0563*	-0.0991*	0.1033*	0.0191*
Turnover	0.1627*	0.0171*	0.2323*	0.2066*	-0.0556*	0.3735*	0.1639*	0.0613*	0.0620*	0.7846*	0.0964	0.0946	0.0462	0.0892	0.0617	0.0158*	0.0563*	-0.0991*	0.1033*	0.0191*
Local Auditor_MSA	0.4175*	0.2194*	0.5155*	0.5242*	-0.1671*	0.3423*	0.1639*	0.0699*	0.0620*	0.7846*	0.0964	0.0946	0.0462	0.0892	0.0617	0.0158*	0.0563*	-0.0991*	0.1033*	0.0191*
Integrated Audit	0.3927*	0.1909*	0.5196*	0.5189*	-0.1671*	0.3423*	0.1639*	0.0699*	0.0620*	0.7846*	0.0964	0.0946	0.0462	0.0892	0.0617	0.0158*	0.0563*	-0.0991*	0.1033*	0.0191*
Accidental Filter	0.1148*	0.0694*	0.1309*	0.1295*	-0.0399*	0.0558*	0.0077*	0.0887*	0.0348*	0.7846*	0.0964	0.0946	0.0462	0.0892	0.0617	0.0158*	0.0563*	-0.0991*	0.1033*	0.0191*
Busy	0.0980*	0.1090*	-0.0155*	0.0891*	-0.0241*	0.0692*	0.0077*	0.0887*	0.0348*	0.7846*	0.0964	0.0946	0.0462	0.0892	0.0617	0.0158*	0.0563*	-0.0991*	0.1033*	0.0191*
Workload Compression	0.0889*	0.4197*	0.0190*	0.0920*	-0.0190*	0.0797*	-0.1014*	0.0854*	0.0419*	0.7846*	0.0964	0.0946	0.0462	0.0892	0.0617	0.0158*	0.0563*	-0.0991*	0.1033*	0.0191*
Auditor Competition_MSA	0.1060*	-0.0365*	-0.1429*	-0.1288*	0.0243*	-0.0806*	-0.0387*	-0.1023*	-0.0988*	-0.0149*	0.0946	0.0172*	0.1383*	0.4464*	0.3338*	0.1418*	0.2475*	0.0536*	0.0364*	0.0743*
Auditor Resignation	0.0059*	0.0292*	0.0192*	-0.1983*	-0.1983*	0.0059*	0.0292*	-0.1983*	-0.1983*	0.0059*	0.0292*	0.0192*	-0.1983*	-0.1983*	0.0059*	0.0292*	-0.1983*	-0.1983*	0.0059*	0.0292*
Audit Fees	0.3078*	0.3078*	0.3112*	0.3390*	-0.1078*	0.3333*	0.3035*	0.3035*	0.2811*	0.0371*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*
Tax Fee	0.3037*	0.2937*	0.3112*	0.3390*	-0.1078*	0.3333*	0.3035*	0.3035*	0.2811*	0.0371*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*	0.0510*
Audit-Rated Fee	0.2101*	0.1861*	0.2235*	0.2356*	-0.0745*	0.1882*	0.1835*	0.1835*	0.1791*	0.0763*	0.0119	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*
Other Audit Fee	0.0888*	0.0888*	0.0854*	0.1138*	-0.0340*	0.0673*	0.0417*	0.0791*	0.0763*	0.0763*	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*	0.0402*
Non-Audit Fee Ratio	-0.2498*	0.0433*	-0.6217*	-0.3299*	0.1138*	-0.0995*	-0.0945*	-0.1410*	-0.0795*	-0.0150*	0.2409*	0.3477*	0.3477*	0.3477*	0.3477*	0.3477*	0.3477*	0.3477*	0.3477*	0.3477*
Abnormal Audit Fee	0.1310*	0.0438*	0.2311*	0.1546*	-0.0722*	-0.0299*	-0.0352*	-0.0666*	-0.0352*	-0.0352*	-0.0449*	-0.0893*	-0.0893*	-0.0893*	-0.0893*	-0.0893*	-0.0893*	-0.0893*	-0.0893*	-0.0893*
Audit Report Lag	-0.3680*	-0.2241*	-0.4189*	-0.4222*	0.1761*	-0.3145*	-0.1435*	-0.4632*	-0.4522*	-0.6606*	-0.5052*	-0.2723*	-0.2723*	-0.2723*	-0.2723*	-0.2723*	-0.2723*	-0.2723*	-0.2723*	-0.2723*
Non-tenure Issuance of IOK Due to Audit	-0.0809*	-0.0448*	-0.0933*	-0.0947*	0.0771*	-0.0669*	-0.0564*	-0.0770*	-0.0768*	-0.0186*	-0.0158*	-0.0240*	-0.0240*	-0.0240*	-0.0240*	-0.0240*	-0.0240*	-0.0240*	-0.0240*	-0.0240*
IOK Due to Audit	-0.3049*	-0.1879*	-0.4025*	-0.3654*	0.1174*	-0.1973*	-0.1668*	-0.3864*	-0.3824*	-0.0223*	-0.0223*	-0.0224*	-0.0224*	-0.0224*	-0.0224*	-0.0224*	-0.0224*	-0.0224*	-0.0224*	-0.0224*
Georg Concern	0.0061	-0.0045	0.0346*	0.0177*	0.0337*	-0.0325*	0.0129*	0.1236*	0.1125*	-0.0077*	-0.0089	-0.0137*	0.0111	0.0776*	0.0085	0.0021	0.0149*	-0.0231*	0.0211*	0.0549*
Internal Control	0.0130*	0.0176*	0.0144*	-0.0262*	0.0176*	0.0176*	0.0138*	0.0138*	0.0143*	0.0133*	0.0212*	0.0245*	0.0096*	0.0128*	0.0073	-0.0082	0.0053	-0.0112	0.0043	0.0112
Disc. Accruals	-0.2044*	-0.1443*	-0.2674*	-0.2408*	-0.1478*	-0.1478*	-0.1220*	-0.2425*	-0.2369*	-0.0217*	-0.0341*	-0.0467*	-0.0566*	-0.3238*	-0.1518*	-0.1413*	-0.0855*	-0.0508*	0.0452*	-0.0220*
Abn (Disc. Accruals)	-0.1973*	-0.1250*	-0.2750*	-0.2391*	-0.1318*	-0.1318*	-0.1387*	-0.2362*	-0.2306*	-0.0328*	-0.0478*	-0.0574*	-0.0574*	-0.3341*	-0.1612*	-0.1411*	-0.0929*	-0.0718*	0.0425*	-0.0076
Abn (Accruals/CFO)	-0.1153*	-0.0713*	-0.1340*	-0.1132*	0.0559*	-0.1013*	-0.0588*	-0.1340*	-0.1341*	-0.0009	-0.0157*	-0.0336*	0.0560*	-0.1410*	-0.0841*	-0.0844*	-0.0548*	-0.0424*	0.0321*	-0.0027
DD Residual	-0.2407*	-0.1615*	-0.2948*	-0.2794*	0.0902*	-0.1654*	-0.1198*	-0.2277*	-0.2277*	-0.0272*	-0.0372*	-0.0372*	-0.0372*	-0.1431*	-0.1643*	-0.1721*	-0.0980*	-0.0757*	0.0549*	-0.0056
Prior ROA Meet	0.0353*	0.0338*	0.0349*	0.0362*	-0.0151*	0.0385*	0.0167*	0.0236*	0.0236*	0.0072	0.0104	0.0068	0.0049	0.0466*	0.0266*	0.0250*	0.0187*	0.0097	0.0187*	0.0187*
Small Profit	0.0056	0.0059	0.0252*	-0.0002	0.0142*	-0.008	0.0066	0.0194*	0.0141*	-0.0208*	-0.0115*	-0.0083	-0.0123*	0.0426*	-0.0044	0.012	0.0187*	-0.0084	0.0026	-0.0008

(Continued)

Audit Report Lag	0.1480*	0.1180*	0.0228*	0.0066	0.3008*	0.4947*	0.2516*	0.1333*	-0.0225*	-0.0543*	0.0222*
Non-tenure Issuance of IOK Due to Audit	0.3952*	0.1180*	-0.0228*	-0.0066	-0.3008*	0.4947*	0.2516*	0.1333*	-0.0225*	-0.0543*	0.0222*
Georg Concern	0.1175*	0.0642*	-0.0228*	-0.0066	-0.3008*	0.4947*	0.2516*	0.1333*	-0.0225*	-0.0543*	0.0222*
Internal Control	-0.0359*	-0.0061	-0.0514*	-0.0167*	-0.3808*	0.4947*	0.2516*	0.1333*	-0.0225*	-0.0543*	0.0222*
Disc. Accruals	0.2518*	0.0561*	0.4126*	-0.0148*	-0.3808*	0.4947*	0.2516*	0.1333*	-0.0225*	-0.0543*	0.0222*
Abn (Disc. Accruals)	0.3011*	0.0903*	0.4767*	-0.0148*	-0.3808*	0.4947*	0.2516*	0.1333*	-0.0225*	-0.0543*	0.0222*
Abn (Accruals/CFO)	0.1888*	0.0459*	0.1555*	-0.0329*	0.1644*	0.4947*	0.2516*	0.1333*	-0.0225*	-0.0543*	0.0222*
DD Residual	0.2953*	0.0919*	0.4543*	-0.0145*	-0.4381*	0.4947*	0.2516*	0.1333*	-0.0225*	-0.0543*	0.0222*
Prior ROA Meet	-0.0492*	-0.0032	-0.0207*	-0.0083	-0.0121*	-0.0081	-0.0081	-0.0201*	-0.0543*	0.0164*	-0.0222*
Small Profit	0.0067	-0.0095	-0.0757*	0.0261*	0.0118	0.0118	-0.0500*	0.0164*	-0.0543*	0.0164*	-0.0222*

Note: * indicates significant at 5% level.

Appendix C. Cost-Sensitive Learning (CSL) and Multi-Subset Observation Under-sampling (OU) Method

CSL works mainly by adjusting the ratio between the number of positive and negative instances in the training dataset (Elkan 2001). Specifically, this adjustment allows us to keep all positive instances and a proportion (i.e., $1/\text{misclassification cost}$) of the negative instances in the training dataset (Thai-Nghe, Gantner, and Schmidt-Thieme 2010). In other words, rendering an algorithm cost-sensitive is equivalent to under-sampling the negative instances in the training dataset (Thai-Nghe et al. 2010). However, merely under-sampling the negative examples can result in the discarding of potentially useful information for the classification task (Perols et al. 2016). To avoid information loss, we follow the example of Perols et al. (2016) and adopt the Multi-Subset Observation Undersampling (OU) method (Chan and Stolfo 1998), which creates n (n equals misclassification cost) training datasets, each of which contains all positive observations as well as a different subsample of negative observations. Perols et al. (2016) document a detailed description of the OU method.

Appendix D. Overall Experiment Procedure for Algorithm Selection

After we identify IAQI, we use them as inputs and MAR as outcomes to compare the performance of the five common machine learning algorithms. For each discrete combination of an algorithm and a misclassification cost, we perform seven sets of rolling-window prediction with cost-sensitive learning (i.e., training 2005-2009 and testing 2011; training 2006-2010 and testing 2012; training 2007-2011 and testing 2013; training 2008-2012 and testing 2014; training 2009-2013 and testing 2015; training 2010-2014 and testing 2016; and training 2011-2015 and testing 2017). In each rolling-window prediction set, we partition the training data into Observation Undersampling (OU) subsets³⁴ (Perols et al. 2016) based on the misclassification cost. For example, if the misclassification cost is 20, we create 20 OU subsets wherein each subset contains all positive observations as well as a different and randomly selected subsample of the negative observations from the training dataset. For each OU subset, we train the machine learning algorithm to build one prediction model, which we then use to predict the outcome in the test set and to generate a probability prediction for each observation in the test set. Therefore, in each

³⁴ Specifically, the negative instances in the training data were partitioned into N parts, where N equals the misclassification cost ratio. We then combined each of the N groups of negative instances with all the positive instances in the training data, resulting in N subsets.

rolling-window prediction set, the number of prediction models forecasting the outcome of observation in the test is equal to the misclassification cost. To aggregate the prediction results from different prediction models, we average the probability predictions for each observation in the test set to obtain a single value of the probability prediction for the respective observation (Perols et al. 2016; hereafter, we refer to this averaged probability prediction as to the “final probability prediction”). With the final probability predictions for the test set, we can calculate one AUC value for the test set in each set of rolling-window predictions. Therefore, for each combination of one algorithm and one misclassification cost, we can obtain 5 AUC values, each from a set of rolling-window predictions.

Since the creation of OU subsets involves random sampling of the negative observations, we repeat the above procedure five times (i.e., rounds) for each combination of an algorithm and a misclassification cost in order to reduce the impact of the randomness on the prediction results. Therefore, through our experiment, we can collect 35 AUC values (5 rounds * 7 rolling-window prediction sets) for each combination of an algorithm and a misclassification cost. Furthermore, for each algorithm, we can collect 560 AUC values (5 rounds * 7 rolling-window prediction sets * 16 misclassification costs). In total, we can collect 2800 AUC values (5 rounds * 7 rolling-window prediction sets * 16 misclassification costs * 5 algorithms).