THE EUROPEAN JOURNAL OF APPLIED ECONOMICS

EJAE 2019, 16(2): 1-58
ISSN 2406-2588
UDK:657.6-051
005.521:657.635
DOI: 10.5937/EJAE16-21832
Original paper/Originalni naučni rad

PREDICTING THE TYPE OF AUDITOR OPINION: STATISTICS, MACHINE LEARNING, OR A COMBINATION OF THE TWO?

Nemanja Stanišić, Tijana Radojević*, Nenad Stanić

Singidunum University,
Belgrade, Serbia

Abstract:
The goal of this study is to overcome the identified methodological limitations of prior studies aimed at predicting the type of auditor opinion and draw definite conclusions on the relative predictive performance of different predictive methods for this particular task. Predictive performance of twelve candidate models from the realms of statistics and machine learning is assessed separately for the two common real-life scenarios: a) when prior information on the client (i.e. types of audit opinion received in the past) is available and can be used in prediction, and b) when such information is not available (e.g. new companies). The results show that, in the first scenario, several methods from both realms achieve comparable predictive performance of around 0.89, as measured by the Area under the curve (AUC). In the second scenario, however, machine learning algorithms, particularly tree-based ones, such as random forest, perform significantly better, achieving the AUC of up to 0.79. Finally, we develop and assess the predictive performance of two hybrid models aimed at combining the strong points of both statistical (i.e. interpretability of results) and machine learning (i.e. handling a large number of predictors and improved accuracy) approaches. The complete procedure is demonstrated in a reproducible manner, using the largest empirical data set ever used in this stream of research, comprising 13,561 pairs of annual financial statements and the corresponding audit reports. The procedures described in this study allow audit and finance professionals around the globe to develop and test predictive models that will aid their procedures of audit planning and risk assessment.

Article info:
Received: January 9, 2019
Correction: March 5, 2019
Accepted: April 11, 2019

Keywords:
auditor opinion,
financial reports,
generalized linear mixed models,
random forest,
guided regularized random forest,
ensembles.

*E-mail: tradojevic@singidunum.ac.rs

INTRODUCTION

According to the international auditing standards, auditors are required to consider a number of factors related to the risk of material misstatements in their clients' financial reports, to provide a basis for designing and performing audit procedures (ASB GAAS Section 315, 2013; IAASB ISA 315, 2013). Namely, they are required to "discuss the susceptibility of the entity's financial statements to material misstatement" (ASB GAAS Section 315, 2013, para. 5; IAASB ISA 315, 2013, para. 10) and "identify and assess the risks of material misstatement at: (a) the financial statement level and (b) the assertion level for classes of transactions, account balances, and disclosures" (ASB GAAS Section 315, 2013, para. 26; IAASB ISA 315, 2013, para. 35). In light of this, developing and employing models that are, given the data obtained from financial statements, being audited and other available company information, predictive of the forthcoming type of auditor opinion, and thus provide timely and reliable assessments of the risk of detectable material misstatements, are of great practical importance to auditors (Bell and Tabor, 1991). By using such models, auditors can screen their client portfolio and direct attention to those clients who have high estimated probability of receiving a qualified audit opinion, and thus not only save time and money (Gaganis, Pasiouras, Spathis et al., 2007), but also reduce the potential litigation cost and reputation risk.

As a result, over the past decades, numerous models for predicting the type of audit opinion have been proposed in the literature (Kirkos et al., 2007). Earlier studies have mainly relied on the use of classical statistical modeling techniques, mainly probit and logistic regression models (Dopuch et al., 1987; Francis and Krishnan, 1999; Krishnan and Krishnan, 1996), while more recent ones have shown an increased preference for modern machine learning techniques, such as decision trees and artificial neural networks (Fernández-Gámez et al., 2016; Pourheydari et al., 2012; Saif et al., 2013; Yasar et al., 2015).[1] Although the abundance of empirical research, conducted on the topic, might be expected to provide a solid basis for practitioners to identify a predictive technique that is best suited for predicting audit qualifications, we argue that this is not possible for two main reasons.

The first reason is that prior studies have not fully employed the respective strong points of the two sets of techniques. Specifically, the studies demonstrating the use of statistical techniques have not taken the advantage of their capacity to model individual effects of clients and auditor firms, essentially ignoring the valuable information that is readily available in prior audit reports. On the other hand, the studies demonstrating the use of machine learning techniques have considered only a limited pool of theoretically-driven predictors, failing to take advantage of their capacity to handle a large number of predictors that can be generated based on the data available from the complete set of financial reports. Therefore, the predictive performance achieved and reported in prior studies can be deemed suboptimal; consequently, any comparison based on these results would be indicative rather than definite.

The second reason is that most studies — with the exception of the study by Doumpos and colleagues (2005) — have not differentiated between the cases where prior information on the client and the auditor (i.e. audit reports from prior periods) is available and cases where such information is not available. This difference is crucial, since the highest achievable predictive performance differs substantially between the two scenarios. By not differentiating between the two scenarios, researchers

1 Even though the distinction between statistical and machine learning methods is not a clear-cut one (statistics is the basis for many modern machine learning methods), we use this terminological demarcation in this study mainly to highlight the fact that statistical methods are better at modeling the systematic effects suggested by the researcher, and machine learning methods are more capable of handling a large number of predictors, which is relevant to the aims of our study. This distinction also highlights the temporal trend in the usage of predictive methods observed in prior literature.

have been not only reported an aggregated accuracy across the two scenarios, but, more importantly, failed to examine a likely possibility that, to achieve optimal results, the two scenarios require two different predictive techniques.

Considering the above-mentioned, the main motivation for conducting this study is to build several candidate models from the realms of both statistics and machine learning, using a single empirical data set, in a way that allows us to draw more valid conclusions on their relative predictive performance than is possible based on the existing body of research. To do so, we first demonstrate how to employ the respective strong points of the two sets of techniques more effectively. Specifically, we: a) demonstrate how client- and auditor-specific individual effects can be modeled using random effects within the logistic regression framework and b) design a feature-generation procedure aimed at making better use of the data available in the complete set of financial reports. Those two methodological amendments are demonstrated to meaningfully improve the predictive performance of statistical and machine learning techniques, respectively. Next, we demonstrate that, when their potentials are more fully employed, each set of techniques is better suited to one real-life scenario. While mixed-effects logistic regression, which is a statistical tool, should be preferred for the interpretability of its output, when prior information on the client is available, machine learning techniques, such as random forest [RF] and guided regularized random forest [GRRF], should be preferred for their superior predictive performance in scenarios where no prior information is available (e.g. new companies). Finally, motivated by this finding, we build two hybrid models that combine the feature selection capability and flexibility of machine learning algorithms with the improved interpretability and the ability to account for the systematic effects present in panel data sets of regression models. The first one is a stacked ensemble that uses predictions from seven different machine learning algorithms as predictors and mixed-effects logistic regression as a meta-learner. This model outperformed each of the individual regular predictive models in both scenarios. The second hybrid model involves the inclusion of a compact set of classification rules selected by the GRRF algorithm as dummy variables in mixed-effects logistic regression. This model is shown to be particularly useful in scenarios where prior information is available, providing the best trade-off between interpretability and accuracy.

The complete model development and validation procedure is demonstrated in a reproducible manner using a large empirical data set on audit opinions and financial reports. As such, the framework presented herein can be used by audit and finance professionals around the globe to develop and test the predictive models that will aid the process of audit planning and risk assessments.

LITERATURE REVIEW

The line of research that this study belongs to is that aimed at forming expectations of the type of auditor opinion.[2] The main challenges are identifying factors, associated with receiving qualified or modified opinion (including qualified opinion, adverse opinion, and disclaimer of opinion), as opposed to unqualified audit opinion, and developing predictive models that, with the highest achievable accuracy, estimate the probability of class membership at the level of the individual financial report.

2 Lines of research which, although somewhat related, should be clearly demarcated from the one under consideration in this study, are those devoted to: detection of earnings management (DeAngelo, 1986; Dechow et al., 1995; DeFond & Jiambalvo, 1994; Healy, 1985; Jones, 1991); detection of fraudulent reporting (Beneish, 1999; Glancy and Yadav, 2011; Humpherys et al., 2011; Ngai et al., 2011; Perols, 2011; Perols et al., 2017; Zhou and Kapoor, 2011); evaluation of earnings quality (Dechow et al., 2010) or prediction of a particular type of auditor qualification, such as going-concern qualification (Bell and Tabor, 1991; Maggina and Tsaklanganos, 2011; Monroe and Teh, 2009; Mutchler and Hopwood, 1997; Yeh et al., 2014).

In Table 1, we present prior studies aimed at modeling qualified audit opinions in chronological order.

Paper	Sample Characteristics	Method	Factors Associated with Qualified/ Modified Audit Opinions
Dopuch et al. (1987)	275 qualified and 441 unqualified opinions from 1969–1980 (US public companies)	Probit regression	◆ Being listed on stock exchange for less than five years ◆ Low stock returns of the stocks in the industry ◆ Increase in variability of stock returns ◆ Decrease in value of beta coefficient ◆ Net income negative in current year ◆ Increase in financial leverage ◆ Decrease in ratio of receivables to total assets
Kinney and McDaniel (1989)	24 qualified and 49 unqualified opinions from 1976–1985 (US public companies)	T-test and regression	◆ Listed over-the-cou nter ◆ Small companies ◆ Low profitability ◆ High financial leverage ◆ Low growth ◆ Negative stock returns
Krishnan et al. (1996)	163 qualified and 1,674 unqualified audit opinions from 1986–1988 (US public companies)	One-stage and two-stage probit model	◆ Low share of receivables in total assets ◆ High financial leverage ◆ Small size (assets) ◆ Negative net income in current year ◆ High volatility (SD) of stock returns ◆ Low stock returns ◆ Higher levels of outsider ownership ◆ Small size (assets) relative to auditor client portfolio ◆ Low growth (in assets) ◆ Being listed on stock exchange for longer time ◆ Probability of litigation as measured by the variable constructed by Stice (1991)
Laitinen and Laitinen (1998)	8 qualified and 103 unqualified audit opinions from 1992–1994 (public companies listed on Helsinki stock exchange)	Kruskal-Wallis test and logistic regression	◆ Low profitability ◆ Low growth ◆ High leverage ◆ High credit risk ◆ Small companies
Francis and Krishnan (1999)	284 modified and 2,324 unmodified from 1986–1987 (US companies)	Probit model	◆ High leverage ◆ Net income negative in current year ◆ Small companies ◆ High stock price volatility ◆ Low stock returns ◆ Being listed on stock exchange longer ◆ High accruals as measured by the difference between net income and cash flow from operations

Bartov et al. (2000)	173 qualified and 173 matched-pair unqualified opinions from the period 1980–1997 (US companies)	Contingency-table test and logistic regression	High discretionary accruals as measured by: ◆ The modified Jones model ◆ The cross-sectional Jones model ◆ The cross-sectional modified Jones model
Spathis et al. (2003)	50 qualified and 50 matched-pair unqualified audit opinions from 1997–1999 (Greek companies)	Multicriteria decision aid classification method (UTADIS), discriminant analysis, and logistic regression	◆ High share of receivables in sales ◆ Low profitability (net profit/total assets) ◆ Low liquidity (working capital/total assets) ◆ Low asset turnover (sales/total assets) ◆ High credit risk (Z score)
Doumpos et al. (2005)	859 qualified and 5,189 unqualified audit opinions from 1998–2003 (UK and Irish companies)	Support vector machines (SVMs)	◆ Poor credit rating ◆ Low liquidity ◆ Low profitability ◆ Low fixed assets turnover ◆ Low growth
Caramanis and Spathis (2006)	162 qualified and 23 unqualified audit opinions from 2001 (Greek public companies)	Ordinary least squares (OLS) regression	◆ Low profitability as measured by profit margin to total assets ◆ Low liquidity as measured by current assets divided by current liabilities
Gaganis and Pasiouras (2006)	114 qualified and 114 unqualified audit opinions from 2003–2004 (UK and Irish companies)	Discriminant analysis and logistic regression	◆ Large size (assets) ◆ Low profitability ◆ Low growth ◆ A non-Big N auditor
Gaganis, Pasiouras, and Doumpos (2007)	264 qualified and 3,069 unqualified audit opinions from 1997–2004 (public companies listed on the London stock exchange)	Probabilistic neural network (PNN), artificial neural network (ANN), and logistic regression	◆ Low profitability ◆ Poor credit rating ◆ Extreme values (both high and low) of days payable outstanding ◆ Small companies
Gaganis, Pasiouras, Spathis, et al. (2007)	980 qualified and 4,296 unqualified audit opinions from 1998–2003 (UK companies)	K-nearest neighbors (k-NN), discriminant analysis, and logistic regression	◆ High credit risk ◆ High liquidity (current ratio) ◆ High leverage ◆ Low growth ◆ Low profitability
Kirkos et al. (2007)	225 qualified and 225 unqualified audit opinions from 1995–2004 (UK and Irish companies)	C4.5 decision tree, multilayer perceptron (MLP), and Bayesian belief network	◆ High credit risk (Z score) ◆ Low profitability ◆ High leverage ◆ Low revenue
Pourheydari et al. (2012)	347 qualified and 671 unqualified audit opinions from 2001–2007 (public companies listed on Tehran stock exchange)	Multilayer perceptron (MLP), radial basis function (RBF), probabilistic neural network (PNN), and logistic regression	◆ Low liquidity as measured by working capital ◆ Low profitability as measured by net income per employee, EBIT margin, and pre-tax profit margin ◆ Higher values of days sales outstanding ◆ High credit risk ◆ Low growth

Saif et al. (2012)	708 qualified and 310 unqualified audit opinions from 2001–2007 (public companies listed on Tehran stock exchange)	Support vector machine (SVM) and decision trees	◆ Low profitability (net profit per employee) ◆ Small size (number of employees, sales) ◆ Low growth ◆ High cash flow from investing activities relative to revenue
Saif et al. (2013)	708 qualified and 310 unqualified audit opinions from 2001–2007 (public companies listed on Tehran stock exchange)	Multilayer perceptron (MLP) and decision trees	◆ Low inventory turnover ◆ High liquidity (current ratio)
Yasar et al. (2015)	55 qualified and 55 unqualified audit opinions from 2010–2013 (public companies listed on Istanbul stock exchange)	Discriminant analysis, logistic regression, and C5.0 decision tree	◆ Low liquidity ◆ High leverage ◆ Low productivity of operations ◆ Low profitability
Zdolšek et al. (2015)	12 qualified and 293 unqualified audit opinions from 2009 (Slovenian companies)	Logistic regression	◆ High leverage ◆ Low liquidity ◆ Low efficiency ◆ Low profitability
Fernández-Gámez et al. (2016)	78 qualified and 369 unqualified audit opinions from 2008–2010 (Spanish public companies)	Probabilistic neural network (PNN) and multilayer perceptron (MLP)	◆ Small size ◆ Low liquidity ◆ Low profitability ◆ Low solvency (EBITDA/total liabilities) ◆ Low productivity

Table 1. Review of Studies Aimed at Modeling Qualified Audit Opinions

Characteristics of clients, both financial and non-financial, are the most important group of predictors used in prior studies, which is expected, given that the main reasons for issuing a modified audit opinion are materially misstated financial reports and substantial uncertainties related to the client's operations.[3] Characteristics of audit firms are somewhat less frequently considered as predictors, with the auditor's membership of Big N being the most researched effect. Most studies have found that the auditor's membership of Big N is not predictive of qualifications (Kirkos et al., 2007; Krishnan and Krishnan, 1996; Mutchler and Hopwood, 1997; Ruiz-Barbadillo et al., 2004).[4]

The main general conclusion, arising from the results of prior research, is that companies with poor financial conditions (i.e. high leverage, low liquidity, high credit risk, small size, low efficiency, low growth) get qualified audit opinions more frequently. There are several potential reasons for this. The first is that poor financial prospects on their own, when inadequately disclosed as a going concern in

3 The potential reasons for issuing a modified audit opinion, in order of decreasing frequency, are: a) materially misstated financial reports (reports do not conform to the international standards of financial reporting) or inadequate disclosure; b) substantial uncertainties exist about the entity's ability to continue operations (going concern) or some other important aspect of client's operations; c) auditor's inability to obtain enough evidence to verify the information presented in financial reports (scope limitation); and d) auditor's lack of independence.

4 One study (Gaganis and Pasiouras, 2006) reports that the auditor's membership of Big N decreases the chances of qualified opinion.

the financial statements, can be a reason for expressing a qualified or adverse opinion (IAASB ISA 570, 2013, para. 20). Secondly, as noted by Beneish (1999), companies facing poor prospects have greater incentives for earnings manipulation. Empirical findings do not contradict this hypothesis, as more earnings management cases are found to be directed towards overstatement than understatement of financial result (Jones et al., 2008, p. 506; Kinney and McDaniel, 1989, p. 84). Thirdly, companies with poor financial conditions may also have low-quality human resources in their accounting departments, resulting in more errors. Finally, the auditors of poorly performing firms are more likely to decide that contingencies of a given magnitude are material (Dopuch et al., 1987, p. 437), conceivably because of a greater risk of stockholder lawsuits (Kinney and McDaniel, 1989, p. 72) and relatively low importance that such clients have in their portfolio (Krishnan et al., 1996).

Regarding the modeling techniques being used, a shift is observable from the classic statistical tools, such as probit and logistic regression, to more contemporary machine learning tools, such as artificial neural networks, decision trees and k-nearest neighbors. It is acknowledged in the literature that the predictive performance of specific classifiers varies widely across different types of classification task, depending mostly on the characteristics of the data, such as size, dimensionality, and structure. From the results of the prior research, it is difficult to conclude which set of tools is more effective for the task of predicting the type of audit opinion. The key factors that hinder the comparability of the reported classification accuracies across the studies are the following:

- Different outcome variables are used. Researchers use both qualified (Dopuch et al., 1987; Krishnan and Krishnan, 1996) and modified (Francis and Krishnan, 1999) types of audit opinion as the target category. Both categories are legitimate choices, but the classification accuracies between the two cannot be directly compared.
- Different predictors are used. Most researchers use only financial metrics, but some use market data (Dopuch et al., 1987; Francis and Krishnan, 1999; Krishnan and Krishnan, 1996), non-financial data, or latent constructs (Fernández-Gámez et al., 2016) as well. Even when nominally using the very same predictors, the calculations are likely to differ due to the differences in national reporting frameworks.
- Samples from different countries and periods and of different sizes are used. The difficulty of prediction is expected to vary across periods and countries. Also, the accuracies reported in the studies that have used small samples may have been overestimated due to overfitting.
- Different model validation methods are employed. Results obtained using bootstrap replications (Spathis et al., 2003; Zdolšek et al., 2015), k-fold cross-validation (Kirkos et al., 2007), and differently constructed hold-out samples (Gaganis, Pasiouras, and Doumpos, 2007; Pourheydari et al., 2012) are not directly comparable. Some studies even report classification accuracy in training set, which is known to be overoptimistic.
- Different measures of predictive performance are reported. Studies report both average (Doumpos et al., 2005; Gaganis, Pasiouras, Spathis et al., 2007) and overall classification accuracy (Gaganis, Pasiouras, and Doumpos, 2007; Kirkos et al., 2007; Pourheydari et al., 2012; Spathis et al., 2003; Yasar et al., 2015), which are not directly comparable. Furthermore, as the proportions of types of audit opinion vary across countries and periods, achieving a high overall predictive accuracy is easier in countries and periods where one type of opinion is dominant — such as in a study done by Zdolšek et al. (2015), where a sample from Slovenia with only a 3.93 percent share of qualified opinions in all opinions is used, or in a study by Caramanis and Spathis (2006), where

a sample from Greece with 87.57 percent of qualified opinions in all opinions is used — than in those where the two types are more evenly distributed, such as in the study by Dopuch et al. (1987), where the share of qualified opinions was 38.41 percent. No studies report the Kappa metric, which corrects for the differences in proportions of types of audit opinion.

When we try to narrow down the focus to studies that have reported comparative predictive performance of statistical and machine learning techniques on the same data sets, using the same outcome variable and the same measure of accuracy, the following two studies appear to be the most relevant:

1. Gaganis, Pasiouras, and Doumpos (2007, p. 122) reported overall accuracies of 84.35 percent, 80.55 percent, and 86.14 percent achieved by probabilistic neural network (PNN), artificial neural network (ANN) and logistic regression, respectively.
2. Pourheydari et al. (2012, p. 11086) reported overall accuracies of 87.75 percent, 84.81 percent, 78.44 percent, and 77.60 per cent achieved by multilayer perceptron (MLP), probabilistic neural network (PNN), radial basis function (RBF) neural networks, and logistic regression, respectively.[5]

The results of the two studies are conflicting: logistic regression achieves higher accuracy in the first study, while artificial neural networks are shown to be more accurate in the second. It should be noted, however, that the confidence intervals for the classification accuracies have not been reported, meaning that the possibility that the observed within-study differences are insignificant has not been ruled out.

Most importantly, even if the findings of the prior studies were consistent, we argue that the conclusions drawn from such studies would not be definite, because the respective strengths of the two sets of tools have not been fully exploited.

Specifically, the studies relying on statistical methods have missed the opportunity to model the systematic effects of individual auditors and clients. Given that the differences in auditors' propensities to issue (e.g. auditor's reporting conservatism) and clients' propensities to receive (e.g. client's propensity to manage earnings or the quality of its accounting department) a modified audit opinion are expected to exist and persist over time, the failure to model the individual effects (i.e. to use the prior information when available) has likely resulted in suboptimal predictive performance of the statistical models. At the root of this suboptimal modeling choice there may be the unfortunate circumstance that most studies — with the exception of a study by Doumpos et al. (2005) — have not differentiated between cases where prior information on client and auditor (i.e. audit and/or financial reports from prior periods) is available from cases where such information is not available.[6] Differentiating between these two real-life scenarios is extremely important for the task of prediction, from both a theoretical and practical perspective. The inclusion of the systematic effects is not only expected to improve the predictive performance of the statistical models, but also to yield unbiased estimates of the coefficients with the explanatory variables.

5 Yasar et al. (2015) reported overall accuracies of 87.3 percent, 92.7 percent, and 98.2 percent, achieved by discriminant analysis, logistic regression, and C5.0 decision tree, respectively. However, this study is disregarded, since the reported accuracies are measured in different samples (in the training sample for the former two, and in a test sample, comprising only 26 observations, for the third one). Also, Spathis et al. (2003, p. 278) reported overall accuracies of 78.83 percent, 74.34 percent, and 74.70 percent based on 200 bootstrap replications, achieved by the multicriteria decision aid classification method (UTADIS), linear discriminant analysis, and logistic regression, respectively. This study is omitted, however, since UTADIS is neither a statistical nor machine learning technique.

6 A transparent and informative model validation technique was conducted by Doumpos et al. (2005, p. 211), who reported overall accuracies of 84.58 percent, 84.84 percent, and 84.84 percent achieved by support vector machines with linear, RBF, and quadratic kernel, respectively.

On the other hand, the studies relying on machine learning techniques have failed to fully employ their feature extraction capabilities. These studies have typically relied on the traditional theory-driven metrics, developed for the task of financial analysis, instead of using all the available information from the complete set of financial reports to boost the predictive performance (machine learning techniques are expected to make relatively better use of the data available in the complete set of financial statements, since they are designed to be capable of handling more predictors and modeling more complex relationships) and extract novel, data-driven predictors that might be more relevant for the task of predicting auditor opinions.

Finally, no study has made an attempt to combine the respective strong points of the two approaches to construct a hybrid model that would make the best use of the data available while preserving the desired level of parsimony and interpretability of the model.

Based on the review of prior research in the field, we recognize the need for a study which will: 1) develop models from the realms of statistics and machine learning that fully exploit their respective strong points, which are modeling of the systematic effects on one side, and flexibility and feature extraction capabilities on the other; 2) consider the possibility of developing hybrid models that would integrate the best of both worlds; 3) test the predictive performance of the models, while differentiating between the setting where prior information on the client is available and the setting where prior information is not available; 4) report confidence intervals for the predictive accuracies, in addition to their point estimates; and 5) do everything in a reproducible manner, and without relying on the assumption that long time series of historical reports or market data are available, to allow researchers and professionals around the globe to build models and expert systems for the assessment of the probability of material misstatements in financial reports that will suit their needs.

MATERIALS AND METHODS

Data

The data from 13,561 complete sets of annual financial statements for 4,701 companies are combined with the data from the corresponding audit reports, forming an unbalanced panel data set. The client companies included in the sample represent a supermajority of medium- and large-sized companies registered in the Republic of Serbia. The information on the auditor firm name and the type of audit opinion is hand-collected from the audit reports issued by 64 audit firms (Big 4 plus 60 other audit firms), which, again, represents a supermajority of all the auditor firms registered in this country. To the best of our knowledge, this is the largest data set used in the literature devoted to predicting the type of audit opinion.

In the total sample of audit opinions (13,561), the following frequencies of the four main types of audit opinions are observed: adverse opinion (71), disclaimer of opinion (644), qualified opinion (3,706), and unqualified opinion (9,140). We present the absolute and relative frequencies of the specific types of audit opinion by period in Table 2.

	Absolute Frequencies by Period				Relative Frequencies by Period			
	2010	**2011**	**2012**	**2013**	**2010**	**2011**	**2012**	**2013**
Adverse opinion	9	15	22	25	0.27%	0.45%	0.60%	0.78%
Disclaimer of opinion	127	140	200	177	3.76%	4.24%	5.44%	5.52%
Qualified opinion	910	900	1,011	885	26.94%	27.26%	27.49%	27.62%
Unqualified opinion	2,332	2,246	2,445	2,117	69.03%	68.04%	66.48%	66.07%
Total	**3,378**	**3,301**	**3,678**	**3,204**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

Table 2. Observed Frequencies for Each Type of Audit Opinion by Period

Since adverse opinions and disclaimers of opinion are relatively rare, we combine them with qualified opinions to establish a 'modified opinion' category. The absolute and relative frequencies of the modified and unmodified opinions are shown in the following Table 3.

	Absolute Frequencies by Period				Relative Frequencies by Period			
	2010	**2011**	**2012**	**2013**	**2010**	**2011**	**2012**	**2013**
Modified opinion	1,046	1,055	1,233	1,087	30.97%	31.96%	33.52%	33.93%
Unmodified opinion	2,332	2,246	2,445	2,117	69.03%	68.04%	66.48%	66.07%
Total	**3,378**	**3,301**	**3,678**	**3,204**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

Table 3. Frequencies of Modified and Unmodified Audit Opinions by Period as Observed in the Sample

The observed frequency of the two main types of audit opinions by industry classification is presented in Appendix A.

In the complete data set, there are 9,140 (67.4%) unmodified and 4,421 (32.6%) modified audit opinions, indicating no significant class imbalance.

The dependent variable in all the predictive models is the dichotomous variable that indicates the type of audit report received by company *i* in period *t*. A value of 0 indicates an unmodified (unqualified) audit opinion while a value of 1 indicates a modified (qualified, disclaimer, or adverse) audit opinion.

Given the above-mentioned definition of the dependent variable, we model the probability that either: a) the auditor issues a disclaimer of opinion (i.e. withdraws) for a valid reason, which occurs in around five percent of all cases in the sample; or b) the financial statements do not meet one or more of the requirements for issuing an unqualified opinion, and hence contain either non-pervasive (for a qualified opinion) or pervasive (for an adverse opinion) material misstatements, which occurs in around 28 percent of all cases in the sample.[7]

7 The national Law on Auditing requires all auditors to apply the International Standards on Auditing, the International Standard on Quality Control, and the related standards published by the International Auditing and Assurance Standards Board of the International Federation of Accountants. Accordingly, an unqualified/unmodified audit opinion has the same meaning as in all previous studies, and implies that the financial statements:
a) Have been prepared in accordance with IFRS;
b) Comply with relevant statutory requirements and regulations;
c) Provide adequate disclosure of all material matters, and
d) Provide adequate disclosure of any changes in the accounting principles or in the method of their application and the effects thereof.
If any of the listed conditions is not met, a modified opinion is issued.

All computations and modeling are conducted within the R software environment (R Core Team, 2017).

The Predictors

Initially, we consider a comprehensive set of theory-driven metrics, derived from prior studies. In addition to that, to make a better use of quantitative data obtainable from clients' complete sets of financial statements, we design and conduct a feature generation procedure, and subsequently use the feature extraction capabilities of machine learning algorithms to identify features (predictors) highly relevant for the task of predicting auditor opinions.

Theory-Driven Predictors

Among the theory-driven predictors, there are 31 financial variables that, based on the previous research, are expected to be associated with qualified audit opinion. The potentially relevant predictors are classified into eight groups: 1) size (Francis and Krishnan, 1999; Kinney and McDaniel, 1989; Kirkos et al., 2007); 2) profitability (Caramanis and Spathis, 2006; Dopuch et al., 1987; Francis and Krishnan, 1999; Kinney and McDaniel, 1989; Kirkos et al., 2007; Pourheydari et al., 2012); 3) liquidity (Caramanis and Spathis, 2006; Pourheydari et al., 2012); 4) leverage (Dopuch et al., 1987; Francis and Krishnan, 1999; Kinney and McDaniel, 1989; Kirkos et al., 2007); 5) cash conversion cycles (Gaganis, Pasiouras, and Doumpos, 2007; Pourheydari et al., 2012); 6) credit risk (Gaganis, Pasiouras, and Doumpos, 2007; Kirkos et al., 2007; Pourheydari et al., 2012); 7) earnings quality/accruals (Bartov et al., 2000; Francis & Krishnan, 1999), and 8) other (Gaganis, Pasiouras, and Doumpos, 2007).[8] Details on the calculation of the theory-driven predictors are presented in Table 4.

Predictor Group	Predictor Name	Calculation [item ids (as assigned in the layout of chart of accounts) in superscript]
Size	Ln total assets	Ln (operating assets[022])
	Ln revenue	Ln (operating revenue[201])
Profitability	Net result	Net profit[229] - Net loss[230]
	EBIT	Profit before tax[223] - Loss before tax[224] + Interest expenses[667]
	EBITDA	EBIT + Depreciation[661]
	Operating result	Operating profit[213] - Operating loss[214]
	Return on assets	EBIT / Operating assets[022]
	Return on equity	Net result / Capital[101]
	Return on invested capital	EBIT × (1 - Corporate tax rate [15%]) / (Capital[101] + Long-term liabilities[113] + Short-term financial liabilities[117] - Excess cash [any cash over 3% of operating revenue])
	Return on capital invested in core business	Operating result / (Operating assets[022] - Long-term financial investments[009] - Short-term financial investments[018] - Excess cash [any cash over 3% of operating revenue])
	Operating margin	Operating result / Operating revenue[201]
	Net margin	Net result / Operating revenue[201]

8 The group named 'other' includes an average net monthly salary (a proxy for employees' quality and motivation level) and the percentage of foreign ownership, both of which have been calculated based on the data presented in the financial statements.

Liquidity	Net working capital	Current assets[012] - Current liabilities[116]
	Net working capital to assets	Net working capital / Operating assets[022]
	Quick ratio	(Receivables[016] + Receivables from overpaid corporate income tax[017] + Cash and cash equivalents[019]) / Current liabilities[116]
Leverage	Equity to total liabilities	Capital[101] / (Long-term provisions and liabilities[111] + Deferred tax liabilities[123])
	Debt ratio	(Long-term loans[114] + Other long-term liabilities[115] + Short-term financial liabilities[117]) / Operating assets[022]
Cash Conversion	Days sales outstanding	Accounts receivables (sales)[639] / Sales revenue[202] × 365 / (1 + VAT rate [18%])
	Days payable outstanding	Operating liabilities (end of the year)[640] / Operating liabilities (credit turnover without opening balance)[643] × 365
	Days sales of inventory	(Raw materials[616] + Work in process[617] + Finished goods[618] + Merchandise inventory[619]) / Operating liabilities (credit turnover without opening balance)[643] × 365 / (1 + VAT rate [18%])
	Cash conversion cycle	Days sales outstanding - Days payable outstanding + Days sales of inventory
Credit Risk	Z score for private companies	Z' = 0.717 × Net working capital to assets + 0.847 × (Retained profits[108] / Operating assets[022]) + 3.107 × Return on assets + 0.420 × Equity to total liabilities + 0.998 × (Operating revenue[201] / Operating assets[022])
	Z score for non-manufacturers and emerging markets	Z'' = 6.56 × Net working capital to assets + 3.26 × (Retained profits[108] / Operating assets[022]) + 6.72 × Return on assets + 1.05 × Equity to total liabilities
	Zmijewski score	- 4.336 - 4.513 × (Net income / Operating assets[022]) + 5.679 × ((Long-term provisions and liabilities[111] + Deferred tax liabilities[123]) / Operating assets[022]) + 0.004 × (Current assets[012] / Current liabilities[116])
	Shumway score	- 6.307 × (Net result / Operating assets[022]) + 4.068 × ((Long-term provisions and liabilities[111] + Deferred tax liabilities[123]) / Operating assets[022]) - 0.158 × (Current assets[012] / Current liabilities[116])
Earnings Quality	Free cash flow to equity (FCFE)	Cash balance at the end of the period[343] - Cash balance at the beginning of the period[340] + Dividends paid[333] + Buy-up treasury shares and stakes[330] - Increase in the capital stock[326]
	Free cash flow to firm (FCFF)	FCFE - Long-term and short-term borrowings (net inflows)[327] + (Interest paid[308] × (1 - corporate tax rate [15%]))
	FCFE minus net result to revenue	(FCFE - Net result) / Operating revenue[201]
	Cash flow from operations minus operating result to revenue	((Net cash inflow from operating activities[311] - Net cash outflow from operating activities[312]) - (Operating result)) / Operating revenue[201]
Other	Percentage of foreign ownership	(Share capital - % held by foreign investors[624] + Limited liability capital - % held by foreign investors[626] + General and limited partnership capital - % held by foreign investors[628]) / Total paid in capital[633]
	Average net monthly salary in EUR	Liabilities for net salaries and wages (credit turnover excluding opening balance)[644] / Average number of employees based on end-of-the-month data[605] / 12 × 1000

Table 4. Theory-Driven Predictors

Descriptive statistics for the theory-driven predictors are presented in Appendix B (Table B1).

Data-Driven Predictors

In the second stage, we generate a set of data-driven predictors. One new predictor variable is calculated from each pair of 391 items available from the complete set of financial reports (i.e. balance sheet, income statement, cash flow statement, statement of changes in equity, and statistical annex) by division. After discarding redundant variables (reciprocals) and constants (variables divided by themselves), the procedure yields 76,245 predictors which, when combined with the original 391 items, add up to a total of 76,636 predictors. Since the calculation of features often involves a division, where the value of the denominator (a numeric item from financial statements) is 0, the resulting values include positive and negative infinite values. This problem is common in calculating theory-driven financial metrics, but is even more pronounced in the case of data-driven features. The positive and the negative infinite values are replaced with the maximum and minimum observed values within the matching predictor. Furthermore, all the predictors with little or no variation (less than 95 percent unique values) are discarded. After applying the described procedure, 1,515 data-driven predictors remain in the data set.

The combined data set, which includes both theory- and data-driven predictors, comprises a total of 1,546 potentially relevant predictors.

Predictive Models

We build twelve predictive models. All the predictors are preprocessed using standardization (centering and scaling). To facilitate a fair comparison of predictive performance, all the models that need tuning are tuned in a consistent way, using 10-fold cross-validation and Area under the receiver operating characteristic curve (AUROC) as the performance metric. The complete preprocessing, model fitting and tuning, as well as comparative analysis of predictive performance are carried using R's 'caret' package (Kuhn, 2017).[9] The information on the predictive models used, including software implementation which is used and the parameters that need to be tuned, are presented in Table 5.

Method	Software Implementation	Tuning Parameters
C5.0	R's package 'C50' (Kuhn and Ross, 2017)	*trials* (the number of boosting iterations) *model* (logical: should the tree be decomposed into a rule-based model?) *winnow* (logical: should feature selection be used?)
Random Forest	R's package 'randomForest' (Liaw and Wiener, 2002)	*mtry* (number of variables randomly sampled as candidates at each split)
Regularized Random Forest	R's package 'RRF' (Deng, 2013; Deng and Runger, 2012, 2013)	*mtry* (number of variables randomly sampled as candidates at each split) *coefReg* (the coefficient(s) of regularization) *coefImp* (importance coefficient)

9 Only the training data set is used for model tuning. The data are preprocessed separately for tuning and evaluation. Each model is tuned for the two training sets separately.

Stochastic Gradient Boosting	R's package 'gbm' (Ridgeway, 2017)	*n.trees* (number of trees used in the prediction) *interaction.depth* (the maximum depth of variable interactions) *shrinkage* (a shrinkage parameter applied to each tree in the expansion) *n.minobsinnode* (minimum number of observations in the trees terminal nodes)
Extreme Gradient Boosting	R's package 'xgboost' (Chen et al., 2017)	*nrounds* (the max number of iterations) *max_depth* (maximum depth of a tree) *eta* (the learning rate) *gamma* (minimum loss reduction required to make a further partition on a leaf node of the tree) *colsample_bytree* (subsample ratio of columns when constructing each tree) *min_child_weight* (minimum sum of instance weight [hessian] needed in a child) *subsample* (subsample ratio of the training instance)
K-Nearest Neighbors	R's package 'class' (Venables and Ripley, 2002)	*k* (number of neighbors considered)
Multilayer Perceptron	R's package 'RSNNS' (Bergmeir and Benitez, 2012)	*size* (number of units in the hidden layer(s))
Support Vector Machine with Radial Basis Function Kernel	R's package 'kernlab' (Karatzoglou et al., 2004)	*sigma* (inverse kernel width) *C* (cost of constraints violation)
Linear Discriminant Analysis	R's package 'MASS' (Venables and Ripley, 2002)	None
Logistic Regression	R's package 'stats' (R Core Team, 2017)	None
Probit Regression	R's package 'stats' (R Core Team, 2017)	None
Mixed-Effects Logistic Regression	R's package 'brms' (Bürkner, 2017)	None

Table 5. Information on the Predictive Models

Seven of the twelve listed models – namely, C5.0, k-nearest neighbors, multilayer perceptron, support vector machine, linear discriminant analysis, logistic regression, and probit regression — have been used for the task of predicting the type of audit opinion in prior studies, while the remaining five — random forest, regularized random forest, stochastic gradient boosting, extreme gradient boosting, and mixed-effects logistic regression — are new to this stream of literature.

We briefly explain the rationale for using mixed-effects logistic regression as a predictive model in this study. Firstly, using a mixed-effects logistic regression framework with the company- and auditor-specific effects included as random effects allows the prior information on client's propensity to receive, and auditor's propensity to issue, a modified opinion to be effectively accounted for when available, which is expected to result in a significant improvement in classification accuracy, especially in the out-of-time sample which comprises clients represented in the training sample. It also allows the model to provide predictions for instances with unseen (new) companies and auditors, which

would be problematic were logistic regression with fixed effects used instead. Mixed-effects logistic regression also demonstrated higher accuracy than fixed-effects logistic regression (the results of the fixed-effects models are discussed in the results section), indicating that regularization of the individual effects, conducted within the random-effect framework, improves model performance. The output of a preliminary variance components model (a model with no predictors included) indicates a significant variability in both company- and auditor-specific effects, justifying the described modeling choice.

The dichotomous dependent variable is modeled using Bernoulli distribution and the logit link function:

$$\text{Modified}_{ijt} \sim \text{Bernoulli}\ (\frac{1}{1+\exp\ (-x_{ijt})}), \text{where}$$

$$x_{ijt} = \alpha + u_i + u_j + \sum_{n=1}^{k} Predictor_{n_{it}} \times \beta_n + e_{i\,jt}$$

In the equation above, α is the population-level intercept, u_i and u_j are company- and auditor-specific varying intercepts respectively, β_n is the regression coefficient with nth predictor, and $e_{i\,jt}$ is the error term. The model has been fitted within the Bayesian framework, with priors on all fixed effects set to a Gaussian $N(0,1)$ distribution.

Model Validation Procedure

To assess the classification accuracy of the described predictive models, we employ a method similar to that described by Doumpos et al. (2005, p. 210). The combined data set is first split into two parts: a) the first set, comprising two thirds (66.66 percent) of the companies, randomly selected; and b) the second set, comprising the remaining third (33.33 percent) of the companies in the sample. Next, the first set is split into a training set, including the observations from the period 2010–2012, and an out-of-time test set (OOT), which includes only the observations from the year 2013. Lastly, the observations from the second set from the year 2013 are used to form the out-of-sample and out-of-time test set (OOS and OOT). The described sub-setting procedure is illustrated in Figure 1.

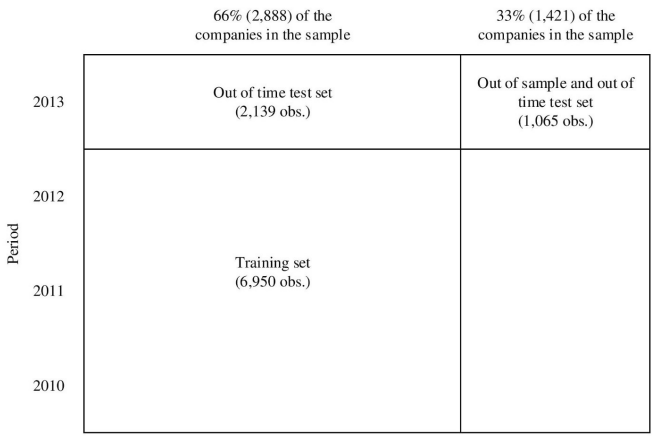

Figure 1. Sub-Setting the Data

Using model validation terminology, both test sets are out-of-time, meaning that they comprise only the observations from later periods than those represented in the training set. The purpose of the out-of-time (or temporal) validation is to emulate the typical situation that occurs in practice. That is to say, classification models are trained on historical data yet need to be applied in a new reporting period, which may be different from the observed periods in multiples respects, such as typical levels of financial metrics (economic cycles), firms' propensities for earnings manipulation, new accounting or auditing standards and regulations, new techniques and patterns of earnings manipulation, etc. In contrast to that, using the whole data set to train the classification models and employing a standard cross-validation procedure to assess their accuracy would result in overly optimistic estimates of accuracy, as the classification models would be allowed to learn from the information presented in financial and audit reports from the current reporting period, and those reports are typically not available at the beginning of a new audit season.

While being out-of-time, the first test set comprises the same companies that are represented in the training set. As having historical data on the client (auditor reports and financial metrics from previous periods) is a significant advantage that can be efficiently used by some models to achieve improved classification accuracy, that test set is employed for assessing the expected model classification accuracy in the setting when auditor reports and financial metrics from previous periods for the client under consideration are available.

On the other hand, in addition to being out-of-time, the second test set is also out-of-sample, in the sense that it exclusively comprises the observations from companies that are not represented in the training set. Consequently, this test set is used to assess the expected classification accuracy of the models in the upcoming periods in the setting when no prior information on the client is available (e.g. a newly established company).

Both scenarios described above occur in real life, and since classification accuracies are expected to differ significantly between the two, it is valuable to have a realistic assessment of the expected classification accuracy for both.

RESULTS

Comparative Analysis of the Predictive Performance of the Individual Models

Table 6 summarizes the comparative performance of all the twelve predictive models in both test sets. Because of the differential baseline (no information) rates in the two test samples (0.6704 vs. 0.6413), in addition to classification accuracy at 50% cut-off, for each predictive model, we report the values of Cohen's Kappa (a chance-corrected measure of proportion agreement; for details, see Cohen, 1960) with the corresponding 95% confidence intervals at 50% cut-off, and the AUROC metric that summarizes overall model performance over all possible cut-offs. For performance comparison, we primarily rely on AUROC.

Model	With Theory-Driven Predictors				With Theory- and Data-Driven Predictors			
	Out-of-time (No information rate: 0.6704)		Out-of-sample and out-of-time (No information rate: 0.6413)		Out-of-time (No information rate: 0.6704)		Out-of-sample and out-of-time (No information rate: 0.6413)	
	Kappa (95%CI)	Accuracy at 50% cut-off / Area under the curve	Kappa (95%CI)	Accuracy at 50% cut-off / Area under the curve	Kappa (95%CI)	Accuracy at 50% cut-off / Area under the curve	Kappa (95%CI)	Accuracy at 50% cut-off / Area under the curve
C5.0	0.4150 (0.3697–0.4602)	0.7691 / **0.7910**	0.3824 (0.3189–0.4459)	0.7455 / **0.7733**	0.4718 (0.4293–0.5143)	0.7831 / **0.8172**	0.4043 (0.3429–0.4657)	0.7465 / **0.7702**
Random Forest	0.4405 (0.3960– 0.4849)	0.7784 / **0.8124**	0.3619 (0.2975–0.4264)	0.7390 / **0.7754**	0.4894 (0.4467–0.5322)	0.7962 / **0.8526**	0.4052 (0.3425–0.4679)	0.7549 / **0.7869**
Regularized Random Forest	0.4377 (0.3930–0.4824)	0.7784 / **0.8130**	0.3476 (0.2831–0.4121)	0.7305 / **0.7704**	0.5020 (0.4603–0.5436)	0.7957 / **0.8436**	0.4109 (0.3496–0.4722)	0.7502 / **0.7907**
Gradient Boosting Machine	0.3988 (0.3529–0.4446)	0.7644 / **0.7903**	0.3540 (0.2890–0.4190)	0.7371 / **0.7737**	0.4457 (0.4026–0.4889)	0.7714 / **0.8052**	0.3909 (0.3295–0.4523)	0.7380 / **0.7800**
Extreme Gradient Boosting	0.3920 (0.3460–0.4382)	0.7620 / **0.7895**	0.3540 (0.2890–0.4190)	0.7371 / **0.7708**	0.5022 (0.4613–0.5431)	0.7901 / **0.8215**	0.3650 (0.3041–0.4259)	0.7183 / **0.7393**
K-Nearest Neighbors	0.3893 (0.3402–0.4297)	0.7466 / **0.7627**	0.2825 (0.2174–0.3475)	0.6948 / **0.6931**	0.3716 (0.3253–0.4178)	0.7508 / **0.7867**	0.2090 (0.1411–0.2769)	0.6714 / **0.6673**
Multilayer Perceptron	0.3515 (0.3039–0.3991)	0.7499 / **0.7885**	0.2808 (0.2127–0.3490)	0.7136 / **0.7574**	0.3365 (0.2879–0.3850)	0.7489 / **0.7808**	0.2744 (0.2066–0.3423)	0.7080 / **0.7485**
Support Vector Machine	0.2987 (0.2487–0.3488)	0.7396 / **0.7474**	0.2379 (0.1680–0.3077)	0.6995 / **0.7269**	Models did not converge because of the large number of predictors			
Linear Discriminant Analysis	0.2914 (0.2413–0.3414)	0.7354 / **0.7717**	0.2615 (0.1919–0.3310)	0.7108 / **0.7642**				
Logistic Regression	0.3166 (0.2672–0.3660)	0.7443 / **0.7759**	0.2855 (0.2170–0.3540)	0.7183 / **0.7657**				
Probit Regression	0.3011 (0.2512–0.3511)	0.7396 / **0.7759**	0.2759 (0.2070–0.3449)	0.7155 / **0.7696**				
Mixed-Effects Logistic Regression	0.5779 (0.5393–0.6165)	0.8233 / **0.8818**	0.3053 (0.2381–0.3726)	0.7221 / **0.7597**				

Table 6. Comparative Predictive Performance of the Twelve Predictive Models

The first thing to be noted is that classification performance indeed differs meaningfully between the two scenarios. Regardless of the technique being used, it is easier to achieve better performance in the out-of-time than in the out-of-sample and out-of-time test set. Even the machine learning techniques, which do not explicitly account for the systematic effects present in the data, achieve better performance in the out-of-time test set than in the out-of-sample and out-of-time test set, which can primarily be attributed to overfitting.

Another key insight is that the comparative performance of classifiers depends heavily on the availability of prior information. When prior information is available, mixed-effects logistic regression effectively accounts for it, providing the most accurate predictions. In the absence of prior information, the machine learning algorithms — particularly the tree-based ones such as random forest, regularized random forest, gradient boosting machine, and C5.0 — provide more accurate predictions by making relatively better use of predictors.

We proceed to examine the significance of the improvements produced by the addition of data-driven predictors. While statistical techniques, which are not designed to handle such a large number of predictors, report problems with model convergence, the machine learning algorithms, particularly the tree-based ones, seem to use the abundance of data-driven predictors effectively to achieve better classification performance. To compare the performance of the machine learning models using theory-driven predictors with the corresponding models using both theory- and data-driven predictors, we use one-tailed bootstrap tests.[10] According to the bootstrap test, the addition of the data-driven predictors gives significant improvement in predictive performance for all the machine learning models except for multilayer perceptron where out-of-time prediction is concerned. The finding that the use of data-driven predictors improves the performance of machine learning algorithms for out-of-time prediction indicates that data-driven predictors can add value in the process of classification. Predictor importance metrics, obtained from the random forest algorithm, confirm the relevance of data-driven predictors. The following lists the ten most important predictors as per the mean-decrease-in-Gini (or Gini importance) criterion:

1. Cash inflow from operating activities / Long-term provisions and liabilities.
2. Liabilities for contributions on salaries and wages paid by the employee (credit turnover without opening balance) / Salaries, salary compensations, and other benefits to employees.
3. Total cash outflow / Current liabilities.
4. Total cash outflow / Long-term provisions and liabilities.
5. Total cash inflow / Long-term provisions and liabilities.
6. Operating revenue / Current liabilities.
7. Net result.
8. Total cash inflow / Current liabilities.
9. Sales and prepayments / Long-term provisions and liabilities.
10. Cash outflow from operating activities / Long-term provisions and liabilities.

The fact that only one theory-driven predictor is among the top ten most important predictors (i.e. net result, ranking seventh) confirms our belief that the theory-driven predictors commonly used in the literature, such as current ratio, quick ratio, conversion cycles, and debt ratio, are not the most relevant predictors of auditor qualifications and that the data provided in the full set of financial statements can be better used to achieve improved predictive performance.

10 Comparison using the DeLong test (DeLong et al., 1988) of AUCs for nested models developed and validated on the same data is problematic and often leads to overly conservative results (Demler et al., 2012).

On the other hand, for out-of-sample and out-of-time predictions, using data-driven predictors improves performance for only the regularized random forest model, worsens the performance of the extreme gradient boosting model, and does not seem to affect the performance of the remaining models.

To conclude, where individual predictive models are concerned, mixed-effects logistic regression is the most effective technique for making out-of-time predictions, whereas other methods — primarily the tree-based machine learning algorithms such as random forest, regularized random forest, gradient boosting machine, and C5.0 — achieve the most accurate out-of-sample and out-of-time predictions.[11] While the addition of data-driven predictors is shown to be advantageous for some models for out-of-time prediction, for the best performing models listed above the addition of data-driven predictors is either unfeasible (mixed-effects logistic regression) or does not result in a statistically significant improvement of the predictive performance (machine learning algorithms). Therefore, these models should be trained using theory-driven predictors only.

Hybrid Models

In view of the finding that machine learning and statistical techniques have their own advantages with regard to predicting the type of audit opinion, we proceeded to specify two hybrid models aimed at combining the respective strengths of the two approaches within a single predictive model.

Stacked Ensemble

The first hybrid model is a stacked ensemble, aimed at achieving the highest possible predictive performance. According to the standard procedure for building ensemble learners, a meta-learner is fitted using the out-of-bag predicted class probabilities estimated by seven different machine learning algorithms and consequently used to generate predictions in the two test sets. Nevertheless, besides using a regular logistic regression as a meta-learner, we used a mixed-effects logistic regression, which is expected to better optimize the values of the parameters in view of the panel structure of the data.

The modified version of the stacked ensemble is of the specification:

$$x_{ijt} = \alpha + u_i + u_j + \beta_1 \times C.50\ prob_{it} + \beta_2 \times RF\ prob_{it} + \beta_3 \times RRF\ prob_{it} + \beta_4 \times GBM\ prob_{it}$$
$$+\beta_5 \times XGBOOST\ prob_{it} + \beta_6 \times KNN\ prob_{it} + \beta_7 \times MLP\ prob_{it} + e_{ijt}$$

The comparative predictive performance of stacked ensembles using regular logistic regression and mixed-effects logistic regression is presented in Table 7. The complete output of the hybrid models is presented in Appendix D.

11 This claim is supported by the results of the principal component analysis presented in Appendix E (see Table E3).

	With Theory-Driven Predictors				With Theory- and Data-Driven Predictors			
	Out-of-Time (No information rate: 0.6704)		Out-of-Sample and Out-of-Time (No information rate: 0.6413)		Out-of-Time (No information rate: 0.6704)		Out-of-Sample and Out-of-Time (No information rate: 0.6413)	
Model	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**
Stacked Ensemble with Logistic Regression as Meta-Learner	0.4556 (0.4119–0.4993)	0.7821 / **0.8207**	0.3725 (0.3091–0.4359)	0.7390 / **0.7692**	0.5538 (0.5142–0.5934)	0.8144 / **0.8628**	0.3867 (0.3251–0.4485)	0.7371 / **0.7772**
Stacked Ensemble with Mixed-Effects Logistic Regression as Meta-Learner	0.6060 (0.5684–0.6435)	0.8350 / **0.8872**	0.3910 (0.3284–0.4536)	0.7455 / **0.7858**	0.6134 (0.5767–0.6502)	0.8340 / **0.8925**	0.4340 (0.3744–0.4934)	0.7540 / **0.7975**

Table 7. Performance of Stacked Ensembles

The results indicate that the stacked ensemble that uses the mixed-effects logistic regression model as the meta-learner may achieve classification performance better than any individual classifier in both test samples (the difference is statistically tested later in this study). This supports the view that in order to achieve the best predictive performance, strong points of both machine learning and statistical tools need to be fully employed in a single predictive model. For this model, the use of data-driven predictors appears to be advantageous, as it leads to slight improvements in classification performance, significant at $\alpha = 0.1$, for both out-of-time prediction and out-of-sample and out-of-time prediction (one-tailed bootstrap test reported p-values of 0.06126 and 0.08471, respectively).

Combining Feature-Selection Capabilities of Tree-Based Models with Mixed-Effects Logistic Regression

Despite achieving a high predictive performance, it can be argued that the models presented so far lack interpretability. Since some researchers and practitioners interested in predicting the type of audit opinion may be willing to trade some of the accuracy for improved interpretability, our second hybrid modeling strategy was aimed at optimizing the interpretability/accuracy trade-off rather than maximizing predictive accuracy. To achieve a good interpretability/accuracy trade-off, we have combined the feature extraction capabilities of the tree-based classification algorithms with the capability of accounting for the systematic (auditor- and company-specific) effects inherent in mixed-effects logistic regression. Firstly, in consideration of the typological redundancy in the composition of the most important data-driven predictors, as indicated earlier by random forest's predictor importance metric, we have employed the guided regularized random forest (GRRF) algorithm developed by Deng and Runger (2012) to extract a non-redundant set of classification rules. In the GRRF algorithm, the

relative importance of each predictor is initially assessed using the ordinary random forest algorithm. The variable importance metric is subsequently used to guide the feature selection process (hence the term 'guided') in the regularized random forest algorithm (Deng and Runger, 2012). The product of this procedure is a compact and non-redundant set of highly interpretable classification rules, presented in Table 8. R's RRF package (Deng, 2013) has been used to build RF and GRRF classifiers; the inTrees (Deng, 2014) package has been used for the extraction of the classification rules.

Rule No.	Rule Specification	Modified Opinion	Rule Interpretation
1	Total cash inflow / Long-term provisions and liabilities > 1.2687	No	Solvency
2	Liabilities for contributions on salaries and wages paid by the employee (credit turnover without opening balance) / Salaries, salary compensations, and other benefits to employees (cash outflows) ≤ 0.1558	No	Liquidity
3	Operating revenue / Current liabilities > 2.0425	No	Liquidity
4	Total cash inflow / Long-term provisions and liabilities ≤ 0.4124	Yes	Solvency

Table 8. The Data-Driven GRRF Classification Rules

Descriptive statistics for the variables used in the GRRF rules are presented in Appendix B (Table B2).

These four classification rules can be further combined with other predictive techniques without any loss of interpretability. To optimize the weights assigned to the classification rules with regard to the panel structure of the data, we included the rules in the form of the following dummy variables into a mixed-effects logistic regression:

$$\text{GRRF rule1}_{it} = \left\{ \begin{array}{c} 0 \; if \; \dfrac{\text{Total cash in flow}_{it}}{\text{Long-term provision and liabilities}_{it}} > 1.2687 \\ 1, \; otherwise \end{array} \right\}$$

$$\text{GRRF rule2}_{it} = \left\{ \begin{array}{c} 0 \; if \; \dfrac{\text{Liabilities for contributions on salaries and wages paid by the employee (credit turnover without opening balance)}_{it}}{\text{Salaries, salary compensations and other benefits to employees (cash outflows)}_{it}} \leq 0.1558 \\ 0, \; otherwise \end{array} \right\}$$

$$\text{GRRF rule3}_{it} = \left\{ \begin{array}{c} 0 if \; \dfrac{\text{Operating revenue}_{it}}{\text{Current liabilities}_{it}} > 2.0425 \\ 1, \; otherwise \end{array} \right\}$$

$$\text{GRRF rule4}_{it} = \left\{ \begin{array}{c} 1 \; if \; \dfrac{\text{Total cash inflow}_{it}}{\text{Long-term provisions and liabilities}_{it}} \leq 0.4124 \\ 0, \; otherwise \end{array} \right\}$$

This resulted in a model of the following specification:

$$x_{ijt} = \alpha + u_i + u_j + \beta_1 \times GRRF\ rule1_{it} + \beta_2 \times GRRF\ rule2_{it} + \beta_3 \times GRRF\ rule3_{it} + \beta_4 \times GRRF\ rule4_{it} + e_{ijt}$$

Since the dummy variables are coded in such a way that the rule-based classification outcomes indicating an increased risk of a modified opinion are assigned the code 1, and the outcomes indicating a lower risk of a modified opinion are assigned the code 0, all the regression coefficients are positive, and their exponentiated values indicate an increase in the odds of receiving a modified audit opinion associated with the unfavorable outcomes of the corresponding classification rules. The regression coefficients are presented in Table 9.

Model	Parameters	Estimate	Std. error	z value	Pr(>|z|)
Mixed-effects Logistic Regression with Outcomes on the GRRF Classification Rules Included as Predictors	Intercept	-2.8638	0.2437	-11.752	< 2e-16 ***
	Rule 1	1.2357	0.1718	7.191	6.45e-13 ***
	Rule 2	1.1177	0.1350	8.279	< 2e-16 ***
	Rule 3	1.2894	0.1518	8.492	< 2e-16 ***
	Rule 4	1.3942	0.2169	6.428	1.29e-10 ***

Table 9. Summary of M-ELR with Outcomes on the GRRF Rules Included as Dummy-Coded Predictors

The results of this model demonstrate that the client's poor solvency and poor liquidity are the main indicators of an increased risk of receiving a modified opinion.

Poor solvency of the client, as indicated by a low value (less than or equal to 1.27) of the ratio of total cash inflow to long-term provisions and liabilities, increases the odds of receiving a modified opinion 3.44 times. A further drop in the value of this ratio (to less than or equal to 0.41) implies an additional four-fold (4.03) increase in the odds of receiving a modified opinion.

The ratio of yearly credit turnover of the account liabilities for contributions on salaries and wages paid by the employee to total yearly cash outflows for salaries, salary compensations, and other benefits to employees is identified as highly indicative of the client's liquidity. Specifically, when the value of this ratio rises to more than 0.16 — which is roughly the proportion of the said contributions in the total salary costs — this indicates that salaries are being calculated but not regularly paid out to employees, leading to a three-fold increase in the odds of the client receiving a modified opinion. Another sign of the client's impaired liquidity is the ratio of operating revenue to current liabilities dropping below 2.04, which is associated with a 3.63-fold increase in the odds. Predictive performance of this model is presented in Table 10.

	Out-of-Time (No information rate: 0.6704)		Out-of-Sample and Out-of-Time (No information rate: 0.6413)	
Model	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**
Mixed-Effects Logistic Regression with Data-Driven Predictors Selected by GRRF Algorithm	*0.5961 (0.5586–0.6336)*	*0.8275 /* **0.8702**	*0.3092 (0.2442–0.3743)*	*0.7108 /* **0.7126**

Table 10. Performance of Mixed-Effects Logistic Regression with Predictors Defined by GRRF Algorithm

The results indicate that where out-of-time predictions are concerned, a good trade-off can be achieved, since mixed-effects logistic regression, which used only four data-driven predictors defined by the GRRF algorithm, was only slightly less precise (0.8702 vs. 0.8818) than the mixed-effects logistic regression that used 31 theory-driven covariates. Even trading the added precision of the stacked ensemble with mixed-effects logistic regression as a meta-learner (0.8702 vs. 0.8925), for the parsimony and the unmatched interpretability of the second hybrid model, may be justified in some cases. Another noteworthy advantage of the hybrid model is that it converges rapidly without preprocessing (standardizing) the predictors.

In contrast, achieving a good model interpretability while maintaining good performance is much more difficult where out-of-sample and out-of-time predictions are concerned, because in this case models cannot rely on the use of prior information but can achieve good performance only by effectively modeling the complex relationships present in the data. Consequently, the added interpretability of the second hybrid model does not compensate for the drop in predictive performance in the out-of-sample and out-of-time test set.

Summary of the Relative Performance of the Predictive Models

To formally test the differences in the observed performance of the predictive models presented in this study, we perform pairwise comparisons of the AUROC values using the method described by DeLong et al. (1988) for all models developed using the same predictors and tested on the same test sets (results shown in Appendix C). As a brief summary of the relative performance of the predictive models, we list the best few performing models (according to the value of the AUROC metric, based on the DeLong test for differences in AUROC) in Table 11.

	With Theory-Driven Predictors		With Theory- and Data-Driven Predictors	
	Out-of-Time	**Out-of-Sample and Out-of-Time**	**Out-of-Time**	**Out-of-Sample and Out-of-Time**
Models for which the predictive performance is not statistically significantly lower (α = 0.05) than that of any other model with the same predictors and for the same test sample	1. Stacked Ensemble with Mixed-Effects Logistic Regression as Meta-Learner 2. Mixed-Effects Logistic Regression	1. Stacked Ensemble with Mixed-Effects Logistic Regression as Meta-Learner 2. Random Forest 3. Gradient Boosting Machine 4. C5.0 5. Extreme Gradient Boosting 6. Regularized Random Forest 7. Probit Regression 8. Logistic Regression 9. Linear Discriminant Analysis	1. Stacked Ensemble with Mixed-Effects Logistic Regression as Meta-Learner	1. Stacked Ensemble with Mixed-Effects Logistic Regression as Meta-Learner 2. Regularized Random Forest 3. Random Forest 4. Gradient Boosting Machine

Table 11. The Best Performing Models According to the AUROC Statistic by Type of Predictors Used and by Test Set

The results clearly indicate that the stacked ensemble with mixed-effects logistic regression as meta-learner is the predictive model with the best overall performance in both test sets, regardless of the predictors used. As stated earlier, for this model, the use of data-driven predictors leads to slight improvements in classification performance, significant at α = 0.1, for both out-of-time prediction and out-of-sample and out-of-time prediction.

Alternative Modeling Strategies Considered in This Study

Alternative modeling options that have been considered in our study are:

a) Modeling the systematic effects of clients and auditors as fixed rather than random, in a logistic regression model;[12]

b) Inclusion of dummy variables for identification of auditor firms, aimed at helping methods other than mixed-effects logistic regression to model the related systematic effects;

c) Use of principal components extracted from the complete set of theory- and data-driven covariates as predictors in the models, and

d) Informing methods other than mixed-effects logistic regression on prior audit opinions for the client using lagged variables (applicable only for OOT prediction).

The results of the three above-listed modeling options are shown in Appendix E. The first and the third options did not result in an improved performance as measured by the AUROC metric. The second option improved the predictive performance of certain machine learning methods (namely,

12 A model with fixed effects for both companies and auditors was fitted and used for OOT prediction, whereas a model with fixed effect only for auditors was fitted and used for OOS and OOT prediction). To make the prediction viable, we had to remove the auditors that were not present in the training sample from the test samples, which resulted in a reduced sample size (a reduction from 2,139 to 1,951 observations in OOT and from 1,065 to 976 observations in OOS and OOT).

C5.0, random forest and regularized random forest) for OOT prediction, but not to a significant degree; anyhow, this also seems to lead to severe overfitting problems in other methods (gradient boosting machine, K-nearest neighbors, and multilayer perceptron); moreover, it renders prediction unfeasible for the remaining methods (extreme gradient boosting, support vector machine, and linear discriminant analysis) when the auditor identification dummies from the training set that have become part of the classification rules are missing in the test set. For these reasons, none of the first three alternative modeling options seems to be advantageous for achieving an improved classification performance.

The fourth modeling option has been shown to significantly improve the predictive performance of the models other than mixed-effects logistic regression but has not been included in our primary analysis since it is known to introduce bias in models. As the circumstance that the predictive models, other than MELR, has not provided information on prior audit opinions, is an important aspect of our study that needs to be carefully considered when interpreting the results presented in our primary analysis, we will briefly explain why this was the case, how this circumstance affected the comparative performance of the predictive models presented so far in the study, and how this methodological limitation can be overcome in future studies.

First, it should be noted that the first 11 algorithms listed in Table 5 were trained using theory- and data-driven variables, as described and listed in sections 3.2.1. and 3.2.2; on the other hand, however, the MELR models (as implied by the model specification presented in section 3.3) were trained using the same sets of variables along with a set of dummy variables identifying auditors and companies. The main reason for the use of different sets of predictors is the fact that the original data set is unbalanced. Namely, when the data set is unbalanced, MELR incorporates the prior information by modeling the client-specific systematic effects based on the identifier variables. Informing algorithms other than MELR of prior opinions is, nonetheless, not trivial; technically, this can be achieved via the inclusion of dummy identifiers for auditors and companies among the predictors. This method does not result in an improved classification performance since the algorithms other than MELR are either incapable of taking advantage of a large number of predictors that carry information pertaining to a tiny subset of the sample (i.e. tree-based algorithms may either omit the dummy identifiers from the rules and lose valuable information, or include them in the rules in an unregularized manner, resulting in overfitting) in the case of machine learning algorithms, or are incapable of handling such a large number of predictors in general, in the case of statistical tools. One alternative method for informing the algorithms on the prior audit opinions is the inclusion of two lagged variables (i.e. Modified_t-1 and Modified_t-2) among the predictors. Due to the unbalanced structure of the dataset, however, the inclusion of the lagged variables results in numerous missing observations for these variables.[13] Moreover, the observations on the lagged variables are not missing completely at random (MCAR), but are rather missing at random (MAR), meaning that the propensity for an observation to be missing is related to the observed data (e.g. the propensity of the client not having prior opinions available for both preceding years is related to the size of the client, as measured by their revenue and total assets, because the client's size is an important criterion for determining whether the external audit is obligatory for the client). The MAR pattern of missingness has implications for the modeling options. Namely, whilst the MELR analysis is robust and fully functional under the MAR assumption of missingness, providing unbiased results

13 Specifically, without considering the lagged variables, there are 6,950 complete observations (4,743 unmodified and 2,207 modified opinions) in the training set. After adding the first lagged variable (Modified_t-1), 4,077 complete observations (2,896 unmodified, 1,181 modified and 2,873 missing values) are left in the training set. After adding the second lagged variable (Modified_t-2), 1,779 observations (1,311 unmodified, 468 modified and 5,171 missing values) remain in the training set. The size of the OOT test set reduces as well – from 6,248 to 3,124 observations.

(Baayen et al., 2008; Gibbons et al., 2010, p. 1) due to the implicit imputation that internally takes place in these types of models (Ashbeck and Bell, 2016, p. 7), the remaining algorithms considered in our research require the use of specific procedures for dealing with missing data before becoming functional. The methods commonly proposed in the literature are complete-case analysis, the missing-indicator method, and the missing data imputation method. Even though these methods make the algorithms functional, most of them, when applied to MAR data, result in the decreased efficiency of the analysis and/or biased predictive models. Complete-case analysis decreases the sample size considerably, thus reducing the efficiency of the analysis, and has been shown to introduce bias when data is not MCAR (see e.g. Pedersen et al., 2017, p. 159; Schafer and Graham, 2002, pp. 155–162). Furthermore, the resulting model is not applicable for predictions in cases where prior data are incomplete (where either Modified_t-1 or Modified_t-2 are missing). Similarly, the missing-indicator and single imputation methods are both expected to introduce bias in the presence of MAR patterns (see e.g. Pedersen et al., 2017, p. 159; Schafer and Graham, 2002, pp. 155–162).

Table 12 shows the predictive accuracies of the models for OOT prediction when lagged variables are added to the set of predictors and the three aforementioned methods for dealing with missing data are employed.

	With Theory-Driven Predictors						With Theory- and Data-Driven Predictors					
	Out-of-Time (No information rate: 0.6665) Complete case analysis		Out-of-Time (No information rate: 0.6704) Missing-indicator method		Out-of-Time (No information rate: 0.6704) Single KNN imputation		Out-of-Time (No information rate: 0.6665) Complete case analysis		Out-of-Time (No information rate: 0.6704) Missing-indicator method		Out-of-Time (No information rate: 0.6704) Single KNN imputation	
Model	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**
C5.0	0.6734 (0.6370-0.7099)	0.8590 / **0.8894**	0.6681 (0.6337-0.7026)	0.8569 / **0.8848**	0.6604 (0.6256-0.6952)	0.8541 / **0.8908**	0.6836 (0.6476 - 0.7195)	0.8633 / **0.8848**	0.6711 (0.6370 - 0.7051)	0.8565 / **0.8899**	0.6720 (0.6379 - 0.7061)	0.8574 / **0.8942**
Random Forest	0.6502 (0.6124 - 0.6880)	0.8511 / **0.9009**	0.6715 (0.6374-0.7057)	0.8574 / **0.8884**	0.6546 (0.6193-0.6899)	0.8532 / **0.8953**	0.6805 (0.6445 - 0.7166)	0.8617 / **0.8884**	0.5376 (0.4966 - 0.5786)	0.8139 / **0.8782**	0.5814 (0.5427-0.6201)	0.8266 / **0.8914**
Regularized Random Forest	0.6859 (0.6502 - 0.7217)	0.8638 / **0.8941**	0.6650 (0.6306-0.6995)	0.8551 / **0.8869**	0.6684 (0.6341-0.7027)	0.8560 / **0.8916**	0.6715 (0.6351 - 0.7079)	0.8574 / **0.8909**	0.6595 (0.6248 - 0.6942)	0.8527 / **0.8934**	0.6606 (0.6261-0.6951)	0.8518 / **0.8944**
Gradient Boosting Machine	0.6829 (0.6470 - 0.7188)	0.8622 / **0.8958**	0.6660 (0.6315-0.7005)	0.8560 / **0.8923**	0.6655 (0.6311-0.7000)	0.8551 / **0.8977**	0.6873 (0.6516 - 0.7230)	0.8644 / **0.8910**	0.6582 (0.6235 - 0.6928)	0.8513 / **0.8929**	0.6798 (0.6463-0.7134)	0.8593 / **0.8945**
Extreme Gradient Boosting	0.6844 (0.6485 - 0.7203)	0.8638 / **0.8873**	0.6553 (0.6203-0.6903)	0.8518 / **0.8909**	0.6592 (0.6243-0.6941)	0.8537 / **0.8947**	0.6786 (0.6425 - 0.7147)	0.8606 / **0.8872**	0.6520 (0.6169 - 0.6871)	0.8499 / **0.8941**	0.6679 (0.6336-0.7023)	0.8560 / **0.8955**
K-Nearest Neighbors	0.6747 (0.6385 - 0.7110)	0.8585 / **0.8824**	0.6364 (0.6007-0.6721)	0.8429 / **0.8626**	0.6437 (0.6083-0.6791)	0.8457 / **0.8728**	0.4015 (0.3529 - 0.4502)	0.7638 / **0.8038**	0.3385 (0.2908 - 0.3861)	0.7429 / **0.7845**	0.3958 (0.3497-0.4419)	0.7644 / **0.8127**
Multilayer Perceptron	0.6787 (0.6429 - 0.7146)	0.8590 / **0.8909**	0.6600 (0.6254-0.6947)	0.8527 / **0.8844**	0.6297 (0.5935-0.6660)	0.8424 / **0.8912**	0.5832 (0.5432 - 0.6231)	0.8181 / **0.8546**	0.2915 (0.2412 - 0.3418)	0.7373 / **0.7923**	0.5480 (0.5082-0.5878)	0.8121 / **0.8718**

Support Vector Machine	0.6307 (0.5919 - 0.6696)	0.8441 / **0.8707**	0.6081 (0.5707- 0.6454)	0.8350 / **0.8656**	0.6529 (0.6178- 0.6880)	0.8509 / **0.8640**	Models did not converge because of the large number of predictors
Linear Discriminant Analysis	0.6886 (0.6530 - 0.7243)	0.8649 / **0.8941**	0.2945 (0.2441- 0.3449)	0.7396 / **0.8252**	0.6727 (0.6387- 0.7068)	0.8579 / **0.8974**	
Logistic Regression	0.6879 (0.6522 - 0.7236)	0.8649 / **0.8908**	0.3115 (0.2616- 0.3614)	0.7452 / **0.8232**	0.6737 (0.6396- 0.7078)	0.8588 / **0.8974**	
Probit Regression	0.6903 (0.6548 - 0.7259)	0.8660 / **0.8931**	0.3056 (0.2555- 0.3558)	0.7438 / **0.8257**	0.6713 (0.6371- 0.7055)	0.8579 / **0.8977**	
Mixed-Effects Logistic Regression	0.6527 (0.6150 - 0.6904)	0.8521 / **0.8858**	0.6493 (0.6137 - 0.6850)	0.8518 / **0.8931**	0.6209 (0.5838 - 0.6581)	0.8424 / **0.8940**	

Table 12. Comparative Predictive Performance of the Twelve Predictive Models for OOT Prediction when Using Lagged Variables to Inform Models on Prior Audit Opinions Along with the Methods for Dealing with Missing Observations

The results suggest that the classification accuracies for OOT prediction of most of the algorithms presented in this study, when ignoring (or being willing to accept for the sake of improved classification accuracy) the likely adverse consequences of the methods used for dealing with missing data, are, in a statistical sense, comparable.

The multiple imputation method that considers the systematic effects present in the data is, arguably, the only methodologically sound approach to dealing with missing data in datasets such as the one analyzed in our study, and it can and should be employed to improve the predictive performance of all the models presented in this study, including MELR. As this modeling procedure was not viable within the software used in the study, we identify our failing to employ this procedure as a major methodological limitation of the study and suggest that future research seriously consider this option.

CONCLUSIONS

Models for predicting the type of audit opinion have numerous applications in the field of finance. Both statistical and machine learning approaches have been used for this task in prior research, but for numerous reasons pointed out in this study, no definite conclusion can be drawn on their relative predictive performance based on the results reported therein. Moreover, no previous study has considered the option of combining the two approaches. To address these issues, we have conducted this study, making several important methodological contributions to this line of research.

Firstly, we have used a hand-collected data set comprising 13,561 pairs of annual financial statements and the corresponding audit reports, which is, to the best of our knowledge, the largest empirical data set ever used in this stream of research.

Secondly, we have exploited more fully the capability of machine learning algorithms to handle a large number of predictors. This option was largely overlooked in prior research, where predictor variables were either selected solely based on theory, or selected among a limited pool of conceivably relevant financial metrics by employing a statistical procedure (e.g. statistical significance test or backward stepwise elimination). To make better use of machine learning algorithms, we have considered using all numeric items presented in financial statements, and the relations between them, as potentially relevant predictors. We have presented an innovative method for feature generation, which has generated a total of 76,636 predictors. That is by far the largest number of potential predictors considered in a single study. The use of data-driven predictors has been shown to improve predictive performance; in addition to that, it has resulted in gaining insights into new types of financial metrics that are highly relevant for predicting the type of audit opinion. These metrics combine the data from cash flow statements and income statements with the data from the balance sheet, which is innovative from the perspective of classical financial analysis. It should be noted that, since the relevance of these metrics is suggested by the empirical evidence, there is no supporting theory or prior literature available. The GRRF algorithm will identify different ratios as having discriminatory power in different countries and different periods. Based on the authors' knowledge of the local economy, the metrics identified in this study, along with their corresponding thresholds, can be considered to be highly relevant as indicators of solvency and liquidity.

Thirdly, we have fully exploited the capacity of statistical techniques to account for the systematic effects present in the data, which has resulted in significant improvements in predictive performance over the statistical models used in prior studies.

Fourthly, we have compared the predictive performance of the models in two common and equally important real-life settings: when prior information on the client is available and when prior information on the client is not available. This aspect of the research design, largely neglected in prior research, has allowed us to assess the differential predictive performance of the models in those two settings, which has proved to be crucial for understanding the potential for their use. The results have shown that machine learning techniques — primarily the tree-based machine learning algorithms such as RF, RRF, GBM, and C5.0 — are the most appropriate predictive models to use when prior information on the client is not available. On the other hand, for the clients on which prior information is obtainable, a properly specified mixed-effects logistic regression has been shown to yield the predictive performance equal to that yielded by the tree-based machine learning algorithms but with improved interpretability and without introducing bias into the estimated probabilities. Importantly, these two approaches respectively have outperformed by a wide margin the machine learning and statistical classifiers used in prior research.

Finally, to explore the possibility of combining the relative strengths of the two regular approaches, we have specified two innovative hybrid models. The stacked ensemble using mixed-effects logistic regression as a meta-learner has yielded predictive performance better than any individual classifier in both test samples, making it the most effective compound modeling strategy for predicting the type of audit opinion. The second hybrid modeling strategy has integrated the classification rules obtained from the GRRF algorithm into a structured panel model, resulting in a parsimonious, interpretable, and computationally easy to fit model that has achieved an excellent interpretability/accuracy trade-off in case of out-of-time predictions.

The main conceptual takeaway from this study is that for the development of effective models for prediction of the type of auditor opinion, the strong points of both statistical and machine learning approaches should be used to the full extent and combined when possible.

Regarding the practical applicability of this study, the procedure described herein can be thought of as a framework for developing and testing models for predicting auditor opinions globally. The procedure itself is described in enough detail to allow complete reproducibility. As regards the availability of data needed to employ this framework, both audit opinions, as the outcome variable, and financial reports, from which most predictors are extracted, are directly observable, easily obtainable, and have relatively consistent form globally. Therefore, the proposed framework can be applied in any country. Moreover, by taking advantage of the random-effects capability of mixed-effects logistic regression (specifically, by the inclusion of random intercepts and random effects by countries), the framework can appropriately handle empirical data coming from multiple countries. The models are highly modular in the sense that additional predictors can be seamlessly incorporated, based on their availability. Importantly, the models are applicable for predicting the type of audit opinion for both existing and new companies: owing to the random-effect capability, prior information on the client is used when available but is not necessary to get a risk assessment that is significantly better than the naive classification.

All the predictive models developed in this study provide probabilistic classifications, which can be of great value during the stage of audit planning. Specifically, auditing firms can use different cut-offs based on their specific misclassification costs or choose cut-off points based on the preferred expected level of predictive accuracy within specific types of audit opinion (see Appendix F for observed predictive accuracies of the models by specific type of audit opinion at different cut-off points). Related to this, we recommend future research to follow the methodological framework presented in this study and examine more extreme modified opinions (i.e. disclaimer of opinion and adverse opinion)

separately, as Appendix F shows that predictive accuracies vary across different types of modified opinions. Additionally, the predictive output from the Bayesian mixed-effects logistic regression, such as that specified in this study, is particularly rich in information and can be used in creative ways to form a basis for decision-making: the risk assessments can be aggregated at the level of the audit firm, and various probabilistic statements regarding risk exposure can be made to help form the price of the audit, etc.

At this point, it needs to be emphasized that the predictive models presented in this study assess the risk of material misstatements due to events that are repeatedly detected through a regular external audit and, as such, are primarily useful for audit planning and resource allocation decisions. The risk of material misstatements, due to events that are difficult to detect through a regular audit, as examined by Dechow et al. (2011), Perols (2011), and Perols et al. (2017) using proxies other than audit opinions, is relatively more difficult to assess reliably, even with greater resource allocation, and is better considered to be a part of making client portfolio decisions aimed at reducing litigation and reputation risk.

It is also important to note that the economic significance of the incremental performance improvements described in the results section (and, thus, the optimal choice of a specific predictive model) depends heavily on the size of the auditing firm. Large auditors (and particularly members of the Big 4) should be expected to use the best performing model, which is stacked ensemble with mixed-effects logistic regression as meta-learner for both OOT predictions and OOS and OOT predictions. They may also consider the use of data-driven predictors, as this seems to give a slight improvement (statistically significantly at $\alpha = 0.1$) in the predictive performance of the stacked ensemble. The rationale for using this more involved modeling strategy is that, for these auditors, the benefits gained from a more optimal resource allocation (even a one percent improvement) should outweigh the (for them) immaterial marginal financial costs of developing the most sophisticated model. On the other hand, for small local auditing firms with more constrained financial and human resources, a properly specified mixed-effects logistic regression that uses only theory-driven predictors may suffice for both OOT predictions and OOS and OOT predictions.

Once the audit season is completed, the audit regulators can compare the predictions produced by the models to the actual opinions issued at the level of individual audit firms, and effectively direct their supervisory inspections toward those auditors with the highest estimated misclassification errors. Furthermore, the estimated probability of receiving a modified opinion correlates strongly with the probability of there being material misstatements in the corresponding financial reports (as stated earlier, non-pervasive and pervasive material misstatements are the leading reasons for a modified opinion being issued; based on our sample, they are present in around 85 percent of cases in which a modified opinion is issued) and, as such, is highly indicative of the credibility of the financial reports, which is of interest to all major stakeholders. Accordingly, the predicted probability obtained from the models can be used by financial institutions as a valuable input (risk metric) for their decision-making processes—for instance, by banks when making decisions on loan approvals, or by credit rating agencies when assigning credit ratings.

A potential limitation of this study is that the relative performance of the models might vary depending on the choices made in the outlier treatment procedure described in the methodology section. Therefore, we recommend that future research experiments with different outlier treatment procedures and reports their effects on the relative performance of the models. Another potential limitation of our study is data imbalance. Given the relatively high frequency of qualified audit opinions in Serbia (32.6%) compared to elsewhere in Europe, e.g. 16% in Spain, 2% in Germany, 0.5% in Austria, 3.7% in Switzerland, 1.3% in France, and 0.5% in the United Kingdom (Blandón and Bosch, 2013; Gassen and Skaife,

2009), and other parts of the world, e.g. 13% in the US and 11% in China (Abad et al., 2017; Dhaliwal et al., 2014), the results of the study may only be generalized to a smaller number of countries with unusually high frequencies of qualified audit opinions. In light of this, future studies should consider using different techniques for dealing with class imbalance. We also encourage researchers building on the framework proposed herein to include additional non-financial predictors, such as employee turnover, board structure, etc., when available, as these may be expected to marginally improve the predictive performance of the models.

REFERENCES

Abad, D., Sánchez-Ballesta, J. P., & Yagüe, J. (2017). Audit opinions and information asymmetry in the stock market. *Accounting & Finance*, *57*(2), 565–595. https://doi.org/10.1111/acfi.12175

ASB GAAS Section 315. (2013). GAAS section 315, Understanding the Entity and Its Environment and Assessing the Risks of Material Misstatement. Retrieved from https://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00315.pdf

Ashbeck, E. L., & Bell, M. L. (2016). Single time point comparisons in longitudinal randomized controlled trials: power and bias in the presence of missing data. *BMC Medical Research Methodology*, *16*(1), 43. https://doi.org/10.1186/s12874-016-0144-0

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/J.JML.2007.12.005

Bartov, E., Gul, F. A., & Tsui, J. S. L. (2000). Discretionary-accruals models and audit qualifications. *Journal of Accounting and Economics*, *30*(3), 421–452. https://doi.org/10.1016/S0165-4101(01)00015-5

Bell, T. B., & Tabor, R. H. (1991). Empirical Analysis of Audit Uncertainty Qualifications. *Journal of Accounting Research*, *29*(2), 350. https://doi.org/10.2307/2491053

Beneish, M. D. (1999). The Detection of Earnings Manipulation. *Financial Analysts Journal*, *55*(5), 24–36. https://doi.org/10.2469/faj.v55.n5.2296

Bergmeir, C., & Benitez, J. M. (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. *Journal of Statistical Software*, *46*(7), 1–26. Retrieved from http://www.jstatsoft.org/v46/i07/

Blandón, J. G., & Bosch, J. M. A. (2013). Audit firm tenure and qualified opinions: New evidence from Spain. *Revista de Contabilidad*, *16*(2), 118–125. https://doi.org/10.1016/j.rcsar.2013.02.001

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Caramanis, C., & Spathis, C. (2006). Auditee and audit firm characteristics as determinants of audit qualifications. *Managerial Auditing Journal*, *21*(9), 905–920. https://doi.org/10.1108/02686900610705000

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2017). xgboost: Extreme Gradient Boosting. Retrieved from https://cran.r-project.org/package=xgboost

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

DeAngelo, L. E. (1986). Accounting numbers as market valuation substitutes: A study of management buyouts of public stockholders. *Accounting Review*, *61*(3), 400–420. Retrieved from /han/GoogleScholar/www.jstor.org/stable/10.2307/247149

Dechow, P., Ge, W., & Schrand, C. (2010). Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of Accounting and Economics*, *50*(2–3), 344–401. https://doi.org/10.1016/j.jacceco.2010.09.001

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting Material Accounting Misstatements. *Contemporary Accounting Research*, *28*(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting Earnings Management. *The Accounting Review*, *70*(2), 193–225. https://doi.org/10.2307/248303

DeFond, M. L., & Jiambalvo, J. (1994). Debt covenant violation and manipulation of accruals. *Journal of Accounting and Economics*, *17*(1–2), 145–176. https://doi.org/10.1016/0165-4101(94)90008-6

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, *44*(3), 837. https://doi.org/10.2307/2531595

Demler, O. V, Pencina, M. J., & D'Agostino, R. B. (2012). Misuse of DeLong test to compare AUCs for nested models. *Statistics in Medicine*, *31*(23), 2577–2587. https://doi.org/10.1002/sim.5328

Deng, H. (2013). Guided Random Forest in the RRF Package. *ArXiv*, 1–2. Retrieved from http://arxiv.org/abs/1306.0237

Deng, H. (2014). Package ' inTrees .' Retrieved from https://cran.r-project.org/package=inTrees

Deng, H., & Runger, G. (2012). Feature selection via regularized trees. In *2012 Annual International Joint Conference on Neural Networks, IJCNN 2012*. Brisbane, Australia. https://doi.org/10.1109/IJCNN.2012.6252640

Deng, H., & Runger, G. (2013). Gene selection with guided regularized random forest. *Pattern Recognition*, *46*(12), 3483–3489.

Dhaliwal, D. S., Liu, Q., Xie, H., & Zhang, J. (2014). Negative Press Coverage, Litigation Risk, and Audit Opinions in China. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2381696

Dopuch, N., Holthausen, R., & Leftwich, R. (1987). Predicting audit qualifications with financial and market variables. *The Accounting Review*, *62*(3), 431–454.

Doumpos, M., Gaganis, C., & Pasiouras, F. (2005). Explaining qualifications in audit reports using a support vector machine methodology. *Intelligent Systems in Accounting, Finance and Management*, *13*(4), 197–215. https://doi.org/10.1002/isaf.268

Fernández-Gámez, M. A., García-Lagos, F., & Sánchez-Serrano, J. R. (2016). Integrating corporate governance and financial variables for the identification of qualified audit opinions with neural networks. *Neural Computing and Applications*, *27*(5), 1427–1444. https://doi.org/10.1007/s00521-015-1944-6

Francis, J. R., & Krishnan, J. J. (1999). Accounting Accruals and Auditor Reporting Conservatism. *Contemporary Accounting Research*, *16*(1), 135–165. https://doi.org/10.1111/j.1911-3846.1999.tb00577.x

Gaganis, C., & Pasiouras, F. (2006). Auditing models for the detection of qualified audit opinions in the UK public services sector. *International Journal of Accounting, Auditing and Performance Evaluation*, *3*(4), 471. https://doi.org/10.1504/IJAAPE.2006.011207

Gaganis, C., Pasiouras, F., & Doumpos, M. (2007). Probabilistic neural networks for the identification of qualified audit opinions. *Expert Systems with Applications*, *32*(1), 114–124. https://doi.org/10.1016/j.eswa.2005.11.003

Gaganis, C., Pasiouras, F., Spathis, C., & Zopounidis, C. (2007). A comparison of nearest neighbours, discriminant and logit models for auditing decisions. *Intelligent Systems in Accounting, Finance and Management*, *15*(1–2), 23–40. https://doi.org/10.1002/isaf.283

Gassen, J., & Skaife, H. A. (2009). Can Audit Reforms Affect the Information Role of Audits? Evidence from the German Market. *Contemporary Accounting Research*, *26*(3), 867–898. https://doi.org/10.1506/car.26.3.10

Gibbons, R. D., Hedeker, D., & DuToit, S. (2010). Advances in Analysis of Longitudinal Data. *Annual Review of Clinical Psychology*, *6*(1), 79–107. https://doi.org/10.1146/annurev.clinpsy.032408.153550

Glancy, F. H., & Yadav, S. B. (2011). A computational model for financial reporting fraud detection. *Decision Support Systems*, *50*(3), 595–601. https://doi.org/10.1016/j.dss.2010.08.010

Healy, P. M. (1985). The effect of bonus schemes on accounting decisions. *Journal of Accounting and Economics*, *7*(1), 85–107.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, *50*(3), 585–594. https://doi.org/10.1016/j.dss.2010.08.009

IAASB ISA 315. (2013). ISA 315, Identifying and Assessing the Risks of Material Misstatement through Understanding the Entity and Its Environment. Retrieved from https://www.iaasb.org/system/files/meetings/files/20130415-IAASB-Agenda_Item_5-D_Disclosures - ISA 315 %28Revised%29 for reference ONLY.pdf

IAASB ISA 570. (2013). ISA 570, Going Concern. Retrieved from http://www.ifac.org/system/files/downloads/a031-2010-iaasb-handbook-isa-570.pdf

Jones, J. J. (1991). Earnings Management During Import Relief Investigations. *Journal of Accounting Research*, *29*(2), 193–228. https://doi.org/10.2307/2491047

Jones, K. L., Krishnan, G. V., & Melendrez, K. D. (2008). Do Models of Discretionary Accruals Detect Actual Cases of Fraudulent and Restated Earnings? An Empirical Analysis. *Contemporary Accounting Research*, *25*(2), 499–531. https://doi.org/10.1506/car.25.2.8

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, *11*(9), 1–20. Retrieved from http://www.jstatsoft.org/v11/i09/

Kinney, W. R., & McDaniel, L. S. (1989). Characteristics of firms correcting previously reported quarterly earnings. *Journal of Accounting and Economics*, *11*(1), 71–93. https://doi.org/10.1016/0165-4101(89)90014-1

Kirkos, E., Spathis, C., Nanopoulos, A., & Manolopoulos, Y. (2007). Identifying Qualified Auditors' Opinions: A Data Mining Approach. *Journal of Emerging Technologies in Accounting*, *4*, 183–197.

Krishnan, J., & Krishnan, J. (1996). The Role of Economic Trade-Offs in the Audit Opinion Decision: An Empirical Analysis. *Journal of Accounting, Auditing & Finance*, *11*(4), 565–586. https://doi.org/10.1177/0148558X9601100403

Krishnan, J., Krishnan, J., & Stephens, R. G. (1996). The Simultaneous Relation Between Auditor Switching and Audit Opinion: An Empirical Analysis. *Accounting & Business Research (Wolters Kluwer UK)*, *26*(3), 224–236.

Kuhn, M. (2017). caret: Classification and Regression Training. Retrieved from https://cran.r-project.org/package=caret

Kuhn, M., & Ross, Q. (2017). C50: C5.0 Decision Trees and Rule-Based Models. Retrieved from https://cran.r-project.org/package=C50

Laitinen, E. K., & Laitinen, T. (1998). Qualified audit reports in Finland: evidence from large companies. *European Accounting Review*, *7*(4), 639–653. https://doi.org/10.1080/096381898336231

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from http://cran.r-project.org/doc/Rnews/

Maggina, A., & Tsaklanganos, A. A. (2011). Predicting audit opinions evidence from the athens stock exchange. *Journal of Applied Business Research*, *27*(4), 53–68.

Monroe, G. S., & Teh, S. T. (2009). Predicting uncertainty audit qualifications in Australia using publicly available information. *Accounting & Finance*, *33*(2), 79–106. https://doi.org/10.1111/j.1467-629X.1993.tb00200.x

Mutchler, J. F., & Hopwood, W. (1997). The Influence of Contrary Information and Mitigating Factors on Audit Opinion Decisions on Bankrupt Companies. *Journal of Accounting Research*, *35*(2), 295–310. https://doi.org/10.2307/2491367

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, *9*, 157–166. https://doi.org/10.2147/CLEP.S129785

Perols, J. (2011). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *AUDITING: A Journal of Practice & Theory*, *30*(2), 19–50. https://doi.org/10.2308/ajpt-50009

Perols, J. L., Bowen, R. M., Zimmermann, C., & Samba, B. (2017). Finding needles in a haystack: Using data analytics to improve fraud prediction. In *Accounting Review* (Vol. 92, pp. 221–245). https://doi.org/10.2308/accr-51562

Pourheydari, O., Nezamabadi-Pour, H., & Aazami, Z. (2012). Identifying qualified audit opinions by artificial neural networks. *African Journal of Business Management*, *6*(44), 11077–11087. https://doi.org/10.5897/AJBM12.855

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Ridgeway, G. (2017). gbm: Generalized Boosted Regression Models. Retrieved from https://cran.r-project.org/package=gbm

Ruiz-Barbadillo, E., Gómez-Aguilar, N., De Fuentes-Barberá, C., & García-Benau, M. A. (2004). Audit quality and the going-concern decision-making process: Spanish evidence. *European Accounting Review*, *13*(4), 597–620. https://doi.org/10.1080/0963818042000216820

Saif, S. M., Sarikhani, M., & Ebrahimi, F. (2012). Finding rules for audit opinions prediction through data mining methods. *European Online Journal of Natural and Social Sciences*, *1*(2), 28–36.

Saif, S. M., Sarikhani, M., & Ebrahimi, F. (2013). An Expert System with Neural Network and Decision Tree for Predicting Audit Opinions. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *2*(4), 151–158. Retrieved from http://iaesjournal.com/online/index.php/IJAI/article/view/3950

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Spathis, C., Doumpos, M., & Zopounidis, C. (2003). Using client performance measures to identify pre-engagement factors associated with qualified audit reports in Greece. *The International Journal of Accounting*, *38*(3), 267–284. https://doi.org/10.1016/S0020-7063(03)00047-5

Stice, J. D. (1991). Using Financial and Market Information to Identify Pre-Engagement Factors Associated with Lawsuits against Auditors. *The Accounting Review*, *66*(3), 516–533. https://doi.org/DOI:

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Yasar, A., Yakut, E., & Gutnu, M. M. (2015). Predicting Qualified Audit Opinions Using Financial Ratios : Evidence from the Istanbul Stock Exchange. *International Journal of Business and Social Science*, *6*(8), 57–67.

Yeh, C.-C., Chi, D.-J., & Lin, Y.-R. (2014). Going-concern prediction using hybrid random forests and rough set approach. *Information Sciences*, *254*, 98–110. https://doi.org/10.1016/j.ins.2013.07.011

Zdolšek, D., Jagrič, T., & Odar, M. (2015). Identification of auditor's report qualifications: An empirical analysis for Slovenia. *Economic Research-Ekonomska Istrazivanja* , *28*(1), 994–1005. https://doi.org/10.1080/1331677X.2015.1101960

Zhou, W., & Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision Support Systems*, *50*(3), 570–575. https://doi.org/10.1016/j.dss.2010.08.007

APPENDIX A: OBSERVED FREQUENCY OF SPECIFIC TYPES OF AUDIT OPINION BY INDUSTRY CLASSIFICATION

ISIC Section	Unmodified	Modified	Total	% modified	Division	Unmodified	Modified	Total
A: Agriculture, forestry and fishing	522	364	**886**	41.08%	**01**	496	347	**843**
					02	12	4	**16**
					03	14	13	**27**
B: Mining and quarrying	46	51	**97**	52.58%	**05**	3	8	**11**
					06	4	0	**4**
					07	12	15	**27**
					08	27	22	**49**
					09	0	6	**6**
C: Manufacturing	2634	1473	**4107**	35.87%	**10**	641	325	**966**
					11	73	59	**132**
					12	16	8	**24**
					13	59	39	**98**
					14	104	62	**166**
					15	52	30	**82**
					16	72	44	**116**
					17	85	23	**108**
					18	77	30	**107**
					19	20	11	**31**
					20	153	59	**212**
					21	28	10	**38**
					22	150	64	**214**
					23	146	108	**254**
					24	66	41	**107**
					25	286	177	**463**
					26	117	33	**150**
					27	119	39	**158**
					28	130	83	**213**
					29	58	105	**163**
					30	19	23	**42**
					31	93	56	**149**
					32	37	24	**61**
					33	33	20	**53**

D: Electricity, gas, steam, and air conditioning supply	131	79	**210**	37.62%	**35**	131	79	**210**
E: Water supply; sewerage, waste management, and remediation activities	358	189	**547**	34.55%	**36**	171	106	**277**
					38	187	82	**269**
					39	0	1	**1**
F: Construction	773	410	**1183**	34.66%	**41**	275	136	**411**
					42	249	165	**414**
					43	249	109	**358**
G - Wholesale and retail trade; repair of motor vehicles and motorcycles	2776	805	**3581**	22.48%	**45**	242	67	**309**
					46	2048	519	**2567**
					47	486	219	**705**
H: Transportation and storage	400	187	**587**	31.86%	**49**	271	142	**413**
					50	10	3	**13**
					51	8	4	**12**
					52	97	35	**132**
					53	14	3	**17**
I: Accommodation and food service activities	165	157	**322**	48.76%	**55**	118	103	**221**
					56	47	54	**101**
J: Information and communication	257	109	**366**	29.78%	**58**	76	48	**124**
					59	10	7	**17**
					60	13	18	**31**
					61	65	19	**84**
					62	74	13	**87**
					63	19	4	**23**
K: Financial and insurance activities	53	53	**106**	50.00%	**64**	47	52	**99**
					65	3	1	**4**
					66	3	0	**3**
L: Real estate activities	120	78	**198**	39.39%	**68**	120	78	**198**
M: Professional, scientific, and technical activities	424	183	**607**	30.15%	**69**	36	6	**42**
					70	119	51	**170**
					71	138	55	**193**
					72	44	25	**69**
					73	74	17	**91**
					74	7	1	**8**
					75	6	28	**34**

N: Administrative and support service activities	163	51	**214**	23.83%	**77**	27	3	**30**
					78	10	0	**10**
					79	5	14	**19**
					80	50	9	**59**
					81	42	19	**61**
					82	29	6	**35**
O: Public administration and defense; compulsory social security	0	4	**4**	100.00%	**84**	0	4	**4**
P: Education	8	15	**23**	65.22%	**85**	8	15	**23**
Q: Human health and social work activities	14	12	**26**	46.15%	**88**	14	12	**26**
R: Arts, entertainment, and recreation	54	26	**80**	32.50%	**90**	1	0	**1**
					91	12	1	**13**
					92	24	12	**36**
					93	17	13	**30**
S: Other service activities	27	27	**54**	50.00%	**94**	1	2	**3**
					95	12	16	**28**
					96	14	9	**23**
N/A	215	148	**363**	40.77%	**NA**	215	148	**363**
Total	9140	4421	13561	32.60%		9140	4421	13561

Table A1. Types of Audit Opinion by ISIC (International Standard Industrial Classification of All Economic Activities) Sections and Divisions

APPENDIX B: DESCRIPTIVE STATISTICS FOR THEORY-DRIVEN AND THE MOST IMPORTANT DATA-DRIVEN PREDICTORS

	2010 (n=3378)			2011 (n=3301)			2012 (n=3678)			2013 (n=3204)		
Predictor	**Pctl (25)**	**Median**	**Pctl (75)**	**Pctl (25)**	**Median**	**Pctl (75)**	**Pctl (25)**	**Median**	**Pctl (75)**	**Pctl (25)**	**Median**	**Pctl (75)**
LN total assets (in 000 EUR)	7.54	8.25	9.12	7.68	8.35	9.22	7.56	8.26	9.14	7.67	8.38	9.25
LN total revenue (in 000 EUR)	7.45	8.22	8.98	7.67	8.39	9.18	7.63	8.33	9.12	7.70	8.37	9.15
Net result (in 000 EUR)	0.00	37.91	285.79	0.95	78.47	421.44	-0.34	63.85	386.23	-0.50	73.30	370.80
EBIT (in 000 EUR)	1.30	94.12	412.23	5.75	148.93	556.23	1.98	127.98	493.30	3.38	136.42	489.80
EBITDA (in 000 EUR)	34.82	210.45	596.39	47.91	281.05	761.20	37.78	225.60	692.17	38.76	249.12	709.55
Operating result (in 000 EUR)	-11.28	106.85	469.53	-22.70	152.30	563.81	-14.14	167.20	582.75	-27.45	141.44	512.74
Return on assets	0.00	0.03	0.08	0.00	0.04	0.10	0.00	0.04	0.10	0.00	0.04	0.09
Return on equity	0.00	0.04	0.19	0.00	0.06	0.21	0.00	0.06	0.22	0.00	0.10	0.20
Return on invested capital	0.00	0.04	0.12	0.00	0.05	0.14	0.00	0.05	0.14	0.00	0.05	0.12
Return on capital invested in core business	-0.01	0.04	0.12	-0.01	0.04	0.12	-0.01	0.05	0.13	-0.01	0.04	0.11
Operating margin	-0.01	0.03	0.09	-0.02	0.03	0.09	-0.01	0.04	0.10	-0.02	0.03	0.08
Net margin	0.00	0.01	0.05	0.00	0.02	0.07	0.00	0.02	0.06	0.00	0.02	0.10
Net working capital (in 000 EUR)	-283.40	213.73	1068.37	-326.72	237.06	1288.93	-284.54	283.85	1303.24	-343.80	261.51	1418.04
Net working capital to assets	-0.10	0.07	0.25	-0.10	0.07	0.27	-0.10	0.09	0.31	-0.11	0.09	0.30
Quick ratio	0.30	0.59	1.03	0.30	0.62	1.06	0.30	0.62	1.10	0.30	0.63	1.14
Equity to total liabilities	0.20	0.68	1.75	0.22	0.67	1.82	0.22	0.69	1.85	0.26	0.77	2.19
Debt ratio	0.04	0.19	0.40	0.04	0.19	0.39	0.04	0.20	0.39	0.03	0.17	0.37
Days sales outstanding (in days)	32.74	60.55	111.08	31.56	58.90	106.91	29.13	57.69	99.32	29.82	58.24	98.89

Days payable outstanding (in days)	41.85	76.03	150.03	39.70	72.87	147.20	34.64	67.82	134.73	35.77	68.12	137.40
Days sales of inventory (in days)	11.24	31.28	73.88	9.79	31.28	71.66	9.67	31.28	70.38	10.71	31.28	72.15
Cash conversion cycle (in days)	-12.68	26.36	78.56	-13.54	26.36	77.10	-10.61	26.36	77.38	-12.04	26.36	82.55
Z score for private companies	0.96	1.91	3.18	1.03	2.11	3.39	1.02	2.19	3.54	1.06	2.22	3.60
Z score for non-manufacturers and emerging markets	0.15	2.21	5.02	0.18	2.38	5.35	0.26	2.55	5.70	0.16	2.63	6.09
Zmijewski score	-2.54	-1.11	0.34	-2.60	-1.15	0.28	-2.69	-1.13	0.29	-2.80	-1.28	0.18
Shumway score	0.81	2.02	3.19	0.76	1.99	3.12	0.67	1.98	3.14	0.62	1.86	3.05
FCFE (in 000 EUR)	-13.33	3.12	103.83	-19.33	3.32	112.17	-11.64	7.18	135.24	-20.15	4.01	126.02
FCFF (in 000 EUR)	-379.74	-7.98	80.71	-432.36	-7.19	83.71	-385.37	-6.05	76.54	-275.84	-1.52	107.29
FCFE minus net result to revenue	-0.05	-0.01	0.03	-0.06	-0.01	0.03	-0.05	-0.01	0.04	-0.05	-0.01	0.04
CFO minus operating result to revenue	-0.07	0.00	0.08	-0.07	0.00	0.09	-0.07	-0.01	0.08	-0.05	0.01	0.10
Percentage foreign	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average net monthly salary (in EUR)	212.59	289.83	417.61	228.44	307.99	457.56	226.81	305.69	450.43	239.22	316.46	468.89

Table B1. Descriptive Statistics for Theory-Driven Predictors for Each Period

	2010 (n=3378)			2011 (n=3301)			2012 (n=3678)			2013 (n=3204)		
Predictor	**Pctl (25)**	**Median**	**Pctl (75)**	**Pctl (25)**	**Median**	**Pctl (75)**	**Pctl (25)**	**Median**	**Pctl (75)**	**Pctl (25)**	**Median**	**Pctl (75)**
Total cash inflow/Long-term provisions and liabilities (used in Rules 1 and 4)	1.11	2.27	4.29	1.15	2.40	4.58	1.21	2.62	4.86	1.21	2.71	5.10
Liabilities for contributions on salaries and wages paid by the employee (credit turnover without opening balance)/Salaries, salary compensations, and other benefits to employees (cash outflows) (used in Rule 2)	0.13	0.14	0.15	0.13	0.14	0.15	0.13	0.14	0.15	0.13	0.15	0.16
Operating revenue/Current liabilities (used in Rule 3)	1.16	2.48	4.50	1.26	2.59	4.83	1.29	2.87	5.28	1.26	2.86	5.22

Table B2. Descriptive Statistics for Data-Driven Predictors for Each Period

APPENDIX C: PAIRWISE COMPARISONS OF AUCS FOR TWO CORRELATED ROCS USING THE METHOD DESCRIBED BY DELONG ET AL. (1988)

	C5.0	RF	RRF	GBM	XG-BOOST	KNN	MLP	SVM	LDA	Logit	Probit	MELR	Ensemble with LR	Ensemble with M-ELR
C5.0		Z = -3.3425, p-value = 0.0008304	Z = -3.3712, p-value = 0.0007484	Z = 0.14935, p-value = 0.8813	Z = 0.31003, p-value = 0.7565	Z = 2.6966, p-value = 0.007005	Z = 0.42464, p-value = 0.6711	Z = 4.6567, p-value = 3.213e-06	Z = 2.7463, p-value = 0.006028	Z = 2.2194, p-value = 0.02646	Z = 2.2387, p-value = 0.02517	Z = -8.4778, p-value < 2.2e-16	Z = -4.4371, p-value = 9.117e-06	Z = -10.099, p-value < 2.2e-16
RF			Z = -0.29492, p-value = 0.7681	Z = 3.8212, p-value = 0.0001328	Z = 3.8516, p-value = 0.0001173	Z = 4.9989, p-value = 5.764e-07	Z = 3.3958, p-value = 0.0006844	Z = 7.6883, p-value = 1.491e-14	Z = 5.0897, p-value = 3.587e-07	Z = 4.5097, p-value = 6.493e-06	Z = 4.5405, p-value = 5.612e-06	Z = -7.0304, p-value = 2.059e-12	Z = -2.5421, p-value = 0.01102	Z = -9.0038, p-value < 2.2e-16
RRF				Z = 3.7505, p-value = 0.0001765	Z = 3.7541, p-value = 0.000174	Z = 5.0421, p-value = 4.606e-07	Z = 3.3882, p-value = 0.0007036	Z = 7.6257, p-value = 2.427e-14	Z = 5.0429, p-value = 4.584e-07	Z = 4.4945, p-value = 6.972e-06	Z = 4.5195, p-value = 6.198e-06	Z = -7.0327, p-value = 2.026e-12	Z = -2.3152, p-value = 0.0206	Z = -9.071, p-value < 2.2e-16
GBM					Z = 0.23321, p-value = 0.8156	Z = 2.5976, p-value = 0.009387	Z = 0.33168, p-value = 0.7401	Z = 4.7937, p-value = 1.637e-06	Z = 2.8115, p-value = 0.004931	Z = 2.2578, p-value = 0.02396	Z = 2.2564, p-value = 0.02404	Z = -8.5642, p-value < 2.2e-16	Z = -4.7892, p-value = 1.674e-06	Z = -10.156, p-value < 2.2e-16
XG-BOOST						Z = 2.524, p-value = 0.0116	Z = 0.15898, p-value = 0.8737	Z = 4.6294, p-value = 3.668e-06	Z = 2.4943, p-value = 0.01262	Z = 1.9431, p-value = 0.05201	Z = 1.9438, p-value = 0.05192	Z = -8.7525, p-value < 2.2e-16	Z = -4.8749, p-value = 1.089e-06	Z = -10.439, p-value < 2.2e-16
KNN							Z = -2.4921, p-value = 0.0127	Z = 1.4145, p-value = 0.1572	Z = -0.81809, p-value = 0.4133	Z = -1.1995, p-value = 0.2303	Z = -1.21, p-value = 0.2263	Z = -11.426, p-value < 2.2e-16	Z = -7.967, p-value = 1.625e-15	Z = -13.229, p-value < 2.2e-16
MLP								Z = 4.5186, p-value = 6.224e-06	Z = 3.0031, p-value = 0.002673	Z = 2.3609, p-value = 0.01823	Z = 2.3609, p-value = 0.01823	Z = -8.8482, p-value < 2.2e-16	Z = -4.4291, p-value = 9.463e-06	Z = -10.092, p-value < 2.2e-16

SVM			Z = -2.3576, p-value = 0.01839	Z = -2.7092, p-value = 0.006744	Z = -2.7246, p-value = 0.006438	Z = -11.379, p-value < 2.2e-16	Z = -8.7168, p-value < 2.2e-16	Z = -13.083, p-value < 2.2e-16
LDA				Z = -1.6235, p-value = 0.1045	Z = -2.0698, p-value = 0.03847	Z = -10.32, p-value < 2.2e-16	Z = -5.9935, p-value = 2.054e-09	Z = -10.987, p-value < 2.2e-16
Logit					Z = -0.0075588, p-value = 0.994	Z = -9.9685, p-value < 2.2e-16	Z = -5.417, p-value = 6.061e-08	Z = -10.605, p-value < 2.2e-16
Probit						Z = -10.013, p-value < 2.2e-16	Z = -5.4682, p-value = 4.546e-08	Z = -10.657, p-value < 2.2e-16
M-ELR							Z = 6.6183, p-value = 3.634e-11	Z = -1.4504, p-value = 0.1469
Ensemble with LR								Z = -8.8592, p-value < 2.2e-16
Ensemble with M-ELR								

Table C1. DeLong's Test for Two Correlated ROC Curves: Performance of the Models with Theory-Driven Predictors in Out-of-Time Test Set

	C5.0	RF	RRF	GBM	XG-BOOST	KNN	MLP	SVM	LDA	Logit	Probit	M-ELR	Ensemble with LR	Ensemble with M-ELR
C5.0		Z = -0.23407, p-value = 0.8149	Z = 0.30975, p-value = 0.7567	Z = -0.04991, p-value = 0.9602	Z = 0.35356, p-value = 0.7237	Z = 5.2109, p-value = 1.879e-07	Z = 1.7658, p-value = 0.07743	Z = 3.1823, p-value = 0.001461	Z = 0.89294, p-value = 0.3719	Z = 0.76484, p-value = 0.4444	Z = 0.38196, p-value = 0.7025	Z = 0.96835, p-value = 0.3329	Z = 0.43162, p-value = 0.666	Z = -1.2132, p-value = 0.225
RF			Z = 1.6042, p-value = 0.1087	Z = 0.22176, p-value = 0.8245	Z = 0.51947, p-value = 0.6034	Z = 5.4498, p-value = 5.042e-08	Z = 1.5754, p-value = 0.1152	Z = 3.7847, p-value = 0.0001539	Z = 0.90421, p-value = 0.3659	Z = 0.78369, p-value = 0.4332	Z = 0.47308, p-value = 0.6362	Z = 1.0218, p-value = 0.3069	Z = 1.1984, p-value = 0.2308	Z = -1.0844, p-value = 0.2782
RRF				Z = -0.38375, p-value = 0.7012	Z = -0.038732, p-value = 0.9691	Z = 5.0532, p-value = 4.345e-07	Z = 1.094, p-value = 0.2739	Z = 3.2781, p-value = 0.001045	Z = 0.48308, p-value = 0.629	Z = 0.36842, p-value = 0.7126	Z = 0.067355, p-value = 0.9463	Z = 0.68285, p-value = 0.4947	Z = 0.22603, p-value = 0.8212	Z = -1.5682, p-value = 0.1168
GBM					Z = 0.52466, p-value = 0.5998	Z = 5.3015, p-value = 1.148e-07	Z = 1.7079, p-value = 0.08766	Z = 3.2272, p-value = 0.00125	Z = 0.90247, p-value = 0.3668	Z = 0.78179, p-value = 0.4343	Z = 0.40181, p-value = 0.6878	Z = 1.0034, p-value = 0.3157	Z = 0.51685, p-value = 0.6053	Z = -1.2589, p-value = 0.2081
XG-BOOST						Z = 5.1073, p-value = 3.268e-07	Z = 1.3662, p-value = 0.1719	Z = 2.9564, p-value = 0.003112	Z = 0.58871, p-value = 0.5561	Z = 0.47284, p-value = 0.6363	Z = 0.11287, p-value = 0.9101	Z = 0.7716, p-value = 0.4404	Z = 0.16931, p-value = 0.8656	Z = -1.5329, p-value = 0.1253
KNN							Z = -4.1339, p-value = 3.566e-05	Z = -2.1151, p-value = 0.03442	Z = -4.3968, p-value = 1.098e-05	Z = -4.4818, p-value = 7.4e-06	Z = -4.7312, p-value = 2.232e-06	Z = -3.8458, p-value = 0.0001201	Z = -6.8889, p-value = 5.622e-12	Z = -6.7025, p-value = 2.049e-11
MLP								Z = 1.9942, p-value = 0.04613	Z = -0.80987, p-value = 0.418	Z = -0.99259, p-value = 0.3209	Z = -1.4511, p-value = 0.1467	Z = -0.17594, p-value = 0.8603	Z = -1.0098, p-value = 0.3126	Z = -2.4073, p-value = 0.01607
SVM									Z = -2.2615, p-value = 0.02373	Z = -2.3381, p-value = 0.01938	Z = -2.569, p-value = 0.0102	Z = -1.7819, p-value = 0.07477	Z = -3.2888, p-value = 0.001006	Z = -4.0461, p-value = 5.207e-05

LDA			Z = -0.40288, p-value = 0.687	Z = -1.8989, p-value = 0.05758	Z = 0.41787, p-value = 0.676	Z = -0.39769, p-value = 0.6909	Z = -1.6947, p-value = 0.09013
Logit				Z = -1.8592, p-value = 0.063	Z = 0.55909, p-value = 0.5761	Z = -0.27995, p-value = 0.7795	Z = -1.5984, p-value = 0.11
Probit					Z = 0.93227, p-value = 0.3512	Z = 0.02592, p-value = 0.9793	Z = -1.2921, p-value = 0.1963
M-ELR						Z = -0.6207, p-value = 0.5348	Z = -2.593, p-value = 0.009514
Ensemble with LR							Z = -1.8364, p-value = 0.0663
Ensemble with M-ELR							

Table C2. DeLong's Test for Two Correlated ROC Curves: Performance of the Models with Theory-Driven Predictors in Out-of-Sample and Out-of-Time Test Set

	C5.0	RF	RRF	GBM	XGBOOST	KNN	MLP	Ensemble with LR	Ensemble with M-ELR	M-ELR with GRRF rules
C5.0		Z = -5.645, p-value = 1.651e-08	Z = -4.1546, p-value = 3.259e-05	Z = 1.988, p-value = 0.04681	Z = -0.61689, p-value = 0.5373	Z = 2.8133, p-value = 0.004903	Z = 3.7931, p-value = 0.0001488	Z = -7.3954, p-value = 1.41e-13	Z = -9.5817, p-value < 2.2e-16	Z = -5.4511, p-value = 5.005e-08
RF			Z = 3.1876, p-value = 0.001435	Z = 8.1076, p-value = 5.164e-16	Z = 4.6142, p-value = 3.947e-06	Z = 6.9548, p-value = 3.53e-12	Z = 8.8324, p-value < 2.2e-16	Z = -3.8005, p-value = 0.0001444	Z = -6.9936, p-value = 2.68e-12	Z = -2.1434, p-value = 0.03208
RRF				Z = 6.4879, p-value = 8.705e-11	Z = 3.2197, p-value = 0.001283	Z = 5.7204, p-value = 1.063e-08	Z = 7.5411, p-value = 4.662e-14	Z = -5.6945, p-value = 1.237e-08	Z = -7.8825, p-value = 3.209e-15	Z = -3.0667, p-value = 0.002164
GBM					Z = -2.5429, p-value = 0.01099	Z = 1.6643, p-value = 0.09605	Z = 3.1048, p-value = 0.001904	Z = -8.3756, p-value < 2.2e-16	Z = -10.398, p-value < 2.2e-16	Z = -6.5032, p-value = 7.864e-11
XGBOOST						Z = 3.2303, p-value = 0.001237	Z = 4.4052, p-value = 1.057e-05	Z = -7.1575, p-value = 8.215e-13	Z = -9.5975, p-value < 2.2e-16	Z = -5.1898, p-value = 2.105e-07
KNN							Z = 0.53676, p-value = 0.5914	Z = -8.9634, p-value < 2.2e-16	Z = -11.597, p-value < 2.2e-16	Z = -8.0493, p-value = 8.325e-16
MLP								Z = -9.5238, p-value < 2.2e-16	Z = -11.308, p-value < 2.2e-16	Z = -8.0783, p-value = 6.567e-16
Ensemble with LR									Z = -5.9764, p-value = 2.281e-09	Z = -0.94275, p-value = 0.3458
Ensemble with M-ELR										Z = 4.7514, p-value = 2.02e-06
M-ELR with GRRF rules										

Table C3. DeLong's Test for Two Correlated ROC Curves: Performance of the Models with Theory- and Data-Driven Predictors in Out-of-Time Test Set

	C5.0	RF	RRF	GBM	XGBOOST	KNN	MLP	Ensemble with LR	Ensemble with M-ELR	M-ELR with GRRF rules
C5.0		Z = -1.8282, p-value = 0.06751	Z = -2.2387, p-value = 0.02517	Z = -1.1285, p-value = 0.2591	Z = 2.6972, p-value = 0.006993	Z = 6.0165, p-value = 1.782e-09	Z = 1.7117, p-value = 0.08696	Z = -0.84645, p-value = 0.3973	Z = -2.6246, p-value = 0.008674	Z = 3.4589, p-value = 0.0005424
RF			Z = -0.86954, p-value = 0.3846	Z = 0.86774, p-value = 0.3855	Z = 3.8082, p-value = 0.00014	Z = 7.3415, p-value = 2.112e-13	Z = 3.3479, p-value = 0.0008143	Z = 1.938, p-value = 0.05262	Z = -1.379, p-value = 0.1679	Z = 4.9311, p-value = 8.175e-07
RRF				Z = 1.4141, p-value = 0.1573	Z = 4.1863, p-value = 2.835e-05	Z = 7.4955, p-value = 6.605e-14	Z = 3.4971, p-value = 0.0004703	Z = 2.3871, p-value = 0.01698	Z = -0.81864, p-value = 0.413	Z = 5.0811, p-value = 3.753e-07
GBM					Z = 4.0254, p-value = 5.687e-05	Z = 6.6888, p-value = 2.249e-11	Z = 2.5698, p-value = 0.01018	Z = 0.29825, p-value = 0.7655	Z = -1.8444, p-value = 0.06512	Z = 4.3705, p-value = 1.24e-05
XG-BOOST						Z = 4.0604, p-value = 4.9e-05	Z = -0.64293, p-value = 0.5203	Z = -3.7499, p-value = 0.0001769	Z = -4.8282, p-value = 1.378e-06	Z = 1.5197, p-value = 0.1286
KNN							Z = -4.8933, p-value = 9.916e-07	Z = -7.6332, p-value = 2.29e-14	Z = -8.3037, p-value < 2.2e-16	Z = -2.3398, p-value = 0.0193
MLP								Z = -2.4675, p-value = 0.01361	Z = -4.0402, p-value = 5.341e-05	Z = 2.2405, p-value = 0.02506
Ensemble with LR									Z = -2.6308, p-value = 0.008518	Z = 4.1161, p-value = 3.853e-05
Ensemble with M-ELR										Z = 7.0902, p-value = 1.34e-12
M-ELR with GRRF rules										

Table C4. DeLong's Test for Two Correlated ROC Curves: Performance of the Models with Theory- and Data-Driven Predictors in Out-of-Sample and Out-of-Time Test Set

APPENDIX D: COMPLETE OUTPUT OF THE HYBRID MODELS

Model	Parameters	**Estimate**	**Std. error**	**z value**	**Pr(>|z|)**
Stacked ensemble with logistic regression as meta learner (theory-driven predictors)	Intercept	-2.78959	0.07024	-39.717	< 2e-16 ***
	C5.0	0.03290	0.31839	0.103	0.91770
	RF	2.01953	0.62237	3.245	0.00117 **
	RRF	2.38909	0.57824	4.132	3.6e-05 ***
	GBM	-0.55581	0.51326	-1.083	0.27885
	XGBOOST	0.49897	0.43499	1.147	0.25135
	KNN	1.59955	0.13773	11.614	< 2e-16 ***
	MLP	-0.26654	0.26603	-1.002	0.31638
Stacked ensemble with logistic regression as meta learner (theory- and data-driven predictors)	Intercept	-3.27834	0.08045	-40.748	< 2e-16 ***
	C5.0	0.88376	0.30336	2.913	0.00358 **
	RF	4.30689	0.56498	7.623	2.48e-14 ***
	RRF	2.04153	0.51107	3.995	6.48e-05 ***
	GBM	-2.39261	0.32336	-7.399	1.37e-13 ***
	XGBOOST	1.57105	0.20341	7.724	1.13e-14 ***
	KNN	0.82217	0.14607	5.629	1.82e-08 ***
	MLP	-0.21596	0.15481	1.395	0.16301
Stacked ensemble with mixed-effects logistic regression as meta learner (theory-driven predictors)	Intercept	-4.2019	0.2391	-17.575	< 2e-16 ***
	C5.0	1.1939	0.5073	2.354	0.01859 *
	RF	0.8296	0.9624	0.862	0.38869
	RRF	2.3983	0.8948	2.680	0.00736 **
	GBM	1.3937	0.8220	1.695	0.08999 .
	XGBOOST	0.9356	0.6751	1.386	0.16574
	KNN	1.3238	0.2241	5.906	3.5e-09 ***
	MLP	0.5433	0.4292	1.266	0.20558
Stacked ensemble with mixed-effects logistic regression as meta learner (theory- and data-driven predictors)	Intercept	-4.099045	0.203777	-20.115	< 2e-16 ***
	C5.0	1.129696	0.406793	2.777	0.005485 **
	RF	3.984554	0.762891	5.223	1.76e-07 ***
	RRF	2.536672	0.689888	3.677	0.000236 ***
	GBM	-0.491317	0.475486	-1.033	0.301466
	XGBOOST	1.152543	0.280092	4.115	3.87e-05 ***
	KNN	0.351922	0.205382	1.714	0.086620 .
	MLP	-0.007802	0.212315	-0.037	0.970686

Mixed-effects logistic regression with GRRF classification rules as predictors	Intercept	-2.8638	0.2437	-11.752	< 2e-16 ***
	Rule 1	1.2357	0.1718	7.191	6.45e-13 ***
	Rule 2	1.1177	0.1350	8.279	< 2e-16 ***
	Rule 3	1.2894	0.1518	8.492	< 2e-16 ***
	Rule 4	1.3942	0.2169	6.428	1.29e-10 ***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table D1. The Complete Output of the Hybrid Models

APPENDIX E: ALTERNATIVE MODELING STRATEGIES

	With Theory-Driven Predictors			
	Out-of-Time (No information rate: 0.6961) Test sample size reduced from 2,139 obs. to 1,951 obs. because of auditors who were not in the training sample		**Out-of-Sample and Out-of-Time (No information rate: 0.6629) Test sample size reduced from 1,065 obs. to 976 obs. because of auditors who were not in the training sample**	
Model	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**
Fixed effects logistic regression	0.5138 (0.4716 -–0.5561)	0.7929 / 0.7586	0.3242 (0.2538–0.3945)	0.7336 / 0.7537

Table E1. Predictive Performance of the Fixed-Effects Logistic Regressions

	With Theory-Driven Predictors and Dummy Variables for Specific Auditor Firms			
	Out-of-Time (No information rate: 0.6704)		**Out-of-Sample and Out-of-Time (No information rate: 0.6413)**	
Model	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**
C5.0	0.4423 (0.3984–0.4863)	0.7756 / 0.8026	0.3716 (0.3096–0.4336)	0.7296 / 0.7517
Random Forest	0.4436 (0.3990–0.4883)	0.7821 / 0.8309	0.3582 (0.2936–0.4227)	0.7371 / 0.7669
Regularized Random Forest	0.4559 (0.4117–0.5002)	0.7863 / 0.8298	0.3388 (0.2738–0.4039)	0.7286 / 0.7702
Gradient Boosting Machine	0.0850 (0.0334 - 0.1367)	0.6396 / 0.5281	0.0159 (-0.0565–0.0882)	0.6000 / 0.4885
Extreme Gradient Boosting	Unable to make predictions due to missing predictors			
K-Nearest Neighbors	-0.0339 (-0.088– 0.0205)	0.6064 / 0.5090	-0.0494 (-0.1245–0.0258)	0.5869 / 0.4744

Multilayer Perceptron	-0.1253 (-0.1772–-0.0734)	0.5423 / 0.5975	-0.1436 (-0.2157–-0.0716)	0.5239 / 0.6371
Support Vector Machine	Unable to make predictions due to missing predictors			
Linear Discriminant Analysis	Unable to make predictions due to missing predictors			
Logistic Regression	0.3380 (0.2899–0.3861)	0.7461 / 0.7566	-0.1795 (0.1111–0.2479)	0.6582 / 0.6136
Probit Regression	0.3382 (0.2900–0.3864)	0.7471 / 0.7541	0.1774 (0.1088–0.2459)	0.6582 / 0.6174

Table E2. Machine Learning Methods with Theory-Driven Predictors and Dummy Variables Identifying Auditor Firms

	With 129 Principal Components Derived from Theory- and Data-Driven Predictors			
	Out-of-Time (No information rate: 0.6704)		**Out-of-Sample and Out-of-Time (No information rate: 0.6413)**	
Model	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**	**Kappa (95%CI)**	**Accuracy at 50% cut-off / Area under the curve**
C5.0	0.3709 (0.3238–0.4180)	0.7574 / 0.7830	0.3220 (0.2556–0.3884)	0.7268 / 0.7512
Random Forest	0.3704 (0.3228–0.4179)	0.7602 / 0.8139	0.3053 (0.2380–0.3728)	0.7230 / 0.7697
Regularized Random Forest	0.4438 (0.3994–0.4881)	0.7798 / 0.8239	0.3538 (0.2896–0.4180)	0.7324 / 0.7607
Gradient Boosting Machine	0.4116 (0.4116–0.4575)	0.7719 / 0.7888	0.3448 (0.2795–0.4100)	0.7333 / 0.7748
Extreme Gradient Boosting	0.3956 (0.3499–0.4413)	0.7606 / 0.7929	0.3335 (0.2691–0.3978)	0.7211 / 0.7433
K-Nearest Neighbors	0.3854 (0.3398–0.4311)	0.7546 / 0.7778	0.2466 (0.1801–0.3131)	0.6836 / 0.6817
Multilayer Perceptron	0.3880 (0.3421–0.4339)	0.7578 / 0.7686	0.3292 (0.2644–0.3940)	0.7211 / 0.7539
Support Vector Machine	0.3222 (0.2733–0.3711)	0.7433 / 0.7655	0.2470 (0.1780–0.3160)	0.6995 / 0.7324
Linear Discriminant Analysis	0.3130 (0.2641–0.3618)	0.7382 / 0.7637	0.2817 (0.2131–0.3503)	0.7164 / 0.7594
Logistic Regression	0.3307 (0.2822–0.3791)	0.7447 / 0.7703	0.2978 (0.2302–0.3655)	0.7202 / 0.7483
Probit Regression	0.3165 (0.2676–0.3654)	0.7405 / 0.7690	0.2855 (0.2173–0.3537)	0.7164 / 0.7507
Mixed-Effects Logistic Regression	0.5839 (0.5456–0.6222)	0.8252 / 0.8766	0.2840 (0.2175–0.3505)	0.7052 / 0.7085

Table E3. Performance of the Models Using 129 Principal Components Driven from Theory- and Data-Driven Predictors as Predictors

APPENDIX F: PREDICTIVE ACCURACY BY SPECIFIC TYPE OF AUDIT OPINION AND BY CUT-OFF POINT

		Cut-off								
Model	**Type of Audit Opinion**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
C5.0	Unqualified	33.33%	54.60%	71.13%	84.80%	93.10%	95.68%	97.84%	99.16%	100.00%
	Qualified	91.38%	82.59%	66.21%	49.83%	37.59%	27.41%	18.10%	8.97%	1.21%
	Disclaimer	100.00%	98.17%	95.41%	90.83%	79.82%	65.14%	45.87%	23.85%	5.50%
	Adverse	87.50%	81.25%	75.00%	62.50%	31.25%	31.25%	6.25%	6.25%	0.00%
RF	Unqualified	20.85%	48.05%	70.29%	86.12%	93.51%	97.56%	98.95%	99.37%	99.72%
	Qualified	96.72%	85.86%	72.59%	55.00%	40.34%	22.59%	11.55%	5.69%	1.21%
	Disclaimer	100.00%	99.08%	97.25%	90.83%	75.23%	58.72%	36.70%	22.02%	5.50%
	Adverse	93.75%	87.50%	75.00%	56.25%	50.00%	37.50%	6.25%	0.00%	0.00%
RRF	Unqualified	21.13%	48.68%	71.13%	85.29%	93.93%	97.42%	98.81%	99.44%	99.79%
	Qualified	96.21%	86.21%	73.45%	54.48%	38.97%	23.45%	12.41%	5.86%	1.21%
	Disclaimer	100.00%	99.08%	94.50%	88.07%	76.15%	60.55%	36.70%	22.94%	6.42%
	Adverse	93.75%	87.50%	75.00%	56.25%	56.25%	37.50%	12.50%	0.00%	0.00%
GBM	Unqualified	11.72%	51.88%	75.45%	86.75%	93.38%	97.14%	98.81%	99.51%	99.93%
	Qualified	97.59%	82.24%	63.62%	48.45%	34.31%	24.14%	12.76%	4.66%	0.17%
	Disclaimer	100.00%	98.17%	94.50%	91.74%	82.57%	66.06%	39.45%	20.18%	0.92%
	Adverse	93.75%	81.25%	75.00%	56.25%	43.75%	37.50%	18.75%	6.25%	0.00%
XG-BOOST	Unqualified	20.01%	53.14%	73.85%	86.19%	93.31%	96.09%	98.74%	99.16%	99.86%
	Qualified	95.86%	81.38%	64.48%	48.62%	34.83%	22.93%	13.79%	7.76%	1.21%
	Disclaimer	100.00%	98.17%	95.41%	90.83%	75.23%	60.55%	46.79%	23.85%	3.67%
	Adverse	87.50%	75.00%	75.00%	62.50%	50.00%	37.50%	18.75%	6.25%	6.25%
KNN	Unqualified	40.73%	40.73%	71.27%	71.34%	87.66%	87.73%	96.93%	96.93%	99.44%
	Qualified	87.07%	87.07%	68.45%	68.45%	45.17%	45.17%	21.03%	21.03%	7.41%
	Disclaimer	96.33%	96.33%	85.32%	85.32%	64.22%	64.22%	42.20%	42.20%	15.60%
	Adverse	81.25%	81.25%	62.50%	62.50%	50.00%	50.00%	25.00%	25.00%	6.25%
MLP	Unqualified	16.53%	57.32%	76.78%	88.42%	93.65%	96.58%	98.33%	99.37%	99.86%
	Qualified	97.07%	79.83%	61.55%	44.48%	30.34%	21.55%	13.62%	5.69%	2.59%
	Disclaimer	100.00%	98.17%	97.25%	88.07%	73.39%	54.13%	37.61%	14.68%	2.75%
	Adverse	87.50%	81.25%	68.75%	43.75%	31.25%	25.00%	18.75%	18.75%	0.00%

SVM	Unqualified	0.00%	6.76%	83.05%	90.45%	96.03%	97.28%	98.19%	99.16%	100.00%
	Qualified	100.00%	98.62%	47.24%	34.66%	24.31%	17.93%	11.38%	6.03%	0.00%
	Disclaimer	100.00%	99.08%	87.16%	69.72%	53.21%	44.04%	33.94%	17.43%	0.92%
	Adverse	100.00%	93.75%	62.50%	56.25%	37.50%	31.25%	12.50%	6.25%	0.00%
LDA	Unqualified	7.88%	40.93%	77.20%	90.93%	95.26%	97.42%	98.68%	99.09%	99.79%
	Qualified	99.14%	84.31%	57.59%	36.21%	23.97%	15.52%	9.83%	5.34%	1.72%
	Disclaimer	100.00%	99.08%	96.33%	74.31%	55.96%	40.37%	24.77%	13.76%	7.34%
	Adverse	100.00%	81.25%	81.25%	50.00%	43.75%	18.75%	12.50%	0.00%	0.00%
Logit	Unqualified	10.60%	40.17%	72.52%	89.33%	95.75%	97.42%	98.47%	99.23%	99.65%
	Qualified	97.76%	86.03%	64.48%	41.55%	24.83%	16.55%	10.17%	5.17%	1.90%
	Disclaimer	100.00%	98.17%	95.41%	86.24%	62.39%	45.87%	28.44%	14.68%	7.34%
	Adverse	100.00%	81.25%	75.00%	62.50%	43.75%	31.25%	6.25%	0.00%	0.00%
Probit	Unqualified	10.32%	38.28%	71.90%	89.40%	95.75%	97.42%	98.54%	99.23%	99.65%
	Qualified	97.41%	87.41%	65.69%	40.00%	23.62%	15.00%	9.31%	4.66%	1.90%
	Disclaimer	100.00%	99.08%	96.33%	83.49%	59.63%	41.28%	23.85%	12.84%	8.26%
	Adverse	100.00%	81.25%	75.00%	56.25%	43.75%	25.00%	6.25%	0.00%	0.00%
M-LER	Unqualified	47.28%	56.47%	67.10%	71.72%	77.02%	77.67%	78.26%	78.52%	78.87%
	Qualified	72.63%	68.36%	62.59%	54.66%	40.95%	38.66%	36.47%	33.58%	29.87%
	Disclaimer	75.69%	71.56%	65.60%	58.94%	49.77%	48.62%	46.33%	42.66%	37.61%
	Adverse	68.75%	59.38%	53.13%	48.44%	35.94%	32.81%	31.25%	31.25%	29.69%
Stacked model with logit	Unqualified	25.31%	60.95%	78.17%	88.21%	92.89%	96.37%	98.26%	99.09%	99.72%
	Qualified	96.03%	81.72%	65.17%	55.00%	42.76%	31.72%	19.83%	11.55%	2.59%
	Disclaimer	100.00%	96.33%	93.58%	87.16%	78.90%	65.14%	51.38%	31.19%	11.93%
	Adverse	93.75%	81.25%	75.00%	62.50%	43.75%	31.25%	25.00%	12.50%	0.00%
Stacked model with M-LER	Unqualified	67.64%	79.57%	85.50%	89.82%	92.75%	94.77%	95.75%	97.77%	99.44%
	Qualified	88.62%	80.69%	73.45%	67.41%	61.38%	53.45%	45.00%	36.72%	21.21%
	Disclaimer	96.33%	92.66%	90.83%	88.07%	85.32%	82.57%	77.06%	67.89%	46.79%
	Adverse	68.75%	56.25%	56.25%	50.00%	43.75%	37.50%	31.25%	31.25%	25.00%
M-LER with GRRF dummies	Unqualified	62.27%	76.29%	83.47%	87.59%	90.38%	93.03%	94.56%	96.72%	98.54%
	Qualified	87.93%	81.55%	75.52%	70.34%	64.14%	57.07%	50.00%	38.79%	24.66%
	Disclaimer	94.50%	92.66%	91.74%	88.07%	86.24%	80.73%	76.15%	66.06%	53.21%
	Adverse	75.00%	62.50%	50.00%	50.00%	50.00%	43.75%	37.50%	31.25%	12.50%

Table F1. Predictive Accuracy by Specific Type of Audit Opinion: Models with Theory-Driven Predictors, OOT Predictions

		Cut-off								
Model	**Type of Audit Opinion**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
C5.0	Unqualified	31.04%	53.00%	70.86%	84.63%	93.41%	96.63%	97.95%	98.98%	99.71%
	Qualified	88.52%	79.67%	64.92%	46.23%	32.79%	21.31%	12.13%	5.25%	0.33%
	Disclaimer	98.53%	98.53%	91.18%	79.41%	70.59%	61.76%	47.06%	26.47%	7.35%
	Adverse	100.00%	100.00%	88.89%	88.89%	88.89%	88.89%	66.67%	33.33%	11.11%
RF	Unqualified	17.13%	43.92%	67.06%	84.04%	93.70%	97.22%	98.54%	99.71%	100.00%
	Qualified	96.07%	85.25%	70.16%	48.20%	30.82%	16.72%	7.54%	3.61%	0.98%
	Disclaimer	100.00%	98.53%	89.71%	79.41%	66.18%	52.94%	29.41%	19.12%	7.35%
	Adverse	100.00%	100.00%	100.00%	88.89%	88.89%	66.67%	44.44%	44.44%	0.00%
RRF	Unqualified	18.59%	43.34%	68.08%	83.75%	91.95%	96.93%	98.39%	99.85%	100.00%
	Qualified	95.41%	85.25%	66.89%	48.85%	30.82%	17.05%	7.21%	4.59%	0.66%
	Disclaimer	100.00%	100.00%	89.71%	82.35%	70.59%	54.41%	29.41%	22.06%	7.35%
	Adverse	100.00%	100.00%	100.00%	88.89%	88.89%	66.67%	44.44%	33.33%	0.00%
GBM	Unqualified	12.45%	51.83%	75.40%	87.41%	94.14%	97.51%	98.54%	98.98%	100.00%
	Qualified	96.72%	79.02%	60.98%	44.26%	29.51%	21.31%	8.85%	2.62%	0.00%
	Disclaimer	100.00%	95.59%	92.65%	80.88%	64.71%	52.94%	45.59%	25.00%	0.00%
	Adverse	100.00%	88.89%	88.89%	88.89%	88.89%	77.78%	44.44%	33.33%	0.00%
XG-BOOST	Unqualified	19.47%	52.56%	74.38%	87.85%	94.14%	96.78%	98.24%	99.27%	99.85%
	Qualified	93.77%	77.70%	61.31%	43.93%	27.87%	16.72%	8.85%	4.59%	2.30%
	Disclaimer	100.00%	98.53%	91.18%	77.94%	72.06%	55.88%	42.65%	26.47%	10.29%
	Adverse	100.00%	88.89%	88.89%	88.89%	88.89%	77.78%	55.56%	44.44%	22.22%
KNN	Unqualified	37.63%	37.92%	67.94%	67.94%	85.80%	85.80%	95.31%	95.31%	98.83%
	Qualified	81.97%	81.97%	56.72%	56.72%	36.07%	36.07%	15.74%	15.74%	2.62%
	Disclaimer	95.59%	95.59%	79.41%	79.41%	58.82%	58.82%	32.35%	32.35%	14.71%
	Adverse	100.00%	100.00%	77.78%	77.78%	44.44%	44.44%	33.33%	33.33%	33.33%
MLP	Unqualified	15.67%	57.98%	76.57%	90.48%	94.73%	97.22%	98.54%	99.27%	99.85%
	Qualified	95.74%	72.46%	55.74%	38.36%	22.30%	15.08%	8.85%	3.93%	0.66%
	Disclaimer	98.53%	92.65%	86.76%	69.12%	55.88%	47.06%	39.71%	14.71%	5.88%
	Adverse	100.00%	100.00%	100.00%	88.89%	77.78%	55.56%	44.44%	22.22%	0.00%
SVM	Unqualified	0.00%	5.12%	81.70%	88.58%	94.73%	96.19%	97.66%	98.83%	100.00%
	Qualified	100.00%	97.05%	43.93%	33.11%	19.34%	13.77%	8.85%	4.92%	0.00%
	Disclaimer	100.00%	98.53%	79.41%	67.65%	50.00%	36.76%	27.94%	19.12%	1.47%
	Adverse	100.00%	100.00%	88.89%	88.89%	55.56%	44.44%	33.33%	22.22%	0.00%

LDA	Unqualified	7.76%	41.73%	79.80%	92.83%	96.34%	98.10%	98.39%	98.98%	99.85%
	Qualified	98.03%	84.59%	57.38%	32.13%	19.34%	11.15%	5.57%	1.97%	0.66%
	Disclaimer	98.53%	95.59%	88.24%	63.24%	50.00%	38.24%	32.35%	20.59%	4.41%
	Adverse	100.00%	100.00%	88.89%	66.67%	66.67%	44.44%	22.22%	0.00%	0.00%
Logit	Unqualified	12.74%	41.14%	74.23%	90.78%	96.19%	97.95%	98.68%	99.12%	99.85%
	Qualified	96.72%	84.59%	60.98%	36.39%	20.98%	11.15%	5.25%	3.28%	0.98%
	Disclaimer	98.53%	94.12%	86.76%	67.65%	55.88%	38.24%	32.35%	20.59%	7.35%
	Adverse	100.00%	100.00%	88.89%	77.78%	66.67%	44.44%	33.33%	0.00%	0.00%
Probit	Unqualified	12.15%	39.24%	74.08%	91.07%	96.34%	97.95%	98.54%	99.12%	99.85%
	Qualified	96.39%	85.25%	62.30%	36.07%	20.00%	10.82%	5.25%	2.95%	0.98%
	Disclaimer	98.53%	95.59%	92.65%	69.12%	54.41%	39.71%	32.35%	16.18%	5.88%
	Adverse	100.00%	100.00%	100.00%	66.67%	66.67%	44.44%	11.11%	0.00%	0.00%
M-LER	Unqualified	29.80%	38.36%	47.29%	58.49%	74.38%	75.15%	75.40%	75.77%	75.77%
	Qualified	73.61%	70.08%	64.18%	55.74%	31.56%	27.79%	26.15%	25.49%	25.08%
	Disclaimer	74.26%	72.79%	69.12%	61.40%	41.18%	36.40%	34.56%	31.25%	27.21%
	Adverse	75.00%	75.00%	72.22%	66.67%	38.89%	30.56%	25.00%	25.00%	25.00%
Stacked model with logit	Unqualified	19.91%	55.34%	75.26%	86.82%	91.80%	95.02%	97.66%	99.27%	99.85%
	Qualified	94.43%	76.72%	60.00%	45.57%	34.75%	25.25%	15.08%	5.25%	0.66%
	Disclaimer	100.00%	95.59%	85.29%	77.94%	69.12%	58.82%	50.00%	30.88%	8.82%
	Adverse	100.00%	100.00%	88.89%	88.89%	77.78%	66.67%	55.56%	44.44%	22.22%
Stacked model with M-LER	Unqualified	52.86%	73.21%	83.02%	89.02%	91.80%	93.41%	95.46%	96.93%	98.98%
	Qualified	80.66%	66.56%	55.08%	46.89%	36.07%	28.85%	23.28%	13.44%	6.23%
	Disclaimer	94.12%	86.76%	79.41%	73.53%	72.06%	70.59%	61.76%	51.47%	38.24%
	Adverse	100.00%	88.89%	88.89%	88.89%	88.89%	77.78%	66.67%	55.56%	22.22%
M-LER with GRRF dummies	Unqualified	44.95%	64.71%	78.18%	84.33%	88.87%	90.63%	93.56%	97.22%	99.27%
	Qualified	75.74%	60.98%	49.18%	40.33%	32.79%	28.52%	22.62%	13.44%	6.89%
	Disclaimer	91.18%	82.35%	79.41%	75.00%	63.24%	63.24%	55.88%	39.71%	26.47%
	Adverse	100.00%	100.00%	88.89%	77.78%	77.78%	66.67%	66.67%	33.33%	22.22%

Table F2. Predictive Accuracy by Specific Type of Audit Opinion: Models with Theory-Driven Predictors, OOS and OOT Predictions

		Cut-off								
Model	**Type of Audit Opinion**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
C5.0	Unqualified	23.15%	45.33%	65.90%	81.38%	90.66%	95.40%	98.12%	99.09%	99.72%
	Qualified	96.90%	89.31%	75.52%	61.72%	47.41%	31.03%	18.45%	8.62%	3.45%
	Disclaimer	99.08%	97.25%	97.25%	94.50%	84.40%	72.48%	55.96%	36.70%	10.09%
	Adverse	87.50%	81.25%	75.00%	50.00%	50.00%	37.50%	18.75%	12.50%	6.25%
RF	Unqualified	15.48%	41.35%	68.69%	85.01%	94.14%	98.12%	98.81%	99.65%	100.00%
	Qualified	98.62%	93.79%	81.38%	64.31%	43.62%	26.03%	14.14%	5.69%	0.86%
	Disclaimer	100.00%	98.17%	96.33%	95.41%	85.32%	72.48%	39.45%	17.43%	2.75%
	Adverse	93.75%	81.25%	81.25%	68.75%	43.75%	37.50%	25.00%	0.00%	0.00%
RRF	Unqualified	11.02%	36.19%	59.55%	79.92%	91.70%	97.00%	98.61%	99.51%	99.86%
	Qualified	99.14%	94.66%	85.52%	70.00%	48.97%	29.66%	17.24%	7.07%	1.72%
	Disclaimer	99.08%	99.08%	96.33%	95.41%	88.07%	70.64%	44.95%	21.10%	8.26%
	Adverse	100.00%	81.25%	81.25%	68.75%	43.75%	31.25%	18.75%	0.00%	0.00%
GBM	Unqualified	13.74%	44.28%	68.34%	82.57%	89.40%	94.70%	96.86%	98.26%	99.72%
	Qualified	97.93%	88.45%	74.14%	58.10%	45.17%	33.28%	23.79%	11.55%	2.41%
	Disclaimer	100.00%	98.17%	96.33%	93.58%	89.91%	82.57%	66.97%	44.95%	12.84%
	Adverse	81.25%	81.25%	81.25%	68.75%	50.00%	43.75%	37.50%	18.75%	6.25%
XG-BOOST	Unqualified	35.70%	57.25%	70.50%	81.03%	88.77%	92.54%	95.19%	97.42%	98.81%
	Qualified	93.79%	83.62%	73.97%	64.14%	53.79%	42.07%	33.28%	21.38%	12.59%
	Disclaimer	98.17%	97.25%	94.50%	89.91%	86.24%	84.40%	76.15%	63.30%	36.70%
	Adverse	81.25%	81.25%	81.25%	75.00%	68.75%	56.25%	43.75%	25.00%	12.50%
KNN	Unqualified	45.05%	45.26%	75.17%	75.17%	91.35%	91.49%	97.91%	97.91%	99.58%
	Qualified	90.00%	90.00%	69.31%	69.14%	39.31%	39.31%	17.07%	17.07%	4.66%
	Disclaimer	90.83%	90.83%	85.32%	85.32%	56.88%	56.88%	33.03%	33.03%	13.76%
	Adverse	87.50%	87.50%	68.75%	68.75%	37.50%	37.50%	31.25%	31.25%	6.25%
MLP	Unqualified	53.00%	79.29%	82.01%	84.45%	95.19%	96.58%	96.79%	97.49%	99.37%
	Qualified	80.69%	55.69%	51.55%	47.93%	26.72%	22.59%	21.03%	16.38%	6.90%
	Disclaimer	98.17%	91.74%	91.74%	90.83%	70.64%	63.30%	61.47%	51.38%	24.77%
	Adverse	87.50%	75.00%	56.25%	50.00%	31.25%	18.75%	18.75%	12.50%	0.00%
Stacked model with logit	Unqualified	32.71%	62.06%	78.52%	86.19%	91.84%	95.12%	97.14%	98.40%	99.51%
	Qualified	96.72%	87.24%	76.21%	65.34%	54.48%	45.00%	32.07%	20.52%	7.93%
	Disclaimer	99.08%	96.33%	95.41%	92.66%	90.83%	81.65%	69.72%	51.38%	23.85%
	Adverse	81.25%	81.25%	75.00%	62.50%	62.50%	43.75%	37.50%	25.00%	0.00%

Stacked model with M-LER	Unqualified	56.97%	75.24%	82.36%	87.59%	90.38%	93.93%	96.03%	98.05%	99.23%
	Qualified	92.76%	86.21%	78.97%	72.41%	65.86%	57.76%	48.79%	37.93%	22.24%
	Disclaimer	97.25%	95.41%	92.66%	91.74%	89.91%	87.16%	84.40%	77.98%	53.21%
	Adverse	81.25%	68.75%	62.50%	56.25%	50.00%	43.75%	31.25%	25.00%	18.75%
M-LER with GRRF dummies	Unqualified	62.27%	76.29%	83.47%	87.59%	90.38%	93.03%	94.56%	96.72%	98.54%
	Qualified	87.93%	81.55%	75.52%	70.34%	64.14%	57.07%	50.00%	38.79%	24.66%
	Disclaimer	94.50%	92.66%	91.74%	88.07%	86.24%	80.73%	76.15%	66.06%	53.21%
	Adverse	75.00%	62.50%	50.00%	50.00%	50.00%	43.75%	37.50%	31.25%	12.50%

Table F3. Predictive Accuracy by Specific Type of Audit Opinion: Models with Theory- and Data Driven Predictors, OOT Predictions

		Cut-off								
Model	**Type of Audit Opinion**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
C5.0	Unqualified	17.42%	43.63%	61.79%	78.48%	89.75%	94.73%	97.22%	98.98%	99.71%
	Qualified	94.75%	84.59%	72.13%	56.72%	41.31%	24.92%	14.43%	7.54%	2.30%
	Disclaimer	100.00%	95.59%	92.65%	85.29%	73.53%	63.24%	51.47%	32.35%	10.29%
	Adverse	100.00%	100.00%	100.00%	77.78%	66.67%	55.56%	55.56%	33.33%	0.00%
RF	Unqualified	10.25%	33.24%	59.59%	79.36%	94.14%	97.66%	98.98%	99.85%	100.00%
	Qualified	97.38%	89.18%	76.39%	57.38%	35.08%	18.69%	7.21%	3.28%	0.33%
	Disclaimer	100.00%	100.00%	95.59%	86.76%	70.59%	52.94%	32.35%	13.24%	2.94%
	Adverse	100.00%	100.00%	100.00%	88.89%	66.67%	55.56%	44.44%	33.33%	0.00%
RRF	Unqualified	8.49%	29.87%	53.59%	72.18%	90.48%	97.22%	98.24%	99.56%	100.00%
	Qualified	98.69%	91.48%	81.31%	64.92%	40.33%	21.97%	12.13%	4.26%	1.31%
	Disclaimer	100.00%	100.00%	97.06%	91.18%	75.00%	52.94%	32.35%	19.12%	5.88%
	Adverse	100.00%	100.00%	100.00%	88.89%	77.78%	66.67%	44.44%	44.44%	11.11%
GBM	Unqualified	14.06%	43.19%	66.47%	79.50%	87.85%	94.44%	96.78%	98.83%	99.27%
	Qualified	97.70%	82.30%	72.13%	58.03%	40.98%	30.82%	20.98%	9.84%	1.97%
	Disclaimer	100.00%	100.00%	94.12%	88.24%	79.41%	72.06%	57.35%	35.29%	11.76%
	Adverse	100.00%	100.00%	88.89%	88.89%	77.78%	77.78%	66.67%	44.44%	11.11%
XG-BOOST	Unqualified	30.60%	52.56%	65.74%	75.99%	82.58%	90.04%	93.70%	97.07%	99.27%
	Qualified	88.20%	73.44%	60.33%	51.80%	44.92%	36.72%	27.87%	19.67%	9.18%
	Disclaimer	98.53%	94.12%	91.18%	88.24%	83.82%	70.59%	58.82%	45.59%	29.41%
	Adverse	100.00%	88.89%	77.78%	77.78%	77.78%	77.78%	66.67%	66.67%	55.56%

KNN	Unqualified	36.16%	36.16%	65.74%	65.74%	86.68%	86.68%	96.34%	96.34%	99.27%
	Qualified	82.62%	82.62%	56.72%	56.39%	29.84%	29.84%	12.79%	12.79%	3.93%
	Disclaimer	94.12%	94.12%	67.65%	67.65%	41.18%	41.18%	20.59%	20.59%	10.29%
	Adverse	88.89%	88.89%	66.67%	66.67%	44.44%	44.44%	22.22%	22.22%	11.11%
MLP	Unqualified	50.37%	77.16%	79.94%	81.84%	93.12%	95.75%	96.78%	97.80%	98.98%
	Qualified	78.03%	55.74%	53.11%	52.79%	25.90%	21.64%	19.34%	14.10%	3.28%
	Disclaimer	97.06%	79.41%	77.94%	76.47%	50.00%	44.12%	44.12%	39.71%	20.59%
	Adverse	100.00%	66.67%	66.67%	66.67%	55.56%	55.56%	55.56%	55.56%	44.44%
Stacked model with logit	Unqualified	23.57%	51.54%	67.79%	79.80%	88.14%	94.29%	97.51%	98.24%	99.71%
	Qualified	94.43%	79.67%	67.87%	53.44%	40.98%	30.82%	21.64%	13.77%	3.61%
	Disclaimer	100.00%	98.53%	92.65%	82.35%	75.00%	64.71%	54.41%	36.76%	14.71%
	Adverse	100.00%	100.00%	88.89%	77.78%	77.78%	77.78%	55.56%	44.44%	33.33%
Stacked model with M-LER	Unqualified	39.82%	62.96%	75.11%	81.70%	87.85%	92.09%	95.90%	97.36%	98.83%
	Qualified	88.20%	77.38%	63.61%	55.08%	46.56%	38.36%	28.20%	20.00%	8.85%
	Disclaimer	100.00%	92.65%	89.71%	83.82%	79.41%	70.59%	66.18%	54.41%	38.24%
	Adverse	100.00%	100.00%	100.00%	77.78%	77.78%	66.67%	55.56%	55.56%	22.22%
M-LER with GRRF dummies	Unqualified	44.95%	64.71%	78.18%	84.33%	88.87%	90.63%	93.56%	97.22%	99.27%
	Qualified	75.74%	60.98%	49.18%	40.33%	32.79%	28.52%	22.62%	13.44%	6.89%
	Disclaimer	91.18%	82.35%	79.41%	75.00%	63.24%	63.24%	55.88%	39.71%	26.47%
	Adverse	100.00%	100.00%	88.89%	77.78%	77.78%	66.67%	66.67%	33.33%	22.22%

Table F4. Predictive Accuracy by Specific Type of Audit Opinion: Models with Theory- and Data Driven Predictors, OOS and OOT Predictions

PREDVIĐANJE VRSTE REVIZORSKOG MIŠLJENJA: STATISTIKA, MAŠINSKO UČENJE ILI KOMBINACIJA NAVEDENIH?

Rezime:

Cilj ovog istraživanja je prevazilaženje metodoloških ograničenja uočenih u prethodnim istraživanjima u oblasti predviđanja vrste revizorskog mišljenja i izvlačenje pouzdanih zaključaka o uporedivim prediktivnim performansama različitih metoda koje se koriste u te svrhe. Prediktivne performanse dvanaest modela iz oblasti statistike i mašinskog učenja su ocenjene u dva različita praktična scenarija: a) kada su prethodne informacije (vrste revizorskih mišljenja) o klijentu dostupne i mogu se koristiti za predikciju i b) kada su prethodne informacije nedostupne (npr. novoosnovana društva). Rezultati pokazuju da, u prvom scenariju, nekoliko metoda iz obe grupe prediktivnih metoda ostvaruju uporedive performanse u vrednosti od 0,89, mereno površinom ispod krive (eng. Area under the curve). U drugom scenariju, međutim, algoritmi mašinskog učenja, posebno oni zasnovani na drveću odlučivanja, kao što je random forest, ostvaruju značajno bolje rezultate od statističkih metoda, i to u vrednosti od 0,79. Razvili smo i ocenili performanse dva hibridna modela, koji za cilj imaju da iskoriste prednosti statističkih metoda (interpretabilnost rezultata) i metoda mašinskog učenja (obrada velikog broja objašnjavajućih varijabli i veća preciznost). Celokupna procedura je prikazana na reproducibilan način, uz korišćenje najvećeg empirijskog skupa podataka korišćenog u dosadašnjim istraživanjima ovog tipa, koji obuhvata 13.561 par godišnjih finansijskih izveštaja i korespondirajućih revizorskih izveštaja. Procedure opisane u ovom članku omogućavaju revizorskim kućama i finansijskim službenicima širom sveta da razviju i testiraju prediktivne modele koji podržavaju procedure revizorskog planiranja i ocenu rizika ispravnosti podataka u finansijskim izveštajima.

Ključne reči:

mišljenje revizora, finansijski izveštaji, generalizovani linearni mešoviti modeli, random forest (drveće odlučivanja), statistički paket GRRF (Guided Regularized Random Forest), skupovi