Sylwestrzak , M. (2022). Using a hybrid model to detect earnings management for Polish public companies . *Journal of International Studies, 15*(3), 158-172. doi:10.14254/2071-8330.2022/15-3/11

Journal of International Studies

Scientific Papers

Using a hybrid model to detect earnings management for Polish public companies

Marek Sylwestrzak
Faculty of Economic Sciences, University of Warsaw, Warsaw, Poland
msylwestrzak@wne.uw.edu.pl
ORCID 0000-0001-8962-8168

Abstract. This paper analyses the role of non-financial variables in the detection of earnings management in Poland. Previous research on earnings management in Poland concentrated on the use of the Beneish and Roxas models. The sample comprises 63 non-financial Polish companies listed on the Warsaw Stock Exchange for the years 2010-2021. The author uses the hybrid model with elements of decision trees and logistic regression as a proxy for earnings management detection. The results indicate that using a hybrid model increases the accuracy more than standard methods such as decision trees and logistic regression do. Accordingly, inclusion of non-financial variables related to the shareholding structure and the audit increases model accuracy and has a significant impact on the construction of the hybrid model. The findings suggest that using only financial variables worsens model accuracy. The author makes a significant contribution to accounting literature by providing new empirical evidence on the importance of non-financial variables in earnings management detection and their impact on model construction.

Keywords: hybrid model, earnings management, Warsaw Stock Exchange, non-financial variables.

JEL Classification: G34, M41, M42

Received: November, 2021
1st Revision: May, 2022
Accepted: September, 2022

DOI: 10.14254/2071-8330.2022/15-3/11

1. INTRODUCTION

Based on research conducted by the Association of Certified Fraud Examiners (ACFE, 2020), the majority of fraud schemes involve asset misappropriation (86%), corruption (43%), and, least commonly, financial statement fraud (10%), although the last one is the most harmful and costliest category of occupational fraud. Financial statement fraud or earnings management is a serious challenge to market participants' confidence in financial information; it is estimated to cost firms a significant amount of money and is viewed as unacceptable, illegitimate, and illegal corporate conduct (Rezaee, 2005). In general, financial statement fraud techniques work by improperly recognising revenue and overstating assets or understating expenses and liabilities (Beasley et al., 2010).

In the Polish legal system, no legal act refers to the definition of financial statement fraud. In such cases, serious objections from auditors or processes initiated by various regulators resulting in the imposition of penalties may be the only clear evidence that the financial statements have been manipulated. The Polish Financial Supervision Authority (UKNF Board) is one of the bodies ensuring proper functioning, stability, security, transparency, and confidence in the financial market and that the interests of market participants are protected. The UKNF Board also imposes financial or legal sanctions in connection with noncompliance with the International Financial Reporting Standards (IFRS) guidelines. In previous research on earnings management in Poland, the authors used the Beneish (1999) or Roxas (2011) models in empirical analyses (Golec, 2019; Comporek, 2020; Hołda, 2020). However previous studies also had some limitations: some excluded the control group selection from the analysis (Comporek, 2020), others included only eight companies in the analysis (Hołda, 2020), while yet others classified companies that received an adverse or disclaimer opinion by the auditors as manipulators (Golec, 2019).

Many authors apply traditional methods to detect earnings management, such as logistic regression. In recent years, many researchers have attempted to use data mining because of its superiority in terms of forecasting after inputting large amounts of data for machine learning. Data mining is an analytical tool used to handle complicated data analysis and can solve the main shortcomings of the traditional statistical analysis methods by the overcoming limitations of data sets and avoiding the high classification error rate (Yao et al., 2019). In my study, I use hybridization of the decision trees model with logistic regression. Using a hybrid model approach provides a higher predictive accuracy than traditional methods (Steinberg & Cardell, 1998; Brezigar-Masten & Masten, 2012; Łapczyński, 2014). In the first step, the decision tree with a 10-fold cross-validation approach was based on the independent variables, and each leaf included the interaction between the ratios. In the second step, logistic regression, a set of statistically significant independent variables from stepwise regression with backward selection and 10-fold cross-validation was complemented by an artificial variable in the category notified for classification from the root node. I use stepwise regression with backward selection because adding too many variables to the logistic regression may cause overfitting of sample data, model instability, or difficulties in applying the model to an external data set. Several studies focused on the identification of significant indicators in fraud detection, while the number of statistically significant variables in the models ranges from 4 to 35. In recent years, several empirical studies have revealed a significant relationship between non-financial indicators and financial statement fraud (Beasley, 1996; Skounsen et al., 2008; Yuan et al., 2008; Brazel et al., 2009; Johnson et al., 2009; Amara et al., 2013; Jan, 2018; Nindito, 2018; Yao et al., 2019; Subair et al., 2020). For instance, Brazel et al. (2009) found that substantial differences between financial statement data and non-financial indicators should serve as a red flag to auditors and a tipping point for assigning forensic experts to the engagement. Skounsen et al. (2008) also discovered that non-financial variables improve the prediction of financial statement fraud models.

My analysis is based on a sample of 63 public companies listed on the Warsaw Stock Exchange (WSE) that were involved, according to the UKNF Board, in alleged instances of earnings management over the period 2010–2021. Each fraudulent company was matched with a control firm based on firm size, financial year, and industry. The classifiers used in the study were logistic regression and a decision tree. Also, I selected R-squared as a measure of the goodness of fit model and accuracy as metrics to evaluate the classification performance of each classifier.

This study contributes to the literature on the detection of financial statement fraud in several ways. First, evidence suggests that a hybrid model improves model accuracy and goodness of fit more than standard models. Therefore, I can determine that combining the elements of the models will give better results than using standard fraud detection methods. Secondly, the results show that to detect financial statement fraud it is necessary to include non-financial indicators. The inclusion of variables representing the company's shares being held by the Management and Supervisory Board, the shareholding ratio of the

largest shareholder, and the use of unqualified audit opinion increase the likelihood of identifying earnings management. Thirdly, when authors build predictive models to detect financial fraud, they should include non-financial variables in the first step, not as additional model parameters. I thus contribute to the literature by providing evidence that more directly explains the impact of non-financial variables under a constructed financial fraud model. Finally, I am the first, to the best of my knowledge, to use methods other than the Beneish or Roxas model to detect financial statement fraud in Polish public companies listed on the WSE.

The rest of this paper is organized as follows. Section 2 presents the literature review. Section 3 describes the methodology. Section 4 presents the empirical results, finally Section 5 concludes.

2. LITERATURE REVIEW

The research on earnings management prediction contributes to understanding factors that can be used to predict fraud. In prior research studies, authors most often used nonlinear regression such as logit and probit models to detect earnings management (Dechow et al., 1996; Beasley, 1996; Beneish, 1999; Spathis et al., 2002; Skounsen et al., 2008; Johnson et al., 2009; Dechow et al., 2011; Amara et al., 2013; Kanapickiene & Grundiene, 2015; Ozcan, 2016; Ozdagoglu et al., 2017; Pazarskis et al., 2017; Nindito, 2018; Mohammadi et al., 2020). In the logit and probit regressions, the coefficients of the explanatory variables do not influence the different values of the indices in the fraudulent and control companies. Logistic regression is most often used if the researchers' goal is only to identify the variables that are important in detecting fraudulent financial statements. Research that uses nonlinear regression more often uses ratio analysis than non-financial variables as the method of determining financial statement fraud.

Table 1 presents studies that use non-linear regression to detect financial fraud. Most of the research concerns the financial markets of the United States and European countries and uses a small number of independent variables in the regression. Moreover, in most analyzed research, the model accuracy reaches between 85 and 93%.

Table 1

Studies using non-linear regression

Author	Country	Number of observations (fraud)	Number of variables (non-financial)	Accuracy
Dechow et al. (1996)	USA	184 (92)	6 (6)	n.d.
Beasley (1996)	USA	150 (75)	12 (9)	n.d.
Beneish (1999)	USA	1 758 (50)	8 (0)	91.8%
Spathis et al. (2002)	Greece	76 (38)	10 (0)	85.6%
Skounsen et al. (2008)	USA	172 (86)	16 (8)	70.9%
Yuan et al. (2008)	China	274 (137)	10 (7)	70.8%
Brazel et al. (2009)	USA	100 (50)	18 (6)	n.d.
Johnson et al. (2009)	USA	90 (45)	5 (4)	n.d.
Dechow et al. (2011)	USA	88 386 (354)	11 (4)	65.9%
Amara et al. (2013)	France	80 (40)	5 (2)	61.3%
Kanapickiene & Grundiene (2015)	Lithuania	165 (40)	5 (0)	92.8%
Ozcan (2016)	Turkey	144 (72)	10 (0)	84.7%
Pazarskis et al. (2017)	Greece	146 (73)	4 (0)	90.9%
Nindito (2018)	Indonesia	28 (14)	10 (5)	90.5%
Mohammadi et al. (2020)	Iran	330 (165)	9 (0)	67.9%

Source: compiled by the authors.

In contrast to nonlinear regressions, the data mining methods enable the analysis of large and complex data sets. Data mining methods allow for analyses of a larger number of independent variables, which

increases the probability of detecting fraudulent financial statements. Moreover, data mining methods are frequently implemented for financial forecasting to identify market trends.

Decision trees, next to logistic regression, are one of the most popular methods of detecting financial statement fraud. Decision trees are a type of data mining tool and can handle continuous data or non-parametric data for classification. The choice of dividing conditions is based on the quantity and attributes of the data as well as the Gini index (Chen et al., 2014). In decision trees, each node represents a test of an attribute and each branch represents an outcome of the test. In this way, the tree attempts to divide observations into mutually exclusive subgroups. The goodness of a split is based on the selection of the attribute that best separates the sample (Kirkos et al., 2007). The attributes are chosen in terms of the goodness of a split and the sample is divided into subsets until all the training data are correctly classified. The biggest advantage of decision trees is the interpretability of the rules generated from the model (Hajek & Henriques, 2017). Decision trees provide a hierarchical decision model and are easy to interpret. However, the decision tree model generated may be complex, which may be due to overfitting and memorizing the training data, which reduces the generalizability of the resulting model (Ata & Seyrek, 2009).

Table 2 describes studies that use decision trees to detect earnings management. Most of the research concerns the financial markets of the United States and Asian countries and uses fewer independent non-financial variables and more financial variables in the analysis than logistic regression analysis. Moreover, in most analyzed research, the model accuracy reaches between 80 and 95%. On the other hand, the number of observations included in the analysis is similar to that of logistic regression research.

Table 2

Studies using decision trees

Author	Country	Number of observations (fraud)	Number of variables (non-financial)	Accuracy
Kotsiantis et al. (2006)	Greece	164 (41)	25 (0)	91.2%
Kirkos et al. (2007)	Greece	76 (38)	10 (0)	75.0%
Ata & Seyrek (2009)	Turkey	100 (50)	15 (0)	67.9%
Pai et al. (2011)	Taiwan	75 (25)	6 (1)	76.0%
Abbasi et al. (2012)	USA	9 000 (815)	12 (0)	66.4%
Gupta & Gill (2012)	USA	114 (29)	35 (0)	94.7%
Chen et al. (2014)	Taiwan	132 (66)	8 (1)	85.7%
Chen (2016)	Taiwan	176 (44)	30 (7)	83.2%
Hajek & Henriques (2017)	USA	622 (311)	32 (3)	85.8%
Ozdagoglu et al. (2017)	Turkey	214 (110)	13 (0)	82.3%
Dong et al. (2018)	USA	128 (64)	12 (0)	69.3%
Jan (2018)	Taiwan	160 (40)	12 (3)	88.0%
Yao et al. (2019)	China	536 (134)	13 (3)	80.6%

Source: compiled by the authors.

3. METHODOLOGY

In the Polish legal system, there are no legal acts that refer to the definition of financial statement fraud. In such a case, the only clear evidence that financial statements have been manipulated may be serious reservations of auditors or proceedings initiated by various regulators resulting in the imposition of penalties. The UKNF Board is one of the bodies ensuring proper functioning, stability, security, transparency, and confidence in the financial market and that the interests of market participants are protected. The UKNF Board also imposes financial or legal sanctions in connection with non-compliance with the IFRS guidelines.

I identified instances of alleged earnings management by companies listed on the WSE that received a monetary fine from the UKNF Board in the context of compliance with International Accounting Standards (IAS) or IFRS principles. My sample includes 63 public companies involved in alleged instances of earnings management during the period 2010–2021. The 63 fraudulent companies are matched with 63 control firms. I use a matched pair of samples whereby each company is matched with a corresponding control firm based on:

- Firm size, where a nonfraudulent firm was considered similar if total assets were within +/– 30% of total assets for the fraudulent firm in the fraud year,
- Financial year, where annual reports for non-fraudulent firms were available for the same time period as the fraudulent firm,
- Industry, where firms were reviewed to identify a non-fraudulent firm within the same three-digit Standard Industrial Classification (SIC) as the fraudulent firm. If no match was found, two-digit codes were used.

The choice of variables used as candidates for participation in the input vector was based on previous research related to earnings management topics. I collected data from the annual reports of the companies. To effectively detect financial statement fraud, researchers use not only financial variables but also non-financial variables that are known to have some predictive ability in detecting financial statement fraud (Johnson et al., 2008; Skousen et al., 2008; Pai et al., 2011; Amara et al., 2013; Chen et al., 2014; Chen, 2016; Jan, 2018; Nindito, 2018; Yao et al., 2019). So, I decided to divide the non-financial variables into three groups related to the company's board of directors, the shareholding structure, and the audit. Table 3 includes definitions of variables used in this paper.

Table 3

Definitions and measurements of variables

Symbol	Definition	Measurement
Dependent Variable		
Fraud	Fraud	Dummy variable equal 1 for fraud firm and 0 for non-fraud company
Financial Variables		
NM	Net margin	Net profit divided by sales
GM	Gross margin	Gross profit divided by sales
OM	Operating margin	Operating profit divided by sales
ROA	Return on assets	Net profit divided by total assets
DR	Debt ratio	Total liabilities divided by total assets
DE	Debt to Equity	Total liabilities divided by total equity
CR	Current ratio	Current assets divided by current liabilities
QR	Quick ratio	Quick assets divided by current liabilities
C_A	Cash to Total assets	Cash divided by total assets
F_A	Fixed to Total assets	Fixed assets divided by total assets
I_A	Inventory to Total assets	Inventory divided by total assets
R_A	Receivables to Total assets	Receivables divided by total assets
WC_A	Working capital to Total assets	Working capital by total assets
IT	Inventory turnover	Inventory divided by sales
RT	Receivables turnover	Receivables divided by sales
TAT	Total assets turnover	Sales divided by total assets
Non-Financial Variables		
CEO	CEO Change	Dummy variable equals 1 if there was the change of CEO and 0 otherwise
Board	Board size	Total number of Management and Supervisory Board

Symbol	Definition	Measurement
Shares_B	Board shares	Dummy variable equals 1 if Management and Supervisory Board hold the company's shares and 0 otherwise
Shares_I	Investors shares	Percentage of shares held by outside investors
Shares_One	Largest shareholder	Shareholding ratio of the largest shareholder
BIG4	Big Four Audit	Dummy variable equals 1 for companies audited by BIG 4 and 0 otherwise
Audit	Audit opinion	Dummy variable equals 1 for unqualified audit opinion and 0 otherwise

Source: Author's own elaboration.

4. EMPIRICAL RESULTS

4.1. Descriptive statistics

The data mining tool used in this paper is R. In Table 4, I report the descriptive statistics of the continuous variables for fraudulent and control firms. All financial variables are winsorised at 5%. Fraudulent companies are less successful than non-fraudulent ones in terms of profitability indicators, high leverage, low liquidity, and asset rotation ratios.

Table 4

Summary statistics by group

Variable	Q1		Mean		Median		Q3	
	Fraud	Control	Fraud	Control	Fraud	Control	Fraud	Control
NM	–0.959	0.009	–0.237	0.032	–3.195	0.042	0.026	0.082
GM	–1.074	0.016	–0.236	0.042	–3.242	0.055	0.036	0.100
OM	–0.300	0.022	–0.107	0.057	–0.436	0.096	0.053	0.114
ROA	–0.266	0.007	–0.103	0.027	–0.177	0.030	0.012	0.054
DR	0.421	0.354	0.658	0.441	0.628	0.444	0.825	0.602
DE	0.361	0.548	1.198	0.790	2.472	1.041	3.032	1.511
CR	0.503	0.997	0.978	1.406	1.593	2.795	1.552	2.144
QR	0.251	0.408	0.544	0.876	1.045	1.724	1.092	1.329
C_A	0.006	0.013	0.015	0.066	0.029	0.077	0.052	0.101
F_A	0.393	0.265	0.592	0.547	0.586	0.539	0.781	0.763
I_A	0.002	0.022	0.070	0.095	0.104	0.161	0.165	0.229
R_A	0.046	0.053	0.149	0.114	0.175	0.151	0.241	0.209
WC_A	–0.236	0.002	–0.009	0.107	–0.031	0.179	0.158	0.299
IT	0.007	0.067	0.107	0.153	0.189	0.174	0.250	0.257
RT	0.125	0.066	0.251	0.159	0.828	0.181	0.427	0.212
TAT	0.207	0.462	0.624	0.752	0.709	0.895	1.102	1.248
Shares_I	0.175	0.229	0.380	0.300	0.423	0.317	0.655	0.384
Shares_One	0.151	0.185	0.286	0.267	0.359	0.365	0.515	0.569
Board	7	7	7	8	8	9	9	11

Source: Authors' results.

In Figure 1, I present a correlation matrix for the continuous variables and show only significant correlations at the 1% level. Overall, most correlation coefficients are either insignificant or have a low significance. However, *RT* is negatively and significantly associated with margin ratios. On the other hand, *CR* and *QR* are highly and positively correlated and have the same relationship as *OM*, *GM*, and *NM*.

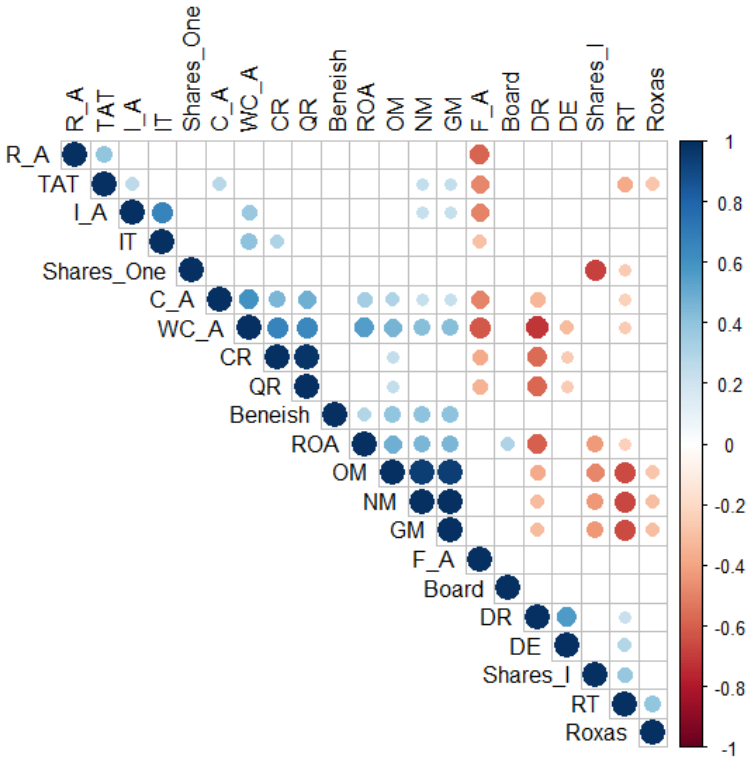

Figure 1. Correlation matrix with statistically significant levels

Blue circles indicate a positive correlation coefficient and red circles indicate a negative correlation that is significantly different from zero at the 1% level.

Source: Authors' results

4.2. Regression results for financial variables

I chose to follow a 10-fold cross-validation approach. Each subset is tested sequentially by adopting the classifier trained on the remaining nine subsets. Cross-validation accuracy is the percentage of data that is correctly classified. I define a Type I error as classifying a fraudulent firm as a non-fraudulent one and a Type II error as classifying a non-fraudulent firm as a fraudulent one. Type I errors may result in unacceptable audits that damage reputation and lead to huge economic losses. Type II errors may lead to additional investigation. I initiate the cost of a type I error as 2 and the cost of a type II error as 1.

Figure 2 shows the statistically significant financial variables with the critical values used in the construction of the decision tree of these rules. The decision tree analysis selected five variables: *OM*, *DR*, *IT*, *RT*, and *WC_A*. It can be observed that the feature of *OM* is the first split point. This means that the relationship between operating profit and sales is critical in predicting financial statement fraud. Twenty-nine percent of the total sample (37 observations) had a lower *OM* value than the critical value equal to –0.059 and were classified as manipulators, with 100% (37 observations) being fraudulent firms. The model correctly classified 92.1% of the total sample, 84.1% of the fraudulent firms, and 100.0% of the non-fraudulent firms.

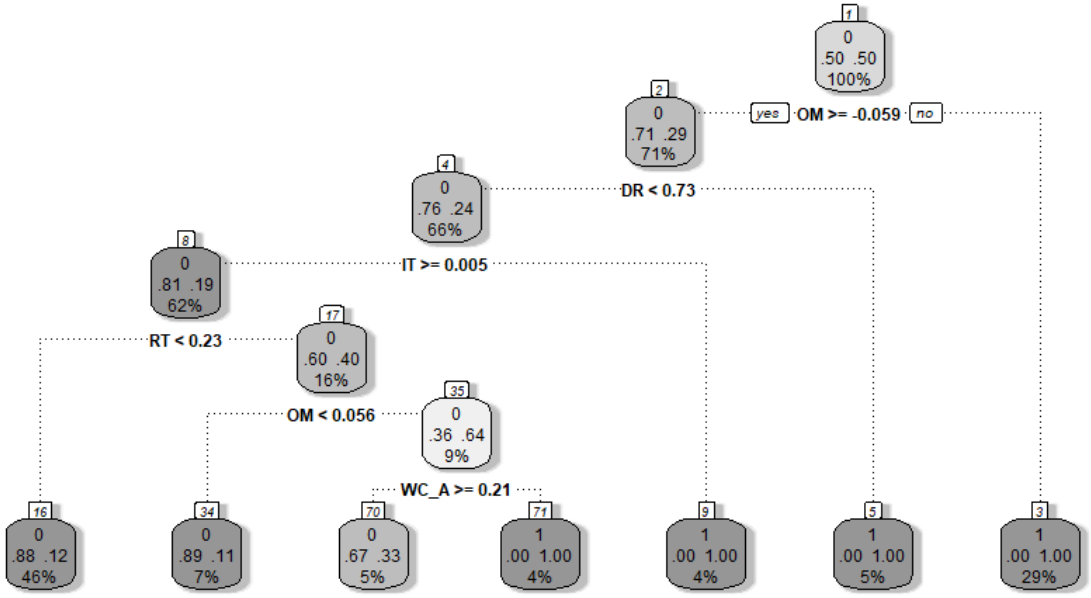

Figure 2. The structure of the decision tree for financial variables

Source: Authors' results

I use stepwise regression with backward selection and 10-fold cross-validation. The stepwise regression shows that the indicators *NM*, *CR*, *C_A*, *F_A*, *I_A*, *IT*, and *TAT* are statistically significant. Table 5 reports the results from logistic regressions with only financial variables. Columns (1) and (3) show the results of variables from the stepwise regression, columns (2) and (4) include variables from the decision tree, column (3) adds *Leaf* as a binary variable equal to 1 if *OM* is smaller than -0.059 and 0 otherwise, which is from the root node, and in column (4) I analyzed only significant variables from decision tree and changed *OM* with *Leaf*.

Table 5

Logistic regression results

Variables	(1)	(2)	(3)	(4)
Constant	2.604 (1.824)	–0.616 (0.752)	2.707 (2.188)	–1.568*** (0.436)
NM	–2.445*** (0.592)		–0.982 (2.086)	
CR	–0.208* (0.124)		–0.182 (0.129)	
C_A	–21.335*** (6.417)		–19.676*** (7.043)	
F_A	–4.273** (1.782)		–4.234** (2.063)	
I_A	–16.489*** (4.570)		–13.230*** (4.734)	
TAT	1.504* (0.777)		1.005 (0.947)	

Variables	(1)	(2)	(3)	(4)
IT	10.172*** (3.302)	1.122 (1.328)	7.013* (3.604)	
OM		-6.462*** (1.548)		
DR		0.082 (1.214)		
WC_A		–2.309* (1.246)		–2.352* (1.275)
RT		2.185** (0.854)		4.364*** (1.669)
Leaf			7.886 (8.562)	11.361 (23.050)
Observations	126	126	126	126
R-squared	39.9%	29.5%	50.4%	45.2%
Accuracy	82.5%	75.4%	82.5%	81.0%
Sensitivity	69.8%	55.6%	68.3%	63.5%
Specificity	95.2%	95.2%	96.8%	98.4%

Standard deviations are in parenthesizes.
Source: Authors' results. * indicates significance level at 0.10 level, ** indicates significance level at 0.05 level, *** indicates significance level at 0.01 level.

The results show that all the variables from the stepwise regression in column (1) are statistically significant. On the contrary, for the decision tree model only the coefficients for *OM*, *WC_A*, and *RT* in column (2) are statistically significant at the 1 or 5% level. However, logistic regression constructed from decision tree indicators has lower accuracy and goodness of fit measured by R-squared than stepwise regression ratios. In addition, adding *Leaf* to the logistic regression (column (3)) improves the model's R-squared and does not change the accuracy of the model, however, the variable was not statistically significant. Also, removing statistically insignificant variables from decision trees and replacing *OM* with *Leaf* (column (4)) affects the model accuracy by increasing the specificity and sensitivity and increasing the R-squared of the model, however, the variable also was not statistically significant. As for the controls, *RT* is positively and significantly related to *Fraud* and the other controls are insignificant; however, this is likely attributable to the use of variables from the decision tree.

4.3. Regression results with non-financial variables

Table 6 reports the results from logistic regressions including non-financial variables. Columns (4-6) show the results of significant variables from stepwise regression with the artificial variable *Leaf*, and columns (7–9) include statistically significant variables from the decision tree replacing *OM* with *Leaf*. The analysis included *Leaf*, even though it was statistically insignificant in regressions with financial variables, but it improved the model's goodness of fit and accuracy.

Table 6

Logistic regression results with non-financial variables

Variables	(4)	(5)	(6)	(7)	(8)	(9)
Constant	2.826 (2.3667)	-0.147 (3.109)	2.788 (2.507)	-1.580 (1.142)	-4.786*** (1.838)	-2.321*** (0.642)
NM	-0.605 (2.188)	-0.662 (2.361)	1.114 (2.307)			
CR	-0.188 (0.132)	-0.126 (0.181)	-0.149 (0.131)			
C_A	-18.439*** (7.020)	-23.842*** (7.922)	-18.326** (7.623)			
F_A	-4.457** (2.177)	-4.741** (2.396)	-4.274* (2.429)			
I_A	-12.900*** (4.866)	-14.596*** (4.839)	-9.189** (4.459)			
TAT	1.064 (0.982)	1.042 (1.020)	0.262 (1.031)			
IT	7.115* (3.723)	7.455** (3.712)	3.822 (3.107)			
WC_A				-1.924 (1.355)	-1.947 (1.273)	-1.681 (1.414)
RT				4.218** (1.715)	5.599*** (1.830)	4.739** (1.899)
Leaf	16.420 (481.362)	15.170 (206.472)	37.798 (66.841)	12.251 (37.997)	7.785** (3.120)	7.532 (5.008)
CEO	0.665 (0.636)			0.640 (0.565)		
Board	-0.043 (0.113)			-0.024 (0.095)		
Shares_B		2.116 (1.487)			2.051* (1.091)	
Shares_I		-0.182 (2.114)			-0.223 (1.969)	
Shares_One		3.874** (1.890)			2.818* (1.557)	
BIG4			-0.748 (0.831)			-0.608 (0.761)
Audit			2.187*** (0.732)			2.633*** (0.684)
Observations	126	126	126	126	126	126
R-squared	51.2%	55.1%	56.3%	46.0%	49.2%	55.3%
Accuracy	82.5%	85.7%	87.3%	81.7%	84.1%	87.3%
Sensitivity	68.3%	73.0%	77.8%	65.1%	69.8%	77.8%
Specificity	96.8%	98.4%	96.8%	98.4%	98.4%	96.8%

Standard deviations are in parenthesizes.

Source: Authors' results. * indicates significance level at 0.10 level, ** indicates significance level at 0.05 level, *** indicates significance level at 0.01 level.

Adding variables related to the company's board of directors (*CEO* and *Board*) to the base models does not change the accuracy of the stepwise regression model but does improve the accuracy of the decision tree model slightly. Also, neither variable is statistically significant, *CEO* is positively correlated and *Board* is negatively correlated with the dependent variable. Including ratios related to the shareholding structure (*Shares_B*, *Shares_I*, and *Shares_One*) in the base models improved the accuracy of the stepwise regression

model and the decision tree model. All shareholder ratios are positively correlated with the dependent variable, and *Shares_B* and *Shares_One* are statistically significant. This means that if the Management and Supervisory Board holds the company's shares or increases the number of shares held by the largest shareholder, the likelihood of earnings management increases. The marginal effect of holding the company's shares by the Management and Supervisory Board and increasing the number of shares held by the largest shareholder by 1 percent will increase the probability of financial fraud by 1.3 to 1.8 percentage points and by 0.5 to 1.3 percentage points, respectively. In addition, the accuracies of the stepwise regression model and the decision tree model have been improved by including ratios related to the audit (*BIG4* and *Audit*) in the base models. *BIG4* is negatively correlated with the dependent variable, *Audit* is positively correlated, and *Audit* is statistically significant. This means that if a company has an unqualified audit opinion, it increases the likelihood of earnings management. The marginal effect of receiving an unqualified audit opinion will increase the probability of financial fraud by 0.01 to 0.4 percentage points.

These results show that including variables related to the shareholding structure (*Shares_B*, *Shares_I*, and *Shares_One*) or ratios related to the audit (*BIG4* and *Audit*) in the hybrid model increases model accuracy compared to the models with only financial ratios.

Furthermore, the regressions performed without *Leaf* showed that the R-squared decreased from 7.0 to 13.2 percentage points, and the accuracy decreased from 2.4 to 4.0 percentage points for stepwise regression. For decision trees, not including *Leaf* reduced the R-squared from 7.9 to 19.2 percentage points and the accuracy from 1.5 to 4.7 percentage points. The regressions without *Leaf* showed two main changes in the analysis results. First, *CEO* was statistically significant and positively correlated with the dependent variable for logistic regression. Second, *Shares_B* turned out to be statistically insignificant.

4.4. Models with non-financial variables

Figure 3 shows the statistically significant financial and non-financial variables with the critical values used in the construction of the decision tree that follows the rules previously described in Section 5.2. The decision tree analysis selected six variables: *OM*, *Shares_I*, *Audit*, *F_A*, *RT*, and *QR*. Ratios *DR*, *WC_A*, and *RT* from the decision tree with only financial variables have been replaced by *Shares_I*, *Audit*, *F_A*, and *QR*. Also, *OM* is the first split point in the decision tree. The model correctly classified 96.0% of the total sample, in particular, 96.8% of the fraudulent firms and 95.2% of the non-fraudulent companies.

The stepwise regression that follows the rules previously described and includes non-financial variables showed that the indicators *NM*, *CR*, *QR*, *C_A*, *IT*, *RT*, *TAT*, *CEO*, *Shares_B*, *Shares_One*, and *Audit* are statistically significant. Table 7 reports the results from logistic regressions for including non-financial variables in the construction of the model. Columns (10) and (12) show the results of variables from the stepwise regression with non-financial variables, columns (11) and (13) include variables from the decision tree with non-financial variables, in column (12), I add *Leaf* as a binary variable equal to 1 if *OM* is lower than -0.059 and 0 otherwise, which is from the root node, and in column (13), I analyzed only significant variables from decision tree and changed of *OM* with *Leaf*.

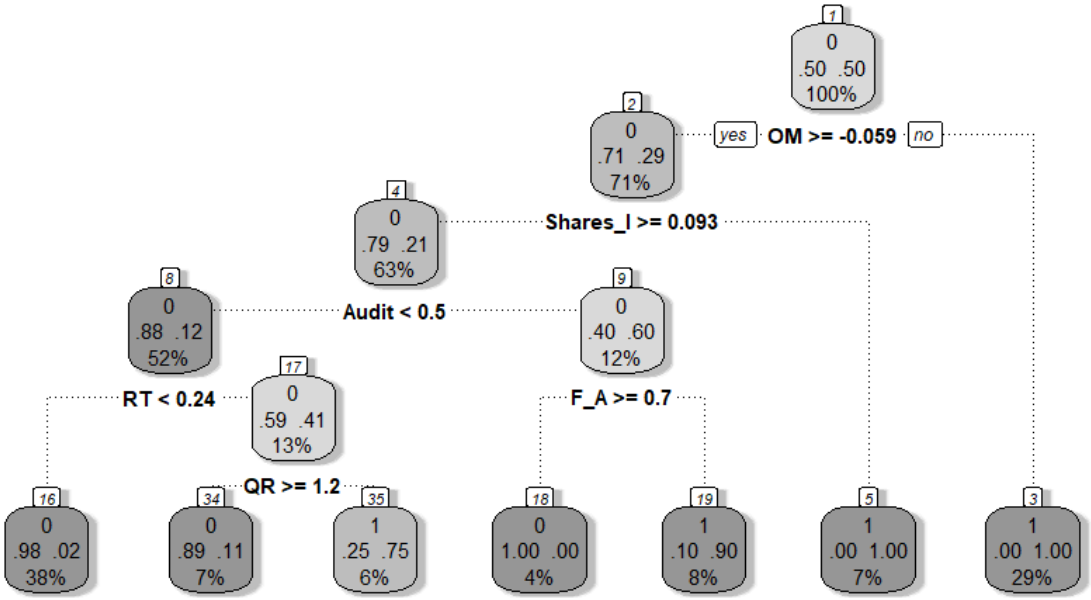

Figure 3. The structure of the decision tree for financial and non-financial variables

Source: Authors' results

Table 7

Logistic regression results with non-financial variables

Variables	(10)	(11)	(12)	(13)
Constant	-7.646*** (1.961)	-1.737* (0.957)	-8.352*** (2.084)	-2.504*** (0.518)
NM	-0.947** (0.450)		0.210 (0.164)	
CR	-1.469** (0.721)		-0.843 (0.888)	
C_A	-20.111** (9.348)		-15.562 (9.920)	
IT	4.992** (2.376)		3.007 (3.093)	
TAT	1.396** (0.695)		1.285* (0.759)	
CEO	1.616*** (0.620)		1.298* (0.718)	
Shares_B	2.100* (1.244)		2.524* (1.400)	
Shares_One	3.697** (1.587)		4.320*** (1.571)	
QR	2.424** (1.137)	-0.033 (0.164)	1.398 (1.391)	-0.178 (0.142)
RT	5.473*** (2.026)	4.839*** (1.590)	6.364*** (2.188)	5.222*** (1.683)
Audit	3.170*** (0.680)	2.712*** (0.579)	2.787*** (0.705)	2.503*** (0.602)

Variables	(10)	(11)	(12)	(13)
OM		-6.832*** (2.223)		
F_A		0.203 (1.176)		
Shares_I		-1.178 (1.331)		
Leaf			6.271** (2.456)	5.536*** (1.906)
Observations	126	126	126	126
R-squared	54.3%	46.9%	62.0%	53.9%
Accuracy	85.7%	84.1%	88.1%	85.7%
Sensitivity	77.8%	79.4%	79.4%	76.2%
Specificity	93.7%	88.9%	96.8%	95.2%

Standard deviations are in parenthesizes.
Source: Authors' results. * indicates significance level at 0.10 level, ** indicates significance level at 0.05 level, *** indicates significance level at 0.01 level.

Including non-financial variables in building the hybrid model increases model accuracy and goodness of fit compared to the base model. Also, adding *Leaf* to the model increases the accuracy and R-squared of the stepwise regression model and the decision tree model. The variable *Leaf* is positively correlated with the dependent variable and statistically significant. This suggests that if a company has an operating margin lower than -0.059, the likelihood of financial statement fraud is increased.

The above results suggest that building a model using only financial variables or adding non-financial variables rather than the first step of model building, decreases model goodness of fit and accuracy.

5. CONCLUSION

Overall, my findings suggest that including an artificial variable related to the decision tree increases the accuracy of the hybrid model. I show that including non-financial variables improves the accuracy of the hybrid models both after adding the variables to the models and during the construction of the models. This empirical evidence means that using only financial variables and logistic regression in the study of financial statement fraud reduces the accuracy of the model and the construction of decision trees. Therefore, my findings may be relevant for other researchers who analyze earnings management. However, the models should include or exclude variables over time to effectively identify the companies likely to report earnings management. Moreover, my study is the first attempt to investigate this type of analysis for Polish public companies and I hope that future research will explore the other variables and models to improve their predictions.

REFERENCES

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud: a meta-learning framework for detecting financial fraud. *MIS QUARTERLY*, *36*(4), 1293-1327. https://doi.org/10.2307/41703508

Amara, I., Amar, A. B., & Jarboui, A. (2013). Detection of fraud in financial statements: French companies as a case study. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, *3*(3), 40-51. https://doi.org/10.6007/IJARAFMS/v3-i3/34

Association of Certified Fraud Examiners (2020). Report to the Nations 2020. Global Study on Occupational Fraud and Abuse. Association of Certified Fraud Examiners. https://acfepublic.s3-us-west-2.amazonaws.com/2020-Report-to-the-Nations.pdf

Ata, H. A., & Seyrek, I. H. (2009). The use of data mining techniques in detecting fraudulent financial statements: an application on manufacturing firms. *Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences*, *14*(2), 157-170.

Beasley, M., Carcello, J., Hermanson, D., & Neal, T. (2010). Fraudulent Financial Reporting 1998–2007: An Analysis of US Public Companies. Committee of Sponsoring Organizations of the Treadway Commission. https://www.coso.org/Shared%20Documents/COSO-Fraud-Study-2010-001.pdf.

Beasley, M. S. (1996). An empirical analysis of the relation between the board of director composition and financial statement fraud. *Accounting Review*, *71*(4), 443-465.

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

Brazel, J. F., Jones, K. I., & Zimbelman, M. F. (2009). Using nonfinancial measures to assess fraud risk. *Journal of Accounting Research*, *47*(5), 1135-1166. https://doi.org/10.1111/j.1475-679X.2009.00349.x

Brezigar-Masten, A., & Masten, I. (2012). CART-based selection of bankruptcy predictors for the logit model. *Expert Systems with Applications*, *39*(11), 10153-10159. https://doi.org/10.1016/j.eswa.2012.02.125

Chen, S. (2016). Detection of fraudulent financial statements using the hybrid data mining approach. *SpringerPlus*, *5*(1), 1-16. https://doi.org/10.1186/s40064-016-1707-6

Chen, S., Goo, Y. J. J., & Shen, Z. D. (2014). A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements. *The Scientific World Journal*, *2014*. https://doi.org/10.1155/2014/968712

Comporek, M. (2020). The effectiveness of the Beneish model in the detection of accounting violations – the example of companies sanctioned by the Polish Financial Supervision Authority. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, *64*(10), 18-30. http://dx.doi.org/10.15611/pn.2020.10.2

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. (1996). Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC. *Contemporary Accounting Research*, *13*(1), 1-36. https://doi.org/10.1111/j.1911-3846.1996.tb00489.x

Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging financial social media data for corporate fraud detection. *Journal of Management Information Systems*, *35*(2), 461-487. https://doi.org/10.1080/07421222.2018.1451954

Golec, A. (2019). Ocena skuteczności modelu Beneisha w wykrywaniu manipulacji w sprawozdaniach finansowych. *Acta Universitatis Lodziensis. Folia Oeconomica*, *2*(341), 161-182. https://doi.org/10.18778/0208-6018.341.10

Gupta, R., & Gill, N. S. (2012). Prevention and detection of financial statement fraud–An implementation of data mining framework. *Editorial Preface*, *3*(8), 150-160. https://dx.doi.org/10.14569/IJACSA.2012.030825

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139-152. https://doi.org/10.1016/j.knosys.2017.05.001

Holda, A. (2020). Using the Beneish M-score model: Evidence from non-financial companies listed on the Warsaw Stock Exchange. *Investment Management & Financial Innovations*, *17*(4), 389-401. http://dx.doi.org/10.21511/imfi.17(4).2020.33

Jan, C. (2018). An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from Taiwan. *Sustainability*, *10*(2), 513. https://doi.org/10.3390/su10020513

Johnson, S. A., Ryan, H. E., & Tian, Y. S. (2009). Managerial Incentives and Corporate Fraud: The Sources of Incentives Matter. *Review of Finance*, *13*(1), 115-145. https://doi.org/10.1093/rof/rfn014

Kanapickiene, R., & Grundiene, Z. (2015). The model of fraud detection in financial statements by means of financial ratios. *Procedia-Social and Behavioral Sciences*, *213*, 321-327. https://doi.org/10.1016/j.sbspro.2015.11.545

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, *32*(4), 995-1003. https://doi.org/10.1016/j.eswa.2006.02.016

Kotsiantis, S., Koumanakos, E., Tzelepis, D., & Tampakas, V. (2006). Forecasting fraudulent financial statements using data mining. *International journal of computational intelligence*, *3*(2), 104-110.

Łapczyński, M. (2014). Hybrid C&RT-Logit Models In Churn Analysis. *Folia Oeconomica Stetinensia*, *14*(2), 37-52. https://doi.org/10.1515/foli-2015-0006

Mohammadi, M., Yazdani, S., Khanmohammadi, M. H., & Maham, K. (2020). Financial reporting fraud detection: An analysis of data mining algorithms. *International Journal of Finance & Managerial Accounting*, *4*(16), 1-12.

Nindito, M. (2018). Financial statement fraud: Perspective of the Pentagon Fraud model in Indonesia. *Academy of Accounting and Financial Studies Journal*, *22*(3), 1-9.

Ozcan, A. (2016). Firm characteristics and accounting fraud: a multivariate approach. *Journal of Accounting, Finance and Auditing Studies*, *2*(2), 128-144.

Ozdagoglu, G., Ozdagoglu, A., Gumus, Y., & Kurt Gumus, G. (2017). The application of data mining techniques in manipulated financial statement classification: The case of Turkey. *Journal of AI and Data Mining*, *5*(1), 67-77. https://doi.org/10.22044/jadm.2016.664

Pai, P. F., Hsu, M. F., & Wang, M. C. (2011). A support vector machine-based model for detecting top management fraud. *Knowledge-Based Systems*, *24*(2), 314-321. https://doi.org/10.1016/j.knosys.2010.10.003

Pazarskis, M., Drogalas, G., & Baltzi, K. (2017). Detecting false financial statements: Evidence from Greece in the period of economic crisis. *Investment Management and Financial Innovations*, *14*(3), 102-112. http://dx.doi.org/10.21511/imfi.14(3).2017.10

Rezaee, Z. (2005). Causes, consequences, and deterence of financial statement fraud. *Critical Perspectives on Accounting*, *16*(3), 277-298. https://doi.org/10.1016/S1045-2354(03)00072-8

Roxas, M. L. (2011). Financial statement fraud detection using ratio and digital analysis. *Journal of Leadership, Accountability, and Ethics*, *8*(4), 56-66.

Skousen, C. J., Smith, K. R., & Wright, C. J. (2009). Detecting and predicting financial statement fraud: The effectiveness of the fraud triangle and SAS No. 99. In M.Hirschey, K. John & A. Makhija (Eds.), *Corporate Governance and Firm Performance (Advances in Financial Economics, Vol. 13),* (pp. 53-81). Emerald Group Publishing Limited. https://doi.org/10.1108/S1569-3732(2009)0000013005

Spathis, C., Doumpos, M., & Zopounidis, C. (2002). Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques. *European Accounting Review*, *11*(3), 509-535. https://doi.org/10.1080/0963818022000000966

Steinberg, D., & Cardell, N. S. (1998). The hybrid CART-Logit model in classification and data mining. *Salford Systems White Paper*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.2179&rep=rep1&type=pdf

Subair, M. L., Salman, R. T., Abolarin, A. F., Abdullahi, A. T., & Othman, A. S. (2020). Board Characteristics and the Likelihood of Financial Statement Fraud. *Copernican Journal of Finance & Accounting*, *9*(1), 57-76. https://doi.org/10.12775/CJFA.2020.003

Yao, J., Pan, Y., Yang, S., Chen, Y., & Li, Y. (2019). Detecting fraudulent financial statements for the sustainable development of the socio-economy in China: a multi-analytic approach. *Sustainability*, *11*(6), 1579. https://doi.org/10.3390/su11061579

Yuan, J., Yuan, C., & Deng, X. (2008). The effects of manager compensation and market competition on financial fraud in public companies: an empirical study in China. *International Journal of Management*, *25*(2), 322-335.

Zhu, M., Philpotts, D., Sparks, R., & Stevenson, M. (2011). A hybrid approach to combining CART and logistic regression for stock ranking. *The Journal of Portfolio Management*, *38*(1), 100-109. https://doi.org/10.3905/jpm.2011.38.1.100