ERASMUS UNIVERSITY ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS
MSc Accounting, Auditing and Control

Can Random Forest improve two-stage linear discretionary accrual models?

The content of this thesis is the sole responsibility of the author and does not reflect the view of either Erasmus School of Economics or Erasmus University.

Author:	Johan Hendrik Prins
Student number:	475043
Thesis supervisor:	dr. Y. Li
Second assessor:	dr. J. Pierk
Date:	24-04-2023

ABSTRACT

The purpose of this study is to improve discretionary accrual models for banks by using random forest to predict the nondiscretionary component of loan loss provisions. Investigation of managerial conduct within the banking sector depends strongly on the accuracy and effectiveness of accrual models to proxy for earnings management. Existing literature favours linear regressions to model the nondiscretionary component of accruals. Based on nonlinearity concerns and overall better prediction power, we argue for random forest regressions as an alternative to linear models. Using United States banking data for the period 2010-2021, we compare the results of linear and random forest models based on R^2, mean (absolute) error, persistence of the discretionary accrual and the ability of the model to identify cases of artificially induced managed earnings. Random forest regressions outperform linear regressions in every test apart from the mean error.

Keywords: earnings management, random forest regression, artificially induced earnings management

Table of Contents

1 Introduction

The objective of this research is to use machine learning to improve existing discretionary loan loss provision models. Quantitative models form the foundation of earnings management literature, with the ability to accurately distinguish companies' (non)discretionary accruals at the core of researchers' concerns and an extensive branch of literature dedicated to this matter alone. Current literature often employs two-stage linear models to first estimate the non-discretionary component of loan loss provisions, and then uses the residual of the first estimation as a proxy for earnings management. We argue that ordinary linear models are inferior to nonlinear models in predicting the nondiscretionary component of the accrual in the first stage of the model and propose the random forest regression as a nonlinear alternative.

Despite arguments that they are just an accounting number, earnings are still an important item on any company's income statements. Earnings are an accrual accounting resolution to quantify a firm's profit or loss over a certain period of time and are used by both internal and external users for performance indication, valuation, or as a metric in debt covenants or bonus plans (Dechow, 1994). Graham, Harvey & Rajgopa (2005) conducts a survey among over 400 executives and finds that financial officers see earnings as the most relevant financial metric to outsiders, rather than cash flows. As so much weight is put on a number that essentially represents the difference between income and expenditure during a period, incentives may compel a firm's management to manage earnings in their favour.

The concept of earnings management is not limited to a single industry, country or method and can happen for a variety of reasons. Prior research has spent a vast amount of time and energy investigating whether management uses their discretion to manage earnings. Many studies have answered this question affirmatively, providing evidence that management conducts earnings management to affect reported earnings. Incentives to manage earnings include higher company valuation in case of imminent public offerings (Teoh, Welch & Wong, 1998) or stock-for-stock mergers (Erickson & Wang, 1999), meeting analysts' forecasts (Graham et al., 2005), smoothing income (Sood, 2012), debt covenants or regulation (Jones, 1991). Alternatively, incentives can be non-financial such as an increased likelihood for career advancement, or personal prestige. Regardless of the underlying motivation, the presence of earnings management indicates a conflict of interest between the numerous stakeholders. Earnings management impacts all stakeholders, as its presence decreases the transparency of financial reports, hindering efficient resource allocation from the worst- to the best performing companies.

This study focusses on banks for several reasons. First, the literature is not yet decided on a single approach to model the discretionary component of loan loss provisions (Beatty & Liao, 2014), meaning existing earnings management studies rest on a precarious foundation. Secondly, due to the way the banking sector and the financial system are intertwined, the safety and soundness of the banking sector is of critical importance to the economy. Monitoring and detecting extreme levels of earnings management is one of the means regulators have to ensure proper reserves in the banking sector (Cornett, McNutt & Tehranian, 2009). On similar note, Cheng, Warfield & Ye (2011) underline the importance of monitoring the presence of earnings management in the banking industry, arguing that the credit crisis of 2008 shows just how critical banks are to a properly functioning economy. Accurate modelling of the discretionary component of banks' loan loss provisions is therefore of fundamental importance to a wide range of stakeholders, including supervisory boards and regulators. Lastly, we argue that the relation between nondiscretionary bank accruals and its explanatory economic variables is non-linear. As of yet, nearly all studies investigating earnings management in the banking industry utilize linear regressions, and not without reason. Linear regressions are a classic analytical method that have proven themselves over time and combine strong analytical power with ease of interpretations. However, such a statement only holds when certain conditions are met. When faced with the task of predicting values for nonlinear relationships or higher order interaction terms, both ease of interpretation and model selection suffer. Nonlinear machine learnings algorithms, such as the regression tree, can outperform classical approaches when the relationship between the independent and dependent variables is not well approximated by a linear model (James, Witten, Hastie & Tibshirani, 2021).

Whereas earnings management research in sectors such as retail or industrial firms commonly focusses on R&D, operating expenses or revenue recognition, financial firms require a different approach due to large differences in balance sheet composition. For banks, the loan loss provision (LLP) is found to be the most significant accrual for earnings management due to its relatively large size and the high degree of discretion that can be exercised by the banks' management. Banks set aside a certain part of their outstanding loans to reflect expected defaults in the future on current debt outstanding. As the estimation of future defaults is not an exact science, bank managers can exert influence over the provision.

This paper uses five tests to examine the effectiveness of tree-based regression algorithms in modelling the (non)discretionary component of loan loss provisions and to compare its effectiveness to existing linear models. To test model performance, we investigate R^2 values, mean (absolute) errors, the persistence of the discretionary component of loan loss provisions, and the ability of the model to recognize cases of earnings management. Due to the limited availability of data on actual cases of earnings management in our study, we artificially inflate the loan loss provision in randomly selected observations with 0-0.15% of total loans and assess to extent which the models are able to recognize these cases. In all cases with the exception of the mean error metric, the random forest model outperforms the classic linear models. Based on these results, we advise future research to consider using the random forest algorithm when estimating the nondiscretionary component of loan loss provisions.

2 Literature review

2.1 Real and accrual earnings management

Current earnings management literature can typically be divided in two categories. The first approach focusses on real earnings management. Through real earnings management, firms can manipulate earnings upwards or downwards not by changing their accounting policy or using judgement, but through a more tangible approach. This approach is common in samples excluding financial institutions (e.g. Bartov (1993), Roychowdhury (2006)), but rarely applied to financial institutions. Arguably because of banks' limited opportunities to cut R&D costs or overproduction, which Roychowdhury (2006) identified as common real earnings management approaches. Ertan (2021) is one of few exceptions to investigate real earnings management in banks.

The second and most common approach is the study of earnings management in financial institutions through the manipulation of accruals. Earnings as reported on the income statement consists of two components: cash flows and total accruals. Total accruals consist of discretionary and non-discretionary accruals. Stubben (2010) suggests three features of specific accrual accounts that make them fit for academic research. The ideal accrual is i) common across industries, ii) leaves room for managements' discretion, and iii) is economically significant. Studies examining non-financial samples tend to focus on aggregate accruals due to an unclear hierarchy in accrual importance. Banking literature on the other hand tends to focus on loan loss provisions whilst excluding other accruals. Using specific accruals rather than aggregate accruals has the added benefit of decreasing concerns raised at aggregate accrual models (McNichols & Stubben, 2018). The preference for loan loss provisions as sole accrual can be explained by strong explanatory power of total accruals. Beatty & Liao (2014) find that the loan loss provision is the largest accrual explaining 56% of the variability of total accruals, nearly double the impact of the second most relevant variable. Second, loan loss provisions are essentially managements' best estimates of future losses, meaning considerable judgement can be exercised over them.

2.2 The loan loss provision

Arguably the most infamous accrual for earnings management in banks is the loan loss provision. The primary purpose of banks is to collect deposits and issue these deposits to individuals, firms, or other entities in return for interest. Apart from liquidity risks and exposure to complicated derivative structures, one of the main risks for banks is its borrowers defaulting on their principal and/or interest. Banks prepare for future expected defaults by expensing a loan loss provision in the income statement, adding that amount to the loan loss reserve on the balance sheet. The loan loss reserve, sometimes referred to as the 'Allowance for Loan and Lease Losses', is a contra asset recorded to represent an estimate of the total value of uncollectible loans and leases and is used to reduce the book value of outstanding loans and/or leases to the amount that is expected to be collected over the lifetime of the loan or lease. When a loan is considered (partly) uncollectable, this amount is charged off against the reserve. Loan loss reserves change frequently, increasing with loan loss provisions and decreasing with net charge-offs. Similar to how retail firms adjust their allowance for doubtful accounts through bad debt expense, banks make adjustments to their loan-loss reserve through loan loss provisions (Basu, Vitanza & Wang, 2020). When loans are deemed uncollectible and thus charged off, the reduction in loans is charged against the loan loss allowance leaving net income unaffected. Effectively, assuming loan write-offs do not exceed the total allowance, the timing of loan losses that become uncollectible bears no significance from the standpoint of net income. The decisive factor lies in the moment at which the management decides to record the provision.

Management estimates the required loan loss provisions based on historical experiences, statistical factors and professional judgement. Bank managers can exercise considerable power over the latter (Ozili & Outa, 2017), creating an environment where opportunistic behaviour by managers could harm the primary goal of accruals: to convey private information based on managements' bests expectations. To balance varying interests, bank regulators periodically review loan loss reserves. If the loan loss reserve falls short of expected losses, the bank's capital ratio overstates its capacity to withstand losses, endangering the safety and soundness of the bank. Regulators adopt a cautious and forward-looking perspective towards the use of loan loss allowances, considering their objective of upholding the security and stability of banks (Cornett et al., 2009).

2.3 One- and two-stage accrual models

As the 'true' value of the (non)discretionary accrual is unobservable, existing literature models the nondiscretionary part of the loan loss provision as a linear combination of credit risk indicators and macroeconomic variables and uses the residual of the estimation as a proxy for the discretionary component. Researchers have adopted two methods when modelling specific accruals such as loan loss provisions. The first method includes the variable of interest among a list of control variables to assess whether the independent variable, generally an accounting measure, significantly affects the dependent variable. The control variables are meant to capture the nondiscretionary part of the accrual. Models taking this approach are called one-stage models and are used in Lobo & Yang (2001), Liu & Ryan (2006) and Alali & Jaggi (2011), among others.

The second method is the two-stage approach, which in the first stage predicts the accrual based on all economic factors deemed relevant for the objective determination of the loan loss provision. This represents the nondiscretionary component of the accrual. The residual of this prediction represents the discretionary part of the loan loss provision and is used as the dependent variable in the second stage. Here, discretionary accruals are regressed on a variable of interest (possibly including a set of control variables) to test a hypothesis on earnings management. Beaver & Engel (1996), Kanagaretnam, Krishnan & Lobo

(2009), Cheng et al. (2011) and Grougiou, Leventis & Dedoulis (2014) among others take this approach to model loan loss provisions. Despite concerns raised throughout time regarding possible attenuation biases (see Beaver (1987), Chen, Hribar & Melessa (2018)), use of the two-stage accrual model is still very popular within the accounting and finance literature. According to Chen et al. (2018), 61 studies published in main accounting papers[1] between 2011 and 2015 employed this procedure, of which 24 studies were focused on separating discretionary and nondiscretionary accruals. The choice of research design depends on the aim of each study. Given the purpose of this study is to improve the predictive power of loan loss provision models, rather than interpret the effect of a certain variable, the two-stage approach will be focused on.

2.4 Existing accrual models

Whereas non-financial accrual studies have conveyed towards a selected few preferred models for two-stage accrual models, loan loss provision studies have not yet reached such a consensus. Beatty & Liao (2014) summarizes models from nine different accrual models between 1994 and 2012 and conducts a factor analysis on the residuals to understand the driving factors. Based on the factor analysis, four models are proposed that capture as much of the relevance as possible. Collectively, these four models, named linear model 1, 2, 3 and 4, will be used as the current literature's most substantiated models.

$$LLP_t = \alpha_0 + \alpha_1 \Delta NPA_{t+1} + \alpha_2 \Delta NPA_t + \alpha_3 \Delta NPA_{t-1} + \alpha_4 \Delta NPA_{t-2} + \alpha_5 SIZE_{t-1} + \alpha_6 \Delta LOAN_t + \alpha_7 \Delta GDP_t + \alpha_8 CSRET_t + \alpha_9 \Delta UNEMP_t + \varepsilon_t \quad (1)$$

$$LLP_t = \alpha_0 + \alpha_1 \Delta NPA_{t+1} + \alpha_2 \Delta NPA_t + \alpha_3 \Delta NPA_{t-1} + \alpha_4 \Delta NPA_{t-2} + \alpha_5 SIZE_{t-1} + \alpha_6 \Delta LOAN_t + \alpha_7 \Delta GDP_t + \alpha_8 CSRET_t + \alpha_9 \Delta UNEMP_t + \alpha_{10} ALW_{t-1} + \varepsilon_t \quad (2)$$

$$LLP_t = \alpha_0 + \alpha_1 \Delta NPA_{t+1} + \alpha_2 \Delta NPA_t + \alpha_3 \Delta NPA_{t-1} + \alpha_4 \Delta NPA_{t-2} + \alpha_5 SIZE_{t-1} + \alpha_6 \Delta LOAN_t + \alpha_7 \Delta GDP_t + \alpha_8 CSRET_t + \alpha_9 \Delta UNEMP_t + \alpha_{10} CO_t + \varepsilon_t \quad (3)$$

$$LLP_t = \alpha_0 + \alpha_1 \Delta NPA_{t+1} + \alpha_2 \Delta NPA_t + \alpha_3 \Delta NPA_{t-1} + \alpha_4 \Delta NPA_{t-2} + \alpha_5 SIZE_{t-1} + \alpha_6 \Delta LOAN_t + \alpha_7 \Delta GDP_t + \alpha_8 CSRET_t + \alpha_9 \Delta UNEMP_t + \alpha_{10} ALW_t + \alpha_{11} CO_t + \varepsilon_t \quad (4)$$

Where *LLP* is the loan loss provision divided by lagged total loans, *NPA* is the change in nonperforming assets over the quarter scaled by lagged total loans, *SIZE* is the natural log of total assets and *LOAN* is the change in total loans over the quarter divided by lagged total loans. *GDP*, *CSRET* and *UNEMP* are the percental change in respectively the gross domestic product, Case Shiller Home Price Index and unemployment over the quarter. *CO* is the net charge-off divided by lagged total loans and *ALW* is the loan loss allowance divided by total loans. Nonperforming assets consist of loans that are no longer paying interest and loans that are at least

[1] *Contemporary Accounting Research, Journal of Accounting and Economics, Journal of Accounting Research, Review of Accounting Studies, and The Accounting Review*

90 days overdue. The *NPA* variable is therefore an indicator of loan quality. Variables *GDP, CSRET and UNEMP* are included because these variables provide information about the macroeconomic environment.

2.5 Nonlinearity concerns in banking accruals

Economic research often uses linear regressions to approximate functions because their coefficients can be interpreted as marginal effects. Our study differentiates from existing literature by applying random forest regression instead of a linear regression to predict the non-discretionary part of loan loss provisions. As the primary goal of the first step in the two-stage approach is to predict the non-discretionary component of loan loss provision rather than to interpret marginal effect, we believe a hypothesized increase in predictive power of the random forest regression outweighs the loss of interpretability. When prediction results are not significantly different, secondary properties like ease of use or interpretability of the models can be considered when choosing a model.

We argue in favour of random forest regressions for several reasons. First, we argue that the economic factors that influence the nondiscretionary component of the loan loss provision might not be strictly linear. Wu (2014) finds that nonlinear accrual models outperform traditional models by incorporating the asymmetric influence of performance. Balboa, Lopez-Espinosa & Rubia (2013) investigates inconclusive evidence regarding income smoothing through loan loss provisions in the banking industry, arguing that main conclusions from previous studies are sensitive to the choice of sample and the model used. As the literature has not decided on a preferred model, differences in research designs can influence conclusions. They argue that the relation between loan loss provisions and earnings is nonlinear because incentives and ability to manipulate earnings is dependent on the relative size of all variables. Standard linear regressions fail to appropriately capture nonlinear responses of this kind, creating a bias in the estimator that could lead to misleading conclusions, generally underestimating earnings management in cases with extreme opportunities and overestimating in cases with no to moderate opportunities. Basu et al. (2020) finds a V-shaped relation between loan loss provisions and changes in non-performing loans and argues that failure to account for this relation can bias the conclusions of studies that assumed linearity, arguing that standard linear models would underpredict at the tails and over predict in the middle of the non-performing loan change distribution. Applying a nonlinear algorithm could increase prediction power by reducing the negative impact from nonlinearity concerns.

3 Methodology

3.1 Machine learning and random forest regression

The following section will discuss the choice of machine learning algorithm. Machine learning algorithms are designed to improve operational performance by learning from experience. As such, machine learning is often seen as a subsection of artificial intelligence. In order to learn from experience, the algorithm is designed to build models based on the data provided to it. Multiple classifications of machine learning algorithms exist, of which supervised and unsupervised algorithms are arguably the most well-known. The difference between supervised and unsupervised learning algorithms lies in the way the algorithm trains on the data provided to it. Supervised systems require labelled training data, whereas unsupervised methods are used when such labels are unavailable (Zhou, 2016). Labelling effectively means that human interaction is needed to indicate (label) the input and output variables. The algorithm in turn uses the provided variables to search for a relationship with the desired output variable. This study utilizes supervised systems as all data is cleaned and labelled. There are three state-of-the-art supervised machine learning algorithms, each of them showing comparable accuracy (Jaiantilal, 2013). These algorithms are support vector machine, boosting and random forest regression. Unfortunately, there is no decisive method to determine which algorithm is best for which individual dataset or research question. Based on this assumption of comparable regression results, we focus on random forest regressions as it is less computationally intensive (Jaiantilal, 2013) and can be used to assess the relative importance of model features, both of which are qualities we deem desirable.

The random forest algorithm is introduced in Breiman (2001) and is a machine learning algorithm that is widely used for classification and regression. The algorithm is an ensemble method based on the combined predictive power of many individual trees. Regression trees, which are a type of decision trees, are simple structures that split the data according to splitting rules. When graphically represented, trees are often drawn upside-down. The top of the tree, referred to as the *root node*, resembles all available data. The algorithm defines a cut-off point and splits the previous node in two new nodes until either the reduction in prediction error gained by splitting data, or the number of observations in sample is lower than a predefined threshold. The nodes that are no longer subdivided are called the terminal nodes (James et al., 2021).

Splitting rules are based on the principles of recursive partitioning which splits the features into groups with similar response values. For continuous variables, the optimal cut-off point to divide the previous node into two more nodes is selected based on the highest mean squared error reduction (Gomes & Jelihovschi, 2020). This is accomplished by a brute force procedure where the algorithm determines the residual of the sum of squares between the observed and mean value for each of the two nodes for multiple cut-off points. This process is repeated until the cut-off point with the lowest residual sum of squares is found. When there is more than one feature, the feature with the highest reduction will be used in the first node.

Singular regression trees have several disadvantages compared to other prediction methods. Individual trees tend to be very sensitive to small changes in data and are likely overtrain on the provided data, resulting in high in-sample prediction power but poor out-of-sample prediction power. Random forest improves on individual regression trees in two different ways: through bagging and random selection of predictors. These two methods will be discussed individually.

Bootstrap aggregation, also known as bagging, is a commonly used technique to reduce variance in learning algorithms. In bagging, the initial training set is used to create multiple smaller training sets with replacements. This means that an observation may be selected either zero or multiple times for each individual training set. The trees are trained on individual bootstrapped training set, and finally aggregated by averaging predictions from all individual regression trees. Bagging is especially beneficial to tree-based methods due to their inherently low bias but high variance. Random forest provides further improvement over the previously described bagged trees by decorrelating the trees. For each split in an individual regression tree, only a specific subset of features is considered by the algorithm, ignoring the others. This process might sound counterintuitive at first, however it offers an important benefit. It prevents the model from building many similar individual trees based on a small set of strong predictors. Averaging nearly identical trees will not decrease variance as much as averaging many uncorrelated trees. For an extensive description of the methodology behind the random forest I refer to Breiman (2001).

3.2 Model assessment

To assess if supervised machine learning models outperform traditional discretionary loan loss provision models, two things are needed. The first is the current literature's most effective and supported loan loss prediction model. As the literature has not yet decided on a preferred model, we make use of the four models suggested in Beatty & Liao (2014). These models combine different features of nine models used in earlier studies. These four models are used in later studies (e.g. Basu et al. (2020) and Tran & Houston (2021)). The residual of the estimation is the proxy for discretionary loan loss provision used to represent earnings management. As the purpose of this study is to compare empirical results, we estimate all models on the same dataset. This means that all firm-specific variables are scaled by lagged total loans to reduce heterogeneity concerns in the linear model, even though random forest does not suffer from potential heterogeneity issues and could alternatively be used to estimate the loan loss provision directly. Effectively, loan loss provision is estimated as a percentage of lagged total loans.

Secondly, a proper set of tests by which compare model performance is needed. To our knowledge, there is not one single test that can unambiguously determine the best model for all occasions. As such, model performance will be assessed by an array of tests. Based on the relative success for each model, we determine if supervised machine learning algorithms improve existing loan loss provision models. In line with Pae (2005), we use three evaluation metrics to test the accuracy of our random forest regressions. The first three metrics are the R^2 measure, absolute forecast errors and mean forecast errors. The R^2 measure is a goodness of fit test that explains the extent to which the variance of the actual value of *LLP* can be explained by the variance of the predicted value of *LLP* in the testing set. Absolute error is the absolute difference between predicted and actual values while mean error is the average of the errors. Given the notion that earnings management reverses over a long enough sample period, the best performing model should produce mean errors closest to zero. Additionally, following model evaluation as discussed in Dechow, Richardson & Tuna (2003), Medeiros, Dantas & Lustosa (2012) and Glen (2015), we perform an analysis on the persistence of discretionary and non-discretionary accruals. Discretionary accruals are a result of internal and external forces exercised on management. Though these effects may be connected to actual events or firm performance, discretionary accruals should, by their very nature, still reverse through time. More so than nondiscretionary part of the loan loss provision, which should change only with economic factors. When regressing future loan loss provisions on current discretionary and non-discretionary accruals, we should expect larger coefficients for the nondiscretionary component and smaller coefficients for discretionary loan loss provisions.

$$LLP_{t+1} = \alpha_0 + NDLLP_t + DLLP_t + \varepsilon_t \tag{5}$$

Where LLP_{t+1} is the loan loss provision in the next quarter, $NDLLP_t$ is the loan loss provision in the current quarter as predicted by the model, and $DLLP_t$ is the discretionary component of loan loss provision, which is the residual of the previous estimation.

Finally, model performance is assessed by the ability of the models to recognize cases of earnings management. Unfortunately, data on predetermined cases of earnings management are notoriously lacking in earnings management literature. Previous studies have attempted to deal with this issue by focussing either on financial restatements and Securities and Exchange Commission (SEC) intervention, or by artificially introducing earnings management into their sample. Dechow, Hutton, Kim & Sloan (2012), for instance, use SEC Accounting and Auditing Enforcement Releases (AAER) to test the power of their model by evaluating how well different models recognize these cases. These releases are a list of enforcement actions related to financial reporting and civil lawsuits and are published by the SEC. The advantage of using such a dataset is that no assumptions need to be made about the timing or the magnitude of the managed provision. Additionally, the SEC arguably only intervenes in the most extreme cases of earnings management. Should a model fall short of identifying such cases, it would be unlikely to perform well in samples where relatively small earnings management is taking place. Unlike prior research that investigated samples excluding financial firms, this study's sample did not encounter substantial intervention by the SEC throughout the sample period. According to Audit Analytics data, there were only three cases of restatements pertaining to fraud, irregularities or misrepresentation related to the balance sheet accounts for the banking sector, which is the category Lo, Ramos & Rogo (2017) identified as the most useful for studies focussing on earnings management. Given the limited number of observations of SEC interventions, this study will focus on artificially inducing earnings management in the sample and evaluate the extent to which the models can identify these cases. Dechow, Sloan & Sweeney (1995), Stubben (2010) and Dechow et al. (2012), among others, implement this approach.

Artificially inducing earnings management faces several challenges. Explicit assumptions need to be made regarding the timing of the accruals. The external validity of the results rest upon how representative the assumptions are. The first assumption is about the timing of the accruals. As discretionary accruals should sum to zero over the lifetime of the firm, an adjustment of similar size but different sign during the observed time must be made. We assume reversal will take place in the quarter following the managed quarter. Secondly, we need to make assumptions about the size of the discretionary component of loan loss provisions. In other words, to what extent do we expect management to have the real ability to manage earnings. While the literature recognizes and accepts the practice of earnings management via loan loss provisions, there are no clear suggestions regarding the size of the issue relative to the size of the company. As such, the loan loss provision of randomly selected quarters will be increased with 0 – 0.15% of outstanding loans. Thirdly, the effect of changes to other accounts needs to be considered. Changes in loan loss provision will affect the allowance for loan and lease losses. These accounts are adjusted accordingly. An increase in loan loss provision also decreases tier 1 capital ratios and shareholder equity. None of these changes however effect the variables used in our models, meaning they do not need to be adjusted.

Combining the beforementioned assumptions leads to the following steps to test the ability of the models to recognize cases of artificially induced earnings.

1) We randomly select 100 observations that have at least one more observation to account for the reversal.
2) The loan loss provision for these observations is increased with 0 - 0.15% of outstanding loans, replicating a scenario where high profits are stored for future periods.
3) A new variable *EM* is created and set to true to indicate that the earnings of this observation have been managed.
4) The loan loss provision is reduced with the exact same amount in the following quarter. Variable *EM* is also set to true. The total allowance for loan and loss leases is adjusted accordingly in both time periods.
5) We repeat the linear regression and random forest regression for each model using all observations including the modified ones in a 75/25 split.
6) Observations are ranked based on the estimated component of the absolute value of the discretionary accrual. The values are divided into deciles.
7) Steps 1-7 are repeated ten times and the number of observations per period are averaged to reduce the influence of chance.
8) A Chi-square test is used to test if the sum of managed observation per decile significantly differs from the distribution of non-managed earnings.

Given the absence of clear evidence regarding the direction and magnitude of actual earnings management in the sample, we cannot simply evaluate model performance by its ability to put a certain number of observations within a certain decile. Instead, we argue that the model that most accurately predicts the nondiscretionary component of the loan loss provision should require a smaller increase to the discretionary loan loss provision to start allocating more observations to the top deciles, without specifically stating which decile. The Chi-square test is used to determine the increase in loan loss provision that is required to change the allocation of observations to deciles from random to non-random. Observations in which the loan loss provision is modified are labelled as earnings management regardless of the direction of the modification. Combined with ranking based on the absolute value of the discretionary accrual, this method allows for model assessment irrespective of the direction of management. This approach enables comparisons between models based not solely on their ability to recognize income-increasing management but also on their ability to detect either direction of management. In total, there are 200 modified observations of which 50 are present in the testing set.

3.3 Hyperparameter tuning of the random forest

The random forest algorithm has decent out of the box performance which can be improved further by hyperparameter tuning. The first one is the number of candidate variables, referred to as *mtry* in R, considered at each split. Higher values of *mtry* result in trees that on average better fit the data, as more variables can be considered at each split. Low values for *mtry* result in more unique and therefore less correlated trees, resulting in more stability when all individual trees are averaged (Probst, Wright & Boulesteix, 2018). The standard method for supervised machine learning algorithm hyperparameter tuning is by using *k-fold cross validation*. Random forest, and other algorithms that use bagging, have the alternative option to use out-of-bag (OOB) observations. When a model bootstraps with replacement, not all observations are used and other observations are used multiple times. Given a large enough dataset, approximately two-third of the data will be utilized, meaning that one-third will not be used. The observations that are not used to construct an individual tree are denoted as OOB observations for that tree. Each tree has several observations that are not used in the creation of that tree and can therefore be used for hyperparameter tuning. This method has significantly less computation time than k-fold cross validation and performs well in all but some specific cases (Probst et al., 2018); none of which apply here. The optimal number of candidate predictors is computed by a trial-and-error method with the objective of finding the smallest out-of-bag estimation error and will then be used on the testing data.

Secondly, the number of trees used in the bagging process needs to be specified. Increasing the number of trees reduce the variance at marginally decreasing rates while increasing computation time. Probst & Boulesteix (2018) find that for random forest classifiers, tuning the number of trees is not required. For regression purposes, they argue on a theoretical basis that the same applies, though further research could extend. Oshiro, Perez & Baranauskas (2012) find that there is no significant difference between the performance of 256, 512, 1024 or more trees for their dataset. Though deviating from the standard number of trees of 500 is not necessary, we use OOB estimation errors to determine the number of trees that corresponds to a stable mean squared error. Finding the number of trees that produces a stable mean squared error balances model performance and computational costs.

The *nodesize* parameter influences tree size by creating a lower bound for the number of observations required to continue growing the tree. Put differently, when all unsplit nodes contain fewer observations than the specified threshold, the tree is fully grown. Increasing this number means this threshold is met more quickly, resulting in smaller trees and lower computation time. Decreasing *nodesize* will increase model complexity and fit to the training data, increasing the risk of overfitting. The standard *nodesize* for random forest regression is set to 5 and is said to offer decent balance and performance for most scenarios. We attempt to improve performance by considering a *nodesize* of 1 to 30 for each model and selecting the number that minimizes the OOB error.

As supervised machine learning algorithms are prone to overtraining, we split the data to ensure out-of-sample performance. As this study uses OOB observations for hyperparameter tuning, there is no requirement to use a separate validation set. The model is trained on 75 percent of the data and tested on the remaining 25 percent.

4 Data

Our sample consists of banks from the Compustat Bank Fundamentals database which contains financial data of the largest and most important banks in the United States since 1950. The Compustat Bank Fundamentals database is accessed through Wharton Research Data Services. Requested quarterly balance sheet and income statement items are based on previous literature and – among others – include loan loss provisions, net charge offs, and non-performing assets. The initial dataset includes observations from Q1 2010 until Q4 2021, totalling 30.235 firm-quarter observation. A total of 9,944 observations are dropped because of missing values. The remaining sample consists of an unbalanced panel of 20,291 observations, representing 789 unique banks. Bank data is merged with quarterly Gross Domestic Product (GDP) data retrieved from the World Bank (World Bank, 2022), unemployment data from the United States Department of Labor (U.S. BUREAU OF LABOR STATISTICS, 2022) and the Case Shiller Home Price Index data from the Federal Reserve Economic data (Federal Reserve Economic Data, 2022). GDP growth rate is calculated relative to the GDP of Q1 2010. All bank variables have been winsorized at 1% and 99% levels to reduce the effect of outliers.

A pooled dataset has several advantages over time-series and cross-sectional oriented models. Time series models, such as the original Jones-model, would force the discretionary accrual to be zero for a specific firm over time. This approach suffers from several drawbacks. The sum of discretionary accruals over a given period does not have to be zero, as this is dependent on the chosen time period and assumes a stable discretionary accrual generating ability. The latter assumption is problematic, based on the findings of Dopuch, Mashruwala & Seethamraju (2012). Many accrual models alternatively opt for cross-sectional analysis. Though arguably fit for the non-banking sector, this method poses risks for loan loss provision models. A popular research topic on earnings management in the banking sector is income smoothing by increasing accruals in good times to create a buffer for when inevitably the bad times arrive. Done to its fullest extent, this means that earnings are not affected by fluctuating loan write-downs through varying business cycles. See Kanagaretnam, Lobo & Yang (2004). Good and bad times for banks are arguably influenced by global factors and in some extent applicable to all banks at the same time, more than for non-financial firms. Forcing the discretionary accrual to be zero for all bank observations in each period, as would be done by a cross-sectional model, would therefore be problematic. When incentives for earnings management are correlated with macroeconomic activities, the banking sector would engage in earnings management at the same time (Beaver & Engel, 1996). Though suffering from its own disadvantages, such as assuming constant coefficients across firms and across years, we argue that pooled data is the still best approach.

Table 1 shows that bank data is heavily skewed to the left, showing large relative differences between the third quartile and the maximum value for all variables caused by the presence of a small number of large banks in the sample. During the sample period, net charge-offs slightly exceed provisions to the loan loss reserve. On average, the loan loss reserve is increased with 0.082% of total loans each quarter.

Table 1. Sample descriptive statistics.

Panel A: Means, quartiles and standard deviation of bank data

Variables	Mean	Std. Dev.	Min	p25	p75	Max
Total assets	26618	117058	177	833	6385	946959
Total loans	14434	60231	102	546	4319	486622
Net charge-offs	15	74	-2	0	2	602
Non-performing assets	143	509	0	7	53	3780
Provision for loan losses	13	61	-13	0	2	480
Loan loss allowance	169	662	1	7	50	5080

Panel B: Means, quartiles and standard deviation of variables used in the estimation

LLP	0.00082	0.00153	-0.00229	0.00008	0.00101	0.00902
LOAN	0.02155	0.05201	-0.07075	-0.00033	0.03230	0.31359
NPA	-0.00085	0.00463	-0.02237	-0.00168	0.00049	0.01512
SIZE	7.90804	1.69738	5.17650	6.72555	8.76164	13.76101
CO	0.00083	0.00167	-0.00111	0.00003	0.00087	0.01001
ALW	0.01437	0.00823	0.00327	0.00941	0.01670	0.05151
CSRET	0.02692	0.03814	-0.0425	0.00345	0.04499	0.15688
UNEMP	-0.00102	0.18316	-0.04500	-0.00400	0.00000	0.11200
GDP	0.01332	0.02803	-0.13572	0.00767	0.01766	0.11251

Note. Panel A depicts statistics of quarterly bank observations from 2010 – 2021. All values are in millions of dollars. Panel B shows descriptive statistics of variables used for linear and random forest regression. Variable definitions: *LLP* is loan loss provision scaled by lagged total loans. *LOAN* is the change in total loans over the quarter divided by lagged total loans. *NPA* is the change in non-performing assets divided by lagged total loans. *SIZE* is the log of total assets. CO refers to the net charge off divided by lagged total loans. *ALW* is the loan loss allowance divided by lagged total loans. *CSRET* is the Case-Shiller real estate index return over the quarter. *UNEMP* is the absolute change in unemployment percentage over the quarter. *GDP* is the growth rate of the gross domestic product over the quarter. Descriptive statistics for *NPA* are the same as for NPA_{t-2}, $NPAt_{-1}$ *and* NPA_{t+1}.

5 Empirical results

5.1 Model tuning

The *mtry*, *ntree* and *nodesize* parameters of the random forest models 1 to 4 are tuned using OOB estimation errors. Based on a grid search using the tuneRF package, setting the *mtry* parameter to 4 produces the lowest error across models 1, 3 and 4. OOB estimation errors stabilize around 350 trees for all four models, suggesting this number of trees as an acceptable balance between predictive performance and computation time. Further results of the hyperparameter tuning are depicted in Table 2.

Table 2. Random forest parameter values.

Paramater	Model 1	Model 2	Model 3	Model 4
mtry	4	3	4	4
ntree	350	350	350	350
nodesize	11	9	16	15

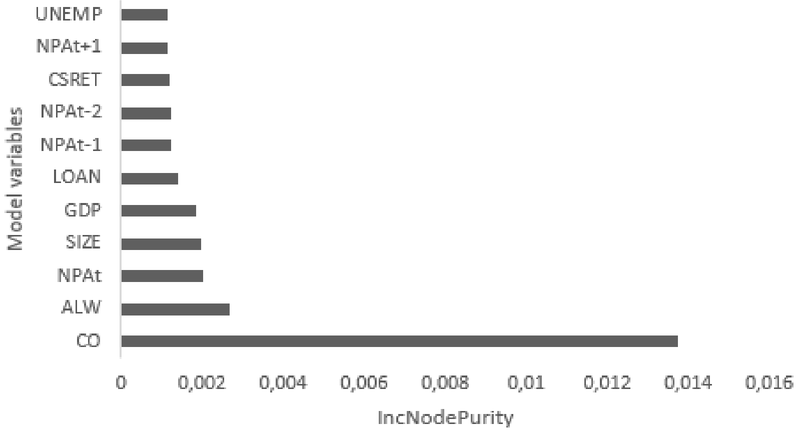
Note. This table depicts the parameter values that minimize out-of-bag estimation error for random forest models 1 to 4. Increasing *ntree* theoretically reduces OOB error even further, but at increasing computational costs. *Mtry* is the number of variables considered at each split of the tree, *ntree* determines how many individual regression trees make up the random forest and *nodesize* determines the minimal number of observations required to perform an additional split.

5.2 Model estimation

Table 3 presents the outcomes of the model estimation for linear models 1 to 4, which were trained on a dataset comprising of 75% of total observations. Pearson correlation values are included in the Appendix. All variables, apart for NPA_{t+1} for model 2 and *SIZE* for model 3 and 4, are significant at the conventional levels. Variable *NPA* has the expected positive coefficient in models 2, 3 and 4 but is unexpectedly negative in model 1. As the value of nonperforming assets divided by total loans is an indicator of loan quality, the negative coefficient suggests that a deteriorating loan portfolio results in a reduction of the loan loss provision, which is counterintuitive. A possible explanation for this phenomenon is that the omission of the *CO* variable from model 1 leads to a distortion of the coefficient for the *NPA* variable as these two variables exhibit a relatively strong correlation and the *CO* variable is strongly correlated with the dependent variable. The *LOAN* coefficient of 0.001 suggests that for model 3 and 4, managers increase their loan loss reserves when total loans increase, while in model 1 the loan loss provision decreases when loans increase ceteris paribus. Variables *GDP* and *CSRET* both have a negative coefficient across all models, indicating that managers on average decrease their loan loss provisions when the economy is growing and home prices are increasing. Conform expectation, changes in the United States unemployment rate are positively related to loan loss provisions across all four models, indicating that managers increase loan loss provisions when unemployment is rising.

For the random forest regression, the relative importance of each of the predictors can be determined using the *VarImp* function in R. See Figure 1. Similar to previous research, we find that net charge offs are by far the most important predictor for our models.

Figure 1. Variable importance plot for random forest regression model 4.



Note. This figure shows the relative importance of the variables used in model 4. *CO* is the net charge off over the last quarter scaled by lagged total loans. *ALW* is the change in loan loss allowance over the quarter divided by lagged total loans. *NPA* is the change in nonperforming assets over the quarter scaled by lagged total loans. *SIZE* is the log of total assets. *GDP, CSRET* and *UNEMP* are respectively the percental changes in GDP, the housing market and the unemployment rate over the quarter. *CO* is the most important variable for the model. *IncNodePurity* uses the Gini impurity index used to determine the cut-off point in trees to calculate relative variable importance. Variables of greater importance have a higher value.

Table 3. Linear regression of models 1 to 4 on scaled loan loss provision.

	Dependent variable:			
	LLP_t			
	(1)	(2)	(3)	(4)
LOAN	-0.001***	0.000**	0.001***	0.001***
	(0.000)	(0.0002)	(0.0002)	(0.000)
NPA_{t+1}	-0.022***	-0.000	0.014***	0.013***
	(0.003)	(0.003)	(0.002)	(0.002)
NPA_t	-0.010***	0.009**	0.034***	0.033***
	(0.003)	(0.003)	(0.002)	(0.002)
NPA_{t-1}	0.005***	0.022***	0.019***	0.018***
	(0.002)	(0.002)	(0.002)	(0.002)
NPA_{t-2}	0.017***	0.029**	0.017***	0.016***
	(0.002)	(0.002)	(0.002)	(0.002)
SIZE	0.000**	0.000***	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
CO			0.631***	0.646***
			(0.006)	(0.007)
ALW		0.065***		0.008***
		(0.002)		(0.002)
CSRET	-0.006***	-0.005***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.0003)	(0.000)
UNEMP	0.014***	0.015***	0.013***	0.013***
	(0.001)	(0.001)	(0.001)	(0.001)
GDP	-0.005***	-0.006***	-0.001***	-0.000***
	(0.001)	(0.001)	(0.000)	(0.000)
intercept	0.001***	-0.0002	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	15,218	15,218	15,218	15,218
R^2	0.067	0.156	0.472	0.472
Adjusted R^2	0.067	0.155	0.471	0.472
F Statistic	122*** (df = 9; 15208)	281*** (df = 10; 15207)	1358*** (df = 10; 15207)	1237*** (df = 11; 15206)

Note: *p<0.1;**p<0.05;p***<0.01. Model estimation of models 1 to 4 based on 75% training data.

LLP is loan loss provision scaled by lagged total loans. *NPA* is the change in non-performing assets divided by lagged total loans. *SIZE* is the log of total assets. *LOAN* represents the change in total loans over the quarter divided by lagged total loans. *CO* refers to the net charge off divided by lagged total loans. *ALW* is the loan loss allowance divided by lagged total loans. *CSRET* is the Case-Shiller real estate index return over the quarter. *UNEMP* is the absolute change in unemployment percentage over the quarter. *GDP* is the growth rate of the gross domestic product over the quarter.

5.3 Percentage of variance explained

In line with Dechow et al. (2003), Pae (2005) and Glen (2015) we start with a comparison of the R^2 of the predicted accruals on actual accruals. Table 4 summarizes the results from the regressions and shows that predictions from the random forest regressions outperform the traditional linear model for all tested variable combinations 1 to 4.

Table 4. R^2 for linear and random forest models.

	R^2	
Model	*Linear model*	*Random forest regression*
1)	0.055	0.275
2)	0.150	0.320
3)	0.482	0.561
4)	0.482	0.567

Note. Variance of the observed scaled loan loss provision explained by the predicated scaled loan loss provision.

Similar to Beatty & Liao (2014), we observe a significant jump in R^2 when comparing models 1 and 2 with models 3 and 4. This jump is attributed to the inclusion of the net charge-off variable and occurs both for the values predicted by the linear model and the random forest regression, though the latter experiences a smaller jump. The increase in performance after the inclusion of the net charge-off variable is also consistent with Figure 1 and Table 3, both of which identified variable *CO* as the most impactful variable. The results from Table 4 indicate higher explanatory performance for estimates originating from the random forest regression.

5.4 Mean (Absolute) Error

The mean absolute error is the second accuracy assessment discussed in this paper, in line with Pae (2005). Table 5 reports the mean absolute error of all four linear and random forest models. The errors are lower for the random forest regression across all models.

Table 5. Mean Absolute error for linear and random forest regression.

	Mean Absolute Error	
Model	*Linear model*	*Random Forest regression*
1)	88.55	74.37
2)	83.03	70.82
3)	62.36	56.23
4)	62.43	55.66

Note. Mean absolute error of the estimation of loan loss provision scaled by total loans. All values have been multiplied by 10^5 for readability.

The results from Table 5 indicate that out-of-sample predictions made by the random forest algorithm are on average more accurate than predictions from its linear counterpart. Together with the improved R^2 among all models, the random forest algorithm makes more accurate forecasts than their linear counterparts.

Evaluation of discretionary accrual estimation models suffers from the issue that actual discretionary accruals are unobservable. The reservable nature of accruals implies that discretionary accruals should, over a large enough sample period, be equal to zero. The closer the mean residual of the model estimate is to zero, the higher the performance. Table 6 indicates that the mean residuals of the linear models are smaller than the residuals of the tree-based models, implicating better performance. Similar to previous results, models 1 and 2 are outperformed by models 3 and 4. Assuming that discretionary accruals are on average equal to zero during the sample period, these results further suggests that all models on average overpredict the nondiscretionary component of the loan loss provision. The inferences drawn from Table 4 and Table 5 are robust to using a balanced data sample, whereas the results from Table 6 show better performance for the random forest regression when using a balanced panel.

Table 6. Mean error for linear and random forest regression.

	Mean Error	
Model	*Linear model*	*Random Forest regression*
1)	-4.651	-5.540
2)	-5.648	-5.592
3)	-2.464	-2.914
4)	-2.301	-2.931

Note. Mean absolute error of the estimation of loan loss provision scaled by total loans. All values have been multiplied by 10^5 for readability.

5.5 Persistence analysis

Results from the linear regression of future loan loss provision on current proxies for discretionary and non-discretionary accruals are shown in Table 7. The results indicate that the estimated coefficient of the discretionary loan loss provision proxy is smaller than the coefficient of the nondiscretionary component proxy for all linear and tree-based models. The effects described by the discretionary component are on average more transitory, which is in line with expectations. An F-test shows that all coefficients for the discretionary and nondiscretionary component are significantly different from each other. Predictions made by the random forest model 4 produce the lowest coefficient for the discretionary component, indicating that this model produces a proxy with the least persistence. Of all models, random forest-based models separate total loan loss provision in discretionary and non-discretionary components that most strongly behave as the true, unobservable, values ought to. Table 7 illustrates that the discretionary component has a smaller influence on the loan loss provision in the next quarter compared to the nondiscretionary component. Unpublished results using a balanced data sample show that linear model 3 produces a more transient proxy than random forest model

3, though the remaining random forest models outperform their linear counterparts when using balanced data.

Table 7. Persistence analysis of discretionary and non-discretionary component.

	Dependent variable: LLP_{t+1}							
	Linear model				*Random Forest regression*			
	1)	2)	3)	4)	1)	2)	3)	4)
$NDLLP_t$	0.986^{***}	0.914^{***}	0.689^{***}	0.684^{***}	0.807^{***}	0.828^{***}	0.697^{***}	0.703^{***}
	(0.043)	(0.027)	(0.016)	(0.016)	(0.020)	(0.019)	(0.016)	(0.016)
$DLLP_t$	0.548^{***}	0.510^{***}	0.468^{***}	0.473^{***}	0.480^{***}	0.453^{***}	0.431^{***}	0.426^{***}
	(0.012)	(0.012)	(0.016)	(0.016)	(0.013)	(0.014)	(0.017)	(0.017)
Intercept	-0.000^{***}	0.000^{***}	0.000^{***}	0.000^{***}	0.000^{***}	0.000^{***}	0.000^{***}	0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	5073	5073	5073	5073	5073	5073	5073	5073
R^2	0.346	0.357	0.345	0.344	0.356	0.365	0.349	0.350
Adjusted R^2	0.345	0.356	0.345	0.344	0.356	0.364	0.349	0.350
F Statistic (df =1)	2208^{***}	1751^{***}	883^{***}	898^{***}	1323^{***}	1118^{***}	638^{***}	615^{***}
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $p^{***}<0.01$. Analysis of coefficients of the discretionary and non-discretionary component of linear and random forest models 1 to 4. F-tests rejects the null-hypothesis that both coefficients are equal to each other.							

5.6 Artificially induced earnings management

The final test evaluates the ability of the models to identify observations with abnormal loan loss provisions. The values in Table 8 are based on linear model 4 and random forest model 4 because these models performed comparable or better than the other three models in the previous tests. The model that most correctly predicts the nondiscretionary component of the loan loss provision should have the highest concentration of managed earnings in its top deciles.

Table 8. Concentration of observations with artificially managed earnings in deciles for model 4.

LLP modification	Decile										
Random Forest	1	2	3	4	5	6	7	8	9	10	p-value
0.0000	4.7	5.2	5.2	5.9	5.1	5.1	5.7	4.0	4.3	4.8	1.000
0.0250	6.1	4.1	5.8	7.1	7.3	6.1	4.4	3.2	2.6	3.3	0.815
0.0375	5.4	4.8	9.4	8.8	7.2	4.1	3.6	3.7	1.8	1.2	0.133
0.0500	6.2	8.1	12.1	9.1	4.2	3.9	1.6	2.0	2.0	0.8	0.002
0.0750	7.1	15.9	12.2	4.9	3.2	2.2	0.8	1.6	0.9	1.2	0.000
0.1000	11.1	23.7	7.9	1.9	1.4	0.9	1.2	0.9	0.4	0.6	0.000
0.1500	30.1	14.9	2.4	1.8	0.1	0.1	0.2	0.3	0.1	0.0	0.000
Linear											
0.0000	4.8	4.2	5.5	4.7	5.6	5.4	5.4	4.1	4.9	5.4	1.000
0.0250	4.9	4.9	7.1	6.1	5.5	3.8	4.2	3.3	5.2	5.0	0.988
0.0375	5.8	5.1	8.9	7.3	5.5	3.7	3.8	3.7	2.6	3.6	0.654
0.0500	5.6	7.9	9.5	6.3	4.0	5.7	2.4	3.4	2.5	2.7	0.298
0.0750	6.7	14.6	9.1	7.3	3.2	2.5	2.3	1.0	1.8	1.5	0.000
0.1000	9.8	20.9	9.1	3.8	2.2	1.3	1.0	0.6	0.6	0.7	0.000
0.1500	24.6	19.4	3.0	1.3	0.6	0.3	0.1	0.1	0.6	0.0	0.000

Note: Concentration of observations with artificially managed earnings per decile. Results depicted are the average number of observations per decile for ten random samples of modified observations. P-value indicates the significance level for rejecting the null-hypothesis of similar true and false distribution among deciles for variable *EM* based on Chi-square statistic with 9 degrees of freedom. Total number of modified earnings in the testing sample was 50, over 5075 observations. Deciles ranked from highest to lowest estimated absolute discretionary accrual. *LLP* modification shows the adjustment to loan loss provision as a percentage of total loans.

The results indicate that both models fail to recognize artificially induced earnings management at 0.025% and 0.050% of total loans, finding no significant difference from random allocation at conventional significance levels. For earnings management higher than 0.0375%, the random forest regression starts to assign significantly more observations with managed earnings to the higher deciles at a significance level of 5% while the linear model does not yet reach the significance level of 10%. Any adjustments above this level are significant at the 1% level for both models. For the entire sample, the mean absolute value for the loan loss provision is approximately 0.11% of total loans, indicating that a material deviation from the average loan loss provision is necessary for both models to reliably identify cases of earnings management. These results are robust to using the median instead of the mean. P-values from the Chi-square test provide strong evidence against the null hypothesis that the distribution of managed earnings is the same across all accrual portfolios for higher levels of earnings management. Using a balanced data panel shows that a loan loss modification of 0.0375% is significant at the 10% level for the random forest model, whereas the Chi-square tests rejects the null hypothesis of equal distribution for a modification of 0.05% at the 5% level of significance for the linear model.

6 Conclusion

This study investigates whether tree-based models can outperform traditional linear models in separating loan loss provisions in its discretionary and nondiscretionary component. Little is published on the application of machine learning algorithms in accounting studies in general, and even less when looking specifically at earnings management studies. Proper discretionary accrual models are of fundamental importance to academics and regulators investigating earnings management in the banking sector. These models serve as a fundamental tool for estimating the proxy for earnings management.

This study contributes to existing literature by providing strong evidence that random forest regressions can outperform linear regressions when prediction of the nondiscretionary component of loan loss provisions is required. Common drawbacks of more complex machine learning algorithms are the loss of interpretability traded for higher predictive power. As interpretation of the first stage of an accrual model is often of lesser importance to the researcher, we argue that this is a favourable trade-off. This paper compared four well-substantiated linear models to their random forest counterparts and found promising results. As the purpose of this paper is to compare the models based on similar data input, variables were scaled and winsorized according to linear requirements. These modifications are not necessarily required for random forest are more robust to issues that plague linear models, such as heteroskedasticity, correlation, unscaled variables or the assumption of linearity.

This study employs various tests to assess comparative performance. Three tests focus on the accuracy of the predictive models. First, the actual accruals were regressed on predicted accruals to determine the R^2. The second and third test compared the absolute and mean errors of the linear and random forest predictions. The random forest regression outperformed the linear model in two out of three tests, achieving the highest R^2 and lowest mean absolute error. The mean error of linear model 4 came closest to zero, which is the expected value when assuming full reversion of discretionary accruals within the sample period. Using a balanced dataset results in better performance for the random forest models in all three tests compared to the linear models. The final two tests focus on testing expected discretionary behaviour within the created proxies. The first of these tests focusses on the persistence of the discretionary component of the loan loss provision. Of all models, the proxy provided by random forest model 4 proves to be the most transient. Comparable results were found for the balanced dataset, with the exception that linear model 3 outperformed random forest model 3 in this subsample. The final test analyses the ability of both methods to recognize discretionary accruals in observations where earnings management has occurred. Due to limitations in the availability of real earnings management data, we artificially induce earnings management in the sample ranging from 0% to 0.15% of total loans. The random forest model assigns significantly more managed observations to the higher deciles when the loan loss provision is increased with 0.05% of total loans, compared to 0.075% for the linear model at a significance level of 1%. Taken together, this study provides evidence that using random forests can improve discretionary accrual estimation and assist academics and regulators in the future. Random forest models are able to more accurately predict loan loss provision values and create proxies that exhibit stronger discretional behaviour than their linear counterparts.

This paper has several limitations. First of all, no distinction is made between different type of loans. Including different categories of loans may improve the predictive power of the model as asset backed loans may prove less risky and therefore require a lower provision. Next, this paper has made several assumptions regarding the timing of reversal of the discretionary component of loan loss provisions. These assumptions may not hold in reality, as managers may reverse accruals slowly over time instead of instantly reversing in the next period. Additionally, this paper also assumes full reversal of all discretionary accruals within the sample period. Future research could improve on this study by considering a broader range of variables, including loan type, consumer confidence or interest rates. Future research could also extend by focussing on different regions, testing different reversal periods or by using different machine learning algorithms such as neural networks.

7 Bibliography

Alali, F., & Jaggi, B. (2011). Earnings versus capital ratios management: role of bank types and SFAS 114. *Review of Quantitative Finance and Accounting*, 105-132.

Balboa, M., Lopez-Espinosa, G., & Rubia, A. (2013). Nonlinear dynamics in discretionary accruals: An analysis of bank loan-loss provisions. *Journal of Banking & Finance*, 5186-5207.

Bartov, E. (1993). The Timing of Asset Sales and Earnings Manipulation. *The Accounting Review*, 840-855.

Basu, S., Vitanza, J., & Wang, W. (2020). Asymmetric loan loss provision models. *Journal of Accounting and Economics, 70*(2-3), 1-21.

Beatty, A., & Liao, S. (2014). Financial accounting in the banking industry: a review of the empirical literature. *Journal of Accounting Economics*, 339-383.

Beaver, W. (1987). The Properties of Sequential Regressions with Multiple Explanatory Variables. *The Accounting Review*, 137-144.

Beaver, W., & Engel, E. (1996). Discretionary behavior with respect to allowances for loan losses and the behavior of security prices. *Journal of Accounting and Economics*, 177-206.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

Chen, W., Hribar, P., & Melessa, S. (2018). Incorrect Inferences When Using Residuals as Dependent Variables. *Journal of Accounting Research*, 751-796.

Cheng, Q., Warfield, T., & Ye, M. (2011). Equity incentives and Earnings Management: Evidence from the Banking Industry. *Journal of Accounting, Auditing & Finance*, 317-349.

Cornett, M. M., McNutt, J. J., & Tehranian, H. (2009). Corporate governance and earnings management at large U.S. bank holding companies. *Journal of Corporate Finance*, 412-430.

Dechow, P. M. (1994). Accounting earnings and cash flows as measures of firm performance: The role of accounting accruals. *Journal of Accounting and Economics*, 3-42.

Dechow, P. M., Hutton, A. P., Kim, J., & Sloan, R. G. (2012). Detecting Earnings Management: A New Approach. *Journal of Accounting Research*, 275-334.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting Earnings Management. *The Accounting Review*, 193-225.

Dechow, P., Richardson, S., & Tuna, I. (2003). Why Are Earnings Kinky? An Examination of the Earnings management Explanation. *Review of Accounting Studies*, 355-384.

Dopuch, N., Mashruwala, R., Seethamraju, C., & Zach, z. (2012). The Impact of a Heterogeneous Accrual Generating Process on Empirical Accrual Models. *Journal of Accounting, Auditing & Finance*, 386-411.

Erickson, M., & Wang, S.-w. (1999). Earnings management by acquiring firms in stock for stock mergers. *Journal of Accounting and Economics*, 149-176.

Ertan, A. (2021). Real earnings management through syndicated lending. *Review of Accounting studies*, 1-42.

Federal Reserve Economic Data. (2022). Case-Shiller U.S. National Home Price Index. Retrieved June 04, 2022, from https://fred.stlouisfed.org/series/CSUSHPINSA

Glen, H. (2015). Predicting Loan Loss Provisions by Including Loan Type Characteristics. *International Journal of Business and Finance Research*, 53-67.

Gomes, C. M., & Jelihovschi, E. (2020). Presenting the Regression Tree Method and its application in a large-scale educational dataset. *International Journal of Research & Method in Education*, 201-221.

Graham, J. R., Harvey, C. R., & Rajgopa, S. (2005). The economic implications of corporate financial reporting. *Journal of Accounting and Economics*, 3-73.

Grougiou, V., Leventis, S., Dedoulis, E., & Owusu-Ansah, S. (2014). Corporate social responsibility and earnings management in U.S. banks. *Accounting Forum*, 155-169.

Jaiantilal, A. (2013). *Feature Selection by Iterative Reweighting: An Exploration of Algorithms for Linear Models and Random Forests.*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning.*

Jones, J. J. (1991). Earnings Management During Import Relief Investigations. *Journal of Accounting Reserach*, 193-228.

Kanagaretnam, K., Krishnan, G., & Lobo, G. (2009). Is the market valuation of banks' loan loss provision conditional on auditor reputation? *Journal of Banking & Finance*, 1039-1047.

Kanagaretnam, K., Lobo, G., & Yang, D.-H. (2004). Joint Tests of Signaling and income Smoothing. *Contemporary Accounting Research*, 843-884.

Kasznik, R. (1999). On the Association between Voluntary Disclosure and Earnings Management. *Journal of Accounting Research*, 57-81.

Liu, C.-C., & Ryan, S. (2006). Income smoothing over the Business Cycle: Changes in Banks' Coordinated Management of Provisions for Loan Losses and Loan Charge-Offs from the Pre-1990 to the 1990s Boom. *The Accounting Review*, 421-441.

Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. *Journal of Accounting and Ecoonmics*, 1-25.

Lobo, G. J., & Yang, D.-H. (2001). Bank Managers' Heterogeneous Decisions on Discretionary Loan Loss Provisions. *Review of Quantitative Finance and Accounting*, 223-250.

McNichols, M. F., & Stubben, S. (2018). Research Design Issues in Studies Using Discretionary Accruals. *Abacus*, 227-246.

Medeiros, R., Dantas, A., & Lustosa, P. (2012). An Extended Model for Estimating Discretionary Loan Loss Provisions in Brazilian Banks.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? *Springer-Verlag Berlin Heidelberg 2012*, 154-168.

Ozili, P., & Outa, E. (2017). Bank loan loss provisions: A review. *Borsa Instanbul Review*, 144-163.

Pae, J. (2005). Expected Accrual Models: The Impact of Operating Cash Flows and Reversals of Accruals. *Review of Quantitative Finance and Accounting*, 5-22.

Probst, P., & Boulesteix, A.-L. (2018). To Tune or Not to Tune the Number of Trees in Random Forest. *Journal of Machine Learning Research*, 1-18.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2018). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(3), 1-15.

Roychowdhury, S. (2006). Earnings management through real activities manipulation. *Journal of Accounting and Economics*, 335-370.

Sood, H. (2012). Loan loss provisioning and income smoothing in US banks pre and post. *International Review of Financial Analysis*, 64-72.

Stubben, S. (2010). Discretionary Revenues as a Measre of Earnings Management. *The Accounting Review*, 695-717.

Teoh, S., Welch, I., & Wong, T. (1998). Earnings Management and the Long-Run Market Performance of Initial Public Offerings. *The Journal of Finance*, 1935-1974.

Tran, D., & Houston, R. (2021). The effects of policy uncertainty on bank loan loss provisions. *Economic Modelling*, 2-15.

U.S. BUREAU OF LABOR STATISTICS. (2022). Labor Force Statistics from the Current Population Survey. Retrieved June 04, 2022, from https://beta.bls.gov/dataViewer/view/timeseries/LNS14000000

World Bank. (2022). GDP (current US$) - United States. Retrieved June 04, 2022, from https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=US

Wu, R.-S. (2014). Predicting earnings management: A nonlinear approach. *International Review of Economics and Finance*, 1-25.

Zhou, Z.-H. (2016). *Machine Learning.* Tsinghua University Press.

8 Appendix

Pearson correlation values

Variables	LOAN	NPAt+1	NPA	NPAt-1	NPAt-2	SIZEt-1	CO	ALW	CSRET	UNEMP	GDP
LLP	-0,024	-0,063	-0,026	0,009	0,040	0,025	0,647	0,294	-0,188	0,118	-0,072
LOAN		0,112	0,171	0,051	0,024	0,038	-0,180	-0,195	0,000	0,201	0,067
NPAt+1			0,092	0,152	0,103	0,079	-0,194	-0,274	0,035	0,022	-0,008
NPA				0,088	0,148	0,075	-0,225	-0,259	0,022	0,026	0,015
NPAt-1					0,082	0,067	-0,099	-0,227	0,018	0,012	0,016
NPAt-2						0,056	-0,058	-0,195	0,017	0,003	0,012
SIZEt-1							0,029	-0,079	0,095	0,012	0,034
CO								0,563	-0,180	-0,009	-0,029
ALW									-0,111	-0,032	0,019
CSRET										-0,021	0,308
UNEMP											0,390
GDP											

Note. Pearson correlation values between model variables. *LLP* is loan loss provision scaled by lagged total loans. *NPA* is the change in non-performing assets divided by lagged total loans. *SIZE* is the log of total assets. *LOAN* represents the change in total loans over the quarter divided by lagged total loans. *CO* refers to the net charge off divided by lagged total loans. *ALW* is the loan loss allowance divided by lagged total loans. *CSRET* is the Case-Shiller real estate index return over the quarter. *UNEMP* is the absolute change in unemployment percentage over the quarter. *GDP* is the growth rate of the gross domestic product over the quarter.