Predicting Profitability Using Machine Learning

Vic Anand
University of Illinois at Urbana-Champaign

Robert Brunner
University of Illinois at Urbana-Champaign

Kelechi Ikegwu
University of Illinois at Urbana-Champaign

Theodore Sougiannis
University of Illinois at Urbana-Champaign

10/28/2019

We thank Laura Li, Matt Lyle, Zach Wang, Wei Zhu, and participants of the archival brownbag at the University of Illinois at Urbana-Champaign for helpful comments and suggestions. We acknowledge gracious financial support from the Gies College of Business, its Department of Accountancy, and the University of Illinois-Deloitte Foundation Center for Business Analytics.

Predicting Profitability Using Machine Learning

Out-of-sample prediction of profitability is a critical step in fundamental analysis and yet even sophisticated regression models do not generate predictions that significantly outperform random walk predictions. We employ random forests with classification trees, a method from machine learning, to generate out-of-sample predictions of directional changes (increases or decreases) in five profitability measures, return on equity (ROE), return on assets (ROA), return on net operating assets (RNOA), cash flow from operations (CFO), and free cash flow (FCF). With a minimum set of independent variables, and out-of-sample, our method achieves classification accuracies ranging from 57 – 64% for our profitability measures, compared to 50% for the random walk. The difference in proportions of accurate classifications is highly significant. Out-of-sample classification accuracy is similar over forecast horizons of 1 to 5 years. We observe better performance on cash flow measures than on traditional, earnings-based profitability measures. Also, accruals show strong incremental ability beyond cash flows in predicting future cash flows. We find predictive accuracy is highest for firms with high and low accruals-to-market and earnings-to-market ratios, exceeding 75% in one instance. Importantly, our method is insensitive to outliers; our method used data that had not been winsorized or standardized. These results suggest that machine learning methods offer better predictive performance than traditional regression-based methods.

1 Introduction

Fundamental analysis requires three steps for its implementation. The first and most critical step is accurate prediction of a firm's future profitability over a forecast horizon. Without accurate forecasts, the estimation of an intrinsic value in the second step will be inaccurate and the wrong investment decision is likely to be made in the third step, where market value is compared to intrinsic value. In addition to accurate forecasts, fundamental analysis requires out-of-sample forecasts of profitability because they are the forecasts investors need in practice.

A long line of empirical research, reviewed by Monahan (2017), has provided evidence that out-of-sample profitability forecasts generated by various regression models are not more accurate than forecasts generated by a random walk model. Furthermore, a recent study by Bradshaw et al. (2012) provides evidence that even forecasts of profitability generated by financial analysts are not more accurate than forecasts generated by a random walk model. Monahan (2017) states that the strong performance of the random-walk model, relative to extant approaches, is problematic from a theoretical and practical perspective. We view the failures of traditional methods to beat a random walk as a motivation for the application of new available methods.

In this study, we apply a method from machine learning (ML), classification trees, to the problem of generating out-of-sample profitability forecasts. Methods from ML are likely better suited to this task since, historically, ML focused on prediction and maximizing prediction accuracy (Mullainathan and Spiess 2017). While the term ML is broad, many methods in ML have been better able to handle econometric problems in the data, such as multicollinearity and nonlinearity, than traditional, regression-based methods. Also, many ML methods do not require the user to specify a functional form ex ante, thereby giving the methods greater flexibility to *discover* a functional form that best fits the data. The tradeoff is that many ML methods are more of a "black box" than a typical regression. Thus, a researcher who chooses ML over regression typically gains out-of-sample forecast accuracy but loses some interpretability. Given the problem at hand, *improving the accuracy of out-of-sample profitability forecasts*, we believe that is a tradeoff worth making.

The forecasting performance of regression models depends on a variety of factors such as functional form, choice of predictors, choice of estimator, and the behavior of the error term. Perhaps some of these restrictive factors cause the underperformance of regression models relative to the random walk model in out-of-sample predictions.[1] In the case of financial analysts, a large literature, reviewed by Ramnath et al. (2008), shows that analysts are systematically optimistic in their forecasts, either because of incentive-driven strategic reporting or inherent cognitive bias. However, ML methods are more flexible than regression methods because they do not rely on restrictive statistical and economic assumptions, and they are not affected by cognitive biases. Instead, they exploit patterns and trends in the historical data to make predictions. Said another way, ML methods rely less on prior assumptions and choices made by the researcher or analyst; instead, they use available data to identify patterns that provide more accurate forecasts.

Since, to our knowledge, this is the first study of profitability prediction using ML, we keep our analyses simple by making the following choices. First, we focus on profitability measures that are both theoretically motivated (Ohlson 1995; Feltham and Ohlson 1995) and widely used in practice (Penman 2012). Thus, we use three accruals-based measures, return-on-equity (ROE), return-on-assets (ROA), and return-on-net operating assets (RNOA), and two cash-based measures, cash flow from operations-to-assets (CFO), and free cash flow-to-assets (FCF). Second, for out-of-sample predictions of these measures, we rely on three well-known properties of the measures: a) persistence in profitability, b) reversion of profitability to a mean value, and c) volatility of profitability. Past values of each measure could potentially allow for the detection of a persistence pattern (Dechow et al. 2010). Industry

[1] Gu et al. (2019) reach a similar conclusion using empirical asset pricing data.

membership and past values of each measure could allow for the detection of a mean reversion pattern (Nissim and Penman 2001). The standard deviation of the historical values of each measure could allow for more accurate predictions (Dichev and Tang 2009). Third, although the prediction of the magnitude of a profitability measure can be more desirable, *we predict the sign of the future directional change in each measure (increase or decrease) and leave the prediction of magnitude for future work*. We do that for two reasons: (a) because we are able to use classifier methods from ML that require less data and are simpler to implement, and (b) the future directional change of profitability can also be a useful indicator as prior studies have shown (e.g., Ball and Brown 1968; Ou and Penman 1989b, 1989a). Fourth, we compare our ML forecasts to two alternative random walk (RW) models. One is a pure RW model that uses no prior profitability information and assumes a 50/50 increase or decrease in future profitability. Another RW model assumes that the current change in profitability (increase or decrease) will also be the future change in profitability.

We create our sample by taking the entire set of publicly-listed firms in the Compustat annual database between 1963 – 2017. We then partition the data into training and testing subsets. The training data subset includes fiscal years 1963 – 2012 (approximately 90% of the data), and the testing data includes fiscal years 2013 – 2017. We then train classification trees (aka decision trees) using the training data and evaluate their accuracy on the testing data. Of note, the classification trees do not train on the testing data; the testing data is a holdout sample. To evaluate the out-of-sample performance of a tree, we evaluate its predictive performance on each firm-year observation in the testing sample. Thus, given the information for firm f in year t, the classification tree predicts whether the profitability for firm f will increase or decrease in year $t + 1$.

Overfitting is a well-known problem in ML (Howard and Bowles 2012; Mullainathan and Spiess 2017; Varian 2014). A single classification tree is likely to over-fit the training data and perform poorly on the out-of-sample testing data. To address this problem, ML researchers have developed “ensemble” methods. We employ one such method, random forests, to reduce the likelihood of over-fitting. A random forest is a collection of classification trees. To train each tree in a forest, the method bootstraps the training data; thus, each tree in the forest trains on a different subset of the training data. The prediction of a random forest is the prediction made by a majority of its constituent trees (for example, if 75% of the trees in a forest predict a firm's ROE will decrease, the forest will predict a decrease). ML researchers have found that random forests reduce over-fitting since no single tree dominates.

We trained and evaluated multiple random forests. For each profitability measure (ROE, ROA, RNOA, CFO, FCF), we created 12 random forests, each with different "features".[2] Our research design begins with a very small set of features and progressively adds more features; we do this to ascertain the effect of incremental information on prediction accuracy. Notably, we were able to add any desired feature since multicollinearity does not pose a problem for classification trees (James et al. 2013).

Our random forests handily beat the random walk. In the first phase of our research design, we use minimal information. We predict the future (one-to-five year ahead) change in a profitability measure using only the firm identifier (PERMNO), fiscal year, and the contemporaneous value of that profitability measure (i.e. we predict the change in ROE over the next five years using only PERMNO, fiscal year, and current ROE). In this first phase, with minimal information, classification accuracy ranges from a low of 57.1% (ROE) to a high of 59.8% (RNOA).[3] This compares to 50% classification accuracy for a random walk model. The difference in proportions is highly significant ($p < 0.001$). We trained three random forests for each profitability measure. For the first, we use un-winsorized data. For the second and third, we winsorize the data at 1% and 2.5% on each side. The results were nearly identical for all three random forests and demonstrate the robustness of our chosen ML methods relative to regression.

[2] In machine learning, the term "feature" means independent variable, and the term "target" means dependent variable.

[3] Classification accuracy is defined as the proportion of firm-year observations in the testing sample (i.e. in the out-of-sample holdout data) for which we correctly predicted the change in future profitability.

We also demonstrate that accounting data does not require preprocessing to generate accurate forecasts.

In the second phase of our research design, we incrementally add features. We add the SIC code, as well as the mean, median, and standard deviation of the profitability measure for each firm's industry in each year. We also add the first two lags of the profitability measure. Surprisingly, we find little improvement in predictive power with the additional information. As we add incremental features, out-of-sample classification accuracy improves to a maximum of 59.1% for ROE, 59.6% for ROA, 60.9% for RNOA, 62.8% for CFO, and 63.7% for FCF. Notably, the cash-flow measures of profitability outperform the earnings measures. This implies that, with the information set we chose, it is easier to predict cash flows than earnings. We also examine the well-debated issue of whether accruals are informative beyond cash flows in predicting out-of-sample future cash flows and we find strong evidence that they are. Another notable finding is that the predictive performance of our random forests does not decrease in the forecast horizon. Our last training year is 2012 and our out-of-sample dataset contains years 2013 – 2017. We examine performance separately in each of the years in the testing data and find that the performance is similar. This is remarkable since it suggests that the process through which profitability measures are generated is relatively stable over time.

In the third phase of our research design, we examine the extent to which predictive accuracy varies across five conditioning variables, book-to-market ratio, size (as measured by market capitalization), debt-to-assets ratio, earnings-to-market ratio, and accruals-to-market ratio. We find that predictive power is highest for firms in the highest and lowest deciles of earnings-to-market and accruals-to-market ratios, in one case exceeding 75% accuracy. In the extreme deciles of all conditioning variables, the accruals-based measures (ROE, ROA, RNOA) often exhibit higher predictability than the cash-flow measures (CFO, FCF). These results suggest that no single profitability measure exhibits high predictability under all conditions.

This paper makes three contributions to the accounting literature. First, to our knowledge, we are the first to be able to demonstrate a method that makes out-of-sample predictions of profitability that beat a random walk. Second, we introduce a method from machine learning, classification trees.[4] And third, we provide evidence that the functional forms used in past, regression-based models may have been too restrictive and thus inhibited the forecasting ability of those models.

2 Methods

2.1 Overview of Machine Learning Methods Used

2.1.1 Comparing Machine Learning and Econometrics

This paper applies two machine learning algorithms, classification trees and random forests, to the problem of earnings forecasting. Since these methods are novel in accounting research, we describe them in some detail. Before doing so, we discuss the differences between machine learning and econometrics to provide context and background.

In econometrics, a researcher typically specifies a model to be fitted. The model is usually based on economic theory and specifies a fixed functional form that includes a dependent variable and one or more independent variables. Models are typically of the form $y = f(\beta, x)$ where y represents the dependent variable, x represents the independent variables, and β represents the parameters of the model with signs expected to be consistent with economic theory.[5] *The econometric problem is to find the best estimates of the parameters* $\hat{\beta}$ *that minimize some loss function (e.g. sum of squared residuals) given all available data* y *and* x (Mullainathan and Spiess 2017). The model is fit by using an estimation sample to estimate the parameters $\hat{\beta}$ (in-sample data) and a holdout sample to derive out-of-sample

[4] In the recent accounting literature, Ding et al. (2018) apply decision trees, a related method, to the prediction of accounting estimates. Also, Zhang (2018) uses decision trees to forecast earnings restatements and audit quality.

[5] In machine learning, the dependent variable y is referred to as the target variable. The independent variables x are referred to as “predictors” or “features” (Varian 2014).

predictions using the parameters $\hat{\beta}$ (out-of-sample data), (Poon and Granger 2003; Elliott and Timmermann 2008).

Machine learning is primarily concerned with *prediction*: produce the best predictions of y given the available data x; informally, "*machine learning belongs in the part of the toolbox marked $\hat{y}$ rather than in the more familiar $\hat{\beta}$ component*" (Mullainathan and Spiess 2017).[6] The goal of producing accurate predictions requires a different approach than that used in econometrics. Instead of assuming a functional form, machine learning (ML) methods attempt to find generalizable patterns in the available data and exploit those patterns to make accurate predictions. Thus, the output of an ML method is a function $y = f(\beta, x)$ that makes predictions of y given new data x (Varian 2014). Note that, in contrast to econometrics, ML methods estimate a function f in addition to the parameters $\hat{\beta}$. This can provide a significant advantage over econometrics because in econometric models the functional form is usually assumed to be linear, non-linearities are difficult to detect or estimate, and therefore the estimated parameters $\hat{\beta}$ often exhibit large errors that reduce the accuracy of the out-of-sample predictions.[7]

Since the objective of machine learning is to make accurate predictions, ML methods must be evaluated differently than econometric methods. The latter are commonly evaluated using metrics that are calculated using in-sample tests (e.g. R^2, p-values) and out-of-sample tests (e.g. bias, accuracy). However, creators and practitioners of ML methods have found that high in-sample performance often implies overfitting and poor out-of-sample performance (Mullainathan and Spiess 2017).[8] This is

[6] In this paper, we use the term "machine learning" to refer to "supervised" learning, wherein an algorithm is trained on data with known outcomes. By contrast, unsupervised learning attempts to uncover patterns in the data without knowing the outcomes.

[7] Gu et al. (2019) apply multiple ML methods to the problem of empirical asset pricing and find ML provides additional predictive power over traditional methods. They trace the predictive gains to "*allowance of nonlinear predictor interactions that are missed by other methods.*"

[8] Overfitting and poor out-of-sample performance are also issues in econometrics.

unacceptable if the goal is high prediction accuracy on new data.[9] In response, architects of ML methods employ different methods than econometricians to fit their models.

As in econometrics, ML methods typically partition the data into training and testing data (in-sample and out-of-sample, respectively). The ML method never sees the testing data when developing its model. The testing data is only used to evaluate the model that has been fit using the training data. ML methods typically employ cross-validation to train a model. For example, in k-fold cross validation, which would be appropriate for cross-sectional data, the training data is randomly partitioned into k equally sized subsets (called "folds"). Multiple models are fit using the first $k-1$ subsets and validated on the k^{th} subset. This process is repeated k times, each time holding aside a different subset for validation while training on the other $k-1$ subsets. When fitting a model using any $k-1$ subsets, the method tries different values of the model's hyperparameters (see the discussion of tuning below) and keeps track of the performance of each model. After fitting k models, the method determines which set of hyperparameters had the best average performance.[10] A final model is then fit using all training data (i.e. all k folds) and using the set of hyperparameters that performed best during the training period. This final model's performance is then evaluated on the testing data that was initially set aside. The final model's performance on the testing (training) data is then reported as out-of-sample (in-sample) prediction accuracy.

Users of ML methods must make many design choices. For example, in the k-fold cross-validation example above, the user must specify k, the number of subsets into which the data is partitioned. If fitting trees, the user must specify the maximum number of levels of the tree, the maximum number of leaves, and a penalty for adding an additional level or leaf. These design choices are called

[9] Consider, for example, facial recognition software. Every time it attempts to recognize a face, it is working with new data. The lighting, camera angle, and background are different with every photo. If predictive accuracy is high with in-sample data, but low with out-of-sample data, then the software is useless.

[10] Mullainathan and Spiess (2017) describe this process as "*an out-of-sample experiment inside the original sample.*"

hyperparameters, and the process of choosing them is called *tuning* (Mullainathan and Spiess 2017). ML researchers have found a tradeoff between complexity and performance of ML methods. Too little complexity and the resulting model will be too general and perform poorly. However, too much complexity and the resulting model will over fit the in-sample data, leading to poor out-of-sample performance. Hence, users of ML methods must specify a loss function that penalizes the resulting model for complexity. For example, in the case of the classification trees that we will describe momentarily, the more levels of the tree, the greater the complexity penalty. The magnitude of the penalty is a design choice that the user must make;[11] users typically rely on intuition, or on studies of such parameter choices (Mullainathan and Spiess 2017).

Finally, a key difference between machine learning and econometrics is the amount of insight that the researcher can extract from the resulting model. In general, econometric models provide more insight. For example, the coefficient of an OLS regression can easily be interpreted as the change in y that is expected to occur given a one-unit change in x. By contrast, many ML methods provide higher out-of-sample prediction accuracy but much less insight. For example, modern facial recognition techniques do not provide easily understandable rules, such as "*look at the distance between the eyes and the curvature of the lips*". Since the goal of this study is to examine whether ML methods can improve the accuracy of profitability forecasts, we are willing to sacrifice some of the insight yielded by traditional econometric methods in exchange for higher prediction accuracy.

[11] Varian (2014) gives an example of a penalty term in LASSO (Least Absolute Shrinkage and Selection Operator) regression, which attempts to choose a subset of independent variables from an available set. As with ordinary least squares, LASSO attempts to minimize sum of squared residuals. However, to the sum of squared residuals, a penalty term $\lambda \sum_{p=1}^{P} \left| b_p \right|$ is added. Thus, every independent variable with a nonzero coefficient adds a penalty, effectively increasing the sum of squared residuals. This encourages the method to choose a value of zero for some coefficients, and small values of coefficients where possible. The parameter λ, the weight given to this penalty, is an example of a hyperparameter.

2.1.2 Classification Trees

Classification trees, also known as decision trees, are used in classification problems to predict a categorical outcome.[12] At each "node" of the tree, the data is partitioned using one of the features (independent variables). At the top of a tree is the root node. At this node, the data is partitioned using one of the independent variables (features). For example, if ROA is less than 0.05, go to the left branch, otherwise go to the right branch.[13] At the terminus of each branch is another node. If a node has incoming and outgoing branches, it is called an interior node, and the data is further partitioned at that node. If a node has no outgoing branches, it is called a leaf node. At each leaf node, a prediction is made. In this study, the prediction is either that a profitability measure will increase or decrease. We provide a sample classification tree in Figure 1.

The algorithm for growing trees works as follows. Create the root node by partitioning the entire training data set into two pieces. The algorithm determines the split by choosing a feature, and a level for that feature, that minimizes the "impurity" that results from partitioning the data.[14] While there are multiple ways of doing this (e.g. Gini criterion and entropy), the idea is that by choosing a feature and a level, the data at each branch will be more homogeneous than at the parent node. For example, say there are 1,000 firm-year observations and 50% of them exhibit an increase in ROA in the following year. If splitting on ROA < 0.02 means that at the left branch, more than 50% will exhibit an increase in ROA, and at the right branch, more than 50% will exhibit a decrease, then impurity is reduced. The algorithm then repeats this process of partitioning the data at each child node until impurity cannot be reduced at any node by further partitioning of the data. In the final tree, every observation in the training data set is assigned to exactly one leaf node.

---

[12] The closest analogs in econometrics are logit and probit regressions.

[13] As is common in machine learning, we focus exclusively on binary trees, i.e. trees in which there are two branches at each node (Varian 2014).

[14] Alternatively, the algorithm chooses a feature and level that provide the most informative split, and therefore the greatest increase in predictive power, at the node.

Classification trees tend to work well in problems in which there are nonlinearities and interactions in the data (Varian 2014). Consider a U-shaped relationship between the dependent variable and one of the independent variables. A linear regression would struggle to capture this, but a tree can easily do so by partitioning on low and high values of the independent variable.

Classification trees are prone to overfitting the data (Varian 2014). That is because, as the tree grows, the number of data points at each leaf node will decrease and overfitting is more likely to occur (in the limit, there could be a single datum per leaf node). To prevent this, a complexity cost is imposed on trees when growing them. A common measure of complexity is the number of leaf nodes (Varian 2014). When growing trees, they are typically tuned by fitting trees with different numbers of leaf nodes in each tree. Eventually, the tree with the best in-sample fit is chosen. Additionally, tree algorithms allow the user to specify hyperparameters such as the minimum number of data points at each leaf node, the maximum depth of the tree, and the maximum number of leaf nodes. Ultimately, the tree algorithm is trying to find the smallest number of samples at each leaf node (to improve performance) while minimizing overfitting.

2.1.3 Random Forests

The process of creating a single tree involves a tradeoff between greater complexity, which will incorporate more of the information in the data, and a risk of overfitting. It is possible to circumvent this tradeoff by creating many trees (a "forest") and averaging their predictions.[15] ML researchers have found that random forests (and ensemble methods in general) outperform methods that produce only a single model, such as a single classification tree (Varian 2014). Howard and Bowles (2012) claim that random forests are "*the most successful general-purpose algorithm in modern times.*"

[15] This is analogous to the Fama-MacBeth procedure that averages regression coefficients estimated in different years of the sample.

A random forest is created by growing many trees. However, individual trees are grown using only a subset of the available training data. For each tree, the method chooses a bootstrap sample of the training data (i.e. randomly chosen rows from the dataset). The method then chooses a random sample of the columns (features). The bootstrap sample with a subset of the available features is used to create an individual tree. This process is repeated many times to grow a "random" forest of trees in which each tree is different.[16] When it is time to make predictions, each tree in the forest is fed the testing data, and each tree makes a prediction. The prediction of a random forest is the majority vote of the predictions of the individual trees.

While random forests tend to outperform individual trees, they offer less insight into the relationships in the data. A user can easily view a single tree, like that shown in Figure 1, and understand the splits in the data. However, it is much more difficult to do that with thousands of trees. While modern machine learning packages allow the user to compute "feature importance," a measure of the frequency and level in the tree at which variables appear, this is a weak substitute for the insight offered by a single tree or by a regression equation, and another example of the tradeoff between insight and prediction accuracy in machine learning.

2.2 Implementing Random Forests to Predict Changes in Profitability

In this paper, we forecast *changes* in profitability for individual firms, where a change is coded as -1 for decreases and $+1$ for increases. Our target variables (i.e. dependent variables) are return on equity (ROE), return on assets (ROA), return on net operating assets (RNOA), cash flow from operations deflated by average total assets (CFO), and free cash flow deflated by average total assets (FCF); see Appendix A for variable definitions.

[16] The user must specify many hyperparameters when growing a random forest: the number of trees to grow, the size of the bootstrap sample, the number of features available to each tree, as well as the loss function for complexity of individual trees.

We use Python's Scikit-Learn package (Pedregosa et al. 2011) to create multiple random forests for each target variable. In each successive forest, we add additional features. We describe the research design and all features used in section 2.3. To train the random forests, we use the *TimeSeriesSplit* cross-validator in Python's scikit-learn machine learning package (Pedregosa et al. 2011). Since our data contains time series for firms and industries, a standard cross-validator, which shuffles the data, is inappropriate since future observations might be used to predict past values (James et al. 2013). Instead *TimeSeriesSplit* partitions the training data into n blocks of years, e.g. years 1-5, 6-10, 11-15, etc. Multiple random forests, each with different values for the hyperparameters, are trained on the first block and validated on the second block. Another set of forests is then trained on the first two blocks and validated on the third block, and so on.[17] When the iterations have completed, the algorithm is able to observe which set of hyperparameters led to the best in-sample performance. A final random forest is trained using all n blocks of years. The final step is to evaluate the performance of the final random forest on the testing subsample.

For each random forest, we vary the following hyperparameters. The number of trees can be 150, 250, 350, or 500. We vary the function used to measure the quality of a split at two levels, Gini impurity and entropy. We vary the minimum number of observations required to split an internal node at 2 and 5. We vary the minimum number of samples required to be at a leaf node at 10, 15, and 20. And finally, we vary the maximum number of features to consider when looking for the best split at two levels: all features, and the square root of the number of features.

[17] This is analogous to rolling-window regressions.

2.3 Research Design

Our research design is to grow multiple random forests per target variable. In each successive forest, we add new features. This allows us to determine whether specific features are informative and increase predictive power.

We perform the analyses in three phases. In the first phase, we test whether our method performs better with winsorized data or un-winsorized data. The winner of the first phase is then used in the remaining phases (i.e. if the method performs equally well, or better, with un-winsorized data, we will only use un-winsorized data in the remaining phases). In phase two, we create a sequence of random forests, each of which adds new features relative to the previous. The purpose of phase two is to learn the incremental value of additional variables used in fundamental analysis. Finally, in phase three, we identify the profitability measures with the strongest prediction performance when we condition on variables common in accounting and finance research.

2.3.1 Phase 1: Effect of Winsorization

The purpose of phase 1 is to test the effect of winsorization on predictive accuracy. Our prior is that winsorization will have no effect on predictive accuracy of random forests. This is because trees are robust to arbitrary scaling of the data (scikit-learn 2019). In phase 1, we use a minimal information set in which each firm-year observation consists of firm identifier (PERMNO), fiscal year, the contemporaneous value of a measure (e.g. ROE), and the change that occurs in the subsequent year (i.e. increase or decrease). We create three random forests using this design. We train the first forest on un-winsorized data and refer to this as analysis 1-1. We train the second (third) forest on data in which the measure (e.g. ROE) has been winsorized at 1% (2.5%) on each side and refer to this as analysis 1-2 (1-3).

2.3.2 Phase 2: Effect of Adding Additional Fundamental Information

In phase 2, we progressively add information about each firm's industry, as well as more of the firm's time series. The specific analyses are:

- Analysis 2-1a: Add SIC code as a feature.
- Analysis 2-1b: Add industry information. For each firm-year observation, add the industry mean, median, and standard deviation for the measure. For example, if the measure is RNOA, the firm is IBM, and the year is 2000, then add the industry mean, median, and standard deviation of RNOA for IBM's industry in 2000. Industry is determined by the firm's 4-digit SIC code.
- Analysis 2-1c: For the earnings-based measures, ROE, ROA, and RNOA, decompose the measure into accrual and cash flow components and add those as features. For the cash flow measures, CFO and FCF, we add accruals deflated by total assets as a feature in addition to CFO or FCF.
- Analyses 2-2a, 2-2b, 2-2c. Repeat analyses 2-1a, 2-1b, and 2-1c and add one lag of the measure. For example, if the measure is ROE, then for IBM in 2000, add IBM's 1999 ROE as a feature.
- Analyses 2-3a, 2-3b, 2-3c. Repeat each of the analyses above and add two lags of the measure.

2.3.3 Phase 3: Conditional Analyses

In phase 3, we examine the prediction performance of each of our five profitability measures by conditioning on levels of variables that reflect key firm characteristics. Specifically, the objective is to identify which of the five profitability measures shows the strongest prediction performance when we condition on variables such as firm size, book-to-market (B/P), earnings-to-price (E/P), accruals-to-price, and debt-to-assets. For this, we form 10 portfolios by computing deciles of one of these variables and then compare the prediction performance of our five profitability measures in each portfolio. For example, when we form 10 book-to-market portfolios, our objective is to discover which of the five profitability measures shows highest prediction accuracy in each portfolio. While we could use more than five conditioning variables, we chose the above five because they are commonly used in prior research and investment analysis (e.g., Fama and French 1992, 1995; Lakonishok et al. 1994; Sloan 1996). That is, in the literature, firms are characterized as small or large, all equity or highly leveraged, high or low E/P (or B/P), and with high or low accruals (and corresponding investment strategies are implemented). Our conditional analysis will reveal which of the five profitability measures shows highest prediction accuracy, for example, in the cases of low versus high accruals.

We do the conditional analysis by building on the best phase 2 model for each measure. Thus, if the best performance for CFO in phase 2 was analysis 2-2b, we use that set of features for CFO in phase 3. This reduces the number of combinations of analyses. We perform the conditional analyses by examining how the predictive performance of our model varies *ex post* with the level of the conditioning variable. This requires no additional analysis. After we run the main analysis, we simply analyze the performance of each profitability measure on the testing data and see if it varies with our conditioning variables.

2.4 Implementing the Random Walk as a Benchmark

The random walk is a model of a stochastic process. It can be described by the following equation (Campbell et al. 1997, 31, hereafter CLM):

$$P_t = \mu + P_{t-1} + \epsilon_t \qquad (1)$$

In (1), P_t is the value of some state variable at time t, μ is a drift term that we will assume is 0, and ϵ_t is a disturbance, or increment. CLM note that a common assumption is that the disturbance term is independent and identically distributed, with mean zero and variance σ^2, i.e. $\epsilon_t \sim IID(0, \sigma^2)$.

The random walk is a mathematical description of a process, *not a forecasting tool.* However, some papers (e.g., Bradshaw et al. 2012) that wish to use the random walk as a benchmark for their results treat it as a forecasting tool by taking the expected value of both sides of equation (1). That yields the following:

$$E[P_{t+1}] = P_t \Leftrightarrow E[P_{t+1} - P_t] = 0 \qquad (2)$$

These papers then reason that, since the expected value of P_{t+1} is P_t, the best forecast of P_{t+1} is also P_t. However, *the random walk does not actually predict* P_{t+1}. It simply says that the *expected* magnitude of change in P at any time is zero. In other words, if you observe changes in P over a long time series, the average change will be zero.

To be consistent with prior literature, we treat the random walk model as a forecasting tool. However, our work introduces a complication. Prior literature forecasts the *magnitude* of some variable. By contrast, we forecast the *direction* of change in our variables. Applying the random walk to our setting therefore requires some additional work. CLM notes that the expected value of ϵ_t is zero, but without additional assumptions, we cannot know the proportion of increases and decreases in ϵ_t.

2.4.1 Random Walk: Benchmark 1

Rearranging equation (1) and setting the drift term to zero yields $P_t - P_{t-1} = \epsilon_t$. Thus, the change in profitability is exactly equal to the disturbance term. That implies that the sign of the change in profitability is equal to the sign of the disturbance term. By assumption, $\epsilon_t \sim IID(0, \sigma^2)$, and CLM (p.32) state that "*[p]erhaps the most common distributional assumption for the innovations or increments ϵ_t is normality.*" Normality implies a 50-50 chance of an up or down movement, since the normal distribution is symmetric. Therefore, *our first random walk prediction should simply be that profitability is equally likely to increase or decrease.*

Under the assumption that the error term is equally likely to increase or decrease, we find that the random walk model has an expected accuracy of 50%. To see this, assume the testing subsample contains N observations and $0 \leq p \leq 1$ is the fraction of increases. Then pN of those observations are increases, and $(1-p)N$ are decreases. For any observation, the random walk will predict increase 50% of the time and decrease 50% of the time. Thus, of the pN increases, on average $0.5pN$ will be classified accurately. Similarly, of the $(1-p)N$ decreases, $0.5(1-p)N$ will be classified accurately. For the overall sample, the fraction that will be accurately classified is:

$$\text{Accuracy} = \frac{0.5pN + 0.5(1-p)N}{pN + (1-p)N} = \frac{1}{2}$$

Note that the classification accuracy of this interpretation of the random walk is invariant to the fraction of increases in the data. For any value of p, if we assume increases and decreases are equally likely, this random walk model will accurately predict the outcome 50% of the time, on average.

2.4.2 Random Walk: Non-normal Error Term

The random walk model requires that the error term has zero mean and finite variance, $\epsilon_t \sim IID(0, \sigma^2)$. This restriction does not imply normality of the error term: ϵ_t could have any arbitrary probability of increase as long as its expected value is zero. For example, if $\frac{2}{3}$ of the time ϵ_t increases by 1, and $\frac{1}{3}$ of the time it decreases by 2, it has mean zero. A distributional assumption other than normality may be appropriate if the actual fraction of increases in the data is not 50%.

In Appendix B, we show that if the fraction of increases in ϵ_t is within the range of 40 – 60%, then the classification accuracy of a non-normal random walk model remains close to 50%. Table 1, panel C shows that, for most measures in most years, the actual fraction of increases is very close to 50%. Even in the most extreme case (in 2017, ROA increases 59.3% of the time), we show in Appendix B that even if we assume ϵ_t will increase 59.3% of the time, classification accuracy will only improve to 51.6%. Thus, random walk model 1 (above) is a reasonable choice.

2.4.3 Random Walk: Benchmark 2

As an alternative interpretation of applying the random walk model to forecasting, consider the following equations:

$$P_t = P_{t-1} + \epsilon_t \tag{3}$$
$$P_{t-1} = P_{t-2} + \epsilon_{t-1} \tag{4}$$

Subtracting equation 4 from equation 3 yields:

$$(P_t - P_{t-1}) = (P_{t-1} - P_{t-2}) + (\epsilon_t - \epsilon_{t-1})$$
$$\Delta P_t = \Delta P_{t-1} + (\epsilon_t - \epsilon_{t-1}) \tag{5}$$

If we take the expected value of both sides, equation (5) reduces to $\Delta P_t = \Delta P_{t-1}$ since, by assumption, $E[\epsilon_t] = 0, \forall t$. Under this interpretation, in expectation an increase in a profitability measure in period $t - 1$ will *persist* into period t. Thus, *our second random walk prediction is that the sign of the change in a measure in one period will persist into the subsequent period.*

3 Sample Selection and Descriptive Statistics

We follow the sample selection process of Hou et al. (2012). We begin with all rows from the Compustat fundamentals annual file with fiscal years 1963 – 2018. We merge this file with the CRSP monthly returns file, including all securities listed on the NYSE, Amex, and Nasdaq with share codes 10 or 11.[18] The merge yields 231,535 firm-year observations. We drop all rows for which total assets, common equity, dividends, income before extraordinary items, or accruals are missing (variables are defined below); this results in 194,172 firm-year observations.[19] Since our machine-learning methods require a minimum of three years of data per firm, and many of our variables are computed with a lagged value, we exclude firms with fewer than 4 years of data. Our final sample contains 166,925 firm-year observations. This sample consists of 13,954 unique firms spanning 1964 - 2017, with a mean (median) of 13.9 (11) years per firm.

We retrieve and compute the following variables for use in our analyses (see Appendix A for details). Earnings is income before extraordinary items. Return on common equity (ROE) is computed as earnings divided by average total common equity. Return on assets (ROA) is computed as earnings plus interest expense divided by average total assets. We follow Li et al. (2014) and compute return on net operating assets (RNOA) as operating income divided by average net operating assets. Cash flow from operations (CFO) is scaled by average total assets. We follow Lev et al. (2009) and compute free cash flow (FCF) as cash flow from operations minus capital expenditures; we scale FCF by average total assets. Finally, we follow Hou et al. (2012) to compute accruals: "*Prior to 1988, accruals are calculated using the balance sheet method as the change in non-cash current assets less the change in current liabilities excluding the change in short-term debt and the change in taxes payable minus depreciation*

[18] Share codes 10 and 11 are ordinary common shares of U.S. companies, and exclude ADR's, closed-end funds, and REIT's.

[19] Below, we describe our procedure for computing accruals. If any of the Compustat variables needed to compute accruals are missing, we drop the observation.

and amortization expense. Starting in 1988, we use the cash flow statement method to calculate accruals as the difference between earnings and cash flows from operations."

Table 1 provides descriptive statistics on variables retrieved from Compustat, and on variables derived from those. Panel A provides summary statistics on the raw (not winsorized, not truncated) Compustat variables and on the intermediate variables for accruals (ACC), net operating assets (NOA), and operating income (OI) we compute from those. Large extreme values are present in most variables and all of them have positively skewed distributions (the mean is greater than the median). Panel B provides summary statistics on our target variables, ROE, ROA, RNOA, CFO, and FCF. Since the raw variables are used to construct them, their values of the targets are also extreme. However, unlike the raw variables, the distributions of the targets are negatively skewed (the mean is smaller than the median). Panel C shows the percent of times the target variables increase in the five-year holdout period 2013 – 2017, again using the raw data. For each firm, we compute the year-over-year change in each target variable; we code increases as +1, decreases as -1, and no change as 0. Increases in all target variables exceed 50% in 2017 reflecting the stronger economic environment relative to the earlier years. Over all five years, the last line of panel C shows that the target variables increase approximately as often as they decrease. For all target variables, this measure is close to 50%, and ranges from 49.3% (RNOA) to 50.9% (ROA).

Panel D of Table 1 shows Pearson correlations between variables that will serve as features in the random forests. Since correlations of raw variables are extreme due to large outliers, we truncate each variable at 2.5% on each side of its distribution. Of note, many of the feature variables are correlated, e.g. in the matrix for ROE, the correlation between ROE-Ind. Median and ROE is 77%. However, we do not expect that to cause problems because our ML method is not sensitive to collinearity. The correlation matrices for CFO and FCF show the well-known negative correlation between these cash flow measures and accruals. Panel E shows the correlations among the (truncated) target variables.

Unsurprisingly, ROE and ROA are highly correlated (77%), as is ROA and RNOA (70%). However, the earnings measures are not as highly correlated with the cash flow measures, e.g. correlation of ROE and FCF is 22%. Finally, panel F shows the first- and second-order autocorrelations in the (truncated) target variables. The autocorrelations are high and range from 0.32 (RNOA, 2^{nd} order) to 0.68 (ROA, 1^{st} order). The second-order autocorrelations are consistently smaller than the first-order autocorrelations but still large especially for ROA (55%) and CFO (52%).

4 Results

4.1 Phase 1 Results: Robustness of Random Forests to Winsorization

Phase 1 tests the predictive power of random forests with minimal information. It also tests the effect of winsorizing the input data, a common practice in accounting research. In phase 1, we create three random forests per profitability measure and use a minimal set of features. For each random forest, the only predictors (features) in each observation are a firm identifier (PERMNO), fiscal year, and the contemporaneous value of the profitability measure. The target variable (dependent variable) is an indicator variable that indicates whether profitability increased or decreased in the subsequent year. The non-categorical measures in the input data are winsorized at three levels, 0%, 1%, and 2.5% on each side.

In Table 2, panels A to E report the results of phase 1 for each of the target variables. The results are remarkably strong. For every profitability measure, the random forests resoundingly beat both random walk models. Out-of-sample predictive accuracy (defined as the fraction of observations properly classified in the testing data) reported in the first line of each panel ranges from a low of 57.1% for ROE (Panel A) to a high of 59.8% for RNOA (Panel C), compared to 50% for the more difficult random walk model 1.[20] A test of proportions (untabulated) reveals that the differences in proportions of accurate classifications between ML and 50% (random walk model 1) are highly significant ($p < 0.001$). Thus, with minimal information, a random forest is better able to forecast changes in profitability than either of the two random walk models we employ.

All five panels show that winsorizing the data appears to have negligible effect: the accuracy of the random forests is essentially invariant to winsorization. For example, the first line in Panel A shows out-of-sample accuracy of 57.1% for all three degrees of winsorization. This is unsurprising given how classification trees work; they recursively partition the data using the available features and choose a threshold level for each feature. Outliers do not affect this process, and Scikit-Learn's documentation claims that trees are robust to arbitrary scaling of the data (scikit-learn 2019).

Our testing subsample spans the years 2013 – 2017. In addition to reporting the overall classification accuracy, we also report accuracy for each year of the testing subsample ("Test Scores by Year" in the middle of each panel). Interestingly, in all five panels the performance does not decline over time. Accuracy is quite stable, even 5 years out of the training data. For example, in Panel A with 0% winsorization, out-of-sample accuracy is 57.9% in 2013 and 58.9% in 2017. In general, the accuracy for individual years is typically between $\pm 1.5\%$ of the overall accuracy. This suggests that the patterns uncovered by the random forests, which used training data from 1963 – 2012, are relatively stable over time.

4.2 Phase 2 Results: Incremental Informativeness of Industry Information and Other Fundamental Variables

In phase 2 of the research design, we create additional random forests for each profitability measure. In each forest, we add additional features that potentially contain incremental information

[20] Random Walk Model 2 never achieves a score above 50%. Therefore, all remaining discussion focuses on model 1, which is equivalent to flipping a coin.

that can be exploited. Specifically, we add industry profitability measures and, in some analyses, one to two lags of the profitability measure, one to two lagged differences of the measure, and the accrual and cash flow components of the contemporaneous measure. In Table 3, panels A to E report the results of phase 2 for each of the target variables. The results are similar across all measures. As in phase 1, all models resoundingly beat both random walk models; all scores in the first line of each panel are a *minimum* of 8% higher than the random walk benchmark of 50%. A comparison of Table 3 to Table 2 (0% winsorization) reveals a marginal improvement in predictive accuracy of phase 2 relative to phase 1. For example, in Panel A of Table 2 accuracy is 57.1% and in Panel A of Table 3 accuracy is 58.9% when all features are included in the analysis (last column). In Table 3 the maximum classification accuracy is higher for the cash flow measures (62.8% for CFO and 63.7% for FCF) than for the earnings measures (59.1% for ROE, 59.6% for ROA, and 60.9% for RNOA).

For return on equity (ROE), out-of-sample classification accuracy improves from 57.1% (best score in phase 1) to 59.1% (Table 3, Panel A, analysis 2-2b). For return on assets (ROA), out-of-sample classification accuracy improves from 57.7% (best score in phase 1) to 59.6% (Table 3, Panel B, analysis 2-3a). For return on net operating assets (RNOA), out-of-sample classification accuracy improves from 59.8% (best score in phase 1) to 60.9% (Table 3, Panel C, analysis 2-2c). For cash flow from operations (CFO), out-of-sample classification accuracy improves from 57.8% (best score in phase 1) to 62.8% (Table 3, Panel D, analyses 2-2c and 2-3c). Finally, for free cash flow (FCF), out-of-sample classification accuracy improves from 58.5% (best score in phase 1) to 63.7% (Table 3, Panel E, analysis 2-3c).

As in phase 1, results of phase 2 are reasonably stable during the individual years 2013 – 2017 of the testing subsample (reported in the middle of each panel of Table 3). Classification accuracy varies by about $\pm 2\%$, but always remains significantly higher than the random walk benchmark of 50% (untabulated test of proportions, $p < 0.001$, for all measures and all years).

4.2.1 Feature Importance

In each panel of Table 2 and Table 3, we report “feature importance” scores. Since the tables are organized by measure, we created a sequence of figures shown in Figure 2 that are organized by analysis; these allow examination of feature importance across measures. Feature importance is a measure of the degree to which features are used across the random forest.[21] Recall that each tree in the forest is randomly created through bootstrapping, and therefore different. Feature importance detects which features are used most often, and at higher levels of the trees. The sum of the feature importance scores for any forest must sum to 100%.

Some patterns emerge from the feature importance results. The contemporaneous value of the measure is consistently the most important feature, across all analyses. This finding suggests that there is a lot of information about future changes in profitability in the current level of profitability. In other words, the best predictor of future profitability is current profitability. Since random forests are more “black box” than regressions, it is not possible to extract more general rules from the above analysis.

The second-most important group of features appear to be the difference between current profitability and the industry mean/median. The features *Measure – Ind. Mean* and *Measure – Ind. Median* are consistently important, and often the second-most important features. Perhaps, these differences help the method detect mean reversion in profitability. Table 1, panel D shows that these two features are highly correlated. The correlations between these measures for ROE, ROA, RNOA, CFO, and FCF are 0.68, 0.8, 0.49, 0.9, and 0.9, respectively. Thus, removing one of these features from the analysis would likely increase the feature importance of the other. Striking feature importance patterns for accruals are present in analyses 2-1c, 2-2c, and 2-3c, where for predicting CFO and FCF, accruals are the second most important feature after CFO and FCF. This evidence, coming from the very demanding

[21] Feature importance is analogous to significance in a regression. Feature importance is also analogous to coefficient magnitude; a feature may appear in a tree, but the higher the placement in the tree (nearer the root node), the greater the impact on the prediction.

out-of-sample prediction tests, highlights the usefulness of accruals in predicting future cash flows, a central issue in financial accounting but without robust evidence from prior studies.

Features that are differences are generally more important than features that are levels. For example, *Measure – Ind. Median* is generally more important than *Ind. Median*. Also, the first and second differences of the measures are generally more important than the first and second lags. Apparently, this is common with trees in machine learning. The tree can figure out how to use differences, but it requires multiple splits. Spoon-feeding the difference to the tree makes the tree's job easier.

As the number of features increases, feature importance tends to spread out across the features without large gains in classification accuracy. We suspect this pattern results from the high correlations among our features. High correlations imply little incremental information. Note, however, that high correlations are not problematic for trees as they are for regressions.

PERMNO and SIC codes are surprisingly useful, especially since Scikit-Learn's tree methods do not recognize categorical variables. Instead, Scikit-Learn treats these variables as continuous![22,23] Thus, there may be information in the order in which the PERMNO's and SIC codes were assigned.

4.2.2 Mean Reversion within Industries

Nissim and Penman (2001), hereafter NP, find that profitability measures are mean reverting across the entire universe of firms. We conjecture that profitability measures revert *within* individual industries and that is why we chose to include industry-level profitability as a feature. We provide prima facie evidence that profitability reverts within industry. We replicated the procedure of NP within five

[22] An implication of this limitation is that these trees do not recognize panel data in the same way that econometric methods do. Machine learning has not yet developed methods for panel data (Varian 2014).

[23] It would be possible to convert PERMNO or SIC to a string and thereby force them to be categorical variables, but the number of possibilities the tree would need to try to find a split would be computationally infeasible. Alternatively, "one hot" encoding would create an indicator variable for each category of a discrete variable. However, since there are over 13,000 unique PERMNO values in our dataset, this technique would greatly expand our feature set while providing little insight into why one model outperforms another.

randomly-chosen industries. All five exhibit evidence of mean reversion of profitability, within the industry. In Figure 3, we present graphs, one for each of our measures of profitability, for two industries, crude oil (SIC 1311) and pharmaceuticals (SIC 2834). The graphs are qualitatively similar to those in NP, and we interpret this as evidence of mean reversion within industries that potentially our method exploits in the prediction process.

4.3 Phase 3 Results: Conditional Analyses

In phase 3 of the research design, we explore whether the predictive performance of our ML varies across the levels of these "conditioning" variables: book-to-market ratio, market capitalization (size), debt-to-assets ratio, earnings-to-market ratio, and accruals-to-market ratio. We computed these variables for every firm year in the *testing* data (fiscal years 2013 – 2017), and then grouped all observations in each fiscal year into deciles. For each decile, we used the best random forest from phase 2 to compute fitted values for each profitability measure (target variable). We then computed the classification accuracy within each decile of each conditioning variable for each target measure.

Figure 4 displays the results by conditioning variable and by profitability measure. For example, the top-left graph of Figure 4, panel A shows how the classification accuracy of ROE varies across book-to-market deciles. In that figure, we observe that accuracy is highest in the first decile (very low book-to-market firms with negative book values), and lowest in the second decile (low book-to-market firms with positive book values)[24]. Looking across all figures, we observe that the highest predictability tends to occur in the first and last decile of the conditioning variable. This trend is particularly strong for earnings-to-market ratio (Figure 4, panel D) and accruals-to-market ratio (Figure 4, panel E), where the classification accuracy often exceeds 70%. We also observe that the classification accuracy for firms with high accruals-to-market ratio exceeds 70% for the *cash-flow variables* CFO and FCF. Since accruals are

[24] The average values of the conditioning variables in each decile are shown in Table 4.

included as features in predicting CFO and FCF, the implication of such strong prediction performance is that high-accruals convey reliable information about future cash flows. Surprisingly, classification accuracy does not vary with leverage (debt-to-assets ratio) and firm size (market capitalization).

Table 4 presents similar information as Figure 4, but adds the mean value of the conditioning variable in each decile. Also, data in Table 4 are organized by decile of the conditioning variable. This allows for comparison of the profitability measures across levels of the conditioning variables. For example, in Table 4, panel A, we can see that, for the firms with the lowest book-to-market ratio (decile 1), ROE demonstrates the highest predictability (looking across the row for decile 1). By contrast, for firms with the highest book-to-market ratio (decile 10), FCF demonstrates the highest predictability. In decile 9 of book-to-market, where the mean value of book-to-market is close to 1.0, CFO is the most predictable. In unreported results we compare the prediction performance of the CFO and FCF models when accruals is not included as a feature to the performance of ROE, ROA, and RNOA models. The comparison reveals that the absence of accruals resulted in weaker performance of the cash flow models than the performance shown in Table 4. This finding confirms the usefulness of accruals in predicting cash flows under various conditions.

In general, the cash-flow measures of profitability exhibit higher predictability than do the earnings-based measures. Some notable exceptions occur within debt-to-assets ratio (Table 4, panel C, decile 10), earnings-to-market ratio (Table 4, panel D, deciles 1 and 10), and (Table 4, panel E, decile 1). In Table 4, panel D, we observe that in the lowest and highest decile of earnings-to-market ratio, the accruals-based measures (ROE, ROA, and RNOA) outperform the cash-flow measures (CFO and FCF). The results are similar in Table 4, panel E, where the accruals-based measures outperform for low accruals firms, but not for high accruals firms.

5 Discussion and Conclusion

We explore whether a method from machine learning (ML), classification trees, is able to generate out-of-sample profitability forecasts that are superior to random walk forecasts. We are motivated to apply the ML method because a) the literature shows that traditional regression methods cannot generate out-of-sample forecasts that are superior to random walk forecasts, and b) ML methods have some advantages over regression methods in generating out-of-sample predictions as they are focused on prediction and maximizing prediction accuracy. Examples of such advantages include insensitivity to econometric issues such as multicollinearity, better handling of nonlinearity in the data than traditional regression-based methods, and discovery of the functional form that best fits the data. We implement the ML method using a large sample of US firms with valid required data over the period 1963-2017 and generate out-of-sample predictions of directional changes (increases or decreases) in five profitability measures: return on equity (ROE), return on assets (ROA), return on net operating assets (RNOA), cash flow from operations (CFO), and free cash flow (FCF). Results based on a minimum set of independent variables show that our ML method achieves classification accuracies ranging from 57 – 64% for our profitability measures, compared to 50% for the random walk, and that the differences in proportions of accurate classifications between ML and random walk are highly significant. Furthermore, we find that the predictive performance of our ML method does not decline even when the forecast horizon is five years long. In addition, we find higher classification accuracy for the two cash-flow measures (CFO and FCF), especially when we include accruals in the prediction of cash flows, than the three earnings-based measures of profitability (ROE, ROA, and RNOA). However, in the extreme portfolios of all conditioning variables, the earnings-based target measures (ROE, ROA, RNOA) often outperform the cash-flow target measures (CFO, FCF). These results suggest that no single profitability measure dominates under all conditions. Overall, our results provide some evidence that ML methods have the potential to become useful in predicting profitability.

We feel that these results can be further improved, especially if we are willing to sacrifice model interpretability. For example, we could add more features, employ different ML methods, explode the feature space to properly handle categorical variables, and expand the set of hyperparameters over which we searched. Future work can also explore the use of accruals as features in the prediction of cash flows. Finally, our results can be compared to the results of a recent study by Vorst and Yohn (2018), who perform out-of-sample predictions using a regression methodology, but without using the random walk as a benchmark.

6 References

Ball, R., and P. Brown. 1968. An Empirical Evaluation of Accounting Income Numbers. *Journal of accounting research*:159-178.

Bradshaw, M. T., M. S. Drake, J. N. Myers, and L. A. Myers. 2012. A re-examination of analysts' superiority over time-series forecasts of annual earnings. *Review of Accounting Studies* 17 (4):944-968.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay. 1997. *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University press.

Dechow, P., W. Ge, and C. Schrand. 2010. Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of accounting and Economics* 50 (2-3):344-401.

Dichev, I. D., and V. W. Tang. 2009. Earnings volatility and earnings predictability. *Journal of accounting and Economics* 47 (1-2):160-181.

Ding, K., B. Lev, X. Peng, T. Sun, and M. A. Vasarhelyi. 2018. Machine Learning Improves Accounting Estimates. Available at SSRN: https://ssrn.com/abstract=3253220.

Elliott, G., and A. Timmermann. 2008. Economic Forecasting. *Journal of economic literature* 46 (1):3-56.

Fama, E. F., and K. R. French. 1992. The Cross-Section of Expected Stock Returns. *the Journal of Finance* 47 (2):427-465.

———. 1995. Size and Book-to-Market Factors in Earnings and Returns. *the Journal of Finance* 50 (1):131-155.

Feltham, G. A., and J. A. Ohlson. 1995. Valuation and Clean Surplus Accounting for Operating and Financial Activities. *Contemporary accounting research* 11 (2):689-731.

Gu, S., B. T. Kelly, and D. Xiu. 2019. Empirical Asset Pricing via Machine Learning. In *Chicago Booth Research Paper No. 18-04; 31st Australasian Finance and Banking Conference 2018; Yale ICF Working Paper No. 2018-09*.

Hou, K., M. A. van Dijk, and Y. Zhang. 2012. The implied cost of capital: A new approach. *Journal of accounting and Economics* 53 (3):504-526.

Howard, J., and M. Bowles. 2012. The Two Most Important Algorithms in Predictive Modeling Today. Paper read at Strata Conference, http://strataconf.com/strata2012/public/schedule/detail/22658, February 28.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112. New York: Springer.

Lakonishok, J., A. Shleifer, and R. W. Vishny. 1994. Contrarian Investment, Extrapolation, and Risk. *the Journal of Finance* 49 (5):1541-1578.

Lev, B., S. Li, and T. Sougiannis. 2009. The usefulness of accounting estimates for predicting cash flows and earnings. *Review of Accounting Studies* 15 (4):779-807.

Li, N., S. Richardson, and İ. Tuna. 2014. Macro to micro: Country exposures, firm fundamentals and stock returns. *Journal of accounting and Economics* 58 (1):1-20.

Monahan, S. J. 2017. Financial Statement Analysis and Earnings Forecasting. *Foundations and Trends in Accounting* 12 (2):105-215.

Mullainathan, S., and J. Spiess. 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* 31 (2):87-106.

Nissim, D., and S. H. Penman. 2001. Ratio Analysis and Equity Valuation: From Research to Practice. *Review of Accounting Studies* 6 (1):109-154.

Ohlson, J. A. 1995. Earnings, Book Values, and Dividends in Equity Valuation. *Contemporary accounting research* 11 (2):661-687.

Ou, J. A., and S. H. Penman. 1989a. Accounting Measurement, Price-Earnings Ratio, and the Information Content of Security Prices. *Journal of accounting research* 27:111-144.

———. 1989b. Financial Statement Analysis and the Prediction of Stock Returns. *Journal of accounting and Economics* 11 (4):295-329.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825-2830.

Penman, S. H. 2012. *Financial Statement Analysis and Security Valuation*. 5th ed: McGraw-Hill/Irwin.

Poon, S.-H., and C. W. Granger. 2003. Forecasting Volatility in Financial Markets: A Review. *Journal of economic literature* 41 (2):478-539.

Ramnath, S., S. Rock, and P. B. Shane. 2008. Financial Analysts' Forecasts and Stock Recommendations: A Review of the Research. *Foundations and Trends in Finance* 2 (4):311-421.

scikit-learn. 2019. *Compare the effect of different scalers on data with outliers* 2019 [cited August 29 2019]. Available from https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#sphx-glr-auto-examples-preprocessing-plot-all-scaling-py.

Sloan, R. G. 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review* 71 (3):289-315.

Varian, H. R. 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives* 28 (2):3-28.

Vorst, P., and T. L. Yohn. 2018. Life Cycle Models and Forecasting Growth and Profitability. *The Accounting Review* 93 (6):357-381.

Zhang, A. C. 2018. Predict Audit Quality Using Machine Learning Algorithms. Available at SSRN: https://ssrn.com/abstract=3449848.

7 Appendix A: Variable Definitions

7.1 Compustat Variables

Variable	Code
Current Assets – Total	ACT
Assets – Total	AT
Capital Expenditures	$CAPX$
Common Equity – Total	CEQ
Cash and Short-Term Investments	CHE
Debt in Current Liabilities (short-term debt)	DLC
Long-Term Debt – Total	$DLTT$
Depreciation and Amortization	DP
Data year – fiscal	$FYEAR$
Earnings (income before extraordinary items)	IB
Earnings (income before extraordinary items, from statement of cash flows)	IBC
Interest and Related Income – Total	$IDIT$
Short-term investments – Total	$IVST$
Current Liabilities – Total	LCT
Liabilities – Total	LT
Operating Activities – Net Cash Flow	$OANCF$
Operating Income before Depreciation	$OIBDP$
CRSP permanent number	$PERMNO$
Property, Plant, and Equipment – Total (Net)	$PPENT$
Standard Industry Classification	SIC
Income Taxes Payable	TXP
Extraordinary Items and Discontinued Operations (Statement of Cash Flows)	$XIDOC$
Interest Expense	$XINT$

7.2 Intermediate Computed Variables

To compute accruals, we follow the procedure described in Hou et al. (2012). The subscripts t and $t-1$ denote an arbitrary year and the year prior. The Δ operator indicates the change between the current and previous year. We compute net operating assets by following the procedure described in Li, Richardson, and Tuna (2014); net operating assets are the difference between operating assets and operating liabilities.

Variable	Formula
Accruals (1963 - 1987)	$ACC_t = (\Delta \mathrm{ACT_t} - \Delta CHE_t) - (\Delta \mathrm{LCT_t} - \Delta DLC_t - \Delta TXP_t) - DP_t$
Accruals (1988 – 2017)	$ACC_t = IBC_t - (OANCF_t - XIDOC_t)$
Net Operating Assets	$NOA_t = (AT_t - CHE_t - IVST_t) - (LT_t - DLC_t - DLTT_t)$
Operating Income	$OI_t = IB_t + XINT_t - IDIT_t$

7.3 Target Variables

Variable	Formula
Return on Common Equity (ROE)	$ROE_t = \frac{IB_t}{0.5(CEQ_t + CEQ_{t-1})}$
Return on Assets (ROA)	$ROA_t = \frac{IB_t + XINT_t}{0.5(AT_t + AT_{t-1})}$
Return on Net Operating Assets (RNOA)	$RNOA_t = \frac{OIBDP_t - DP_t}{0.5(NOA_t + NOA_{t-1})}$
Cash Flow from Operations (CFO), scaled	$CFO_t = \frac{OANCF_t}{0.5(AT_t + AT_{t-1})}$
Free Cash Flow (FCF), scaled	$FCF_t = \frac{OANCF_t - CAPX_t}{0.5(AT_t + AT_{t-1})}$

8 Appendix B: Classification Accuracy of Random Walk Model with Non-Normal Error Term

Consider a general random walk model in which the error term is positive q percent of the time, and negative $1 - q$ percent of the time. Assume the testing data contains N observations, and $0 \leq p \leq 1$ is the fraction of increases. Then pN of those observations are increases, and $(1-p)N$ are decreases. For any observation, the random walk will predict increase q percent of the time and decrease $1-q$ percent of the time. Thus, of the pN increases, on average qpN will be classified accurately. Similarly, of the $(1-p)N$ decreases, $(1-q)(1-p)N$ will be classified accurately. From this, we can write the following confusion matrix.

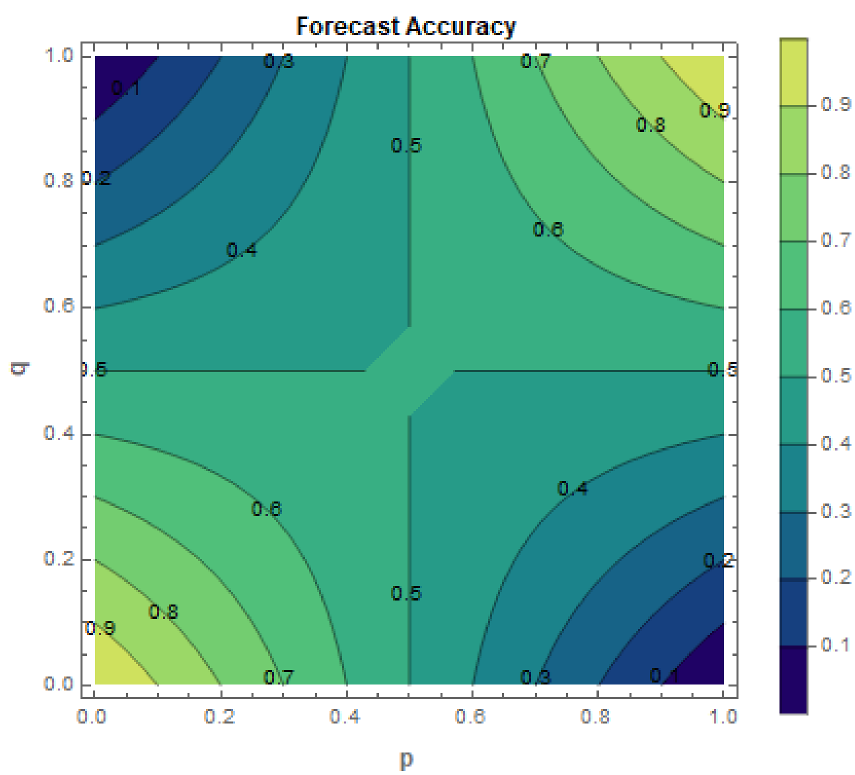
	Actual: Decrease	**Actual: Increase**
Forecasted: Decrease	$(1-p)(1-q)N$	$p(1-q)N$
Forecasted: Increase	$(1-p)qN$	pqN

From this matrix, we can compute the expected accuracy as follows:

$$Accuracy = \frac{\text{True Increases} + \text{True Decreases}}{N} = \frac{qpN + (1-q)(1-p)N}{N} = 1 - p - q + 2pq$$

Thus, if the actual fraction of increases in the data is p, and the error term of the random walk increases q percent of the time, then the expected classification accuracy of the random walk model is $1 - p - q + 2pq$.

Plotting the accuracy over the space $0 \leq p, q \leq 1$ yields:



The figure shows that if p and q are within the range 40 – 60%, classification accuracy ranges from 48 – 52%, which is very close to the 50% achieved by assuming a normally distributed error term.

9 Figures

Figure 1: Sample classification tree.

The tree has three layers. The root node indicates that firms with ROA in a given year greater than 0.05 are predicted to experience an increase in ROA in the subsequent year (right branch). Firms with ROA less than or equal to 0.05 (left branch) and less than -0.03 (left branch) are predicted to experience a decrease in ROA in the next year. However, if ROA is in the range (-0.03, 0.05], then industry mean ROA is the final deciding factor. If industry median ROA is less than or equal to (greater than) 0.02, the firm is predicted to experience an increase (decrease) in ROA.

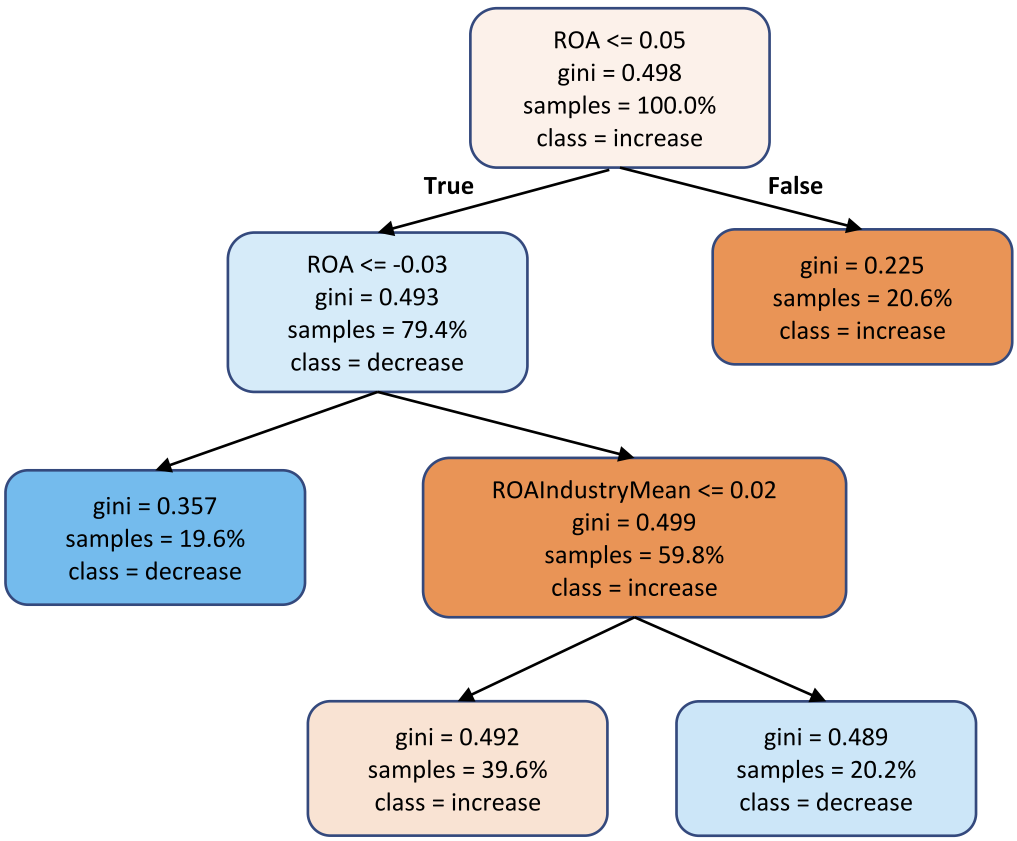
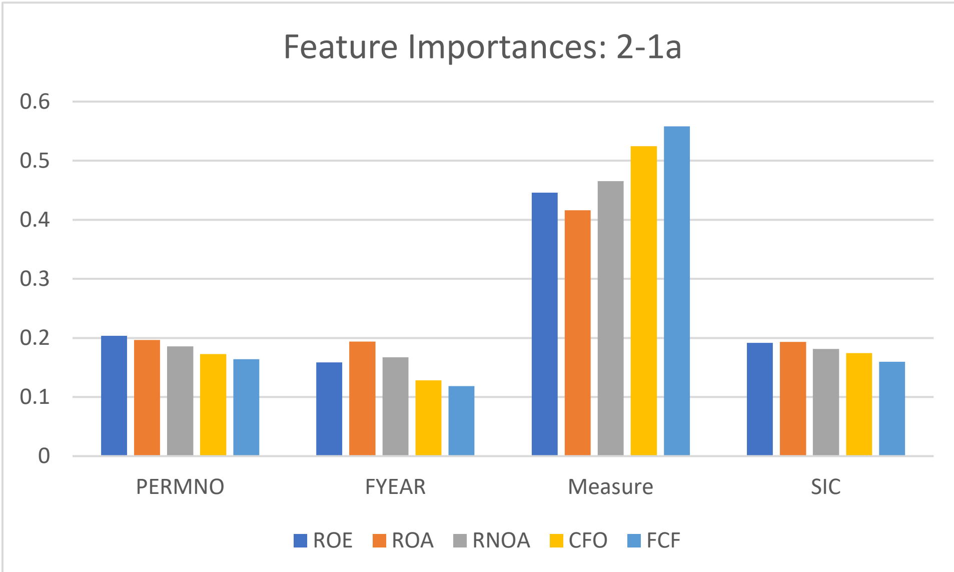
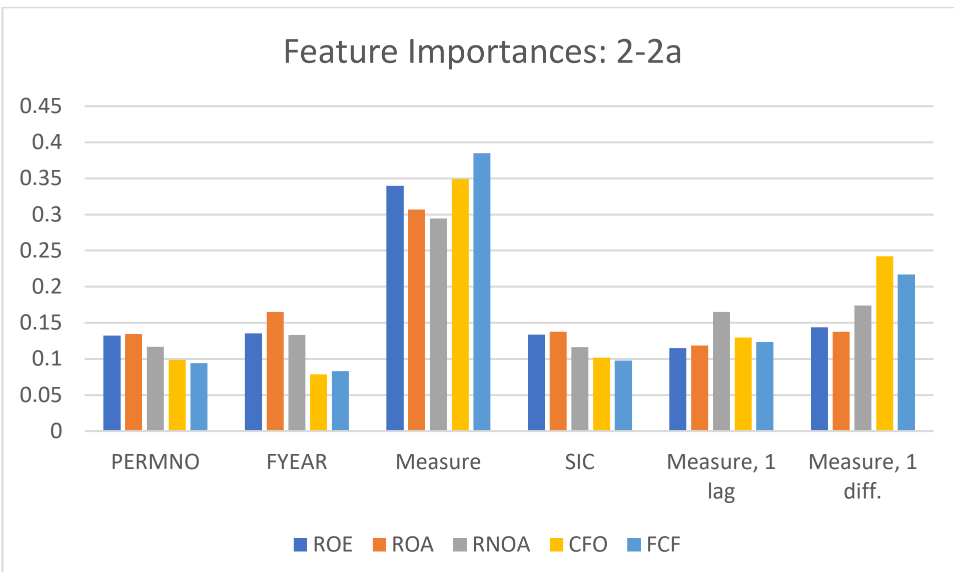
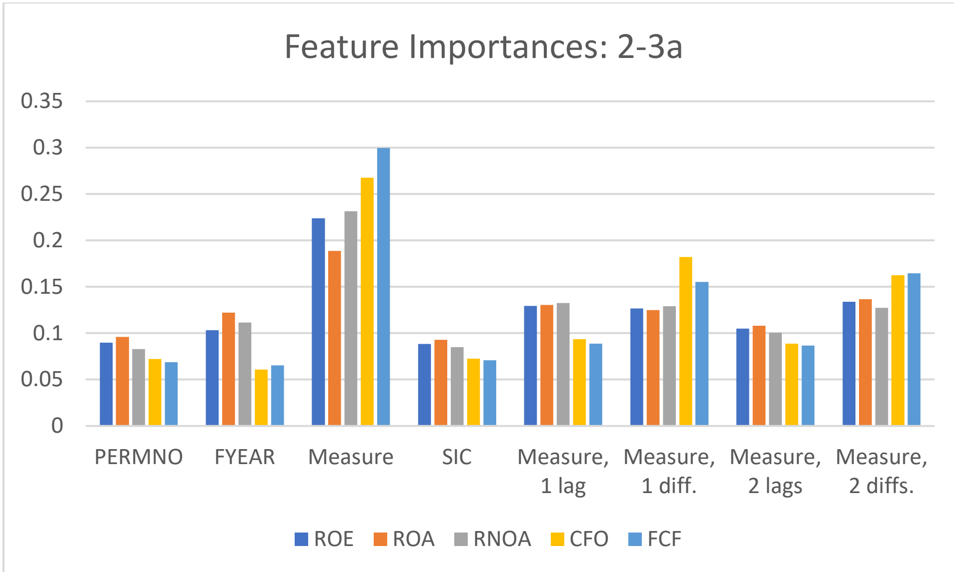


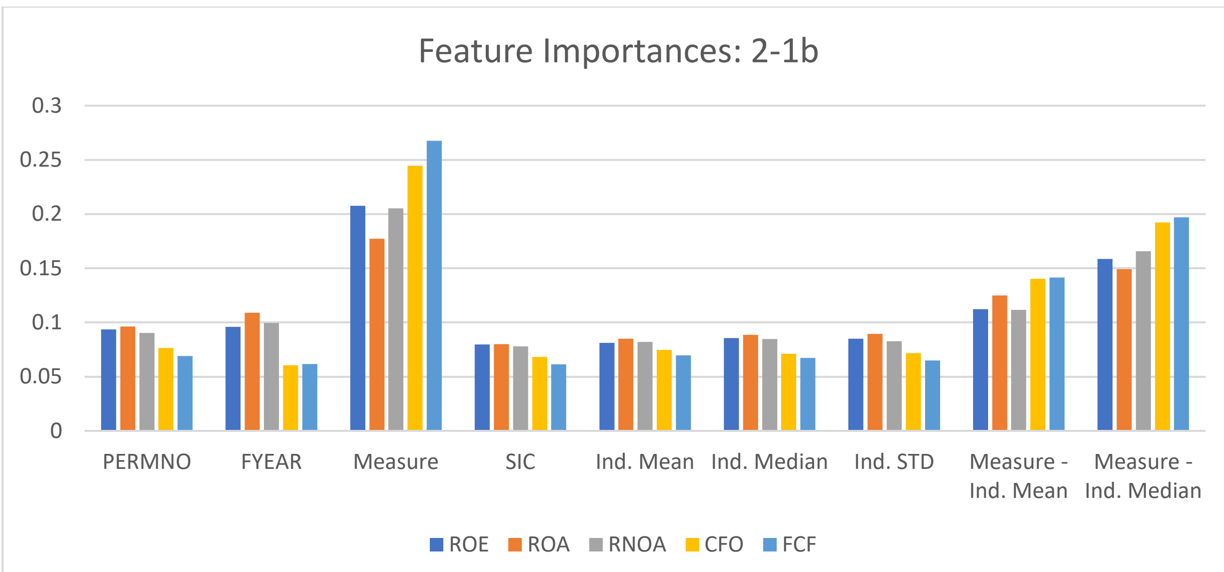
Figure 2: Feature Importance

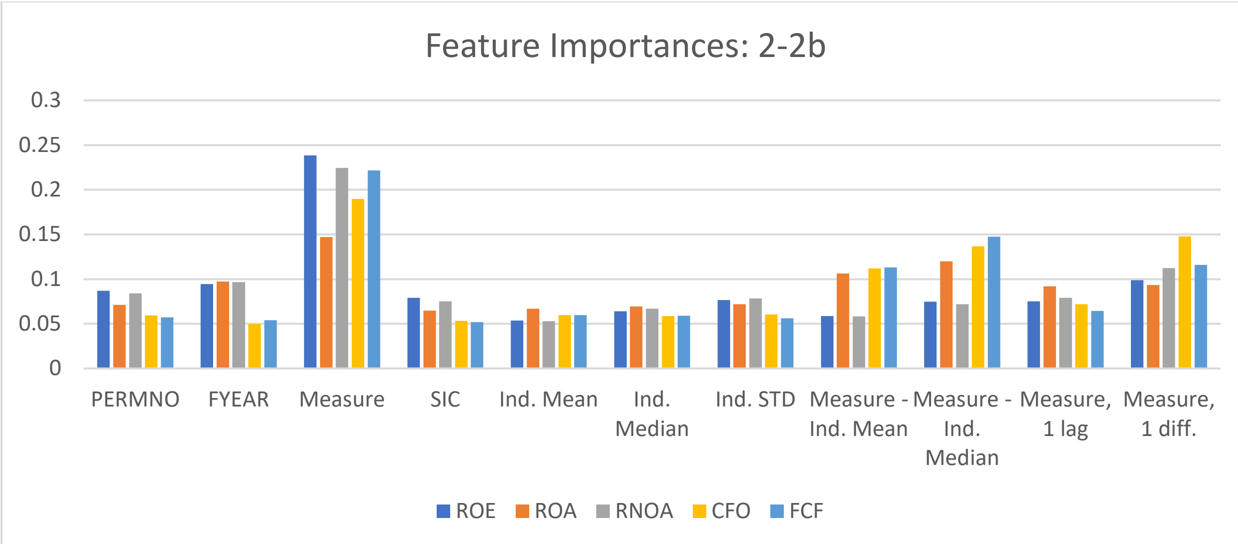
For each of the models reported in Table 3, we show feature importance in a bar chart, grouped by feature. Feature importance is a measure of each feature's importance in determining splits. The higher the feature importance, the more frequently the feature was used in the forest's trees, and the higher the feature was, on average, in the tree hierarchy.

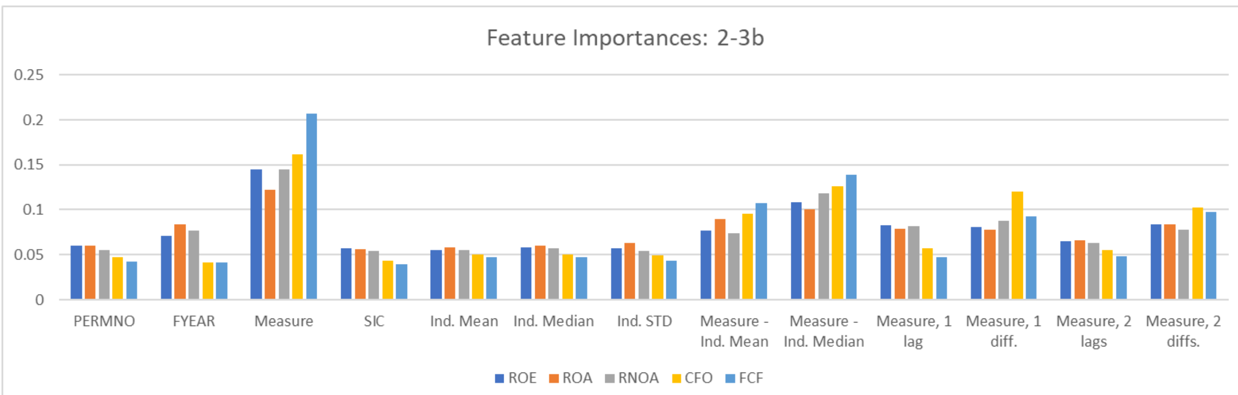


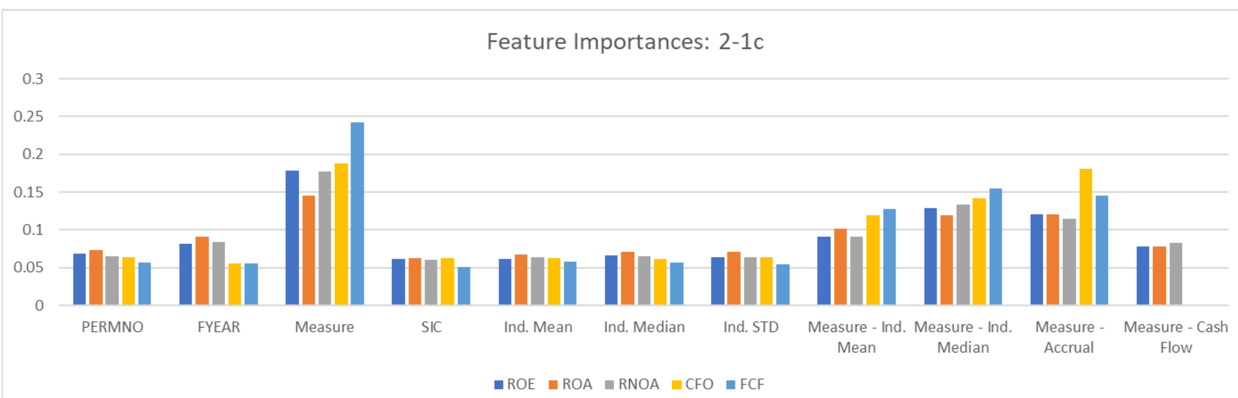


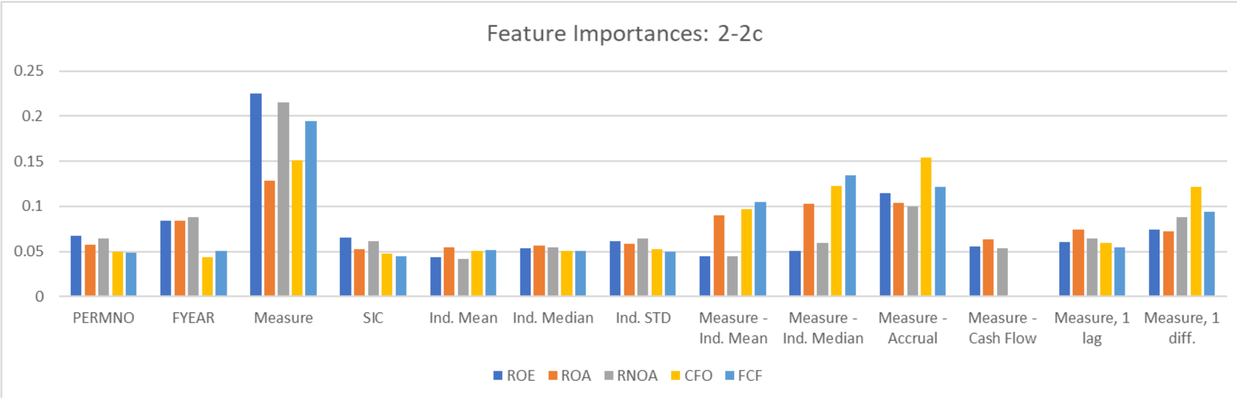

Feature Importances: 2-3a
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
PERMNO
FYEAR
Measure
SIC
Measure, 1 lag
Measure, 1 diff.
Measure, 2 lags
Measure, 2 diffs.
ROE
ROA
RNOA
CFO
FCF


Feature Importances: 2-1b
0.3
0.25
0.2
0.15
0.1
0.05
0
PERMNO
FYEAR
Measure
SIC
Ind. Mean
Ind. Median
Ind. STD
Measure - Ind. Mean
Measure - Ind. Median
ROE
ROA
RNOA
CFO
FCF


Feature Importances: 2-2b
0.3
0.25
0.2
0.15
0.1
0.05
0
PERMNO
FYEAR
Measure
SIC
Ind. Mean
Ind. Median
Ind. STD
Measure - Ind. Mean
Measure - Ind. Median
Measure, 1 lag
Measure, 1 diff.
ROE ROA RNOA CFO FCF


Feature Importances: 2-3b
0.25
0.2
0.15
0.1
0.05
0
PERMNO
FYEAR
Measure
SIC
Ind. Mean
Ind. Median
Ind. STD
Measure - Ind. Mean
Measure - Ind. Median
Measure, 1 lag
Measure, 1 diff.
Measure, 2 lags
Measure, 2 diffs.
ROE ROA RNOA CFO FCF


Feature Importances: 2-1c
0.3
0.25
0.2
0.15
0.1
0.05
0
PERMNO
FYEAR
Measure
SIC
Ind. Mean
Ind. Median
Ind. STD
Measure - Ind. Mean
Measure - Ind. Median
Measure - Accrual
Measure - Cash Flow
ROE ROA RNOA CFO FCF


Feature Importances: 2-2c
0.25
0.2
0.15
0.1
0.05
0
PERMNO
FYEAR
Measure
SIC
Ind. Mean
Ind. Median
Ind. STD
Measure - Ind. Mean
Measure - Ind. Median
Measure - Accrual
Measure - Cash Flow
Measure, 1 lag
Measure, 1 diff.
ROE ROA RNOA CFO FCF

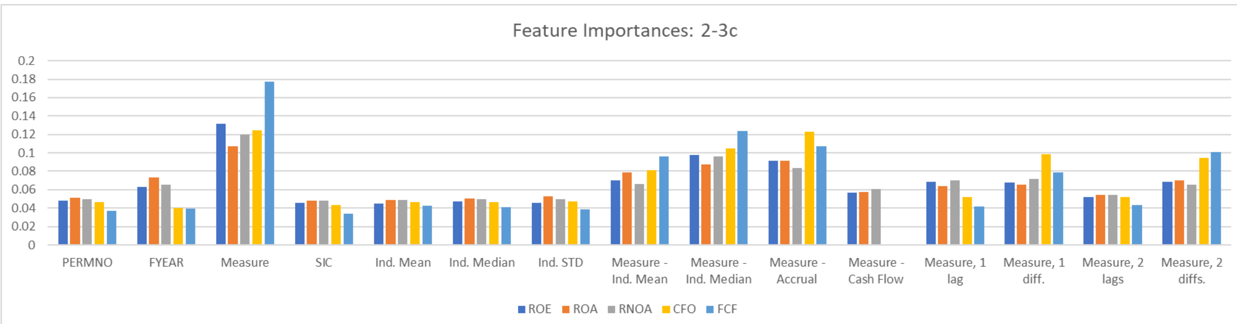
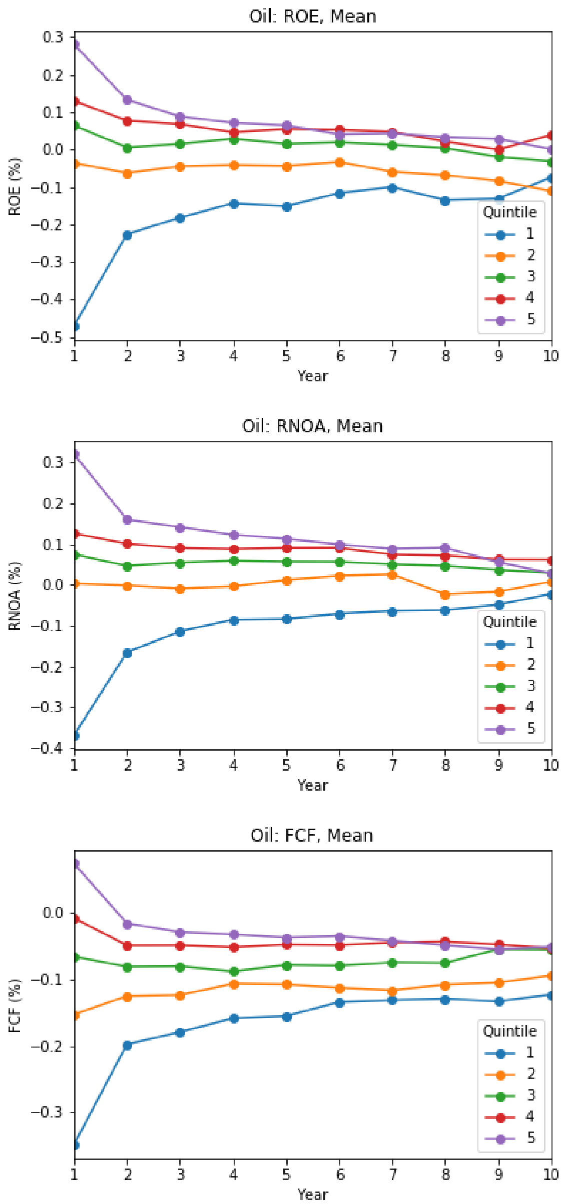

Feature Importances: 2-3c
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
PERMNO
FYEAR
Measure
SIC
Ind. Mean
Ind. Median
Ind. STD
Measure - Ind. Mean
Measure - Ind. Median
Measure - Accrual
Measure - Cash Flow
Measure, 1 lag
Measure, 1 diff.
Measure, 2 lags
Measure, 2 diffs.
ROE ROA RNOA CFO FCF

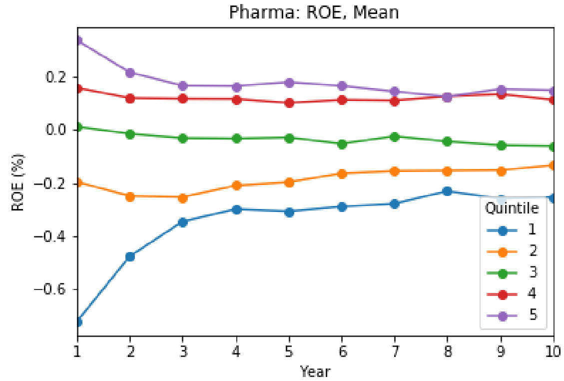
Figure 3: Mean Reversion within Industries

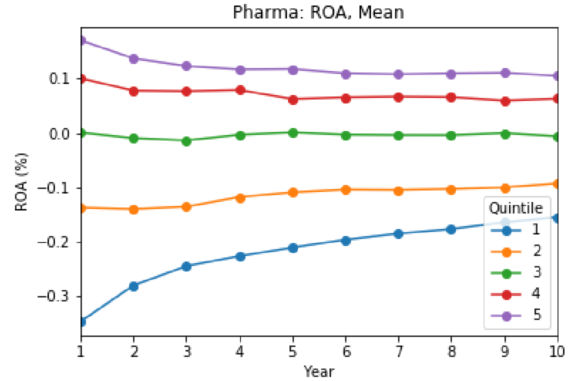
For two randomly chosen industries, crude oil (SIC: 1311, Panel A) and pharmaceuticals (SIC: 2834, Panel B), we test for mean reversion of each of our target variables within industry. The graphs are generated following the method in Nissim and Penman (2001). In each year, firms are grouped into portfolios based on quintile of the target variable. The performance of each portfolio is tracked over the next 9 years. The average value of the target variable is then computed in each year.

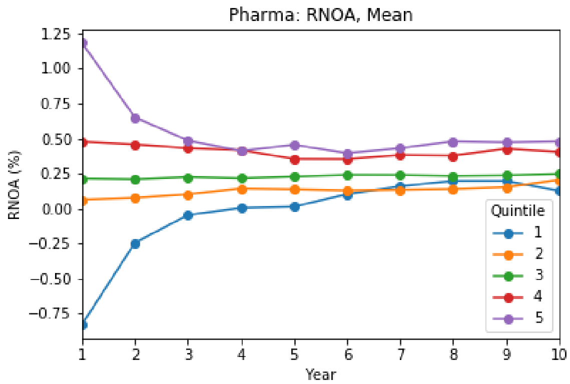
Panel A: Crude Oil (SIC: 1311)

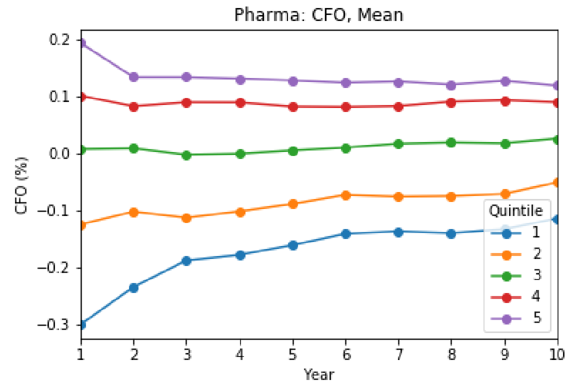


Panel B: Pharmaceuticals (SIC: 2834)









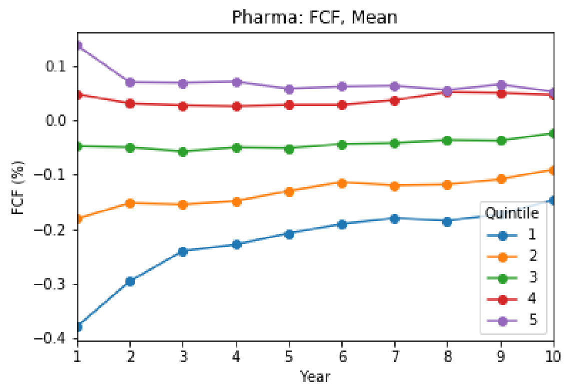


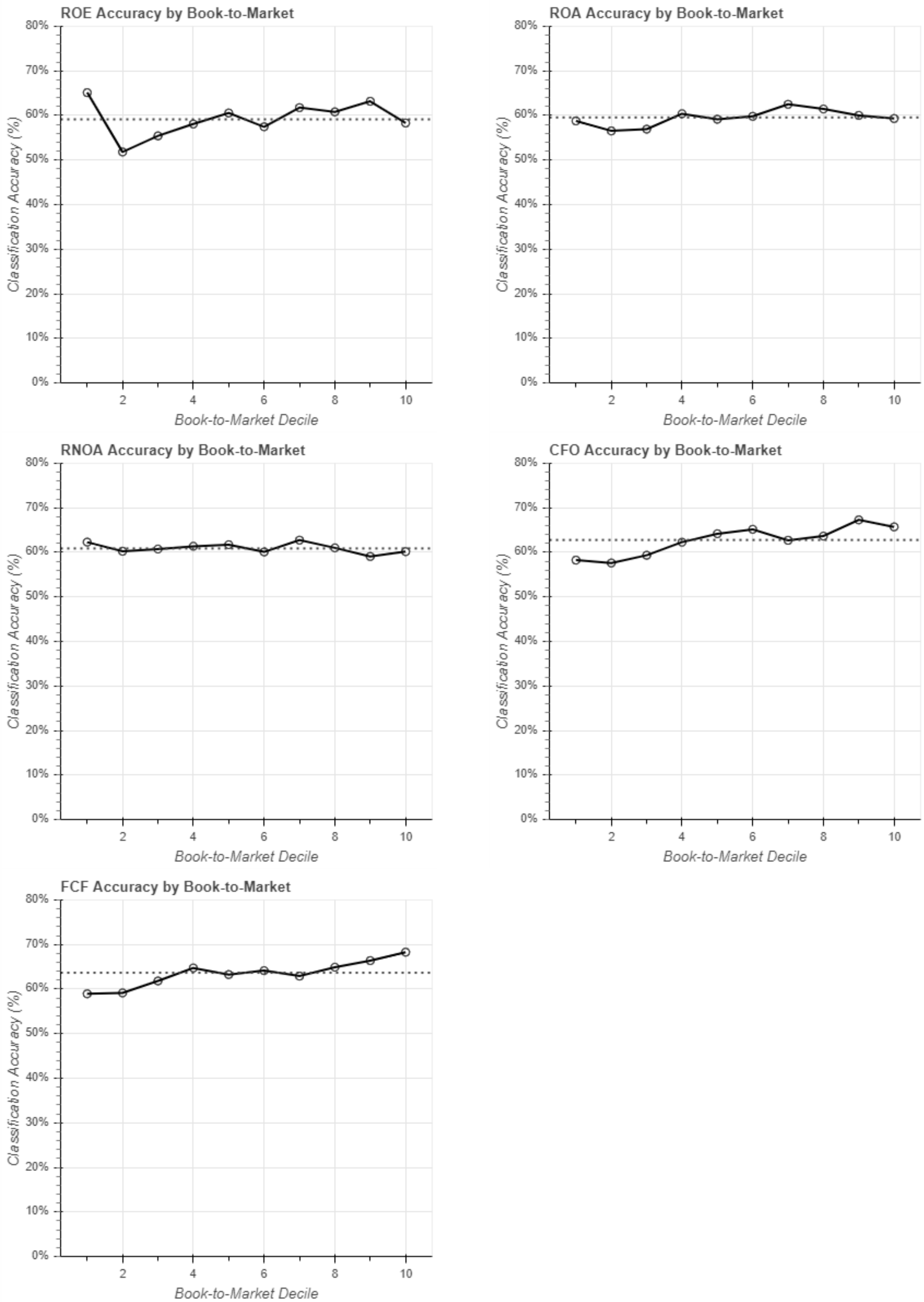
Figure 4: Conditional Analysis

This figure shows the results of phase 3 of the research design. For each firm-year observation in the testing sample (fiscal years 2013 – 2017), we computed the book-to-market ratio, market capitalization, debt-to-assets ratio, earnings-to-market ratio, and accruals-to-market ratio. Book-to-market ratio is computed as common equity (Compustat: CEQ) divided by market capitalization. Market capitalization is computed as closing price at the end of the fiscal year (PRCC_F) times common shares outstanding (CSHO). Debt-to-assets ratio is computed as [long-term debt (DLTT) plus debt in current liabilities (DLC) plus preferred stock (PSTK)] divided by total assets (AT). Earnings-to-market ratio is computed as income before extraordinary items (IB) divided by market capitalization. Accruals-to-market ratio is computed as accruals (as defined in section 3) divided by market capitalization.

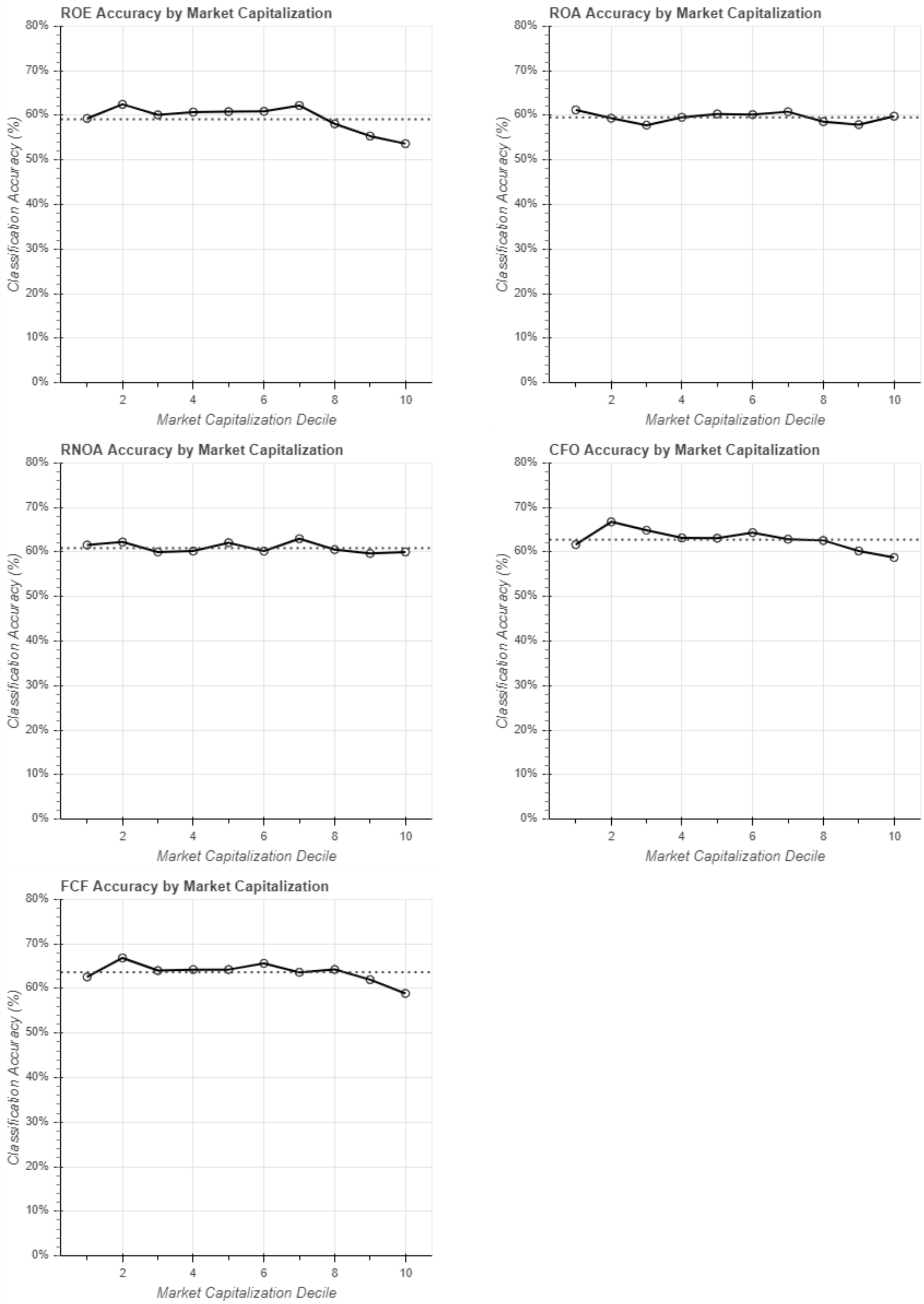
Within each fiscal year, and for each conditional variable, we formed deciles of each conditioning variable. For each decile of each conditioning variable, we generated fitted values for each target variable (ROE, ROA, RNOA, CFO, and FCF). We used the best-performing random forest grown in phase 2 to generate the fitted values. As shown in Table 3, the best-performing random forest for ROE was grown in Analysis 2-2b; the best performing random forests for ROA, RNOA, CFO, and FCF were grown in Analyses 2-3a, 2-2c, 2-3c, and 2-3c, respectively.

We used the fitted values to compute the classification accuracy within each decile of each conditioning variable for each profitability measure (target variable). We then averaged the classification accuracy across all 5 years of the testing data. In each figure below, we plot the classification accuracy (% of changes in the target variable predicted correctly) by decile of a conditioning variable. In each figure, the dotted line is the overall classification accuracy for that target variable; the average of the decile scores will equal the overall average. Panels A – E report the results for book-to-market ratio, market capitalization, debt-to-assets ratio, earnings-to-market ratio, and accruals-to-market ratio, respectively.

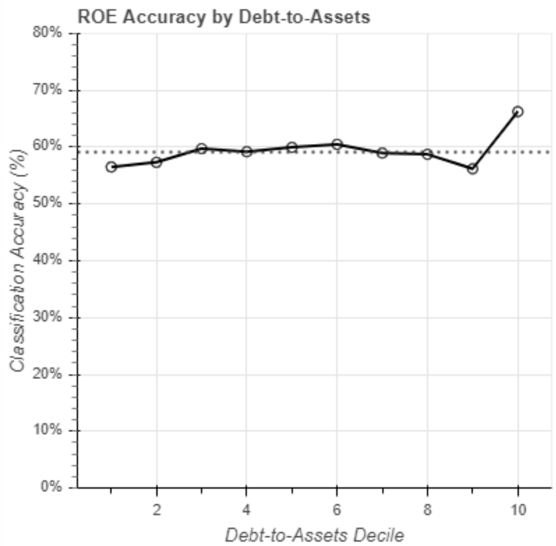
Panel A: Book-to-Market Ratio

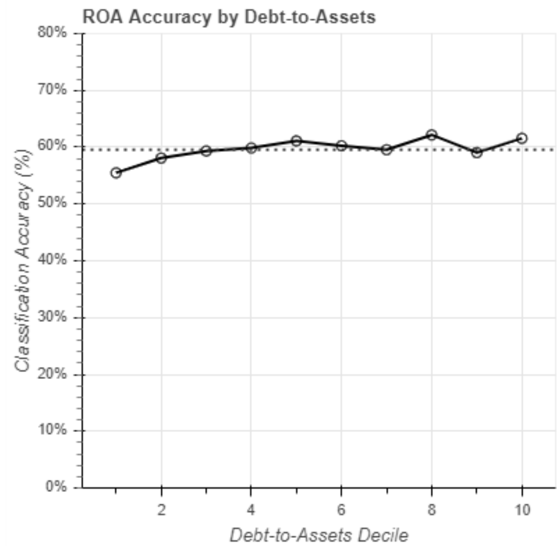


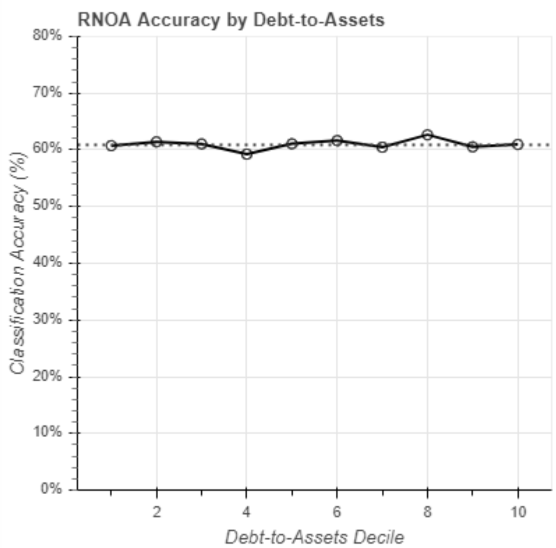
Panel B: Market Capitalization

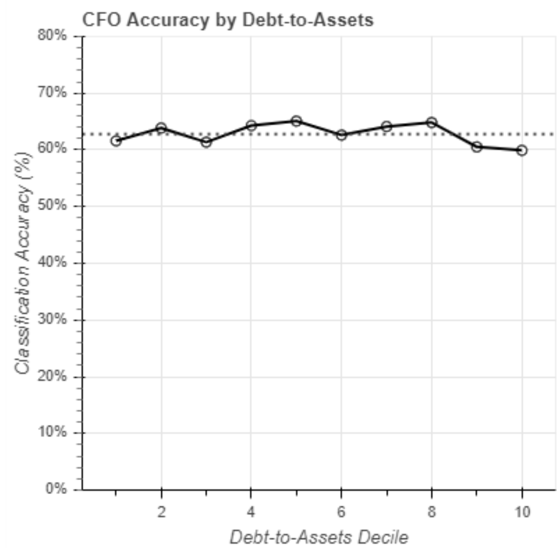


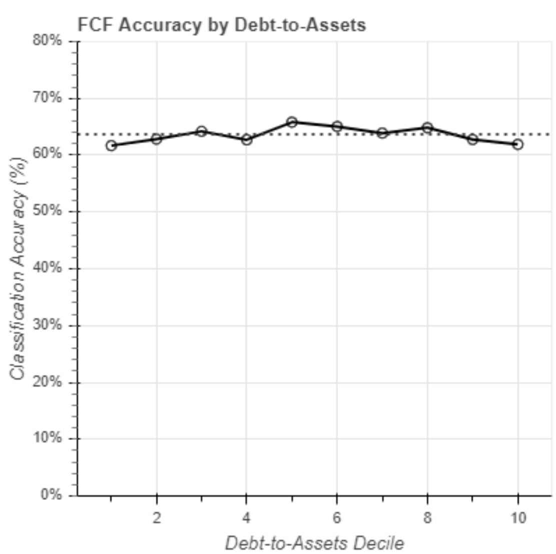
Panel C: Debt-to-Assets Ratio



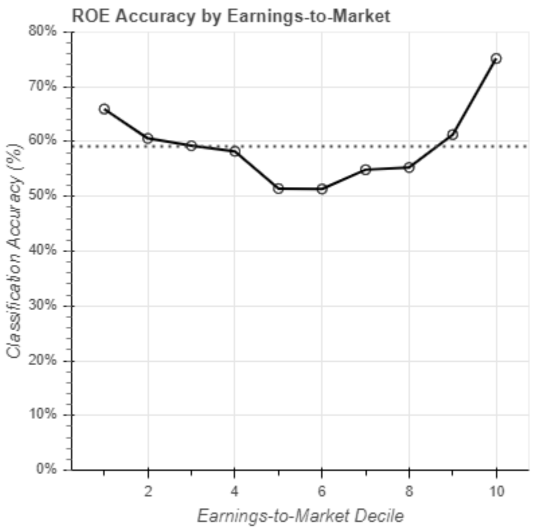


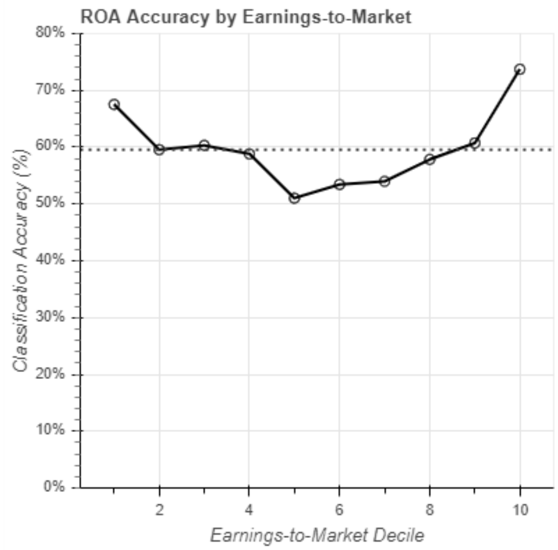


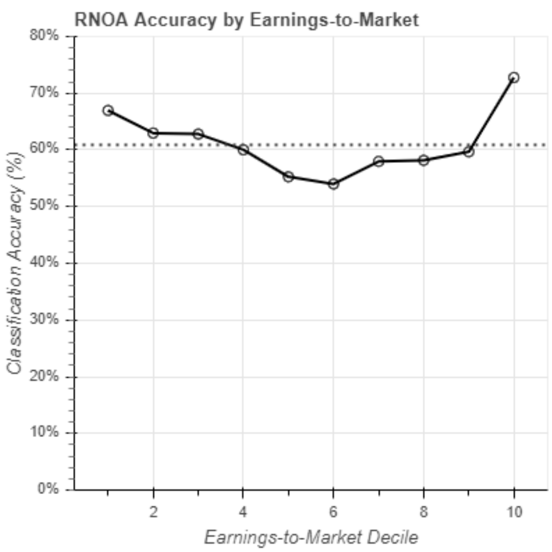


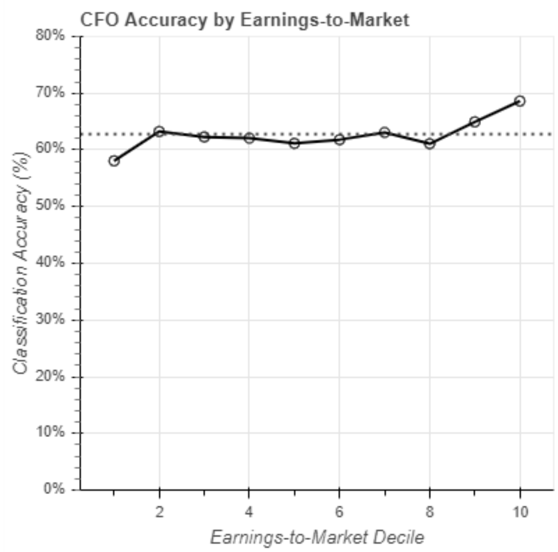


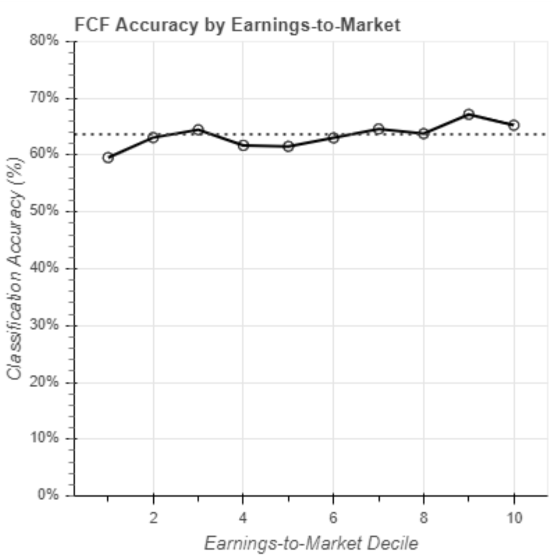
Panel D: Earnings-to-Market Ratio



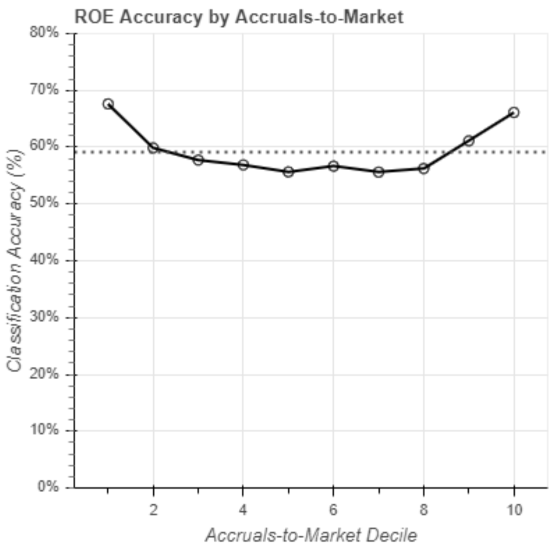


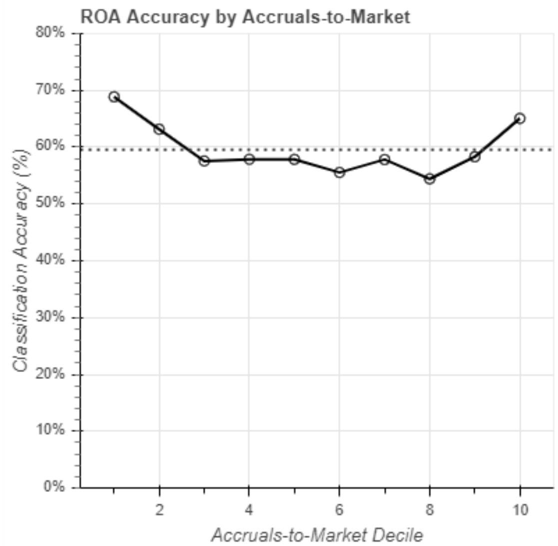


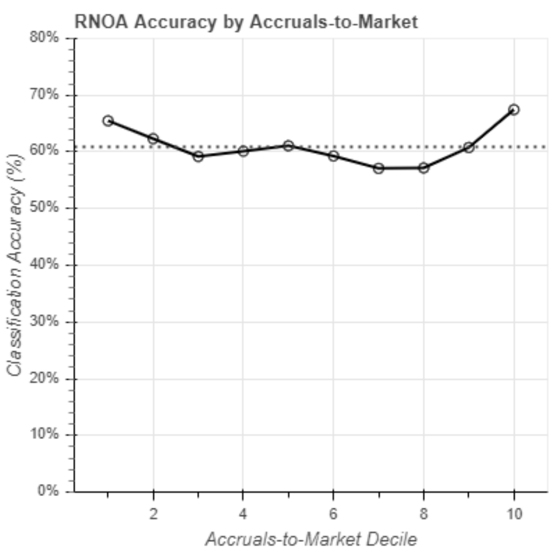


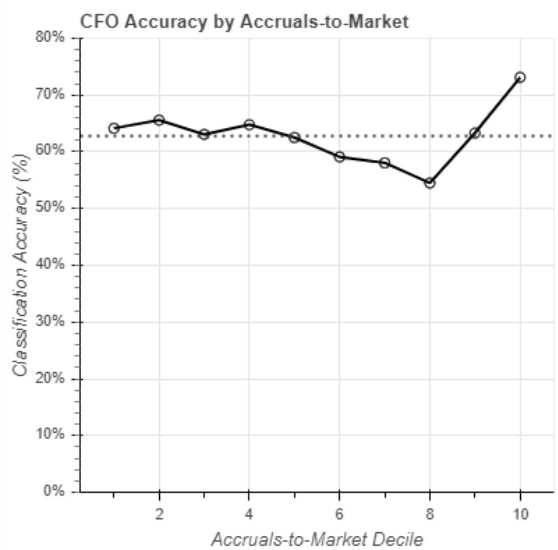


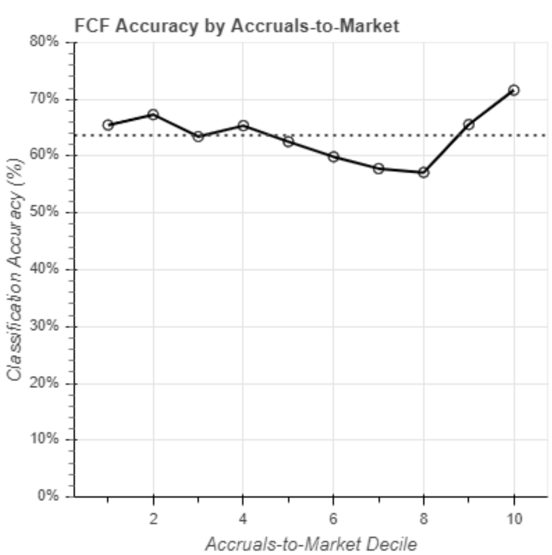
Panel E: Accruals-to-Market Ratio











10 Tables

Table 1: Descriptive Statistics

Panel A: Compustat and Computed Variables (not winsorized or truncated)

	Count	Mean	St. Dev.	Min	1%	25%	50%	75%	99%	Max
					Compustat Variables					
ACT	173,901	548.4	2,823.2	0.0	0.6	15.6	59.8	247.8	8,688.0	169,662.0
AT	190,816	3,719.5	41,519.1	0.0	2.0	33.3	148.4	828.7	50,421.8	2,622,532.0
CAPX	189,067	108.0	688.2	-401.6	0.0	1.0	5.3	31.2	1,943.6	37,985.0
CEQ	190,816	800.0	5,406.0	-96,620.0	-81.2	13.8	61.5	290.6	13,050.5	348,703.0
CHE	190,805	416.2	7,236.0	-9.2	0.0	2.2	12.1	63.7	4,364.1	728,111.0
DLC	190,582	403.5	8,675.1	-882.0	0.0	0.1	2.1	15.6	2,504.2	562,857.0
DLTT	190,410	701.2	7,468.2	0.0	0.0	0.5	12.6	139.7	10,294.6	486,876.0
DP	189,240	79.1	502.2	-7.9	0.0	1.0	4.4	24.6	1,323.6	28,430.0
IB	190,816	92.5	942.6	-99,289.0	-270.8	-0.7	3.2	26.8	2,013.7	59,531.0
IBC	187,537	94.8	959.1	-99,289.0	-274.0	-0.7	3.3	27.4	2,059.6	59,531.0
IDIT	190,816	3.3	31.9	-5.7	0.0	0.0	0.0	0.5	57.0	2,184.9
IVST	190,816	256.8	6,257.8	0.0	0.0	0.0	0.0	7.5	1,679.6	700,280.0
LCT	174,482	402.1	2,838.8	0.0	0.3	6.6	26.3	126.5	6,740.0	329,795.0
LT	190,537	2,880.2	37,223.2	0.0	0.4	12.5	65.2	470.0	36,700.3	2,366,017.0
OANCF	190,816	196.0	1,663.2	-110,560.0	-93.8	0.0	6.9	55.4	3,646.4	129,731.0
OIBDP	189,966	281.8	1,837.4	-76,735.0	-65.3	1.4	13.4	86.3	4,892.5	81,730.0
PPENT	189,699	654.5	4,138.7	0.0	0.0	4.8	25.1	158.6	13,095.1	252,668.0
TXP	181,255	18.2	161.1	-1,346.0	0.0	0.0	0.1	2.6	323.0	12,727.0
XIDOC	184,995	0.9	64.3	-6,343.0	-16.2	0.0	0.0	0.0	21.0	8,035.0
XINT	190,816	50.6	626.1	-0.9	0.0	0.1	1.3	11.1	675.0	57,302.0
					Computed Variables					
ACC	190,816	-101.6	1,302.5	-128,634.0	-2,022.6	-30.5	-4.0	0.1	151.0	125,925.0
NOA	190,165	1,275.1	11,984.9	-630,823.0	-197.9	10.6	63.1	384.6	20,601.9	981,882.0
OI	190,816	139.8	1,190.7	-88,882.0	-192.8	0.0	5.3	38.9	2,576.7	62,771.0

Panel B: Target Variables (not winsorized or truncated)

	count	mean	std	min	1%	25%	50%	75%	99%	max
ROE	183,917	-0.06	22.70	-5,286.00	-3.30	-0.03	0.09	0.16	1.76	4,534.50
ROA	183,920	0.00	0.27	-30.93	-0.97	0.01	0.06	0.10	0.27	22.00
RNOA	183,089	-0.12	34.09	-11,114.80	-5.13	0.02	0.10	0.17	5.01	910.48
CFO	183,920	0.03	0.21	-13.90	-0.77	0.00	0.06	0.12	0.35	3.55
FCF	182,222	-0.03	0.21	-13.90	-0.84	-0.07	0.01	0.06	0.29	3.55

Panel C: Percent increases in target variables in the holdout sample (not winsorized or truncated)

	N	ROE	ROA	RNOA	CFO	FCF
2013	3,127	49.0%	49.6%	47.7%	43.2%	42.5%
2014	3,117	46.5%	45.5%	44.6%	49.2%	51.6%
2015	3,135	47.8%	50.4%	50.0%	50.8%	53.5%
2016	3,081	49.6%	51.0%	50.5%	46.7%	46.6%
2017	2,631	55.4%	59.3%	55.0%	55.4%	54.8%
Overall	**15,091**	**49.4%**	**50.9%**	**49.3%**	**48.9%**	**49.7%**

Panel D: Correlations in Features (Independent Variables)

The variables were truncated at 2.5% on each side before computing the Pearson correlations. In the correlation matrices, Ind. mean (median) is the mean (median) value of the measure for each firm's industry in a given year. Ind. Std. is the standard deviation of the measure for each firm's industry in a given year. Industry is defined by the 4-digit SIC code. We also take the differences between the measure and the industry mean (median) and include those as a separate feature (e.g. ROE – Ind. Mean). The suffixes _L1, _L2, _D1, and _D2 indicate the first and second lags and first and second differences, respectively, of the measure. The suffixes _ACC and _CF are the measure decomposed into its accrual and cash flow components; for example ROE_ACC (ROE_CF) is the accruals (cash flow) component of earnings deflated by common equity.

Return on Equity (ROE)

	ROE	Ind. mean	Ind. Median	Ind. Std.	ROE - Ind. Mean	ROE - Ind. Median	ROE_L1	ROE_D1	ROE_L2	ROE_D2	ROE_ACC	ROE_CF
ROE	1.00	0.28	0.44	-0.14	0.42	0.77	0.71	0.35	0.59	0.46	0.33	0.43
Ind. mean		1.00	0.66	-0.42	-0.75	-0.16	0.21	0.08	0.17	0.13	0.09	0.12
Ind. Median			1.00	-0.31	-0.32	-0.23	0.34	0.12	0.27	0.19	0.15	0.18
Ind. Std.				1.00	0.30	0.07	-0.12	-0.02	-0.11	-0.04	-0.09	-0.02
ROE - Ind. Mean					1.00	0.68	0.29	0.16	0.25	0.20	0.14	0.18
ROE - Ind. Median						1.00	0.53	0.29	0.45	0.36	0.25	0.34
ROE_L1							1.00	-0.41	0.68	0.05	0.21	0.33
ROE_D1								1.00	-0.14	0.54	0.14	0.12
ROE_L2									1.00	-0.44	0.14	0.31
ROE_D2										1.00	0.21	0.14
ROE_ACC											1.00	-0.71
ROE_CF												1.00

Return on Assets (ROA)

	ROA	Ind. mean	Ind. Median	Ind. Std.	ROA - Ind. Mean	ROA - Ind. Median	ROA_L1	ROA_D1	ROA_L2	ROA_D2	ROA_ACC	ROA_CF
ROA	1.00	0.39	0.54	-0.07	0.53	0.75	0.73	0.30	0.63	0.41	0.31	0.49
Ind. mean		1.00	0.77	-0.68	-0.58	-0.15	0.31	0.08	0.27	0.13	0.15	0.17
Ind. Median			1.00	-0.20	-0.23	-0.16	0.44	0.10	0.39	0.16	0.14	0.29
Ind. Std.				1.00	0.56	0.07	-0.05	-0.03	-0.03	-0.05	-0.12	0.06
ROA - Ind. Mean					1.00	0.80	0.36	0.19	0.32	0.24	0.14	0.28

ROA - Ind. Median						1.00	0.51	0.28	0.44	0.35	0.25	0.35
ROA_L1							1.00	-0.43	0.70	0.02	0.19	0.40
ROA_D1								1.00	-0.14	0.51	0.15	0.10
ROA_L2									1.00	-0.45	0.12	0.38
ROA_D2										1.00	0.22	0.12
ROA_ACC											1.00	-0.68
ROA_CF												1.00

Return on Net Operating Assets (RNOA)

	RNOA	**Ind. mean**	**Ind. Median**	**Ind. Std.**	**RNOA - Ind. Mean**	**RNOA - Ind. Median**	**RNOA_L1**	**RNOA_D1**	**RNOA_L2**	**RNOA_D2**	**RNOA_ACC**	**RNOA_CF**
RNOA	1.00	0.17	0.43	-0.04	0.32	0.80	0.72	0.30	0.60	0.41	0.32	0.49
Ind. mean		1.00	0.44	-0.11	-0.88	-0.11	0.13	0.04	0.09	0.08	0.08	0.05
Ind. Median			1.00	-0.11	-0.22	-0.19	0.33	0.10	0.27	0.16	0.18	0.17
Ind. Std.				1.00	0.09	0.04	-0.04	0.01	-0.03	-0.01	-0.08	0.04
RNOA - Ind. Mean					1.00	0.49	0.22	0.11	0.20	0.12	0.07	0.18
RNOA - Ind. Median						1.00	0.56	0.27	0.47	0.34	0.23	0.42
RNOA_L1							1.00	-0.45	0.68	0.01	0.20	0.38
RNOA_D1								1.00	-0.17	0.52	0.13	0.12
RNOA_L2									1.00	-0.48	0.14	0.34
RNOA_D2										1.00	0.19	0.15
RNOA_ACC											1.00	-0.67
RNOA_CF												1.00

Cash Flow from Operations (CFO)

	CFO	**Ind. mean**	**Ind. Median**	**Ind. Std.**	**CFO - Ind. Mean**	**CFO - Ind. Median**	**CFO_L1**	**CFO_D1**	**CFO_L2**	**CFO_D2**	**ROA_ACC**
CFO	1.00	0.39	0.48	-0.01	0.71	0.79	0.61	0.36	0.57	0.39	-0.44
Ind. mean		1.00	0.82	-0.44	-0.38	-0.14	0.28	0.08	0.26	0.11	-0.20
Ind. Median			1.00	-0.09	-0.14	-0.15	0.35	0.10	0.33	0.13	-0.28
Ind. Std.				1.00	0.33	0.05	0.00	-0.02	0.01	-0.02	-0.07
CFO - Ind. Mean					1.00	0.90	0.40	0.29	0.38	0.31	-0.30
CFO - Ind. Median						1.00	0.44	0.33	0.41	0.35	-0.30
CFO_L1							1.00	-0.53	0.53	0.03	-0.15
CFO_D1								1.00	-0.02	0.39	-0.30
CFO_L2									1.00	-0.53	-0.18
CFO_D2										1.00	-0.25
ROA_ACC											1.00

Free Cash Flow (FCF)

	FCF	Ind. mean	Ind. Median	Ind. Std.	FCF - Ind. Mean	FCF - Ind. Median	FCF_L1	FCF_D1	FCF_L2	FCF_D2	ROA_ACC
FCF	1.00	0.40	0.50	-0.07	0.69	0.79	0.56	0.38	0.49	0.42	-0.28
Ind. mean		1.00	0.83	-0.55	-0.38	-0.12	0.29	0.08	0.25	0.11	-0.05
Ind. Median			1.00	-0.20	-0.16	-0.14	0.34	0.11	0.30	0.14	-0.12
Ind. Std.				1.00	0.36	0.06	-0.06	-0.01	-0.06	-0.01	-0.08
FCF - Ind. Mean					1.00	0.90	0.33	0.32	0.29	0.34	-0.24
FCF - Ind. Median						1.00	0.39	0.35	0.34	0.38	-0.23
FCF_L1							1.00	-0.56	0.48	0.01	-0.01
FCF_D1								1.00	-0.05	0.40	-0.27
FCF_L2									1.00	-0.59	-0.01
FCF_D2										1.00	-0.24
ROA_ACC											1.00

Panel E: Correlations in Target Variables

The variables were truncated at 2.5% on each side before computing the Pearson correlations.

	ROE	ROA	RNOA	CFO	FCF
ROE	1.00	0.77	0.57	0.34	0.22
ROA		1.00	0.70	0.44	0.26
RNOA			1.00	0.31	0.20
CFO				1.00	0.71
FCF					1.00

Panel F: Autocorrelations in Target Variables.

The variables were truncated at 2.5% on each side before computing the autocorrelations.

	1 Lag	2 Lags
ROE	0.58	0.44
ROA	0.68	0.55
RNOA	0.47	0.32
CFO	0.59	0.52
FCF	0.55	0.45

Table 2: Tests of Winsorization (Phase 1)

This table reports results of Phase 1 of the research design. For each dependent variable, ROE, ROA, RNOA, CFO, and FCF, we created a random forest using raw, un-winsorized data. We then created two copies of the dataset. In the copies, we winsorized non-categorical variables at 1% and 2.5% on each side, respectively. For each copy of the dataset, we created five random forests, one for each profitability measure. The trees in each forest had three features: PERMNO, SIC, and the contemporaneous value of the dependent variable. For example, to create a forest for ROE, we created a target variable ROE_CHG which was coded as +1 (-1) if ROE for a given firm year increased in the subsequent year.

Before creating each random forest, we partitioned the data into training and testing data. The training data contains approximately 90% of the data (fiscal years 1963 – 2012), and the training data contains fiscal years 2013 – 2017. The forest was trained on the training (in-sample) data. We then tested the performance of the random forest on the testing (out-of-sample) data.

Out-of-Sample Accuracy (Testing) and *In-Sample Accuracy (Training)* are measures of classification accuracy. These metric are computed as (*TI* + *TD*) / (*TI* + *FI* + *TD* + *FD*), where *TI* is the count of true increases (the forest predicted an increase when an increase occurred), *FI* is the count of false increases (the forest predicted a decrease when an increase occurred), *TD* is the count of true decreases, and *FD* is the count of false decreases. We also report the out-of-sample classification accuracy by fiscal year in the holdout data.

The *Random Walk Model* benchmark scores report the expected classification accuracy of two random walk models. Model 1 assumes that increases and decreases in the target variable are equally likely. Model 2 assumes that increases are persistent; if the target measure increases this year, it will increase next year.

For each random forest model, we report the feature importance, which measures the degree to which a feature is used to make predictions. For a given model, feature importance will sum to 100%.

Panels A – E report the results for each of our target variables, ROE, ROA, RNOA, CFO, and FCF, respectively.

Panel A: Return on Equity (ROE)

	Degree of Winsorization		
	None	**1% per side**	**2.5% per side**
Out-of-Sample Accuracy (Testing)	57.1%	57.1%	57.1%
In-Sample Accuracy (Training)	58.8%	58.8%	58.8%
Random Walk Model 1	50.0%		
Random Walk Model 2	45.5%		
Test Scores by Year			
2013	57.9%	58.0%	57.9%
2014	56.6%	56.1%	56.4%
2015	56.4%	56.3%	56.5%
2016	55.8%	56.1%	56.0%
2017	58.9%	59.3%	59.3%
Feature Importance			
PERMNO	28.9%	29.2%	28.6%
Fiscal Year	16.9%	17.0%	17.5%
ROE	54.2%	53.8%	54.0%

Panel B: Return on Assets (ROA)

	Degree of Winsorization		
	None	1% per side	2.5% per side
Out-of-Sample Accuracy (Testing)	57.7%	57.7%	57.6%
In-Sample Accuracy (Training)	58.1%	58.1%	58.1%
Random Walk Model 1	50.0%		
Random Walk Model 2	45.2%		
Test Scores by Year			
2013	57.8%	57.7%	57.7%
2014	59.0%	59.1%	58.8%
2015	56.7%	57.0%	56.8%
2016	56.4%	56.4%	56.7%
2017	58.5%	58.3%	58.3%
Feature Importance			
PERMNO	29.8%	28.8%	30.5%
Fiscal Year	19.4%	20.0%	19.4%
ROA	50.8%	51.2%	50.0%

Panel C: Return on Net Operating Assets (RNOA)

	Degree of Winsorization		
	None	1% per side	2.5% per side
Out-of-Sample Accuracy (Testing)	59.8%	59.8%	59.6%
In-Sample Accuracy (Training)	60.1%	60.1%	60.1%
Random Walk Model 1	50.0%		
Random Walk Model 2	45.6%		
Test Scores by Year			
2013	59.5%	59.5%	59.2%
2014	60.4%	60.2%	60.7%
2015	59.5%	59.5%	58.9%
2016	59.8%	59.8%	59.5%
2017	59.8%	59.9%	59.7%
Feature Importance			
PERMNO	28.2%	28.5%	27.4%
Fiscal Year	16.7%	16.9%	17.5%
RNOA	55.0%	54.6%	55.1%

Panel D: Cash Flow from Operations (CFO)

	Degree of Winsorization		
	None	1% per side	2.5% per side
Out-of-Sample Accuracy (Testing)	57.8%	57.8%	57.8%
In-Sample Accuracy (Training)	62.8%	62.8%	62.8%
Random Walk Model 1	50.0%		
Random Walk Model 2	41.1%		
Test Scores by Year			
2013	58.2%	58.1%	58.2%
2014	58.7%	58.9%	58.6%
2015	57.0%	57.1%	57.0%
2016	58.8%	58.7%	58.9%
2017	55.9%	55.9%	55.9%
Feature Importance			
PERMNO	24.8%	24.9%	25.3%
Fiscal Year	13.1%	13.1%	13.0%
CFO	62.1%	62.0%	61.7%

Panel E: Free Cash Flow (FCF)

	Degree of Winsorization		
	None	1% per side	2.5% per side
Out-of-Sample Accuracy (Testing)	58.5%	58.3%	58.5%
In-Sample Accuracy (Training)	64.7%	64.7%	64.7%
Random Walk Model 1	50.0%		
Random Walk Model 2	40.8%		
Test Scores by Year			
2013	58.1%	57.5%	58.1%
2014	59.6%	59.3%	59.7%
2015	58.3%	58.3%	58.4%
2016	60.0%	59.9%	60.0%
2017	56.1%	56.1%	56.2%
Feature Importance			
PERMNO	21.2%	23.2%	21.4%
Fiscal Year	11.3%	12.1%	11.3%
FCF	67.5%	64.7%	67.3%

Table 3: Effect of Incremental Information on Prediction Accuracy (Phase 2)

This table reports the results of phase 2 of the research design. For each target measure, we created multiple random forests, each with more features than the previous. All random forests were created in the manner described in Table 2, with the same 90/10 training/testing split of the data, and all forests used un-winsorized data. Each column of each table reports the results of a different random forest model. The features used in each model can be ascertained by noting which rows have non-empty feature importance scores in that column. We highlight the model with the highest out-of-sample classification accuracy score.

SIC Code is the 4-digit industry classification code for a firm-year observation. Industry mean / median / stdev is the mean / median / standard deviation of the target variable for that observation's SIC code. Target minus Industry Mean (Median) is the difference between the target variable (e.g. ROE) and the mean (median) value for the industry in that year. For ROE, ROA, and RNOA, we decompose the measure into its accrual and cash flow components and use those components as features. We also include the 1st and 2nd lags and differences.

Panel A: Return on Equity (ROE)

	Analysis								
	2-1a	2-1b	2-1c	2-2a	2-2b	2-2c	2-3a	2-3b	2-3c
Out-of-Sample Accuracy (Testing)	58.5%	59.0%	58.8%	58.8%	59.1%	58.5%	59.0%	58.6%	58.9%
In-Sample Accuracy (Training)	59.2%	59.7%	60.8%	60.2%	60.1%	61.3%	60.2%	60.4%	61.4%
Random Walk Model 1	50.0%								
Random Walk Model 2	45.5%								
Test Scores by Year									
2013	58.4%	59.4%	58.9%	58.2%	58.5%	58.5%	58.7%	58.0%	58.9%
2014	57.7%	57.1%	57.4%	58.3%	58.3%	57.3%	58.7%	58.7%	57.8%
2015	58.2%	59.9%	58.9%	59.3%	60.6%	58.2%	59.9%	59.7%	58.7%
2016	58.6%	59.1%	59.9%	57.5%	58.7%	59.3%	58.6%	59.2%	60.5%
2017	59.6%	59.7%	59.2%	60.9%	59.7%	59.2%	59.0%	57.2%	58.8%
Feature Importance									
PERMNO	20.4%	9.4%	6.9%	13.2%	8.7%	6.8%	9.0%	6.0%	4.8%
Fiscal Year	15.9%	9.6%	8.2%	13.5%	9.4%	8.4%	10.3%	7.1%	6.3%
ROE	44.6%	20.8%	17.9%	34.0%	23.9%	22.5%	22.4%	14.5%	13.2%
SIC Code	19.2%	8.0%	6.1%	13.4%	7.9%	6.5%	8.8%	5.7%	4.6%
Industry mean ROE		8.1%	6.2%		5.3%	4.4%		5.5%	4.5%
Industry median ROE		8.6%	6.6%		6.4%	5.4%		5.8%	4.7%
Industry stdev. ROE		8.5%	6.4%		7.7%	6.1%		5.7%	4.6%
ROE minus Industry Mean		11.2%	9.1%		5.8%	4.5%		7.6%	7.0%
ROE minus Industry Median		15.9%	12.9%		7.5%	5.1%		10.9%	9.8%
ROE: Accrual Component			12.1%			11.5%			9.1%
ROE: Cash Flow Component			7.8%			5.6%			5.7%
ROE: 1st Lag				11.5%	7.5%	6.0%	12.9%	8.2%	6.9%
ROE: 1st Difference				14.4%	9.9%	7.4%	12.7%	8.1%	6.7%
ROE: 2nd Lag							10.5%	6.5%	5.2%
ROE: 2nd Difference							13.4%	8.4%	6.9%

Panel B: Return on Assets (ROA)

	Analysis								
	2-1a	2-1b	2-1c	2-2a	2-2b	2-2c	2-3a	2-3b	2-3c
Out-of-Sample Accuracy (Testing)	59.2%	58.0%	58.9%	59.5%	58.1%	58.6%	59.6%	58.2%	58.5%
In-Sample Accuracy (Training)	58.7%	59.3%	60.0%	59.4%	59.6%	60.3%	59.1%	59.4%	59.9%
Random Walk Model 1	50.0%								
Random Walk Model 2	45.2%								
Test Scores by Year									
2013	58.3%	59.4%	59.3%	58.6%	59.7%	59.6%	60.2%	59.8%	59.6%
2014	59.9%	58.6%	59.4%	60.7%	59.2%	60.4%	60.8%	59.1%	60.3%
2015	59.5%	57.5%	58.2%	59.4%	57.6%	57.5%	59.3%	57.7%	57.9%
2016	58.8%	58.6%	59.9%	58.8%	58.9%	59.2%	58.3%	59.4%	58.8%
2017	59.4%	55.5%	57.6%	59.9%	54.7%	55.8%	59.0%	54.7%	55.7%
Feature Importance									
PERMNO	19.7%	9.6%	7.3%	13.4%	7.1%	5.8%	9.6%	6.0%	5.1%
Fiscal Year	19.4%	10.9%	9.1%	16.5%	9.7%	8.4%	12.2%	8.3%	7.3%
ROA	41.6%	17.7%	14.5%	30.7%	14.7%	12.8%	18.9%	12.2%	10.7%
SIC Code	19.3%	8.0%	6.3%	13.8%	6.5%	5.3%	9.3%	5.6%	4.8%
Industry mean ROA		8.5%	6.8%		6.7%	5.5%		5.8%	4.9%
Industry median ROA		8.9%	7.0%		6.9%	5.7%		6.0%	5.0%
Industry stdev. ROA		9.0%	7.1%		7.2%	5.9%		6.3%	5.3%
ROA minus Industry Mean		12.5%	10.2%		10.6%	9.0%		8.9%	7.9%
ROA minus Industry Median		14.9%	12.0%		12.0%	10.3%		10.0%	8.7%
ROA: Accrual Component			12.0%			10.4%			9.2%
ROA: Cash Flow Component			7.8%			6.3%			5.7%
ROA: 1st Lag				11.9%	9.2%	7.4%	13.0%	7.9%	6.4%
ROA: 1st Difference				13.8%	9.3%	7.3%	12.5%	7.8%	6.5%
ROA: 2nd Lag							10.8%	6.6%	5.4%
ROA: 2nd Difference							13.7%	8.4%	7.0%

Panel C: Return on Net Operating Assets (RNOA)

	Analysis								
	2-1a	2-1b	2-1c	2-2a	2-2b	2-2c	2-3a	2-3b	2-3c
Out-of-Sample Accuracy (Testing)	59.9%	60.3%	60.8%	60.1%	60.1%	60.9%	60.0%	60.3%	60.8%
In-Sample Accuracy (Training)	60.2%	60.8%	61.6%	60.5%	60.5%	61.5%	60.4%	60.6%	61.5%
Random Walk Model 1	50.0%								
Random Walk Model 2	45.6%								
Test Scores by Year									
2013	58.7%	59.2%	59.5%	59.2%	58.8%	59.4%	59.7%	58.7%	60.1%
2014	60.8%	61.9%	61.8%	61.2%	61.8%	61.9%	61.5%	62.5%	62.1%
2015	58.9%	59.8%	59.8%	60.7%	59.6%	60.3%	59.7%	59.8%	59.6%
2016	60.3%	58.9%	59.9%	59.1%	59.6%	60.7%	59.3%	60.1%	60.6%
2017	61.1%	61.7%	63.1%	60.4%	61.1%	62.3%	59.7%	60.4%	62.0%
Feature Importance									
PERMNO	18.6%	9.0%	6.5%	11.7%	8.4%	6.5%	8.3%	5.5%	5.0%
Fiscal Year	16.8%	10.0%	8.4%	13.3%	9.7%	8.9%	11.1%	7.6%	6.5%
RNOA	46.5%	20.5%	17.7%	29.5%	22.4%	21.5%	23.1%	14.5%	12.0%
SIC Code	18.1%	7.8%	6.1%	11.6%	7.5%	6.1%	8.5%	5.4%	4.8%
Industry mean RNOA		8.2%	6.3%		5.3%	4.1%		5.5%	4.8%
Industry median RNOA		8.5%	6.5%		6.7%	5.4%		5.7%	5.0%
Industry stdev. RNOA		8.3%	6.3%		7.8%	6.4%		5.5%	4.9%
RNOA minus Industry Mean		11.2%	9.1%		5.8%	4.5%		7.4%	6.6%
RNOA minus Industry Median		16.6%	13.4%		7.2%	6.0%		11.8%	9.6%
RNOA: Accrual Component			11.4%			10.0%			8.4%
RNOA: Cash Flow Component			8.3%			5.4%			6.1%
RNOA: 1^{st} Lag				16.5%	7.9%	6.4%	13.2%	8.2%	7.0%
RNOA: 1^{st} Difference				17.4%	11.2%	8.8%	12.9%	8.8%	7.1%
RNOA: 2^{nd} Lag							10.0%	6.3%	5.4%
RNOA: 2^{nd} Difference							12.8%	7.8%	6.6%

Panel D: Cash Flow from Operations (CFO)

	Analysis								
	2-1a	2-1b	2-1c	2-2a	2-2b	2-2c	2-3a	2-3b	2-3c
Out-of-Sample Accuracy (Testing)	60.0%	61.1%	62.3%	61.6%	61.3%	62.8%	61.9%	61.8%	62.8%
In-Sample Accuracy (Training)	63.7%	64.5%	65.5%	65.2%	65.5%	66.2%	65.5%	65.8%	66.5%
Random Walk Model 1			50.0%						
Random Walk Model 2			41.1%						
Test Scores by Year									
2013	61.7%	62.0%	63.0%	63.5%	63.7%	64.8%	64.4%	64.8%	64.8%
2014	60.0%	62.7%	63.7%	61.6%	62.9%	64.2%	62.9%	63.1%	63.9%
2015	58.8%	61.2%	63.4%	61.1%	61.0%	63.0%	60.7%	61.4%	62.7%
2016	60.7%	60.9%	61.9%	63.0%	62.0%	63.0%	63.8%	62.2%	64.8%
2017	58.8%	58.0%	59.0%	58.1%	56.6%	58.3%	57.4%	56.8%	57.2%
Feature Importance									
PERMNO	17.3%	7.6%	6.4%	9.9%	5.9%	5.0%	7.2%	4.8%	4.6%
Fiscal Year	12.9%	6.0%	5.6%	7.9%	5.0%	4.4%	6.1%	4.1%	4.0%
CFO	52.4%	24.5%	18.7%	34.9%	19.0%	15.1%	26.8%	16.1%	12.4%
SIC Code	17.5%	6.8%	6.3%	10.2%	5.3%	4.8%	7.2%	4.4%	4.3%
Industry mean CFO		7.5%	6.3%		6.0%	5.0%		5.0%	4.7%
Industry median CFO		7.1%	6.1%		5.9%	5.1%		5.0%	4.6%
Industry stdev. CFO		7.2%	6.4%		6.0%	5.2%		5.0%	4.7%
CFO minus Industry Mean		14.0%	11.9%		11.2%	9.7%		9.6%	8.1%
CFO minus Industry Median		19.2%	14.2%		13.7%	12.2%		12.6%	10.5%
ROA: Accrual Component			18.1%			15.4%			12.3%
CFO: 1st Lag				13.0%	7.2%	6.0%	9.3%	5.7%	5.2%
CFO: 1st Difference				24.2%	14.8%	12.2%	18.2%	12.0%	9.8%
CFO: 2nd Lag							8.9%	5.5%	5.2%
CFO: 2nd Difference							16.2%	10.3%	9.5%

Panel E: Free Cash Flow (FCF)

	Analysis								
	2-1a	2-1b	2-1c	2-2a	2-2b	2-2c	2-3a	2-3b	2-3c
Out-of-Sample Accuracy (Testing)	60.5%	61.1%	62.6%	62.5%	62.4%	63.5%	62.5%	63.3%	63.7%
In-Sample Accuracy (Training)	65.4%	65.9%	66.4%	66.0%	66.4%	66.9%	66.4%	66.6%	67.1%
Random Walk Model 1			50.0%						
Random Walk Model 2			40.8%						
Test Scores by Year									
2013	60.8%	62.0%	64.2%	65.5%	64.8%	65.6%	65.0%	65.4%	66.2%
2014	61.4%	63.1%	63.9%	63.3%	64.3%	65.5%	63.4%	64.3%	64.7%
2015	60.3%	60.8%	62.9%	63.4%	62.7%	63.3%	62.4%	63.3%	63.5%
2016	61.0%	62.2%	62.7%	62.2%	62.3%	63.4%	63.4%	63.6%	64.7%
2017	58.5%	57.0%	58.6%	57.7%	57.3%	58.9%	57.9%	59.1%	58.7%
Feature Importance									
PERMNO	16.4%	6.9%	5.6%	9.4%	5.7%	4.9%	6.9%	4.2%	3.7%
Fiscal Year	11.9%	6.2%	5.5%	8.3%	5.4%	5.0%	6.5%	4.1%	3.9%
FCF	55.8%	26.8%	24.2%	38.5%	22.1%	19.4%	30.0%	20.7%	17.7%
SIC Code	16.0%	6.1%	5.0%	9.8%	5.2%	4.5%	7.1%	3.9%	3.4%
Industry mean FCF		7.0%	5.8%		6.0%	5.2%		4.7%	4.2%
Industry median FCF		6.7%	5.6%		5.9%	5.1%		4.7%	4.1%
Industry stdev. FCF		6.5%	5.4%		5.6%	4.9%		4.3%	3.8%
FCF minus Industry Mean		14.2%	12.8%		11.3%	10.5%		10.8%	9.6%
FCF minus Industry Median		19.7%	15.4%		14.7%	13.4%		13.9%	12.4%
ROA: Accrual Component			14.5%			12.2%			10.7%
FCF: 1st Lag				12.4%	6.4%	5.5%	8.9%	4.8%	4.2%
FCF: 1st Difference				21.7%	11.6%	9.4%	15.5%	9.2%	7.8%
FCF: 2nd Lag							8.7%	4.9%	4.3%
FCF: 2nd Difference							16.5%	9.8%	10.1%

Table 4: Conditional Analysis (Phase 3)

This table reports the results of phase 3 of the research design. For each firm-year observation in the testing sample (fiscal years 2013 – 2017), we computed the book-to-market ratio, market capitalization, debt-to-assets ratio, earnings-to-market ratio, and accruals-to-market ratio. Book-to-market ratio is computed as common equity (Compustat: CEQ) divided by market capitalization. Market capitalization is computed as closing price at the end of the fiscal year (PRCC_F) times common shares outstanding (CSHO). Debt-to-assets ratio is computed as [long-term debt (DLTT) plus debt in current liabilities (DLC) plus preferred stock (PSTK)] divided by total assets (AT). Earnings-to-market ratio is computed as income before extraordinary items (IB) divided by market capitalization. Accruals-to-market ratio is computed as accruals (as defined in section 3) divided by market capitalization.

Within each fiscal year, and for each conditional variable, we formed deciles of each conditioning variable. For each decile of each conditioning variable, we generated fitted values for each target variable (ROE, ROA, RNOA, CFO, and FCF). We used the best-performing random forest grown in phase 2 to generate the fitted values. As shown in Table 3, the best-performing random forest for ROE was grown in Analysis 2-2b; the best performing random forests for ROA, RNOA, CFO, and FCF were grown in Analyses 2-3a, 2-2c, 2-3c, and 2-3c, respectively.

We used the fitted values to compute the classification accuracy within each decile of each conditioning variable for each profitability measure (target variable). We then averaged the classification accuracy across all 5 years of the testing data. In each table below, *Accuracy* is the classification accuracy (% of changes in the target variable predicted correctly) for the given decile (row) of the target variable (column). *Decile Mean* is the average value of conditioning variable for that decile. Highlighted cells show the profitability measure (column) with the highest classification accuracy in each decile (row). Panels A – E report the results for book-to-market ratio, market capitalization, debt-to-assets ratio, earnings-to-market ratio, and accruals-to-market ratio, respectively.

Panel A: Book-to-Market Ratio

Decile	ROE Accuracy	ROA Accuracy	RNOA Accuracy	CFO Accuracy	FCF Accuracy	Decile Mean
1	65.1%	58.8%	62.3%	58.3%	59.0%	-0.401
2	51.8%	56.5%	60.3%	57.6%	59.1%	0.134
3	55.4%	56.9%	60.8%	59.4%	61.8%	0.219
4	58.1%	60.4%	61.4%	62.3%	64.7%	0.299
5	60.6%	59.2%	61.7%	64.2%	63.3%	0.390
6	57.5%	59.8%	60.1%	65.2%	64.1%	0.490
7	61.8%	62.5%	62.8%	62.7%	62.9%	0.606
8	60.8%	61.5%	61.0%	63.7%	64.9%	0.742
9	63.2%	60.0%	59.1%	67.3%	66.4%	0.931
10	58.3%	59.3%	60.2%	65.7%	68.3%	1.846

Panel B: Market Capitalization

Decile	ROE Accuracy	ROA Accuracy	RNOA Accuracy	CFO Accuracy	FCF Accuracy	Decile Mean ($mil)
1	59.3%	61.2%	61.6%	61.7%	62.6%	23
2	62.5%	59.4%	62.3%	66.8%	66.9%	71
3	60.1%	57.8%	60.0%	64.9%	64.0%	158
4	60.8%	59.6%	60.2%	63.2%	64.3%	307
5	60.9%	60.3%	62.1%	63.1%	64.3%	562
6	60.9%	60.2%	60.2%	64.4%	65.6%	990
7	62.2%	60.8%	63.0%	62.9%	63.6%	1,737
8	58.1%	58.6%	60.5%	62.6%	64.3%	3,202
9	55.4%	57.9%	59.7%	60.3%	62.0%	7,170
10	53.7%	59.8%	60.0%	58.8%	58.9%	53,322

Panel C: Debt-to-Assets Ratio

Decile	ROE Accuracy	ROA Accuracy	RNOA Accuracy	CFO Accuracy	FCF Accuracy	Decile Mean
1	56.5%	55.5%	60.8%	61.6%	61.7%	0.000
2	57.3%	58.1%	61.4%	63.9%	62.8%	0.004
3	59.7%	59.3%	61.1%	61.4%	64.2%	0.034
4	59.2%	59.9%	59.3%	64.3%	62.7%	0.085
5	60.0%	61.1%	61.1%	65.1%	65.8%	0.141
6	60.5%	60.2%	61.7%	62.6%	65.0%	0.207
7	59.0%	59.6%	60.5%	64.1%	63.9%	0.278
8	58.8%	62.2%	62.7%	64.8%	64.8%	0.355
9	56.2%	59.0%	60.5%	60.5%	62.8%	0.463
10	66.3%	61.6%	61.0%	59.9%	61.9%	0.826

Panel D: Earnings-to-Market Ratio

Decile	ROE Accuracy	ROA Accuracy	RNOA Accuracy	CFO Accuracy	FCF Accuracy	Decile Mean
1	66.0%	67.5%	67.0%	58.1%	59.6%	-1.027
2	60.6%	59.6%	63.0%	63.2%	63.1%	-0.160
3	59.2%	60.3%	62.8%	62.3%	64.5%	-0.048
4	58.2%	58.8%	60.0%	62.1%	61.7%	0.002
5	51.4%	51.0%	55.3%	61.2%	61.6%	0.024
6	51.4%	53.4%	54.0%	61.8%	63.0%	0.037
7	54.9%	54.0%	58.0%	63.1%	64.6%	0.047
8	55.3%	57.8%	58.2%	61.1%	63.8%	0.056
9	61.3%	60.7%	59.7%	64.9%	67.2%	0.069
10	75.2%	73.7%	72.8%	68.6%	65.3%	0.157

Panel E: Accruals-to-Market Ratio

Decile	ROE Accuracy	ROA Accuracy	RNOA Accuracy	CFO Accuracy	FCF Accuracy	Decile Mean
1	67.6%	68.9%	65.5%	64.1%	65.5%	-0.969
2	59.9%	63.2%	62.3%	65.6%	67.3%	-0.172
3	57.7%	57.6%	59.2%	63.1%	63.4%	-0.098
4	56.9%	57.9%	60.1%	64.8%	65.4%	-0.066
5	55.6%	57.8%	61.1%	62.5%	62.6%	-0.046
6	56.7%	55.6%	59.3%	59.1%	59.9%	-0.033
7	55.6%	57.8%	57.1%	58.1%	57.8%	-0.022
8	56.2%	54.4%	57.2%	54.5%	57.1%	-0.012
9	61.1%	58.3%	60.8%	63.3%	65.6%	0.001
10	66.1%	65.1%	67.5%	73.1%	71.6%	0.206