Can Machines Learn Capital Structure?

Shahram Amini[a], Ryan Elmore[a], and Jack Strauss[*,a]

[a]*Daniels College of Business, University of Denver*

October 28, 2019

Abstract

Yes, they can! Machine learning models that exploit big data can consistently predict corporate leverage better than classical methods over time and identify its determinants. Using a large sample of U.S. firms from 1972 to 2018, we apply random forests, gradient boosting machines, neural networks, and generalized additive models to predict corporate leverage and analyze its determinants. Results show that machine learning models that allow for nonlinearities and complex interactions boost the out-of-sample R^2 from 36% to 56% over linear methods such as LASSO. The superior predictive performance occurs every year of the out-of-sample period at the aggregate level as well as subsamples such as firms undergoing a major capital restructuring. Additionally, machine learning methods consistently identify the determinants of corporate leverage over time. Our best performing model, a random forest, selects market-to-book, industry median leverage, cash & equivalents, Z-Score, profitability, stock returns, and firm size as robust and reliable predictors of market leverage. These findings suggest that despite leverage climbing to record highs, firm fundamentals drive corporate leverage because the determinants and predictability of corporate leverage have remained stable. Our work offers promise for applying machine learning models to the repertoire of capital structure studies seeking to predict corporate leverage with high precision; further, machine learning methods can change our perception of what variables are reliably important.

JEL Code: G0, G17, G30, G32, C10, C50

Keywords: Machine Learning, Capital Structure, Corporate Leverage, Random Forest, Gradient Boosting Machine, Neural Network, Generalized Additive Model, LASSO

[*]Shahram Amini (shahram.amini@du.edu, 303-871-2703), Ryan Elmore (ryan.elmore@du.edu, 303-871-3026), and Jack Strauss (jack.strauss@du.edu, 303-871-3504) are from the Daniels College of Business at the University of Denver. Corresponding author: Shahram Amini.

1. Introduction

Changes in corporate leverage convey new information to the market about the prospects of the firm. These innovations can affect stock returns, and consequently firm value, by increasing the risk of cash flow to equity holders as well as by influencing investment decisions leading to higher business risk (Ozdagli, 2012). As a result, finding a reliable set of factors that accurately and consistently predict a firm's capital structure is critical to both managers and outside investors, and has been the focus of a vast number of empirical studies over decades (see Titman and Wessels, 1988; Harris and Raviv, 1991; Rajan and Zingales, 1995; Frank and Goyal, 2009; Öztekin, 2015, among others). Despite extensive debate and empirical analysis, however, there is no consensus on which factors robustly determine corporate leverage, and whether these factors evolve over time.

Our work employs a series of machine learning models to evaluate the importance of potential determinants of corporate leverage suggested in the literature, and to compare the predictive performance of these models to those of conventional models. More specifically, our study applies random forests (RF), gradient boosting machines (GBM), neural networks (NNET), generalized additive models (GAM), and least absolute shrinkage and selection operator (LASSO) models to both predict capital structure and analyze its determinants. Using firm-level data from 1972 to 2018, we document substantial gains in predictability measured by out-of-sample R-squared (R^2_{OS}) in forecasting corporate leverage; e,g., R^2_{OS} increases from 36.2% for LASSO to 56.2% for RF. We also identify the most informative predictors of capital structure. To avoid omitted variable bias in estimating a model, including the appropriate set of variables is critical. Our results show that machine learning models draw information from a broader set of characteristics compared to prior studies (e.g., Rajan and Zingales, 1995; Frank and Goyal, 2009).

We believe that machine learning methods are particularly effective for predicting leverage, evaluating its determinants, and assessing the stability of the determinants over time for several reasons. First, accurately evaluating the extent of predictability of corporate leverage has implications for stock returns and firm value. Gomes and Schmid (2010) and Bhamra et al. (2010) show that leverage and asset returns are linked through complex interactions related to growth, equity issuance cost, and macroeconomic risk. Machine learning models have been utilized to improve forecasting accuracy relative to conventional statistical models by leveraging the nonlinear and collinear relationships that are often inherent in large data sets. Hence, because machine learning methods capture hidden interactions and avoid multicollinearity, they are likely better able to predict corporate leverage than conventional linear models.

Second, nonlinearities are relevant as well as intrinsic in the relationship between capital structure and its predictors. Graham and Leary (2011) find that common variables used to determine capital structure appear to have nonlinear relations with the dependent variables, and yet few empirical tests explicitly model these nonlinearities for prediction or quantifying a determinant's importance. Machine learning models take into account nonlinearities and complex interactions between predictors, which are endemic in affecting a firm's financing and borrowing decisions (Childs et al., 2005).

Third, in the presence of a large number of potential determinants of corporate leverage (Harris and Raviv, 1991), a few studies use linear model selection techniques to parse through the pool of variables that empirically test capital structure theories (e.g., Frank and Goyal, 2009). Machine learning estimators, on the other hand, are equipped with efficient algorithms that are capable of searching among the entire model space and removing the redundant variation among potentially highly correlated predictors (Gu et al., 2019). Further, nonlinearities and complex interactions may affect the variable selection process; thus, linear models by ignoring complex interactions among variables and nonlinear effects may lead to biases in their identification of the relevant predictors.

Given these features of machine learning estimators, the goals of this paper are threefold: 1) implementing out-of-sample predictions for corporate leverage and estimating the gain in accuracy when machine learning estimators are applied, 2) using machine learning models to identify the robust determinants of corporate leverage, and 3) assessing the stability of the determinants and predictability over time (Peyer and Shivdasani, 2001; Frank and Goyal, 2009). This paper applies machine learning models to a sample comprising 233,225 firm-year observations from 1972–2018. Following Gu et al. (2019), we choose RF, GBM and NNET, and compare these models to LASSO and OLS linear benchmark models. We also add GAM, since our focus is on nonlinear models. We use a training and cross-validation period from 1972–2000 to estimate parameters and tune the machine learning models, and then assess out-of-sample root mean squared forecast error (RMSE) and R^2_{OS} from 2001–2018.

A brief preview of our results shows that nonlinear machine learning methods achieve substantial gains in predictive performance relative to linear OLS and LASSO models. Aggregate R^2_{OS} statistics are 39.8% and 36.2% for OLS and LASSO over the testing period (2001–2018) compared to R^2_{OS} statistics of 56.2%, 55.3%, 52.0%, and 46.6% for RF, GBM, NNET, and GAM, respectively. On a year-to-year analysis, RF's R^2_{OS} range from 45–58% over the testing period while OLS and LASSO methods R^2_{OS} range from 28–44% and 30–39%. Additionally, a plot of differences in cumulative sum of squared errors highlights that RF and GBM particularly outperform the linear model, and lead to lower errors every year in the 2001–2018 out-of-sample period. We show further that machine learning models significantly reduce the RMSE. For example, RF leads to a 14.7% decline in RMSE compared to the OLS model and a more than 15% decline compared to LASSO. The differences in RMSE values are significant at the 1% level. Thus, our findings collectively imply that machine learning models more accurately predict leverage.

Throughout the paper, we focus particularly on RF, since this method excels at modeling linear combinations of a large number of covariates with complex interactions and outperforms linear models when covariate shifts are likely. RF's ensemble nature and random element correct for decision trees' habit of overfitting to their training set even when the process is nonlinear or involves complex high-order interaction effects (Strobl et al., 2008; Altman and Krzywinski, 2017). Thus, we believe that machine learning models, such as RF can capture the data generating process of corporate leverage better than conventional models.

We then turn to identifying the determinants of corporate leverage. RF and GBM both show that market-to-book, industry median leverage, cash, Altman's (1968) Z-Score, profitability, stock returns, and firm size are the primary determinants of market-based corporate leverage. In contrast, NNET highlights the importance of research and development expenditure, profitability, cash, stock variance, and market-to-book in determining market-based leverage. LASSO demonstrates that industry median leverage and cash dominate as determinants of corporate leverage, and growth in GDP, profitability, and market-to-book as less influential determinants of leverage; hence, by ignoring nonlinearities and complex interactions, LASSO selects fewer variables and exhibits poor out-of-sample performance. Overall, while our results provide evidence for the prominence of the core leverage factors identified in Frank and Goyal (2009) and Öztekin (2015), we further demonstrate the importance of additional factors as well. These variables include risk factors proxied by Z-Score, stock market conditions proxied by cumulative annual stock returns and annual stock variance, liquidity proxied by cash, and discretionary expenditures proxied by R&D, and selling, general, and administrative cost (SGA).

Next, we investigate the stability of predictability and variable importance over time; this is relevant as Graham and Leary (2011) demonstrate that there has been considerable variation in the predictors of leverage across time. Since the determinants of leverage may evolve over time due to regulations and tax changes or business cycles, our paper further focuses on whether machine learning models can consistently predict leverage and outperform linear models. This analysis is particularly timely due to current record-high leverage among U.S. firms. We find no evidence of changes in firm determinants or a breakdown in predictability. The same seven variables selected by RF or GBM consistently are important over time in determining and predicting corporate leverage. The stability of both determinants and predictability suggests that there is no mismatch between firm fundamentals and leverage. This implies that today's high leverage is driven by firm factors, and does not exhibit traits of overborrowing.

We also investigate whether the superior predictive performance of machine learning methods carries over into subsamples constructed based on different firm attributes, including size, growth opportunities, information opacity, and whether the firm is implementing a major capital restructuring. The results show that RF outperforms linear models in predicting the capital structure of small and large firms, value and growth firms, and high-tech and non high-tech firms. Further, RF substantially outperforms linear models in predicting the debt ratio of firms that are moving towards their optimal targets. More specifically, we examine times of significant capital restructuring when both the net debt issuance and net payouts in a year increase by at least 3% relative to total assets. Compared to a linear model, the gains for predicting leverage for the refinancing firms range from R^2_{OS} of 9% in 2009 to 36% in 2012. This implies that machine learning models more accurately predict leverage for firms with different size, growth and refinancing characteristics.

The main results in this study are based on market leverage defined as total debt (the sum of short-term and long-term debt) scaled by the market value of assets (TDM). As a robustness check, we also provide results for book leverage defined as total debt scaled by total assets (TDA), long-term debt scaled by total assets (LDA), and long-term debt scaled by the market value of assets (LDM). Results for book leverage and market leverage are relatively similar, except cash and Z-Score rank more important for determining book leverage. Further, RF and GBM also possess higher R^2_{OS} than OLS and LASSO for TDA, LDA and LDM every year of the out-of-sample period;

their superior predictability ranges from 8%–27%.

Our study contributes to the expansive literature on the determinants of capital structure. Frank and Goyal (2009) apply a linear model selection method using Akaike information criterion (AIC) and Bayesian information criterion (BIC) to identify the determinants of corporate leverage. From a large pool of potential candidates, the study finds support for industry median leverage, market-to-book, profitability, tangibility, firm size, and expected inflation as the key factors for U.S. firms. Öztekin (2015) extends this study by using a large sample of international data and linear models and finds similar results. Both studies, however, do not consider nonlinearities that are present in the relation between corporate leverage and its determinants. We complement their analysis by using machine learning models that are able to more accurately model the complex and nonlinear determinants of corporate leverage with better statistical properties. As a result, our results exhibit a substantial gain in the accuracy of predicting corporate leverage.

Our work also adds to a new and growing literature that applies machine learning estimators to various areas of research in finance. Gu et al. (2019), for instance, apply machine learning methods to predict stock returns as well as the most informative predictors of the returns. Han et al. (2018) introduce a combination LASSO method to improve out-of-sample forecasts of cross-sectional expected stock returns. With the help of machine learning algorithms, Brogaard and Zareei (2018) aim to detect mispricing in the stock market using past prices while controlling for data snooping. Erel et al. (2018) exploit machine learning's superior predictive performance to determine which potential director would be the best fit for a given firm. Li et al. (2018) use machine learning techniques such as a neural network model to calculate the association between words appearing in earnings call transcripts in order to quantify corporate culture. Our study contributes to this growing literature by using machine learning models to predict managers' financial leverage policies, and to identify the determinants of leverage after allowing for possible nonlinearities and complex interactions between predictors.

The remainder of this paper is organized as follows. Section 2 outlines the econometric methodology of the paper, detailing our machine learning methods. Section 3 presents the sample selection and summary statistics. Section 4 presents the main results. Section 5 offers additional tests of attributes including size and market-to-book, while Section 6 concludes.

2. Methodology

In this section, we describe the specific models employed in this study for predicting capital structure, including our strategy for training the machine learning methods.

The basic regression problem in predicting capital structure is to estimate a function $g(\mathbf{x}_{i,t}) = E(y_{i,t+1}|\mathbf{x}_{i,t})$ where

$$y_{i,t+1} = g(\mathbf{x}_{i,t}) + \varepsilon_{i,t+1} \quad (1)$$

is the i^{th} firm's leverage ratio in year t and $\varepsilon_{i,t}$ is a random error component. The regression function $E(y_{i,t+1}|\mathbf{x}_{i,t})$ is the conditional expectation of $y_{i,t+1}$ conditioned on the vector of covariates described in the next section. As the focus of this paper is on the prediction of $y_{i,t+1}$, we will omit any discussion involving assumptions on the random errors, ε_i.

Our goal is to estimate the function $g(\cdot)$ using various methods including standard multiple linear regression to highly nonlinear, discontinuous methods such as a random forest. The first class of models presented in Section 2.1 are variants of a standard linear model. Note, however, that they are not necessarily linear in the covariates. In fact, the generalized additive model (GAM) in Section 2.1.3 can be used to estimate highly nonlinear relationships between the covariates and the response of interest using splines or any suitable nonparametric function estimation tool. The second class of models in Section 2.2, referred to here as machine learning prediction functions, allows for nonlinear and discontinuous relationships between y_i and the associated covariates. Further, these machine learning models may capture complex interaction effects that are present in the covariates. See Hastie et al. (2009) and Kuhn and Johnson (2013) for additional information related to all of the models considered here.

2.1. Linear Models

2.1.1. Multiple Regression Model

Our baseline model for comparison purposes is the standard, multiple regression model of the form

$$g(\mathbf{x}_{i,t}; \boldsymbol{\beta}) = \mathbf{x}'_{i,t}\boldsymbol{\beta}. \quad (2)$$

The regression parameters $\boldsymbol{\beta}$ are estimated using ordinary least squares (OLS) and are defined by

$$\hat{\boldsymbol{\beta}}^{ols} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{x}\boldsymbol{\beta}\|_2^2, \tag{3}$$

where $\|\mathbf{a} - \mathbf{b}\|_2$ is the distance (Euclidean) between the vectors $\mathbf{a}$ and $\mathbf{b}$, *i.e.*, the l_2-norm.

2.1.2. LASSO

Drawbacks to using a standard multiple regression model in this scenario are overfitting and the potential of multicollinearity given a large number of predictor variables. In order to mitigate these concerns, we will utilize the least absolute shrinkage and selection operator model, also known as the LASSO (Tibshirani, 1996). This work demonstrates that LASSO is an effective method in shrinking parameters associated with insignificant covariates to zero. Thus, the LASSO acts as both a shrinkage *and* model selection tool, and the end result is a sparse version of the standard multiple regression model defined in Equation (2). As a result, LASSO should be able to select the appropriate variables in a more general capital structure model, by including only a subset of the original predictor variables that are important for predicting corporate leverage.

The parameter estimates associated with the LASSO model for a given value of λ is defined by

$$\hat{\boldsymbol{\beta}}_{\lambda}^{lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{x}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{4}$$

for some $\lambda > 0$. The effect of λ and the l_1-penalty term ($\lambda\|\boldsymbol{\beta}\|_1 = \lambda\sum_{j=1}^{p}|\beta_j|$) is straightforward. If $\lambda = 0$, then $\hat{\boldsymbol{\beta}}^{ols} = \hat{\boldsymbol{\beta}}^{lasso}$ and as $\lambda \to \infty$, the penalty term forces the coefficients to zero. Data-driven choices of the tuning parameter λ are usually obtained using some form of cross validation, as we discuss in Section 2.3.

2.1.3. Generalized Additive Model

Unfortunately, the OLS and LASSO models are not well equipped to adequately model nonlinear relationships between the covariates and the response without pre-specified manual intervention. The generalized additive models, or GAM, is a family of nonparametric models that can handle the nonlinearities that may be present (James et al., 2013). In the case of a GAM, the regression function $g(\cdot)$ defined in Equation (1) can be written as

$$g(\mathbf{x}_{i,t}) = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij,t}) \tag{5}$$

for smooth, nonlinear functions f_j. Spline functions are often used in estimating the functions f_j, however, almost any standard nonparametric function estimator can be used, e.g., local polynomial regressions estimators.

One of the main advantages of using a GAM as opposed a classic machine learning algorithm for modeling nonlinear relationships is that the additive, functional effects of the covariates x_j on y are readily available and, in fact, are often interpretable. Therefore, when interpretability and predictive accuracy are both desired, a GAM is a highly attractive option. It should also be noted that some people might refer to both LASSO and GAM as machine learning methods rather than a variant of a linear model.

2.2. Machine Learning Models

Given that prediction and identification of variable importance are our central goals in this research and that nonlinear models might be more apt in the capital structure domain, we investigate the utility of several popular machine learning (ML) models in this scenario. As mentioned in the introduction, ML algorithms are becoming increasingly popular in, for example, asset pricing, and as prediction tools in general. They are particularly useful in modeling nonlinear relationships among dependent and independent variables, as well as capturing hidden interactions in these variables. And, although nonlinear relationships and interaction effects can be modeled using the methods described in Section 2.1, they must be noted *a priori*. In contrast, the ML methods described here are fully nonparametric and sufficiently flexible to capture these complex structures.

The specific ML models utilized here and described in Sections 2.2.1, 2.2.2, and 2.2.3 are a random forest (Breiman, 2001), a gradient boosting machine (Friedman, 2001), and a feed-forward neural network with a single hidden layer (Venables and Ripley, 1997), respectively. Additional algorithmic details related to these methods can be found in Kuhn and Johnson (2013) and James et al. (2013). To be consistent with the ML literature, we adopt their terminology and refer to independent variables as "features" and the collection of feature vectors as the feature space.

2.2.1. Random Forests

In order to motivate a random forest (RF), we first describe two key aspects of a simple regression tree. First, suppose we split the feature space $\mathcal{X}$ into J unique, non-overlapping regions: $R_1, R_2, \ldots, R_J$. The predicted value of y for any value within R_j is simply the average overall response values in R_j, or more formally:

$$\hat{g}^{rf}(\mathbf{x}) = \sum_{j=1}^{J} \bar{y}_j \mathbb{1}_{\{\mathbf{x} \in R_j\}}. \tag{6}$$

The residual sum of squares (RSS) can be then easily computed. Next, we describe the process of "growing" a tree informally and point the interested reader to James et al. (2013) or Hastie et al. (2009) for more formal treatments including the popular *recursive binary partitioning* algorithm. We initially place all observations into the same bin so that every predicted value will be $\bar{y}$. We then find a cutpoint t_o in our feature space, resulting in two non-overlapping regions, that provides the maximal reduction in RSS. Next, we find a cutpoint in one of the two regions (resulting in three regions) such that the RSS is reduced by the maximum amount. The process is repeated until a suitable limit is reached, e.g., only a marginal reduction in the RSS is attained at further splits.

Unfortunately, predictions made from growing a single tree are notorious for exhibiting high variance. That is, predictions made from a tree may change substantially (highly variable) from sample to sample. The method of bootstrap aggregation, or bagging, is often employed in order to alleviate this potential problem. In a regression context, bagging starts by taking a bootstrap (Efron and Tibshirani, 1994) sample and growing a regression tree on this sample. As is usually the case in utilizing a bootstrap methodology, a large number of bootstrap samples are generated, say B, and trees are grown on each sample. This results in an ensemble (or forest) of trees from which to make predictions. The bagged estimate at $\mathbf{x}$ is the average estimate over all trees

$$\hat{g}^{bag}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{g}^{b}(\mathbf{x}), \tag{7}$$

where $\hat{g}^b$ is the estimator defined in Equation (6) on the b^{th} bootstrap sample. The practical effect of averaging over this ensemble is to reduce the variation in the final estimate. See Breiman (1996) for additional details related to bagging.

The variance of the bagged estimate in Equation (7), however, can still be higher than desired because we are essentially averaging over (potentially) highly correlated estimates. Each tree, when using the entire collection of features, will tend to be correlated to the trees grown from the other bootstrap samples, particularly when several features tend to dominate the trees. Fortunately, the *random forest* was introduced by Breiman (2001) as a method to "decorrelate" the bootstrap-grown trees.

Each tree in a random forest is grown using the following algorithm. First, a bootstrap sample is selected from the original data set, upon which a tree is grown. The random forest differs from a bagged approach in the following way. At each step in the growing process, only a random subset, say $m < p$, of the features is considered as candidates for partitioning the feature space. Note that a bagged estimate is produced if $m = p$. A practical feature of the RF is that we attenuate the effect of dominant features and promote features that tend to be masked. This will usually result in a collection of trees that are less correlated than their bagged analogues. The final estimate is still constructed as defined in Equation (7).

2.2.2. Gradient Boosting Regression Trees

Similar to bagging and the random forest, a gradient boosting regression tree, or a gradient boosting machine (GBM), is based on growing a large number of trees. In contrast to bagging and the RF approaches that use bootstrapping, the GBM grows trees in a sequential manner by updating the data used to grow the tree after each tree is fit. More specifically, the algorithm starts by fitting a tree using the original data set. Subsequent trees are grown using the fitted residuals, updated after each fit. The final estimate of $\hat{g}_{gbm}(\mathbf{x})$ is the weighted sum of the individual estimates from each tree. The weighting is controlled by a parameter λ that determines how fast the model "learns" and is usually found via cross validation.

2.2.3. Neural Networks

The final machine learning algorithm that we use in our investigation is a single layer, feed forward neural network (1-FFNL) (see Venables and Ripley, 1997). Essentially, the predictor variables are fed into a hidden layer that transforms the predictors in a (possibly) nonlinear and interactive manner. The hidden layer is then combined in order to estimate the response variable. Gu et al. (2019) provide an excellent, high-level overview of neural networks in the context of empirical asset pricing.

2.3. Model Tuning

Machine learning models (as well as other types of models) usually have one or more parameters that are not directly estimable from the data. Such parameters are often referred to as tuning parameters. In this situation, a candidate set of tuning parameters can be selected *a priori* to evaluate for their effectiveness. That is, for each candidate parameter, the model is fit to a subset of the data and then evaluated on the model's performance at predicting new observations in which the user knows the true response variables. The final tuning parameters are chosen based on pre-specified optimality criteria.[1] Kuhn and Johnson (2013) provide a detailed discussion of the entire tuning process.

One important question in ML is how to evaluate a parameter's effectiveness on model performance. The answer often arrives in the form of subsampling from the original data set. In the case of this paper, we choose to divide our data into training (1972–2000) and testing (2001–2018) sets. Within the training set, we create ten sub-samples, or folds in the ML parlance, upon which we tune and validate each model's performance on a validation set.

As an example of the tuning/validation process used in our analysis, consider Figure 1. The figure's horizontal axis represents the years contained in our training set. Each row in this figure highlights a fold, or sub-sample, and illustrates the data by year that are used to tune our models (blue squares) followed by the validation year (red squares). The data in red squares are often referred to as a validation sample.

The process evolves by training each model using their respective candidate tuning parameters on the individual ten subsamples. The tuning parameters' performance is evaluated based on how well the trained model predicts the response variables in the validation sets averaged across the subsamples. For example, a researcher might choose the tuning parameters which yield the lowest mean squared forecast error (MSFE) on average.

[1] The specific tuning parameters used in our final models can be obtained upon request from the corresponding author.

2.4. Model Fitting

Once the appropriate set of tuning parameters is found for its respective model, we can begin the model fitting process in earnest. That is, we start by fitting each candidate model to the entire training set from 1972–2000 using the appropriately chosen tuning parameters. We use this model to make out-of-sample predictions in the year 2001. The model is updated in each subsequent year, e.g., using data from 1972–2001, in order to make predictions on the following year 2002. We repeat this process through 2018, however, the tuning parameters remain the same. The results are presented in Section 4.

3. Sample Selection and Summary Statistics

3.1. Variable Selection and Data Sources

The empirical literature that tests the determinants of capital structure applies different sets of potential variables implied by each theory. Whether these factors affect corporate leverage positively or negatively depends on the assumptions behind the underlying theory. The potential determinants of corporate leverage roughly belong in three categories: 1) firm-level characteristics including proxies for size, profitability, growth opportunities, bankruptcy risk, and nature of assets, 2) industry-level characteristics including proxies for industry leverage, industry growth, and whether the firm operates in a regulated industry, and 3) macro-level characteristics including proxies for macroeconomic conditions, stock market conditions, debt market conditions, tax policies, and accessibility to the debt market. For the empirical tests in this paper, we collect data from multiple sources to construct different proxies within each category.

Our primary sample includes U.S. firms listed on the NYSE, AMEX, or NASDAQ with CRSP share codes of 10 and 11, that are covered by CRSP and Compustat between 1972 and 2018. We exclude firms for which sales and total assets are either missing or negative, as well as utility (SIC codes 4900-4949) and financial (SIC codes 6000-6999) firms. Following Hou et al. (2015), we use CRSP SIC codes whenever Compustat SIC codes are not available. Data on the expected inflation rate are from the Livingston Survey conducted by the Federal Reserve Bank of Philadelphia. Growth in corporate profits, and growth in GDP are from the Federal Reserve Bank of St. Louis Economic Data (FRED). Table A1 provides a description of the variables used in this paper and the sources.

We also refer the reader to Frank and Goyal (2009) and Kayo and Kimura (2011) who provide a detailed discussion on these proxies and the prediction of each theory on the sign of the proxies. At times, some of the data possess outlier problems. To cope with extreme values and mitigate their effect, we follow Giroud and Mueller (2011) and Valta (2012) and winsorize all ratios at the top and bottom 0.5 percentiles of their underlying distributions.

3.2. Descriptive Statistics

Our sample closely follows Frank and Goyal (2009) with a few changes. We extend their sample by 15 years. Our sample, however, begins in 1972, not 1950, due to the high degree of missing data prior to 1972. The earliest year in which we have observations for the entire set of potential determinants of capital structure is 1972. We also consider an additional measure of firm risk by including Altman's (1968) Z-Score following Graham and Leary (2011). Further, we add cash as a proxy that captures the firm's liquidity. Lastly, we exclude net operating loss carryforward (NOLCF) from our sample due to a high percentage of missing observations. The final sample consists of 233,225 firm-year observations. After keeping the observations that are nonmissing across all the control variables and dependent variables, we are left with 128,417 firm-year observations. We lag all the right hand-side variables by one year in our empirical models.

Table 1 reports the summary statistics. An average firm in our sample has a total debt to market ratio of 27.4%, and enjoys a 5.1% profit rate, where profit is operating income before depreciation scaled by total assets. Capital expenditures account for 6.2% of the total assets, R&D expenditures are 21.3% of total sales, and cash & equivalents account for 15.9% of the total assets. Investing in an average firm in our sample yields an annual return of 15.2%. Empirical capital structure literature uses alternative measures of debt ratio, and these proxies differ in the choice of scaling factor and whether total debt or long-term debt is considered. Throughout this paper, we use the most common measure of capital structure defined as the ratio of total debt to the market value of assets (see for instance, Frank and Goyal, 2009). To save space, we do not report all the results using alternative measures of leverage, however, they are available upon request.

4. Main Results

4.1. Predictive Performance

For the main results, we choose total debt to the market value of assets (TDM) as the dependent variable and the 26 variables listed in Panels A, B, and C of Table 1 (starting by "Profit" and ending by "MacroGr") as the independent control variables. We use 1972–2000 as the training period to tune the machine learning models, and test the out-of-sample performance of these models over the 2001–2018 period. To evaluate the predictive performance of each machine learning model at time $t+1$ in the testing period, we estimate two closely related metrics: mean squared forecast error (MSFE) and the out-of-sample R-squared (R^2_{OS}) defined as

$$MSFE = \frac{1}{N}\sum_{i=1}^{N}(y_{i,t+1} - \hat{y}_{i,t+1})^2, \text{ and} \tag{8}$$

$$R^2_{OS} = 1 - \frac{\sum_{i=1}^{N}(y_{i,t+1} - \hat{y}_{i,t+1})^2}{\sum_{i=1}^{N} y^2_{i,t+1}}, \tag{9}$$

where N is the number of observations at time $t+1$, $y_{i,t+1}$ is the ith actual response at time $t+1$ and $\hat{y}_{i+1}$ is the ith fitted value based on the selected model.

Goyal and Welch (2003) introduce a simple method of evaluating and diagnosing the out-of-sample forecasting ability of predictive regressions. Using only 'then available' (real-time) data, they estimate the difference over time between forecast accuracy of a benchmark and alternative model. Their graphical method reports accuracy performance over the entire out-of-sample period; this contrasts to a statistic such as an MSFE or R^2_{OS} which are point estimates and hence mask the stability of the predictive performance over time. Figure 2 plots the percentage difference between the cumulative sum of the squared forecast errors of the linear benchmark and our machine learning models over the test period (2001–2018); an upward slope indicates that the errors for the benchmark model exceed the machine learning models. The steady positive slopes of RF, GBM, NNET, and GAM imply that these models consistently outperform the linear model (LM). In fact, all four methods have a positive slope every year, indicating that their forecasts generate lower out-of-sample errors every year compared to the LM. In contrast, LASSO's relatively flat and then downward slope demonstrates this method leads to a more parsimonious model at the expense of similar or even larger errors than the LM.

To further highlight the degree of predictability over the past eighteen years, Figure 3 presents R^2_{OS} statistics for each method over the out-of-sample period. This figure illustrates out-of-sample predictability relative to the historical average constant, in contrast to the prior figure that illustrates performance relative to the LM. The linear model generates R^2_{OS} statistics ranging from approximately 28.0% in 2004 to 44.0% in 2015. LASSO's performance is roughly similar to the LM, albeit with less variation, ranging from 30.6% to 39.4%. Overall, RF has the highest consistent forecasting power, with R^2_{OS} ranging between 45.6% in 2009 to 58.7% in 2012. The relatively low predictability in 2009 likely is driven by the financial crisis and its recovery, which drove large changes in investment, borrowing and market values; nonetheless, predictability over 45% during a financial crisis indicates relatively robust predictive performance.

Predictability for GBM nearly matches RF's R^2_{OS} throughout the sample, and by 2018, it reaches 55.3%. GBM's R^2_{OS} also takes a modest hit during the financial crisis, dipping to approximately 46%. NNET R^2_{OS} range from 42.4% in 2010 to 56.1% in 2014, while GAM R^2_{OS} range from 37.1% in 2004 to 48.9% in 2015. Overall, the results show that all the machine learning models improve on the performance of the linear OLS and LASSO models, and R^2_{OS} remains relatively stable from 2001–2018.[2] In particular, the RF model emerges as the best predictor and generates substantial forecasting gains over conventional linear models. The difference between the R^2_{OS} of the RF and LM model ranges from 7.8% in 2009 to 23.7% in 2004. One characteristic leading to the superior performance of machine learning models including RF is that they are designed to allow for complex covariate interactions and nonlinear relationships.

Table 2 presents the average out-of-sample predictability over the entire testing period. The LM generates an R^2_{OS} of 39.8%, and it outperforms LASSO's performance by 3.6%. The two best performers are RF and GBM with R^2_{OS} of 56.2% and 55.3%, respectively. Compared to the linear model, RF and GBM R^2_{OS} represent an economically large increase in out-of-sample predictability. The performance compares favorably with Frank and Goyal (2009) that use backward stepwise regressions; their in-sample R^2 that minimizes the AIC/BIC criteria is 26.6%. Typically, R^2_{OS} are lower than their in-sample counterpart due to less over-fitting. Further, low R^2_{OS} statistics may

[2] In general, the results do not support the earlier work by Graham and Leary (2011) that find declining ability to forecast leverage.

indicate model failure due to parameter instability; however, since predictability is relatively high, and the R^2_{OS} stable over time, there is little evidence that the machine learning models of corporate leverage lead to predictive failure arising from parameter instability.

Table 2 also presents the root mean squared error (RMSE). LM and LASSO have RMSE of 0.190 and 0.196, respectively, while GBM and RF's RMSE are 0.164 and 0.162, respectively, representing sizeable reductions of 13.7% and 14.7%. To evaluate the significance of our predictive accuracy, we use a panel Diebold Mariano test (Ductor et al., 2014). The null is equal predictability, and the alternative is the machine learning model contains useful information beyond the LM, since there is a large reduction in RMSE. The test statistic is distributed $N(0, 1)$, and the large number of observations imply that even modest reductions in RMSE will produce statistical significance. Results document that RF, GBM, NNET and GAM possess significantly lower RMSE than the benchmark LM. Thus, nonlinear machine learning models generate statistically significant increases in forecasting performance.

4.2. Variable Importance

We now turn our attention to identifying covariates that have substantial power in explaining the cross-sectional variation in corporate leverage. We illustrate the contribution of each variable in predicting TDM with variable importance plots for four of our machine learning models estimated for the first year of the out-of-sample period (2001). GAM is not presented because it does not measure variable importance similar to these models. Additionally, to examine whether the determinants of leverage have evolved over time, we also present variable importance for the last year of the out-of-sample period (2018) to evaluate potential changes in importance. Garson (1991) and Goh (1995) identify the relative importance of explanatory variables by deconstructing the model weights. The intuition is that the relative importance or strength of association of a particularly covariate can be determined by identifying all weighted connections between the nodes of interest. Variables with relatively high importance are drivers of the forecasts and their estimates have a larger impact on leverage than values with low importance. For comparison, each model's variable importance measures are scaled so that the most important variable is normalized to 100. All other variables are then scaled relative to this maximum.

Our main focus is the RF model given that it is generally the most accurate model in terms of predictability. Archer and Kimes (2008) demonstrate that RF outperforms other machine learning methods in providing insight regarding the covariates that best contribute to the predictive structure. Similarly, Strobl et al. (2008) find that tree-based methods such as RF excel in identifying relevant predictor variables even in high dimensional settings involving complex interactions. Moreover, RF models make use of a technique called feature bagging, which has the advantage of significantly decreasing the correlation between each decision tree and thus increasing its predictive accuracy, on average (Breiman, 2001).

Turning to the contribution of each control variable to explaining corporate leverage, Figure 4 presents the variable importance fitted to the training data (1972–2000). Panel A documents that RF picks market-to-book as the most important predictive variable, and this value is then normalized to 100. Industry median leverage, cash, Z-Score, profitability, stock returns, and firm size are also important, with weights of at least 20% relative to the most important variable. This implies that RF identifies more variables as important than the work by Rajan and Zingales (1995); their work uses market-to-book, profitability, tangibility, and size as drivers of corporate leverage. We find that industry conditions, stock market conditions, and risk factors are also relevant determinants of corporate leverage. Altogether, the RF model has 20 variables with weights exceeding 5% relative to the most important variable.

Similarly, Panel B of Figure 4 shows that GBM and RF generate relatively similar variable importance results. GBM weights market-to-book as the most important factor, although industry median leverage, cash, Z-Score, and profitability also have weights higher than 20%. GBM weights 11 variables at least 5%. The key difference to RF is that the GBM weights tend to drop off more rapidly. The equitably of weights is even more relevant for NNET. The NNET variable importance plot shows that R&D expenditure is the most important leverage determinant, although profitability, cash, stock variance, and market-to-book also have at least 50% weights. Further, another seven variables have weights exceeding 20% including: depreciation and amortization, growth in GDP, expected inflation, debt ratings, regulated industry, investment tax credit, and selling, general, and administrative expenses. NNET hence more equally assigns variable importance than RF, GBM and LASSO.[3] The LASSO procedure, which is inherently a linear model, weights industry leverage

[3]It is worth noting that the black-box nature of neural networks can make interpreting the individual independent variable's contribution to prediction difficult. Several attempts have been made to rectify this problem, and we specifically employ the R Software package NeuralNetTools' Garson method (Garson, 1991) in our attempt to estimate

as the most important followed by cash with at least 20% weight. Other non-zero-weight variables picked by LASSO include profitability, growth in GDP, profitability, market-to-book, stock returns, and Z-Score. The other variables receive a weight of zero.

Graham and Leary (2011) provide evidence that some of the suggested determinants affect the debt ratio in a nonlinear fashion, yet few studies explicitly model these nonlinearities for prediction or identification of variable importance. Our study takes a step to address these issues and shows that incorporating nonlinearities into capital structure models (here through machine learning estimators) changes our perception of what factors are reliably important. While linear models such as LASSO select only a few factors as important explanatory variables, machine learning models draw predictive information from a much broader set of covariates.

To evaluate whether variable importance evolves over time, we present two different methods. In the first method, we plot variable importance in Figure 5 for data fitted over the years 1972–2017 to predict TDM in 2018; we then compare these plots to Figure 4. The RF model displays only modest reordering of the top three variables; for instance, industry median leverage is now the most important factor over the full sample and market-to-book moves from first place to third place. The graph shows that cash, Z-Score, profitability, stock returns, firm size, selling, general, and administrative cost, and tangibility still emerge as important determinants of leverage seventeen years later with weights of at least 20%. GBM, plotted in Panel B, also displays remarkable variable stability; almost similar to seventeen years earlier, 12 variables are weighted over 5%. Additionally, similar to the prior figure, GBM ranks fewer variables as 'very' important relative to RF; e.g., only four variables possess an importance rank of over 20% and a dozen variables have a relative importance between 5%–20%.

Turning to Panel C of Figure 5, NNET displays modest instability of variable importance with respect to the ranking of a number of variables including profitability. This variable is the second important variable in the prior figure, but then falls to seventh by the end of the sample period. However, the other variables do not change substantially in importance; stock variance, cash, R&D expenditure, depreciation and amortization, expected inflation, and market-to-book are the top six variables in Figure 5, and these variables are among the top eight in Figure 4. Lastly, LASSO's

importance. See Beck (2018) and R Core Team (2019) for additional information related to the NeuralNetTools package and the R Software, respectively.

ranks in 2018 remain identical to the ranks at the beginning of the testing period with only a few exceptions. Industry leverage is ranked as the most important, followed by cash, profitability, market-to-book, stock returns, Z-Score, and tangibility with nonzero weights. All other variables possess zero weights.

Remarkably, the top ten variables in all panels of Figure 4 are also ranked in the top ten in the panels of Figure 5. There is again only moderate reordering within the top ten and top twenty. Industry median leverage is now ranked as the most important factor by three out of four models. Compared to the other three methods, LASSO assigns most variables with low weights and hence leverage is determined by only a few variables. However, LASSO's relatively low predictability compared to RF and GBM documented in Table 2 suggests that leverage is determined both by a wide array of variables and nonlinearities.

In the second method, we illustrate the stability of variable importance in Figure 6, and focus on RF, our top performing model. For clarity and conciseness, we plot variables with importance over 20%. Results show that market-to-book has the highest predictive power from 2001–2012, and then its relative importance declines as both industry median leverage and cash climb in their importance. This implies that managers over time tend to push the firm debt ratio closer to the debt ratio of peer firms that are in the same industry. The figure conveys another central message: variable importance is relatively stable over time. The top three performers are the same for the full out-of-sample period, and the other four variables, Z Score, profitability, stock returns, and firm size possess relatively flat slopes, indicating their relative contribution remains similar over the out-of-sample period.

How does our variable importance results compare to other studies? Rajan and Zingales (1995) analyze the determinants of capital structure using data from major industrialized countries. They suggest using market-to-book, profitability, tangibility, and firm size as the more important factors explaining the observed cross-sectional variation in corporate leverage. Frank and Goyal (2009) use a model selection approach using Akaike information criterion (AIC) and Bayesian information criterion (BIC) to identify a set of reliable corporate leverage determinants. They find six core factors that determine 26.6% of the variation in leverage: industry median leverage, tangibility, market-to-book, profitability, firm size and expected inflation. Results for RF also highlight the importance of these variables. However, RF additionally weights cash, Z-Score, stock returns, and selling, general, and administrative expenses with weights exceeding 20% in Figures 5 and 6. GBM further highlights the relevance of cash, Z-Score, and stock returns with high weights. Work by Kayo and Kimura (2011) emphasizes fewer variables including growth opportunities, profitability, distance from bankruptcy, size and tangibility.

Figure 7 plots the stability of variable importance over time for core factors identified by Rajan and Zingales (1995) and Frank and Goyal (2009). Results show that these core factors remain influential over most of the out-of-sample period and the magnitude of their effect on corporate leverage is robust and stable. The importance of market-to-book and industry median leverage fluctuates between 80% and 100%, while the importance of remaining variables stays between approximately 10% and 35% over the sample.

4.3. Additional Tests

Although many studies advocate market-based leverage as a more managerially relevant measure of corporate leverage due to its forward-looking nature (e.g., Welch, 2004), others support the use of book-based measures of corporate leverage. Since a firm's book values exclude growth opportunities, these values primarily reflect tangible assets, which can be used as debt collateral. Therefore, managers are more likely to set their debt issuance policies based on book leverage. A survey by Graham and Harvey (2001) provides support for this view. Additionally, Barclay et al. (1995) find that practitioners typically use book leverage as the proxy for corporate leverage. Given support in the literature for the use of book leverage, and that most of our focus has been on market leverage (TDM) thus far, this section investigates how predictability and variable importance differ for our three other measures of leverage: LDM, TDA, and LDA.

We begin by presenting the predictive performance of these three alternative metrics of leverage. Figure 8 plots the percentage difference between the cumulative sum of the squared forecast errors of the LM benchmark and our machine learning models over the test period (2001–2018). The three plots tell a remarkably consistent story. Similar to Figure 2, we observe an upward slope for all the nonlinear models indicating that the errors for the benchmark model exceed the machine learning models. The steady positive slopes of RF, GBM, NNET, and GAM imply that these models generate forecasts with lower out-of-sample errors every year compared to the LM irrespective of how corporate leverage is quantified. In contrast, LASSO's relatively flat line in all three panels implies this linear method leads to approximately the sample forecast errors as the OLS linear model. LASSO's poor performance again compared to the other machine learning models further highlights the importance of allowing for nonlinearities and interactions, as this method imposes linear parameters after it chooses its variables, and its modeling does not explicitly allow for interactions between variables.

To further assess the performance of the models, we report predictability (R^2_{OS}) across our out-of-sample period for each machine learning method in Figure 9. The panels in this figure highlight several salient observations. RF dominates in all three panels. Focusing on book leverage in Panel B, for most years, RF averages around 20% greater than the LM and around 4% more than GBM. GBM and NNET for all years outperform the LM by around 15%. LASSO and the LM show similar levels of underperformance. Lastly, there is a modest downward trend in predictability for all models after 2012; e.g., RF predictability decreases about 12% for TDA between 2012 and 2018. However, despite the decline in predictive performance, RF remains the best performing model over this period.

Figure 10 displays the variable importance for LDM, TDA, and LDA; we also include TDM for ease of comparison. To save space, we report the importance of only RF. The relevant comparison to TDM is TDA, where the total debt is scaled by the book value of assets not the market value of assets. Cash emerges as the most important determinants of book leverage followed by Z-Score. These factors represent the third and fourth most important determinants of market leverage, TDM. Overall, the top five and ten drivers of TDA are the same drivers for TDM. Inspection of LDM and LDA also highlights the similarity of drivers of variables. Industry median leverage, market-to-book, cash, and Z-Score are the main factors driving these variables.

5. Predicting Leverage for Alternative Attributes

5.1. Size, Growth, and Technology Sector Attributes

Thus far, our study provides insights on the importance of variables that determine capital structure and demonstrates that machine learning methods, e.g. the RF, that potentially accommodate nonlinear and interaction effects, outperform linear models in predicting corporate leverage. The predictive performance of capital structure theories, however, depends on different settings and conditions (Myers, 2003). The degree of factors such as risk, growth potential or informational opacity may affect the performance of these theories in predicting debt ratios. We therefore classify firms every year to 1) small, medium and large firms based on their size, 2) low, medium and high growth firms based on their market-to-book, or 3) high-tech and non high-tech firms based on their industry. We report the mean squared forecast error (MSFE) and R^2_{OS} every year for the RF and LM.

Panels A and B in Figure 11 and Panel A in Figure 13 present MSFE values for size, growth and technology portfolios, respectively. Results show that the superior predictive performance of machine learning methods carries over into our examined subsamples. RF, which by design accommodates nonlinear functional forms, consistently generates lower out-of-sample forecast errors relative to the linear model for each classification each year. Panels A and B in Figure 12 and Panel A in Figure 14 present R^2_{OS} statistics. The figures demonstrate substantial gains in predicting corporate leverage for RF. For large cap firms, the difference between the R^2_{OS} of the RF and the LM model ranges from 8% in 2003 to 26% in 2004. The difference for small cap firms ranges from 7% in 2003 to 24% in 2013. Similarly, Panels B of Figure 12 and Panel A of Figure 14 document significant and large improvements in R^2_{OS} for growth and value firms as well as high-tech and non high-tech firms. Overall, these results provide substantial evidence that improvement gains in predicting corporate leverage by using machine learning methods are not only an average firm effect. Instead, machine learning methods can be reliably applied to small firms, high-growth firms, or high-tech firms as well as their counterparts.

5.2. Predicting Leverage During Capital Structure Rebalancing

The classic trade-off theory of capital structure predicts that a firm's goal is to maintain an optimal range of debt ratio where the tax benefits and bankruptcy costs of raising debt are balanced. Empirical studies (Fama and French, 2002; Leary and Roberts, 2005; Flannery and Rangan, 2006, among others), however, document that firms rebalance their leverage infrequently and not immediately when their debt ratio deviates from the optimal level. Dynamic trade-off theories such as Leary and Roberts (2005) reconcile this observation and argue that the cost of issuing debt prevents firms from continuously rebalancing their capital structure. Firms readjust their debt ratio when the benefits of recapitalization outweigh the issuance costs. As a result, significant capital restructuring in practice occurs infrequently, but debt changes are relatively large when they do occur. Consequently, it is meaningful to predict debt ratios during capital refinancing periods, because firms are moving at this time towards their optimal target.

In this section, we examine how accurately machine learning models predict leverage ratios of firms that implement a major capital adjustment. We follow Danis et al. (2014) to identify years in which firms adjust their capital structure by a large amount. More specifically, we flag a firm in our sample as refinancing if its net long-term debt issuance relative to assets exceeds 3%, and also its net payouts (cash dividends plus net share repurchases) relative to total assets exceed the 3% threshold. We then focus on the out-of-sample fit for the two subsamples (refinancing and nonrefinancing firms) using our RF and LM estimators. The descriptive statistics for net debt issuance (NetDebt) and net payouts measures (NetPay) are in Panel D of Table 1.

Panel B of Figure 13 and Panel B of Figure 14 report the MSFE and R^2_{OS}, respectively. Results show that RF consistently dominates the LM estimator in predicting the debt ratio of both refinancing and non-refinancing firms. RF predictability measured by R^2_{OS} peaks at over 73% for refinancing firms in 2015. Further, the R^2_{OS} gains relative to the LM range from 9% in 2009 to 36% in 2012 for firms that are adjusting their capital structure. Overall, over the entire sample, RF's average R^2_{OS} for both refinancing and nonrefinancing firms are relatively similar, approximately 55% and 53%, respectively. The takeaway from these results is that machine learning models such as RF have superior ability than LM to predict debt ratios of firms when they undergo major recapitalizations of their capital structure and move closer towards their optimal leverage ratio.

6. Conclusion

Machine learning models that allow for nonlinear and interactions outperform linear OLS models and LASSO models in predicting capital structure. Results show that random forests and gradient boosting models significantly and substantially beat linear models in terms of out-of-sample predictability and lower mean squared error. The better predictive performance by nonlinear machine learning models relative to linear OLS and LASSO models occurs every year of the out-of-sample 2001–2018 period. Results further highlight that the LASSO method underperforms relative to other machine learning methods. Its consistently poor performance relative to other machine learning models reinforces the importance of models that allow for interactions and nonlinearities in identifying variables and predicting corporate leverage.

We find that both random forest and gradient boosting models select market-to-book as the most important predictor of corporate leverage. Industry median leverage, cash, Z-Score, profitability, stock returns, and firm size are also important in determining market leverage. Further, results reveal that the determinants of leverage identified by the machine learning models such as RF are stable for nearly all the out-of-sample period. We also demonstrate that machine learning methods can be reliably applied to small firms, high-growth firms, high-tech firms and firms experiencing capital restructuring as well as their counterparts. Hence, machine learning models consistently outperform OLS in predicting and identifying corporate leverage.

Predicting corporate leverage is salient for managers and investors due to its effect on stock returns and firm value. Machine learning models demonstrate robust and consistent performance gains compared to traditional linear predictive methods; therefore, they should be in the toolbox of scholars and practitioners examining and predicting capital structure. Further, machine learning models can alter our understanding of what variables really matter in explaining corporate leverage. Evaluating the determinants of a firm's leverage over time is more relevant than ever. Corporate leverage relative to GDP is at historic highs, markedly above pre-financial crisis levels. Our results show that the relatively robust predictive performance of random forests coupled with the stability of its determinants suggests that the same firm fundamentals consistently contribute to predicting leverage.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609.

Altman, N. and Krzywinski, M. (2017). Ensemble methods: Bagging and random forests. *Nature Methods*, 14(10):933–934.

Archer, K. J. and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(5):2249–2260.

Barclay, M. J., Smith, C. W., and Watts, R. L. (1995). The determinants of corporate leverage and dividend policies. *Journal of Applied Corporate Finance*, 7(4):4–19.

Beck, M. W. (2018). NeuralNetTools: Visualization and analysis tools for neural networks. *Journal of Statistical Software*, 85(11):1–20.

Bhamra, H. S., Kuehn, L.-A., and Strebulaev, I. A. (2010). The aggregate dynamics of capital structure and macroeconomic risk. *The Review of Financial Studies*, 23(12):4187–4241.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brogaard, J. and Zareei, A. (2018). Machine learning and the stock market. Unpublished working paper, University of Utah.

Childs, P. D., Mauer, D. C., and Ott, S. H. (2005). Interactions of corporate financing and investment decisions: The effects of agency conflicts. *Journal of Financial Economics*, 76(3):667–690.

Danis, A., Rettl, D. A., and Whited, T. M. (2014). Refinancing, profitability, and capital structure. *Journal of Financial Economics*, 114(3):424–443.

Ductor, L., Fafchamps, M., Goyal, S., and van der Leij, M. J. (2014). Social networks and research output. *The Review of Economics and Statistics*, 96(5):936–948.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, Boca Raton, FL.

Erel, I., Stern, L. H., Tan, C., and Weisbach, M. S. (2018). Selecting directors using machine learning. Unpublished working paper, National Bureau of Economic Research.

Fama, E. F. and French, K. R. (2002). Testing trade-off and pecking order predictions about dividends and debt. *The Review of Financial Studies*, 15(1):1–33.

Flannery, M. J. and Rangan, K. P. (2006). Partial adjustment toward target capital structures. *Journal of Financial Economics*, 79(3):469–506.

Frank, M. Z. and Goyal, V. K. (2009). Capital structure decisions: Which factors are reliably important? *Financial Management*, 38(1):1–37.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Garson, D. G. (1991). Interpreting neural-network connection weights. *Artificial Intelligence Expert*, 6(4):46–51.

Giroud, X. and Mueller, H. M. (2011). Corporate governance, product market competition, and equity prices. *The Journal of Finance*, 66(2):563–600.

Goh, A. T. C. (1995). Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, 9(3):143–151.

Gomes, J. F. and Schmid, L. (2010). Levered returns. *The Journal of Finance*, 65(2):467–494.

Goyal, A. and Welch, I. (2003). Predicting the equity premium with dividend ratios. *Management Science*, 49(5):639–654.

Graham, J. R. and Harvey, C. R. (2001). The theory and practice of corporate finance: Evidence from the field. *Journal of Financial Economics*, 60(2):187–243.

Graham, J. R. and Leary, M. T. (2011). A review of empirical capital structure research and directions for the future. *Annual Review of Financial Economics*, 3(1):309–345.

Gu, S., Kelly, B. T., and Xiu, D. (2019). Empirical asset pricing via machine learning. *The Review of Financial Studies*, Forthcoming.

Han, Y., He, A., Rapach, D. E., and Zhou, G. (2018). Firm characteristics and expected stock returns. Unpublished working paper, Emory University.

Harris, M. and Raviv, A. (1991). The theory of capital structure. *The Journal of Finance*, 46(1):297–355.

Hastie, T., Tibshirani, R. J., and Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Series in Statistics. Springer, New York, NY.

Hou, K., Xue, C., and Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3):650–705.

James, G., Witten, D. M., Hastie, T., and Tibshirani, R. J. (2013). *An introduction to statistical learning: With applications in R*. Springer Texts in Statistics. Springer, New York, NY.

Kayo, E. K. and Kimura, H. (2011). Hierarchical determinants of capital structure. *Journal of Banking & Finance*, 35(2):358–371.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer, New York, NY.

Leary, M. T. and Roberts, M. R. (2005). Do firms rebalance their capital structures? *The Journal of Finance*, 60(6):2575–2619.

Li, K., Feng, M., Shen, R., and Yan, X. (2018). Measuring corporate culture using machine learning. Unpublished working paper, University of British Columbia.

Myers, S. C. (2003). Financing of corporations. In Constantinides, G. M., Harris, M., and Stulz, R. M., editors, *Corporate Finance*, volume 1A of *Handbook of the Economics of Finance*, chapter 4, pages 215–253. Elsevier.

Ozdagli, A. K. (2012). Financial leverage, corporate investment, and stock returns. *The Review of*

Financial Studies, 25(4):1033–1069.

Öztekin, Ö. (2015). Capital structure decisions around the world: Which factors are reliably important? *Journal of Financial and Quantitative Analysis*, 50(3):301–323.

Peyer, U. C. and Shivdasani, A. (2001). Leverage and internal capital markets: Evidence from leveraged recapitalizations. *Journal of Financial Economics*, 59(3):477–515.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rajan, R. G. and Zingales, L. (1995). What do we know about capital structure? Some evidence from international data. *The Journal of Finance*, 50(5):1421–1460.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307–317.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Titman, S. and Wessels, R. (1988). The determinants of capital structure choice. *The Journal of Finance*, 43(1):1–19.

Valta, P. (2012). Competition and the cost of debt. *Journal of Financial Economics*, 105(3):661–682.

Venables, W. N. and Ripley, B. D. (1997). *Modern applied statistics with S-PLUS*. Statistics and Computing. Springer, New York, NY.

Welch, I. (2004). Capital structure and stock returns. *Journal of Political Economy*, 112(1):106–132.

Table 1: Summary statistics for the aggregate sample

The sample includes U.S. firms listed on NYSE, AMEX, or NASDAQ with CRSP share codes of 10 and 11, that are covered by CRSP and Compustat between 1972 and 2018. We exclude utility (SIC codes 4900-4999), financial (SIC codes 6000-6999) firms as well as firms for which total assets are either missing or negative. The table reports the summary statistics for the sample including the number of observations, first quartile (Q1), mean, median, third quartile (Q3), and standard deviations (Std.). All ratios are winsorized at the 0.5th and 99.5th percentiles of their empirical distributions. The description of variables is provided in Table A1.

	Observations	Q1	Mean	Median	Q3	Std.
Panel A: Firm-Level Characteristics:						
Leverage Measures:						
TDM	202,061	0.033	0.274	0.205	0.455	0.261
TDA	231,648	0.052	0.244	0.202	0.374	0.224
LDM	202,061	0.007	0.205	0.130	0.344	0.221
LDA	232,130	0.011	0.180	0.122	0.289	0.195
Profitability:						
Profit	226,831	0.020	0.051	0.095	0.162	0.240
Firm Size:						
Assets	233,225	3.590	5.193	5.101	6.704	2.245
Mature	233,225	0.000	0.747	1.000	1.000	0.434
Growth:						
Mktbk	202,061	0.704	1.603	1.021	1.716	1.867
ChgAsset	227,710	-0.024	0.126	0.072	0.199	0.445
Capex	216,700	0.014	0.062	0.039	0.079	0.075
Nature of Assets:						
Tang	227,820	0.056	0.264	0.192	0.400	0.247
R&D	229,170	0.000	0.213	0.000	0.023	1.452
Unique	225,021	0.000	0.235	0.000	0.000	0.424
SGA	190,409	0.141	0.404	0.236	0.374	0.886
Cash	232,420	0.024	0.159	0.070	0.205	0.207
Taxes:						
TaxRate	233,225	0.350	0.380	0.350	0.460	0.058
Depr	215,060	0.022	0.046	0.037	0.057	0.043
InvTaxCr	209,254	0.000	0.001	0.000	0.000	0.004
Risk:						
StockVar	230,480	0.000	0.002	0.001	0.002	0.009
Z-Score	187,768	0.638	1.017	1.793	2.764	3.793
Supply-Side Factors:						
Rating	233,225	0.000	0.110	0.000	0.000	0.312
Stock Market Conditions:						
StockRet	217,712	-0.241	0.152	0.048	0.358	0.835
CrspRet	217,712	-0.008	0.115	0.130	0.253	0.173

Continued on next page

Table 1: Summary statistics: Continued from previous page

	Observations	Q1	Mean	Median	Q3	Std.
Panel B: Industry-Level Characteristics:						
Industry:						
IndustLev	216,178	0.069	0.242	0.210	0.381	0.197
IndustGr	233,035	0.029	0.086	0.076	0.121	0.163
Regultd	225,021	0.000	0.034	0.000	0.000	0.181
Panel C: Macro-Level Characteristics:						
Debt Market Conditions:						
TermSprd	233,225	-0.045	0.019	0.008	0.075	0.079
Macroeconomic Conditions:						
Inflation	233,225	0.024	0.041	0.033	0.048	0.023
MacroProf	233,225	-0.060	0.058	0.090	0.163	0.131
MacroGr	233,225	0.019	0.028	0.031	0.041	0.019
Panel D: Other Characteristics:						
Refinancing Proxies:						
NetDebt	200,540	-0.008	-0.041	0.000	0.019	0.199
NetPay	206,945	-0.016	0.012	0.000	0.027	0.103

Table 2: Out-of-sample prediction of machine learning models

This table reports the out-of-sample R-squared, R^2_{OS}, and the root mean squared error, RMSE, for linear models (LM), least absolute shrinkage and selection operator (LASSO) models, generalized additive models (GAM), neural networks (NNET), gradient boosting machines (GBM), and random forests (RF). The out-of-sample period is from 2001–2018. The dependent variable is total debt scaled by the market value of assets (TDM), and the control variables are listed in panels A, B, and C of Table 1. The last column reports the percentage difference between the corresponding model's RMSE and the LM as the benchmark. Symbols $***$, $**$, and $*$ indicate significance at the 1%, 5%, and 10% levels, respectively, for the panel Diebold Mariano test.

Method	R^2_{OS}	RMSE	RMSE Differential Percentage
LM	0.398	0.190	0.0
LASSO	0.362	0.196	+3.15***
GAM	0.466	0.179	-5.78***
NNET	0.520	0.170	-10.52***
GBM	0.553	0.164	-13.68***
RF	0.562	0.162	-14.73***

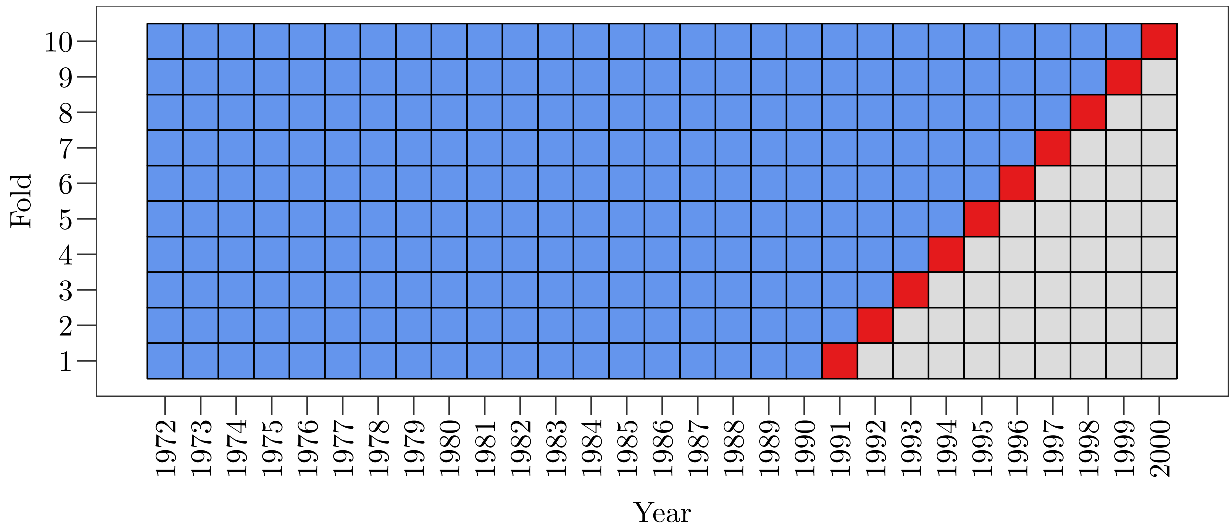


Figure 1: Ten “folds” in the cross validation study

This figure shows the ten folds used for tuning the various machine learning models, where we use an expanding window. The first fold shows that a model is trained on data using the years 1972–1990 (blue boxes) and tested on year 1991 (red box). Similarly, fold ten is trained on years 1972–1999 (blue) and is tested on year 2000 (red). The tuning parameters are then selected by averaging across all ten models.

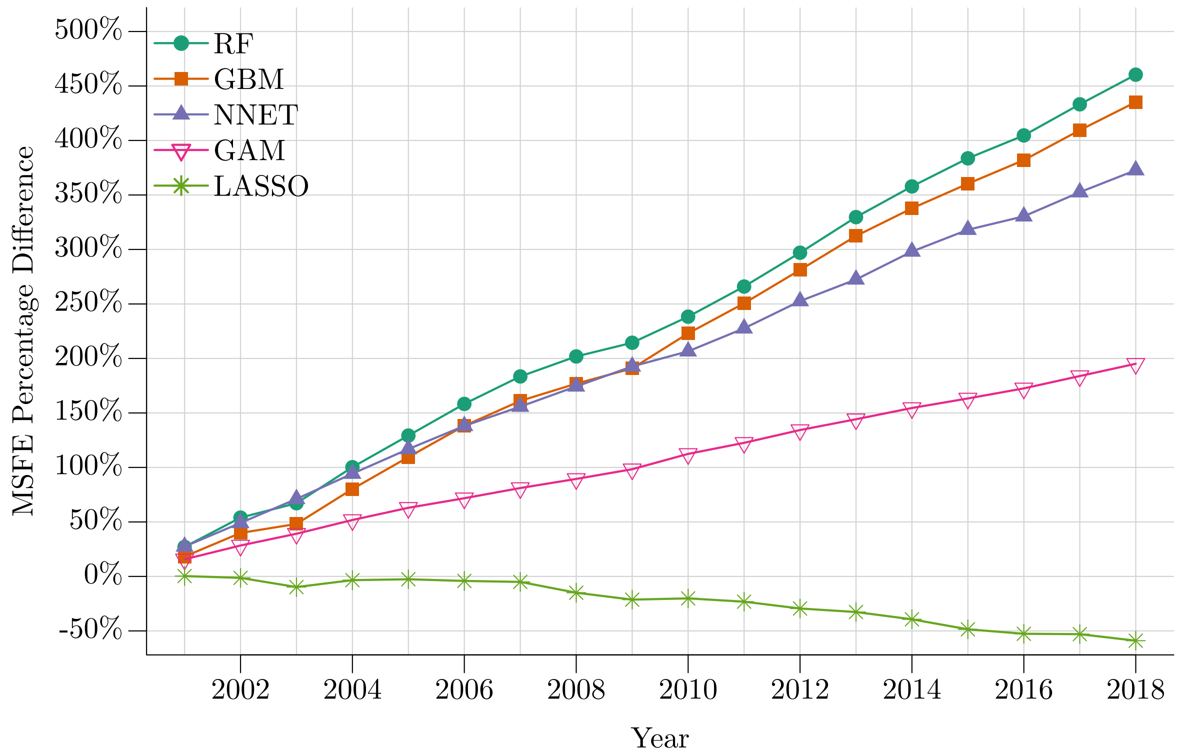


Figure 2: Out-of-sample performance for TDM

This figure plots the percentage change in mean squared forecast error (MSFE) for different machine learning (ML) models relative to a benchmark linear model (LM): $(\text{MSFE}_{\text{LM}} - \text{MSFE}_{\text{ML}})/\text{MSFE}_{\text{LM}}$. The out-of-sample period is from 2001–2018, and the dependent variable is total debt scaled by the market value of assets (TDM). Machine learning models used in the analysis are random forests (RFs), gradient boosting machines (GBMs), neural networks (NNETs), general additive models (GAMs), and least absolute shrinkage and selection operator (LASSO) models. A positive slope indicates lower MSFE for the corresponding machine learning model relative to the benchmark linear model.

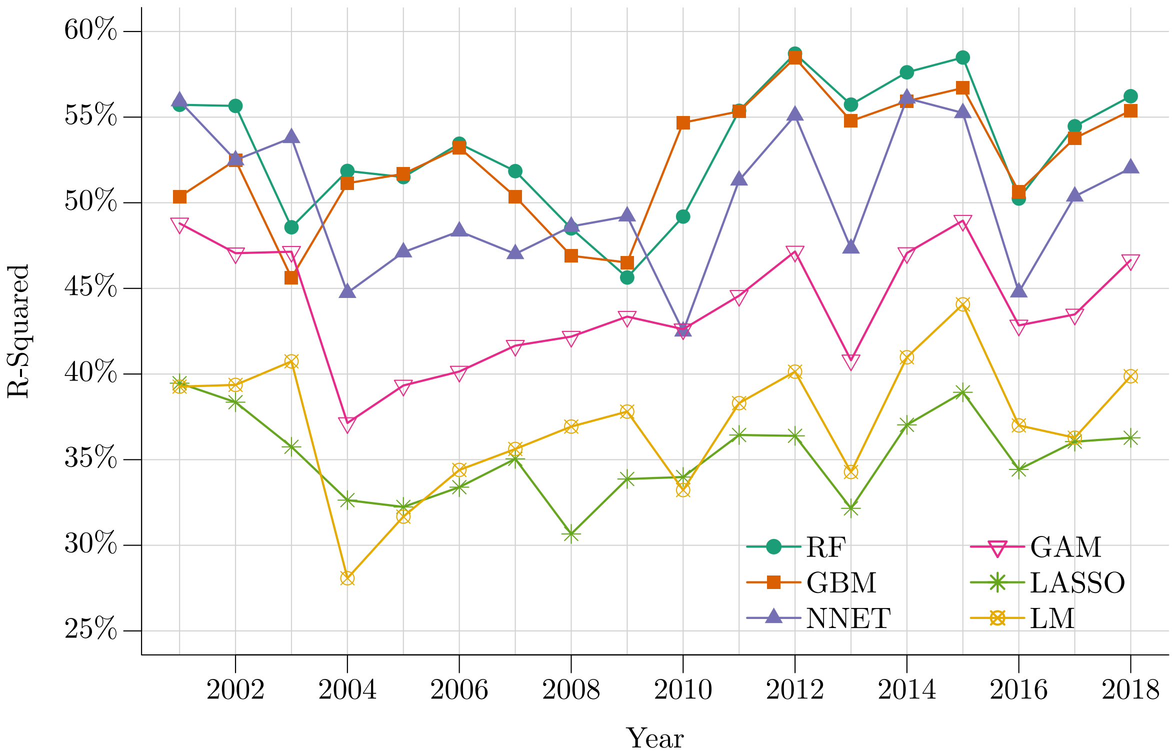


Figure 3: Out-of-sample performance for TDM

This figure plots the out-of-sample R-squared, R^2_{OS}, for different machine learning models and a linear model (LM). The out-of-sample period is from 2001–2018, and the dependent variable is total debt scaled by the market value of assets (TDM). Machine learning models used in the analysis are random forests (RFs), gradient boosting machines (GBMs), neural networks (NNETs), general additive models (GAMs), and least absolute shrinkage and selection operator (LASSO) models.

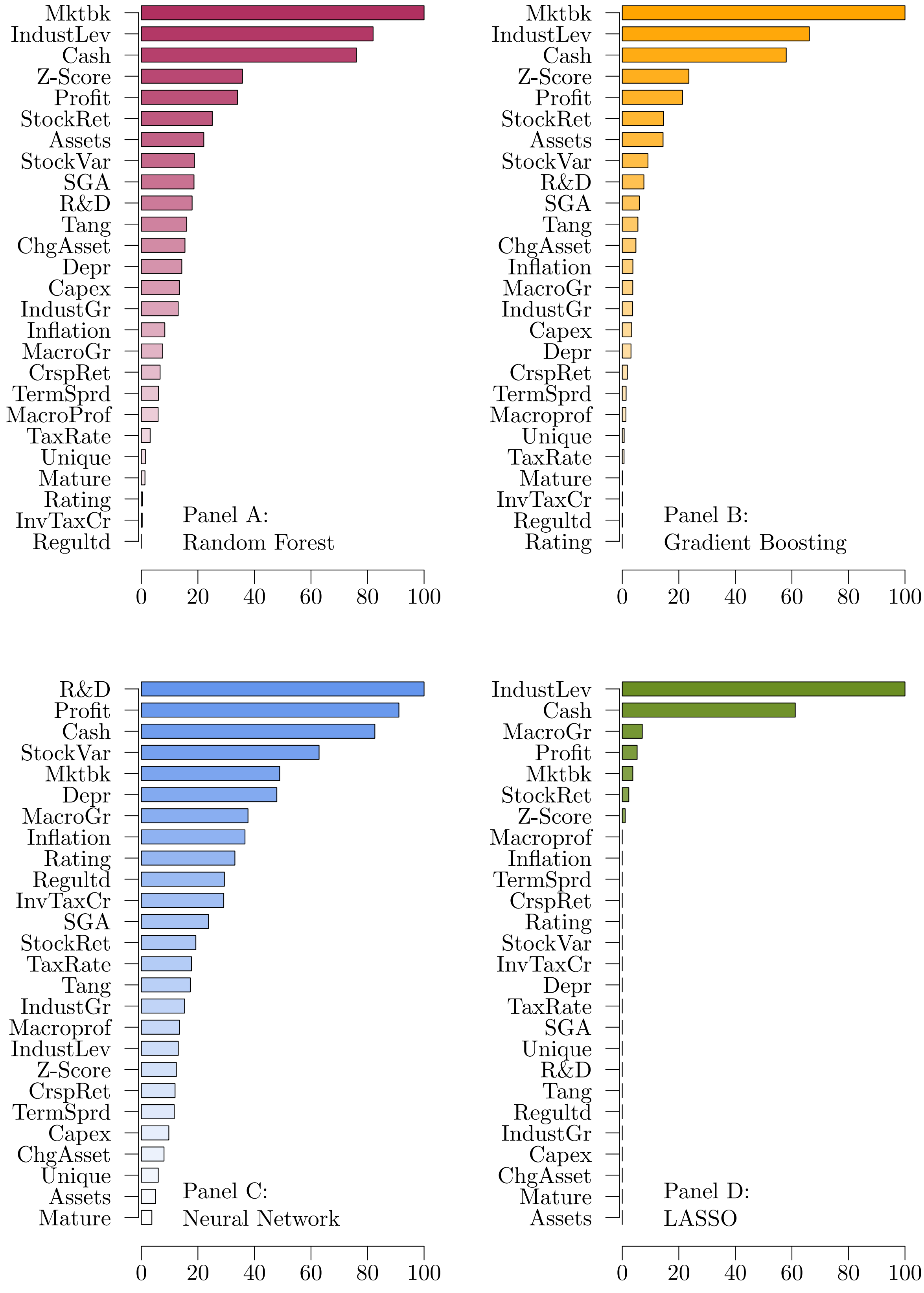


Figure 4: Initial variable importance obtained from different machine learning models

This figure plots the importance of the explanatory variables for predicting corporate leverage using random forest (RF), neural network (NNET), gradient boosting machine (GBM), and least absolute shrinkage and selection operator (LASSO) estimators. The variable with the highest importance is normalized to 100. These estimates are obtained from our training sample over 1972–2000, and used to calculate variable importance for year 2001. The dependent variable is total debt scaled by the market value of assets (TDM).

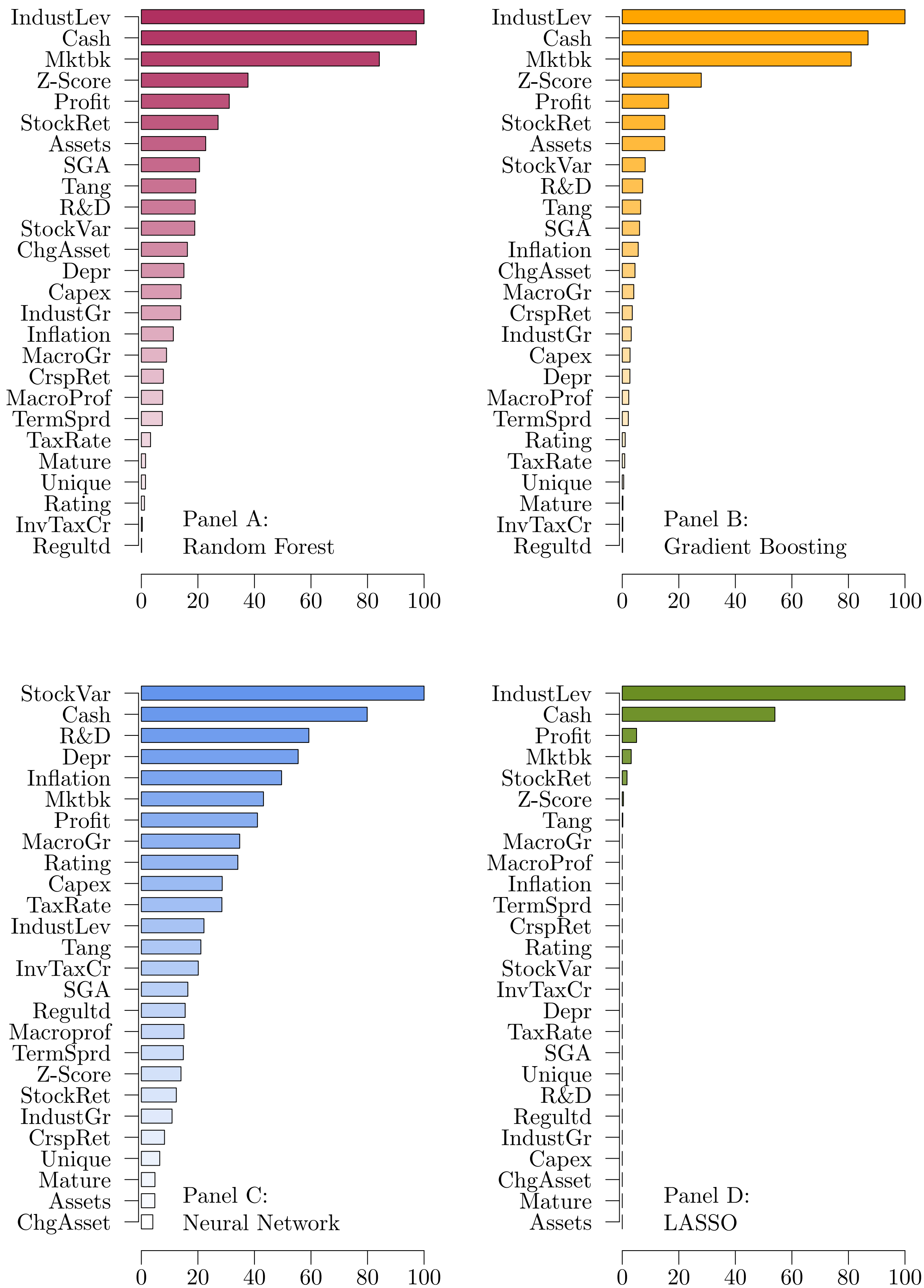


Figure 5: Final variable importance obtained from different machine learning models

This figure plots the importance of the explanatory variables using a random forest (RF), a neural network (NNET), a gradient boosting machine (GBM), and a least absolute shrinkage and selection operator (LASSO) model as estimators. The variable with the highest importance is normalized to 100. The estimates are obtained from the period 1972–2017, and used to calculate variable importance for the final year in the sample, 2018. The dependent variable is total debt scaled by the market value of assets (TDM).

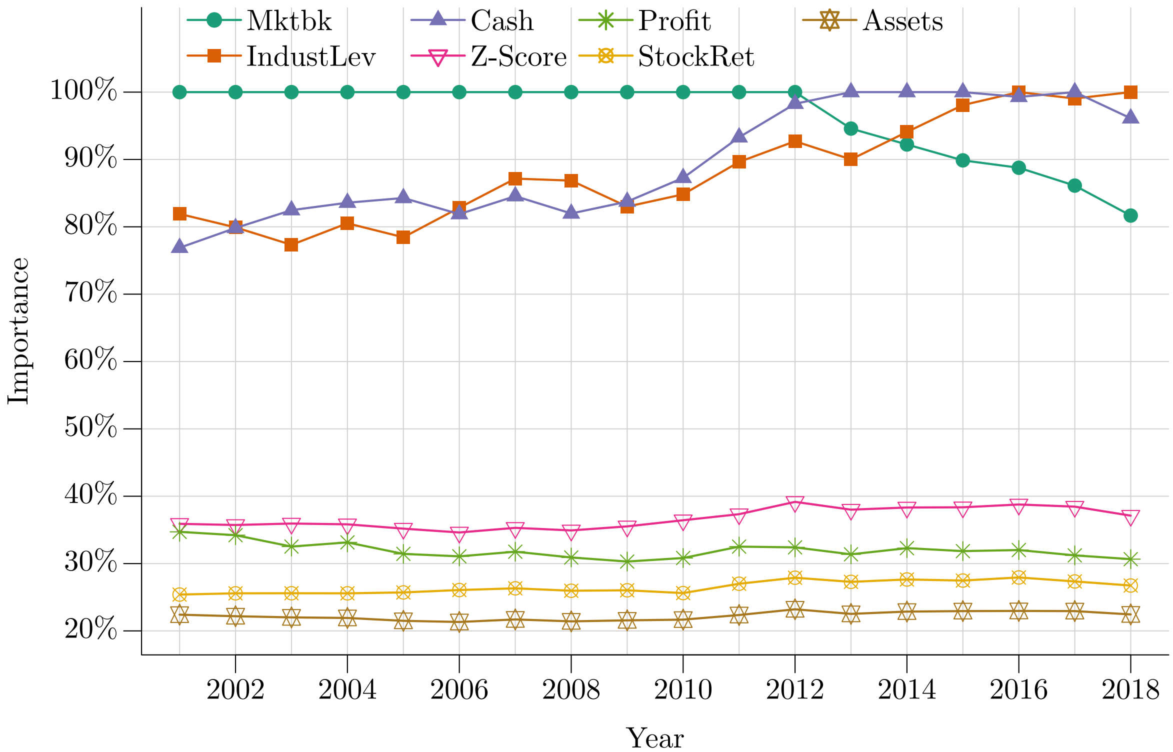


Figure 6: Variable importance over time for TDM

This figure plots the variable importance that exceed 20% over the testing period, 2001–2018, using random forests (RFs). The variable with the highest importance is normalized to 100. The dependent variable is total debt scaled by the market value of assets (TDM). The description of the selected variables is given in Table A1.

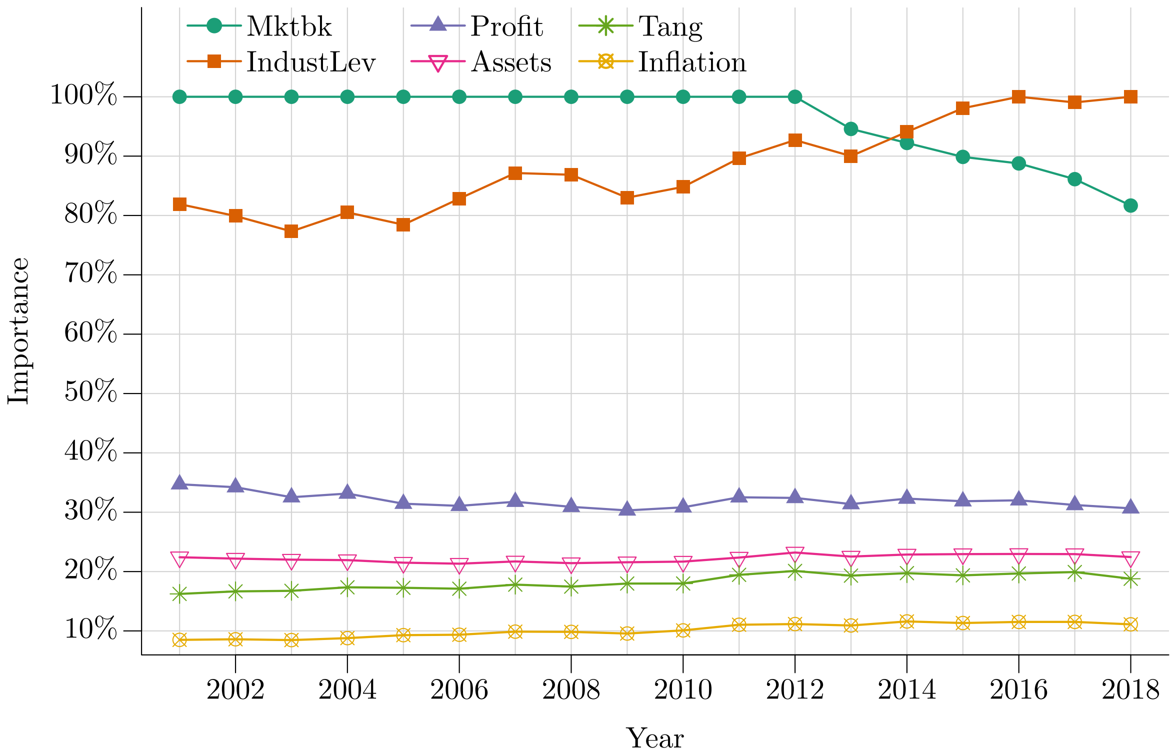


Figure 7: Variable importance over time for TDM

This figure plots the importance of Rajan and Zingales (1995) and Frank and Goyal (2009) core capital structure factors over the testing period, 2001–2018, using random forests (RFs). The variable with the highest importance is normalized to 100. The dependent variable is total debt scaled by the market value of assets (TDM). The description of the selected variables is given in Table A1.

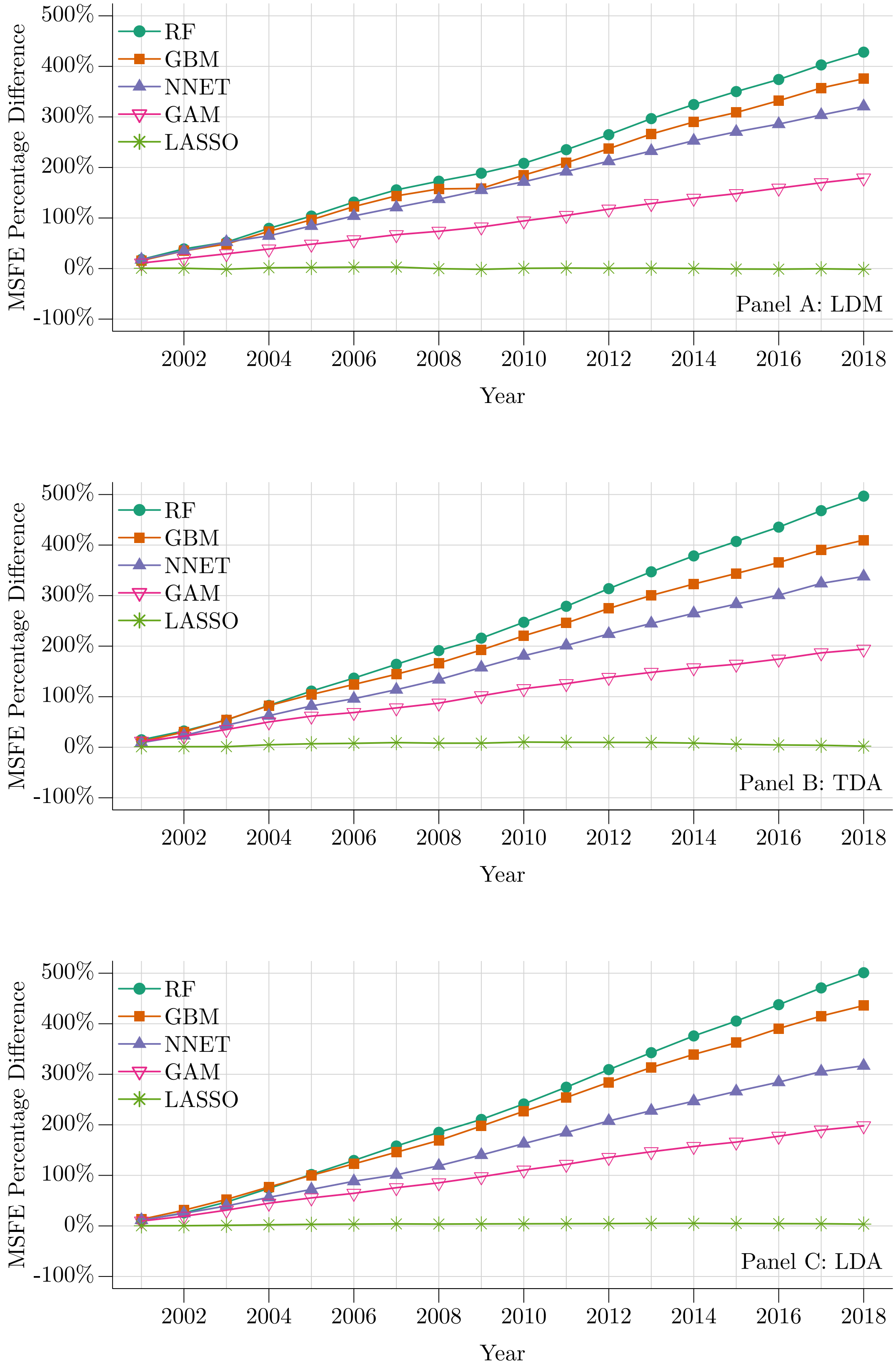


Figure 8: Out-of-sample performance for LDM, TDA, and LDA

This figure plots the percentage change in mean squared forecast error (MSFE) for different machine learning models relative to a linear model. The dependent variable is LDM, TDA, or LDA (for descriptions, see Table A1). Machine learning models used in the analysis are random forests (RFs), gradient boosting machines (GBMs), neural networks (NNETs), general additive models (GAMs), and least absolute shrinkage and selection operator (LASSO) models.

Figure 9: Out-of-sample performance for LDM, TDA, and LDA

This figure plots the out-of-sample R-squared, R^2_{OS}, for different machine learning models and a linear model (LM). The dependent variable is LDM, TDA, or LDA (for descriptions, see Table A1). Machine learning models used in this analysis are random forests (RFs), gradient boosting machines (GBMs), neural networks (NNETs), general additive models (GAMs), and least absolute shrinkage and selection operator (LASSO) models.

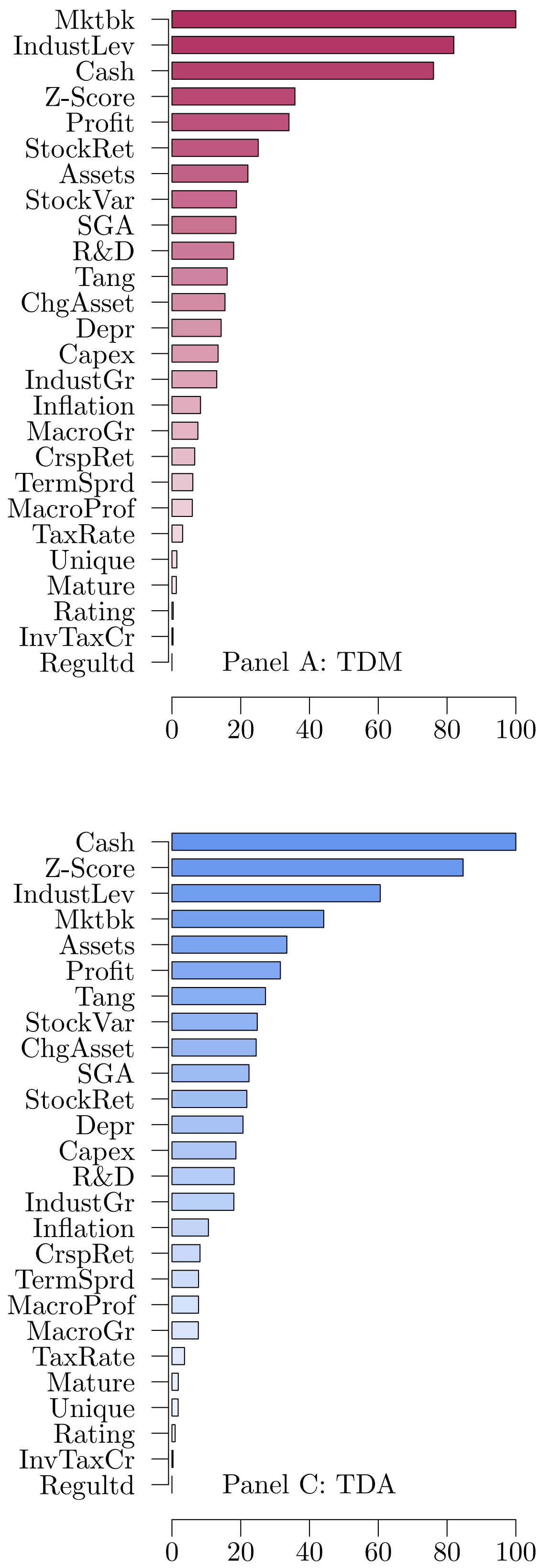
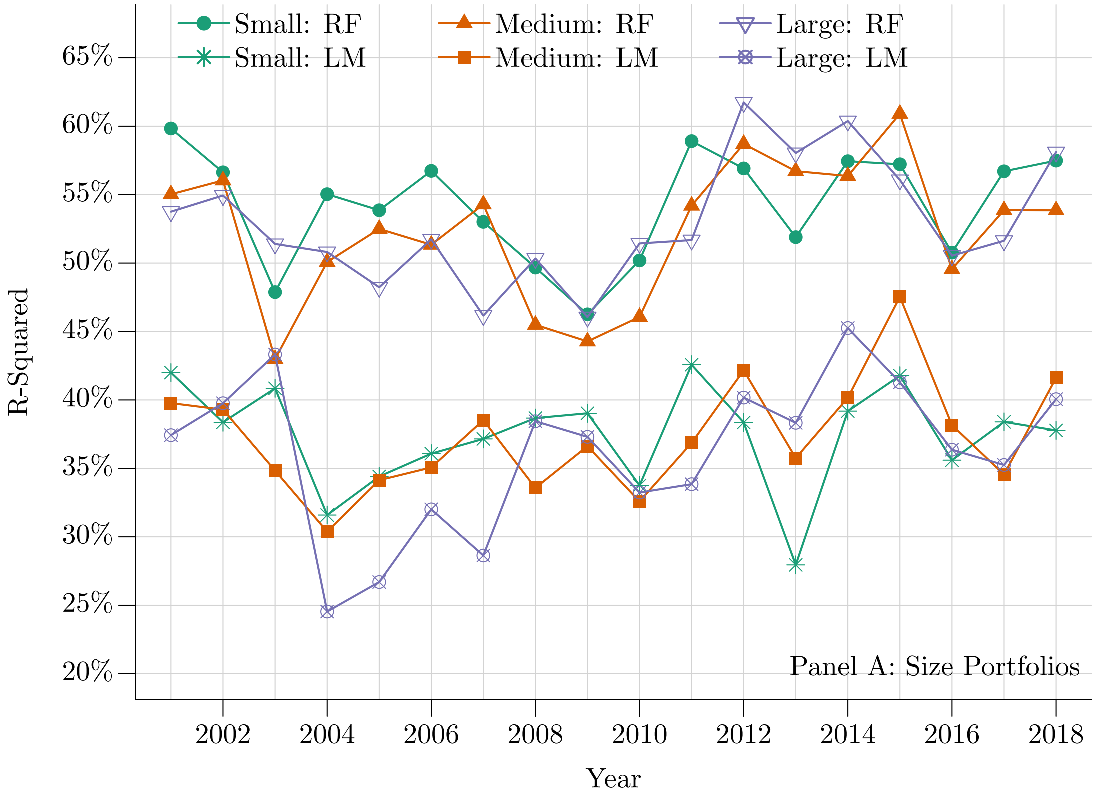


Figure 10: Initial variable importance obtained from a random forest (RF)

This figure plots the importance of the explanatory variables using a random forest (RF) model. The highest importance value is normalized to 100. The estimates are obtained from our training sample over 1972–2000, and used to calculate variable importance for year 2001. The dependent variable is TDM, LDM, TDA, or LDA (for descriptions, see Table A1).

Figure 11: Out-of-sample performance for TDM

This figure plots the mean squared forecast error (MSFE) for size and growth portfolios using a random forest (RF) and a linear model (LM). The out-of-sample period is from 2001–2018, and the dependent variable is total debt scaled by the market value of assets (TDM). See Table A1 for the construction of the portfolios.



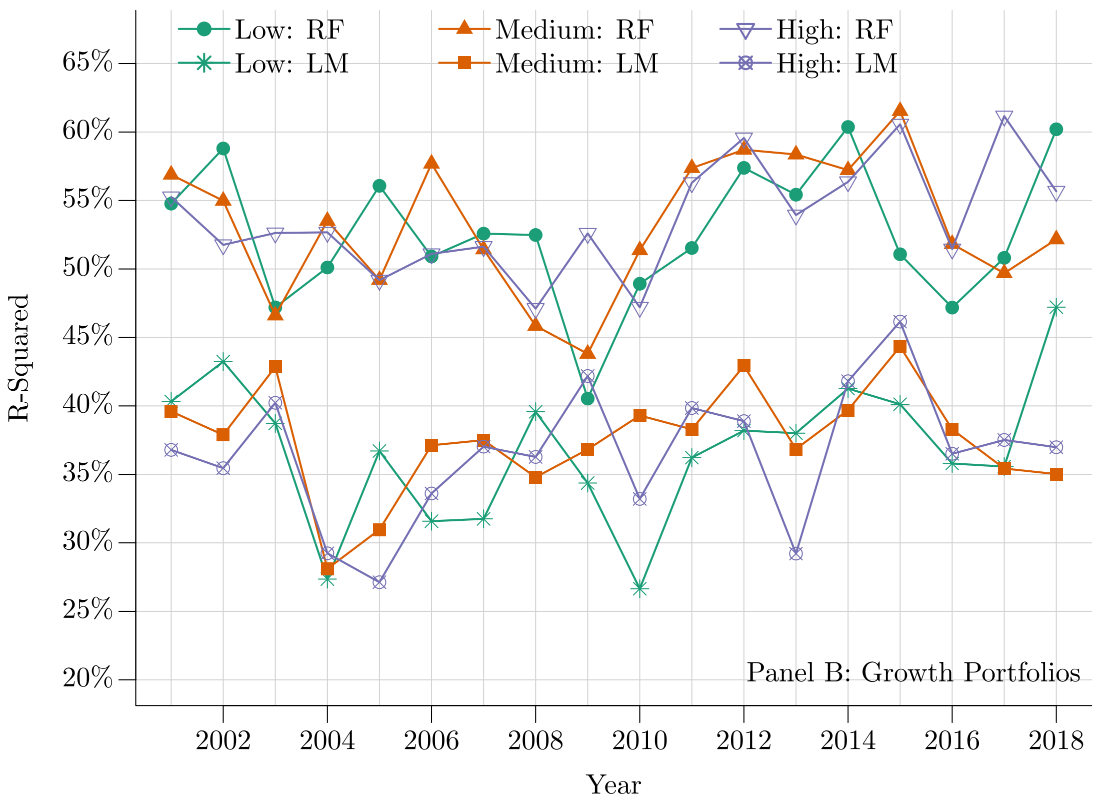


Figure 12: Out-of-sample performance for TDM

This figure plots the out-of-sample R-squared, R^2_{OS}, for size and growth portfolios using random forests (RFs) and linear models (LMs). The out-of-sample period is from 2001–2018, and the dependent variable is total debt scaled by the market value of assets (TDM). See Table A1 for the construction of the portfolios.

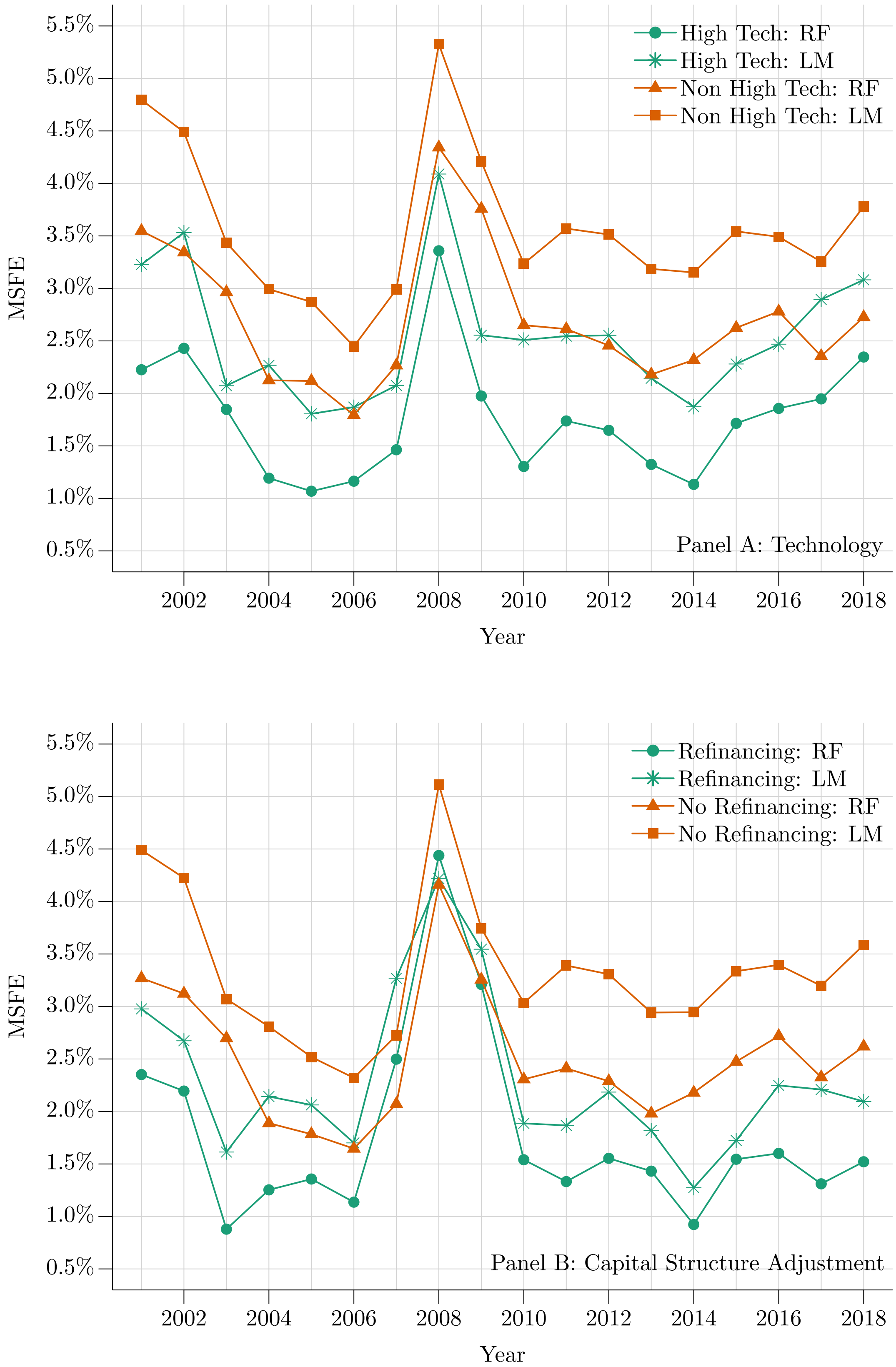
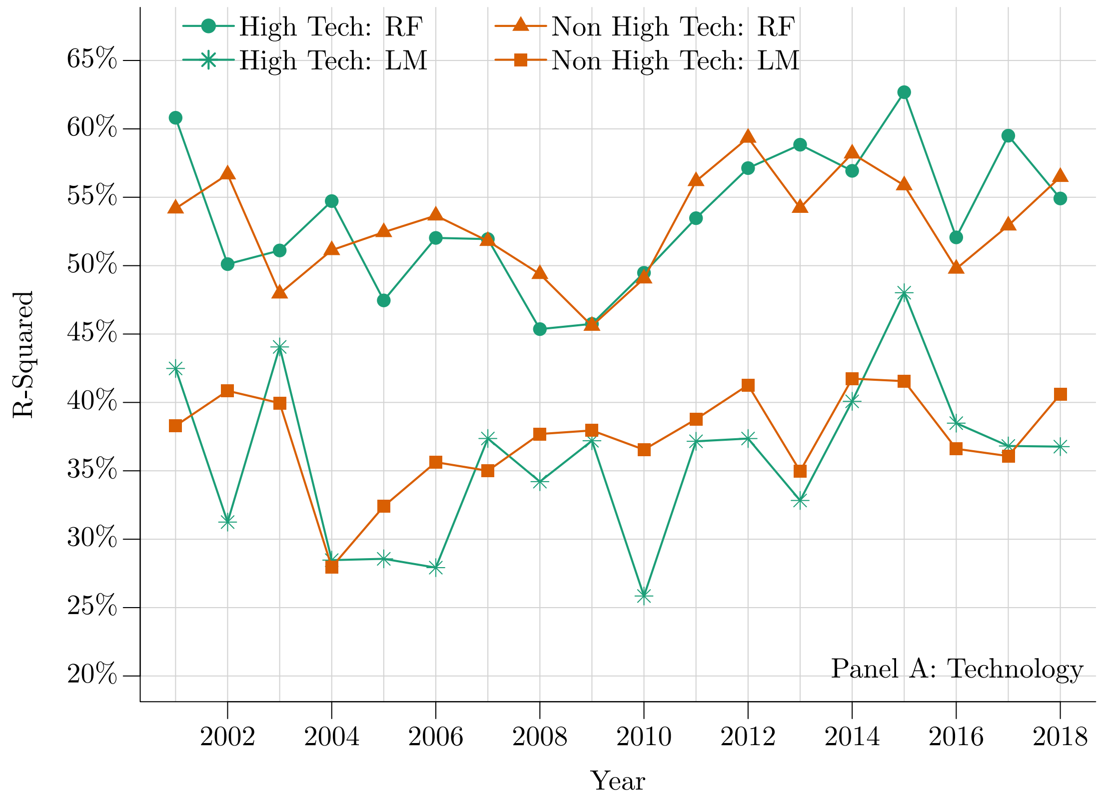


Figure 13: Out-of-sample performance for TDM

This figure plots the mean squared forecast error (MSFE) for technology and refinancing portfolios using random forests (RFs) and linear models (LMs). The out-of-sample period is from 2001–2018, and the dependent variable is total debt scaled by the market value of assets (TDM). See Table A1 for the construction of the portfolios.



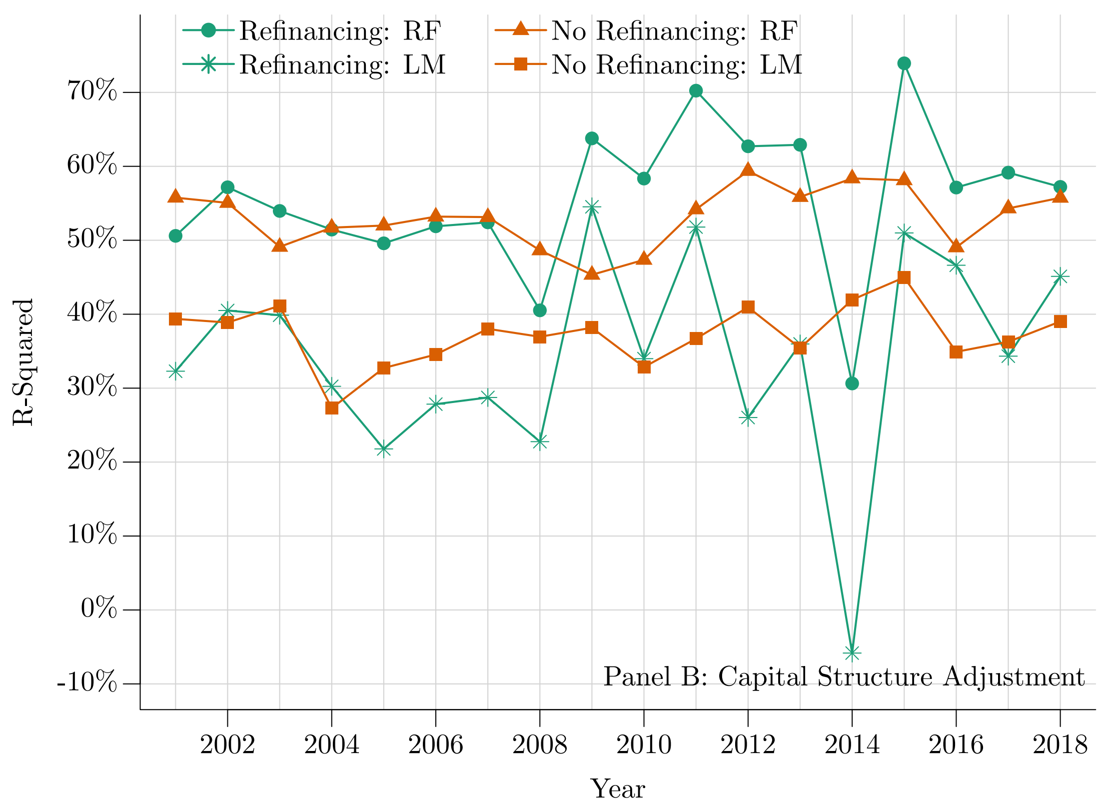


Figure 14: Out-of-sample performance for TDM

This figure plots the out-of-sample R-squared, R^2_{OS} for technology and refinancing portfolios using random forests (RFs) and linear models (LMs). The out-of-sample period is from 2001–2018, and the dependent variable is total debt scaled by the market value of assets (TDM). See Table A1 for the construction of the portfolios.

Table A1: Appendix: Variable definitions and sources

Variable	Description and Source
Market value of equity (MVE)	The stock's fiscal year close price (PRCC_F) times common shares outstanding (CSHPRI). Data source: Compustat.
Market value of assets (MVA)	Debt in current liabilities (DLC) plus long-term debt (DLTT) plus preferred stock liquidating value (PSTKL) minus differed taxes and investment tax credit (TXDITC) plus the market value of equity (MVE). Data source: Compustat.
Leverage (TDM)	Debt in current liabilities (DLC) plus long-term debt (DLTT) scaled by the market value of assets (MVA). Data source: Compustat.
Leverage (TDA)	Debt in current liabilities (DLC) plus long-term debt (DLTT) scaled by total assets (AT). Data source: Compustat.
Leverage (LDM)	Long-term debt (DLTT) scaled by the market value of assets (MVA). Data source: Compustat.
Leverage (LDA)	Long-term debt (DLTT) scaled by total assets (AT). Data source: Compustat.
Profitability (Profit)	Operating income before depreciation (OIBDP) scaled by total assets (AT). Data source: Compustat.
Firm size (Assets)	The logarithm of total assets (AT) deflated to 1992 dollars. Data source: Compustat.
Mature firm (Mature)	A dummy variable which equals 1 if the firm has been in the Compustat database more than 5 years at time t, and equals 0 otherwise. Data source: Compustat.
Market-to-book (Mktbk)	Market value of assets (MVA) scaled by total assets (TA). Data source: Compustat.
Assets growth (ChgAsset)	Change in the logarithm of total assets (AT). Data source: Compustat.
Physical investment (Capex)	Capital expenditures (CAPX) scaled by total assets (AT). Data source: Compustat.
Assets tangibility (Tang)	Net property, plant, and equipment (PPENT) scaled by total assets (AT). Data source: Compustat.

Continued on next page

Table A1: Appendix: Continued from previous page

Variable	Description and Source
Innovation investment (R&D)	Research and development expenses (XRD) scaled by total sales (SALE). Following the standard practice in the literature, we set the R&D expenses to zero whenever it is missing in the Compustat database. Data source: Compustat.
Uniqueness (Unique)	A dummy variable which equals 1 if the firm's 4-digit SIC code is between 3400 and 4000, and equals 0 otherwise. These codes include industries producing sensitive products such as space vehicles, guided missiles, aircraft, computers, semiconductors, chemical & allied, etc. Data source: Compustat.
Non-production cost (SGA)	Selling, general, and administrative expenses (XSGA) scaled by total sales (SALE). Data source: Compustat.
Cash holdings (Cash)	Cash and short-term investments (CHE) scaled by total assets (AT). Data source: Compustat.
Top tax rate (TaxRate)	The top statutory tax rate in the U.S. The rates are 48% from 1971 to 1978, 46% from 1979 to 1986, 40% in 1987, 34% from 1988 to 1992, 35% from 1993 to 2017, and 21% in 2018. Data source: Internal Revenue Service Data Book.
Depreciation (Depr)	Depreciation and amortization (DPC) scaled by total assets (AT). Data source: Compustat.
Investment tax credit (InvTaxCr)	Investment tax credit (ITCB) scaled by total assets (AT). Data source: Compustat.
Stock variance (StockVar)	The annual variance of daily stock returns. The variance value is set to missing if there are less than 100 valid daily returns on CRSP in a fiscal year. Data source: CRSP.
Bankruptcy probability (Z-Score)	Altman's (1968) Z-Score is defined as 3.3 times the difference in operating income before depreciation (OIBDP) and depreciation & amortization (DP) plus sales (SALE) plus 1.4 times retained earnings (RE) plus 1.2 times the difference in total current assets (ACT) and total current liabilities (LCT) scaled by total assets (TA). Data source: Compustat.
Debt rating (Rating)	A dummy variable which equals 1 if a firm's long-term credit rating (SPLTICRM) or subordinated debt rating (SPSDRM) is one of the AAA, AA+, AA, AA-, A+, A, A-, BBB+, BBB, BBB-, BB+, or BB ratings. The dummy variable is 0 if the firm's rating is lower than BB or if the rating is missing. Data source: ADSPRATE database in Compustat.

Continued on next page

Table A1: Appendix: Continued from previous page

Variable	Description and Source
Stock returns (StockRet)	Cumulative annual stock returns using monthly raw returns. Data source: CRSP.
Market returns (CrspRet)	Cumulative annual market returns using monthly value-weighted CRSP returns. Data source: CRSP.
Industry leverage (IndustLev)	The median of corporate leverage (TDM) by 4-digit SIC code and by year. Data source: Compustat.
Industry growth (IndustGr)	The median of assets growth (ChgAsset) by 4-digit SIC code and by year. Data source: Compustat.
Regulated industry (Regultd)	A dummy variable which equals 1 if the firm operates in regulated industries, and equals 0 otherwise. Regulated industries include gas and electric utilities (SIC codes between 4900 and 4939), telecommunications (SIC codes 4812 and 4813) through 1982, airlines (SIC code 4512) through 1978, railroads (SIC code 4011) through 1980, and trucking (SIC codes between 4210 and 4213). Data source: Compustat.
Term spread (TermSprd)	The difference between the 10 year bond returns (B10RET) and the 1 year bond returns (B1RET). Data source: ACTI database in CRSP.
Expected inflation (Inflation)	Expected one-year change in the consumer price index. More specifically, the expected inflation rate is calculated as (Forecast12Month-BasePeriod)/BasePeriod using the December values for expected consumer price index. Data source: Livingston Survey conducted and maintained by Federal Reserve Bank of Philadelphia. The link to the data is here.
Macro profit growth (MacroProf)	Change in logarithm of annual corporate profits with inventory valuation and capital consumption adjustments for nonfinancial firms. Data source: Federal Reserve Bank of St. Louis Economic Data. The link to the series is here.
Growth in GDP (MacroGr)	Change in logarithm of real gross domestic product in 1996 dollars. Data source: Federal Reserve Bank of St. Louis Economic Data. The link to nominal GDP is here and the link to real GDP is here.
Net debt issuance (NetDebt)	Long-term debt issuance (DLTIS) minus long-term debt reduction (DLTR) scaled by total assets (AT). Data source: Compustat.
Net payout (NetPay)	Cash dividends (DV) plus purchase of common and preferred stock (PRSTKC) minus sale of common and preferred stock (SSTK) scaled by total assets (TA). Data source: Compustat.

Continued on next page

Table A1: Appendix: Continued from previous page

Variable	Description and Source
Size dummies	A firm is labeled small, medium, or large in a given year if the size of the firm (Assets) lies in the bottom, middle, or top tercile of its empirical distribution in that year, respectively. Data source: Compustat.
Growth dummies	A firm is labeled low-growth, medium-growth, or high-growth in a given year if the market-to-book of the firm (Mktbk) lies in the bottom, middle, or top tercile of its empirical distribution in that year, respectively. Data source: Compustat.
High tech dummy	A dummy variable which is 1 if a firm offers technology products and services, and equals 0 otherwise. More specifically, a firm is defined as a high-tech firm if its corresponding 4-digit SIC code equals one of the 3571, 3572, 3575, 3577, 3578, 3661, 3663, 3669, 3671, 3672, 3674, 3675, 3677, 3678, 3679, 3812, 3823, 3825, 3826, 3827, 3829, 3841, 3845, 4812, 4813, 4899, 7371, 7372, 7373, 7374, 7375, 7378, or 7379 values. Data source: Compustat.
Refinancing dummy	A dummy variable which is 1 if a firm's both net long-term debt issuance (NetDebt) and net payout (NetPay) relative to total assets exceed the 3% threshold, and equals 0 otherwise. Data source: Compustat.