Title
Improving Inference with Machine Learning: Application to CEO Turnover

Author
Zafirov, Athanasse

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Improving Inference with Machine Learning :

Application to CEO Turnover

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Management

by

Athanasse Zafirov

2022

ABSTRACT OF THE DISSERTATION

Improving Inference with Machine Learning :

Application to CEO Turnover

by

Athanasse Zafirov

Doctor of Philosophy in Management

University of California, Los Angeles, 2022

Professor Henry L. Friedman, Chair

In this paper I estimate the risk of CEO dismissal using a variety of machine learning algorithms. I show that linear regression tree methods significantly outperform the logit and linear models used in prior literature, as well as other algorithms, notably neural networks, who perform poorly in this setting. Taking these superior predictions to applications from prior studies, I find that relationships change. Peters and Wagner (2014) found that increases in forced turnover risk were related to a material increase in CEO pay, the more accurate risk estimate remains statistically significant, but becomes less than a percent of its previous size. As well, decreases in pay-performance-sensitivity found in Bushman, Dai, and Wang (2010) are no longer economically significant. Furthermore, using the likelihood of dismissal to address sample selection bias in Huson, Malatesta, and Parrino (2004), I find even stronger evidence of a positive link between CEO firing and future firm-level performance.

The dissertation of Athanasse Zafirov is approved.

Amit Goyal

Ivo I. Welch

Constanca Esteves-Sorenson

Henry L. Friedman, Committee Chair

University of California, Los Angeles

2022

Contents

List of Figures

List of Tables

Vita

Education

2012 - 2015	Master's of Science, Applied Economics, HEC Montréal
2008 - 2011	Bachelor's of Commerce, Finance, Concordia University
2005 - 2008	Diploma of College Studies, Computer Science, Dawson College

Professional Experience

2014 - 2016	Financial Analyst, Letko Brosseau
2013	Intern, Caisse de dépôt et placement du Québec (CDPQ)
2010 - 2011	Intern, Scotiabank, Capital Markets
2009	Intern, Healthcare of Ontario Pension Plan
2008	Intern, Centre d'Etude et de Développement de l'Informatique

Research Assistance

Coviello, Decio, Erika Deserranno, and Nicola Persico. "Minimum wage and individual worker productivity: Evidence from a large US retailer." Journal of Political Economy 130.9 (2022): 2315-2360.

Faraglia, Elisa, et al. "Government debt management: The long and the short of it." The Review of Economic Studies 86.6 (2019): 2554-2604.

Oikonomou, Rigas. "Unemployment insurance with limited commitment wage contracts and savings." The BE Journal of Macroeconomics 18.1 (2018).

Mankart, Jochen, and Rigas Oikonomou. "Household search and the aggregate labour market." The Review of Economic Studies 84.4 (2017): 1735-1788.

Publications

Xu, Yuancheng, et al. "FREEtree: a tree-based approach for high dimensional longitudinal data with correlated features." arXiv preprint arXiv:2006.09693 (2020).

Goyal, Amit, Ivo Welch, and Athanasse Zafirov. "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction II." Available at SSRN 3929119 (2021).

Projects

Earnings Sentiment Analysis: A Comparison of Transfer Learning Methods (BERT, ALBERT, XLNet)

PropCast: Deep Inception-Convolutional LSTM Networks for Forecasting Spatio-Temporal Transitions in Property Index Heatmaps

Teaching Assistance

MGMT 439: Tools and Analysis for Business Strategy, UCLA
MGMT 264A: Market Research, UCLA
MGMT 403: Financial Accounting, UCLA
ECO 680109: Applied Econometrics, HEC Montréal
ECO 600112A: Economic Corporate Organization HEC Montréal
ECO 380607: Introduction to Econometrics, HEC Montréal

Improving Inference with Machine Learning: Application to CEO Turnover

Athanasse Zafirov
UCLA Anderson

December 6, 2022

Abstract

In this paper I estimate the risk of CEO dismissal using a variety of machine learning algorithms. I show that linear regression tree methods significantly outperform the logit and linear models used in prior literature, as well as other algorithms, notably neural networks, who perform surprisingly poorly in this setting. Taking these superior predictions to applications from prior studies, I find that relationships change. Peters and Wagner (2014) found that increases in forced turnover risk were related to a material increase in CEO pay, the more accurate risk estimate remains statistically significant, but becomes less than a percent of its previous size. As well, decreases in pay-performance-sensitivity found in Bushman, Dai, and Wang (2010) are no longer economically significant. Furthermore, using the likelihood of dismissal to address sample selection bias in Huson, Malatesta, and Parrino (2004), I find even stronger evidence of a positive link between CEO firing and future firm-level performance.

Keywords: CEO Turnover, Machine Learning, Regression, Linear Model Trees, Hyperparameter Tuning.

I would like to thank Henry Friedman, Ivo Welch, Amit Goyal, Costanca Esteves-Sorenson, Velibor Misic, Florian Peters and Christina Ramirez for all the help and support.
E-mail: dimitri.zafirov.phd@anderson.ucla.edu

1 Introduction

The decision to retain or dismiss a CEO is one of the most visible and important choices a board of directors can make. The likelihood of forced CEO turnover has been studied as a determinant of compensation and performance, but previous work has pointed to contradictory impacts of the chance of dismissal on chief executive pay (Peters and Wagner (2014); Bushman, Dai, and Wang (2010)). Identification is exacerbated by the difficulty in differentiating the risk of forced turnover from the event itself, and further complicated by the large number of factors that have been shown in the literature to be potential determinants of CEO removal relative to the small fraction of firm-years featuring forced turnover.

Thus far, logistic regression has been the dominant method to estimate the probability of CEO turnover, often as a first-stage to estimating its impacts on other company dynamics, such as CEO compensation. However, recent advances have shown alternative machine learning algorithms to be better suited to producing more accurate forecasts for a wide range of uncommon firm-level phenomenon such as fraud, litigation and restatements (Chen, Cho, Dou, and Lev (2021); Bao, Ke, Li, Yu, and Zhang (2020); Bertomeu, Cheynel, Floyd, and Pan (2021)). Here, I make use of machine learning methods to form the best prediction of forced executive turnover risk and test it against previous findings in the extant literature. A generalized linear mixed-effects model tree (GLMM tree, Fokkema, Smits, Zeileis, Hothorn, and Kelderman (2018)) outperformed a comprehensive list of alternative algorithms, including neural networks, by a wide margin. A review of machine learning applications in finance and accounting reveals this class of model-based tree methods is rarely employed in the literature.

My paper is the first application of machine learning to CEO turnover. Beyond providing more accurate predictions, it tests the robustness of previously estimated effects of CEO turnover probability on compensation and firm performance. As is typical practice in machine learning, I rank the validity and generalizability of the attempted models by comparing their in-sample (training sample) prediction performance to their out-of-sample (testing sample) performance.

A feature selection algorithm was used to select regressors, called features in the machine learning domain. Random Forest Recursive Feature Elimination (RF-RFE) selects the top most

important variables by rank. This reduces the dimensionality problem and the risk of overfitting relative to using all available series.

For the sample without compensation variables, the best performing GLMM tree applied to the entire data sample achieved prediction performances that far surpassed those obtained by Peters and Wagner (2014)'s replicated model. Depending on the metric, we see between 50% and 950% approximate improvement from the best performing previous model results in the literature over Peters and Wagner (2014).

Using a linear probability model to estimate the risk of dismissal, Peters and Wagner (2014) find a large and significant impact on CEO compensation of $233,678 for every 1% increase in the likelihood of dismissal. Using my more accurate risk measure I instead find evidence of a statistically significant but relatively minuscule change in compensation.

Bushman, Dai, and Wang (2010) used a logit regression to estimate the likelihood of forced CEO turnover and found a negative effect on the change in total compensation and pay-performance-sensitivity (PPS), which is the change in the dollar value of executive stock and option holdings for a 1% change in the stock price in a given year. In the replication of Bushman, Dai, and Wang (2010)'s model, the effect on the change in total compensation was a negative $46,166 for a 1% increase in the risk of dismissal, and it was only marginally statistically significant. This shrunk in absolute terms to a negative $2,469 impact using the more accurate prediction of forced turnover, and was no longer statistically significant. Meanwhile the effect on PPS in the replication was significant at the 5% level. Using my more accurate predictions, the coefficient remained negative and statistically significant, but fell to 8.6% of its previous size. Both of these results suggest a small drop in compensation resulting from an increase in dismissal risk.

However not all second-stage relationships in the previous literature weakened when using the machine learning derived predictions. When used as Propensity Score Weightings (PSW) to address potential sample selection bias in a second-stage regression model to estimate the effects of forced CEO turnover on firm performance, the better first-stage favors the "improved management hypothesis". Whereas Huson, Malatesta, and Parrino (2004) found a positive but

not significant relationship between dismissal and future profit growth, I find a materially larger and statistically significant effect by using the predicted turnover probabilities as propensity score weights to address potential sampling bias issues.

The paper proceeds as follows. Section 1 reviews the raw variable list, called features in machine learning, which has been gathered from previous literature on CEO turnover determinants. Section 2 describes the chosen regressors. Section 3 contains a description of each machine learning model and details on its implementation. Raw performance results of the models, including optimal tuning parameters, are in section 4. Section 5 then studies the implications of these more accurate estimates for previously studied relationships in the literature, before I conclude in section 6.

2 Literature Review

2.1 CEO Turnover

Previous studies have shown a relation between CEO turnover and company performance (Murphy and Zimmerman (1993); Khurana and Nohria (2000)) and investor outcomes (Adams and Mansi (2009); Clayton, Hartzell, and Rosenberg (2005)). Past literature looking at the determinants of CEO turnover have linked it to a wide number of potential factors: from raw and relative firm performance (Jenter and Lewellen (2021); Jenter and Kanaan (2015); Huson, Parrino, and Starks (2001); Kaplan and Minton (2012); Engel, Hayes, and Wang (2003)), to expectations (Puffer and Weintrop (1991); Farrell and Whidbee (2002); Lee, Matsunaga, and Park (2012)), to the ability of the board to monitor performance (Bushman, Dai, and Wang (2010)), as well as conflicts of interests among directors (Weisbach (1988); Laux (2008)). Other dynamics include inertia from tenure (Taylor et al. (2010); Dikolli, Mayew, and Nanda (2014)), the quality of corporate governance (Fisman, Khurana, Rhodes-Kropf, and Yim (2014); Fiordelisi and Ricci (2014)), institutional factors (Dah, Frye, and Hurst (2014); Defond and Hung (2004)), the executive job market (DeFond and Park (1999); Eisfeldt and Kuhnen (2013)), and even the press (Farrell and Whidbee (2002)). The vast majority of these mostly empirical studies make use of logistic regression to estimate the impact of their variables of interest on the forced turnover

outcome.

A large part of the literature models the turnover decision as a signaling game in which the BoD is updating their beliefs about the CEO's ability on which firm performance directly depends (eg: earnings are a direct function of CEO type). As such much of the tension in these models stems from the board's estimation of the CEO's type given what is being observed, such as various performance metrics like stock returns (Warner, Watts, and Wruck (1988)), earnings growth (Farrell and Whidbee (2003)) and return-on-assets (Huson, Malatesta, and Parrino (2004)). Beyond absolute metrics of performances, studies have looked into measures relative to industry (Eisfeldt and Kuhnen (2013); Goyal and Park (2002)), analyst forecasts (Farrell and Whidbee (2003)), management forecasts (Lee, Matsunaga, and Park (2012)) and past performance (Krupa and Minutti-Meza (2021)). Event related studies have also looked into the incidence of class action lawsuits or insider trading (Niehaus and Roth (1999)) and restatements (Desai, Hogan, and Wilkins (2006); Hennes, Leone, and Miller (2008); Burks (2010)). Particular CEO characteristics of interest include age (particularly indicators of being near-retirement) and education (Bhagat, Bolton, and Subramanian (2010)). Certain papers explore factors that can affect how accurate and readily measurable these performance metrics are, such as: Earnings management (Hazarika, Karpoff, and Nahata (2012)), earnings volatility (Engel, Hayes, and Wang (2003)), disclosure informativeness (Bochkay, Chychyla, and Nanda (2019)) and accounting regulations (Burks (2010); Meng (2019)). The current CEO's salary is a consideration for dismissal decisions as it increases the cost of retaining the incumbent. Campbell, Gallmeyer, Johnson, Rutherford, and Stanley (2011) and Laux (2012) studied the impacts of particular options schemes on turnover. Others looked into the effects that the presence of the CEO on the compensation committee had (Fiordelisi and Ricci (2014)), evidence of overcompensation (Coughlan and Schmidt (1985)), as well as the existence of severance pay agreements (Inderst and Mueller (2010)).

Studies that look at varying ways in which a CEO can be entrenched in his position have looked at factors such as founder status (see for instance Mobbs (2013); Beneish, Marshall, and Yang (2017)) or relation to the founding family (Huson, Parrino, and Starks (2001)), the share of CEO ownership in the firm (e.g.: Bhagat, Bolton, and Subramanian (2010); Hazarika,

Karpoff, and Nahata (2012)), voting power (Guo and Masulis (2015)), and whether the CEO sits on the board of directors (eg: Fiordelisi and Ricci (2014)) or is the active chairman, often referred to as *duality* (Helwege, Intintoli, and Zhang (2012)). Board of director characteristics that influence turnover decisions include its level of independence (Guo and Masulis (2015)), the presence of outside directors (Laux (2008); Fiordelisi and Ricci (2014)), and staggered board terms (Laux (2008)). What could also be thought of as retaining costs (or negative dismissal costs) are shareholder pressures in the form of institutional or block share ownership (Kaplan and Minton (2012)) and the presence of activist investors (Helwege, Intintoli, and Zhang (2012)), and anti-takeover provisions (Bushman, Dai, and Wang (2010)). Another component of turnover costs includes the ease or challenge of finding and hiring a new CEO. Factors include the executive availability in the firm's industry (DeFond and Park (1999)) and the presence of a suitable internal replacement (Mobbs (2013)). Other considerations such as the loss of human capital associated with executive dismissal (Sliwka (2007)) and laws regarding CEO termination (Cornelli, Kominek, and Ljungqvist (2010)) have also been explored. Table 1 presents a list of these determinants of CEO turnover in past literature.

[Table 1 around here]

Aside from looking at the factors influencing CEO turnover, certain papers have studied the risk of CEO turnover and its impacts on firm and executive level variables. Of particular interest, the probability of forced CEO turnover has been associated with a higher absolute level of pay (Peters and Wagner (2014)), a lower pay-performance sensitivity and lower pay growth (Bushman, Dai, and Wang (2010)). Huson, Malatesta, and Parrino (2004) show a positive impact of termination on firm level performance. I revisit these findings using a more accurate measure of CEO turnover risk, as an alternative independent variable (in the case of Peters and Wagner (2014) and Bushman, Dai, and Wang (2010)) and as a way to mitigate sample selection bias (in the case of Huson, Malatesta, and Parrino (2004)).

2.2 Machine Learning Predictions

Easy to interpret linear regression models are the standard analytical tools used in this literature. However, large numbers of inputs have shown to overfit the data in accounting research when interactions are included (Barth, Li, and McClure (2019)). Past popular solutions in domain-adjacent research, as in Bernanke, Boivin, and Eliasz (2005), sought to remedy this issue through the use of dimensionality reduction techniques such as principal component analysis to distill the most important factors driving variance in the data and use these factors as regressors.

Machine learning techniques have previously been applied to large datasets to gain insight on firm-level phenomena such as litigation (Lee, Naughton, Zheng, and Zhou (2020)) and stock returns (Harvey, Liu, and Zhu (2016)). By taking advantage of deep neural networks' ability to estimate complex non-linear relations in the data that are not pre-specified, these studies were able to achieve greater predictive accuracy than regression techniques typically used in their fields. However, machine learning techniques have not previously been used in the CEO turnover literature.

In the last few years, neural networks have proven themselves to be flexible and powerful machine learning algorithms for large data applications. Their use in financial and accounting literature has seen a steady increase in recent years as well. In his review, Jiang (2021) covers 124 papers that use deep learning in stock market prediction applications. Gu, Kelly, and Xiu (2019) used neural networks in the context of asset pricing.

Other popular machine learning algorithms include Ridge regression (Hoerl and Kennard (1970)), LASSO (Zou (2006)), elastic net (Zou and Hastie (2005)), gradient boosted trees (Ridgeway (2007)) and random forests (Ho (1995)). These are often featured in machine learning oriented papers where a number of methods are attempted in order to find the best performing one, or from which to create ensemble predictions (derived from averaging or otherwise aggregating predictions from a number of algorithms) (Lee, Naughton, Zheng, and Zhou (2020); Gu, Kelly, and Xiu (2019); Makridakis, Spiliotis, and Assimakopoulos (2018)). Model-based trees such as linear regression trees are less well-known methods which, like decision trees, create decision rules to partition observations in order to make a prediction, but in addition run a full OLS

regression at each final node. Generalized linear mixed-model (GLMM) trees are the particular implementation used in this paper, and were introduced in Fokkema, Smits, Zeileis, Hothorn, and Kelderman (2018).

3 Data

3.1 Data Sources

I considered a wide variety of data sources to obtain a large initial pool of explanatory variables to predict forced CEO turnovers for public firms listed in the United States. The forced turnover sample is an updated version of the one used in Peters and Wagner (2014) and follows the most commonly employed method for determining forced, rather than voluntary, executive turnover.[1] To determine executive dismissals, Peters and Wagner (2014) use a method based on Parrino (1997), who joined newspaper reports of forced dismissals to departures of CEOs under the age of 60 which were not due to death, health reasons or taking on another position. Among other factors, it also includes only retirements which were not announced at least six months beforehand. The final sample runs from 1991 to 2019 and includes 906 forced turnovers. To align with the release of financial statements and other corporate information, the frequency is annual with each observation's timing based on its firm's fiscal year.

Therefore, data which did not span this almost 3 decade long period could not be considered.[2] Explanatory variable data was gathered through the WRDS portal from Execucomp, Compustat and CRSP. Market and analyst data from CRSP were aligned with fiscal year-end data obtained from Compustat and Execucomp.

[Table 2 around here]

The final dataset has 40,098 firm-year observations and includes 78 variables across 3,521 firms. Table 2 contains the names and definitions of these series.

[1] While the forced CEO turnover sample I use is partly available on WRDS, an updated longer sample was provided to me by Florian Peters.

[2] For instance, this excluded data from otherwise popular sources such as Institutional Shareholder Services (ISS).

3.2 Data preparation

It is best practice when preparing data for most neural methods to normalize independent variables (called features in the machine learning literature) to be within a certain boundary, such as –1 to 1, in order to set all features to a common scale (Ioffe and Szegedy (2015)). The features were standardized to have mean 0 and standard deviations of 1. Outlier adjustment was treated as a hyperparameter selecting across the option to winsorize or truncate the top and bottom 1% of observations, or to otherwise leave the data unadjusted.

Dropping observations containing missing data points would mean the loss of a large percentage due to disparate data availability after merging various data sources together. Data point imputation techniques sometimes used in machine learning applications, such as decision tree imputation using other available datapoints, are not typically used in the Finance and Accounting literature and would be plagued by unreliable data availability (randomly missing other variables). The baseline method for dealing with missing data was to replace these observations by a fixed value of 0.[3] The inclusion of binary series indicating the presence of missing observations was also tested.

3.3 Empirical observations

To gain further insight into the degree of differentiation between the regressors being used, I use factor analysis and clustering to investigate if the variables are being governed by similar data trends and whether they can be convincingly grouped together based on similarity. Hierarchical cluster analysis was unstable for clusters larger than two, as more than 20% of observations changed cluster membership when compared to a non-hierarchical k-means method. As well, both Caliński and Harabasz (1974) variance ratio criterion and Duda, Hart, and Stork (1973)) stopping rule both showed the optimal number of clusters as being two, which only resulted in a coarse and unintuitive split between the series.

Through the use of a correlation based clustering method, weighted gene co-expression net-

[3] Given that series are standardized to have mean 0 and a standard deviation of 1, missing observations are thus essentially replaced by the mean of the series.

work analysis (WGCNA) (Langfelder and Horvath (2008)), I derive 9 modules.[4] Each are color coded in Figure 1's network linking series exhibiting pairwise correlations above 65%.

[Figure 1 around here]

WGCNA clusters features within each module that are highly correlated whereas features from different modules are approximately uncorrelated. The residual grey module includes variables which are not assigned to any group, meaning that they are uncorrelated to any other features.

My chief insight from this exercise leads me to believe that, after excluding series which are by construction slight variations of each other (eg: stock return, stock return including dividends, excess stock return, etc), the predictors are rather unique, with the exception of a few that are intuitively close siblings, such as book to market ratios (pink module) and measures of company size (green module). This alleviates some concerns of negative potential effects on attempted methods vulnerable to multicollinearity, such as linear regression, logit regression and linear regression trees.

4 Methodology

The implementation first makes use of a regressor (feature) selection algorithm, Random Forest Recursive Feature Elimination (RF-RFE), to select the 20 most important features that can explain forced turnover. These variables are added to the same 7 variables in Peters and Wagner (2014)'s main regression.[5]

4.1 Feature selection (RF-RFE)

The relatively limited number of firm-year observations increases the risk of overfitting. Running a feature selection algorithm in an initial step has been shown to help avoid issues of overfitting

[4] This was run using the WGCNA (version 1.70-3) package in R, through the use of the `blockwiseModules`, with the following relevant parameters: `power = 6`, `minModuleSize` of 1, and a `mergeCutHeight` of 0.25.

[5] A set of CEO turnover likelihood estimations were produced without compensation variables to properly study their effects on pay given that two of the applications study the effects of turnover risk on compensation.

(Bermingham, Pong-Wong, Spiliopoulou, Hayward, Rudan, Campbell, Wright, Wilson, Agakov, Navarro, et al. (2015)). It improves hold-out-sample performance as well as mechanically lowering training time significantly. In their neural network analysis of litigation using a similar number of observations, Lee, Naughton, Zheng, and Zhou (2020) used RF-RFE (Gregorutti, Michel, and Saint-Pierre (2017)), a well-known feature filtration method, to lower the number of independent variables from 68 to 10.

RF-RFE uses a variable importance metric derived by how prominent a variable's presence is in the random forest's constituent decision trees' splitting nodes to determine which features play a larger role in determining the outcome (forced turnover in this case).

4.2 Dealing with imbalanced class issue (SMOTE)

The relative scarcity of forced turnover events in the sample (842 of 40,098, or around 2.1%) is commonly known as an imbalanced class problem in classification machine learning applications such as this one. When optimizing a typical accuracy based prediction error function for non-deterministic data, this can lead to the algorithm strongly favoring the over-represented class in certain machine learning algorithms. An extreme case would have it always predict the absence of forced turnover to be awarded a prediction accuracy of 97.9%. A wide variety of well-studied phenomena in Accounting face this unbalanced class issue, including fraud, litigation, bankruptcy, and misstatements.

Common solutions include oversampling the uncommon outcome (e.g.: SMOTE), undersampling the common one, increasing the cost of misclassification or otherwise targeting a prediction error function which prioritizes performance across both classes. In their neural network application to predict corporate litigation events (5.6% of total observations), Lee, Naughton, Zheng, and Zhou (2020) make use of synthetic minority oversampling technique (SMOTE) (Chawla, Bowyer, Hall, and Kegelmeyer (2002)) to create enough artificial observations of the underrepresented class to balance the training data set. This improves performance across both classes for a machine learning algorithm that is trained to maximize accuracy (the percentage of correctly predicted training observations).

Simply duplicating forced turnover observations could lead to overfitting. SMOTE instead increases the sample by creating amalgamations of existing minority class observations using their same-class nearest neighbor. Each feature of the newly created observation is chosen at random from its two constituent observations. This resulted in the number of training observations (which account for 80% of the full sample) being fed into the machine learning algorithm to increase by 96%, from 32,078 to 62,858, resulting in a perfectly balanced training sample. Note that given the relatively small dataset (for typical machine learning applications) and the extent of the class imbalance, undersampling the larger class was not an option.

SMOTE was tested on all machine learning algorithms attempted in this paper, and implemented when it was beneficial to the performance of that method. While predicting forced CEO turnover can easily be interpreted as a classification problem, I instead interpret the problem as a regression, where the target variable is the probability of dismissal. Once the probabilities are estimated, the threshold which optimizes prediction performance, F1 in this case (refer to section 4 for more details), is then selected to determine classification. In practice, this process tended to decrease the benefits of SMOTE, especially for OLS, Logit and GLMM, where it either had no benefit or a slight negative effect on performance.

4.3 Ridge Regression, LASSO and Elastic Net

In comparison to ordinary least squares, which minimizes the sum of squared residuals, Ridge regression's loss function also considers the size of the regression coefficients by including the sum of the squares of coefficients (L2 normalization):

$$\arg\min_{\beta} \quad \sum_{i} \left(y_i - \boldsymbol{\beta}'\boldsymbol{x}_i\right)^2 + \lambda \sum_{k=1}^{K} \beta_k^2$$

Here λ is a hyperparameter that influences the relative weight of these two terms in the loss function. This makes slope coefficients smaller by penalizing their size, which addresses potential multicollinearity and model complexity issues, allowing for better generalization and less overfitting.

LASSO aims to achieve similar goals to Ridge Regression by instead including the sum of

the absolute value of coefficients (L1 normalization):

$$\arg\min_{\beta} \quad \sum_{i} \left(y_i - \boldsymbol{\beta}'\boldsymbol{x}_i\right)^2 + \lambda \sum_{k=1}^{K} |\beta_k|$$

Elastic Net combines L1 and L2 normalization by adding both the sum of the absolute value of coefficients and the sum of the squares of coefficients to the loss function:

$$\arg\min_{\beta} \quad \sum_{i} \left(y_i - \boldsymbol{\beta}'\boldsymbol{x}_i\right)^2 + \lambda_1 \sum_{k=1}^{K} |\beta_k| + \lambda_2 \sum_{k=1}^{K} \beta_k^2$$

For each of these models, 10 different values of λ were tested with cross-validation to select the best tuning, with all permutations tested for Elastic Net (1e-15, 1e-10, 1e-8, 1e-5,1e-4, 1e-3, 1e-2, 1, 5, 10).

4.4 Random Forest

Random Forest is a popular ensemble method that trains a cluster of decision trees applied to different bootstraps of the training dataset. For each split, a random selection of features is considered. Predictions are made by taking the average outcome of the wide variety of different trees (or majority in the case of classification), making the method more effective than an individual decision tree.

Random forests contain a number of hyperparameters, including the number of trees in the forest (100, 250, 500, 750, 1000, 1250, 1500 and 1750 were selected), the maximum number of features considered at each split (2, 3, 4, 5, 6, 7), the maximum depth of each decision tree (2, 3, 4, 5, 6, 7), the minimum number of data points placed in a node before the node is split (2, 5, 10) and the minimum number of data points allowed in a leaf node (1, 2, 4). Popular values for each of these tuning parameters bring the number of possible different configurations for the random forest into the thousands. Thus random search training was used with cross validation to find the best candidate for the most performing set of hyperparameter values. Another cross validation is made more comprehensively using tuning parameter values close to the ones found in the random search.

4.5 Gradient Boosted Tree

Gradient boosted trees is an ensemble of decision trees. However, instead of working independently as in random forests, they are built in a forward stage-wise manner, where one tree attempts to improve the prediction results of the previous tree.

For regression, the method begins by taking the average of the dependent variable in the training data as the first prediction. It then builds a tree to explain the residuals of this prediction using the available features (the size of these trees is a tuning parameter). This creates a new regression tree which will be further improved by explaining its residuals in a similar manner, the process is then repeated until a maximum number of trees is reached or the loss stops decreasing passed a certain threshold.

- Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable Loss Function $L\left(y_i, F(x)\right)$
- Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L\left(y_i, \gamma\right)$
- Step 2: for $m = 1$ to M :
  - Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1, \ldots, n$
  - Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$
  - For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L\left(y_i, F_{m-1}\left(x_i\right) + \gamma\right)$
  - Update $F_m(x) = F_{m-1}(x) + \lambda \sum_{j=1}^{J_m} \gamma_{jm} I\left(x \in R_{jm}\right)$
- Step 3: Output $F_M(x)$

To handle the overfitting issue that would arise from explaining residuals at each step, Gradient Boost scales the contribution of each tree by a learning rate λ. This is treated as a hyperparameter (values of 0.15, 0.1, 0.05, 0.01, 0.005 and 0.001 were selected for testing), along with the maximum depth of each tree, the maximum number of trees (100, 250, 500, 750, 1000, 1250, 1500 and 1750 were selected), the maximum number of features considered for each split (2, 3, 4, 5, 6, 7), the minimum number of samples to split (2, 4, 6, 8, 10, 20, 40, 60, 100) and the minimum number of samples to form a leaf (1, 3, 5, 7, 9).

Once again, popular values for each of these tuning parameters bring the number of possible different configurations into the thousands. Thus random search training was used with cross validation with a more comprehensive search around the best candidate was used to pinpoint the best tuning parameters.

4.6 Neural Network

Neural networks are often better at learning arbitrary input-output functions than competing machine learning algorithms. Their power stems from the flexibility of their architectures. Taking the neuron as the most basic building block of the neural network, each node accepts a vector of weight adjusted values, and imposes a transformation using an activation function in order to return a scalar. Most commonly used activation functions will be sigmoidal or piecewise linear (flat then angled) such that they will "activate" when given a large enough input. See Appendix A for further description and implementation details.

4.7 Generalized linear mixed-model (GLMM) tree

Generalized linear mixed-model (GLMM) tree (Fokkema, Smits, Zeileis, Hothorn, and Kelderman (2018)) is a tree-based method that employs model based recursive partioning (Zeileis, Hothorn, and Hornik (2008)) to create a decision tree with individual linear model estimations for each leaf subgroup. After an initial estimation of the linear model, the algorithm tests for parameter instability to determine the splitting variables for a user determined tree depth level. This section describes the steps of this algorithm.

Steps of the GLMM Recursive Partitioning Algorithm (glmertree)

1. **Fit OLS model to all observations in the current node by maximizing the objective function, such as the sum of squared errors or log-likelihood.**

   Consider the objective function $\Psi(Y, \theta)$ for observations Y with a k-dimensional vector of parameters θ is minimized to yield parameter estimate $\hat{\theta}$:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \Psi\left(Y_i, \theta\right).$$

with First Order Conditions:

$$\sum_{i=1}^{n} \psi\left(Y_i, \hat{\theta}\right) = 0$$

where

$$\psi(Y, \theta) = \frac{\partial \Psi(Y, \theta)}{\partial \theta}$$

is the score function, or the derivative of the objective function with respect to the parameters.

2. **Evaluate whether there is significant parameter instability with respect to partitioning variables. Select partitioning variable associated with the highest parameter instability, otherwise stop. To assess whether splitting of the node is necessary, a fluctuation test for parameter instability is performed. If there is significant instability with respect to any of the partitioning variables Z_j, split the node.**

   The score function evaluated at the estimated parameters $\hat{\psi}_i = \psi\left(Y_i, \hat{\theta}\right)$ is inspected for systematic deviations from its mean 0 with respect to partitioning variables Z_j. These deviations can be captured by the empirical fluctuation (partial sum) process:

$$W_j(n) = \hat{J}^{-1/2} n^{-1/2} \sum_{i=1}^{n} \hat{\psi}_{\sigma(Z_{ij})}$$

   where $\sigma\left(Z_{ij}\right)$ is the ordering permutation which gives the antirank (index of the smallest to the largest) of the observation Z_{ij} in the vector $Z_j = (Z_{1j}, \ldots, Z_{nj})^\top$ and $\hat{J}$ is an estimate of the covariance matrix $COV(\psi(Y, \hat{\theta}))$, e.g., $\hat{J} = n^{-1} \sum_{i=1}^{n} \psi\left(Y_i, \hat{\theta}\right) \psi\left(Y_i, \hat{\theta}\right)^\top$.

   Under the null hypothesis of parameter stability, W_j converges to a Brownian bridge with expected value of zero, as seen in Figure 2.

   [Insert Figure 2 & Figure 3 around here]

   To assess whether $W_j n$ significantly diverges from this random process (as exemplified in Figure 3), the algorithm takes the following scalar function $\lambda(\cdot)$ of $W_j(n)$ as a test statistic:

$$\lambda\left(W_j\right) = \max_{i=\underline{i},\ldots,\bar{i}} \frac{\left\| W_j\left(\frac{i}{n}\right) \right\|_2^2}{\left(\frac{i}{n} \cdot \frac{n-i}{n}\right)}$$

which is the maximum of the squared Euclidean norm of the empirical fluctuation process scaled by its variance function over the interval $[\underline{i}, \bar{i}]$ with minimum segment size $\underline{i}$, where $\bar{i} = n - \underline{i}$ (see Zeileis and Hornik (2007) for more details).

For the corresponding critical values, the limiting distribution is given by $\sup_{t \in \Pi}(t(1-t))^{-1} \left\| W^0(t) \right\|_2^2$ and approximate asymptotic p-values can be computed using the same methodology as in Hansen (1997). If no instabilities are detected, the node is a terminal node (a leaf).

3. **Compute the split point that locally optimizes the objective function.**

   In this step of the algorithm the fitted model has to be split with respect to the partitioning variable most systematically effecting parameter changes Z_{j^*} into a segmented model. Two rival segmentations can be easily compared by evaluating the segmented objective functions of each for every possible split point (Fokkema, Smits, Zeileis, Hothorn, and Kelderman (2018)). The optimal split on Z_{j^*} will be the one that minimizes the downstream segmented objective functions.

5 Results

5.1 Variable Importance

Table 3 shows the feature rankings of the RF-RFE algorithm in descending order for the top performing feature count specification of 20 for both the complete data sample and the one excluding compensation variables. Classical stock and income performance metrics had larger effects on the model than other variables, including market capitalization, leverage, total sales, and stock return.

[Table 3 around here]

In terms of interpreting variable importance for explaining forced turnover, I opt to augment RF-RFE with a different method. While RF-RFE will tend to rank regressors according to their importance in predicting the target variable, it does not take correlation into account. This means that in the case where two important variables are very similar (correlated), it is highly possible that one, say the first, ranks high while the second ranks low, given the redundancy of information. If this feature selection algorithm was to be run with only the second of these two variables, it can likely jump from being low ranked to very high ranked, for instance. This resulting instability hinders variable importance interpretability, leading me to use it in conjunction with another method that takes feature correlation into account, based on the WGCNA method used to create color modules in the data section.

[Table 4 around here]

The key here is to consider that module membership for a high ranking variable other than grey means that it can likely be substituted for another member of its group. This places the group, and not the individual series, as the relevant unit of comparison, again only for series which are members of colored groups, meaning that they have similar series in the data set. For example, stock return, total stock return and excess stock return are all highly correlated and part of the same colored module (red). While stock return (ret) appears in the top rankings for variable importance, its ranking should be interpreted as a group based one, which would imply that any other member of its highly correlated module could replace it should it be removed, such as total stock return (tot_ret). Table 4 shows the exact importance rankings derived by this process.

5.2 Performance

Raw accuracy figures can be misleading for an unbalanced class problem such as forced CEO turnover, which sees only 2.1% of its observations falling into the second of two classes. As previously mentioned, a model can simply learn to never predict forced turnovers in order to achieve 97.9% accuracy. To appreciate how well a model performed from the perspective of both

classes, researchers look at a variety of different metrics defined below, such as Precision, Recall and AUC.

Davis and Goadrich (2006) argue that, particularly for strongly imbalanced classification problems, common machine learning performance metrics such as area under the receiver operating characteristic curve (AUC) are inferior relative to metrics that provide more sensitivity to changes in the true positive rate, such as the F-score, which is based on the trade-off between Precision and Recall. Precision is the ratio of true positives (TP) over true positives (TP) plus false positives (FP) (Precision = TP / [TP + FP]), and can be interpreted as the accuracy of a model's prediction conditional on a positive forecast. Recall, also known as the total positive rate, is equal to the ratio of true positives to true positives plus false negatives (FN) (Recall = TP / [TP + FN]), and can be interpreted as the share of sample positives accurately predicted by the algorithm. ROC AUC methods on the other hand are based on the trade off between Recall and the False Positive Rate (FPR), which is equal to the ratio of false positives relative to true negatives (TN) plus false positives (FPR = FP / [FP + TN]). Finally, the familiar accuracy metric is simply the sum of true predictions over total predictions ([TP + TN]/ N).

Table 8 contains the results of each machine learning algorithm chosen for evaluation. Training Accuracy, Precision and Recall for the testing sample are found in columns (1), (2) and (3) respectively, from the perspective of a forced turnover being a positive result. The F1 score in column (4) balances these last two measures using their harmonic mean.

[Table 8 around here]

Metrics such as Recall and the False Positive Rate are calculated using the resulting performance at a particular threshold. The ratio of these two measures mapped at varying thresholds is called the Receiver operating characteristic (ROC) line, with the area under the curve (AUC) being an overall measure of performance. Analogous to this is the Precision-Recall Curve which also maps the trade-off between Precision and Recall (the same components as the F-score), with Precision-Recall Area Under the Curve (PRAUC) quantifying overall performance under relevant thresholds.

Thus both the AUC and PRAUC (columns (5) and (6) respectively) take into consideration Recall. However, for a binary data set with a small number of positive events, the Precision metric is much more sensitive to changes in the algorithm's ability to accurately predict positive events relative to the FPR. This is due to the much larger relative importance of negatives in the data set, which makes TN large in FPR's ratio and renders it less sensitive to performance improvements. This dynamic favors the F-score and PRAUC as the metrics of choice to measure performance in a highly imbalanced classification problem. These were thus the metrics used to rank model specifications.

Using PRAUC as our performance metric of choice, the best performing machine learning algorithm was the GLMM tree by a wide margin, followed by Gradient Boosted Tree (GBT), Random Forests (RF), Logit and OLS. Surprisingly, neural networks lagged severely behind these simpler methods, even when using a wide variety of specifications and optimizing on hyperparameters (Table A1 & Table A2).

By comparing testing performance (columns 7 through 12) versus training performance (columns 1 through 6), we can see how out-of-sample results deviate from in-sample to detect potential issues with overfitting. Overfitting issues are observed when algorithms explain (training) data they have already seen significantly better than new (testing) data. Table 8 has a 80-20 training to testing split. Judging from AUC and PRAUC, while certain low performing neural networks (CONV, HG, FF) see larger gaps between training and testing, the remaining algorithms show only a mild decrease in performance. The only exception being Random Forest, which performed very well in training but lagged GBT and GLMM in testing. GLMM's training F1 of 34.8% and PRAUC of 30.5% decreased to 30.6% and 25.9% respectively in testing. This was robust to boostrapping 1000 iterations using 80-20 splits. The average showed a similarly small decrease in out-of-sample performance relative to in-sample.

The neural networks estimated in Table 8 were stopped after 500 training epochs. Increasing epochs would consistently lead to large overfitting issues, as can be seen in Table A1 where the best performing neural networks would only achieve testing AUC/PRAUC of 53% and 2.6% but 92.2% and 18.7% in-sample results (Table A2 shows a similar problem).

As can be seen in Table 8,[6] the best performing GLMM tree trained on the same sample as Peters & Wagner (2014) achieved an F1 score of 38.6%, AUC of 94.1% and PRAUC of 32.9%. This compares to an F1 score of 24.2%, AUC of 87.6% and PRAUC of 15.5% for the OLS model run with the same features and sample. Performance for the best performing network on the complete data sample without compensation variables included was similar though lower (in general this was true for all models and specifications attempted), with an F1 score of 33.7%, AUC of 96.4% and PRAUC of 31.5%. Note the very marginal improvement in performance from the addition of two compensation variable between the last two panels of the table.

[Table 8 around here]

To compare, the best performing neural network trained on the entire sample achieved an F1 score of 16.2%, AUC of 92.2% and PRAUC of 18.7%. However this neural network suffered from major overfitting issues, as the aforementioned in-sample results greatly outstripped the hold-out sample's, such as 6.9%/60.7%/3.6% for F1/AUC/PRAUC as seen in Table 8. In comparison, testing sample results for GLMM tree (20.5%/93.8%/23.8% for F1/AUC/PRAUC) were very in line with training sample performance. In addition to performance, GLMM proved to have a much better generalizeability in this application.

In addition to superior performance, GLMM tree also maintains a great advantage over neural networks in the interpretability of its results. The resulting linear model-based tree is displayed in Figure 4, for both trees at depths of 2 (1 split and 2 linear models, one for each subgroup), and depths of 3 (3 splits and 4 linear models. one for each subgroup).

The precise node level regressions are detailed in Table 7. In general, performance metrics (such as EBT, revenue growth, EPS, etc) with significant coefficients had a negative impact on forced turnover, where better performances lower the likelihood of dismissal. Outside CEO hires were more likely to be fired for CEOs with a short tenure (nodes 1 and 2), and CEO age also had a positive effect in node 2. CEO tenure significantly decreased the chance of forced turnover

[6] The results in Table 5 were derived using the extended sample without the use of compensation variables or industry fixed effects. Table 6 results were produced using the entire sample, while Table 5 uses a training and testing split.

in nodes 1 and 2, with a coefficient of -0.55 for CEO with a tenure one year over the ones with no tenure (node 1), and a coefficient of -0.15 for every year beyond that (node 2). Stock market returns (both idiosyncratic and industry adjusted) paradoxically increased the likelihood of dismissal in node 2, while lowering it in node 4. Assets decreased the probability of forced turnover in node 2, while the book-to-market ratio (btm) increased it within the same node. Within node 3 (CEOs with over a year of tenure for companies with bottom percentile weighted earnings-before-tax), return on assets (roa) decreased the probability of dismissal, while ebt increased it. Within node 4 was the node with the highest number of observations (CEOs with over a year of tenure for companies with percentile weighted earnings-before-tax above the third percentile), effect sizes on significant regressors were small.

[Table 7 around here]

To illustrate the difference in optimal GLMM tree performance to the baseline Peters and Wagner (2014) OLS model a bit further, Table 8 and Table 9 contain the confusion matrices of both for the full sample. Here we can see that the optimal GLMM tree's additional accuracy relative to the OLS is driven by large drops in false positives (3,320 vs 7,835) and false negatives (22 vs 161), leading to a performance gain in true negatives (17,847 vs 13,532) and true positives (402 vs 264).

[Table 8 and Table 9 around here]

The confusion matrices above are evaluated at a particular threshold, such as 0.5, meaning that only predicted likelihoods above 0.5 are assumed to be forced turnovers. Lowering the threshold will tend to increase the true positive rate (the ratio of true positives to the sum of true positives and false negative forecasts), as less certain (lower) predictions get interpreted as positive. Decreasing the threshold will also increase the false positive rate (the ratio of false positives to true negatives + false positives).

In other words, lowering the boundaries at which a prediction is deemed to be positive correctly captures more true positive observations, at the cost of also capturing more false

positives. To account for this, ROC (Figure 5) and the PR-Curve (Figure 6) are calculated at varying thresholds, with the area under each curve being an overall measure of performance. Performances from model replications of Peters and Wagner (2014) and Bushman, Dai, and Wang (2010) were added to each figure for comparison.

[Figure 5 and Figure 6 around here]

6 Implications

I take advantage of the more accurate estimates of CEO turnover risk from section 4 by using them to test the robustness of some contradictory findings in the literature regarding the effects of dismissal likelihood on CEO compensation and firm performance. Peters and Wagner (2014) found that increases in forced turnover risk were related to increases in CEO pay. Meanwhile, Bushman, Dai, and Wang (2010) find that a higher estimated risk of dismissal decreases executive pay-performance-sensitivity.

Finally, Huson, Malatesta, and Parrino (2004) showed a positive but not statistically significant effect of forced CEO turnover on future firm-level performance. I use the improved measure of CEO turnover risk as a Propensity Score Weighting (PSW) to see if less selection bias in selecting treatment and control groups impact this result.

6.1 Bootstrapping standard errors

For Peters and Wagner (2014), the linear probability model yielded consistent second-stage estimates. This does not hold true for the herein proposed first stage estimated through GLMM (Bennedsen, Nielsen, Pérez-González, and Wolfenzon (2007), Oyer (2008)). To provide asymptotically accurate estimates of the second stage coefficients' standard errors, I bootstrap them using the following steps for 1000 iterations:

- Draw $B = 1000$ bootstrap samples of the first stage data set.
- For each $b = 1, \ldots, B$:

- Estimate the GLMM model using the bootstrapped first stage data set for replicate b; call this model m_b.
- Estimate the second stage OLS model using the prediction from m_b as a regressor, and using the bootstrapped second stage data set for replicate b. Let β_b be the estimated slope of this prediction in bootstrap replicate b.
- Given the set of estimated slopes (β_1, ..., β_B), compute the standard error of this set by taking the sample standard deviation and dividing by $\sqrt{B}$).

These boostrapped standard errors are reported for each model that used first stage forced turnover probability estimates generated by GLMM (column 1 in Table 10, columns 2 and 4 in Table 11).

6.2 Relationship between probability of forced turnover and compensation

6.2.1 Peters & Wagner (2014)

Peters and Wagner (2014) use the predicted probabilities of forced CEO turnover estimated in a first-stage linear regression (1) as an input in explaining CEO compensation in a second-stage regression (2). They find that the increased risk of dismissal is related to an increase in pay, and interpret it as being used to compensate the executive for the inherent risk of accepting the position due to potential forced turnover.

$$Forced\,Turnover_{it} = \alpha_1 + \beta_1' X_{it} + \epsilon_{1it} \tag{1}$$

$$Compensation_{it} = \alpha_2 + \beta_2' X_{it} + \gamma_2 \widehat{Forced\,Turnover}_{it} + \epsilon_{2it} \tag{2}$$

Table 10's first column displays Peters and Wagner (2014)'s basic result from Table IV, column 1. I replicate their findings in column 2 using the entire available sample. When producing the probability of forced turnover using the best performing GLMM tree specification, the intersection of the merged databases resulted in a somewhat smaller sample than the replication's (21,792 vs 24,453). Column 3 shows the replication's results are robust to this. Forced turnover probability is shown to have a highly statistically significant and positive relationship with

total CEO compensation. Finally, column 4 uses the likelihood of forced turnover estimated by the GLMM, and shows a still significant negative coefficient that is however not economically material at less than 1% of its former size. To give additional context, while a 1% increase in the likelihood of forced turnover would have resulted in a $233,678 total salary increase using Peters and Wagner (2014)'s coefficient, the coefficient obtained from the GLMM tree predictions result only in a $806 increase in pay for each percentage point increase in the risk of dismissal.

[Table 10 around here]

This result is less consistent with the authors' suggested dynamic: that volatile firms must pay CEOs more generously to compensate for decreased executive employment stability. The exact gain in prediction performance over Peters and Wagner (2014) can be clearly seen in the confusion matrices in Tables 5 & 6, namely an almost 1,000% gain in PRAUC and 300% gain in the F1 score.

Peters and Wagner (2014) focus on causal identification via two-stage least-squares regression, by using industry stock return volatility (computed from monthly equal-weighted returns of 48 industries) as a way to instrument a CEO's risk of being fired. That executive compensation would be correlated with industry volatility only through the risk of dismissal, even with the added control variables, could easily be argued against — a weak instrument affects results and any inferences made therefrom. I do not use this type of IV, which departs from this inferential perspective, however, the associative analysis remains similar.

6.2.2 Bushman et al. (2010)

Bushman, Dai, and Wang (2010) similarly study the relationship between forced turnover probability and the change in total compensation, but instead find a significant negative relationship between the two. This is consistent with a dynamic in which poor firm-level performance, leading to a higher estimated risk of forced CEO turnover, is met with lower compensation in lieu of dismissal. The first stage regression from Bushman, Dai, and Wang (2010) follows the same general form as equation (1) in Peters and Wagner (2014) but with different regressors and makes use of a logit model instead of a linear probability model. The second stage regression

in the first two columns of Table 8 uses the year-over-year change in compensation rather than the natural logarithm of total CEO compensation, which appears in the last two columns for comparison's sake.

$$YoY\Delta\ \ Compensation_{it} = \alpha_2 + \beta_2' X_{it} + \gamma_2 \widehat{Forced\ Turnover}_{it} + \epsilon_{2it} \tag{3}$$

I did not have access to their control and treatment groups, as the authors augmented the sample derived from the Parrino (1997) method with their additional discretionary judgments of which retirements below the age of 60 were deemed suspect enough to be categorized as forced turnovers. Running the same regression on the same sample as Peters and Wagner (2014) but using Bushman, Dai, and Wang (2010)'s time period of 1994-2005, I similarly find a negative relationship between pay-performance-sensitivity (PPS) (Table 11, column 1) and the risk of forced turnover, as well as a marginally significant ($p < 0.10$) negative relationship for the change in total compensation (column 3). As in their paper, these were derived using a logit model in a first step. When using the likelihoods derived through the GLMM tree, while the coefficient on PPS remains negative and statistically significant, it shrinks to less than 9% of its former size (column 2). Meanwhile, the coefficient on the change in total compensation shrinks to around 5% of its former size, while remaining statistically not significant at conventional levels (column 4).

Peters and Wagner (2014) interpret the positive relationship between risk of dismissal and pay as a risk premium paid to executives to compensate for the likelihood of being fired. However Bushman, Dai, and Wang (2010) and Gao, Harford, and Li (2009) posit that smaller pay increases for a CEO can act as a substitute for asking them to step down.

[Table 11 around here]

6.3 Improving Inference through ML derived Propensity Score Weightings

Propensity Score Weighting (PSW) (Abadie and Imbens (2002)) is a method which uses the likelihood of a given observation to have been treated to correct for sample selection bias, with control observations estimated to have had a higher propensity for having been treated being

weighted more heavily in the regression. PSW weights have often been estimated using logistic regression models in an initial step, although they are not limited to using that methodology by design.

A more accurate measure of forced turnover probability can assist in providing cleaner control and treatment sample weightings when doing causal inference. I illustrate this in practice in a context inspired by Huson, Malatesta, and Parrino (2004). The authors study the impact of forced CEO turnover on firm performance from one year before the executive shuffle to three years after, using voluntary turnovers as the control sample. They find a positive though not statistically significant relation at common levels ($p > 0.10$). I find very similar results using a comparably sized but more recent sample of 2006-2016 when running a similar regression (Table 12, column 1). I run this regression keeping only observations which experienced a CEO turnover the same year, thus also using voluntary turnovers as the control sample.

[Table 12 around here]

In column 2, I run the same regression using the probability of forced turnover estimated through a logit regression as propensity score weight, using the same previously determined 20 features as regressors. The resulting coefficient is marginally larger but still not statistically significant. In column 3, I use the predicted probability of forced turnover estimated from the GLMM tree, again using the same independent variables. The relationship between forced turnover and firm performance is materially larger but this time also statistically significant ($p < 0.05$). This shows that the accuracy of treatment likelihood can significantly affect estimation inference when using PSW, potentially reducing selection bias in selecting treatment and control samples, helping to mitigate the classical problem in causal inference. In this case, the GLMM tree based PSW strengthens Huson, Malatesta, and Parrino (2004)'s finding that forced management turnover is motivated by a desire to find a more able executive that will increase firm performance, or what the authors refer to as the improved management hypothesis.

7 Conclusion

Beyond providing improved estimates of CEO turnover likelihood through the use of generalized linear mixed-effects model (GLMM) trees, showing this method as superior to other machine learning algorithms, this paper uses them to improve upon the understanding of their relationships to executive compensation and firm performance. Such an approach can be used as an example of a constructive way to make use of more accurate machine learning predictions to test the robustness of previously studied inferences, particularly for probability based theoretical constructs, such as the risk of bankruptcy, misstatements, fraud and litigation.

Although neural networks rank among the most powerful machine learning models, simpler methods outperformed them by a wide margin in this setting. They also did not suffer from the same overfitting issues that surfaced when training neural networks, and did not require the use of synthetic data to balance the training data set to make competitive predictions. In addition, optimal neural network implementations hinge on a very large permutation of hyper parameters, including a near infinite number and variety of layers and node configurations, as well as unclear but not trivial decisions on the input order of series for company level data. This is further exacerbated by their lengthy training run times. All this significantly complicates a researcher's task when implementing deep learning methods.

Finally, relative to neural networks, these simpler methods also have the net advantage of providing very interpretable results to researchers, which helps to further theoretical and empirical knowledge about the determinants of forced CEO turnover.

However methods such as recurrent neural networks (RNN) and inverse reinforcement learning (IRL) can potentially provide further insights through their ability to interpret temporal data and model not only complex interactions between features, but over time as well. IRL is particularly well suited to estimating utility functions, but requires a relatively small and discrete feature state space to obtain meaningful estimations. RNN, a method often used in speech recognition and language modeling, is more flexible and well-suited to predicting the next element of a series by studying how their recent explanatory features behave over time.

While improving predictions of some of uncommon accounting events have inherent value

to practitioners, their potential for improving empirical understanding in research applications is so far untapped. In addition, well-described machine learning implementations contribute to the eventual establishment of best-practices in such settings, providing important guidance and time-saving resources to future researchers.

8 Figures & Tables

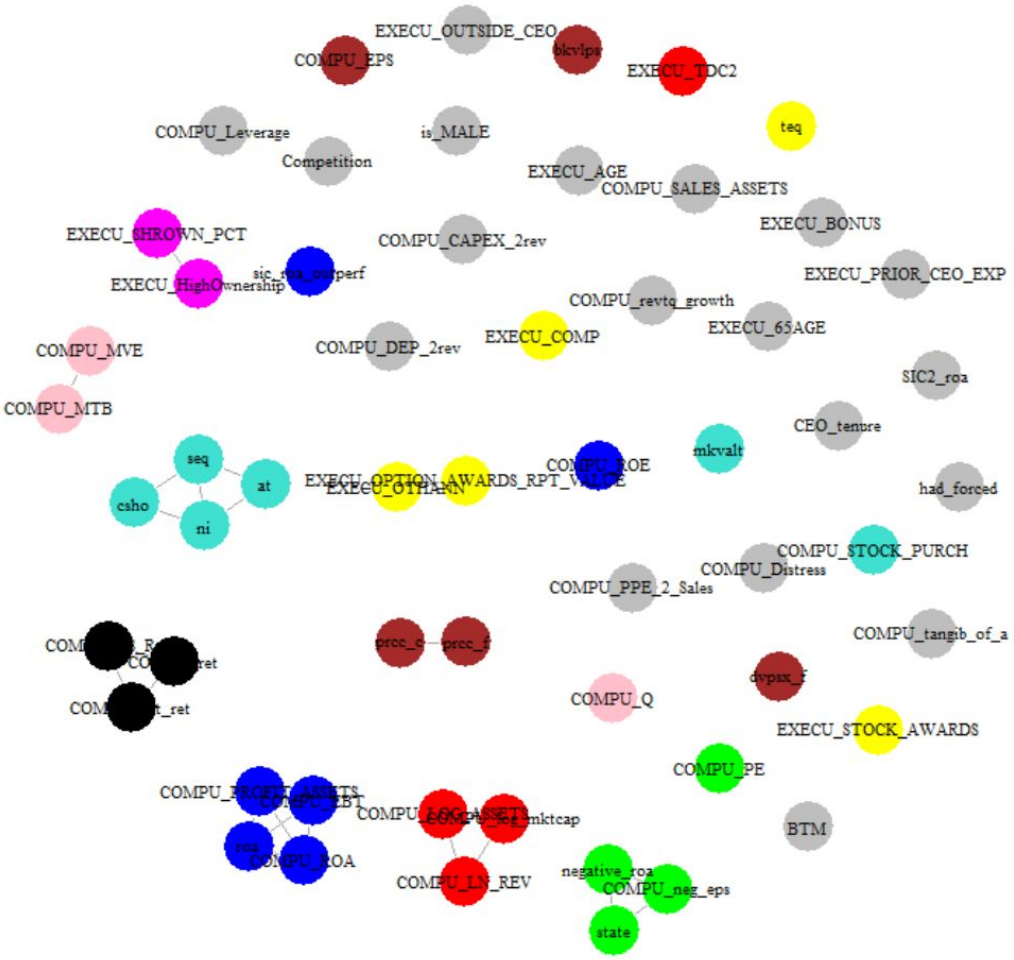

Figure 1: WGCNA Color Module Network

Note: The color network displays the network of series by their module membership. Series are placed into modules using a correlation based method, with those joined by links exhibiting correlations above 65%. Series in grey do not exhibit any strong correlation with other series in the data set.

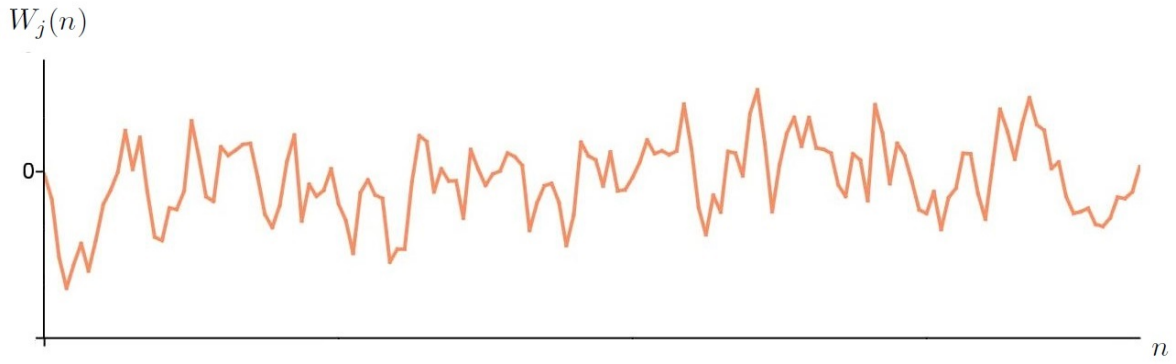

Figure 2: Score function fluctuations with no systematic relationship to the partitioning variable (X axis)

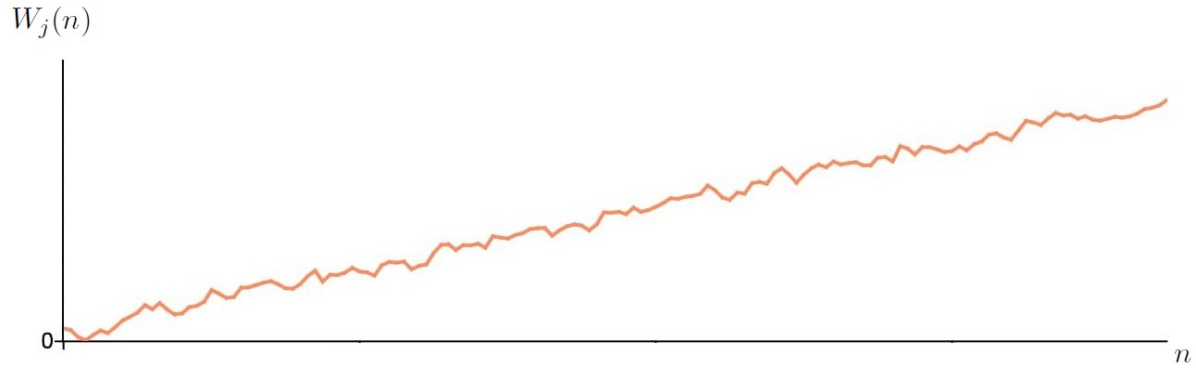

Figure 3: Score function fluctuations with systematic relationship to the partitioning variable (X axis)

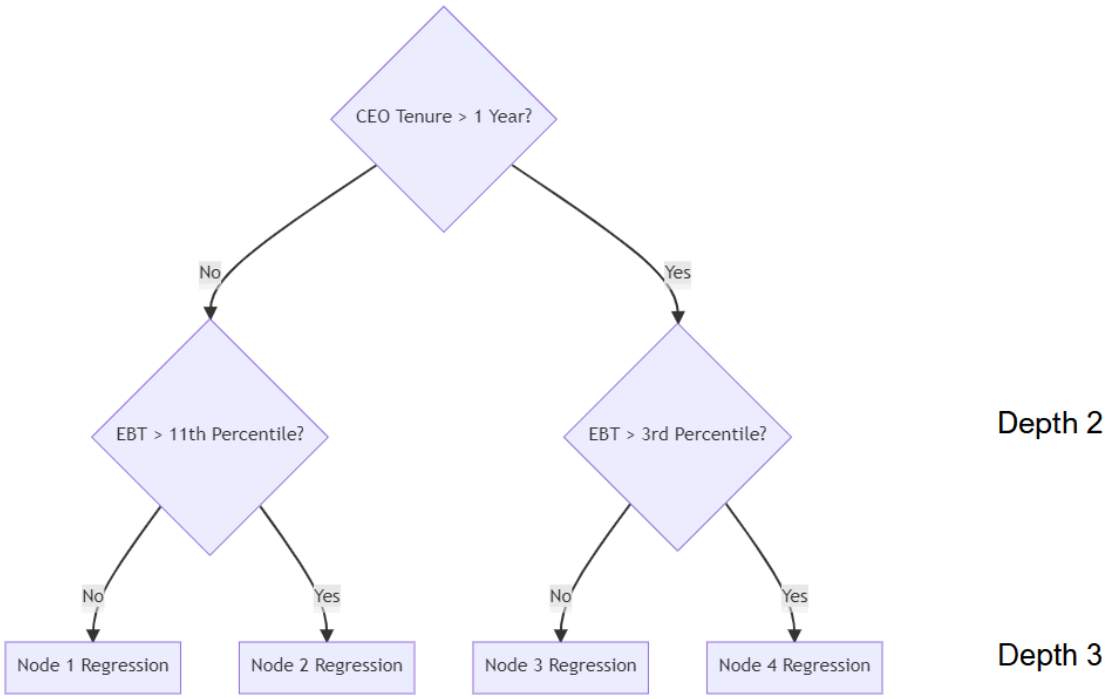

Figure 4: GLMM Tree

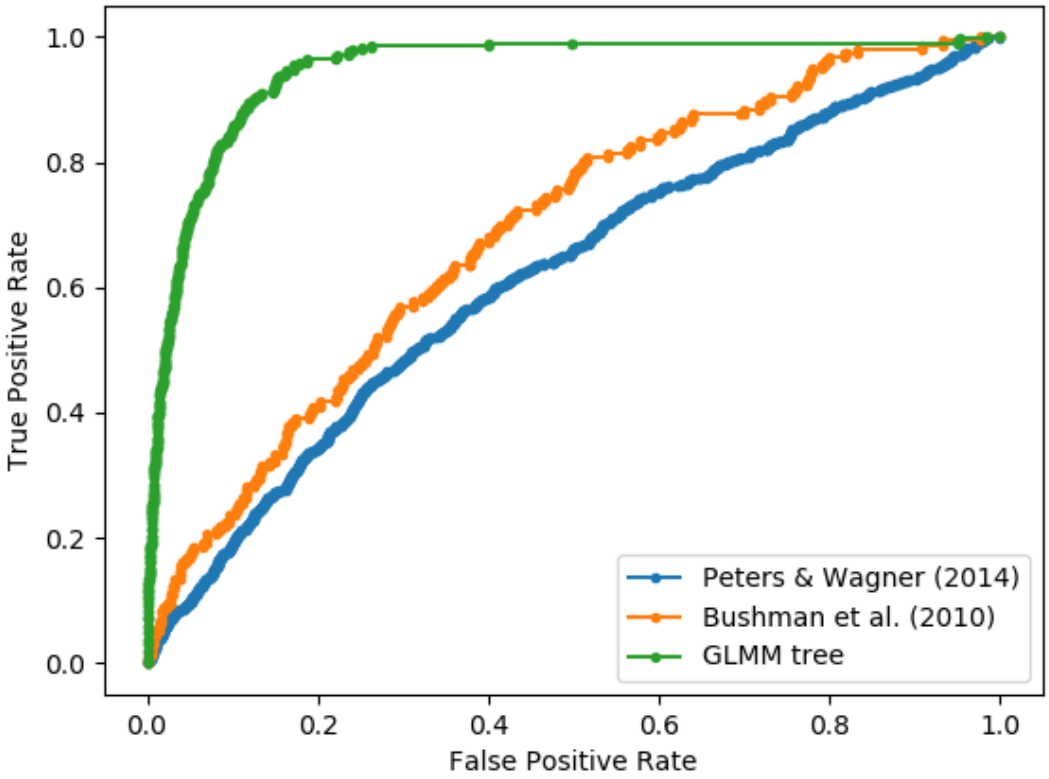

Figure 5: ROC Curve

Note: *ROC*, or Receiver operating characteristic, is a line representing the model's diagnostic performance under different prediction thresholds. *True positive rate*, also known as the Recall, is the ratio of true positives to true positives + false negatives. *False positive rate* is the ratio of false positives to true negatives + false positives. *No skill* represents a model that would consistently predict a random class or a constant class. *GLMM tree* is the ROC curve of the best performing GLMM tree trained on the same sample period as Peters & Wagner (2014) with the addition of 20 extra explanatory variables *Peters & Wagner (2014)* and *Bushman et al. (2010)* are the ROC curves of the each of these paper's replicated models using their respective samples.

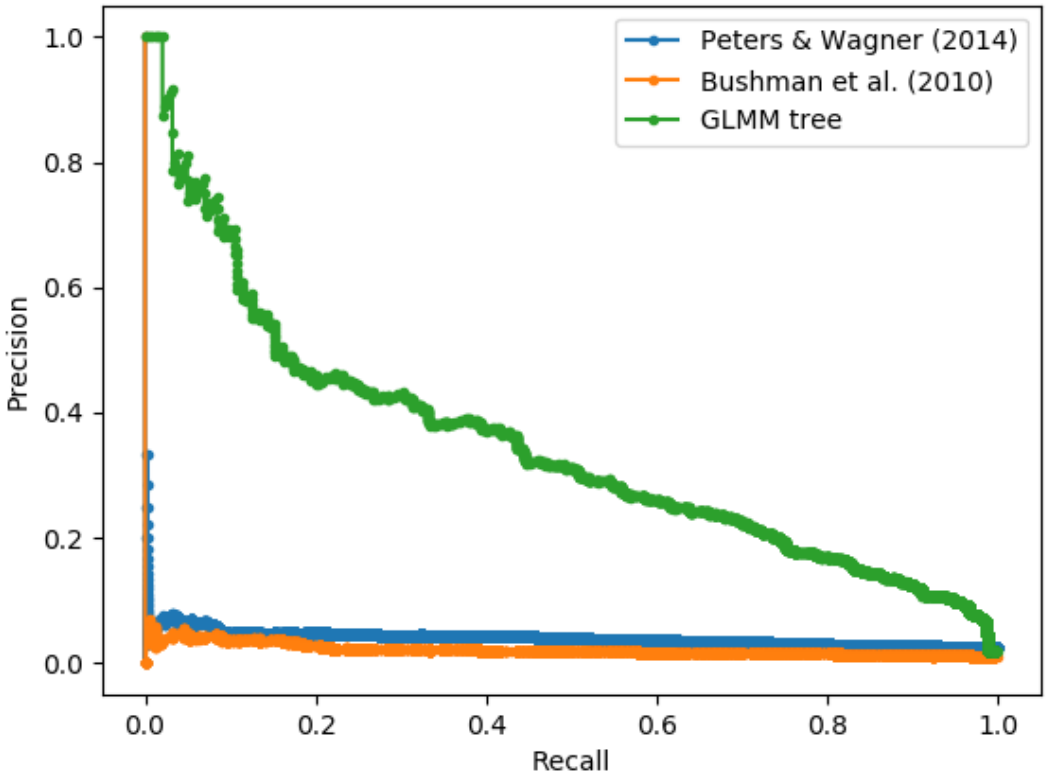

Figure 6: Precision-Recall Curve

Note: *Precision-Recall Curve*, is a line representing the model's diagnostic performance under different prediction thresholds. *Recall*, also known as the true positive rate, is the ratio of true positives to true positives + false negatives. *Precision* is the ratio of true positives over true positives + false positives. *GLMM tree* is the Precision-Recall curve of the best performing GLMM tree trained on the same sample period as Peters & Wagner (2014) with the addition of 20 extra explanatory variables. *Peters & Wagner (2014)* and *Bushman et al. (2010)* are the Precision-Recall curves of the each of these paper's replicated models using their respective samples.

Table 1: Determinants of CEO Turnover in Past Literature

Type	Variable	Paper
Firm Performance	Stock return	Warner et al. (1988)
	Earnings	Farrell & Whidbee (2003)
	ROA	Huson et al. (2004)
	Relative to industry	Eisfeldt & Kuhnen (2013)
	Relative to analyst forecasts	Farrell & Whidbee (2003)
	Relative to management forecasts	Lee et al. (2012)
	Relative to past performance	Minutti-Meza et al. (2020)
	Incidence of lawsuits	Niehaus & Roth (1999)
	Media reports	Farrell & Whidbee (2002)
CEO Characteristics	Age	Bushman et al. (2010)
	Tenure	Dikolli et al. (2014)
	Education	Bhagat et al. (2010)
	Founder	Mobbs (2013); Beneish et al. (2017)
	Related to founding family	Huson et al. (2001)
	Voting Power	Guo & Masulis (2015)
	Chairman CEO	Helwege et al. (2012)
	Part of BoD	Fiordelisi & Ricci (2014)
	Part of Compensation Committee	Fiordelisi & Ricci (2014)
	Severance pay agreements	Inderst & Mueller (2010)
	Option grants	Campbell et al. (2011); Laux (2012)
	Ease of replacement	DeFond & Park (1999)

Firm Characteristics	Earnings management	Hazarika et al. (2012)
	Earnings volatility	Engel et al. (2003)
	Disclosure informativeness	Bochkay & Chychyla (2019)
	Corporate governance	Defond & Hung (2004)
	Anti-takeover provisions	Bushman et al. (2010)
	Misreporting and restatements	Desai et al. (2006); Hennes et al. (2008); Burks (2010)
	Presence of suitable internal replacement	Mobbs (2013)
	Competition	Goyal & Park (2002)
	Firm Age	Bushman et al. (2010)
Board Characteristics	Independence	Guo & Masulis (2015)
	Monitoring Incentives	Fich & Shivdasani (2006)
	Presence of Outside Directors	Laux (2008); Fiordelisi & Ricci (2014)
	Staggered terms	Laux (2008)
Regulatory Environment	Accounting regulations	Burks (2010); Meng (2020)
	Laws regarding CEO termination	Cornelli et al. (2013)
Other	Loss of human capital	Sliwka (2007)

Table 2: Data Series: Variable Definitions

Variable	Definition
3y_stock_return	3 year total stock return
ad_chg	Change in Advertising expense
age	CEO age
atlman_z	Atman Z-Score
bkvlps	Book value per share
bonus	CEO bonus
capex_2rev	Capex / Sales
capex_chg	Change in Capex
cf_per_share	Cash flow to stock price
cfo	Cash from operations
comp	CEO Compensation (Salary + Bonus)
competition	Number of firms in the same 2-digit SIC code
coo	Indicator that a COO is present
dep_2rev	Depreciation / Revenue
disc_op	Ln(Discontinued operations)
distress	Indicator that ROA has been falling for 2 years straight
dvpsx_f	Dividend per share
ebit	Ln(Earnings before interest and tax)
ebitda	Ln(Earnings before interest, tax, depreciation and amortization)
ebt	Ln(earnings-before-taxes)
employees	Number of employees
eps	Earnings-per-share
firm_age	Firm age (based on year firm first appears in CRSP database)
gender	CEO gender indicator
historical_earnings_chg	Historical average earnings growth rate (expanding window)
ind_adj_ebt	Year-over-year net income growth - 2-digits SIC net income growth
leverage	Liabilities / Shareholder's Equity
ln_rev	Ln(Total Sales)
log_mktcap	Ln(Market Capitalization)
mtb	Market Capitalization / (Total Assets - Total Liabilities)
nb_falling_earnings_q	Number of quarters with greater than a 50% decline in earnings
neg_eps	Indicator if earnings-per-share is below zero
negative_roa	Indicator if ROA is negative
ni	Ln(Net Income)
option_awards_rpt_value	Options Granted (As Reported by Company)
othann	CEO: Other annual income
outside_ceo	Indicator that CEO is outside hire
pe	Price-to-earnings ratio
ppe_2_sales	Property, Plant, and Equipment / Total Sales
prior_ceo_exp	Indicator that CEO has prior CEO experience
profit_assets	Net Income / Total Assets
q	Market Capitalization / Total Assets
ret	Year-over-year stock return (excluding dividends)
retire_age	Indicator that CEO is above the age of 64
revtq_growth	Year-over-year revenue growth
rnd_chg	Change in R&D
rnd_to_a	R&D Expenditure / Total Assets
roa	Net Income / Total Assets
roe	Net Income / Shareholder's Equity
sales_assets	Total Sales / Total Assets
shrown_pct	Percentage of shares owned by CEO
sic_roa_outperf	Firm stock performance - 2-digit SIC industrial stock performance performance (YoY)
sic2_roa	2-digit SIC industrial stock performance performance (YoY)
std earnings	Standard deviation of 4 years of quarterly earnings
std returns	252 day standard deviation of daily stock returns
stock_awards	Value of stock awards
stock_purch	Ln(stock puchases)
tangib_of_a	Intangible Assets / Total Assets
tdc2	Total Compensation (Salary + Bonus + Other Annual + Restriced Stock Grants + LTI)
tenure	CEO tenure (in years)
teq	Ln(Total Shareholder's Equity)
tot_ret	Year-over-year stock return (including dividends)
xs_ret	Year-over-year stock return (including dividends) - year-over-year S&P 500 return
youngfirm	Equals 1 if the firm has been listed on Compustat for five or fewer years, 0 otherwise.

Table 3: RF-RFE: Top 20 Variable Rankings

Rank	**Variable**	
	No Compensation Data	*All Data*
1	log_mktcap	capex_2rev
2	roa	leverage
3	roe	dep_2rev
4	leverage	mtb
5	dep_2rev	log_mktcap
6	tangib_of_a	ln_rev
7	ln_rev	ret
8	ret	tot_ret
9	sales_assets	xs_ret
10	revtq_growth	sales_assets
11	eps	revtq_growth
12	ebit	ebit
13	pe	pe
14	ebt	log_assets
15	log_assets	at
16	ppe_2_sales	bkvlps
17	bkvlps	ni
18	ni	age
19	shrown_pct	tdc2
20	btm	comp

Table 4: RF-RFE: Top 10 Variable Rankings

Rank	**Mnemonic**	**Description**
1	Leverage	firm leverage
2	DEP_2rev	depreciation to sales
3	revt_growth	revenue growth
4	black eigengene	stock returns module
5	blue eigengene	profitability ratios module
6	brown eigengene	earnings-per-share module
7	green eigengene	negative profit indicators module
8	pink eigengene	market valuation ratios module
9	red eigengene	firm size module
10	yellow eigengene	executive compensation module

Table 5: Machine Learning Model Performances

	Training Sample Performance						**Testing Sample Performance**					
Algo.	Acc. (1)	Prec. (2)	Recall (3)	F1 (4)	AUC (5)	PRAUC (6)	Acc. (7)	Prec. (8)	Recall (9)	F1 (10)	AUC (11)	PRAUC (12)
RIDGE	91.8%	7.0%	50.9%	12.4%	88.1%	6.0%	90.3%	6.1%	52.9%	11.0%	84.9%	4.8%
Enet	93.0%	7.2%	42.9%	12.3%	86.9%	5.8%	93.3%	6.7%	40.6%	11.5%	85.5%	5.5%
LASSO	93.4%	7.6%	44.4%	13.0%	87.4%	6.4%	90.9%	6.0%	43.8%	10.5%	83.6%	4.5%
CON NN	51.4%	2.7%	65.5%	5.2%	61.3%	2.8%	50.5%	2.2%	54.7%	4.2%	51.6%	2.0%
HG NN	89.2%	5.9%	57.0%	10.7%	86.1%	5.0%	89.3%	4.8%	45.6%	8.6%	80.9%	3.5%
FF NN	86.8%	5.5%	65.9%	10.2%	82.2%	7.0%	87.5%	5.1%	58.8%	9.4%	74.4%	4.2%
IHG NN	95.5%	11.0%	41.9%	17.4%	74.4%	5.7%	95.5%	9.3%	35.3%	14.8%	74.0%	4.8%
GEO NN	85.4%	4.0%	52.3%	7.5%	75.6%	6.1%	84.6%	4.5%	63.2%	8.3%	79.0%	7.9%
OLS	97.6%	15.4%	28.8%	20.1%	89.9%	11.6%	98.0%	13.0%	20.3%	15.9%	86.9%	8.3%
Logit	92.6%	14.6%	49.7%	22.6%	87.3%	14.1%	94.0%	12.4%	39.0%	18.8%	84.0%	9.8%
RF	100.0%	99.6%	100.0%	99.8%	100.0%	100.0%	97.7%	21.9%	16.7%	18.9%	89.5%	11.3%
GBT	98.4%	29.4%	36.6%	32.6%	96.3%	22.3%	96.6%	18.2%	47.0%	26.2%	94.2%	18.1%
GLMM	98.1%	28.5%	44.8%	34.8%	96.2%	30.5%	99.2%	34.7%	27.4%	30.6%	97.9%	25.9%

Note: *Acc.* is accuracy, the percentage of correctly predicted sample observations. *Prec.* is precision, the ratio of true positive to total positives. *Recall* is the ratio of true positive to true positives + false negatives. *F1* score is the harmonic mean of *Precision* and *Recall*. *AUC* is the Area Under the Curve of the Receiver operating characteristic (ROC) line. *PRAUC* is the Area Under the Curve of the Precision-Recall line. Training and testing samples were based on an 80-20 split. All algorithms were run using SMOTE to synthetically balance the sample, except for *OLS*, *Logit* and *Lmertree*.

Table 6: GLMM Tree Model Full Sample Performances

Nb of features* (1)	Tree Depth (2)	Year FEs (3)	Industry FEs (4)	Accuracy (5)	Precision (6)	Recall (7)	F1 (8)	AUC (9)	PRAUC (10)
Peters & Wagner (2014) variables (1996-2009)									
7	1	Yes	No	95.1%	6.4%	11.0%	8.1%	70.3%	4.1%
7	2	Yes	No	91.7%	6.8%	25.0%	10.6%	72.7%	5.0%
7	3	Yes	No	91.3%	6.9%	27.4%	11.0%	73.0%	5.3%
Peters & Wagner (2014) variables with CEO tenure (1996-2009)									
8	1	Yes	No	90.2%	10.1%	50.9%	16.9%	86.7%	10.3%
8	2	Yes	No	94.9%	17.1%	41.1%	24.2%	91.4%	16.3%
8	3	Yes	No	94.6%	18.7%	51.2%	27.4%	92.5%	18.4%
Peters & Wagner (2014) variables with extra 20 (1996-2009)									
27	1	Yes	No	96.0%	18.5%	29.8%	22.8%	87.2%	14.4%
27	1	Yes	Yes	97.1%	25.3%	23.2%	24.2%	87.6%	15.5%
27	2	Yes	No	96.1%	24.6%	47.3%	32.4%	92.7%	24.8%
27	3	Yes	No	96.7%	28.7%	47.0%	35.7%	92.9%	29.8%
27	2	Yes	Yes	96.0%	26.0%	54.8%	35.3%	94.0%	28.9%
27	3	Yes	Yes	97.0%	32.5%	47.6%	38.6%	94.1%	32.9%
Expanded sample (1996-2019)									
27	2	Yes	Yes	97.9%	26.4%	45.7%	33.5%	95.8%	27.4%
27	3	Yes	Yes	98.0%	26.6%	45.8%	33.7%	96.4%	31.5%
Expanded sample (1996-2019) with compensation variables									
29	2	Yes	Yes	98.5%	33.3%	34.2%	33.8%	95.9%	27.5%
29	3	Yes	Yes	98.6%	33.4%	34.4%	33.9%	96.5%	31.7%

Note: *Accuracy* is the percentage of correctly predicted sample observations. *Precision* is the ratio of true positive to total positives. *Recall* is the ratio of true positive to true positives + false negatives. *F1* score is the harmonic mean of *Precision* and *Recall*. *AUC* is the Area Under the Curve of the Receiver operating characteristic (ROC) line. *PRAUC* is the Area Under the Curve of the Precision-Recall line. *Nb of features* excludes fixed effects.

Table 7: Node level regression for GLMM tree

roa	0.0539*	-0.0252	-0.0146**	0.0016**
	(0.0265)	(0.0127)	(0.0059)	(0.0007)
roe	0.0127	0.0023	0.0037*	0.00001
	(0.0102)	(0.0063)	(0.0019)	(0.0003)
capex 2rev	-0.0138	0.0121	-0.0025	-0.00002
	(0.0139)	(0.0055)	(0.0023)	(0.0002)
mtb	0.0180	-0.0021	0.0009	0.0000
	(0.0168)	(0.0053)	(0.0031)	(0.0002)
ln rev	0.0046	0.0269	-0.0059	-0.0002
	(0.0596)	(0.0189)	(0.0086)	(0.0010)
xs ret	0.0175	-0.0052	0.0022	0.0001
	(0.0172)	(0.0074)	(0.0043)	(0.0003)
sales assets	0.0012	-0.0071	-0.0018	0.0004
	(0.0296)	(0.0085)	(0.0056)	(0.0005)
revtq growth	-0.0169	-0.0140***	-0.0004	-0.0004*
	(0.0120)	(0.0043)	(0.0019)	(0.0002)
eps	-0.0108	-0.0077	0.0061	0.0001
	(0.0253)	(0.0058)	(0.0058)	(0.0002)
ni	-0.0356	0.0036	0.0021	-0.0010***
	(0.0607)	(0.0060)	(0.0248)	(0.0003)
ebt	-0.0773**	-0.0020	0.0195***	-0.0016**
	(0.0302)	(0.0123)	(0.0075)	(0.0007)
log assets	-0.0211	-0.0432**	0.0110	-0.0002
	(0.0706)	(0.0206)	(0.0133)	(0.0011)
seq	-0.0119	0.0031	-0.0279	0.0022***
	(0.0385)	(0.0093)	(0.0279)	(0.0005)
teq	0.0439	0.0018	-0.0084	-0.0010*
	(0.0561)	(0.0082)	(0.0288)	(0.0005)
dvpsx f	0.0053	-0.0028	-0.0015	-0.0004
	(0.0302)	(0.0044)	(0.0103)	(0.0002)
sic2 roa	-0.0083	-0.0053	-0.0036	0.0002
	(0.0189)	(0.0057)	(0.0040)	(0.0003)
shrown pct	0.0561	-0.0079	0.0013	0.0002
	(0.0598)	(0.0088)	(0.0045)	(0.0002)
age	0.0062	0.0208***	0.0060*	-0.0001
	(0.0157)	(0.0040)	(0.0036)	(0.0002)
outside ceo	0.0559***	0.0160***	-0.0004	-0.0002
	(0.0137)	(0.0049)	(0.0028)	(0.0002)
btm	-0.0039	0.0184***	-0.0065	-0.0001
	(0.0160)	(0.0058)	(0.0035)	(0.0003)
competition	-0.0053	-0.0012	0.0024	0.0007*
	(0.0305)	(0.0070)	(0.0072)	(0.0004)
firm value	-0.0004	0.0001	0.0012	-0.00001
	(0.0020)	(0.0060)	(0.0012)	(0.0001)
ceo tenure	-0.5495***	-0.1552***	-0.0002	-0.0001
	(0.2015)	(0.0500)	(0.0042)	(0.0002)
tobins q in t-1	-0.0099	-0.0009	-0.0018	0.0001
	(0.0059)	(0.0040)	(0.0016)	(0.0001)
idiosyn. ret.	0.0255	0.0467***	0.0064	-0.0003
	(0.0309)	(0.0163)	(0.0093)	(0.0005)
mkt adj. ind. ret.	-0.0627	0.0704***	-0.0317*	-0.0027**
	(0.0683)	(0.0217)	(0.0180)	(0.0011)
vol 10y ff48 in t-1	0.0655	0.0502	-0.0382	-0.0025
	(0.2129)	(0.0509)	(0.0501)	(0.0029)
ind.-adj. vol. in t-1	0.0002	0.0176	0.0018	-0.0002
	(0.0281)	(0.0108)	(0.0055)	(0.0006)
Node	1	2	3	4
CEO Tenure > 1 year	NO	NO	YES	YES
EBT > 11th pctile	NO	YES		
EBT > 3rd pctile			NO	YES
Observations	670	3,959	645	16,518
R-squared	22%	8%	23%	1%
Prob(Forced)	0.148%	0.051%	0.007%	0.001%

Note: Regressions include year and industry fixed effects. Standard errors are reported in parenthesis below the regression coefficient. Please refer to the variable description table for variable names. *, ** and *** indicate coefficients which are statistically significant at the 10, 5 and 1 percent level respectively.

Table 8: Confusion Matrix for Peters & Wagner (2004) Replication

		Predicted		
		Routine	**Forced**	**Total**
	Routine	20,673	690	21,363
True Class	**Forced**	382	47	429
	Total	21,055	737	21,792

Note: *Routine* refers to observations where the CEO retains their position or experiences routine turnover. *Forced* is the event in which the CEO is dismissed. Performance is based on complete sample with 20 variable specification and no compensation variables.

Table 9: Confusion Matrix for GLMM Tree (same sample)

		Predicted		
		Routine	**Forced**	**Total**
	Routine	20,939	424	21,363
True Class	**Forced**	225	204	429
	Total	21,164	628	21,792

Note: *Routine* refers to observations where the CEO retains their position or experiences routine turnover. *Forced* is the event in which the CEO is dismissed. Performance is based on best performing GLMM tree specification (depth of 2) and run on the same sample period as Peters & Wagner (2004) with 20 extra explanatory variables.

Table 10: CEO Compensation and Probability of Forced Turnover

Outcome Variable	Ln(total compensation)			
Model	1	2	3	4
Forced turnover probability	18.78***	18.77***	20.01***	0.0806***
	(6.09)	(6.39)	(5.65)	(14.21)
Ln(assets) in t-1	0.39***	0.4320***	0.4624***	0.3737***
	(24.76)	(29.94)	(21.31)	(3123.7)
Tobin's Q in t-1	0.19***	0.1397***	0.1173***	0.1299***
	(10.09)	(11.86)	(10.04)	(456.7)
Idiosyncratic return	0.84***	1.4735***	1.4786***	0.1210***
	(7.56)	(6.85)	(6.02)	(192.4)
Market-adj. industry return	0.79***	0.8353***	0.9149***	0.2955***
	(7.20)	(9.33)	(8.01)	(360.1)
Industry-adj. volatility in t-1	-0.96***	-0.2307***	-0.2338***	0.0588***
	(-3.35)	(-4.15)	(-3.74)	(110.6)
Year fixed effects	YES	YES	YES	YES
Sample	Peters & Wagner	replication	Tree sample	Tree sample
1st step estimation method	OLS	OLS	OLS	GLMM
Observations	24,919	24,453	21,792	21,792
R-squared		32%	34%	34%

Note: *Forced turnover probability* is the probability of forced CEO turnover estimated in a first step (method specified lower in the table). *Tobin's Q* is market value of assets to book value of assets. *Idiosyncratic return*: firm's stock return minus equally weighted industry return. *Market-adj. industry return*: Equally weighted industry return. *Industry-adj. volatility* Stock return volatility (see Peters & Wagner (2014) for details). *Ln(Assets)* is the natural logarithm of the firm's assets taken from Compustat. Standard errors in column 4 were obtained through bootstrapping.*, ** and *** indicate coefficients which are statistically significant at the 10, 5 and 1 percent level respectively. T-statistic reported in parenthesis for this table to match presentation style of Peters & Wagner (2014).

Table 11: CEO Compensation Change and Probability of Forced Turnover

Outcome Variable	ln(1+PPS)		Total Compensation YoYΔ	
Model	1	2	3	4
Forced turnover probability	-15.912***	-1.377***	-46166*	-2469.9***
	(4.3789)	(0.036)	(27755.9)	(160.28)
Ln(assets) in t-1	.46246***	.5115***	77.0518	9.1825
	(.05418)	(0.002)	(66.424)	(9.497)
BTM in t-1	-.10478***	-.1100***	9.3795	25.990***
	(.02328)	(0.001)	(25.830)	(2.049)
CEO age	-.0271***	-.00991***	-22.4585*	-.83661*
	(.0071)	(0.0003)	(13.213)	(1.075)
CEO tenure	.00382	.00391***	6.4642	-1.0595
	(.0067)	(0.0003)	(8.1827)	(1.885)
Firm age	-.0025	-.00118***	.1109	3.8246***
	(.0038)	(0.0002)	(3.0301)	(0.444)
Idiosyncratic risk	-9.670***	-10.136***	1012.36	-21903.1***
	(3.614)	(0.1973)	(9771.7)	(611.5)
Peer risk	32.381***	31.015***	-7866.92	-26922***
	(6.324)	(0.3419)	(17817)	(1552)
12 month return in t	.6086***	.5840***	1729.47***	1718.31***
	(.05407)	(0.0036)	(338.76)	(23.629)
12 month return in t-1	.1177***	.2218***	656.09*	1399.84***
	(.04169)	(0.0031)	(349.69)	(29.021)
CEO Equity Holdings	-.0663***	-.0747***		
	(.0069)	(0.0003)		
Year fixed effects	YES	YES	YES	YES
Industry effects	YES	YES	YES	YES
1st step estimation method	Logit	GLMM	Logit	GLMM
Observations	6,652	6,652	13,874	13,874
R-squared	21%	23%	1%	1%

Note: *Forced turnover probability* is the probability of forced CEO turnover estimated in a first step (method specified lower in the table). *CEO age* and *CEO Tenure* were acquired from Execucomp and are expressed in years. *Firm age* is also in years and based on the earliest date the firm appears in CRSP. *Peer risk* is is the standard deviation stock returns due to industry effects, estimated at the two digit SIC level. *Idiosyncratic risk* is the standard deviation of the idiosyncratic portion of stock returns after removing industry returns. *12 month return* was calculated using CRSP stock price information. *Ln(Assets)* is the natural logarithm of the firm's assets taken from Compustat. *CEO Equity Holdings* is the percentage of outstanding shares owned by the CEO. Standard errors in columns 2 and 4 were obtained through bootstrapping. *, ** and *** indicate coefficients which are statistically significant at the 10, 5 and 1 percent level respectively.

Table 12: Firm Performance after Forced Turnover

Outcome Variable	ROA(t-1,t+3)Δ%		
Model	1	2	3
Forced	0.0766	0.0954	0.3580**
	(0.069)	(0.078)	(0.1609)
CEO Outsider in t-1	-0.0373	-0.0417	0.0434
	(0.052)	(0.052)	(0.1479)
Outsider Board	0.0019	0.0247	-0.0086
	(0.111)	(0.105)	(0.4104)
Industry-adj. ROA in t-1	1.1111***	1.0880***	4.51615***
	(0.246)	(0.293)	(1.0673)
Ln(Assets) in t-1	0.0443***	0.0439**	0.1409***
	(0.014)	(0.015)	(0.0456)
Propensity Score Weighting	none	Logit	GLMM
Observations	669	669	669
R-squared	4.4%	3.0%	5.0%

Note: *Forced* is the binary event of a forced CEO turnover. *CEO Outsider in t-1* is a variable indicating whether the CEO joined the firm less than 1 year before he become its CEO. *Outsider Board* is a binary variable indicating whether at least 60% of the board of directors were outsiders (classified as "Independent" in ISS). *Industry-adj. ROA* is the firm's ROA in the previous fiscal year less the its industry's median ROA. *Ln(Assets)* is the natural logarithm of the firm's assets taken from Compustat. *, ** and *** indicate coefficients which are statistically significant at the 10, 5 and 1 percent level respectively.

9 Appendix

9.1 Additional details on neural networks

Faced with a simple supervised learning problem, such as determining whether an animal is a mouse or a dog, a single neuron trained with a sufficient amount of input data (heights) and outcomes (1 for dog, 0 for mouse) would learn to activate past a certain height level to predict that the weighed animal is indeed a dog. Now let us suppose a slightly more complex problem, such as predicting if an animal is a dog when the input series contains the weights of a variety of mice, dogs and elephants. A single neuron's activation function will not offer anywhere near as much accuracy as two neurons will, one activating when a height is tall enough to discard the possibility the animal is a mouse, and the other activating when the height is not tall enough to be an elephant.

The order in which these two neurons are organized, whether sequentially in different layers (where one has a chance to activate based on the raw datapoint and feeds its output to the second which decides whether to also activate in response), or parallel in a single layer (both decide whether to activate based on the raw data input), may not make as much of a difference in the previous example. However these architecture decisions have proven crucial in improving prediction accuracy in far more demanding problems, such as classifying animals using images instead of heights. A large number of neurons will be required to allow different parts of the neural network to specialize in learning various identifying features of the animal in the image, such as color, ear shape, relative body length, etc. However in this case the layout of the nodes will play a substantial role in the accuracy of the network. For instance, if all nodes were placed sequentially, one per layer, such that they are given an opportunity to activate one after the other, this will limit the amount of complex interactions they can have. This will lead to far inferior predictions relative to an architecture where the same number of nodes were fully-interconnect a dozen per layer, allowing for more complex interactions between the inputs. Thus the architecture of a neural network can greatly impact its accuracy, with particular architectures being well-suited for certain tasks.

Beyond the way in which the nodes are organized into layers and linked, architectures can

include a variety of transformations that have shown useful in encoding information or encouraging interactions between nodes (such as a convolution). With theoretical underpinnings largely unavailable, machine learning researchers have mainly relied on purely empirical findings and heuristics based on what produces the best outcomes for a particular task. For instance, convolutional layers were found to be particularly useful in image recognition, whereas recurrent architectures were found to be best suited for text and speech recognition.[7]

Given the relatively nascent use of neural networks in finance and accounting, established heuristics and best-practices are still being determined. This puts more onus on the researcher to attempt to determine the best suited architecture for their particular task. For the problem of predicting forced CEO turnover, I will test the following network layouts: (i) forward feed (Figure 2), (ii) geometric shrinking (Figure 3), (iii) hourglass-shaped (Figure 4), (iv) inverse hourglass-shaped (Figure 5), and (v) convolutional.

The convolutional neural network follows a feed forward style architecture but makes use of one-dimensional convolutional layers given the nature of the data (instead of the 2D convolution used for images, for instance). They apply a filter through the length of the feature vector in order to identify meaningful behavior in the data. While there is no natural ordering to the features, sufficient layers are added to each architecture such that all features interact with each other. The choice of which neural network architectures to test was based on the ones attempted in Lee, Naughton, Zheng, and Zhou (2020), given their success in improving historical litigation prediction performance in a likewise heavily imbalanced class optimization problem using similar accounting data (see Table A3 in Appendix for more details).

After identifying the best performing architecture, two popular network features were tested to potentially improve performance: dropout and batch normalization. Dropout refers to adding a probability of dropping nodes in a neural network, making it less reliant on particular individual nodes to make predictions and thus alleviating overfitting (Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014)). This is implemented with a random 30% dropout rate for each update of the training phase.

[7] A convolutional layer encodes adjacent data points using a convolution function. Recurrent neural networks work with temporal data and are, for instance, able to produce the next observation in a chain, such as the predictive text to complete a sentence.

Batch size refers to how many observations are fed into the network at one time before the learning algorithm adjusts neuron weights. The various neural networks were run with a batch size of 16 as a baseline. Batch sizes of 8 and 32 were also tested. While data normalization is known to increase the speed of training, batch normalization takes this further by normalizing the activations of layers within the neural network itself. This feature, along with dropout, was applied on the best performing network layouts, but did not improve performance in testing.

9.2 Other implementation details

To reiterate, neuron activations in one layer determine activations in the next through the use of randomly initialized node weightings which are then trained to improve prediction performance of the target forced turnover series. Layers in all models tested were densely connected, meaning that the nodes in a single layer communicated all of their outputs to each and every node in the next layer. The learning algorithm adjusts the node weights in response to each forward and backward pass through the network.[8]

Each neuron in the hidden layers has a rectified linear activation unit (ReLU) commonly used in neural network applications (Jiang (2021)), such that

$$x_i = f\left(\sum_j w_{ij} x_j\right)$$

where f is the ReLU function and x_i is node i's output given the outputs of j nodes that feed into it. The models were implemented using the Tensorflow and Keras Regressor frameworks in Python, and run on a Google Colab cloud processing server. Each model was trained using 500 epochs (iterations through the entire training data set), by which point the error rate was observed to have leveled off and no further significant training performance improvements were being made.

[8] The Adam Optimization Algorithm was used as the learning algorithm in all neural network implementations.

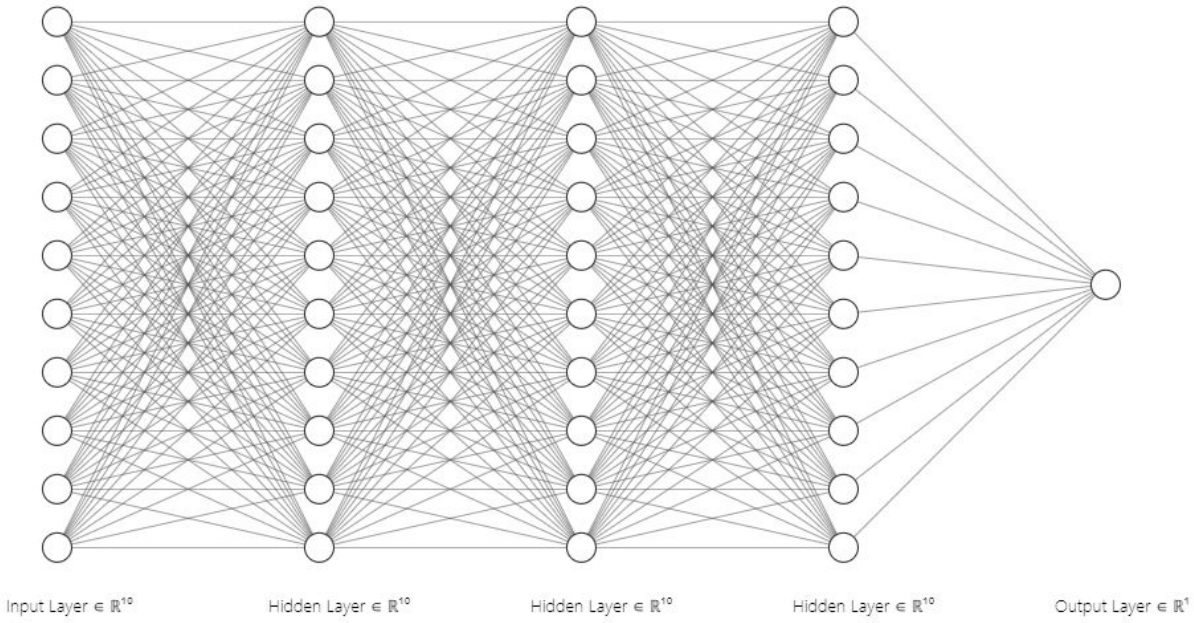

Figure produced using NNSVG.

Figure 7: Feedforward Neural Network Architecture

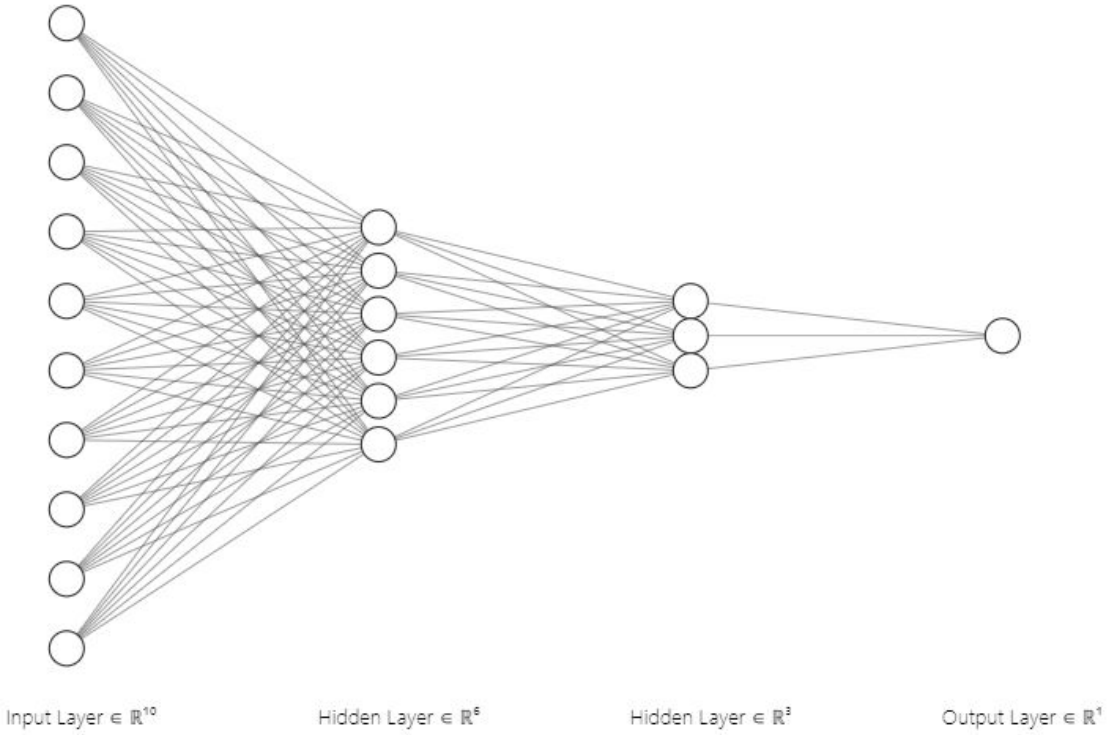

Figure produced using NNSVG.

Figure 8: Geometric Shrinking Neural Network Architecture

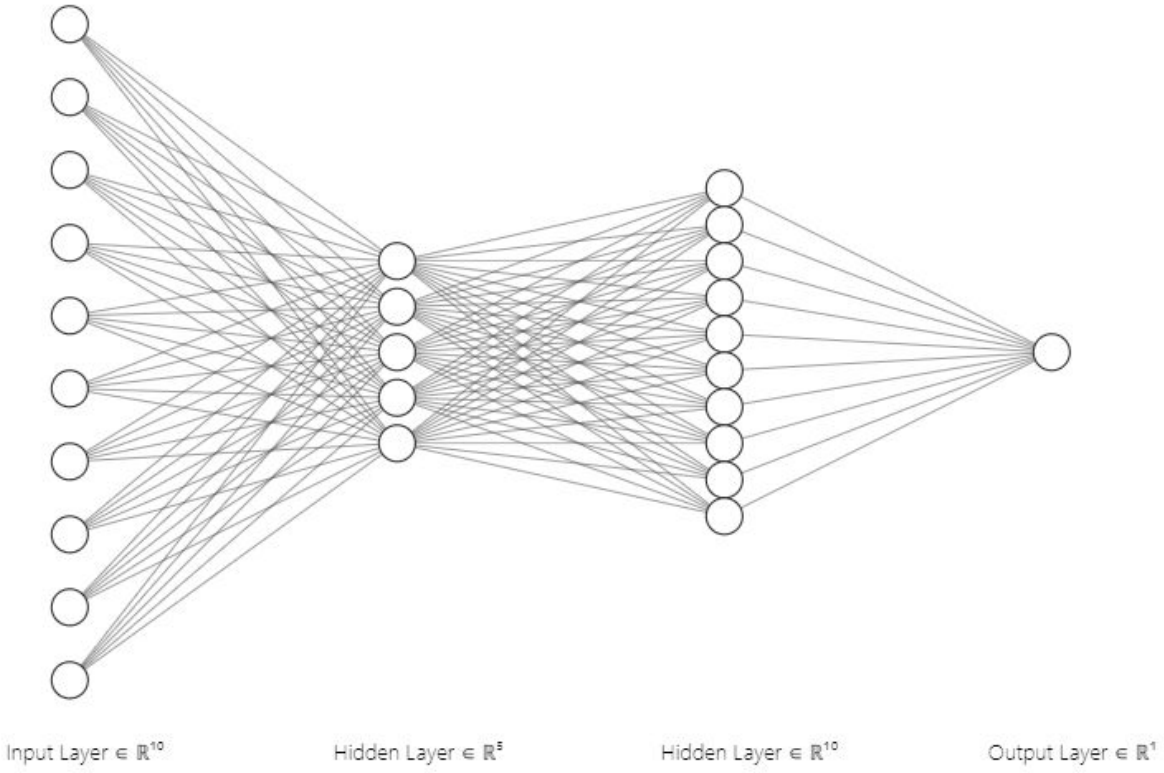

Figure produced using NNSVG.

Figure 9: Hourglass Neural Network Architecture

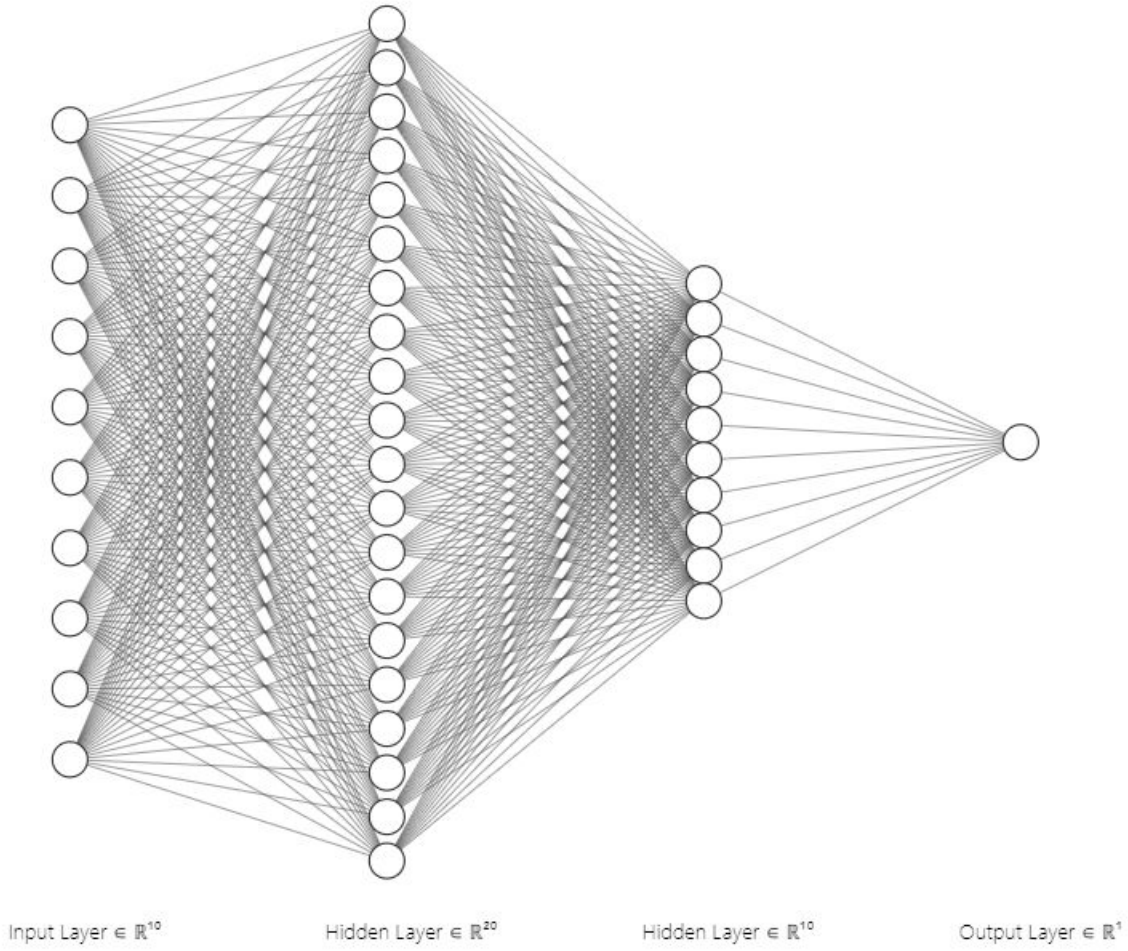

Figure produced using NNSVG.

Figure 10: Inverse Hourglass Neural Network Architecture

Table A1: Model Performances for Sample Without Compensation Variables

Model (1)	Outlier Adjust. (2)	Missing Data (3)	Feature Count (4)	Batch Size (5)	Imbalanced Class Mgmt (6)	Testing Sample Performance Precision (7)	Recall (8)	F1 (9)	AUC (10)	PRAUC (11)
Logit Regression	Winsor.	0 mean	20	n/a	None	0	0	0	49.9%	2.1%
Logit Regression	Winsor.	0 mean	20	n/a	SMOTE	2.5%	51.1%	4.7%	47.1%	2.0%
Feedforward (FF)	Winsor.	0 mean	20	16	None	0	0	0	50.0%	1.8%
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	2.2%	28.3%	4.1%	48.7%	2.1%
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	2.6%	30.6%	4.9%	57.0%	2.8%
Geo. Shrink. (GEO)	Winsor.	0 mean	20	16	SMOTE	1.6%	24.1%	3.0%	48.4%	1.9%
Hourglass (HG)	Winsor.	0 mean	20	16	SMOTE	2.4%	35.3%	4.5%	56.0%	2.7%
Inv.Hourglass (IGH)	Winsor.	0 mean	20	16	SMOTE	2.2%	28.3%	4.1%	48.7%	2.1%
Convolution (CONV)	Winsor.	0 mean	20	16	SMOTE	2.4%	78.6%	4.7%	54.4%	2.4%
FF + Dropout	Winsor.	0 mean	20	16	SMOTE	1.8%	39.9%	3.5%	49.2%	1.9%
FF + BatchNorm	Winsor.	0 mean	20	16	SMOTE	2.1%	36.1%	4.0%	52.0%	2.1%
Feedforward (FF)	raw	0 mean	20	16	SMOTE	2.6%	30.6%	4.9%	57.0%	2.8%
Feedforward (FF)	Winsor.	{0,1} series	20	16	SMOTE	2.3%	30.4%	4.3%	56.2%	1.9%
Feedforward (FF)	raw	0 mean	20	16	SMOTE	1.9%	24.3%	3.6%	47.7%	2.0%
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	2.6%	30.6%	4.9%	57.0%	2.8%
Feedforward (FF)	Trunc.	0 mean	20	16	SMOTE	1.1%	2.2%	1.5%	49.5%	2.2%
Feedforward (FF)	Winsor.	0 mean	16	16	SMOTE	1.9%	27.2%	3.5%	48.2%	1.9%
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	2.6%	30.6%	4.9%	57.0%	2.8%
Feedforward (FF)	Winsor.	0 mean	30	16	SMOTE	2.0%	22.6%	3.7%	49.0%	2.0%
Feedforward (FF)	Winsor.	0 mean	20	8	SMOTE	2.1%	26.6%	3.9%	50.0%	2.4%
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	2.6%	30.6%	4.9%	57.0%	2.8%
Feedforward (FF)	Winsor.	0 mean	20	32	SMOTE	2.9%	35.0%	5.4%	52.8%	2.6%
						Complete Sample Performance				
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	8.9%	90.6%	16.2%	92.2%	18.7%
Logit Regression	Winsor.	0 mean	20	n/a	SMOTE	2.3%	54.4%	4.4%	46.7%	1.8%
Peters & Wagner						4.7%	26.4%	8.0%	61.6%	4.1%
Bushman et al.						3.8%	16.0%	6.1%	68.7%	2.2%

Note: *Accuracy* is the percentage of correctly predicted sample observations. *Precision* is the ratio of true positive to total positives. *Recall* is the ratio of true positive to true positives + false negatives. *F1* score is the harmonic mean of *Precision* and *Recall*. *AUC* is the Area Under the Curve of the Receiver operating characteristic (ROC) line. *PRAUC* is the Area Under the Curve of the Precision-Recall line. *Winsor.* manages outliers by winsorizing the data at 1%. *Trunc.* manages outliers by truncating the data at 1%. *raw* indicates no outlier management. *0 mean* replaces missing data with 0, which is the mean of the standardized features. *{0,1} series* adds a binary series for every feature that is equal to 1 when an observation is missing.

Table A2: Model Performances for Sample With Compensation Variables

Model	Outlier Adjust.	Missing Data	Feature Count	Batch Size	Imbalanced Class Mgmt	**Testing Sample Performance** Precision	Recall	F1	AUC	PRAUC
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Logit Regression	Winsor.	0 mean	20	n/a	None	0	0	0	50.0%	1.9%
Logit Regression	Winsor.	0 mean	20	n/a	SMOTE	2.1%	53.9%	4.0%	51.3%	2.3%
Feedforward (FF)	Winsor.	0 mean	20	16	None	0	0	0	50.0%	1.8%
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	2.3%	26.9%	4.2%	52.5%	2.4%
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	2.3%	26.9%	4.2%	52.5%	2.4%
Geo. Shrink. (GEO)	Winsor.	0 mean	20	16	SMOTE	1.9%	29.4%	3.6%	48.2%	2.1%
Hourglass (HG)	Winsor.	0 mean	20	16	SMOTE	2.1%	33.1%	3.9%	48.6%	2.0%
Inv.Hourglass (IGH)	Winsor.	0 mean	20	16	SMOTE	2.3%	28.8%	4.2%	50.6%	1.9%
Convolution (CONV)	Winsor.	0 mean	20	16	SMOTE	0	0	0	50.0%	1.8%
FF + Dropout	Winsor.	0 mean	20	16	SMOTE	2.0%	32.6%	3.7%	51.4%	1.9%
FF + BatchNorm	Winsor.	0 mean	20	16	SMOTE	2.0%	27.5%	3.7%	50.6%	1.8%
Feedforward (FF)	raw	0 mean	20	16	SMOTE	1.9%	26.6%	3.5%	52.2%	1.9%
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	2.3%	26.9%	4.2%	52.5%	2.4%
Feedforward (FF)	Trunc.	0 mean	20	16	SMOTE	1.6%	26.8%	3.0%	52.1%	1.8%
Feedforward (FF)	Winsor.	0 mean	16	16	SMOTE	2.3%	31.3%	4.3%	51.4%	2.2%
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	2.3%	26.9%	4.2%	52.5%	2.4%
Feedforward (FF)	Winsor.	0 mean	30	16	SMOTE	2.1%	22.0%	3.8%	45.6%	2.1%
Feedforward (FF)	Winsor.	0 mean	20	8	SMOTE	2.6%	25.3%	4.7%	51.7%	2.4%
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	2.3%	26.9%	4.2%	52.5%	2.4%
Feedforward (FF)	Winsor.	0 mean	20	32	SMOTE	1.9%	20.1%	3.5%	50.8%	2.3%
						Complete Sample Performance				
Feedforward (FF)	Winsor.	0 mean	20	16	SMOTE	8.6%	91.5%	15.7%	91.7%	18.9%
Logit Regression	Winsor.	0 mean	20	n/a	SMOTE	2.2%	58.1%	4.2%	46.7%	1.9%

Note: *Accuracy* is the percentage of correctly predicted sample observations. *Precision* is the ratio of true positive to total positives. *Recall* is the ratio of true positive to true positives + false negatives. *F1* score is the harmonic mean of *Precision* and *Recall*. *AUC* is the Area Under the Curve of the Receiver operating characteristic (ROC) line. *PRAUC* is the Area Under the Curve of the Precision-Recall line. *Winsor.* manages outliers by winsorizing the data at 1%. *Trunc.* manages outliers by truncating the data at 1%. *raw* does no outlier management. *0 mean* replaces missing data with 0, which is the mean of the standardized features. {*0,1*} *series* adds a binary series for every feature that is equal to 1 when an observation is missing.

Table A3: Model Performances for Sample With Compensation Variables

Label	Model Descriptions
Feedforward (FF)	Feedforward neural network model with architecture: input-16-16-16-16-16-16-16-ouput
Geo. Shrink. (GEO)	Geometric shrinking shaped neural network with architecture: input-16-8-output
Hourglass (HG)	Hourglass-shaped neural network model with architecture: input-16-ouput
Inv.Hourglass (IHG)	Inverse hourglass-shaped neural network model with architecture: input-32-ouput
Convolution (CONV)	Convolutional neural network with architecture: 30 filters of length 4

References

Abadie, Alberto and Guido Imbens (2002). *Simple and bias-corrected matching estimators for average treatment effects.*

Adams, John C and Sattar A Mansi (2009). "CEO turnover and bondholder wealth". In: *Journal of Banking & Finance* 33.3, pp. 522–533.

Bao, Yang, Bin Ke, Bin Li, Y Julia Yu, and Jie Zhang (2020). "Detecting accounting fraud in publicly traded US firms using a machine learning approach". In: *Journal of Accounting Research* 58.1, pp. 199–235.

Barth, M, K Li, and CG McClure (2019). *Evolution in value relevance of accounting.* Tech. rep. Working paper (SSRN abstract 2933197).

Beneish, Messod D, Cassandra D Marshall, and Jun Yang (2017). "Explaining CEO retention in misreporting firms". In: *Journal of Financial Economics* 123.3, pp. 512–535.

Bennedsen, Morten, Kasper Meisner Nielsen, Francisco Pérez-González, and Daniel Wolfenzon (2007). "Inside the family firm: The role of families in succession decisions and performance". In: *The Quarterly Journal of Economics* 122.2, pp. 647–691.

Bermingham, Mairead L, Ricardo Pong-Wong, Athina Spiliopoulou, Caroline Hayward, Igor Rudan, Harry Campbell, Alan F Wright, James F Wilson, Felix Agakov, Pau Navarro, et al. (2015). "Application of high-dimensional feature selection: evaluation for genomic prediction in man". In: *Scientific reports* 5.1, pp. 1–12.

Bernanke, Ben S, Jean Boivin, and Piotr Eliasz (2005). "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach". In: *The Quarterly journal of economics* 120.1, pp. 387–422.

Bertomeu, Jeremy, Edwige Cheynel, Eric Floyd, and Wenqiang Pan (2021). "Using machine learning to detect misstatements". In: *Review of Accounting Studies* 26.2, pp. 468–519.

Bhagat, Sanjai, Brian J Bolton, and Ajay Subramanian (2010). "CEO education, CEO turnover, and firm performance". In: *Available at SSRN 1670219*.

Bochkay, Khrystyna, Roman Chychyla, and Dhananjay Nanda (2019). "Dynamics of CEO disclosure style". In: *The Accounting Review* 94.4, pp. 103–140.

Burks, Jeffrey J (2010). "Disciplinary measures in response to restatements after Sarbanes–Oxley". In: *Journal of Accounting and Public Policy* 29.3, pp. 195–225.

Bushman, Robert, Zhonglan Dai, and Xue Wang (2010). "Risk and CEO turnover". In: *Journal of Financial Economics* 96.3, pp. 381–398.

Caliński, Tadeusz and Jerzy Harabasz (1974). "A dendrite method for cluster analysis". In: *Communications in Statistics-theory and Methods* 3.1, pp. 1–27.

Campbell, T Colin, Michael Gallmeyer, Shane A Johnson, Jessica Rutherford, and Brooke W Stanley (2011). "CEO optimism and forced turnover". In: *Journal of Financial Economics* 101.3, pp. 695–712.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Chen, Xi, Yang Ha Cho, Yiwei Dou, and Baruch Lev (2021). "Fundamental Analysis of Detailed Financial Data: A Machine Learning Approach". In: *Available at SSRN 3741015*.

Clayton, Matthew C, Jay C Hartzell, and Joshua Rosenberg (2005). "The impact of CEO turnover on equity volatility". In: *The Journal of Business* 78.5, pp. 1779–1808.

Cornelli, Francesca, Zbigniew W Kominek, and Alexander Ljungqvist (2010). "Monitoring managers: Does it matter?" In: *Journal of Finance, Forthcoming, European Corporate Governance Institute (ECGI)-Finance Working Paper* 271.

Coughlan, Anne T and Ronald M Schmidt (1985). "Executive compensation, management turnover, and firm performance: An empirical investigation". In: *Journal of accounting and economics* 7.1-3, pp. 43–66.

Dah, Mustafa A, Melissa B Frye, and Matthew Hurst (2014). "Board changes and CEO turnover: The unanticipated effects of the Sarbanes–Oxley Act". In: *Journal of Banking & Finance* 41, pp. 97–108.

Davis, Jesse and Mark Goadrich (2006). "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240.

Defond, Mark L and Mingyi Hung (2004). "Investor protection and corporate governance: Evidence from worldwide CEO turnover". In: *Journal of Accounting Research* 42.2, pp. 269–312.

DeFond, Mark L and Chul W Park (1999). "The effect of competition on CEO turnover". In: *Journal of Accounting and Economics* 27.1, pp. 35–56.

Desai, Hemang, Chris E Hogan, and Michael S Wilkins (2006). "The reputational penalty for aggressive accounting: Earnings restatements and management turnover". In: *The Accounting Review* 81.1, pp. 83–112.

Dikolli, Shane S, William J Mayew, and Dhananjay Nanda (2014). "CEO tenure and the performance-turnover relation". In: *Review of accounting studies* 19.1, pp. 281–327.

Duda, Richard O, Peter E Hart, and David G Stork (1973). *Pattern classification and scene analysis*. Vol. 3. Wiley New York.

Eisfeldt, Andrea L and Camelia M Kuhnen (2013). "CEO turnover in a competitive assignment framework". In: *Journal of Financial Economics* 109.2, pp. 351–372.

Engel, Ellen, Rachel M Hayes, and Xue Wang (2003). "CEO turnover and properties of accounting information". In: *Journal of Accounting and Economics* 36.1-3, pp. 197–226.

Farrell, Kathleen A and David A Whidbee (2002). "Monitoring by the financial press and forced CEO turnover". In: *Journal of Banking & Finance* 26.12, pp. 2249–2276.

– (2003). "Impact of firm performance expectations on CEO turnover and replacement decisions". In: *Journal of accounting and economics* 36.1-3, pp. 165–196.

Fiordelisi, Franco and Ornella Ricci (2014). "Corporate culture and CEO turnover". In: *Journal of Corporate Finance* 28, pp. 66–82.

Fisman, Raymond J, Rakesh Khurana, Matthew Rhodes-Kropf, and Soojin Yim (2014). "Governance and CEO turnover: Do something or do the right thing?" In: *Management Science* 60.2, pp. 319–337.

Fokkema, Marjolein, Niels Smits, Achim Zeileis, Torsten Hothorn, and Henk Kelderman (2018). "Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees". In: *Behavior research methods* 50.5, pp. 2016–2034.

Gao, Huasheng, Jarrad Harford, and Kai Li (2009). "Incentive effects of extreme CEO pay cuts". In: *Unpublished working paper, University of British Columbia*.

Goyal, Vidhan K and Chul W Park (2002). "Board leadership structure and CEO turnover". In: *Journal of Corporate finance* 8.1, pp. 49–66.

Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre (2017). "Correlation and variable importance in random forests". In: *Statistics and Computing* 27.3, pp. 659–678.

Gu, Shihao, Bryan T Kelly, and Dacheng Xiu (2019). "Empirical asset pricing via machine learning". In: *Chicago Booth Research Paper* 18-04, pp. 2018–09.

Guo, Lixiong and Ronald W Masulis (2015). "Board structure and monitoring: New evidence from CEO turnovers". In: *The Review of Financial Studies* 28.10, pp. 2770–2811.

Harvey, Campbell R, Yan Liu, and Heqing Zhu (2016). "âŠ and the cross-section of expected returns". In: *The Review of Financial Studies* 29.1, pp. 5–68.

Hazarika, Sonali, Jonathan M Karpoff, and Rajarishi Nahata (2012). "Internal corporate governance, CEO turnover, and earnings management". In: *Journal of Financial Economics* 104.1, pp. 44–69.

Helwege, Jean, Vincent J Intintoli, and Andrew Zhang (2012). "Voting with their feet or activism? Institutional investorsâ impact on CEO turnover". In: *Journal of Corporate Finance* 18.1, pp. 22–37.

Hennes, Karen M, Andrew J Leone, and Brian P Miller (2008). "The importance of distinguishing errors from irregularities in restatement research: The case of restatements and CEO/CFO turnover". In: *The Accounting Review* 83.6, pp. 1487–1519.

Ho, Tin Kam (1995). "Random decision forests". In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE, pp. 278–282.

Hoerl, Arthur E and Robert W Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1, pp. 55–67.

Huson, Mark R, Paul H Malatesta, and Robert Parrino (2004). "Managerial succession and firm performance". In: *Journal of Financial Economics* 74.2, pp. 237–275.

Huson, Mark R, Robert Parrino, and Laura T Starks (2001). "Internal monitoring mechanisms and CEO turnover: A long-term perspective". In: *the Journal of Finance* 56.6, pp. 2265–2297.

Inderst, Roman and Holger M Mueller (2010). "CEO replacement under private information". In: *The Review of Financial Studies* 23.8, pp. 2935–2969.

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR, pp. 448–456.

Jenter, Dirk and Fadi Kanaan (2015). "CEO turnover and relative performance evaluation". In: *the Journal of Finance* 70.5, pp. 2155–2184.

Jenter, Dirk and Katharina Lewellen (2021). "Performance-induced CEO turnover". In: *The Review of Financial Studies* 34.2, pp. 569–617.

Jiang, Weiwei (2021). "Applications of deep learning in stock market prediction: recent progress". In: *Expert Systems with Applications* 184, p. 115537.

Kaplan, Steven N and Bernadette A Minton (2012). "How has CEO turnover changed?" In: *International review of Finance* 12.1, pp. 57–87.

Khurana, Rakesh and Nitin Nohria (2000). "The performance consequences of CEO turnover". In: *Available at SSRN 219129*.

Krupa, Jake and Miguel Minutti-Meza (2021). "Regression and Machine Learning Methods to Predict Discrete Outcomes in Accounting Research". In: *Available at SSRN 3801353*.

Langfelder, Peter and Steve Horvath (2008). "WGCNA: an R package for weighted correlation network analysis". In: *BMC bioinformatics* 9.1, pp. 1–13.

Laux, Volker (2008). "Board independence and CEO turnover". In: *Journal of Accounting Research* 46.1, pp. 137–171.

– (2012). "Stock option vesting conditions, CEO turnover, and myopic investment". In: *Journal of Financial Economics* 106.3, pp. 513–526.

Lee, Gene Moo, James P Naughton, Xin Zheng, and Dexin Zhou (2020). "Predicting litigation risk via machine learning". In: *Available at SSRN 3740954*.

Lee, Sam, Steven R Matsunaga, and Chul W Park (2012). "Management forecast accuracy and CEO turnover". In: *The Accounting Review* 87.6, pp. 2095–2122.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos (2018). "The M4 Competition: Results, findings, conclusion and way forward". In: *International Journal of Forecasting* 34.4, pp. 802–808.

Meng, Xiaojing (2019). "Information, incentives and CEO replacement". In: *NYU Stern School of Business*.

Mobbs, Shawn (2013). "CEOs under fire: The effects of competition from inside directors on forced CEO turnover and CEO compensation". In: *Journal of Financial and Quantitative Analysis* 48.3, pp. 669–698.

Murphy, Kevin J and Jerold L Zimmerman (1993). "Financial performance surrounding CEO turnover". In: *Journal of Accounting and Economics* 16.1-3, pp. 273–315.

Niehaus, Greg and Greg Roth (1999). "Insider trading, equity issues, and CEO turnover in firms subject to securities class action". In: *Financial Management*, pp. 52–72.

Oyer, Paul (2008). "The making of an investment banker: Stock market shocks, career choice, and lifetime income". In: *The Journal of Finance* 63.6, pp. 2601–2628.

Parrino, Robert (1997). "CEO turnover and outside succession a cross-sectional analysis". In: *Journal of financial Economics* 46.2, pp. 165–197.

Peters, Florian S and Alexander F Wagner (2014). "The executive turnover risk premium". In: *The Journal of Finance* 69.4, pp. 1529–1563.

Puffer, Sheila M and Joseph B Weintrop (1991). "Corporate performance and CEO turnover: The role of performance expectations". In: *Administrative Science Quarterly*, pp. 1–19.

Ridgeway, Greg (2007). "Generalized Boosted Models: A guide to the gbm package". In: *Update* 1.1, p. 2007.

Sliwka, Dirk (2007). "Managerial turnover and strategic change". In: *Management Science* 53.11, pp. 1675–1687.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1, pp. 1929–1958.

Taylor, Lucian et al. (2010). *CEO pay and CEO power: Evidence from a dynamic learning model.* Citeseer.

Warner, Jerold B, Ross L Watts, and Karen H Wruck (1988). "Stock prices and top management changes". In: *Journal of financial Economics* 20, pp. 461–492.

Weisbach, Michael S (1988). "Outside directors and CEO turnover". In: *Journal of financial Economics* 20, pp. 431–460.

Zeileis, Achim, Torsten Hothorn, and Kurt Hornik (2008). "Model-based recursive partitioning". In: *Journal of Computational and Graphical Statistics* 17.2, pp. 492–514.

Zou, Hui (2006). "The adaptive lasso and its oracle properties". In: *Journal of the American statistical association* 101.476, pp. 1418–1429.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.