MAN VS. MACHINE:

COMPARING MACHINE LEARNING AND ANALYSTS' PREDICTIONS FOR EARNINGS

By

Jessica Nguyen
njessica@wharton.upenn.edu

An Undergraduate Thesis submitted in partial fulfillment of the requirements for the

WHARTON RESEARCH SCHOLARS

Faculty Advisor:

Daniel Taylor
dtayl@wharton.upenn.edu

Associate Professor, Accounting

THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA

MAY 2020

Acknowledgements: The author would like to thank Dr. Daniel Taylor for his extensive guidance on navigating the accounting literature and advice on conducting academic research. The author is grateful to Bradford Lynch for his guidance on machine learning methods and technical implementation. Finally, the author expresses gratitude to Dr. Catherine Schrand for her help on designing a thesis project and her advice on communicating research.

ABSTRACT:

The main goal of this study is to determine whether machine learning can outperform analysts in forecasting earnings. Using gradient boosted regression trees (a recursive regression tree-building method), this paper concludes that machine learning is unable to beat analysts' predictions for earnings, when comparing median absolute percentage error. The model was trained on firms with Wall Street analyst coverage for earnings between years 2013 to 2016. Predictors from existing earnings forecasting literature were input for the model's consideration. The model's performance was compared to analysts' forecasts on out-of-sample earnings for years 2017 to 2019. The results suggest that analysts hold some incremental information that is useful for forecasting earnings. This incremental information is either not contained in financial statements or has not been researched in existing literature.

1. INTRODUCTION

The existing literature on earnings forecasts has used two approaches: time series modeling and cross-sectional forecasts. Both approaches require users to specify and fit a model, a priori. This paper offers a different approach from existing literature – machine learning.

For the purpose of this research, a gradient boosted regression tree (GBRT) is trained on historical public data to determine whether machine learning can outperform analysts or whether analysts offer additional useful information that is not contained in financial statements.

A GBRT is chosen because of its ubiquitous use in industry for a variety of applications. GBRTs forecast by recursively building a series of regression trees that build off the residuals of previous trees. In contrast to other machine learning methods, GBRTs cannot consider all possible relationships between all predictors; the user must specify features to input into the model for consideration. Variables found in existing literature that were predictive of earnings are input into the model. The model is trained on firms found in Compustat that are covered by Wall Street analysts. Analysts' forecasts are found in the IBES summary dataset. Due to machine limitations, the training data is limited to earnings from years 2013 to 2016. These years were arbitrarily chosen by the RAM limits on a 256 GiB machine.

It is hypothesized that machine learning will not outperform analysts in forecasting earnings because analysts have opportunities to learn different information from firms that machines cannot learn from a financial statement. For example, analysts may talk to people within firms – something a machine cannot do. Additionally, the GBRT model represents the best predictors that exist in the literature. It is unlikely that the existing literature has extracted as much information for predicting earnings as analysts have.

In out-of-sample forecasts (years 2017 to 2019), this research found that the GBRT model does not outperform analysts, as determined by median absolute percentage error (MdAPE), in predicting earnings. This confirms the initial hypothesis and suggests several important implications: (1) analysts still offer incremental-value to forecasting earnings beyond information that is available in historical financial statements, and (2) as machine learning becomes more widely adopted by industry, stock prices will more efficiently reflect financial statement information.

The rest of the paper is organized as follows. Section 2 will offer a literature review of earnings forecasts and machine learning methods used with financial statement data. Section 3 will discuss the theory and implementation of GBRT and provide a brief discussion of the data. Section 4 will present results and offer discussion. Section 5 will highlight the limitations of the analysis. Finally, Session 6 will provide future areas of research to consider.

2. LITERATURE REVIEW

The literature relevant to the analysis can be categorized into three areas: time-series models for predicting earnings, financial statement models for predicting earnings, and (3) machine learning models.

2.1 Time-Series Models

The literature for predicting earnings spans decades. Early research of methodologies for predicting earnings consist of autoregressive integrated moving average (ARIMA) models combined with the Box-Jenkins (B-J) method to predict quarterly earnings (Foster 1977). After these models were established, papers such as Brown and Rozeff (1979) sought to optimize the various parameters of the B-J model and recommend them for benchmarking analysts' forecasts. However, these B-J time series models have strict assumptions (survivorship and age requirements). Practically speaking, this limits the sample size to firms with sufficient historical data. Additionally, these time series models have shown to be less accurate than analysts' forecasts (Brown, Hagerman, Griffin, and Zmijewski [1987]).

One potential explanation as to why B-J models cannot beat analysts is because analysts are able to incorporate information more frequently into their forecasts. One solution was proposed in Ball and Ghysels (2017), which employed mixed data sampling (MIDAS) regression methods to predict earnings. This method allows models to use time series data sampled at different frequencies. Ball and Ghysels (2017) built their model and compared it to analysts' forecasts. They found that for smaller sized firm and higher forecasts dispersions, their model outperformed analysts. Overall, when they combined their model with analysts' forecasts, they were able to outperform analysts alone. However, these alternatives modeling approaches still do not employ machine learning.

2.2 Financial Statement Models

A large body of literature studies the ability of fundamental analysis to predict performance. Lev and Thiagarajan (1993) identified twelve fundamental signals that analysts *claimed* to use and determined whether these variables were useful for predicting persistent earnings (measured by ERC and future earnings growth). The signals were: (1) accounts receivable, (2) inventory, (3) Capital Expenditure, (4) R&D, (5) Gross Margin, (6) S&A, (7) Provision for Doubtful Receivables, (8) Effective Tax, (9) Order Backlog, (10) Labor Force, (11) LIFO Earnings, and (12) Audit Qualification. Among their findings, the authors found that fundamentals were associated with these two measures. Their analysis also revealed that an interaction effect exists between fundamentals and macroeconomic conditions when predicting earnings. On their own, several variables were weakly relevant; however, when conditioned under macroeconomic variables (e.g. accounts receivables during high inflation), they were strongly correlated with returns.

Abarbanell and Bushee (1997) responded to Lev and Thiagarjan (1993) by questioning the extent to which analysts actually use the signals that they claim. To accomplish this, they determined whether analysts effectively use information from fundamental signals. This paper concluded that while analysts' forecasts revisions were aligned with many fundamentals, the revisions did not incorporate *all* the information available from fundamentals. Therefore, this paper found that in general, analysts underreact to accounting information.

To solve for the shortcomings of analysts' forecasts, recent research uses cross-sectional regression models of financial statement data to forecast earnings. The most popular such model was built by Hou, Van Dijk, and Zhang (HVZ) (2012). This model estimated pooled regression coefficients (using ten years of lagged data). The cross-sectional model regressed *total assets*, *dividends*, *current period's earnings*, an *indicator variable of loss*, and *working capital accruals* on future earnings (1 to 5 years horizon). This model is significant because its cross-sectional approach allows researchers to bypass the strict requirements of time series models. Numerous papers critique and extend the HVZ model.

One such paper is Li and Mohanram (2014, LM). LM attempted to build a model that could beat HVZ. They used a different approach, a Residual Income (RI) model, to predict future EPS. This model emphasized book value and total accruals. The RI model was 28-38% more accurate than the HVZ model. Another such paper is So (2013). So (2013) showed that the model in HVZ could be extended to predicting analysts EPS forecast error. So (2013) concluded that analysts are slow to incorporate historical financial statement information, and that investors overweight analysts' forecasts and consequently ignore considerable amounts of information imbedded in financial statements.

Gerakos and Gramacy (2013, GG) evaluated various methodological choices in these papers. GG found that the best performing model (defined as the one with the least mean-squared predictive error) hinged critically on whether the researcher scaled the variables, winsorized the variables, and the forecast horizon. In general, they found that parsimonious time-series models (random walk and AR(1)) are more robust and generally performed better than cross-sectional regressions.

2.3 Machine Learning with Financial Statement Data

This paper builds upon recent literature that uses machine learning (ML) to predict financial statement fraud. Perols (2011) compares various machine learning to logistic regression to predict fraud. The various machine-learning methods studied include neural networks and support vector machines (SVMs). Surprisingly, Perols (2011) found that logistic regression and SVMs perform the best. Similarly, Bertomeu, Cheynel, Floyd, and Pan (2019) extend Perols (2011) by comparing logistic regression and gradient-boosted regression trees. They find gradient-boosted regression trees provide considerably more accurate fraud predictions than logistic regression. The research in this paper extends those in the literature by applying similar machine-learnings techniques to the prediction of earnings.

The most recent research uses machine learning to determine which fundamentals influence performance. Binz (2019) applies a neural network to Nissim and Penman (2001)'s equity valuation framework. Binz compares the ability of the neural network to predict fundamental values, with the ability of the HVZ earnings forecasts to predict fundamental values. Anand, Brunner, Ikegwu, and Sougiannis (2019) use yet another machine learning tool, random forests, to predict profitability. They find their model is significantly more accurate than a random walk. Neither of these studies compared their models to analysts' forecasts.

This paper builds upon but is different from the current literature in several ways. First, this research employs newer ML methods – gradient-boosted regression trees. These methods are widely used in industry. Second, this paper offers a comparison between the performances of analysts' forecasts ('human forecasts') and machine ('AI forecasts').

This design and comparison to analysts enables several novel insights into the maximum predictive value of financial statements for future earnings and the corresponding value of analyst forecasts. Can we produce forecasts at least as accurate as analysts using only historical financial statement data? Are human analysts still-value added? Can their forecasts provide informational-value beyond that which a machine can extract from historical public data alone? If machine learning becomes widely adopted by industry, will that lead to stock prices more efficiently reflecting fundamental or less reflecting fundamentals?

3. DATA AND METHODS

The primary goal of this study is to explore whether machines can outperform humans in forecasting earnings. As such, the main response variable is the realization of the earnings number being forecasted by analysts. This statistic is commonly referred to as "street earnings" as it includes adjustments such as excluding special items. The actual earnings number and consensus estimates will come from the IBES summary dataset which provides observations from 1976 to 2019. The machines will be trained on the corpus of historical financial statement data available on Compustat.

3.1 Predictor Variables

For a complete list of predictor variables, see table 1. Each predictor variable from existing literature was included as well as their value scaled by total assets. For variables with ratios, both their numerators and denominators were included. For example, for *Current Ratio,* both *Current Assets* (the numerator) and *Current Liabilities* (the denominator) were included on their own in addition to the ratio. Finally, for variables representing a percent change in some value, the lagged raw value was included. For example, for *Percent Change in Gross Margin,* both the current period's gross margin and lagged gross margin were included. All these transformations for predictors were included to be extensive and provide the algorithm with a wide selection to determine which features were most important. Since this research is focused on forecasting and machine learning, multicollinearity or other issues relating to causal interpretation are not of importance.

Predictors in the literature with too many missing values were excluded from the model. These variables were excluded because too much sparsity (and not enough variation among a variable) within the dataset would not add incremental value to the model. This analysis opts for parsimony to save on memory limitations of the machine. In total, after all variable transformations, there were 268 predictors for the algorithm's consideration.

Since this was time series data, in order to prevent future information from being predictors of past earnings, all 268 predictors from the past were lagged to the current time. This meant that to predict earnings for firms in 2016, all information from before 2016 (but not after 2016) were included in the model.

The model was trained on all firms that had both Compustat information as well as analysts' predictions in the IBES summary dataset between years 2013 and 2016. This totaled 33,925 observations. For the out-of-sample data, there were 6,536 observations.

3.2 Gradient Boosted Regression Trees

Gradient Boosted Regression Trees (GBRT) are an extension of regression trees. Each "tree" represents a partition of the sample space into non-overlapping regions based on predictor variables (or nodes).[1] Nodes are built by minimizing the residual sum of squares which equals

$$\sum_{j=1}^{J} \sum_{i \epsilon R_j} (y_i - \hat{y}_i)^2$$

where J is the number of nodes, and nodes are $R_1, \ldots, R_j$. For each node, the prediction is the average of the all response values for training observations in that node.

GBRT extends regression trees by *recursively* building one tree after another. Each subsequent tree that is built by GBRT uses information from previous trees. The first tree will be fit according to the training data. The second tree will then fit to the residuals of the first tree. The third tree will then fit to the residuals of the second tree, and so on.

There are a variety of tuning parameters for GBRTs: 1) nodes per tree, 2) number of trees, 3) shrinkage rate (λ), 4) minimum number of observations within a leaf, 5) fraction of observations used to build a tree, etc. However, for this analysis, a model will be initially built on a default set of 4 parameters (rules of thumb):[2]

- $\lambda = 0.01$

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. "An Introduction to Statistical Learning with Applications in R" (2017), pg. 312

[2] A guide to building generalized boosted models by Greg Ridgeway (although XGBoost is a different package from GBM, many of the model building techniques are applicable) : https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf

- Number of Trees = 500 (will be tuned by cross-validation)
- Nodes per tree (also known as depth of tree) = 5
- Min. Child Weight (minimum number of instances required in a child node) = 5

The optimal number of trees is usually selected first by performing cross-validation (usually with three folds) to minimize the in-sample Mean Absolute Error (MAE). After the number of trees is chosen, other optimal parameter values will be chosen by sweeping over a grid of potential parameter values (see Table 2) and choosing the combination of values that minimizes in-sample MAE. While this is not an exhaustive search over every possible combination of parameters (because the tuning design table only has discrete values for parameters), due to current computational limitations, this is common practice for tuning GBRTs. To summarize, our GBRT model is represented by:

$$\hat{\boldsymbol{f}}(\boldsymbol{x}) = \sum_{\boldsymbol{b}=\mathbf{1}}^{\boldsymbol{B}} \lambda \hat{f}^{b}(x)$$

λ is the shrinkage rate and will determine how much each subsequent tree learns from the previous tree. The shrinkage rate is used to prevent overfitting; therefore, new trees that are added will generally be smaller. B represents the number of trees, and $\hat{f}^{b}$ represents the collection of trees. Each subsequent tree will update the residuals (r_i):

$$r_i - \lambda \hat{f}^{b}(x_i) \rightarrow r_i$$

A small version of each subsequent tree will be added to the collection of trees:

$$\hat{\boldsymbol{f}}(\boldsymbol{x}) + \lambda \hat{f}^{b}(x) \rightarrow \hat{f}^{b}(\boldsymbol{x})$$

One potential disadvantage of using GBRT, at least relative to neural networks, is that GBRT method will not consider non-linear relationships (ratios and interaction effects) automatically. It will only consider what the user inputs. Therefore, there is a need to select variables from the existing literature and not every single variable from financial statements.

3.2 Technical Implementation

For implementation purposes, the GBRT model will be built using the XGBoost package for R.[3] This package will automatically use parallelization to take advantage of 32 cores, deal with sparse matrices (data sets with lots of missing values) and impose regularization. XGBoost handles missing values internally. Any missing values are inferred from any trends in the dataset (grouped for a given firm). This allows us to still make some use of predictors with missing values. Variables with many missing values are still omitted to retain some accuracy in predictions.

A major limitation in using R is its handling of data frames. To transform variables, R would store copies of data frames multiple times – exhausting memory. For example, to transform a variable, R makes a copy of the data frame in a new location, modifies the copy, and then refers to the new copy each time the old copy is called.[4] This inefficient use of memory limited the ability to consider the full range of data (years 1980 to 2019).

4. RESULTS AND DISCUSSION

The optimal tuning parameters for the model were 500 trees, a tree depth of 5, a minimum child weight of 5, and shrinkage of 0.2.

4.1 Comparison to Analysts

For the out-of-sample data, analysts had a mean absolute percentage error (MAPE) of 5.31%. In contrast, the GBRT model had a 1.92% MAPE. While this could suggest that the GBRT model is superior to analysts, we should consider the median absolute percentage error

[3] See the documentation for XGBoost: https://cran.r-project.org/web/packages/xgboost/xgboost.pdf
[4] See Hadley Wickam's explanation on Memory in R: http://adv-r.had.co.nz/memory.html#memory

(MdAPE) to be a better indicator of accuracy because it disregards outliers that could be skewing the MAPE. The MdAPE for analysts was 1.80% and 4.48% for the model. Therefore, from this metric, the model does not outperform analysts. It is interesting to note that the analysts seem to be inferior with outliers but are superior when these outliers are disregarded. It is unclear whether this says something about analysts' ability to predict surprises (whether they are unable to forecast that outliers could exist or whether they prefer not to make such risky predictions) or whether this result says something about the model's regularization methods. Despite tuning and having shrinkage parameters, it is still possible that the model is overfitting and getting into the nook and crannies of all the outliers. Further research would need to be conducted to determine why this result exists.

However, it is interesting to note that the difference in MdAPE between analysts and the machine was less than 3%. While the analysts do outperform the model, it is not by much, relatively. This is a surprising result as this model only incorporates in the best predictors from the current literature. Given that the current literature still has much left to explore, it is surprising that the model would come so close to analysts' forecasts. However, it is unclear whether this difference is significant and what the confidence intervals surrounding the MdAPE are. Further research should investigate whether this result can be replicated on other time periods of data. The 3% difference could be attributable to specific characteristics of this subset of the data. However, overall, this implies that while analysts are inefficient, they are still able to offer value-added over historical public data. However, if a GBRT could come so close to predicting earnings, it might be worthwhile to build a "cyborg" model that combines both analysts' forecasts and machine learning. This cyborg model could overcome the problems associated with outlier values for analysts in addition to offering improvement over the machine's forecasts.

4.2 Decomposing Variations in APE

Since this paper is only interested in predictions, learning what variables the model considers to be important is not of primary interest. However, learning why the model more accurately predicts for some firms over others could be useful. Knowing this information could allow for a cyborg model to determine what weights to put on analysts' forecasts versus machine forecasts for certain types of firms. From the model's feature importance (Figure 1), accruals are the most important feature. Since accruals heavily dominates all other feature, the relationship between it and APE are examined (Figure 2). There are no obvious relationships because the spread of accruals for firms is quite small. Future research should look more into this relationship as well as relationships with other features.

4.3 Comparison to Hou, van Dijk, and Zhang (2012)

To offer further insight into the model's performance, the HVZ model is replicated on the out-of-sample data. Recall the HVZ model is a pooled cross-sectional regression built on ten years of data (Hou, van Dijk, and Zhang 2012, 507):

$$E_{i,t+\tau} = \alpha_0 + \alpha_1 A_{i,t} + \alpha_2 D_{i,t} + \alpha_3 DD_{i,t} + \alpha_4 E_{i,t} + \alpha_5 NegE_{i,t} + \alpha_6 AC_{i,t} + \varepsilon_{i,t+\tau}$$

HVZ defined the following variables:

- Response variable (E): Future Profitability, income before extraordinary items (NOT scaled by total assets). This is not the "street" earnings predicted for by the GBRT model.
- Accruals (AC): Post-1998, defined by cash flow statement method, the difference between earnings and cash flows from operation
- Total Assets (A)

- Dividend Payment (D)
- Dummy variable for Dividend Payers (DD): equals 1 for dividend payers, 0 otherwise
- Dummy variable for Negative Earnings (NegE): equals 1 for negative earnings, 0 otherwise
- Current period's earnings (E)

Since HVZ is not built to forecasts pro forma earnings, while the GBRT model and analysts' forecasts are, there must be caution for comparisons between the HVZ and the GBRT. The HVZ was replicated on the out-of-sample data to predict Compustat (GAAP) earnings. On this dataset, it had a MdAPE of 29.5%. While comparisons cannot directly be made, the HVZ's performance is worse than the GBRT and analysts' forecasts for pro forma earnings. This result indicates that different models may perform differently based on definitions of earnings. The differences between the two models could also be driven by the differences between how GAAP and pro forma earnings are defined. However, based on the large differences in MdAPE, it is still plausible that the GBRT model could outperform the HVZ model on predicting GAAP earnings. Further research would have to be conducted to reach this conclusion.

5. LIMITATIONS

Feature importance can also yield insight into the model's robustness. This model suggests that nearly all the predictions can be made by differences in firms' accruals. While accruals have been shown to be good predictors of earnings in the literature (HVZ 2012, Gerakos and Gramacy 2013), it does offer some concern. Even slight differences in accruals could drastically change predictions. This indicates a lack of a model's robustness because it could easily change given a different dataset. A possible reason for why the model places too much emphasis on accruals may be the sparsity of the data. For many of the predictors, there are many observations with missing values. A high number of missing values may leave many variables to be too sparse and have too little variation. This could lead the model to rely on a variable (like accruals) that has significant variation among observations. The next most important features are pretax income scaled by total assets and then amortizations.

With bigger RAM capacity or more memory-efficient coding languages, a model should be built on a wider range of data (years 1980 to 2019). This will allow us to better analyze the robustness of our model. If our model, built on years 2013 to 2016 are truly robust, we should find similar results when we build our model on the entire dataset.

Another limitation in this research is that it does not consider whether this model could perform well for firms without analysts' coverage. One practical reason for developing a machine learning model would be to forecast earnings for companies without analysts' coverage. To test this, researchers would need to test this model on such companies and compare how the model performs relative to actual earnings.

6. CONCLUSION AND FURTHER AREAS OF RESEARCH

This paper built a machine learning GBRT model to compete against analysts' forecasts for earnings. The model was trained on public historical financial statements data. Variables found to be predictive of earnings in the literature were used as inputs. While machines could beat analysts for earnings that are outliers, overall, the analysts still outperform machine learning. This indicates that analysts are still value-added beyond financial statement information. However, a combination of machine learning and analysts may perform better overall (to capture accuracy for both outliers and non-outliers).

Further extensions of this research should explore whether a purely "machine" model (as opposed to a model that requires user input of predictors) could outperform analysts. For example, a convolutional neural network that could consider deep and non-linear relationships between predictors could be used. This model would extract the maximum amount of information from financial statements – rather than just considering predictors that already exist in the literature. Another model to consider would be a hybrid combination that could combine and average both the GBRT and the convolution neural network. This model would offer additional insight into which types of machine learning work best for earnings forecasts. It would be interesting to understand why such algorithms work better than others.

Other possible avenues of exploration could look at which industries and what characteristics (firms with higher accruals or higher depreciation) machines perform better than analysts and vice versa. It would be insightful to understand not only which industries analysts are better at but also possible reasons why.

Table 1: Predictor Variables

Variable	Compustat Formula	Literature
c	DVC	HVZ
Common Dividend scaled by total assets	DVC / AT	
Dividend Payers Indicator	Dummy variable: 1 - dividend payers, 0 - o/w (DVP)	HVZ
Dividend Payers	DVP	
Total Assets	AT	HVZ, Gerakos and Gramacy
Negative Earnings	Dummy Variable: 1 - negative earnings, 0 - o/w; earnings = income before extraordinary items (IB in COMPUSTAT)	HVZ, EP, RI (Li and Mohanram)
Lagged Negative Earnings	Dummy Variable: 1 - negative earnings, 0 - o/w	So
Accruals	Δ(ACT-CHE)-Δ(LCT-DLC-TXP)-DP	HVZ, Gerakos and Gramacy
Current Assets - Total	ACT	Part of Accruals (HVZ, Gerakos and Gramacy)
Current Assets - Total scaled by total assets	ACT / AT	
Lagged Current Assets - Total	ACT at t-1	
Lagged Current Assets - Total scaled by total assets	(ACT / AT) at t-1	
Cash and Short-Term Investments	CHE	
Cash and Short-Term Investments scaled by total assets	CHE / AT	
Lagged Cash and Short-Term Investments	CHE at t-1	
Lagged Cash and Short-Term Investments scaled by total assets	(CHE / AT) at t-1	
Current Liabilities - Total	LCT	
Current Liabilities - Total scaled by total assets	LCT / AT	
Lagged Current Liabilities - Total	LCT at t-1	
Lagged Current Liabilities - Total scaled by total assets	(LCT / AT) at t-1	
Debt and Current Liabilities - Total	DLC	
Debt and Current Liabilities - Total scaled by total assets	DLC / AT	

Lagged Debt and Current Liabilities - Total	DLC at t-1	
Lagged Debt and Current Liabilities - Total scaled by total assets	(DLC / AT) at t-1	
Income Taxes Payable	TXP	
Income Taxes Payable scaled by total assets	TXP / AT	
Lagged Income Taxes Payable	TXP at t-1	
Lagged Income Taxes Payable scaled by total assets	(TXP / AT) at t-1	
Depreciation and Amortization	DP	
Depreciation and Amortization scaled by total assets	DP / AT	
Lagged Depreciation and Amortization	DP at t-1	
Lagged Depreciation and Amortization scaled by total assets	(DP / AT) at t-1	
Investment and Advances - Other	IVAO	
Investment and Advances - Other scaled by total assets	IVAO / AT	
Lagged Investment and Advances - Other	IVAO at t-1	
Lagged Investment and Advances - Other scaled by total assets	(IVAO / AT) at t-1	
Liabilities - Total	LT	
Liabilities - Total scaled by total assets	LT / AT	
Lagged Liabilities - Total	LT at t-1	
Lagged Liabilities - Total scaled by total assets	(LT / AT) at t-1	
Long-Term Debt - Total	DLTT	
Long-Term Debt - Total scaled by total assets	DLTT / AT	
Lagged Long-Term Debt - Total	DLTT at t-1	
Lagged Long-Term Debt - Total scaled by total assets	(DLTT / AT) at t-1	
Short-Term Investments - Total	IVST	

Short-Term Investments - Total scaled by total assets	IVST / AT	
Lagged Short-Term Investments - Total	IVST at t-1	
Lagged Short-Term Investments - Total scaled by total assets	(IVST / AT) at t-1	
Preferred/Preference Stock (Capital) - Total	PSTK	
Preferred/Preference Stock (Capital) - Total scaled by total assets	PSTK / AT	
Lagged binary variable indicating negative accruals per share; where accruals = ΔACT + Δ DLC - Δ CHE - ΔLCT	Dummy variable: 1 - negative lagged accruals per share, 0 o/w	So
Lagged binary variable indicating positive accruals per share; where accruals = where accruals = ΔACT + Δ DLC - Δ CHE - ΔLCT	Dummy variable: 1 - positive lagged accruals per share, 0 o/w	So
Interaction term of Negative Earnings Dummy and Earnings	Negative Earnings*Earnings in year t	EP (Li and Mohanram)
Earnings in year t scaled by shares outstanding	(IB – SPI) / CSHO	Part of Interaction term of Negative Earnings Dummy and Earnings (Li and Mohanram)
Book value of equity divided by number of shares outstanding	CEQ / CSHO	RI (Li and Mohanram)
Common/Ordinary Equity - Total	CEQ	Part of Book value of equity (Li and Mohanram)
Common/Ordinary Equity - Total scaled by total assets	CEQ / AT	
Common Shares Outstanding	CSHO	
Common Shares Outstanding scaled by total assets	CSHO / AT	
Inventory	Δ inventory (INVT) - Δ SALE	Abarbanell and Bushee, Lev and Thiagarajan,

		Gerakos and Gramacy
Inventories - Finished Goods	INVFG	Part of Inventory (Abarbanell and Bushee, Lev and Thiagarajan, Gerakos and Gramacy)
Inventories - Finished Goods scaled by total assets	INVFG / AT	
Lagged Inventories - Finished Goods	INVFG at t-1	
Lagged Inventories - Finished Goods scaled by total assets	(INVFG / AT) at t-1	
Inventories - Total	INVT	
Inventories - Total scaled by total Assets	INVT / AT	Ou and Penman
Lagged Inventories - Total	INVT at t-1	Part of Inventory
Lagged Inventories - Total scaled by total Assets	(INVT / AT) at t-1	
Sales/Turnover (Net)	SALE	Gerakos and Gramacy
Sales / Turnover (Net) scaled by total assets, end-of-year values	SALE / AT (Ou and Penman calculated using end of year value)	Ou and Penman, Holthausen and Larcker
Sales / Turnover (Net) scaled by total assets, averaging	SALE / AT (Holthausen and Larcker calculated using average of total assets -- beginning and end of year)	Holthausen and Larcker
Change in Accounts Receivable - Change in Sales	Δ RECT - Δ SALE	Abarbanell and Bushee, Gerakos and Gramacy, Lev and Thiagarajan
Accounts Receivable	RECT	Part of Change in Accounts Receivable - Change in Sales
Accounts Receivables scaled by total assets	RECT / AT	
Lagged Accounts Receivable	RECT at t-1	
Lagged Accounts Receivables scaled by total assets	(RECT / AT) at t-1	
Lagged Sales/Turnover (Net)	SALE at t-1	
Lagged Sales/Turnover (Net) scaled by total assets -- Ou and Penman way	SALE at t-1/ AT (Ou and Penman calculated using end of year value)	
Lagged Sales/Turnover (Net) scaled by total assets -- Holthausen and Larcker way	SALE t-1 / AT (Holthausen and Larcker calculated using average of total assets -- beginning and end of year)	

Capital Expenditures (Firm)	CAPXV	Part of % Change in Capital Expenditure / Total Assets(Ou and Penman, Holthausen and Larcker(
Capital Expenditures (Firm) scaled by total assets	CAPXV / AT	
Lagged Capital Expitures (Firm)	CAPXV at t-1	
Lagged Capital Expenditures (Firm) scaled by total assets	(CAPXV / AT) at t-1	
Change in Sales Minus Change in Gross Margin	Δ SALE- Δ Gross Margin (SALE - COGS); Δ SALE = [$SALE_t$ - E($SALE_t$)] / E($SALE_t$) where E($SALE_t$) = ($SALE_{t-1}$ + $SALE_{t-2}$)/2	Abarbanell and Bushee, Lev and Thiagarajan
Cost of Goods Sold	COGS	Part of Change in Sales Minus Change in Gross Margin
Cost of Goolds Sold Scaled by Total Assets	COGS / AT	
Lagged Cost of Goods Sold	COGS at t-1	
Lagged Cost of Goolds Sold Scaled by Total Assets	COGS / AT at t-1	
Change in SG&A Expenses - Change in Sales	Δ XSGA - Δ SALE	Abarbanell and Bushee, Lev and Thiagarajan, Gerakos and Gramacy
Selling, General and Administrative Expense	XSGA	Part of Change in SG&A expenses minus Change in Sales
Selling, General and Administrative Expense, scaled by total assets	XSGA / AT	
Lagged Selling, General and Administrative Expense	XSGA at t-1	
Lagged Selling, General and Administrative Expense, scaled by total assets	(XSGA / AT) at t-1	
Effective Tax Rate	TXT / (PI + AM)	Abarbanell and Bushee, Lev and Thiagarajan
Pretax Income	PI	Part of Effective Tax Rate
Pretax Income scaled by total assets	PI / AT	
Lagged Pretax Income	PI at t-1	
Lagged Pretax Income scaled by total assets	(PI / AT) at t-1	
Amortization of Intangibles	AM	

Amortization of Intangibles scaled by total assets	AM / AT	
Lagged Amortization of Intangibles	AM at t-1	
Lagged Amortization of Intangibles scaled by total assets	(AM / AT) at t-1	
Labor Force	$(\frac{SALE_{t-1}}{EMP_{t-1}} - \frac{SALE_t}{EMP_t})/\frac{SALE_{t-1}}{EMP_{t-1}}$	Abarbanell and Bushee, Lev and Thiagarajan
Lagged Employees	EMP at t-1	Part of Labor Force
Lagged Employees scaled by total assets	(EMP / AT) at t-1	
Employees	EMP at t	
Employees scaled by total assets	EMP at t / AT	
Indicator variable for dividends paid	=1 if dvt > 0; = 0 o/w	Gerakos and Gramacy
R&D Expense	XRD	Gerakos and Gramacy
R&D Expense scaled by total assets	XRD / AT	
Total Liabilities	LT	Gerakos and Gramacy
Total Liabilities scaled by total assets	LT / AT	
Shareholder's equity	SEQ	Gerakos and Gramacy
Shareholder's equity scaled by total assets	SEQ / AT	
Advertising	XAD	Gerakos and Gramacy
Advertising expense scaled by total assets	XAD / AT	
Extraordinary items and discontinued operations	XIDO	Gerakos and Gramacy
Extraordinary items and discontinued operations scaled by total assets	XIDO / AT	
Interest expense	XINTD	Gerakos and Gramacy
Interest expense scaled by total assets	XINTD / AT	
Market Value of Equity	PRCC_F*CSHO	Gerakos and Gramacy
Provision for Doubtful Receivables	Δ Gross Receivables (RECT+RECD) - Δ Doubtful Receivables (RECD)	Lev and Thiagarajan
Gross Receivables	RECT+RECD	Part of Provision for Doubtful
Gross Receivables scaled by total assets	(RECT+RECD) / AT	

Lagged Gross Receivables	RECT+RECD at t-1	Receivables (Lev and Thiagarajan)
Lagged Gross Receivables scaled by total assets	(RECT+RECD) / AT at t-1	
Change in Sales minus Change in Order Backlog	Δ Sales - Δ Order Backlog (OB)	Lev and Thiagarajan
Order Backlog	OB	Part of Change in Sales minus Change in Order Backlog (Lev and Thiagarajan)
Order Backlog scaled by total assets	OB / AT	
Lagged Order Backlog	OB at t-1	
Lagged Order Backlog scaled by total assets	(OB / AT) at t-1	
Flag for Positive Change in Return on Assets	=1 if ΔROA > 0, = 0 otherwise (where ROA = IB / AT)	Piotroski
Cash flow from operations	OANCF	Piotroski
Cash flow from operations scaled	OANCF / AT	
Cash flow from operations lagged	OANCF at t-1	
Cash flow from operations scaled, lagged	(OANCF / AT) at t-1	
Flag for Positive Return on Assets -- IB / AT = return on assets, ROA = return on assets	=1 if ROA >0; = 0 o/w (where ROA = IB / AT)	Piotroski
Flag for positive cash flows from operation	=1 if CFO >0; = 0 o/w (where CFO = OANCF / AT)	Piotroski
ACCRUAL	Accrual = current year's net income before extraordinary items - cash flow from operations, scaled by beginning-of-the-year total assets	Piotroski
Indicator of Positive Accruals (F_ACCRUAL)	=1 if CFO>ROA; = 0 o/w	Piotroski
Ratio of Long-Term debt to average assets (ΔLEVER)	DLTT / AT (historical average)	Piotroski
Indicator Variable for change in long-term debt to average assets ratio (F_ΔLEVER)	=1 if ΔLEVER >0 in year preceding; = 0 o/w	Piotroski, Ou and Penman, Holthausen and Larcker
Change in firm's current ratio between current and prior year; where current ratio is ratio of current assets to current liabilities at fiscal year end (ΔLIQUID)	(ACT/LCT) at t - (ACT/LCT) at t-1	Ou and Penman, Holthausen and Larcker

Indicator Variable for chane in firm's current ratio (F_ΔLIQUID	= 1 if ΔLIQUID >0; =0 o/w	Piotroski
Ratio of Long-Term debt to average assets	DLTT / AT (historical average)	Piotroski
Current Ratio	ACT/LCT	
Lagged Current Ratio	(ACT/LCT) at t-1	
Indicator Variable of whether common equity was issued	=1 if firm **did** NOT issue common equity in the year before, = 0 otherwise CSHI = common stock issuance	Piotroski
Current gross margin ratio (gross margin scaled by total sales) less prior year's gross margin ratio (ΔMARGIN	[(SALE - COGS)/SALE at t] - [(SALE - COGS)/ SALE at t-1]	Piotroski, Ou and Penman, Holthausen and Larcker
Current Gross Margin Ratio	(SALE - COGS)/SALE	Ou and Penman, Holthausen and Larcker
Prior Year's gross margin ratio	(SALE - COGS)/ SALE at t-1	
Indicator Variable for change in gross margin ratio (F_ΔMARGIN)	=1 if current gross margin ratio less prior year's gross margin ratio is positive, = 0 otherwise	Piotroski
Current year asset turnover ratio (total sales scaled by beginning-of-the-year total assets) less prior year's asset turnover ratio (ΔTURN)	(SALE / AT at t) - (SALE / AT at t-1)	Piotroski
Indicator Variable (F_ΔTURN)	=1 if ΔTURN is positive, = 0 otherwise	Piotroski
Composite Score created by Piotroski	= F_ROA + F_ΔROA + F_CFO + F_ACCRUAL + F_ΔMARGIN + F_ΔTURN + F_ΔLIQUID + F_ΔLEVER + EQ_OFFER	Piotroski
Quick Ratio	(ACT - INVT) / LCT	Ou and Penman, Holthausen and Larcker
Current Assets - Current Inventory	ACT - INVT	Numerator of Quick Ratio
%Δ in Quick Ratio	([(ACT - INVT) / LCT at t] - [(ACT - INVT) / LCT at t-1]) / [(ACT - INVT) / LCT at t-1]	Ou and Penman, Holthausen and Larcker
Lagged Quick Ratio	(ACT - INVT) / LCT at t-1	
Days Sales in Accs. Receivable	RECT*(365/SALE)	Ou and Penman, Holthausen and Larcker

%Δ in Days Sales in Accs. Receivable	([RECT*(365/SALE) at t] - [RECT*(365/SALE) at t-1]) / [RECT*(365/SALE) at t-1]	Ou and Penman, Holthausen and Larcker
Lagged Days Sales in Accs. Receivable	RECT*(365/SALE) at t-1	
Inventory Turnover	COGS / INVT	Ou and Penman, Holthausen and Larcker
Lagged Inventory Turnover	(COGS / INVT) at t-1	
%Δ in Inventory Turnover	[(COGS / INVT at t) - (COGS / INVT at t-1)] / (COGS / INVT at t)	Ou and Penman, Holthausen and Larcker
%Δ (INVT / at)	[(INVT / AT at t) - (INVT / AT at t-1)] / (INVT / AT at t)	Ou and Penman, Holthausen and Larcker
%Δ in Inventory	[(INVT at t) - (INVT at t-1)] / (INVT at t)	Ou and Penman, Holthausen and Larcker
%Δ in sales	[(SALE at t) - (SALE at t-1)] / (SALE at t-1)	Ou and Penman, Holthausen and Larcker
%Δ in depreciation	[(DP at t) - (DP at t-1)] / (DP at t-1)	Ou and Penman, Holthausen and Larcker
Depreciation lagged	DP at t-1	
Dividends per share	DVT / CSHO	So
Dividends per share lagged	(DVT / CSHO) at t-1	
Δ in dividend per share	[(DVT / CSHO) – (DVT / CSHO at t-1)] / (DVT / CSHO at t-1)	Ou and Penman, Holthausen and Larcker
Depreciation / Plant Assets	DP / PPEGT	Ou and Penman, Holthausen and Larcker
Depreciation / Planet Assets lagged	(DP / PPEGT) at t-1	
%Δ in Depreciation / Plant Assets	(DP / PPEGT at t) - (DP / PPEGT at t-1) / (DP / PPEGT at t-1)	Ou and Penman, Holthausen and Larcker
Return on opening equity	IB at t / SEQ at t-1	Ou and Penman, Holthausen and Larcker

Δ in Return on Opening Equity	[(IB at t / SEQ at t-1) – (IB at t - 1 / SEQ at t-2)] / (IB at t-1 / SEQ at t – 2)	Ou and Penman, Holthausen and Larcker
%Δ in (capital expenditure / total assets)	[(CAPXV / AT at t) - (CAPXV / AT at t-1)] / (CAPXV / AT at t-1)	Ou and Penman, Holthausen and Larcker
%Δ in (capital expenditure / total assets), lagged	[(CAPXV / AT at t – 1) - (CAPXV / AT at t-2)] / (CAPXV / AT at t-2)	Ou and Penman, Holthausen and Larcker
Debt-Equity Ratio	DLC / SEQ	Ou and Penman, Holthausen and Larcker
%Δ in debt to equity ratio	[(DLC / SEQ at t) - (DLC / SEQ at t-1)] / (DLC / SEQ at t)	Ou and Penman, Holthausen and Larcker
Debt-Equity Ratio Lagged	(DLC / SEQ) at t-1	Part of change in debt to equity ratio
LT debt to equity	DLTT / SEQ	Ou and Penman, Holthausen and Larcker
LT debt to equity lagged	(DLTT / SEQ) at t-1	
%Δ in LT debt to equity	[(DLTT / SEQ at t) - (DLTT / SEQ at t-1)] / (DLTT / SEQ at t - 1)	Ou and Penman, Holthausen and Larcker
Equity to fixed assets	SEQ / PPEGT	Ou and Penman, Holthausen and Larcker
Gross PPE	PPEGT	
Gross PPE scaled by total assets	PPEGT / AT	
%Δ in Equity to fixed assets	[(PPEGT /AT at t) - (DLTT / SEQ at t-1)] / (DLTT / SEQ at t)	Ou and Penman, Holthausen and Larcker
Times interest earned	IB / XINT	Ou and Penman, Holthausen and Larcker
times interest earned lagged	(IB / XINT) at t - 1	
%Δ in times interest earned	[(IB / XINT) – (IB at t -1 / XINT at t-1)] / (IB at t-1 / XINT at t-1)	Ou and Penman,

		Holthausen and Larcker
%Δ in sales / total assets	[(SALE / AT at t) - (SALE / AT at t-1)] / (SALE / AT at t-1)	Ou and Penman, Holthausen and Larcker
Return on total assets	IB / AT	Ou and Penman, Holthausen and Larcker
Return on closing equity	IB / SEQ	Ou and Penman, Holthausen and Larcker
Op. profit (before dep.) to sales	OIBDP / SALE	Ou and Penman, Holthausen and Larcker
Op. profit (before dep.) to sales lagged	(OIBDP / SAL)E at t-1	
%Δ in Op. profit (before dep.) to sales	[(OIBDP / SALE) – (OIBDP at t – 1 /SALE at t – 1)] / (OIBDP at t – 1 / SALE at t – 1)	Ou and Penman, Holthausen and Larcker
Pretax income to sales	PI / SALE	Ou and Penman, Holthausen and Larcker
Pretax income to sales lagged	(PI / SALE) at t-1	
%Δ in pretax income to sales	[(PI/SALE) – (PI at t-1 / SALE at t-1)] / (PI at t-1/SALE at t-1_	Ou and Penman, Holthausen and Larcker
Net profit margin	SALE / IB	Ou and Penman, Holthausen and Larcker
Net profit margin lagged	(SALE / IB) at t-1	
%Δ in net profit margin	[(SALE / IB) – (SALE at t-1 /IB at t-1)] / (SALE at t-1/IB at t-1)	Ou and Penman, Holthausen and Larcker
Sales to total cash	SALE / CHE	Ou and Penman, Holthausen and Larcker
Sales to accs. Receivable	SALE / RECT	Ou and Penman, Holthausen and Larcker
Sales to Inventory	SALE / INVT	Ou and Penman,

		Holthausen and Larcker
%Δ in Sales to Inventory	[(SALE / INVT at t) - (SALE / INVT at t-1)] / (SALE / INVT at t-1)	Ou and Penman, Holthausen and Larcker
Sales to Inventory lagged	(SALE / INVT) at t-1	
Sales to Working Capital	SALE/WCAP	Ou and Penman, Holthausen and Larcker
Sales to Working Capital at t-1	(SALE/WCAP) at t-1	
%Δ in Sales to Working Capital	[(SALE/WCAP) – (SALE at t-1/WCAP at t-1)] / (SALE at t-1/WCAP at t-1)	Ou and Penman, Holthausen and Larcker
Sales to fixed assets	SALE / PPEGT	Ou and Penman, Holthausen and Larcker
%Δ in R&D	[XRD-(XRD at t-1)] / (XRD at t-1)	Ou and Penman
R&D lagged	XRD at t-1	part of change in R&D expense below
%Δ in (R&D / sales)	[(XRD / SALE) – (XRD at t-1/ SALE at t-1)] / (XRD at t-1/ SALE at t-1)	Ou and Penman
R&D / sales	XRD / SALE	
R&D / sales lagged	(XRD / SALE) at t-1	
%Δ in advertising expense	[XAD -(XAD at t-1)] / (XAD at t-1)	Ou and Penman
advertising expense lagged	XAD at t-1	
%Δ in (advertising/sales)	[(XAD / SALE) – (XAD at t-1/ SALE at t-1)] / (XAD at t-1/ SALE at t-1)	Ou and Penman
advertising / sales	XAD / SALE	
advertising / sales lagged	XAD / SALE at t-1	
%Δ in total assets	[AT -(AT at t-1)] / (AT at t-1)	Ou and Penman, Holthausen and Larcker,, So
total assets lagged	AT at t-1	
Cash flow to total debt	(OANCF + IVNCF + FINCF) / DLC	Ou and Penman, Holthausen and Larcker

Cash Flow – Financing Activities	FINCF	
Cash Flow – Investing Activities	IVNCF	
Working capital / total assets	WCAP / AT	Ou and Penman, Holthausen and Larcker
Working capital / total assets lagged	(WCAP / AT) at t-1	
%Δ in (working capital / total assets)	[(WCAP / AT) – (WCAP at t-1/ AT at t-1)] / (WCAP at t-1/ AT at t-1)	Ou and Penman, Holthausen and Larcker
Operating Income / total assets	OIBDP / AT	Ou and Penman, Holthausen and Larcker
operating income scaled by total assets lagged	(OIBDP / AT) at t-1	
%Δ in (operating income / total assets)	[(OIBDP / AT) – (OIBDP at t-1/ AT at t-1)] / (OIBDP at t-1/ AT at t-1)	Ou and Penman
total uses of fund	FUSET	
total uses of funds lagged	FUSET at t-1	
%Δ in total uses of fund	[FUSET -(FUSET at t-1)] / (FUSET at t-1)	Ou and Penman
total sources of funds	FSRCT	
total sources of funds lagged	FSRCT at t-1	
%Δ in total sources of fund	[FSRCT -(FSRCT at t-1)] / (FSRCT at t-1)	Ou and Penman
Repayment of LT debt as % of total LT debt	DLTR / DLTT	Ou and Penman, Holthausen and Larcker
Reduction of long-term debt	DLTR	part of repayment of LT Debt
Reduction of long-term debt, issued by total assets	DLTR / AT	
Issuance of LT debt as % of total LT debt	DLTIS / DLTT	Ou and Penman, Holthausen and Larcker
LT debt issued	DLTIS	part of Issuance of LT debt as %

		of total LT debt
LT debt issued scaled by assets	DLTIS / AT	
Purchase of treasury stock as % of stock	(TSTK at t - TSTK at t-1) / (CSTK + PSTK); amount of treasury stock / (common stock + preferred stock)	Ou and Penman
Amount of treasury stock	TSTK	Part of purchase of treasury stock as % of stock
Lagged amount of treasury stock	TSTK at t-1	
Amount of treasury stock scaled by total assets	TSTK / AT	
Funds from operations	FOPO	
funds from operations lagged	FOPO at t-1	
Funds from operations scaled by total assets	FOPO / AT	
Funds from operations scaled by total assets lagged	(FOPO / AT) at t-1	
%Δ in funds	[FOPO -(FOPO at t-1)] / (FOPO at t-1)	Ou and Penman, Holthausen and Larcker
%Δ in LT debt	[DLTT -(DLTT at t-1)] / (DLTT at t-1)	Ou and Penman, Holthausen and Larcker
Cash dividend as % of cash flows	DV / (OANCF + IVNCF + FINCF)	Ou and Penman, Holthausen and Larcker
Cash Dividend	DV	Part of cash dividend as % of cash flows
Cash Dividend scaled by total assets	DV / AT	
working capital	WCAP	
working capital at t-1	WCAP at t-1	
%Δ in working capital	[WCAP -(WCAP at t-1)] / (WCAP at t-1)	Ou and Penman, Holthausen and Larcker
Net income over cash flows	IB / (OANCF + IVNCF + FINCF)	Ou and Penman, Holthausen and Larcker
Book-to-market	PRCC_C * CSHO / CEQ	So

End of year fiscal share price	PRCC_F	So

Table 2: Grid of tuning parameters Searched

All possible combinations of the following features and levels were searched:

Learning Rate	Max Depth	Minimum Child Weight	Number of Trees
0.01	3	1	100
0.05	4	3	300
0.10	5	5	500
0.15	6	7	1000
0.20	8		
0.25	10		
0.3			

\

Figure 1: Feature Importances from GBRT Model

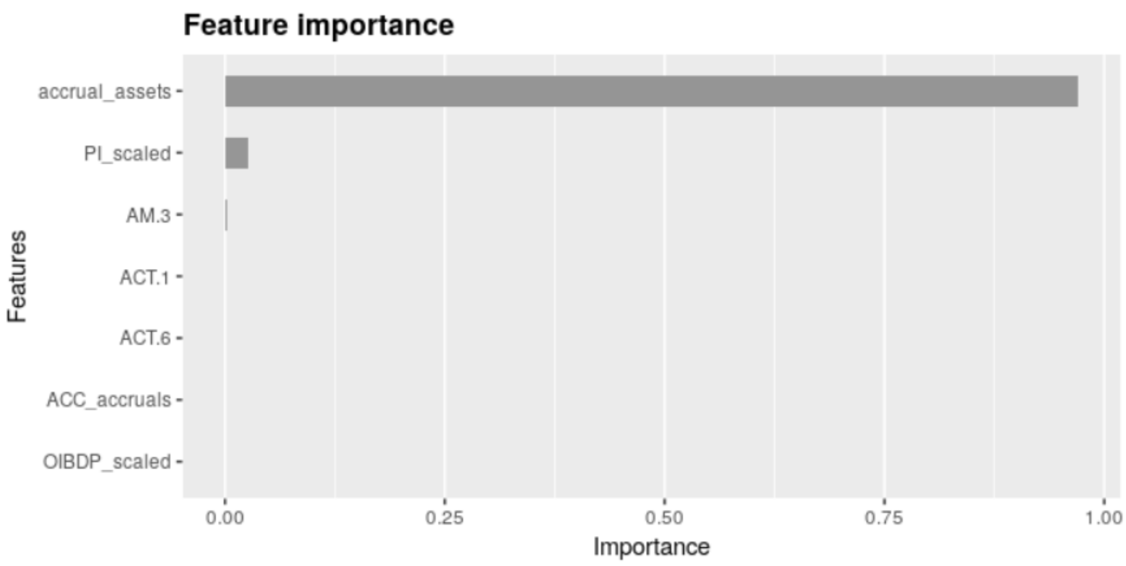

Figure 2: Absolute Percentage Error for Model vs. Accruals

References

Abarbanell, Jeffrey S. and Brian J. Bushee. "Fundamental Analysis, Future Earnings, and Stock Prices." *Journal of Accounting Research* 35, no. 1 (1997). 1-24.

Anand, Vikrant, Robert Brunner, Kelechi Ikegwu, Theodore Sougiannis. "Predicting Profitability Using Machine Learning." (2019).

Ball, Ryan T. and Eric Ghysels. "Automated Earnings Forecasts: Beat Analysts or Combine and Conquer?" *Management Science, Forthcoming* (2017).

Bertomeu, Jeremy, Edwige Cheynel, Erica Floyd, and Weimin Pan. "Ghost in the Machine: Using Machine Learning to Uncover Hidden Misstatements." (2018).

Binz, Oliver. "The Use of Accounting Information in Fundamental Analysis." (2019).

Brown, Lawrence D. and Michael S. Rozeff. "Univariate Time-Series Models of Quarterly Accounting Earnings per Share: A Proposed Model." *Journal of Accounting Research* 17, no. 1 (1979). 179-189.

Brown, Lawrence D., Robert L. Hagerman, Paul A. Griffin, and Mark E. Zmijewski. "Security analyst superiority relative to univariate time-series models in forecasting quarterly earnings." *Journal of Accounting and Economics* 9, no. 1 (1987). 61-87.

Foster, George. "Quarterly Accounting Data: Time-Series Properties and Predictive-Ability Results." *The Accounting Review* 52, no. 1 (1977). 1-21.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. "An Introduction to Statistical Learning with Applications in R" (2017), pg. 312

Gerakos, Joseph and Robert Gramacy. "Regression-Based Earnings Forecasts." Chicago Booth Research Paper No. 12-26 (2013).

Hou, Kewei, Mathijs A. van Dijk, and Yinglei Zhang. "The implied cost of capital: A new approach." *Journal of Accounting and Economics* 53, no. 3 (2012). 504-526.

Lev, Baruch and S. Ramu Thiagarajan. "Fundamental Information Analysis." *Journal of Accounting Research* 31, no. 2 (1993). 190-215.

Li, Kevin K., and Partha Mohanram. "Evaluating cross-sectional forecasting models for implied cost of capital." *Review of Accounting Studies* 19, no. 3 (2014). 1152-1185.

Perols, Johan. 2011. Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms.", *Auditing: A Journal of Practice & Theory* 30 (2): 19-50.

So, Eric C. "A new approach to predicting analyst forecast errors: Do investors overweight analysts forecasts?” *Journal of Financial Economics* 108, no. 3 (2013). 615-640.