Aggregating Artificially Intelligent Earnings Forecasts*

Vidhi Chhaochharia[a], Alok Kumar[a], Swathi Murali[a], Ville Rantala[a]

[a]Department of Finance, Miami Herbert Business School, Coral Gables, FL 33124, USA.

Abstract

This paper proposes a new artificial intelligence-based framework for aggregating the opinions of sell-side equity analysts. Our key conjecture is that analyst-level individual errors are more predictable and can be corrected more effectively than the aggregate consensus forecast. Artificial neural network models are able to learn the individual-level biases of sell-side equity analysts and can compensate for their biases more effectively. Using the bias-corrected consensus earnings forecasts and a neural network-based aggregation method, we find that the accuracy of the consensus forecast improves by 24% relative to a benchmark linear model. The neural network-based model can predict the correct sign for the earnings surprise 68% of the time. A portfolio strategy that uses the improved consensus forecasts earns a monthly abnormal return of 0.50% on a risk-adjusted basis.

Keywords: Sell-side equity analysts, bias correction, artificial intelligence, neural-network, return predictability.

*We would like to thank the seminar participants at University of Miami for useful suggestions and helpful comments. We are responsible for all remaining errors and omissions.

Email addresses: vidhi@miami.edu (Vidhi Chhaochharia), akumar@miami.edu (Alok Kumar), sxm2003@miami.edu (Swathi Murali), vrantala@miami.edu (Ville Rantala)

1. Introduction

Sell-side analysts are important financial intermediaries who use a mix of private and public information to forecast earnings and other firm-level variables. The analyst forecast is often used as a proxy for firm earnings and serves as a fundamental input in various valuation models used by market participants. Unfortunately, analyst forecasts are known to be biased, which can produce biased estimates of earnings and potentially inaccurate valuations (e.g., So, 2013).

In this study, we develop a new artificial intelligence based framework for aggregating the opinions of sell-side equity analysts. Our key insight is that analyst forecast errors are more predictable than the aggregate consensus forecast. Therefore, analyst-level "mistakes" can be corrected more effectively, especially when we utilize flexible and analyst-specific models to capture the heterogeneity in analyst forecasting behavior.

Specifically, motivated by the re-emergence of artificial intelligence and machine learning methods, we use multi-layered, feed-forward neural network (NN) models that learns about the behavior of each analyst. These models learn about the analyst-level biases iteratively and compensate for the mistakes before aggregating the bias-corrected forecasts.

The analyst setting is an ideal laboratory to implement neural network models for a number of reasons. First, we can observe analyst forecasts for a number of years as analysts typically follow a number of firms and typically make forecasts at least four times a year. Additionally, multiple analysts often follow the same firm. The large cross-section and the long time dimension gives us the breadth in data needed to define neural network models that need a training and predictive period for greater predictive power.

Our insight that analyst-level biases are more predictable than the consensus builds upon the recent accounting literature that has recognized this possibility and has developed methods that focus on predicting aggregate future earnings rather than modeling past analyst-level forecast errors (e.g., Hughes et al., 2008; So, 2013). We are also motivated by the recent accounting and finance studies that provide evidence of various analyst-level

biases that are persistent (e.g., Lin and McNichols, 1998; Hong and Kubik, 2003). These biases are likely to be induced by strategic factors (Das et al., 1998 and Lim, 2001), skewness-related optimization (Gu and Wu, 2003), or uncertainty regarding the firm's earnings process (Linnainmaa et al., 2016).

Our contribution in this paper is to extend these insights and explicitly correct for these biases using analyst-specific neural network models. We assume that a forecast consists of the "true" estimate added with an analyst-specific bias that can be related to agency, strategic, or behavioral factors. Our neural network models systematically de-bias analyst earnings estimates and allows us to obtain the "true" earnings estimate.

The choice of neural network models to capture analyst forecasting behavior and improve the aggregation process is motivated by the growth of machine learning techniques and predictive statistical methods in both business and economics research (e.g., Mullainathan and Spiess, 2017; Agrawal et al., 2018). We add to this small but growing literature by demonstrating that artificial intelligence and machine learning techniques can be applied in the analyst setting to make better forecasts.

We show that a neural network can be trained to learn about individual level analyst biases and create a more accurate consensus after compensating for the bias. We also find that an artificially intelligent model can learn about the aggregate biases of analysts without specifically accounting for them one at a time. Both approaches allow us to better predict individual or aggregate forecast errors.

In our empirical tests, we use both panel models and analyst-level models to predict analysts' forecast errors. The panel models use data on firm characteristics and the analyst-wise models use data on analysts' previous forecast errors and forecasts. The panel tests build upon the estimation framework in So (2013), and create a linear panel regression model and a neural network model to predict errors in analysts' earnings forecasts. The linear model is our benchmark model and the linear model and the neural network model both use the same data when predicting forecast errors. As an alternative model specifica-

tion, we define analyst-specific linear models and neural network models that use only the past analyst errors and forecast horizon to predict the analyst's forecast error. To predict consensus-level earnings surprises, we form adjusted consensus forecasts that account for individual analysts' predicted forecast errors based on each model.

Out of the four models, the analyst-wise neural network model performs the best in our tests. The model is able to predict both individual and consensus forecast errors better than the linear benchmark model, and the relative improvement in forecast accuracy is higher with consensus forecasts.

Specifically, by using the analyst-wise neural network model, we are able to improve earnings predictability by 24% in comparison to the linear regression model. Our analyst-wise neural network model is also much better at predicting the sign of the earnings surprise. It makes the correct prediction 68% of the time, while the linear model is able to predict the sign correctly only 54% of the time. The relative difference in performance persists also in out-of-sample estimation, where we use pre-2010 observations as training data, and 2010–2019 as an out-of-sample period.

To interpret these results economically, we develop a Long−Short trading strategy based on the difference between the regular and neural network consensus. We sort firms into decile portfolios based on their neural network-predicted earnings surprise captured by the difference between the actual and adjusted consensus forecasts. The long portfolio consist of firms with the most positive predicted earnings surprises and the short portfolio consists of firms with the most negative predicted earnings surprises. This trading strategy generates an abnormal return of 0.50% per month on a risk-adjusted basis.

These results contribute to the literature on analyst biases (Easterwood and Nutt, 1999, Lys and Sohn, 1990 and Mendenhall, 1991). Our paper is also related to how investors incorporate analyst information while forming expectations of firm value (Malmendier and Shanthikumar, 2007 and Mikhail et al., 2007). Our results indicate that investors could use artificially-intelligent earnings consensuses to de-bias analyst forecasts and improve the

performance of their trading strategies.

Our contribution not only lies in using a neural network to create a more accurate earnings consensus but also in the fact that the network itself has an economic interpretation. Typically, analyst biases are identified and captured one at a time. Additionally, we know that analysts learn from each other (Kumar et al., 2021). Our neural network is able to capture this inter-relationships without explicitly modeling them.

The rest of the paper is organized as follows. Section 2 discusses our data and benchmark linear regression models. Section 3 introduces our neural network models and linear benchmark models. Section 4 examines how well these models predict individuals analysts' forecast errors and earnings surprises. Section 5 tests a portfolio strategy based on predicted earnings surprises. Section 6 concludes with a brief discussion.

2. Data Sources and Variable Definitions

Our main data source are quarterly earnings announcements and associated earnings forecasts from the Institutional Broker Estimates System (I/B/E/S) detail history file. Additionally, we use share price data from the Center for Research in Security Prices (CRSP), financial statement data from Compustat. Monthly data on the risk-free rate, market excess return (MKT), size factor (SMB), value factor (HML) are obtained from the Kenneth French Data Library.

We include all firms with CRSP share codes 10 or 11. The sample starts in 1984 and ends in the third quarter of 2019. We exclude analysts coded as anonymous because it is not possible to track their earnings forecasts across quarters.

Consistent with prior studies, we only include the latest forecast or revision before the earnings announcement date for each analyst. To address potential errors and data quality concerns in I/B/E/S, we require that the date when an analyst forecast becomes effective (ACTDATS) must be on or after the analyst forecast announcement date (ANNDATS). We also eliminate observations where a forecast review date (REVDATS) precedes the forecast

announcement date (ANNDATS).

We define an analysts' forecast error as the difference between the earnings per share forecast and the actual earnings per share in I/B/E/S. Following prior literature, we divide this difference with the share price. We use share price ten trading days prior to the announcement. We winsorize the top and bottom 1% of the forecast errors to limit the effect of outliers. We also analyze consensus forecasts and consensus forecast errors. The consensus forecast is defined as the median of all outstanding analysts' earnings per share estimates on the same earnings announcement.

Table 1, Panel A provides descriptive statistics on analyst forecasts and forecast errors. The average forecast error in the sample is 0.056 and the standard deviation is 1.14. On average, an earnings announcement is covered by 5.7 analysts and the median number of analysts is four.

3. Bias Prediction Models

Our goal is to examine whether artificially-intelligent earnings forecasts from a neural network can potentially perform better than linear prediction models. If the neural network models are able to de-bias analysts more effectively, we expect earnings consensus created using neural networks to be closer to the realized earnings consensus.

As a first step in this comparison, we specify a benchmark model that can predict analysts' forecast errors. We rely on the extant analyst literature in both accounting and finance that have created models to predict analyst behavior. Forecast prediction models can be broadly categorized into two groups: panel models and analyst-wise models. Panel models, like the one used in So (2013) model forecast errors as function of firm characteristics. Alternatively, we can model each analyst's forecast error as a function of analyst characteristics, as in Clement and Tse (2005).

We follow both approaches and create both panel and analyst-wise regression models. In the next section, we describe our panel and analyst-wise regression models that serve as

benchmarks for evaluating our neural network based models.

3.1. Linear Panel Benchmark Regressions

We start with linear panel regressions explaining analyst-level forecast errors using earnings forecast observations. Our explanatory variables include firm characteristics and forecast-specific characteristics following closely the model specification of So (2013). Specifically, we include log(market capitalization), book-to-market, momentum, accruals and consensus long term growth forecast.

Table 2 reports descriptive statistics for these explanatory variables. Market capitalization is calculated as share price (CRSP Item PRC) ten days before the earnings announcement multiplied with shares outstanding (Item SHROUT). Book-to-market is book equity divided market equity and we define book equity as in Davis et al. (2000). Momentum is calculated as the return 6 months prior to the earnings announcement data less the S&P 500 returns. Similar to So (2013), accruals is measured as total accruals scaled by total assets and long term growth forecast is the long term consensus from I/B/E/S. We measure quarterly earnings growth based on the realized earnings in I/B/E/S (Item ACTUAL). We winsorize the top and bottom 1% of the non-logarithmic explanatory variables. Table 2, Panel A presents the summary statistics for independent variables in our linear panel regression model.

Table 3 shows results from linear regressions explaining analysts' forecast errors with these variables. The regression models cumulatively add explanatory variables one by one. The coefficients on all variables except earnings growth are statistically significant. R^2 values in specifications that include firm and earnings announcement characteristics range between 1.2% and 1.7%. Similar to So (2013) all variables are statistically significant at the 1% level.

3.2. Panel Benchmark Regressions: Boosted Regression Trees

We verify our findings from the linear regression model using a Boosted Regression Trees (BRT) model, as this serves as the input in the neural network. There are several advantages to using the BRT machine learning technique. First, it allows us to verify and assess the importance of all the inputs in our linear model by provide a relative weighting importance to each independent variable. Additionally, the technique allows for non-linearity and conditioning on large data information sets without the overfitting problem.[1].

We estimate a BRT model with 10,000 boosting iterations. The dependent variable, like before, is the forecast error and the independent variables are market capitalization, book-to-market, momentum, total accruals and long term growth consensus forecast. As a first step we verify that increasing the number of trees increases the fit of the model. Next, we check the relative importance of the independent variables used in the model. Figure 1 presents our results from the BRTs. We present the results for fit and mean squared errors with 25 trees. We extend our results to 50 and 100 trees and find that our findings are similar and relatively stable. Figures 2c and 1d shows that all our 5 explanatory variables have a relative influence higher than 1% with 25 or 50 trees. Based on this evidence, we conclude that a model with all five explanatory variables is the appropriate baseline model for evaluating our neural network models.

We verify our findings by dividing our sample into a test sample ($<$ year 2010) and an out-of-sample test period (2010–2019). We find a constant learning rate of 0.1. Additionally, we are reasonably assured that overfitting is not a problem as the performance in the out of sample period does not go down relative to the test period (see Figure A.1). Further, as we increase the number of trees, the performance during the test period does not decrease.

[1]The use of BRTs is becoming popular in finance. For more details see Rossi and Utkus (2020).

3.3. Linear Analyst-wise Benchmark Regressions

In our panel model specifications described above, we assume that there is a single representative analyst. However, the extant analyst literature suggests that analysts can have different biases that might lead to very different predictions. To capture this heterogeneity, we specify a linear analyst-wise regression model to predict individual forecast errors.

Specifically, we regress forecast errors on up to four lagged forecast errors, the forecast horizon which is defined as the number of days until the end of the forecast period, the number of years for which each analyst has been submitting EPS forecasts, both for each individual firm and overall, and define them as firm and general experience respectively. To capture information about the analyst's firm portfolio, we also include the number of companies the analyst follows and the 2 digit-SIC code industries followed by the analyst as independent variables. Our choice of independent variables is guided by Clement and Tse (2005) and previous studies that have documented a positive autocorrelation in forecast errors ((Butler and Lang (1991), Mendenhall (1991), Abarbanell and Bernard (1992) and Linnainmaa et al. (2016)).

Table 2, Panel B presents the summary statistics for all control variables used in the analyst-wise regression model. The analyst-wise linear regression model suggests varied significance of the independent variables. Similar to our panel model, we verify our choice of control variables in the analyst-wise models using BRTs.

3.4. Analyst-wise Benchmark Regressions: Boosted Regression Trees

We run a BRT model with 10,000 boosting iterations. The explanatory variables are again the lagged forecast errors (up to 4 lags), forecast horizon, the number of firm an analyst covers in a year, the number of years the analyst has provided EPS forecasts, the number of unique 2-digit SIC industry codes for which the analyst has made EPS forecasts and the number of prior years for which the analyst has provided EPS forecasts. Our results

are shown in Figure A.1c where we present the results with 25 trees.[2]

As expected our fit gets better as we increase the number of trees. Of all our independent variables only lagged errors (up to 4 lags) and forecast horizon have relative influence higher than 1%. Therefore, we re-specify our analyst-wise model to include the 4 lagged forecast errors and forecast horizon as independent variables.

3.5. Panel and Analyst-wise Neural Network Models

We estimate two sets of neural network models: panel and analyst-level. Our panel neural network model uses all independent variables including market capitalization, book-to-market, momentum, total accruals, and long term consensus forecasts. In the analyst-wise neural network, the input data includes lagged forecast errors (up to 4 lags) and forecast horizon. Our target data (or dependent variable) consists of forecast errors.

We use a single hidden layer neural network containing 15 neurons. The neural network models for both the panel and analyst-wise settings divide the data set randomly into a training set, validation set, and a test set. The training set is used to estimate the network weights iteratively. The training continues as long as the network continues to improve the performance with the validation set. The test set or the out-of-sample test examines the estimated model after the training is complete.

We use a supervised learning algorithm to train the neural network. Since the output is known, the machine learns and iteratively adjusts its parameters (i.e., the connection weights) using an error-based feedback system. We use the commonly used Levenberg-Marquardt learning method to adjust the model parameters based on the error.

4. Prediction Model Performance

In this section, we present our key results. We compare the performance of linear and neural network models in predicting individual analysts' forecast errors and consensus

[2]We verify the robustness of our results using 50 and 75 trees.

earnings surprises. Our four test models are the linear panel model, the neural network panel model, the analyst-wise linear model, and the analyst-wise neural network model. We report full sample estimates as well as out-of-sample estimates where we use 2010–2019 as the out-of-sample test period.

4.1. Prediction Model Performance with Individual Forecasts

To begin, we examine how well different models are able to predict the forecast errors of individual analysts. We compare the linear model to the neural network model in panel estimation and analyst-wise estimation.

We first calculate the percentage of observations where the model predicts the sign of the forecast error correctly. The neural network models perform relatively better in both panel and analyst-wise settings. The best-performing model is the analyst-wise neural network model, which predicts the sign correctly in 64% of the observations. In comparison, the analyst-wise linear model makes the correct prediction 54% of the time. In the panel models, the neural network predicts the sign correctly in 59% of observations and the linear model in 51% of the observations.

Next, we focus on the absolute deviations between the predicted forecast error and the actual I/B/E/S forecast error of the analyst. The smaller the absolute deviation, the closer the prediction is to the actual forecast error. Panel A of Table 4 reports statistics based on the full sample.

We find that the analyst-wise models have lower absolute deviations than the panel models and the analyst-wise neural network model outperforms the analyst-wise linear model both based on the mean and the median. The neural network mean is 0.318, which is below the linear model mean of 0.332 and the medians are 0.107 and 0.133, respectively. The neural network mean is 4% lower than the linear model mean.

In panel models, the linear model performs relatively better based on the mean deviation, but the medians are almost identical. Mean absolute deviation of the linear model is 0.386 and the neural network mean is 0.446. The medians are 0.166 and 0.163, respectively. Table

5 complements these results by reporting distributional statistics of the predicted and actual forecast errors for the panel models and the neural network models.

Panels B and C of Table 4 report statistics based on an out-of-sample estimation where we use pre-2010 observations as the in-sample period and 2010-2019 as our out-of-sample period. Panel B reports the in-sample statistics for this estimation and Panel C reports the out-of-sample statistics. Since we are using a different version of the algorithm, these in-sample estimates and the previous full sample estimates are not completely identical.

The relative performance of the models is very similar in the out-of-sample statistics. This finding suggests that overfitting is not a problem in the neural-network estimation. The analyst-wise neural network model has the lowest mean and median absolute deviation from the actual forecast error. Its mean is 0.259, and the analyst-wise linear model mean is 0.338. In panel models, the neural network model does relatively better with a mean of 0.280 compared to the linear model mean of 0.297. As expected, the out-of-sample medians are higher than the in-sample medians for all models.

4.2. Prediction Model Performance with Consensus Forecasts

Our results thus far suggest that earnings predictions based on the analyst-wise neural network outperform alternative models in terms of accuracy. If this is indeed the case, then compared to the standard consensus forecasts, the artificially intelligent consensus should be able to better predict earnings and earnings surprises. In this section, we test this conjecture.

By convention, the consensus forecast is defined as the median forecast among analysts covering an earnings announcement. We follow the same procedure when defining our consensus forecast error predictions for the four models. We first form adjusted forecasts by deducting the forecast error predicted by the model from the actual analyst forecast. In other words, we calculate what the forecast would be without the predicted forecast error. Then, we define the consensus prediction as the median prediction among the analysts covering the same earnings announcement.

Table 6 reports the ability of various models to predict the correct sign for the earnings surprise based on the model-predicted consensus error. Column 2 reports the percentage of observations where the sign of the consensus forecast error is predicted correctly by each model, and for reference, Column 1 reports the corresponding statistics based on individual forecasts discussed in the previous subsection. The results show that the analyst-wise models' ability to predict the sign correctly in consensus forecasts is better than their ability to predict the sign of individual forecasts. In contrast, the performance of the linear models is relatively weaker with consensus forecasts.

The linear panel model predicts the sign of the consensus error correctly only 46.4% of the time, indicating that it has no predictive ability with the sign of the consensus forecast error. The panel neural network model predicts the sign correctly in 55.1% of the cases. The analyst-wise linear model predicts the sign correctly in 54.6% of the observations, and the best-performing model is the analyst-wise neural network model, which predicts the correct sign 68.0% of the time. Interestingly, the aggregation of forecast error predictions appears to improve the relative performance of neural network models, but not the performance of linear models.

Table 7 reports the ability of these models to predict the consensus forecast error based on the absolute deviation between the predicted consensus error and the actual consensus error. The full sample results in Panel A show that the analyst-wise neural network model again obtains the lowest mean absolute deviation among the models. Its mean is 0.295, and the mean of the analyst-wise panel model is 0.386. These results are also illustrated in Figure 3.

The difference between these two means indicates that the absolute difference to the actual earnings surprise is 24% lower with the neural network model. Another way to evaluate the economic significance of the improvement is to compare the improvement in accuracy to the size of the typical earnings surprise in the data. Average absolute earnings surprise scaled by price among the observations used in the analyst-wise tests is 0.378,

and the difference between the accuracy of the neural network model and the linear model corresponds to 24% of the absolute error. The median improvement corresponds to 19% of the median absolute error.

In the panel models, the panel neural network model performs marginally better with a mean absolute difference of 0.373 and the panel linear model has a means absolute difference of 0.380. The medians are 0.120 and 0.143, respectively.

Table 8 reports the statistics for the actual and predicted consensus forecast errors. The average actual error is positive, but the median is negative. All the model-predicted errors have positive means, but only the neural network models have a negative sign for the median predicted forecast error. These statistics are consistent with our previous findings suggesting that the neural network models have better ability to predict the sign of the error.

Table 7 also reports out-of-sample statistics based on an estimation where we use 2010–2019 as the out-of-sample period. Panel B shows the in-sample estimates and the out-of-sample estimates are in Panel C. The analyst-wise neural network model again performs the best, and the relative difference between the linear and neural network models is almost identical to the full sample difference. In panel models, the neural network model performs relatively better than the linear model both based on the mean (0.280 vs. 0.297) and the median (0.098 vs. 0.133).[3]

Altogether, these results suggest that treating the analyst as an individual network and training a model to learn from its past errors is more effective in de-biasing earnings estimates without assuming specific biases or functional forms. The analyst-wise estimates are more accurate than the panel estimates and the analyst-wise neural network model performs better than the linear benchmark model that uses the same data. It is also notable that the relative performance of the analyst-wise neural network model improves

[3]Figure 4 presents histograms for our out-of sample results comparing the linear and neural-netowrk models.

significantly when we move from individual forecasts to consensus forecasts.

5. Predicted Earnings Surprises and Stock Returns

The results so far indicate that our neural network model can predict earnings surprises better than linear models. A natural follow-up question is whether the predicted surprises contain information that is not fully incorporated into stock prices immediately. To examine this possibility, we develop a portfolio strategy where we sort stocks on their neural-network predicted earnings surprise. We use the analyst-wise neural network model to predict earnings because our results show that its predicted consensus forecast errors are closest to the actual earnings surprises. It also has the best ability to predict the sign of the earnings surprise.

We define the predicted earnings surprise as the analyst-wise neural network consensus minus the I/B/E/S consensus scaled by end-of-month stock price. Specifically, at the end of each month, we compute adjusted EPS forecasts for all stocks with a scheduled earnings announcement in the next month. As before, we obtain the adjusted EPS forecast using analyst-wise neural network models where we use lagged forecast errors and the forecast horizon to predict earnings. We form the adjusted forecast by adding the model-predicted forecast error to individual analyst forecasts and we then define the adjusted consensus EPS forecast as the median adjusted forecast among analysts covering the earnings announcement.

We sort all stocks with a scheduled quarterly earnings announcement during the next month on the predicted earnings surprise and allocate them to decile portfolios. The Short portfolio is a value-weighted portfolio of the decile of stocks predicted to have the largest negative earnings surprises and the Long portfolio is a value-weighted portfolio of the decile of stocks predicted to have the largest positive earnings surprises. The Long portfolio contains stocks with systematically downward-biased analyst consensus earnings forecasts and the Short portfolio contains stocks with systematically upward-biased analyst consensus

earnings forecasts. Further, the Long–Short portfolio captures the difference in the returns of the Long and Short portfolios.

Portfolios 2 to 9 are value-weighted portfolios of the remaining deciles of stocks sorted on the predicted earnings surprise. All stocks are held in their respective decile portfolios until the month before their next anticipated earnings announcement or for 3 months, whichever is less. The portfolios are rebalanced each month.

Table 9 presents the average value-weighted monthly returns and return standard deviations of these eleven portfolios during the sample period. The returns are generally increasing across the decile portfolios, indicating that stocks with more positive predicted earnings surprises earn higher returns. The Long portfolio has the highest average return (1.13%) and the Short portfolio has the lowest average return (0.59%) among the decile portfolios.

The average monthly performance differential between the Long and the Short portfolio is 0.54% per month which translates into an annual performance differential of over 6%. The monthly returns of the Long−Short portfolio are statistically different from zero with a t-value of 2.50.

We then test whether the performance of the Long−Short portfolio can be explained by standard asset pricing factors. We estimate the Fama-French three factor model using the monthly excess returns of the Long−Short portfolio, the Long portfolio, and the Short portfolio as the dependent variables.

The results in Table 10 show that the Long−Short portfolio has a statistically significant monthly alpha of 0.50% with a Newey-West adjusted standard error of 0.254. The alphas of the Short and the Long portfolio are negative, but only the alpha of the Short portfolio is statistically significant. The MKT coefficient of the Long−Short portfolio is -0.230 and statistically significant, indicating that its returns have a negative covariance with the market portfolio.

These results show that the neural network model captures information that is relevant

to the market and that it can potentially be used as the basis for a trading strategy. It is notable that the analyst-wise neural network algorithm only uses data on previous analyst forecasts and analyst characteristics, but it does not utilize any information about the firm whose earnings the analyst is forecasting. One interpretation for these abnormal portfolio returns is that analysts' forecast errors have a predictable component and the market does not fully account for the bias.

6. Summary and Conclusions

The individual forecasts of sell-side equity analysts as well as their consensus forecasts guide the information aggregation process in financial markets. Various market participants use these forecasts as inputs when forming their market expectations. Unfortunately, analysts exhibit many biases that make their individual forecasts and the consensus forecast noisy.

In this paper, we propose a new artificial intelligence and machine learning-based framework for effectively aggregating these noisy information signals. The flexible and iterative nature of machine learning algorithms allow the models to adapt as the market conditions and the behavior of market participants change.

Our key insight is that analyst-level individual errors are more predictable and can be corrected more effectively than the aggregate consensus forecast. Consistent with this expectation, we show that neural network models are able to learn the individual-level biases of sell-side equity analysts and can compensate for their biases more effectively. These models also allow us to aggregate the corrected forecasts more effectively.

By using the analyst-wise neural network model, we are able to improve earnings predictability by 24% in comparison to the linear regression model. Additionally, our analyst-wise neural network model is better at predicting the sign of the earnings surprise. Specifically, it makes the correct prediction 68% of the time, while the linear model is able to predict the sign correctly only 54% of the time. The relative difference in performance per-

sists also in out-of-sample estimation, where we use pre-2010 observations as training data, and 2010–2019 as an out-of-sample period. Lastly, our improved consensus forecasts earns a monthly abnormal return of 0.5% on a risk-adjusted basis.

In future work, it would be useful to examine whether a similar machine learning framework can be used to learn the biases of other market participants such as institutional investors. If the trading activities of institutional investors influence market prices and their biases can be predicted using various machine learning models, these models can be used as effective prediction and information aggregation tools.

References

Abarbanell, J. Y. S. and Bernard, V. L. (1992). Tests of analysts' overreaction/underreaction to earnings information as an explanation for anomalous stock price behavior. *Journal of Finance*, 47(3):1181–1207.

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press, Boston, MA, USA.

Butler, K. C. and Lang, L. H. P. (1991). The Forecast Accuracy of Individual Analysts: Evidence of Systematic Optimism and Pessimism. *Journal of Accounting Research*, 29(1):150–156.

Clement, M. B. and Tse, S. Y. (2005). Financial Analyst Characteristics and Herding Behavior in Forecasting. *Journal of Finance*, 60(1):307–341.

Das, S., Levine, C. B., and Sivaramakrishnan, K. (1998). Earnings predictability and bias in analysts' earnings forecasts. *The Accounting Review*, 73(2):277–294.

Davis, J. L., Fama, E. F., and French, K. R. (2000). Characteristics, covariances, and average returns: 1929 to 1997. *Journal of Finance*, 55(1):389–406.

Easterwood, J. C. and Nutt, S. R. (1999). Inefficiency in analysts' earnings forecasts: Systematic misreaction or systematic optimism? *Journal of Finance*, 54(5):1777–1797.

Gu, Z. and Wu, J. S. (2003). Earnings skewness and analyst forecast bias. *Journal of Accounting and Economics*, 35(1):5–29.

Hong, H. and Kubik, J. D. (2003). Analyzing the analysts: Career concerns and biased earnings forecasts. *The Journal of Finance*, 58(1):313–351.

Hughes, J., Liu, J., and Su, W. (2008). On the relation between predictable market returns and predictable analyst forecast errors. *Review of Accounting Studies*, 13:266–291.

Kumar, A., Rantala, V., and Xu, R. (2021). Social learning and analyst behavior. *Journal of Financial Economics*, forthcoming.

Lim, T. (2001). Rationality and analysts' forecast bias. *The Journal of Finance*, 56(1):369–385.

Lin, H.-W. and McNichols, M. F. (1998). Underwriting relationships, analysts' earnings forecasts and investment recommendations. *Journal of Accounting and Economics*, 25(1):101–127.

Linnainmaa, J. T., Walter, T., and James, Y. (2016). Reading the tea leaves: Model uncertainty, robust forecasts, and the autocorrelation of analysts' forecast errors. *Journal of Financial Economics*, 122(1):42–64.

Lys, T. and Sohn, S. (1990). The association between revisions of financial analysts' earnings forecasts and security-price changes. *Journal of Accounting and Economics*, 13(4):341–363.

Malmendier, U. and Shanthikumar, D. (2007). Are small investors naive about incentives? *Journal of Financial Economics*, 85(2):457–489.

Mendenhall, R. (1991). Evidence On The Possible Underweighting Of Earnings-Related Information. *Journal of Accounting Research*, 29(1):170–179.

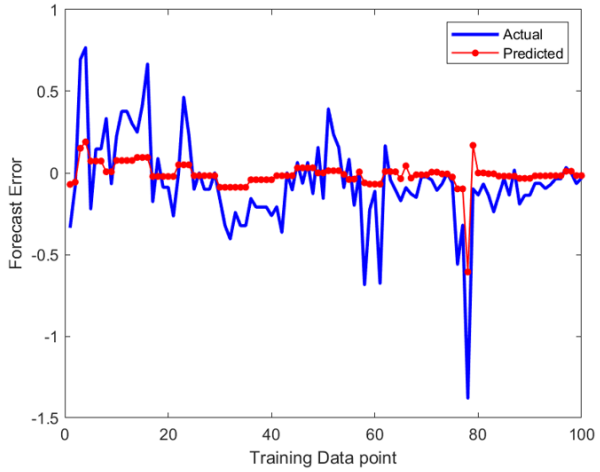
Mikhail, M. B., Walther, B. R., and Willis, R. H. (2007). When Security Analysts Talk, Who Listens? *The Accounting Review*, 82(5):1227–1253.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

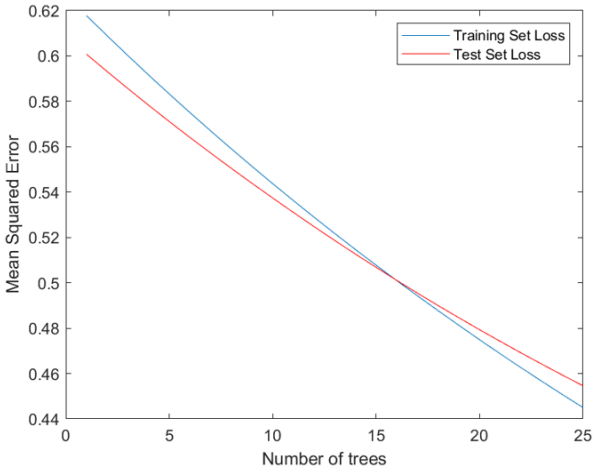
Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Rossi, A. G. and Utkus, S. P. (2020). Who benefits from robo-advising? evidence from machine learning. Available at SSRN: https://ssrn.com/abstract=3552671.

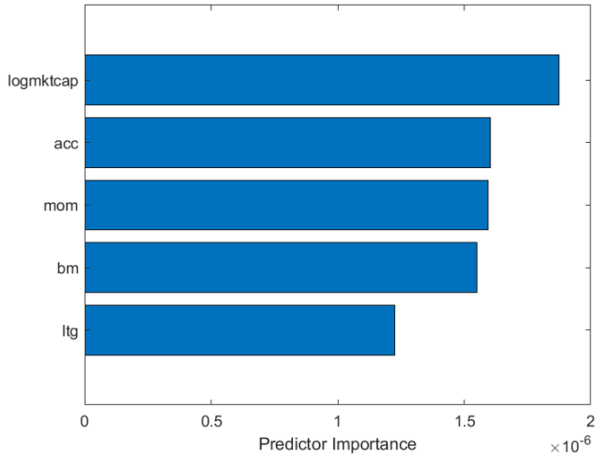
So, E. C. (2013). A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? *Journal of Financial Economics*, 108(3):615 – 640.



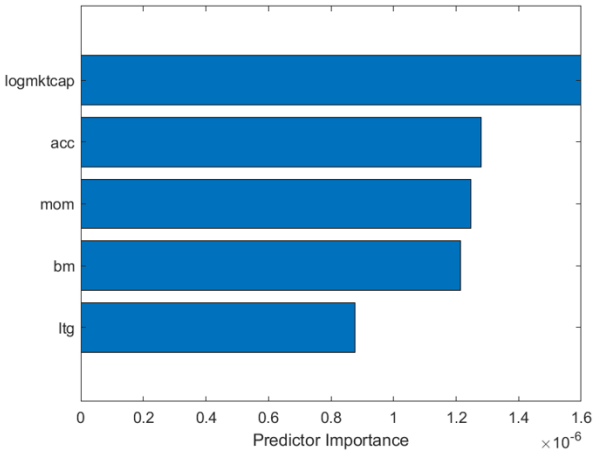
(a) Fit with 25 trees



(b) Fit: Mean Squared Error with 25 trees

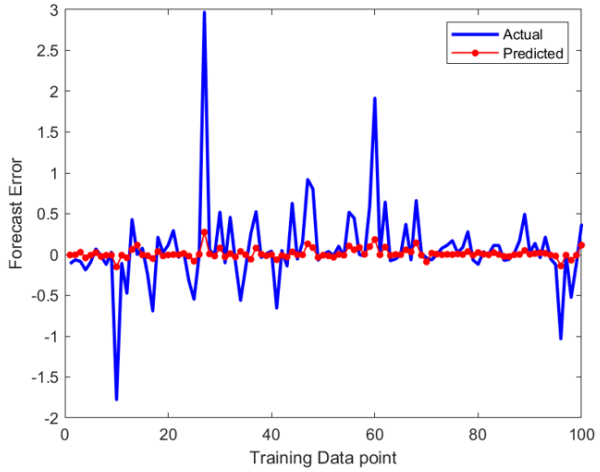


(c) Relative Importance of variables in Panel Model (25 trees)

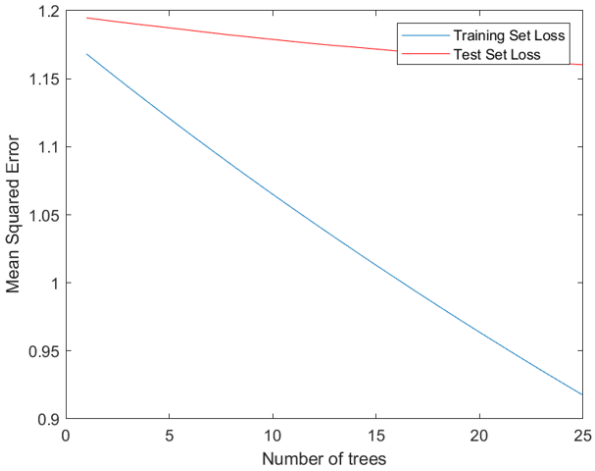


(d) Relative Importance of variables in Panel Model (50 trees)

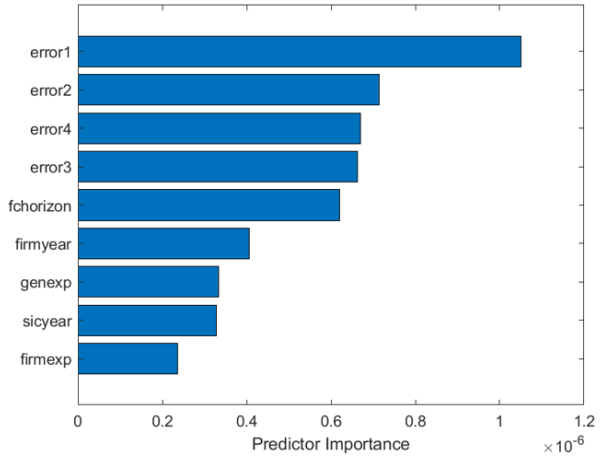
Figure 1: Boosted Regression Trees: Panel Model. The graphs present the results from a BRT with 10,000 boosted iterations. The dependent variable is forecast error.The independent variables includes log of market capitalization (log (mktcap)), book-to-market(bm), accruals (acc), momentum (mom) and long term earnings growth consensus (ltg) as explanatory variables. Figure (a) presents the fit of the panel model with 25 trees. Figure (b) presents the mean squared errors with 25 treees. Figure (c) and (d) present relative importance of all independent variables with 25 and 50 trees respectively.



(a) Fit with 25 trees

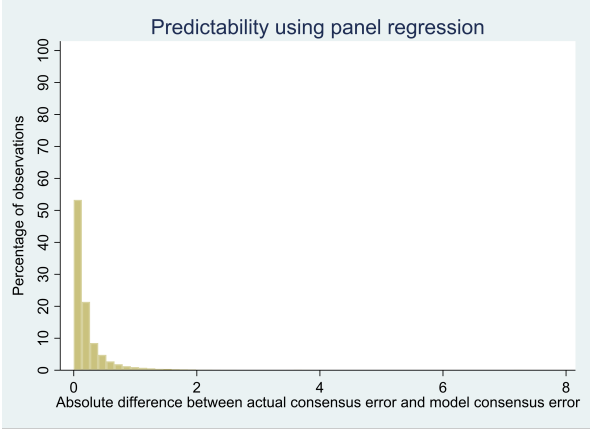


(b) Fit: Mean Squared Error with 25 trees

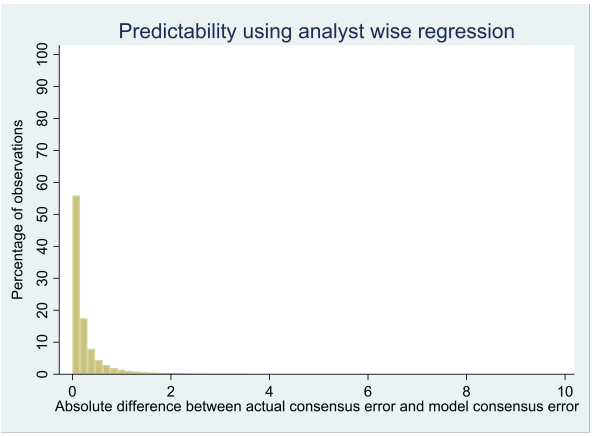


(c) Relative Importance of variables in Panel Model (25 trees)

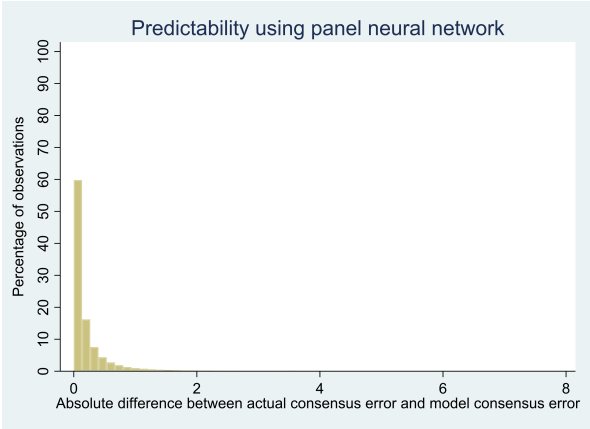
Figure 2: Boosted Regression Trees : Analyst-wise Model. The graphs present the results from a BRT with 10,000 boosted iterations. The dependent variable is forecast error.The independent variables includes upto 4 lags of forecast error, forecast horizon (fchorizon), number of firm that an analyst covers in a year (firmyear), number of years the analyst has provided EPS forecasts (genexp), number of unique 2 digit SIC industry codes for which the analyst has made EPS forecasts (sicyear) and number of prior years for which the analyst has provided EPS forecasts for the firm (firmexp) as explanatory variables. Figure (a) presents the fit of the panel model with 25 trees. Figure (b) presents the mean squared errors with 25 treees. Figure (c) presents relative importance of all independent variables with 25 trees.



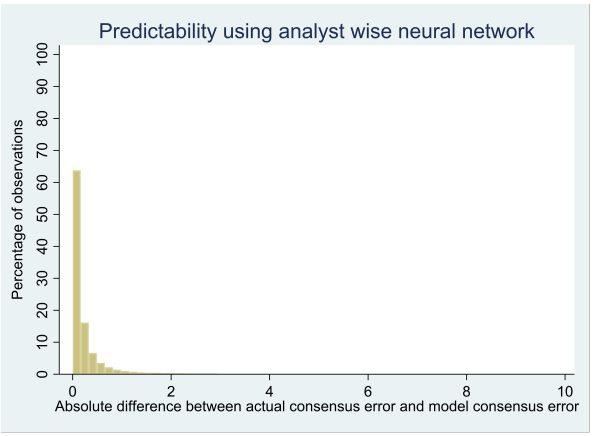
(a) Linear Panel Model



(b) Linear Analyst-wise Model

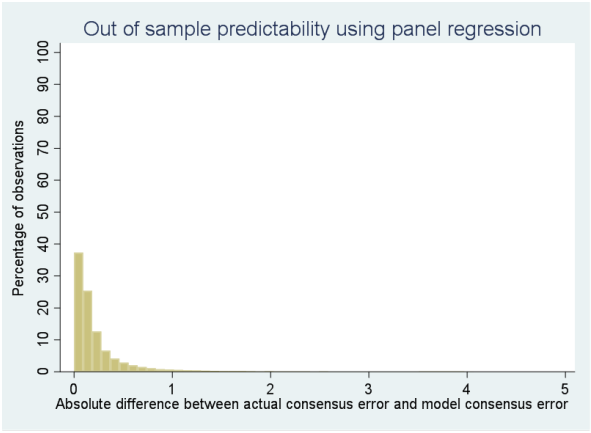


(c) Panel Neural Network

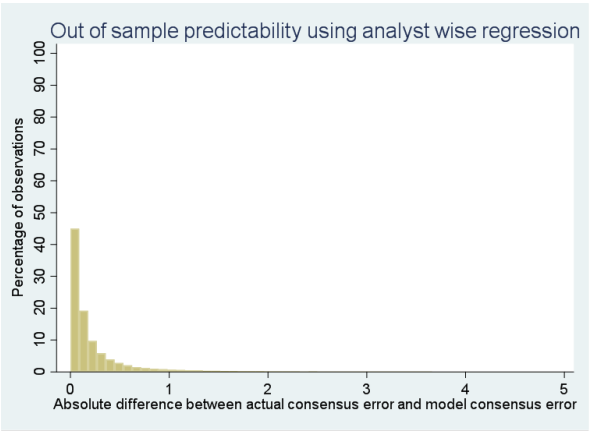


(d) Analyst-wise Neural Network

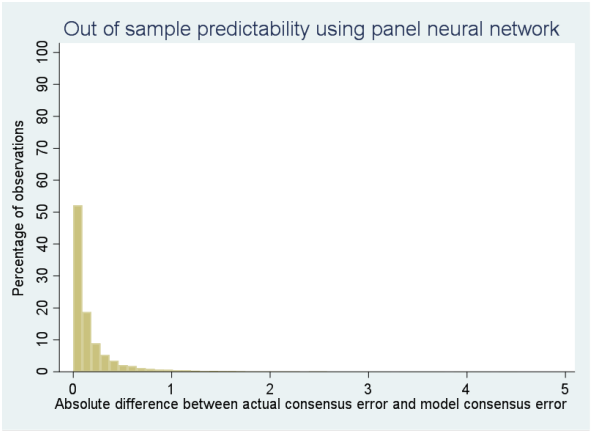
Figure 3: Predictability of absolute difference in earnings consensus error and model consensus error. The graphs present histograms for the absolute difference between the actual consensus error and and the model consensus error where we use linear panel and analyst-wise models and panel and analyst-wise neural network models. The 4 different models are presented as follows: (a) Linear Panel Model, (b) Linear Analyst-wise Model, (c) Panel Neural Network, and (d) Analyst-wise Neural Network.



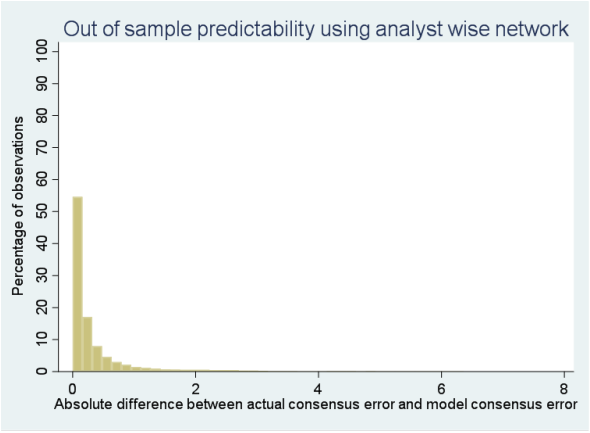
(a) Linear Panel Model



(b) Linear Analyst-wise Model



(c) Panel Neural Network



(d) Analyst-wise Neural Network

Figure 4: Out of Sample Predictability of absolute difference in earnings consensus error and model consensus error . The graphs present histograms of out of sample predictability of the absolute difference in earnings consensus error and model consensus error. We use linear panel and analyst-wise models and panel and analyst-wise neural network models. The model is trained on the data before 2010 and the out of sample test period is after 2010. The 4 different models are presented as follows; (a) Linear Panel Model, (b) Linear Analyst-wise Model, (c) Panel Neural Network, and (d) Analyst-wise Neural Network.

Table 1: Summary Statistics

This table provides summary statistics on analysts' quarterly earnings forecasts and forecast errors. Forecast error is calculated as the difference between the earnings per share forecast and the actual earnings per share in I/B/E/S. We divide this difference with the share price of the firm ten trading days prior to the announcement. Consensus Forecast is the median forecast. Statistics on consensus forecasts are based on earnings announcement observations and the other statistics are based on earnings forecast observations. N of Analysts in the Consensus reports the number of analysts covering the earnings announcement in I/B/E/S.

	Mean	St. Dev.	p5	p50	p95	N
Forecast Error	0.056	1.140	-0.944	-0.039	1.273	2,199,096
Absolute Forecast Error	0.472	1.039	0.000	0.140	2.136	2,199,096
Forecast Minus Consensus	-0.086	0.560	-0.854	0.000	0.407	2,015,638
Absolute Deviation from the Consensus	0.247	0.510	0.000	0.069	1.198	2,015,638
Consensus Forecast Error	0.158	1.440	-1.217	-0.030	2.270	386,169
Absolute Consensus Forecast Error	0.643	1.298	0.000	0.178	3.462	386,169
N of Analysts in the Consensus	5.695	5.426	1.000	4.000	17.000	386,169

Table 2: Statistics on Firm, Analyst, and Forecast Characteristics Variables

This table provides statistics on variables used in the linear benchmark models. Panel A reports on firm characteristics that are used in the panel model. Market capitalization is calculated as share price (CRSP Item PRC) ten days before the earnings announcement multiplied with shares outstanding (Item SHROUT). Book-to-market is book equity divided market equity. Momentum is calculated as the return 6 months prior to the earnings announcement data less the SP 500 returns. Accruals is measured as total accruals scaled by total assets and long term growth forecast is the long term consensus from I/B/E/S. Panel B reports on forecast errors and analyst characteristics that are used in the analyst-wise models. All statistics are based on individual forecast observations.

Panel A: Firm Characteristics						
	N	Mean	Std. Dev.	p10	Median	p90
Log(Market Cap)	759,940	15.167	1.723	12.967	15.135	17.387
Book-to-Market	759,940	0.501	0.375	0.140	0.406	0.968
Momentum	759,940	0.003	0.125	-0.123	-0.001	0.132
Accruals	759,940	0.001	0.035	-0.027	0.001	0.030
Long-Term Growh	759,940	15.161	44.234	4.000	13.500	30.000
Panel B: Analyst and Forecast Characteristics						
	N	Mean	Std. Dev.	p10	p50	p90
Forecast Error Lag 1	1,349,301	0.000	0.009	-0.005	0.000	0.004
Forecast Error Lag 2	1,349,301	0.000	0.008	-0.005	0.000	0.004
Forecast Error Lag 3	1,349,301	0.000	0.008	-0.005	0.000	0.004
Forecast Error Lag 4	1,349,301	0.000	0.008	-0.005	0.000	0.004
Forecast Horizon	1,349,301	40.157	25.890	7.000	45.000	69.000
Firms Covered Per Year	1,349,301	14.081	7.393	6.000	13.000	22.000
SIC Codes Covered Per Year	1,349,301	7.510	4.379	3.000	7.000	13.000
Firm Experience	1,349,301	4.212	3.403	1.000	3.000	9.000
General Experience	1,349,301	7.887	5.188	2.000	7.000	15.000

Table 3: Panel OLS Regressions Explaining Forecast Errors

This table reports results from linear panel regressions explaining forecast errors with firm characteristics. The dependent variable is forecast error calculated as the difference between the earnings per share forecast and the actual earnings per share in I/B/E/S. We divide this difference with the share price ten trading days prior to the announcement. The explanatory variables are defined as in Table 2. Standard errors are reported in parentheses below the coefficients. ***, **, and * denote significance levels at 1%, 5%, and 10% level, respectively.

	(1)	(2)	(3)	(4)	(5)
log(Market Cap)	-0.050***	-0.044***	-0.044***	-0.044***	-0.044***
	(0.000834)	(0.000761)	(0.000759)	(0.000759)	(0.000761)
Book-to-Market		0.099***	0.095***	0.092***	0.091***
		(0.00524)	(0.00524)	(0.00523)	(0.00524)
Momentum			-0.117***	-0.115***	-0.115***
			(0.0121)	(0.0121)	(0.0121)
Accruals				-1.259***	-1.259***
				(0.0523)	(0.0523)
Long-Term Growth Forecast					-0.000
					(0.0000)
Constant	0.748***	0.611***	0.611***	0.618***	0.619***
	(0.0133)	(0.0122)	(0.0122)	(0.0122)	(0.0123)
N	759,940	759,940	759,940	759,940	759,940
R-squared	0.012	0.014	0.014	0.017	0.017

Table 4: Performance of Prediction models with Individual Forecasts

This table provides statistics on the performance of linear and neural network models in predicting analysts' forecast errors. The four models the table compares are a linear regression model, a panel neural network model, an analyst-wise linear model, and an analyst-wise neural network model. The reported statistic is the absolute difference between the actual forecast error and the predicted forecast error. Panel A reports statistics based on the full sample and, Panel B reports statistics based on in-sample estimates using observations before 2010, and Panel C reports statistics using out-of-sample observations starting from 2010.

Panel A: Full Sample						
	N	Mean	Std. Dev.	p25	p50	p75
Panel Linear Model	759,940	0.386	0.706	0.064	0.166	0.401
Panel Neural Network Model	759,940	0.446	0.884	0.064	0.163	0.411
Analyst-Wise Linear Model	1,349,301	0.332	0.710	0.059	0.133	0.302
Analyst-Wise Neural Network Model	1,349,301	0.318	0.702	0.040	0.107	0.291
Panel B: In Sample (<2010)						
	N	Mean	Std. Dev.	p25	p50	p75
Panel Linear Model	474,406	0.370	0.761	0.068	0.155	0.343
Panel Neural Network Model	474,406	0.269	0.602	0.040	0.096	0.244
Analyst-Wise Linear Model	808,922	0.383	0.922	0.074	0.137	0.459
Analyst-Wise Neural Network Model	807,806	0.309	0.742	0.069	0.079	0.424
Panel C: Out of Sample (2010-2019)						
	N	Mean	Std. Dev.	p25	p50	p75
Panel Linear Model	210,103	0.386	0.706	0.064	0.166	0.401
Panel Neural Network Model	210,103	0.268	0.567	0.056	0.122	0.249
Analyst-Wise Linear Model	540,379	0.374	0.720	0.056	0.140	0.351
Analyst-Wise Neural Network Model	438,092	0.249	0.597	0.031	0.086	0.224

Table 5: Predicted Forecast Errors

This table provides statistics on the predicted forecast errors based on different prediction models. The four models the table compares are a linear regression model, a panel neural network model, an analyst-wise linear model, and an analyst-wise neural network model. Panels A and B report statistics on the predicted and actual forecast errors for panel models and analyst-wise models, respectively.

Panel A: Predicted Forecast Errors in Panel Models						
	N	Mean	Std. Dev.	p25	p50	p75
Actual Error	759,940	-0.003	0.791	-0.160	-0.041	0.039
Prediction Based on Linear Model	759,940	-0.003	0.103	-0.070	-0.007	0.057
Prediction Based on Neural Network	759,940	-0.002	0.165	-0.048	-0.039	-0.020
Panel B: Predicted Forecast Errors in Analyst-Wise Models						
	N	Mean	Std. Dev.	p25	p50	p75
Actual Error	1,349,301	0.033	1.087	-0.185	-0.042	0.057
Prediction Based on Linear Model	1,349,301	0.033	0.448	-0.103	-0.017	0.091
Prediction Based on Neural Network	1,349,301	0.029	0.605	-0.150	-0.024	0.135

Table 6: Predicting the Sign of the Forecast Error

This table reports how well different models are able to predict the sign of the forecast error based on the percentage of observations where the sign of the predicted forecast error is same as the sign of the actual forecast error. Column 1 reports statistics based on individual forecasts and column 2 based on consensus forecasts.

Percentage of Observations Predicting the Correct Sign for the Forecast Error		
	Individual Forecasts	Consensus Forecasts
	(1)	(2)
Panel Linear Model	51.01	46.42
Panel Neural Network Model	58.53	55.14
Analyst-Wise Linear Regression Model	54.35	54.62
Analyst-Wise Neural Network Model	63.95	67.95

Table 7: Performance of Prediction Models with Consensus Forecasts

This table provides statistics on the performance of linear and neural network models in predicting consensus forecast errors. The four models the table compares are a linear regression model, a panel neural network model, an analyst-wise linear model, and an analyst-wise neural network model. The predicted consensus forecast error for the models is defined as the median prediction for the earnings announcement. The reported statistic is the absolute difference between the actual consensus forecast error and the predicted consensus forecast error. Panel A reports statistics based on the full sample and, Panel B reports statistics based on in-sample estimates using observations before 2010, and Panel C reports statistics using out-of-sample observations starting from 2010.

Panel A: Full Sample						
	N	Mean	Std. Dev.	p25	p50	p75
Panel Linear Model	104,766	0.380	0.024	0.062	0.143	0.339
Panel Neural Network Model	104,766	0.373	0.016	0.042	0.120	0.348
Analyst-Wise Linear Model	283,836	0.386	0.018	0.047	0.124	0.330
Analyst-Wise Neural Network Model	283,836	0.295	0.014	0.038	0.099	0.258
Panel B: In Sample (<2010)						
	N	Mean	Std. Dev.	p25	p50	p75
Panel Linear Model	76,436	0.214	0.860	0.072	0.126	0.373
Panel Neural Network Model	76,436	0.285	0.633	0.040	0.097	0.258
Analyst-Wise Linear Model	203,166	0.319	0.882	0.055	0.102	0.366
Analyst-Wise Neural Network Model	203,090	0.322	0.690	0.043	0.061	0.285
Panel C: Out of Sample (2010-2019)						
	N	Mean	Std. Dev.	p25	p50	p75
Panel Linear Model	28,090	0.297	0.612	0.060	0.133	0.276
Panel Neural Network Model	28,090	0.280	0.642	0.036	0.098	0.253
Analyst-Wise Linear Model	80,670	0.338	0.749	0.042	0.107	0.289
Analyst-Wise Neural Network Model	77,892	0.259	0.636	0.024	0.063	0.170

Table 8: Predicted Consensus Forecast Errors

This table provides statistics on the predicted consensus forecast errors based on different prediction models. The four models the table compares are a linear regression model, a panel neural network model, an analyst-wise linear model, and an analyst-wise neural network model. The predicted consensus forecast error is defined as the median prediction for the earnings announcement. Panels A and B report statistics on the predicted and actual consensus forecast errors for panel models and analyst-wise models, respectively.

Panel A: Predicted Consensus Forecast Errors in Panel Models						
	N	Mean	Std. Dev.	p25	p50	p75
Actual Error	104,766	0.046	0.918	-0.150	-0.034	0.053
Prediction Based on Linear Model	104,766	0.037	0.109	-0.033	0.031	0.099
Prediction Based on Neural Network	104,766	0.051	0.267	-0.045	-0.032	0.031
Panel B: Predicted Consensus Forecast Errors in Analyst-Wise Models						
	N	Mean	Std. Dev.	p25	p50	p75
Actual Error	283,836	0.115	1.336	-0.195	-0.033	0.110
Prediction Based on Linear Model	283,836	0.084	0.503	-0.075	0.004	0.120
Prediction Based on Neural Network	283,836	0.075	0.644	-0.110	-0.009	0.135

Table 9: Portfolios Based on Artificially-Intelligent Earnings Consensus

This table reports returns on portfolio formed using artificially intelligent earnings consensuses. We sort all stocks with a scheduled quarterly earnings announcement during the next month on the predicted earnings surprise and allocate them to decile portfolios. The Short portfolio is a value-weighted portfolio of the decile of stocks predicted to have the largest negative earnings surprises and the Long portfolio is a value-weighted portfolio of the decile of stocks predicted to have the largest positive earnings surprises. Portfolios 2 to 9 are value-weighted portfolios of the remaining deciles of stocks sorted on the predicted earnings surprise. All stocks are held in their respective decile portfolios until the month before their next anticipated earnings announcement or for 3 months, whichever is less, and portfolios are rebalanced each month. We report average monthly returns and their standard deviations.

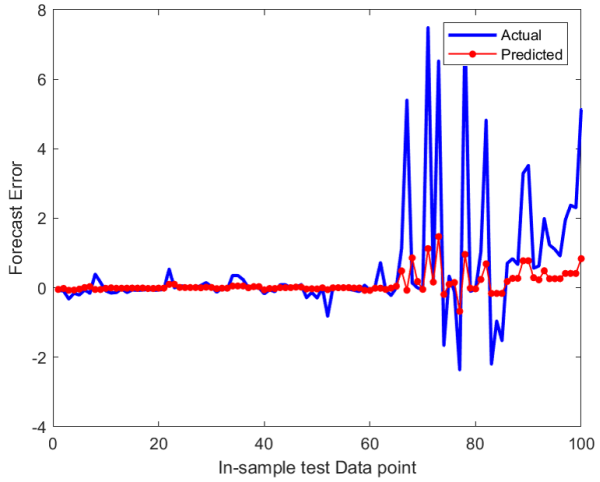
Portfolio	Monthly Return	Standard Deviation
Long (10)	1.1127	5.6681
9	1.1132	5.1254
8	1.0128	5.0011
7	1.0813	4.9019
6	1.0653	4.7688
5	1.0470	4.5320
4	1.0676	4.8528
3	0.8962	4.7527
2	0.8019	5.4795
Short (1)	0.5912	6.7497
Long-Short (10 - 1)	0.5355	5.6701
Number of Months	402	

Table 10: Factor Model Estimates

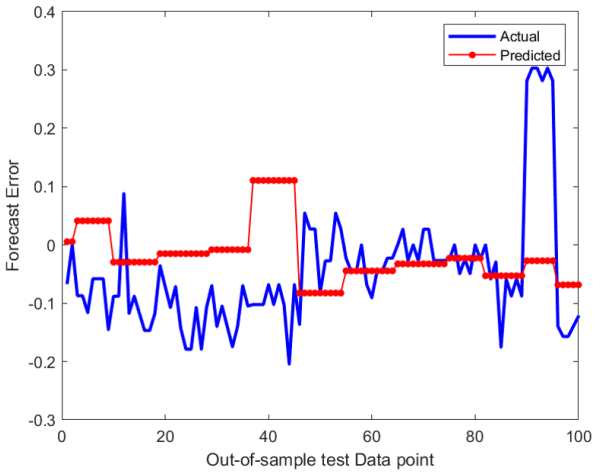
This table reports factor model risk-adjusted performance estimates for the Long, Short, and Long–Short portfolios. The "Long" portfolio is a value-weighted portfolio of the decile of stocks predicted to have the most positive positive earnings surprises based on our analyst-wise neural network model and the "Short" portfolio is a value-weighted portfolio of the decile of stocks predicted to have the most negative earnings surprises. The "Long–Short" portfolio captures the difference in the returns of the Long and Short portfolios. All stocks are held in their respective decile portfolios until the month before their next anticipated earnings announcement or for 3 months, whichever is less, and portfolios are rebalanced each month. The regressions contain the following factors: the market excess return (MKT), the size factor (SMB), and the value factor (HML). Newey-West (1987) adjusted standard errors are reported in parentheses below the estimates. ***, **, and * denote significance levels at 1%, 5%, and 10% level, respectively.

	Long - Short	Short	Long
	(1)	(2)	(3)
Alpha	0.00503**	-0.00998***	-0.00235
	(0.00254)	(0.00197)	(0.00150)
MKT	-0.230***	1.342***	1.114***
	-0.0756	(0.0620)	(0.0406)
SMB	0.207*	0.0565	0.256***
	(0.115)	(0.0686)	(0.0704)
HML	-0.171	0.458***	0.289***
	(0.135)	(0.0993)	(0.0725)
N	400	400	400
R-squared	0.068	0.736	0.776

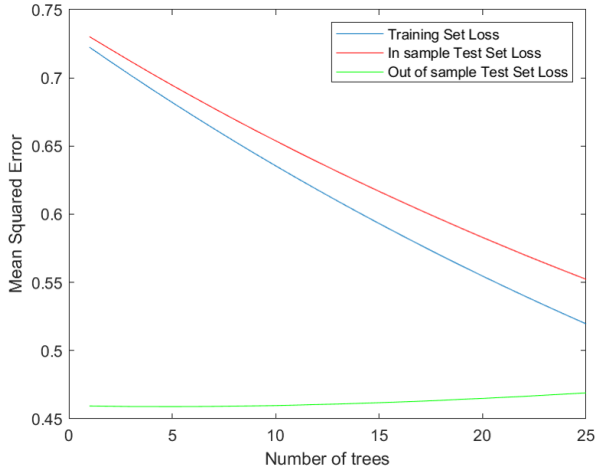
Appendix A.



(a) Fit with 25 trees in test period (< 2000)



(b) Fit with 25 trees in out of sample period (> 2000)



(c) Mean Squared Error in out of sample test (25 trees)

Figure A.1: Boosted Regression Trees: Panel Model. The graphs present the results from a BRT with 10,000 boosted iterations. The dependent variable is forecast error.The independent variables includes log of market capitalization (log (mktcap)), book-to-market(bm), accruals (acc), momentum (mom) and long term earnings growth consensus (ltg) as explanatory variables. Figure (a) presents the fit of the panel model with 25 trees in an a test period defined as the period before the year 2000. Figure (b) presents the out of sample test fit where the out of sample test period is after the year 2000. Figure (c) presents the out of sample mean squared errors.